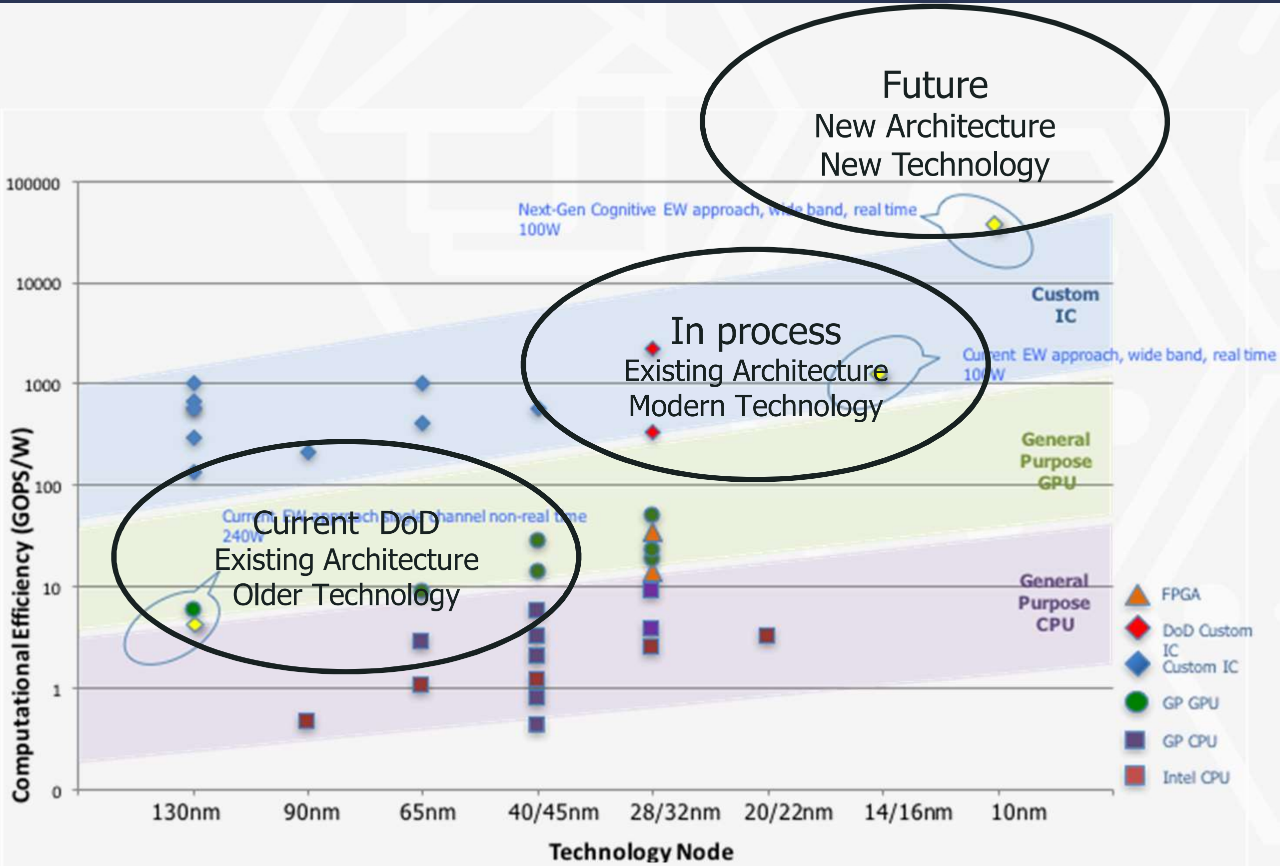
LINTON SALMON

PROGRAM MANAGER
DARPA/MTO

JOINT UNIVERSITY MICROELECTRONICS PROGRAM (JUMP)

DOD TECHNOLOGY NEEDS

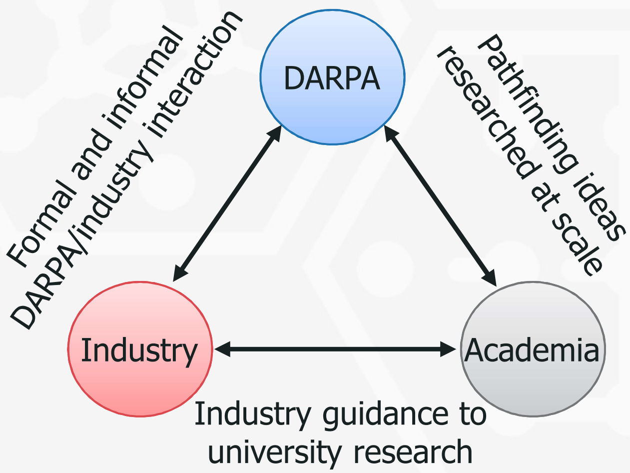


JOINT UNIVERSITY MICROELECTRONICS PROGRAM (JUMP)

JUMP Vision

To drive pathfinding research efforts at scale in new computing and communication technologies through enhanced DARPA-industry collaboration, funding, and guidance of university center research

- Objectives
  - Drive long-term research in microelectronics with key players in industry and from academia
  - Develop long-range ideas that will drive formation of new DARPA programs
- Program overview
  - 6 centers focused on 6 major long-range microelectronic research themes
  - DARPA + 11 industrial sponsors
  - $40M/year anticipated funding for 5 years ($24M/year: industry and $16M/year: DARPA)
  - DARPA program managers as a liaison for every center



OFFICIAL SPONSORS OF JUMP

Micron

DARPA

Intel

IBM

Lockheed Martin

Analog Devices

Raytheon

ARM

EMD Performance Materials

Northrop Grumman

TSMC

Samsung

STARNET AND JUMP EVOLUTION

FCRP | 2013 STARnet 2017 | 2018 JMUP 2022

STARnet Mission

The STARnet mission provides long-term breakthrough research that results in paradigm shifts and multiple technology options. STARnet is a U.S. based university research program that is guided strategically by industry and the U.S. government, but managed by the U.S. university community. It provides a multi-university, multi-disciplinary, collaborative research environment that is highly leveraged by both industry and U.S. Department of Defense funding. STARnet focuses on beyond CMOS technology options and systems integration and discovery to enable both CMOS and beyond CMOS components. The program also provides access to highly trained university graduate students.

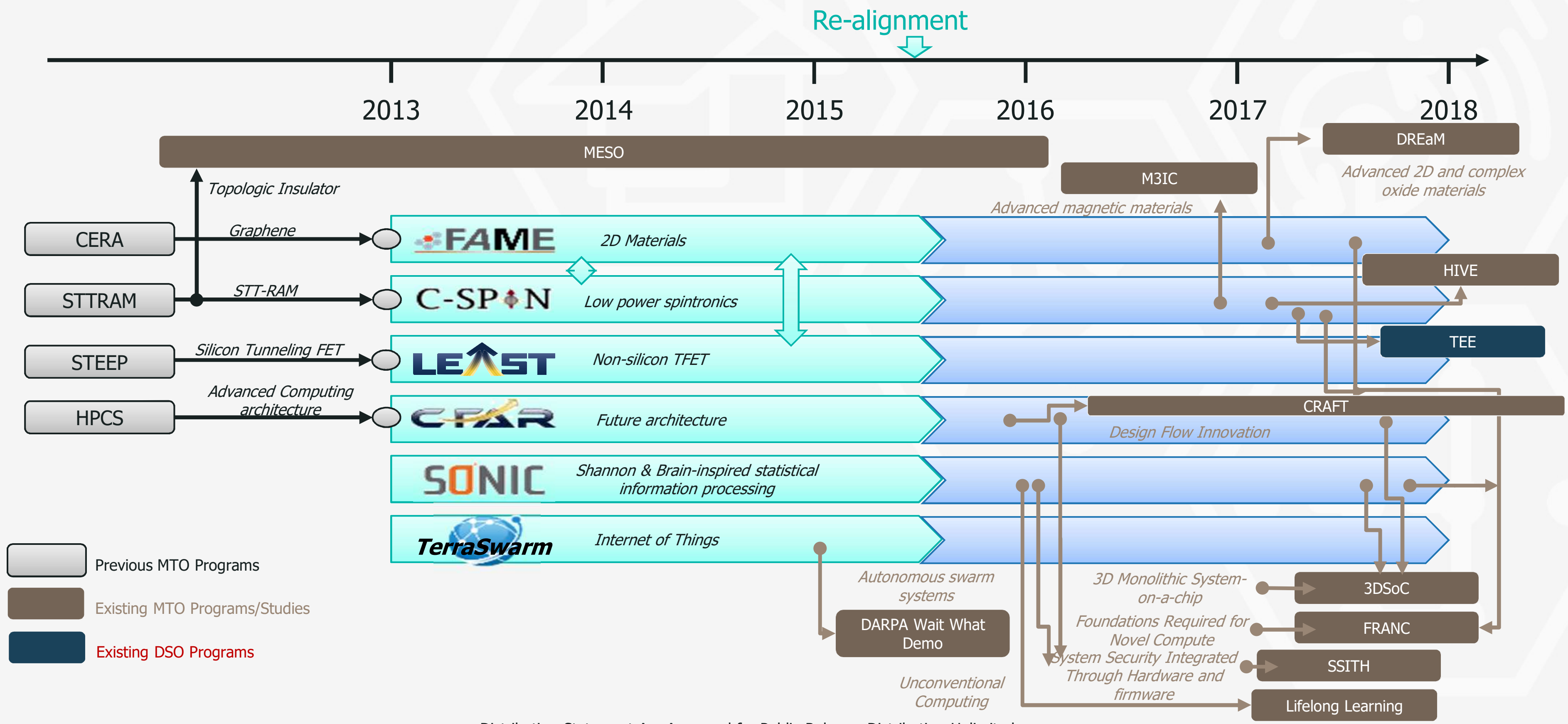
https://www.src.org/program/starnet/about/mission/

JUMP Mission

The mission of JUMP is to look beyond today's technology horizon and lay the scientific groundwork that extends the viability of Moore's Law economics through 2040. This program must create new general purpose architectures and system designs that relax device constraints and provide opportunities for new device types and novel, heterogeneous integration solutions. It must invent new devices and designs that are capable of the performance achievable today at a power consumption that is 1-3 orders of magnitude lower. Finally, it must train tomorrow's workforce to deliver "smart, autonomous, safe, connected, efficient, and affordable" electronics that meet our sensing, actuation, communication, computing, and storage needs for 2025 and beyond. In addition to providing enabling technologies, the research scope for each Center represents a critical component in the development of systems for both the semiconductor and defense industries and the Department of Defense.

https://www.src.org/program/jump/about/mission/

STARNET RESULTS FUELED DARPA PROGRAMS



LAUNCH OF JUMP IN JANUARY 2018

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY | ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US / EXPLORE BY TAG

Defense Advanced Research Projects Agency › News And Events

U.S. Electronics Innovation Leaps Forward Via Joint University Microelectronics Program

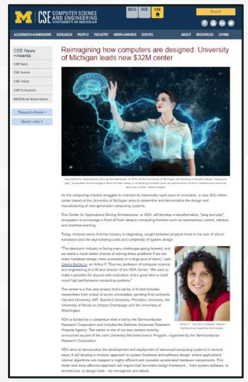
Under the five-year, up to $200 million, cost-shared program with industry partners, six university research centers are poised to JUMP start the path for the microsystems of 2025 and beyond

OUTREACH@DARPA.MIL
1/17/2018

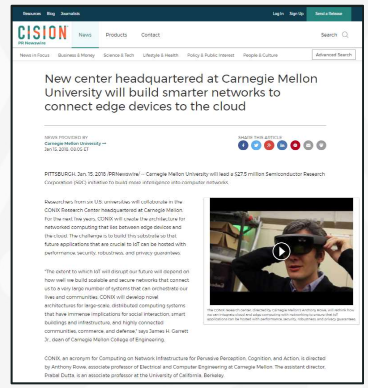
DARPA is all about developing advanced technologies that could underpin decisive national security capabilities in the years to come. A typical sequence that leads to new technology starts with fundamental science and engineering research, which, in turn, opens new pathways toward greatly improved technology by way of applied and goal-directed engineering and product development. In a bid to power up the front end of this sequence in the vast and complex area of microelectronics, DARPA, and a consortium of industry partners in the Joint University Microelectronics Program (JUMP), have completed the search for U.S. university collaborators to undertake high-risk, high-payoff research that addresses existing and emerging challenges in microelectronic technologies. As of January 1, six JUMP research centers comprised of academic researchers from over 30 U.S. universities began exploratory research initiatives that JUMP organizers hope will impact defense and commercial opportunities in the coming decades.

"The point of JUMP and its six thematic centers is to drive a new wave of fundamental research with the potential to deliver the disruptive microelectronics-based technologies required by the Department of Defense and national security in the 2025-2030 timeframe," said Linton Salmon, DARPA's program manager for JUMP. "Through these university teams, we're seeking innovative solutions to tough technical challenges so that we can overcome today's limitations in the performance and scalability of electronic systems. This in turn will open the way to technologies that dramatically boost the warfighter's abilities to sense the environment, process information, and communicate."

With initial efforts starting in 2016, DARPA, in collaboration with the non-profit Semiconductor Research Corporation (SRC), recruited a consortium of cost-sharing industry partners—among them Analog Devices, ARM, EMD Performance Materials (a Merck KGaA affiliate), IBM Corporation, Intel Corporation, Lockheed Martin Corporation, Micron Technology, Inc., Northrop Grumman Corporation, Raytheon Company, TSMC, and Samsung—forming the foundation of JUMP. The consortium, for which SRC serves as the administrative hub, conducted a search for university research proposals throughout 2017 with the goal of uncovering innovative approaches to solving tough development challenges around microelectronics. Funding for the five-year effort is expected to total approximately $200 million, with DARPA providing about 40 percent of the funding and consortium partners collectively kicking in about 60 percent.



Reimagining how computers are designed: University of Michigan leads new $32M center



CISION

New center headquartered at Carnegie Mellon University will build smarter networks to connect edge devices to the cloud

UVA Engineering Tapped to Lead $27.5 Million Center to Reinvent Computing

The investment will fund research to eliminate barriers to mining "big data" for breakthroughs in medicine, national security and many other areas.

The Current SCIENCE + TECHNOLOGY

Beyond 5G

UCSB is the lead institution for a research center that will explore terahertz-range communications and sensing

PURDUE UNIVERSITY News

January 16, 2018

New C-BRIC center will tackle brain-inspired computing

WEST LAFAYETTE, Ind. – Purdue University will lead a new national center to develop brain-inspired computing for intelligent autonomous systems such as drones and personal robots capable of operating without human intervention.

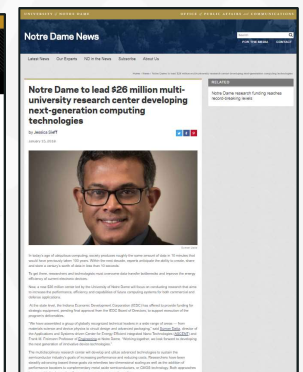
The Center for Brain-inspired Computing Enabling Autonomous Intelligence, or C-BRIC, is a five-year project supported by $27 million in funding from the Semiconductor Research Corp. (SRC) via its Joint University Microelectronics Program, which provides funding from a consortium of industrial sponsors as well as from the Defense Advanced Research Projects Agency. The SRC operates research programs in the United States and globally that connect industry to university researchers, deliver early results to enable technological advances, and prepare a highly-trained workforce for the semiconductor industry. Additional funds include $3.96 million from Purdue and support from participating universities, pending final contracts, which include: Arizona State University, Georgia Institute of Technology, Pennsylvania State University, Portland State University, Princeton University, University of Pennsylvania, and the University of Southern California. At the state level, the Indiana Economic Development Corporation will be providing funds, pending board approval, to establish an intelligent autonomous systems laboratory at Purdue.

Kaushik Roy, Purdue's Edward G. Tiedemann Jr. Distinguished Professor of Electrical and Computer Engineering, is director of the C-BRIC. (Purdue University image/ Rebecca Wilcox) Download image

Research News

- Study: Hormone keys plant growth or stress tolerance, but not both
- Fast-moving electrons create current in organic solar cells
- Babies stir up clouds of bio-gunk when they crawl
- Study uncovers distinctions in major crop genome evolutions
- 'Holostream' allows high-quality wireless 3-D video communications

More Research News



Notre Dame News

Notre Dame to lead $26 million multi-university research center developing next-generation computing technologies

JUMP CENTERS

Systems/Applications

ComSenTer
RF to THz Sensors and Communications
UC/Santa Barbara
Mark Rodwell, Dir.
Ali Niknejad, Asst. Dir.

CRISP
Intelligent Memory and Storage
U/Virginia
Kevin Skadron, Dir.
Yuan Xie, Asst. Dir.

CONIX
Distributed Computing and Networking
CMU
Anthony Rowe, Dir.
Prabal Dutta, Asst. Dir.

C-BRIC
Cognitive Computing
Purdue
Kaushik Roy, Dir.
Anand Raghunathan, Asst. Dir.

Core Technologies

ADA
Advanced Architecture and Algorithms
U/Michigan
Valeria Bertacco, Dir.
David Brooks, Asst. Dir.

ASCENT
Advanced Devices, Packaging, and Materials
Notre Dame
Suman Datta, Dir.
Sayeef Salahuddin, Asst. Dir.

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

ANTHONY ROWE

CARNEGIE MELLON UNIVERSITY

THE CONIX RESEARCH CENTER

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

CONIX RESEARCH CENTER

Computing on Network Infrastructure for Pervasive Perception, Cognition, and Action

Jeff Bilmes
Ras Bodik
David Culler
Prabal Dutta (Co-Director)
Ramesh Govindan
Rajesh Gupta
Chris Harrison
James Hoe
Hao Li
Brandon Lucia
Bryan Parno
Jan Rabaey
Anthony Rowe (Director)
Vyas Sekar
Mani Srivastava
Deian Stefan
Paulo Tabuada
Claire Tomlin
John Wawrzynek

Carnegie Mellon University
University of California, Berkeley
University of California, Los Angeles
University of California, San Diego
University of Southern California
University of Washington

A NEW *COMPUTING* TIER IS EMERGING

Cloud Computing

Internet

Edge Devices

Mobile Devices: the edge morphs

Home routers: dumb becomes smart

Droids and drones: mobility helps connectivity

Autonomous Cars

Drones

Smart Cities

Smarter Grid

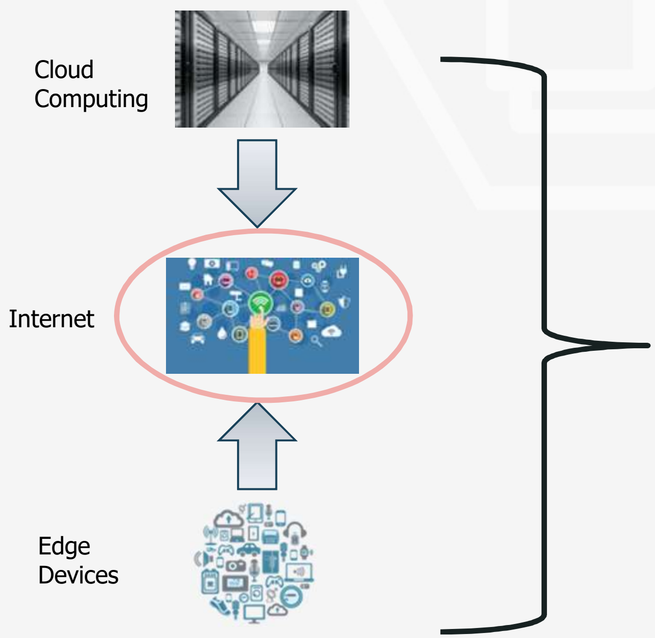
Mixed Reality

Approved for Public Release, Distribution Unlimited

THE NEW "NETWORK INFRASTRUCTURE" WILL BE CHARACTERIZED BY

- Extreme scale
- Unstable dynamics
- Variability and heterogeneity
- Time and location awareness
- Low-latency communications
- Intermittent resource availability
- Fragility to attacks

OUR JOB: BRING STRUCTURE TO CHAOS



"A Run to The Middle"

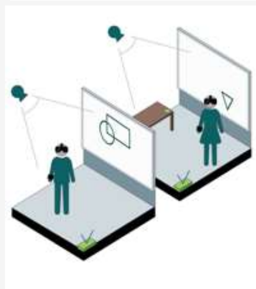
WHY COMPUTE ON THE NETWORK?

- **Network is central to all distributed applications**
  - Naturally tolerant to single point failures
- **Established ecosystems are hard to change**
  - Network is the natural intercept point for complex systems
- **Casts the widest net for security and privacy**
  - Access to per application data and control information
- **Captures spatial and temporal locality**
  - Latency sensitive applications
  - Intermittent connectivity

CONIX TECHNICAL PILLARS

Smart Cities

Mixed Reality



Autonomous Systems

- Machine learning *for* resilience
- Safe, secure, smart, and scalable programming
- Spatio-temporal awareness
- In-network coordination and control
- Cognitive architectures and accelerators for the edge

19 PIs across 6 Universities

UC Berkeley
- Prabal Dutta — CO-DIRECTOR
- David Culler
- Jan Rabaey
- Claire Tomlin
- John Wawrzynek

University of Washington
- Jeff Bilmes
- Ras Bodik

Carnegie Mellon University
- Anthony Rowe — DIRECTOR
- James Hoe
- Chris Harrison
- Vyas Sekar
- Brandon Lucia
- Bryan Parno

UC Los Angeles
- Mani Srivastava
- Paulo Tabuada

University of Southern California
- Ramesh Govindan
- Hao Li

UC San Diego
- Delan Stefan
- Rajesh Gupta

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

VALERIA BERTACCO

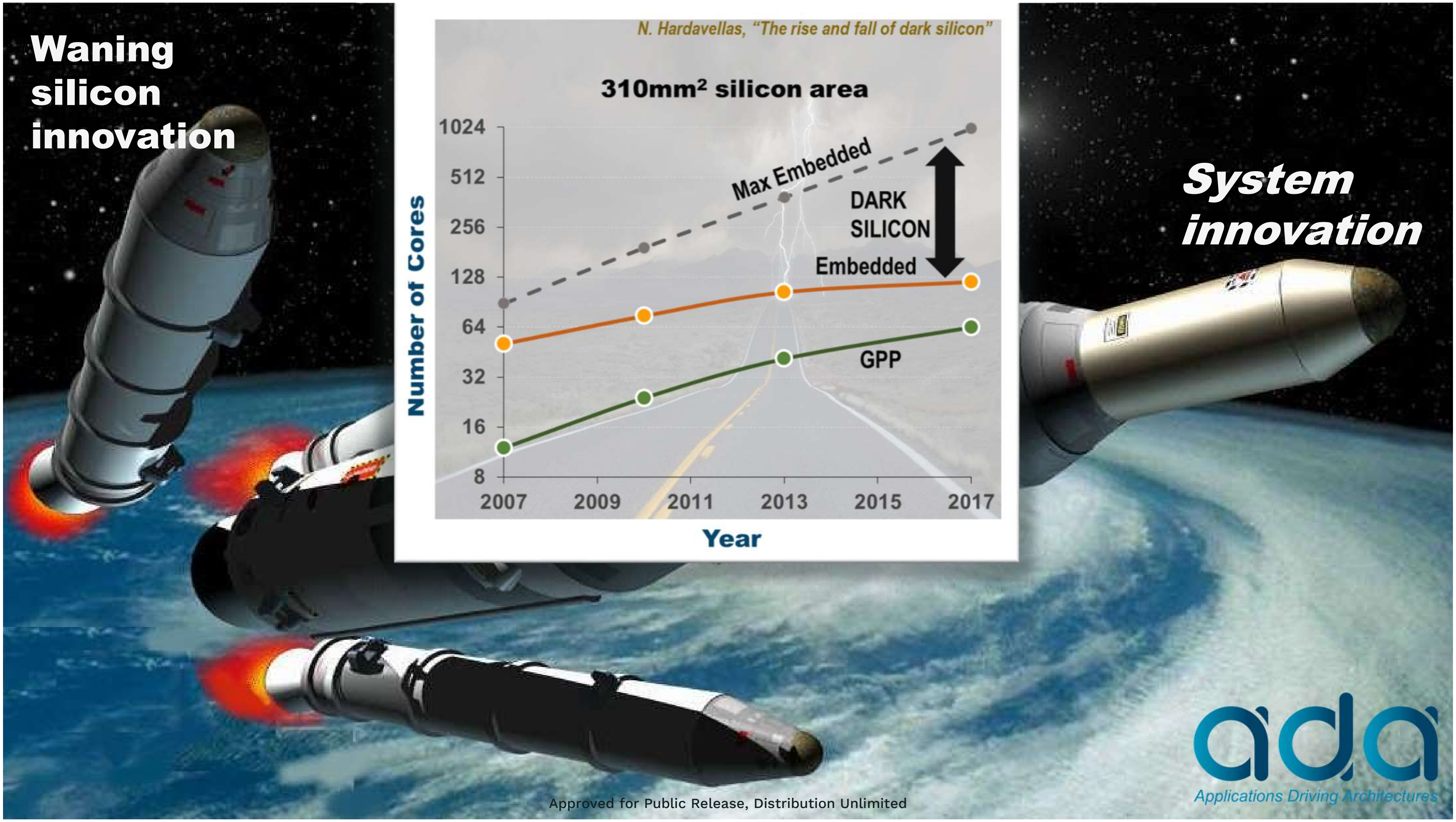
UNIVERSITY OF MICHIGAN

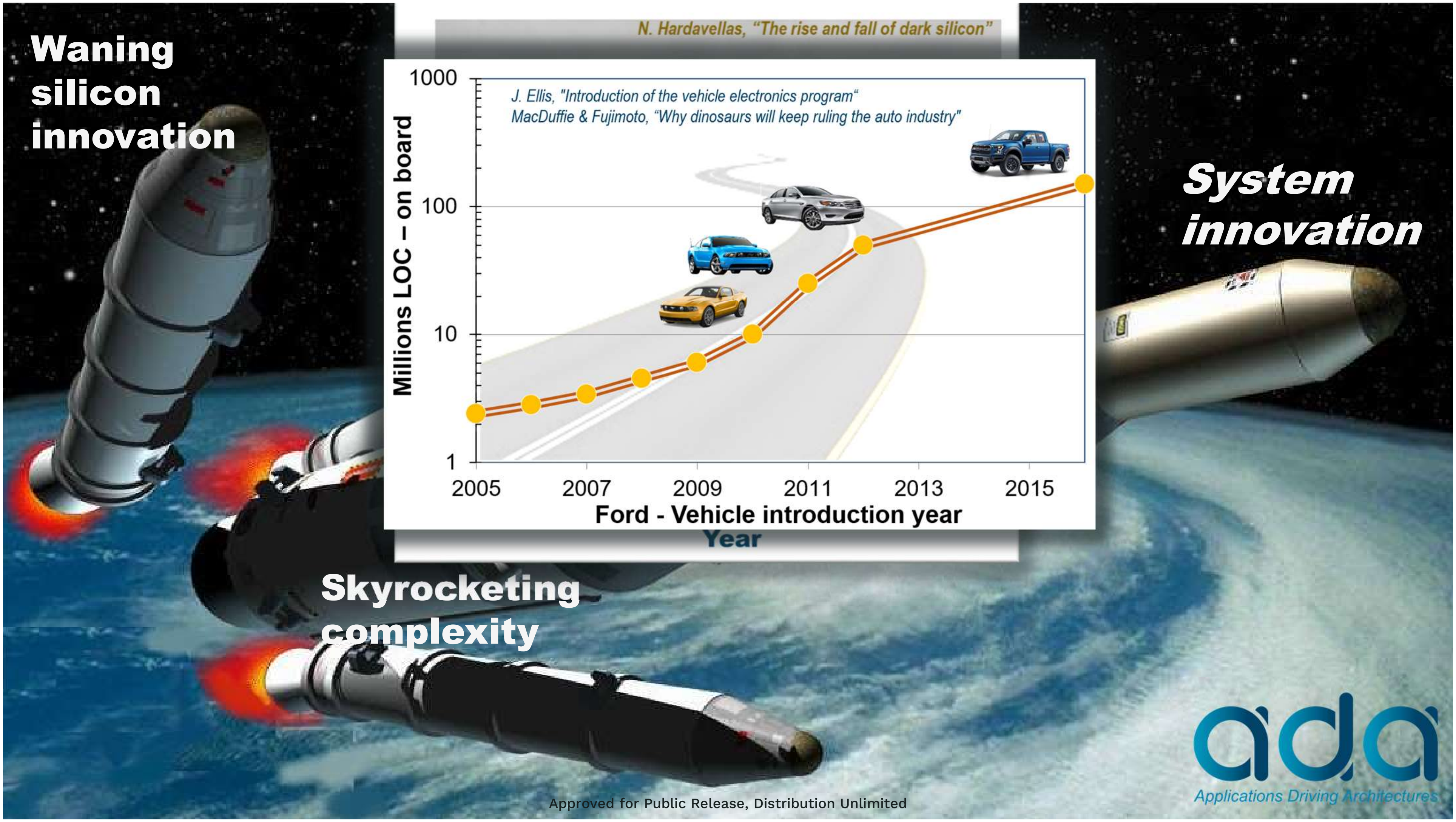
ADA

Applications Driving Architectures

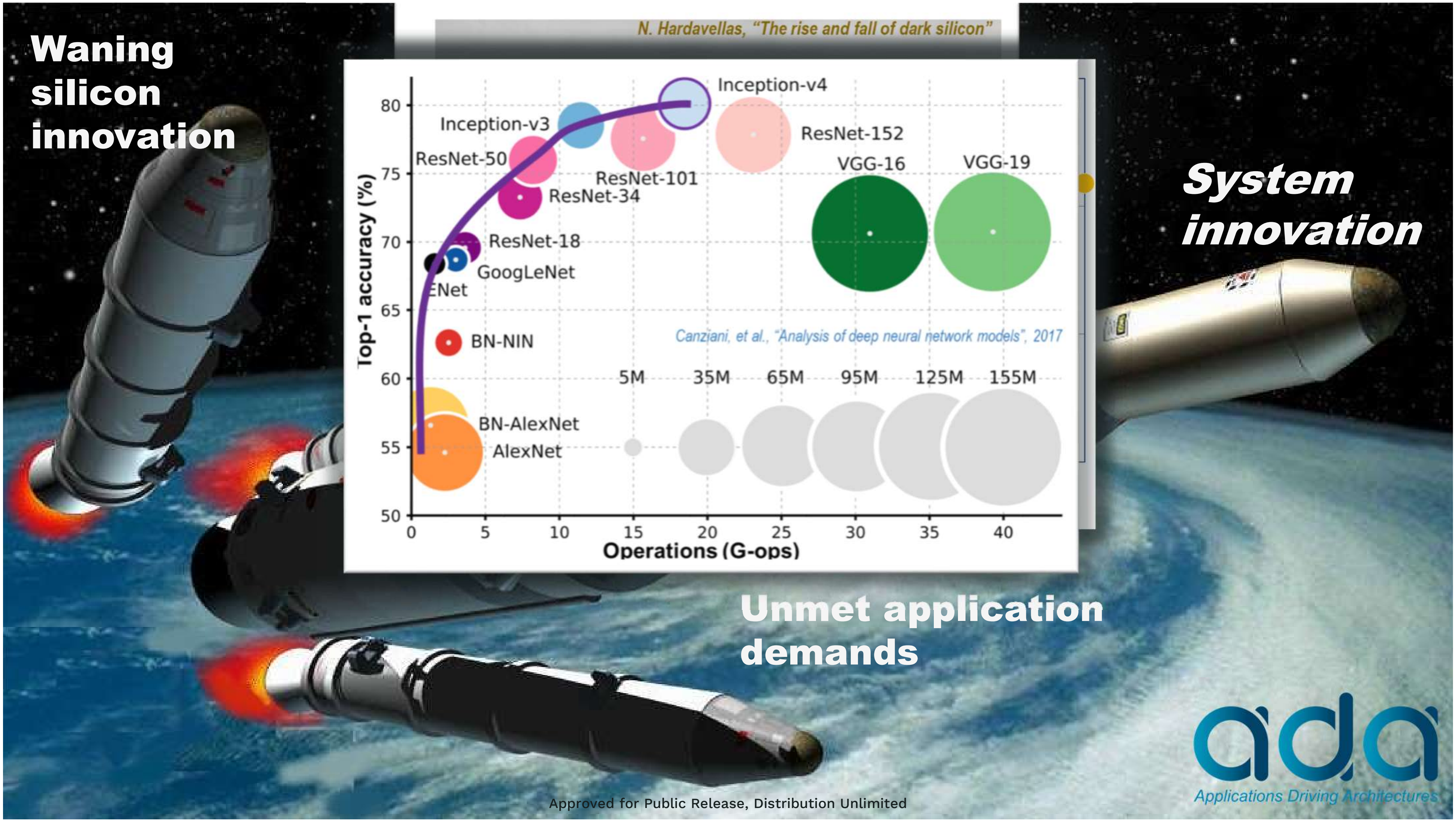
This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Images may come from public domains which are used for research purposes with limitations of Section 107 of the CopyrightAct.

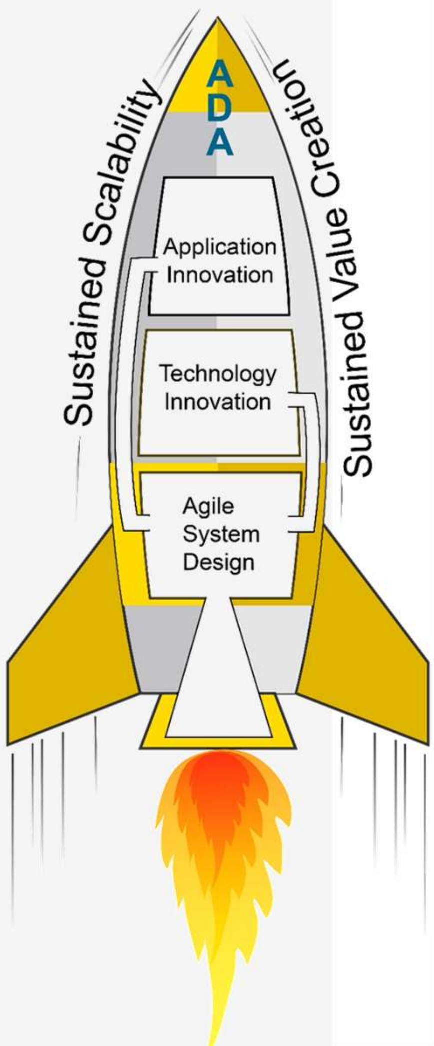

System
innovation
ada
Applications Driving Architectures


Waning silicon innovation
N. Hardavellas, "The rise and fall of dark silicon"
310mm² silicon area
Number of Cores
1024
512
256
128
64
32
16
8
Max Embedded
DARK SILICON
Embedded
GPP
2007
2009
2011
2013
2015
2017
Year
System innovation
ada
Applications Driving Architectures



Waning silicon innovation
N. Hardavellas, "The rise and fall of dark silicon"
J. Ellis, "Introduction of the vehicle electronics program"
MacDuffie & Fujimoto, "Why dinosaurs will keep ruling the auto industry"
1000
100
10
1
Millions LOC – on board
2005
2007
2009
2011
2013
2015
Ford - Vehicle introduction year
Year
System innovation
Skyrocketing complexity
ada
Applications Driving Architectures
Approved for Public Release, Distribution Unlimited


Waning silicon innovation
N. Hardavellas, "The rise and fall of dark silicon"
Inception-v4
Inception-v3
ResNet-152
ResNet-50
ResNet-101
ResNet-34
VGG-16
VGG-19
ResNet-18
GoogLeNet
ENet
BN-NIN
Canziani, et al., "Analysis of deep neural network models", 2017
5M
35M
65M
95M
125M
155M
BN-AlexNet
AlexNet
Top-1 accuracy (%)
Operations (G-ops)
System innovation
Unmet application demands
Approved for Public Release, Distribution Unlimited
ada
Applications Driving Architectures



Applications **D**riving **A**rchitectures

Reigniting system design innovation *by:*

1) *Identifying* ***new sources of application and technology innovation***
2) *Accelerating the adoption of these new solutions with uniquely* ***agile system development frameworks****.*

Approved for Public Release, Distribution Unlimited

ada
Applications Driving Architectures

APPLICATION-DRIVEN INNOVATION: APPLICATION-INSPIRED ARCHITECTURES

GOAL: Create components that slash cost and time-to-market for future designs

KEY ENABLING TECHNOLOGIES:

- Algorithm-inspired, eco-system of reusable accelerators
- Infrastructure to enable flexible, application-enhancing on-chip communication
- Supporting composable integration in the design framework to serve scales from edge to cloud

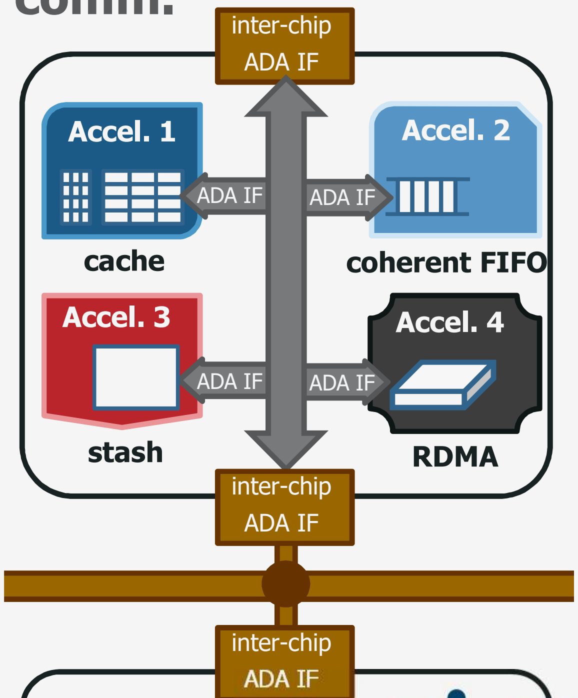
The Coliseum: an application-inspired architecture

ada
Applications Driving Architectures

APPLICATION-CUSTOMIZED ACCELERATOR COMMUNICATION ARCHITECTURE

Problem: integration of accels. + memory systems + comm.

- Friction among app-specific specializations
- Inefficiencies due to deep memory hierarchy
- Multiple scales: on-chip to cloud
- Exploiting solution quality tradeoffs



Goal: New accelerator communication architecture

- Standardized ADA-comm. interface (**ADA IF**)
- Accelerators that implement ADA IF
- Compile time, runtime, synthesis tool chains

Planned approach:

- Leverage prior work: coherence (Spandex, DeNovo), consistency (DRF), parallel IR (HPVM)

PI: Sarita Adve

ada
Applications Driving Architectures

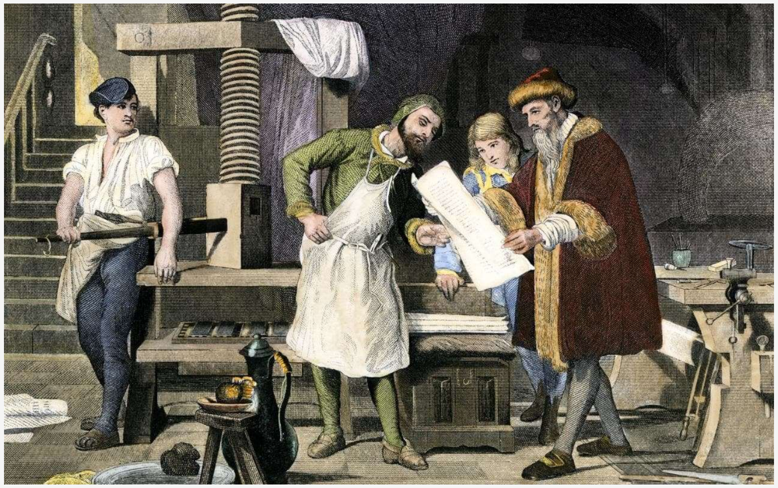
TECHNOLOGY INNOVATION

GOALS:

1) Exploit technology to reduce design cost
2) Direct silicon & non-silicon advances toward compelling system-level benefits

KEY ENABLING TECHNOLOGIES:

- Low-cost system-in-package solutions
- Integration of novel silicon technologies in the design flow (e.g., durable NVM writes)
- Expose technologies advances as first-class design capabilities for maximum ROI



The printing press: the most important technology innovation of the modern era

ada
Applications Driving Architectures

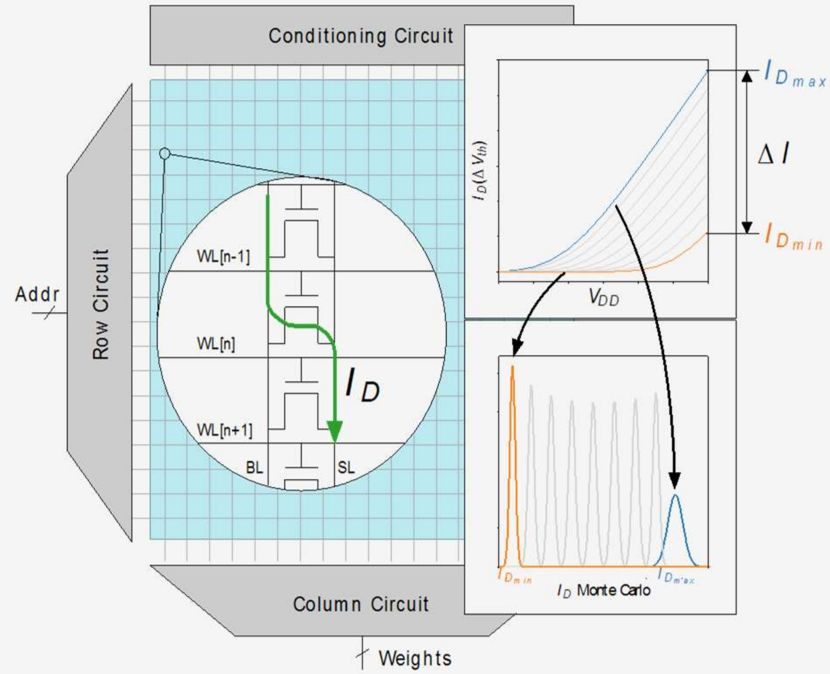
ENVM W/ DIGITAL CMOS FOR LONG-TERM DNN WEIGHT STORAGE

Problem: Emerging applications (e.g., neural network accelerators) are inefficiently
served by conventional memories

Goal: 10X density over 6T SRAM

Planned approach:

- multi-level embedded nonvolatile memories (eNVM) using vanilla CMOS devices via *aging*
- architectural applications: *weights* memory in artificial DNN accelerators, synapse in analog NN arrays (analog weights)



PIs: David Brooks and Gu-Yeon Wei

ada Applications Driving Architectures

AGILE SYSTEM DESIGN

GOALS: Reduce design effort – complexity, cost and time-to-market

KEY ENABLING TECHNOLOGIES:

- Language-level system specifications, amenable to SW developers
- Algorithmic application decomposition
- Language-level support for technology-level innovation
- Correctness through re-use and composability

Cheetah: nature's most agile design

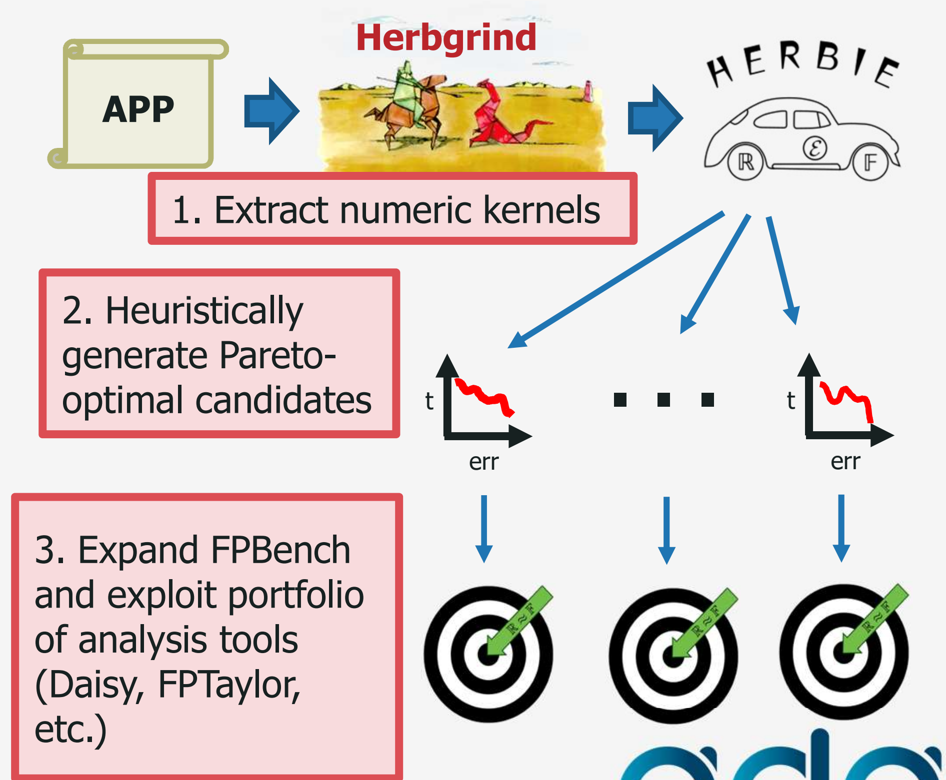
ada
Applications Driving Architectures

MODULAR NUMERIC ACCURACY / PERF TOOLS

Problem: Interesting modern applications depend on IEEE-754, but programmers think in real numbers;
→ hinders accuracy, debugging, optimization, accelerator-adaptivity.

Goal: Enable non-experts to effectively trade-off floating point performance (precision) vs. accuracy (fidelity to reals) for customization.

Approach: Leverage strengths of recent specialized tools by building compositional framework.



ada
Applications Driving Architectures

PI: Zachary Tatlock

Approved for Public Release, Distribution Unlimited

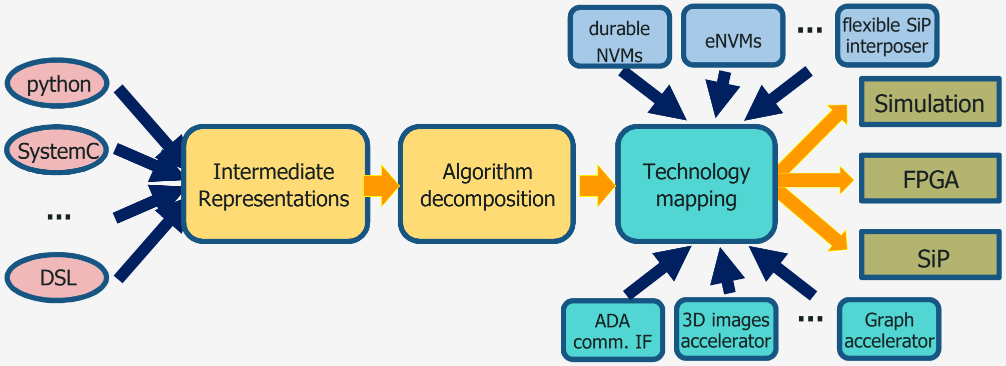
EVALUATION INFRASTRUCTURE

Initial driver applications:

- Natural language processing
- Visual computing
  (AR, VR, visual analytics,
  computational imaging, etc.)

Design flow framework:

center-wide collaborative effort to build a shareable infrastructure



ada
Applications Driving Architectures

THE ADA TEAM

ada

Applications Driving Architectures

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

SUMAN DATTA

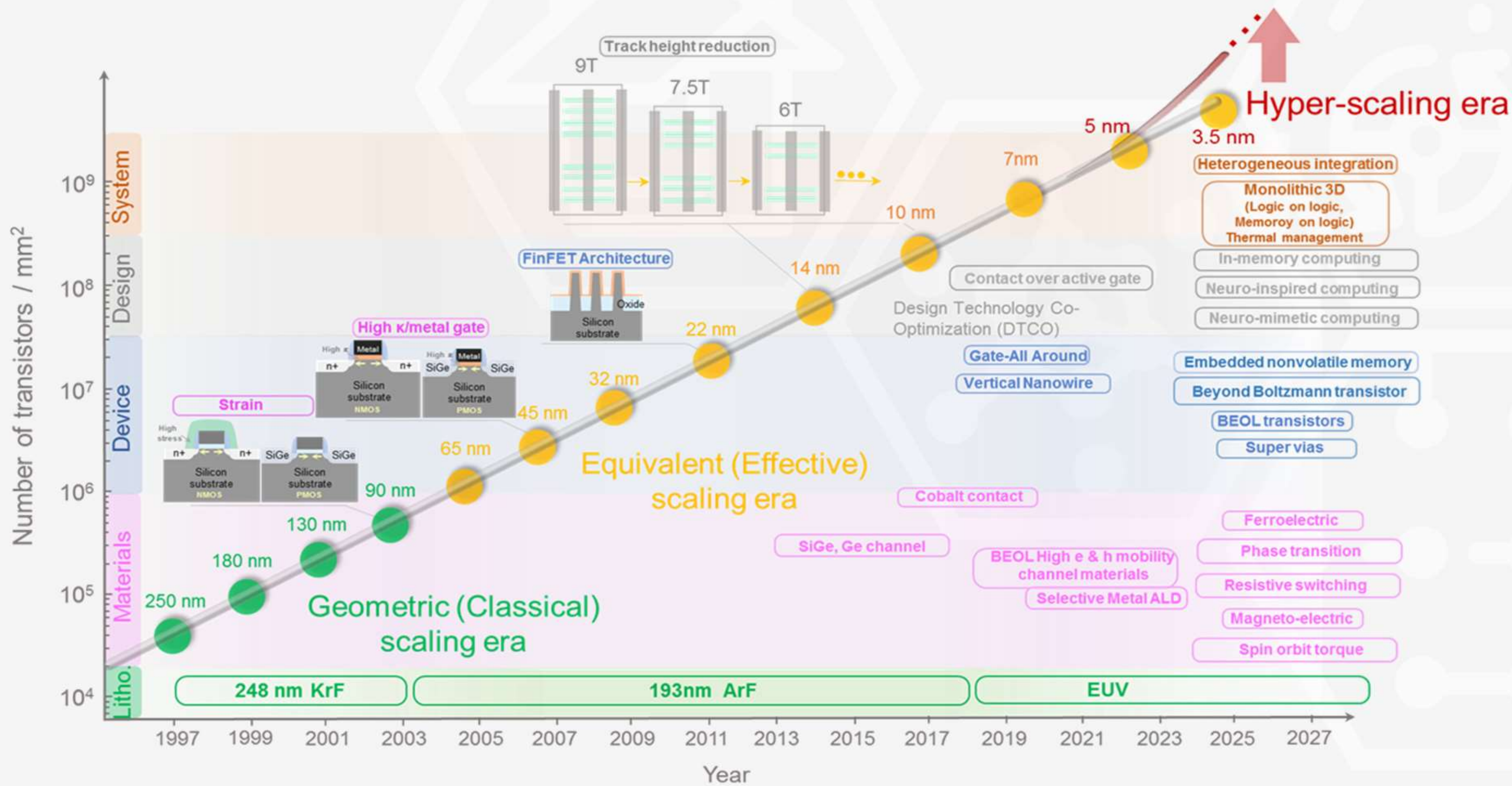
UNIVERSITY OF NOTRE DAME

ASCENT:

APPLICATIONS AND SYSTEMS DRIVEN CENTER FOR ENERGY-EFFICIENT INTEGRATED NANOTECHNOLOGIES

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

GRAND CHALLENGES IN SEMICONDUCTOR INDUSTRY



ASCENT GOALS

- "….drive foundational developments around specific disciplines with the goal of creating disruptive breakthroughs …."

  - three-dimensional integration of device technologies (theme 1)
  - spin-based memory and logic (theme 2)
  - heterogeneous integration of functionally diverse components (theme 3)
  - hardware accelerators for data intensive cognitive workloads (theme 4)

ASCENT ORGANIZATION

Suman Datta
Notre Dame
Director

Sayeef Salahuddin
UC Berkeley
Assistant Director

Bob Dunn
Managing Director

Barbara Walsh
Admin Support

Theme 1 Vertical CMOS	Theme 2 Beyond CMOS	Theme 3 Heterogeneous Integration Fabric	Theme 4 Merged Logic-Memory Fabric	Sponsors
Darrell Schlom LEAD, Cornell	Ramesh Ramamoorthy LEAD, UC Berkeley	Arijit Raychowdhury LEAD, Georgia Tech	H.-S. Philip Wong LEAD, Stanford	SRC
Kyeongjae Cho, UT Dallas	Jeff Bokor, UC Berkeley	Patrick Fay, Notre Dame	Suman Datta, Notre Dame	DARPA
Steve George, U Colorado	Supriyo Datta, Purdue	Subu Iyer, UCLA	Mike Niemier, Notre Dame	NIST
Andrew Kummel, UC San Diego	Sayeef Salahuddin, UC Berkeley	Umesh Mishra, UC Santa Barbara	Shimeng Yu, Arizona State	Analog Devices
Eric Pop, Stanford	J.P. Wang, U Minnesota	Madhavan Swaminathan, Georgia Tech		ARM Limited
		Peide Ye, Purdue		EMD Performance Materials
				IBM Corp.
				Intel Corp.
				Lockheed Martin Corp.
				Micron Technol.
				Northrop Grumman
				Raytheon
				Samsung
				TSMC

ARIZONA STATE • CORNELL • GEORGIA TECH • NOTRE DAME • PURDUE • STANFORD
U-COLORADO • U-MINNESOTA • UC BERKELEY • UCLA • UC SANTA BARBARA • UC SAN DIEGO • UT DALLAS

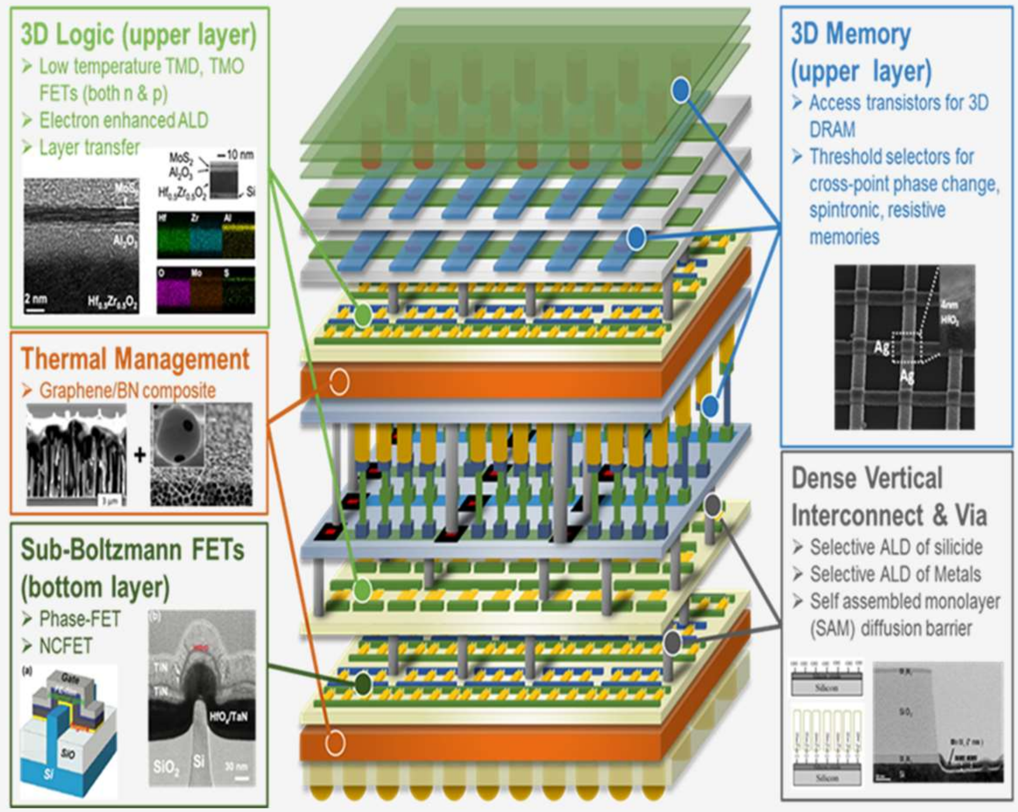
THEME 1: VERTICAL CMOS

Why Vertical CMOS:

1. Reduce interconnect bottleneck
2. Increase # gates/mm^2

Grand Challenges:

1. Protect bottom layer transistors
2. Align top layer with bottom layer
3. Low resistivity inter-layer vias
4. Thermal management
5. Cost of layering logic and memory in a single die



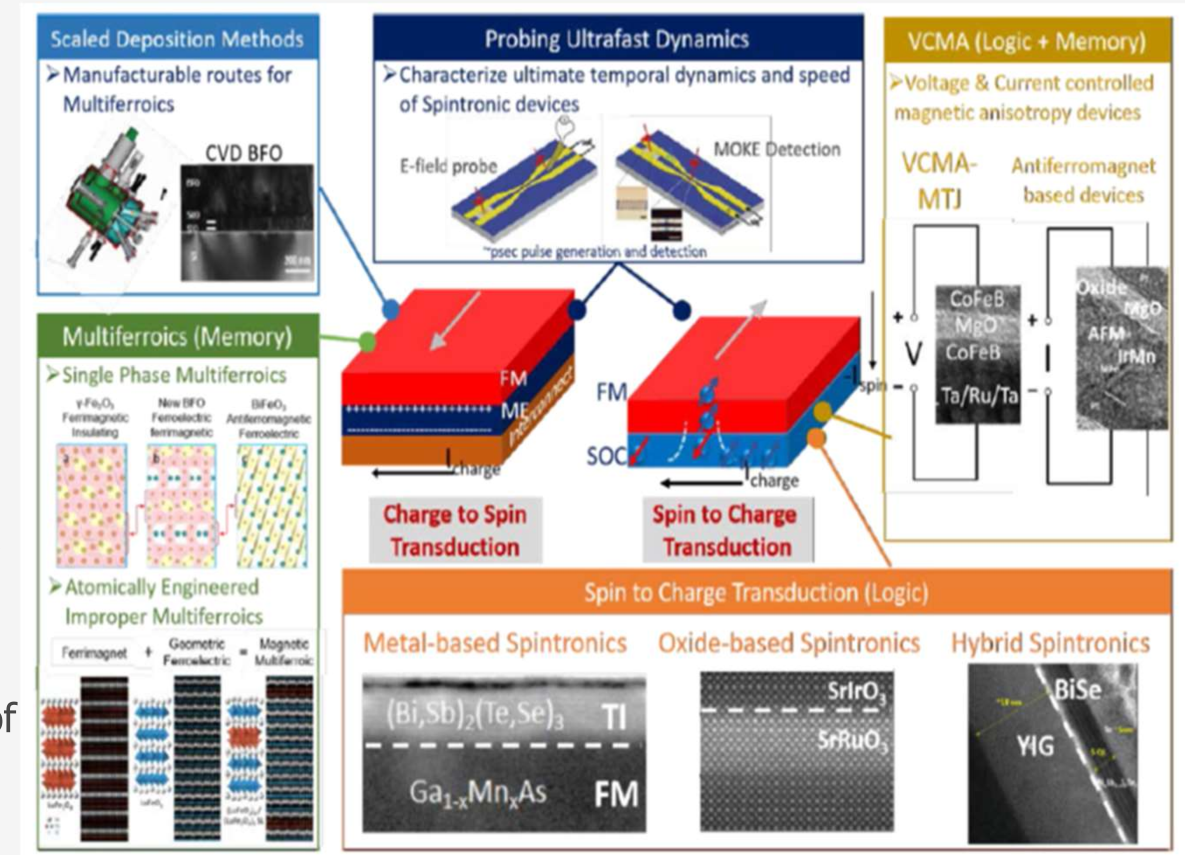
THEME 2: BEYOND CMOS

Why Beyond CMOS:

1. Non-volatile, ultra-high speed, unlimited endurance
2. Energy dissipation approaching thermodynamic limit

Grand Challenges:

1. Low voltage (100mV) driven manipulation of magnetic information
2. Convert magnetic information into voltage signal of 100mV
3. Switch magnetic order at 10s of picosecond
4. Low current driven manipulation of magnetic memory



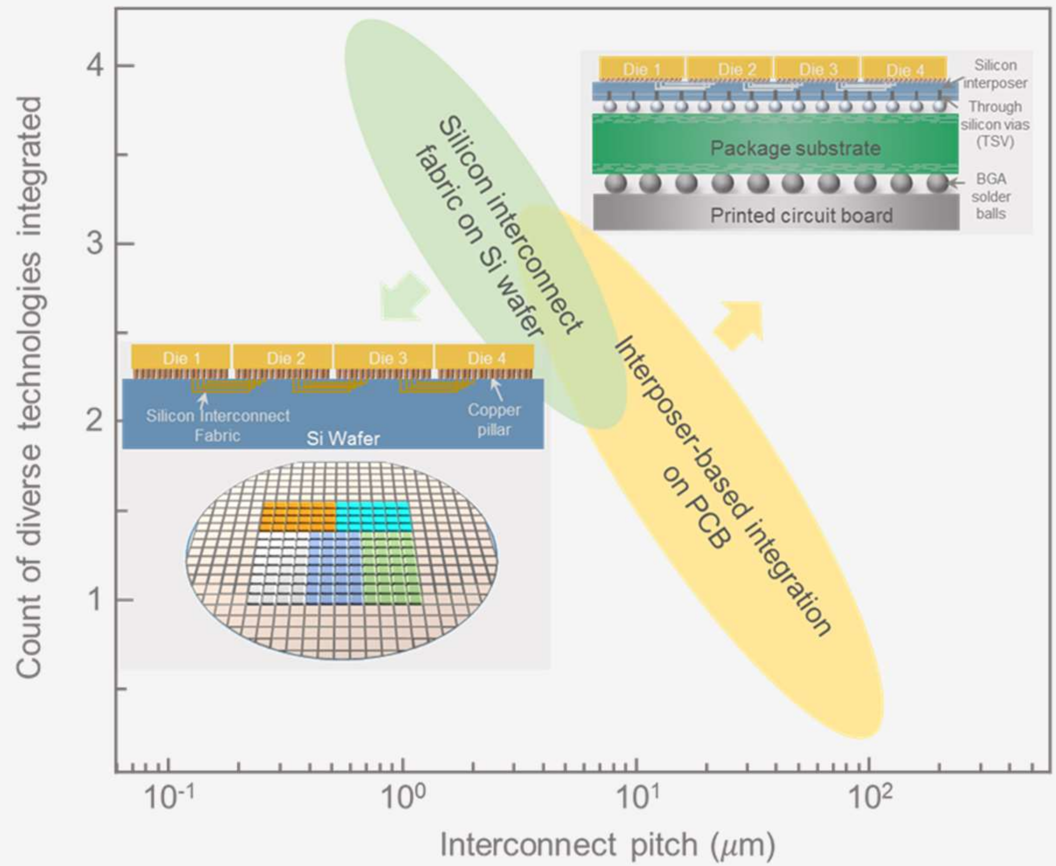
THEME 3: HETEROGENEOUS CMOS

Why Heterogeneous Integration:

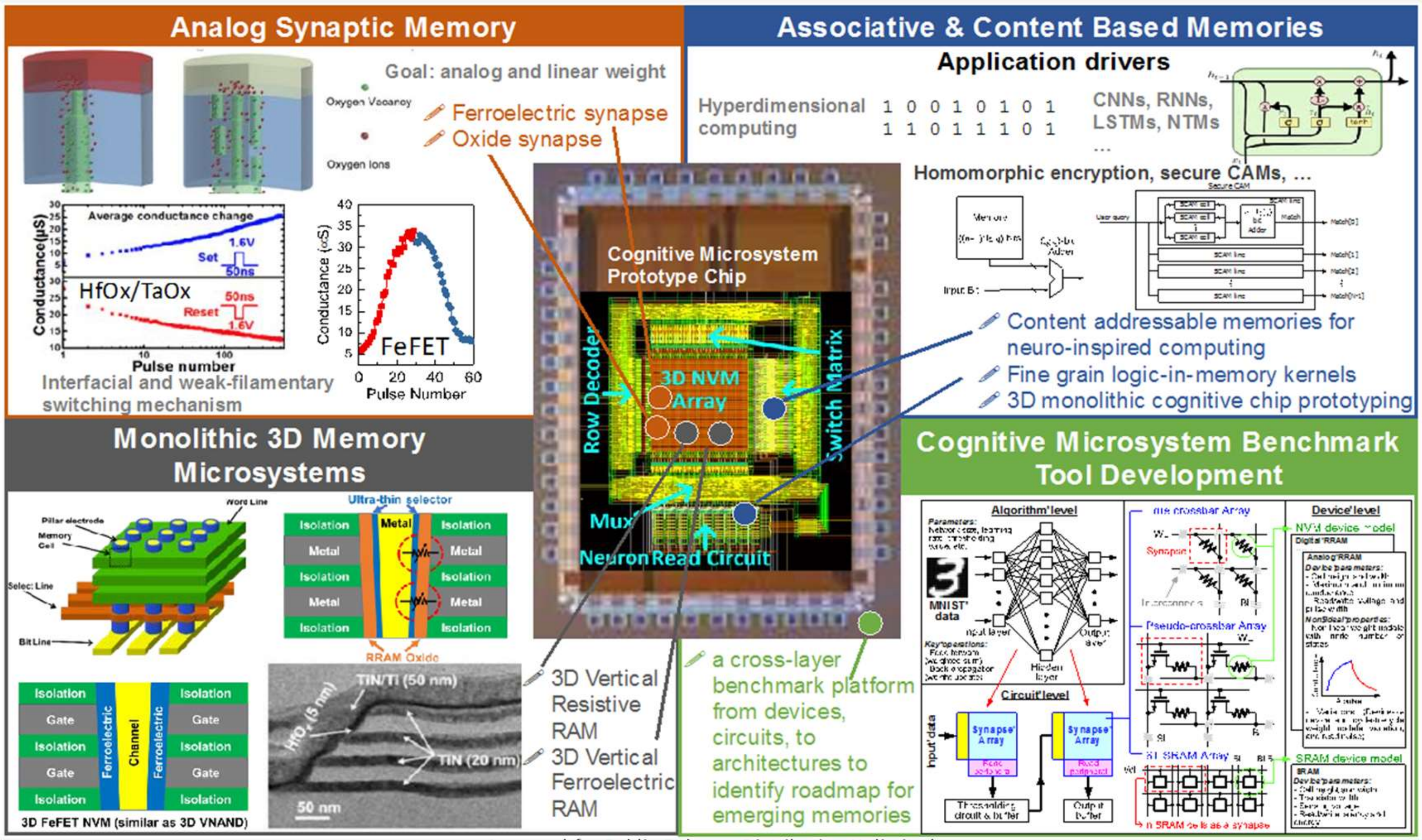
1. PCB ultimately limits size, weight, area, performance of microsystems
2. Silicon IP reuse

Challenges:

1. Fine pitch interconnect, micro-aligned integration of functionally, technologically diverse ICs on a universal substrate
2. Reach 2um interconnect spacing and <50 um die to die spacing
3. Achieve aggregate data transfer rate of 1Tb/s/mm at < 0.1pJ/bit



THEME 4: MERGED LOGIC-MEMORY FABRIC



SUMMARY

Vertical CMOS

Enable hyper scaling by stacking logic and memory layers in the vertical dimension with ultra-dense connectivity

Beyond CMOS

Combine logic and memory functions and operate spintronic units near thermodynamic limit

Multi-function Heterogeneous Fabric

Combine the best of chip technologies and design IPs into a heterogeneous microsystem by tiling dielets together on an ultra-dense and energy-efficient interconnect fabric

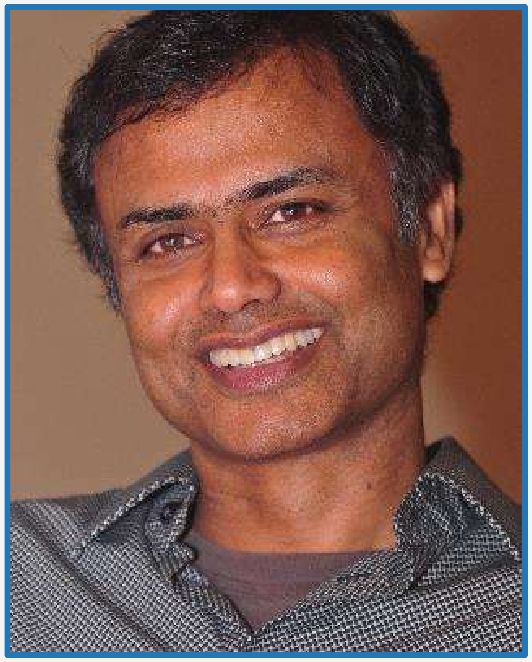
Merged Logic-Memory Fabric

Leverage innovations in vertical 3D memory technologies to create merged logic-memory fabrics to accelerate cognitive and secure computing workloads

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | JULY 23-25



KAUSHIK ROY

PURDUE UNIVERSITY
WEST LAFAYETTE, IN

CENTER FOR BRAIN-INSPIRED COMPUTING (C-BRIC)

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Images may come from public domains which are used for research purposes with limitations of Section 107 of the Copyright Act

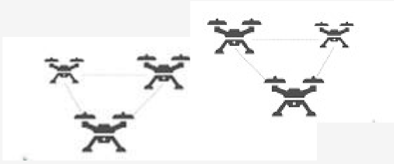
TOWARDS AN AI-DRIVEN WORLD

- Current applications of AI / machine learning are only the tip of the iceberg
  - Few large "killer apps"
- Tremendous potential for economic and societal impact if AI can be applied to a much broader range of applications

Sources: Images may come from public domains which are used for research purposes with limitations of Section 107 of the Copyright Act

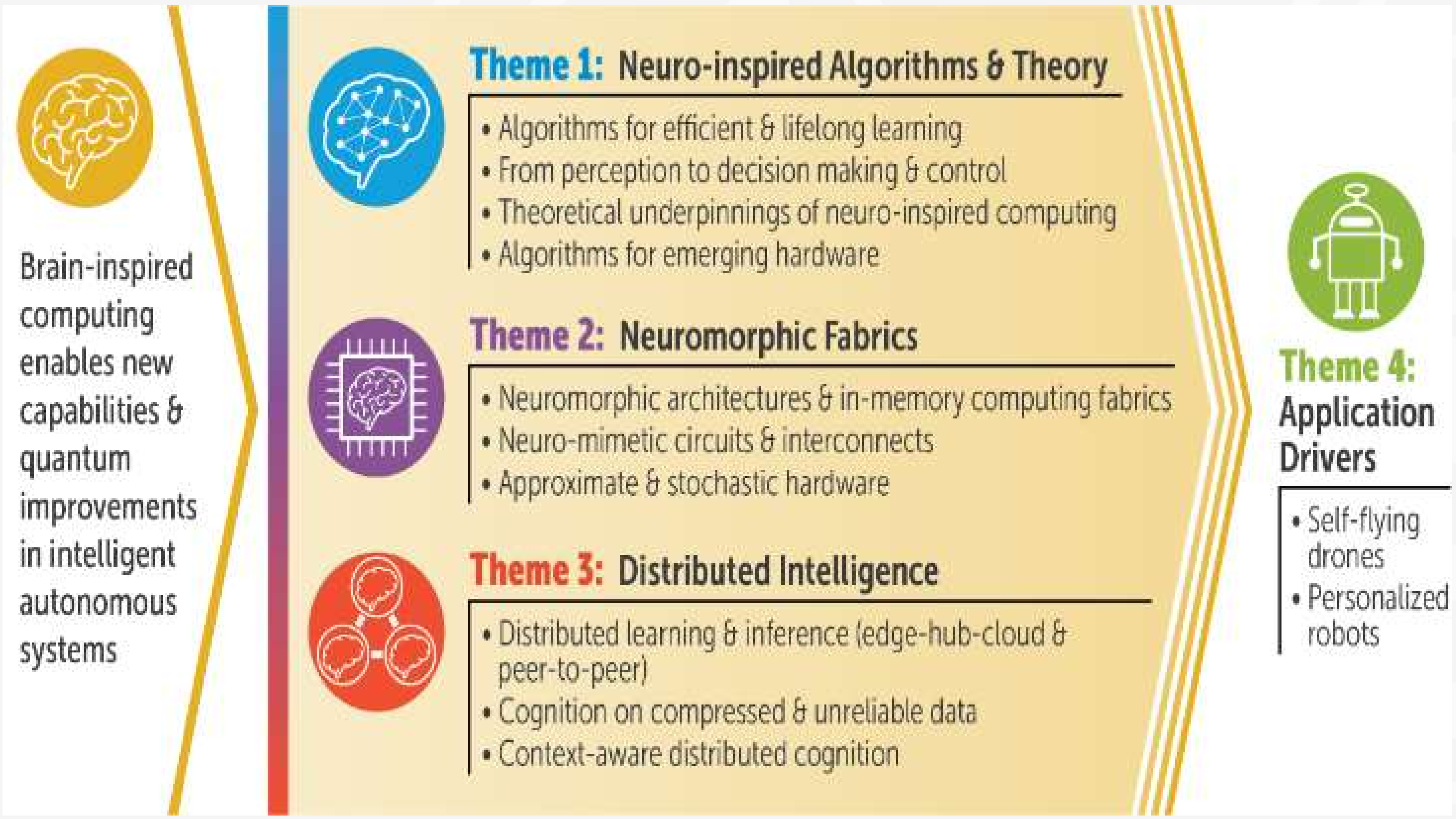
C-BRIC VISION

- Enable next generation of intelligent autonomous systems
  - Narrow the orders-of-magnitude computing efficiency gap between current computing systems and the brain
  - Drive improvements in the robustness of cognitive computing systems
  - Explore distributed intelligence across edge/hub/cloud and peer-to-peer networks
  - Demonstrate the impact of these advances in end-to-end systems such as autonomous drones and personal robotics



Sources: Images may come from public domains which are used for research purposes with limitations of Section 107 of the Copyright Act

C-BRIC ORGANIZATION



WHAT SHOULD BE BRAIN-INSPIRED?

Discrete Spiking (LIF)?
HH?

Long-range connections?
Feedback?

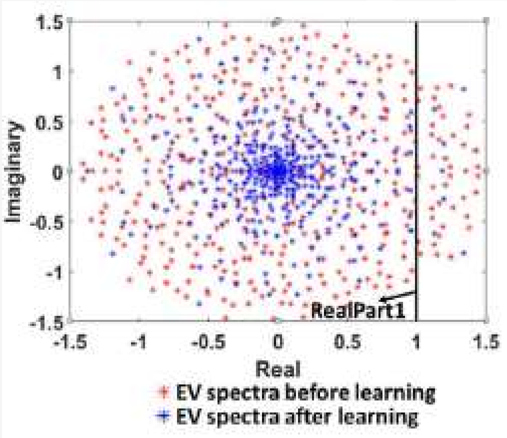
STDP?

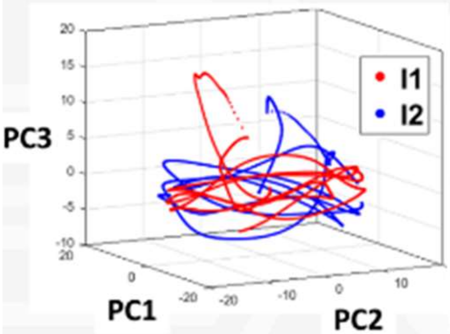
Tightly integrated compute & memory?
Asynchronous, event-driven?
Time-based coding?

THEME 1: NEURO-INSPIRED ALGORITHMS AND THEORY

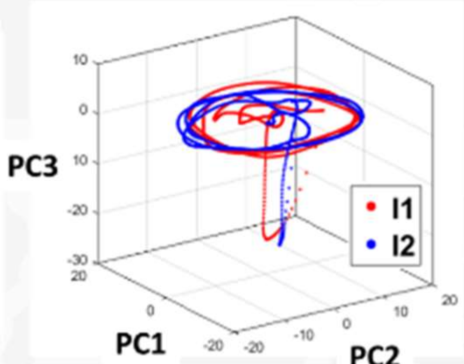
State-of-the-Art: Deep Neural Nets		Proposed Neuro-Inspired Algorithms
• Largely supervised learning • Static (one-time) learning • Training requires global updates (Backpropagation / SGD) • Perception (speech, images, text) • Unknown generalization behavior • Manually designed network topologies	C-BRIC Theme 1 →	• **Computationally efficient algorithms** • **Theory of neural computing from DNN to emerging models** • **Learning with less data** • **Incremental and lifelong learning** • **Algorithms that leverage stochastic and approximate computation** • **Learning and inference on emerging computing fabrics**

THEORETICAL UNDERSTANDING OF LEARNING





Network Activity
before Learning

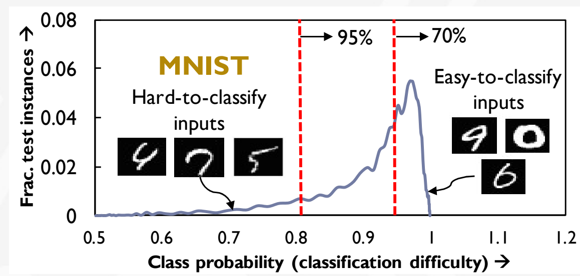


Network Activity
after Learning

- Shrinking of the EigenValue spectral circle represents the stabilizing effect of the learning mechanism
- Understanding network behavior from Random Matrix theory and Principal Component Analysis
- Quantification of stabilizing hyper parameters from network activity

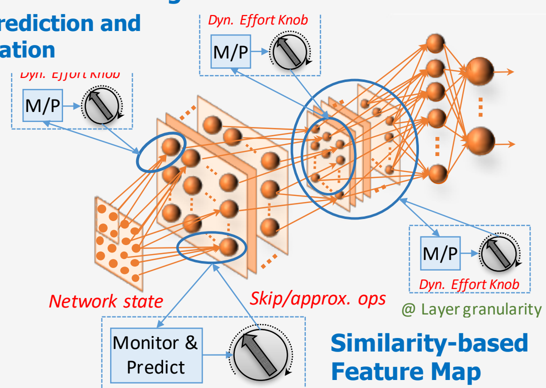
DYNAMIC, VARIABLE-EFFORT DEEP NETWORKS

- Deep nets are fixed-effort and static
- Inputs differ greatly in their difficulty
- Mechanisms to dynamically modulate computational effort of neural nets



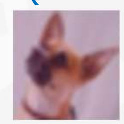
Significance-driven Feature Evaluation

Saturation Prediction and Early Termination

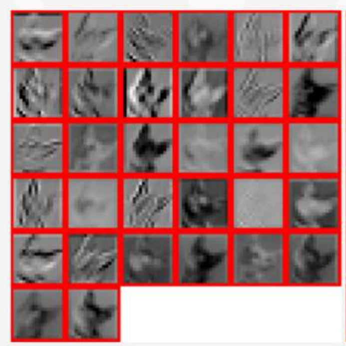


Similarity-based Feature Map Approximation

Input (CIFAR-10)



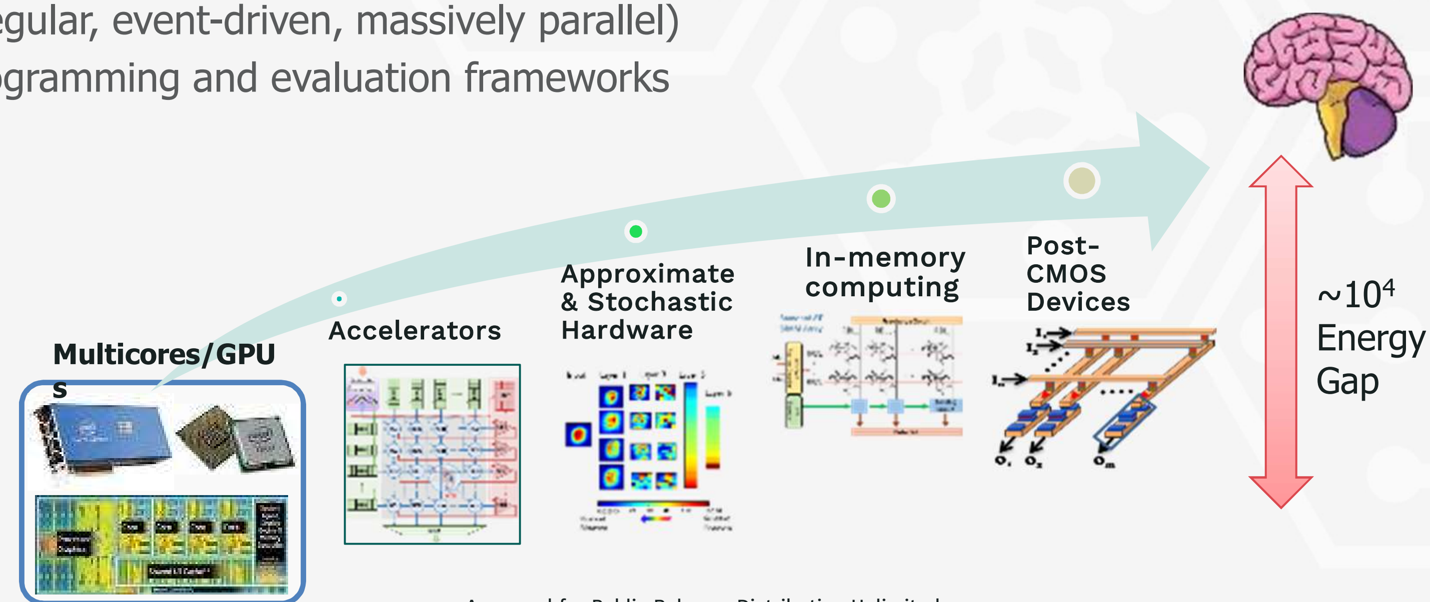
Feature maps (C1 layer)



Computational effort

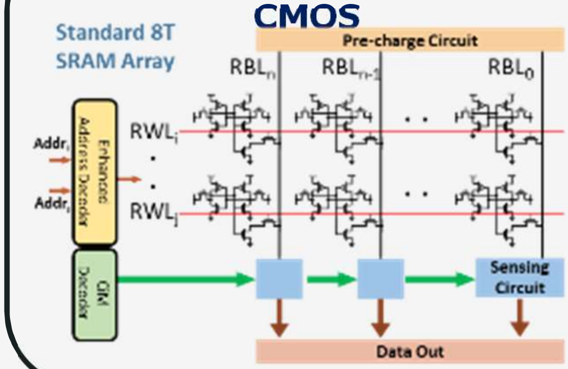
THEME 2: NEUROMORPHIC FABRICS

- CMOS and Post-CMOS neuro-mimetic devices and interconnects
- Compute-near-memory / Compute-in-memory
- Approximate and stochastic neuronal and synaptic hardware
- Architectures that embody computing principles from the brain (sparse, irregular, event-driven, massively parallel)
- Programming and evaluation frameworks

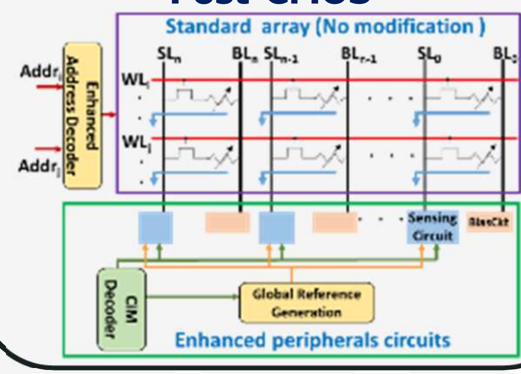


COMPUTE-IN-MEMORY

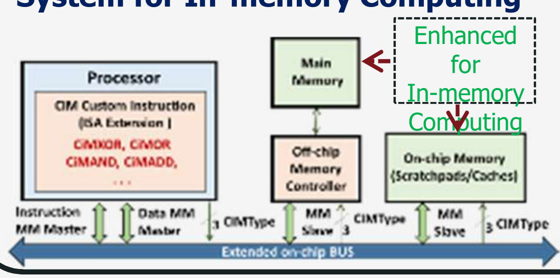
Compute-in-memory with CMOS



Compute-in-memory with Post-CMOS



System for In-memory Computing



Load R_{ADDR1} R_{DEST0}
Load R_{ADDR2} R_{DEST1}
ADD R_{DEST0} R_{DEST1} R_{OUT}

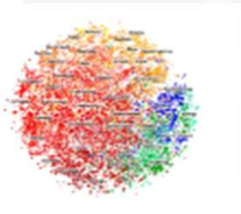
Conventional Computing

CiMADD R_{ADDR1} R_{ADDR2} R_{OUT}

In-memory Computing

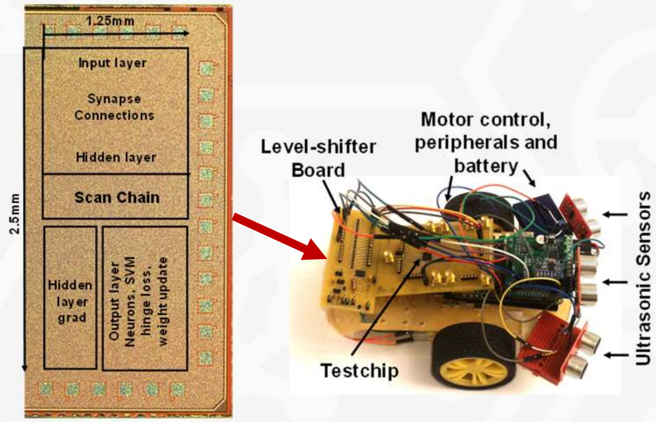
HARDWARE DEMONSTRATION OF AUTONOMOUS DECISION MAKING VIA REINFORCEMENT LEARNING

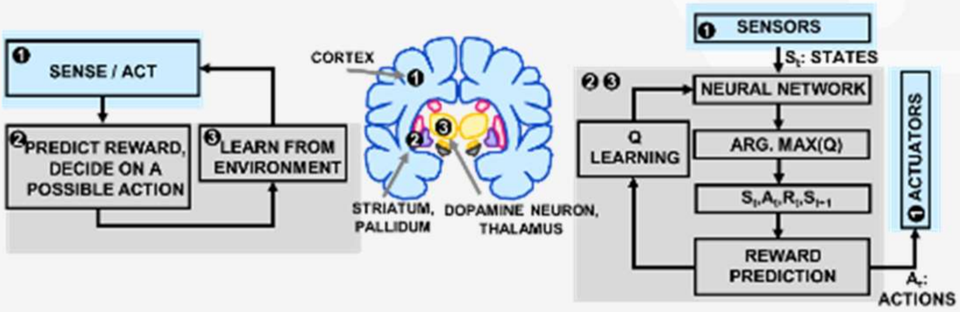
Supervised
Learning known patterns



Unsupervised
Learning unknown patterns

Reinforcement
Generating data
Learning patterns





THEME 3: DISTRIBUTED INTELLIGENCE

State-of-the-Art:
Cloud-enabled Intelligence

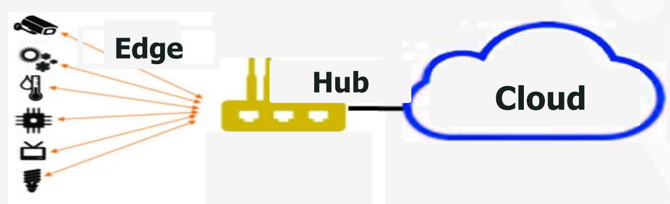
- Centralized training in cloud
- Inference entirely in cloud or entirely on edge device
- Algorithms agnostic to distributed context require high communication

C-BRIC Theme 3

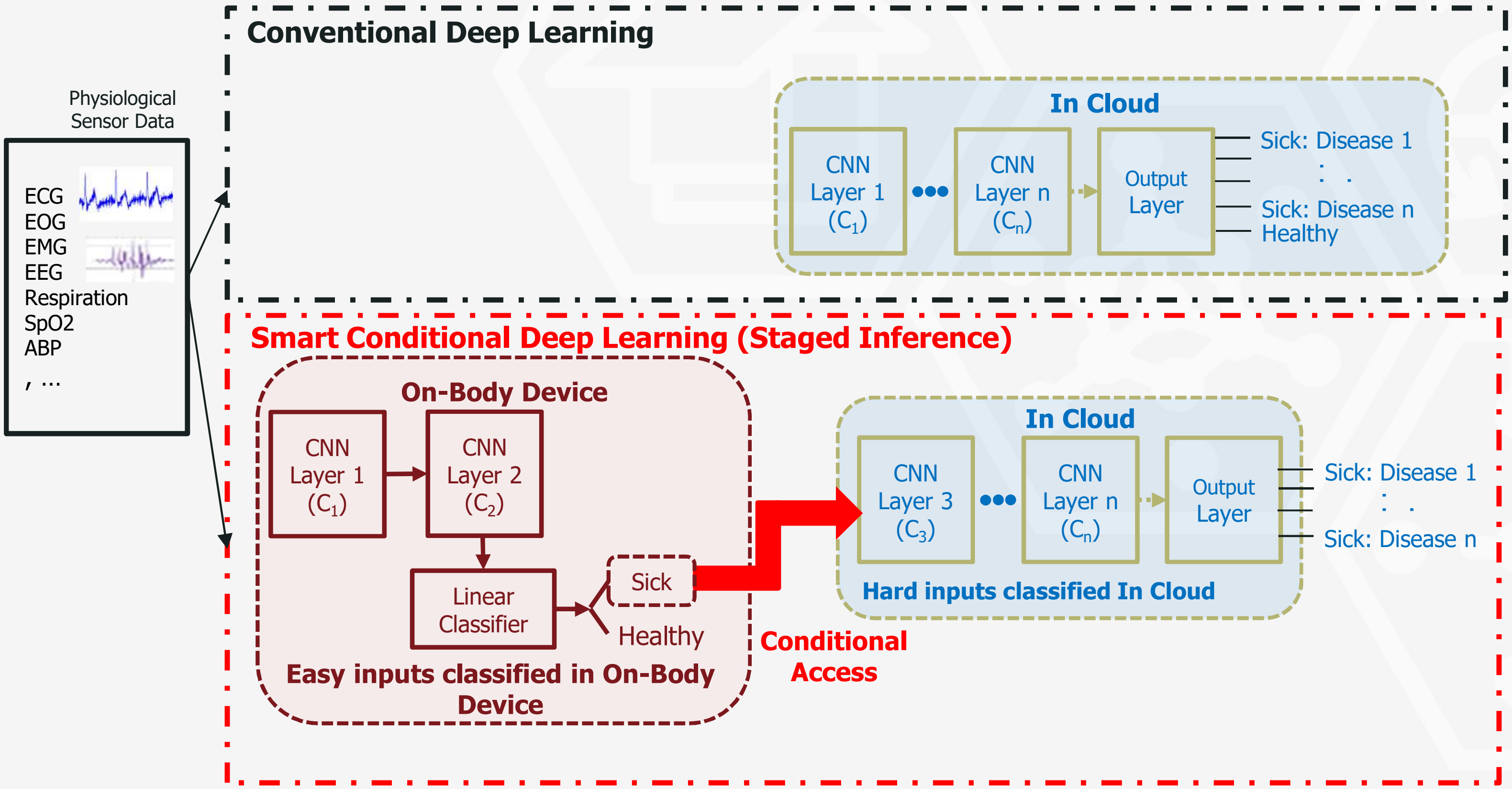
Proposed Distributed Intelligence

- **Partitioned learning and inference**
  - **Algorithms for hierarchical (edge/hub/cloud) and peer-to-peer networks**
- **Cognition on compressed and unreliable data**
  - **Event-driven sensors, data fusion, learning from incomplete/ unsynchronized/noisy data**
- **In-sensor analytics**
  - **Low-complexity algorithms and**

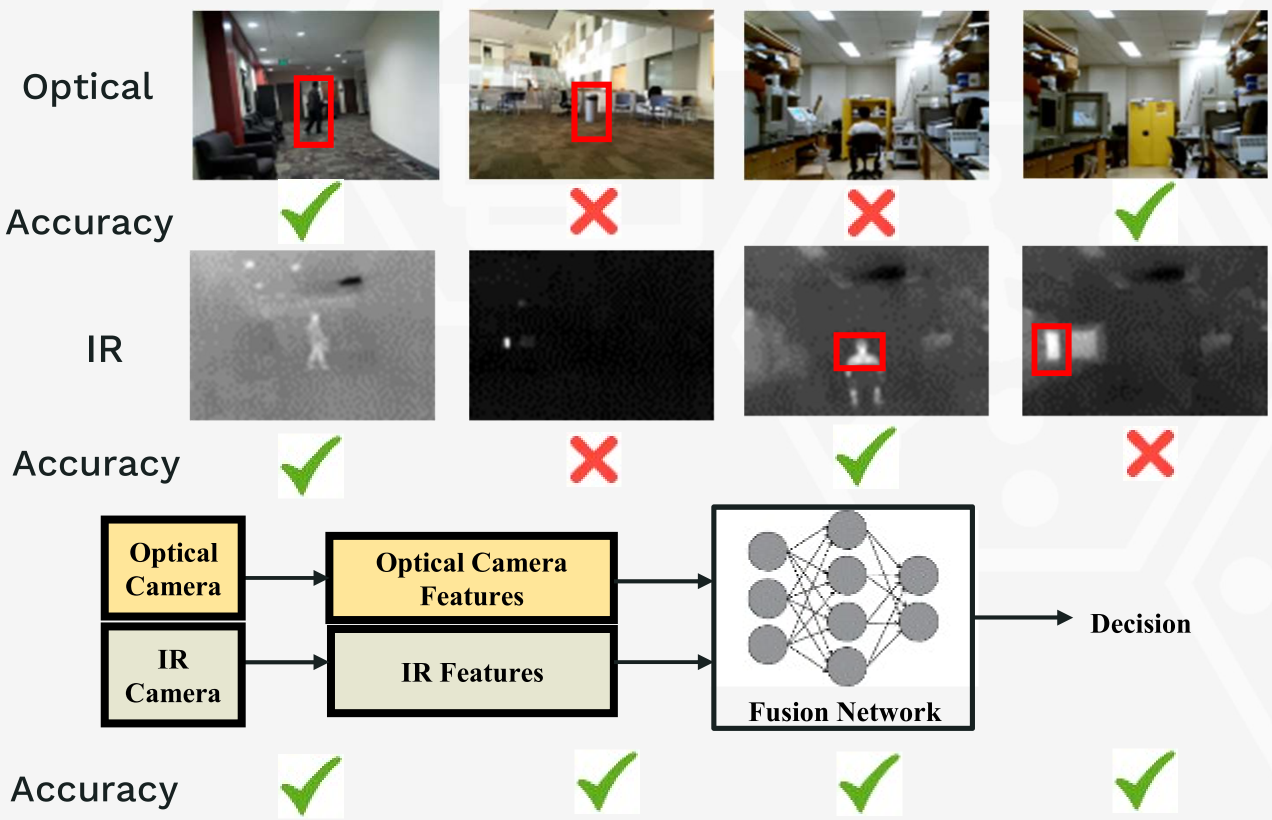
Peer-to-peer



STAGED CONDITIONAL LEARNING/INFERENCE



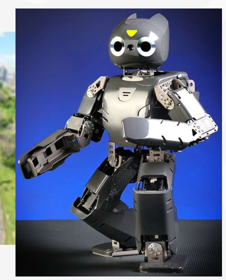
MULTI-SENSOR COGNITION IN SMART BUILDINGS



THEME 4: APPLICATION DRIVERS

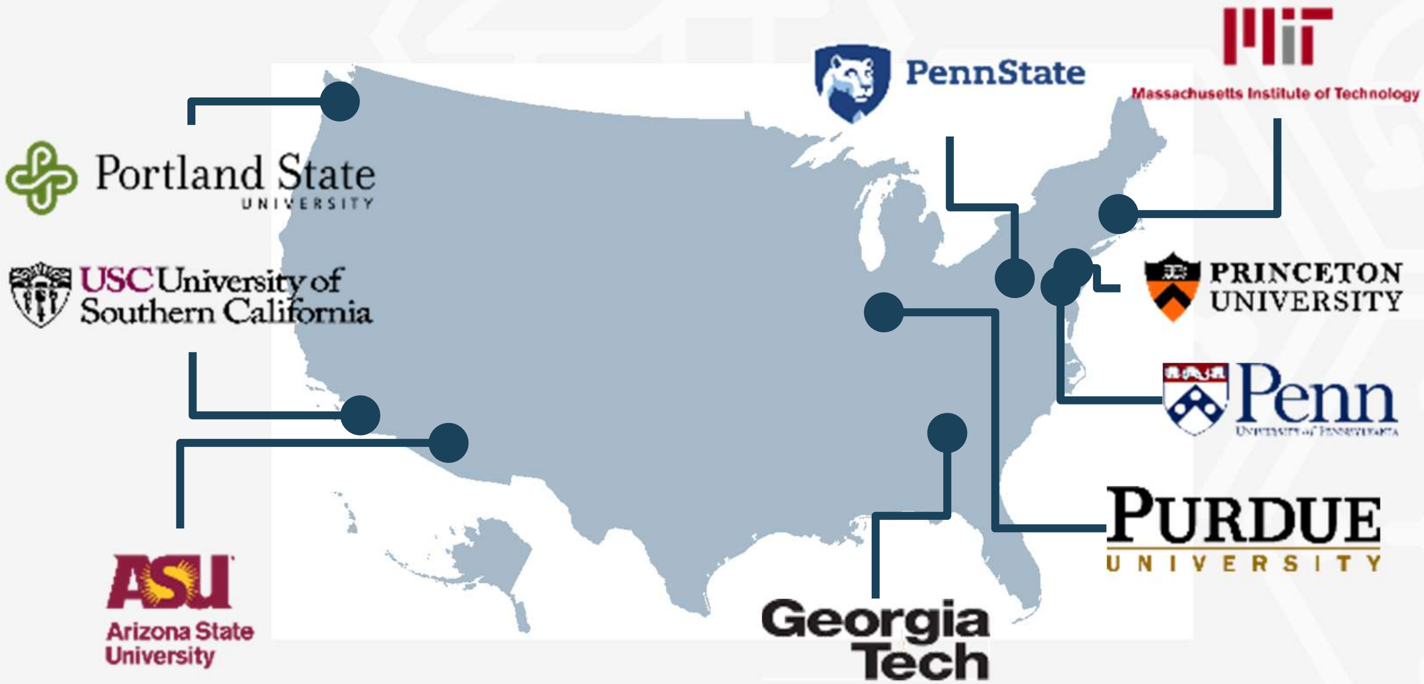
- Autonomous drones and drone swarms
- Personal robotic assistants
- Technologies from Themes 1-3 enable new capabilities with real-time, autonomous operation



Sources: Images may come from public domains which are used for research purposes with limitations of Section 107 of the Copyright Act

C-BRIC UNIVERSITIES



C-BRIC

Center for Brain Inspired Computing

Approved for Public Release, Distribution Unlimited

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

MARK RODWELL

UNIVERSITY OF CALIFORNIA,
SANTA BARBARA

A CENTER FOR
CONVERGED TERAHERTZ COMMUNICATIONS & SENSING

Mark Rodwell, Ali Niknejad
University of California, Santa Barbara
University of California, Berkeley

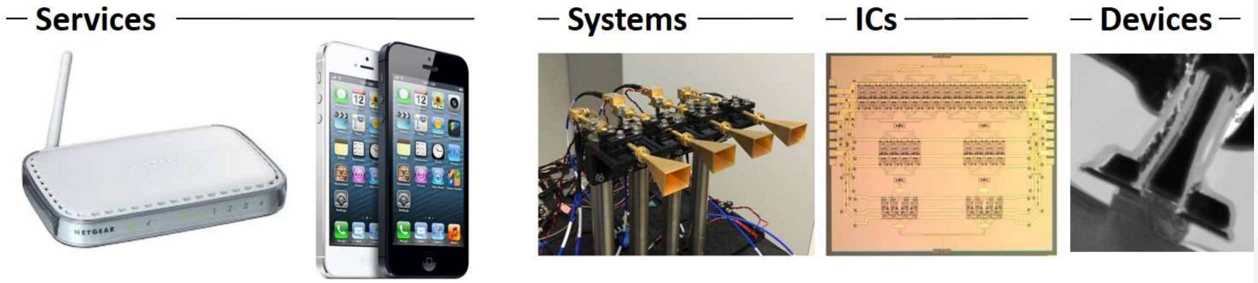
***Debdeep Jena, Alyosha Molnar, Christoph Studer, Huili Xing**: Cornell University*
***Dina Katabi**: MIT*
***Sundeep Rangan**: New York University*
***Amin Arbabian**: Stanford*
***Elad Alon, Ali Niknejad, Borivoje Nikolic, Vladimir Stojanovic**: University of California, Berkeley*
***Srabanti Chowdhury**: University of California, Davis*
***Gabriel Rebeiz**: University of California, San Diego*
***Jim Buckwalter, Upamanyu Madhow, Umesh Mishra, Mark Rodwell**: University of California, Santa Barbara*
***Andreas Molisch**: University of Southern California*
***Kenneth O**: University of Texas, Dallas*

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Sources: Images may come from public domains which are used for research purposes with limitations of Section 107 of the Copyright Act

Joint University Microelectronics Program

COMMUNICATIONS SENSING TERAHERTZ

WHY 100+ GHZ WIRELESS ?



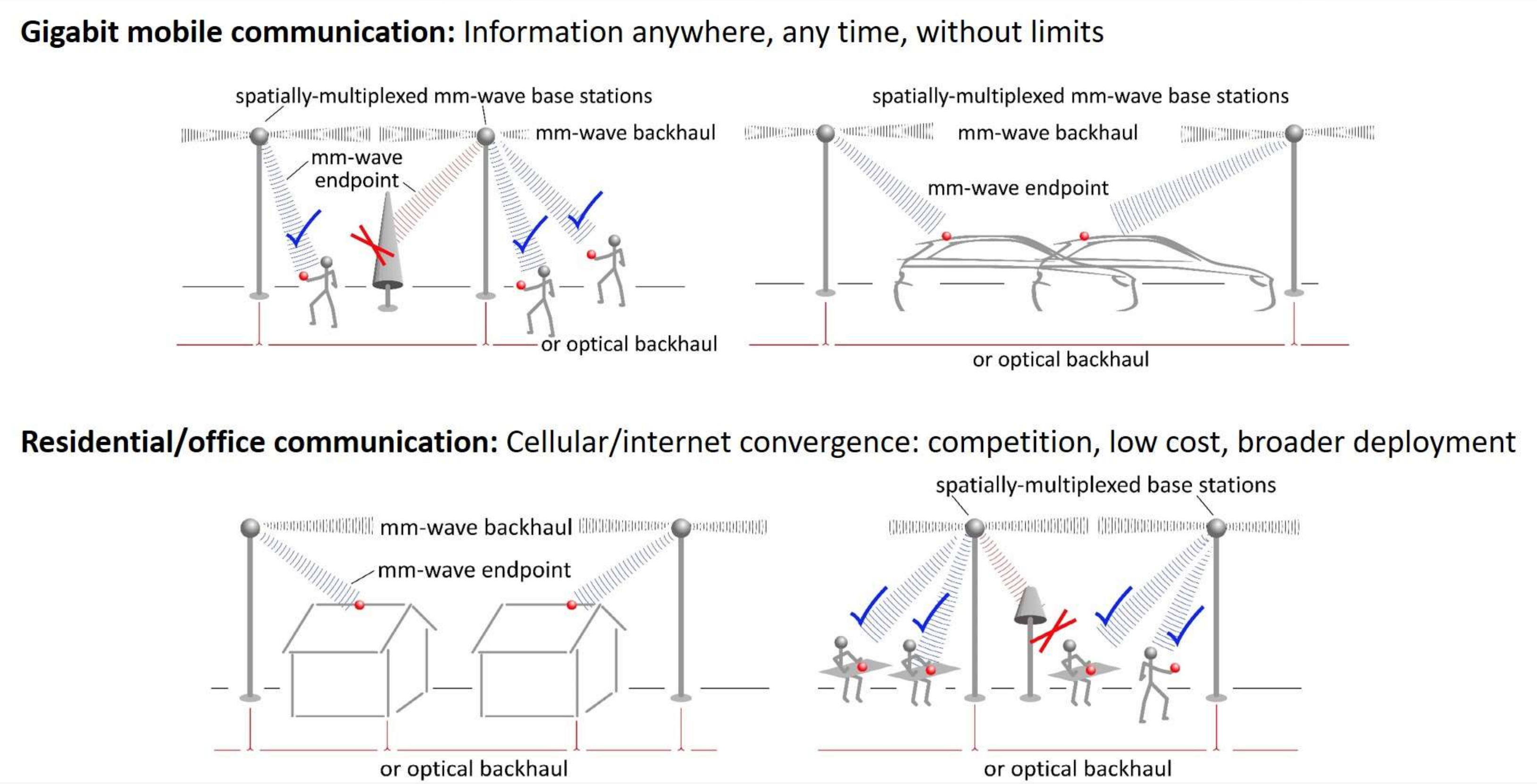
Wireless networks: exploding demand.

Immediate industry response: 5G.
28, 38, 57-71(WiGig), 71-86GHz
increased spectrum, extensive beamforming

Next-generation: above 100GHz.
greatly increased spectrum, massive spatial multiplexing

DOD Applications: Imaging/sensing/radar, comms.

140-1080GHZ: A REVOLUTION IN COMMUNICATIONS



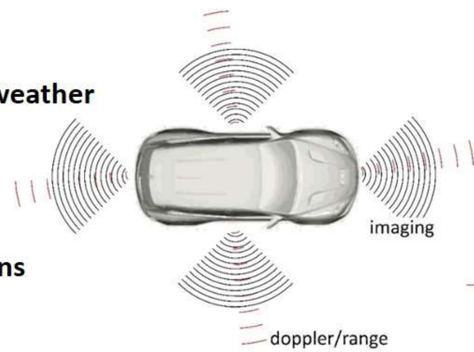
140-1080GHZ IMAGING: FOG/CLOUDS/SMOKE/DUST

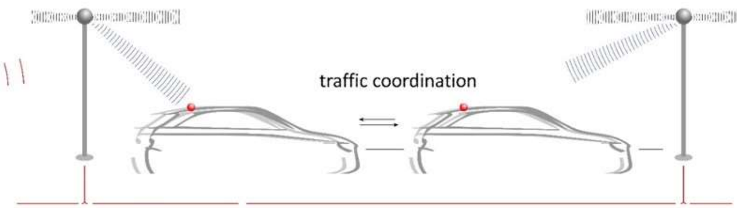
Automatic car, intelligent highway

340 GHz HDTV-resolution radar
drive safely in fog at 100 km/hr
self-driving: complements LIDAR, works in bad weather

Complements 70 GHz Doppler / ranging radar.
object near ? approaching ? Can't tell what.

Intelligent highway: coordinate traffic
anticipate & manage interactions, avoid collisions

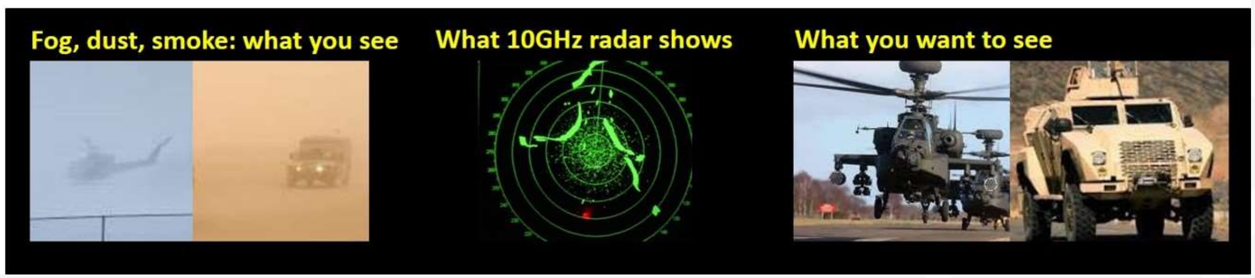




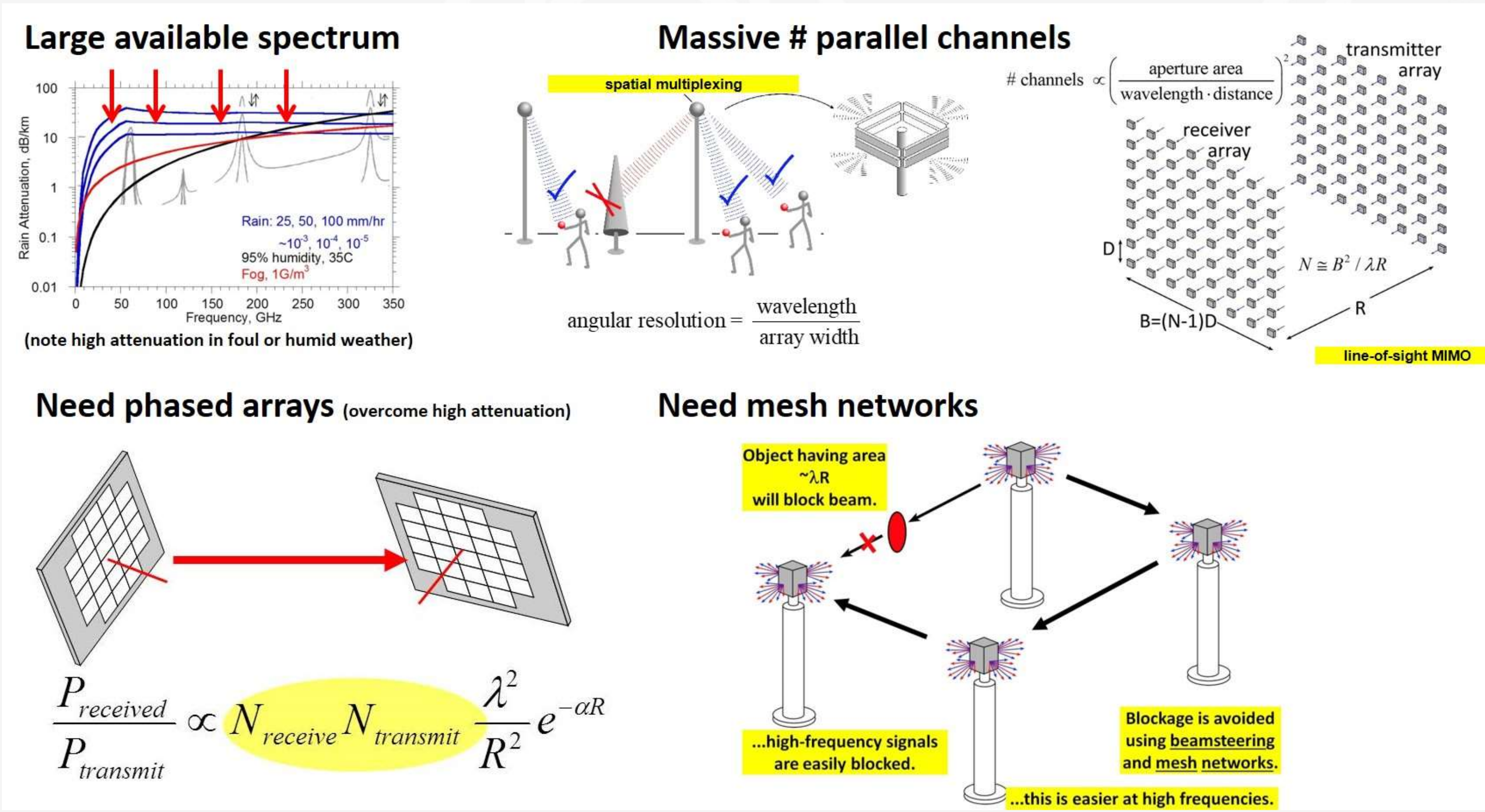
Sensing/imaging for national security

20/70/ 94 GHz radar: is something there ?
Long-range, low-resolution: can't tell what.

140-340GHz imaging radar: what is it ?
shorter range, TV-like resolution
small, light: jeep, helicopter, UAV.

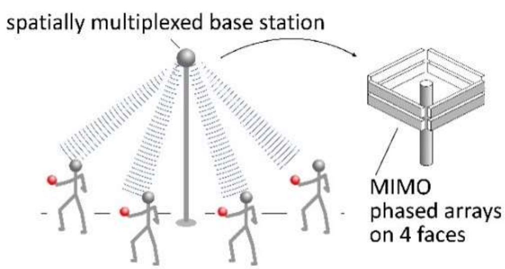


140-1000 GHZ: BENEFITS & CHALLENGES

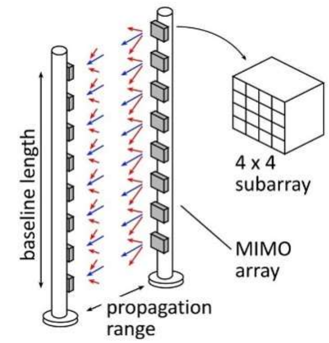


140-1080GHZ DEMONSTRATIONS

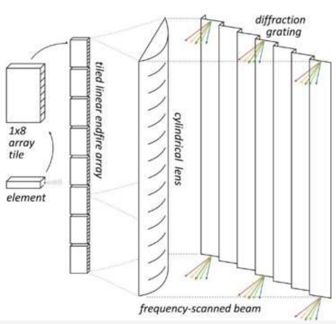
MIMO hub: 256 beams/face, 10Gb/s/beam
140GHz, 220GHz



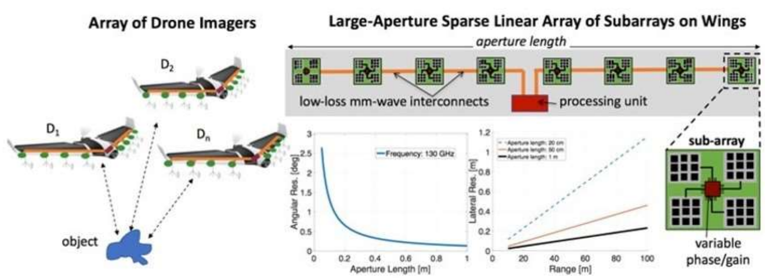
Point-point MIMO: 340GHz: 640Gb/s (650GHz: 1.3Tb/s)



Hardware-efficient 340GHz imaging
300 meters, 512× 64 image, 60Hz, 15 dB SNR



Cooperative / sparse 220 GHz imaging



ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

TAJANA ROSING

UNIVERSITY OF CALIFORNIA, SAN DIEGO

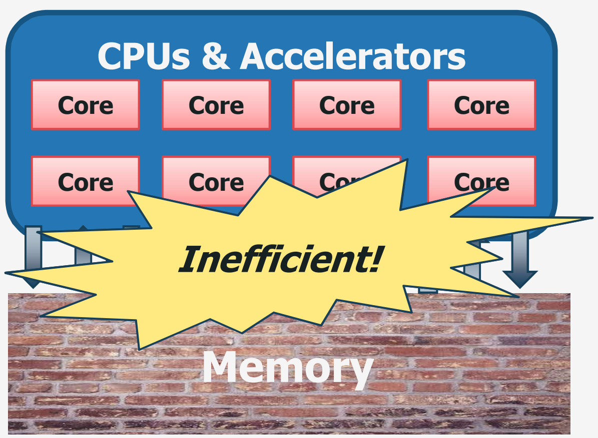
CRISP

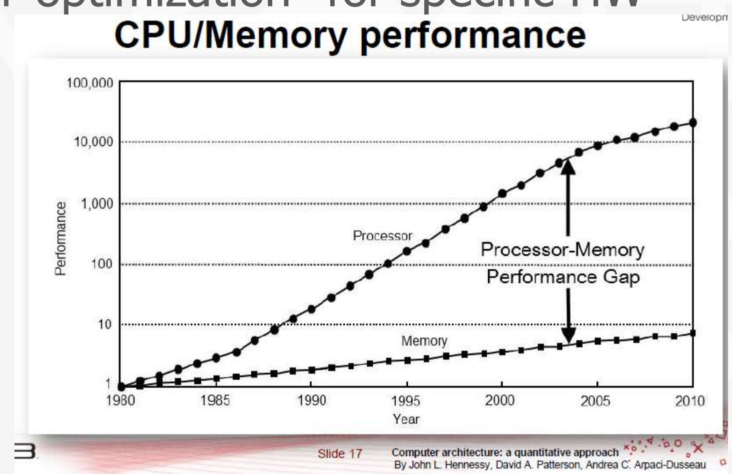
Center for Research on Intelligent Storage and Processing in Memory

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

TODAY'S SYSTEMS ARE HITTING THE MEMORY WALL

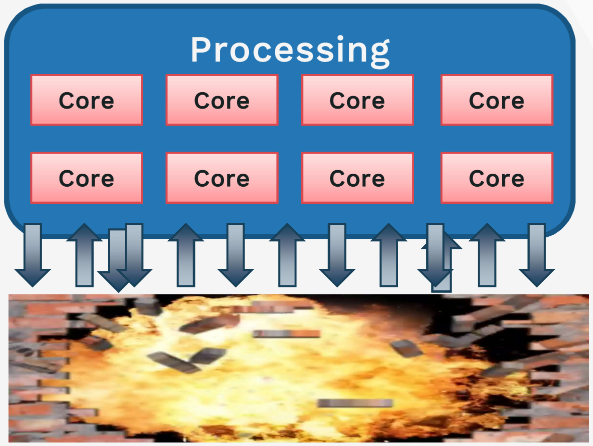
- Stems from separation of processing and memory/storage – von Neuman
  - Prefetching and caches used to hide it, but not anymore!
- Pervades the entire system design
  - Instruction sets hide programmer intent & higher-level data structures
  - Hardware and OS don't know what programmer really wants to do
  - Programming languages encourage "over-optimization" for specific HW

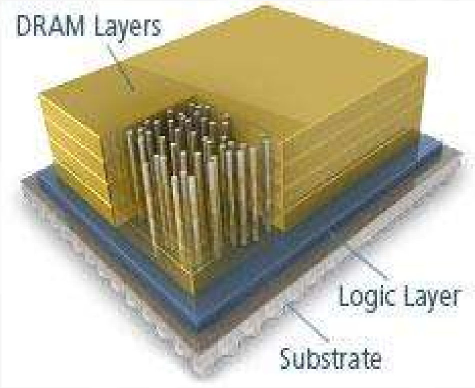




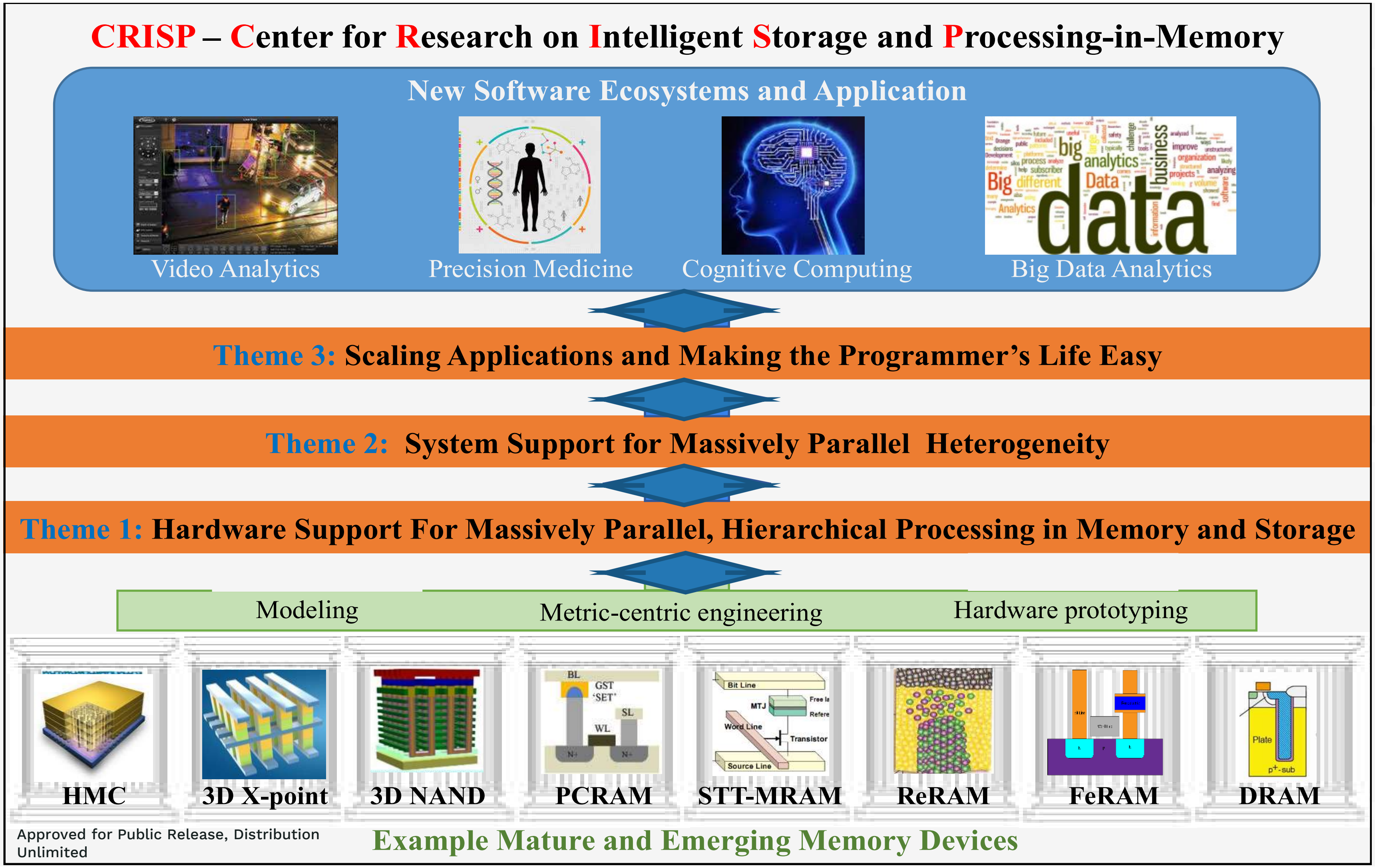
CRISP GOAL: BREAK DOWN THE MEMORY WALL!

- Integration of processing with mem/storage can provide dramatic increases in bandwidth & lower latencies
  - 2.5/3D stacking, logic in DRAM at edge of arrays
  - Emerging devices for processing in memory
- This will require full-stack solutions
  - New hardware, OS, & programming abstractions





Micron Hybrid Memory Cube

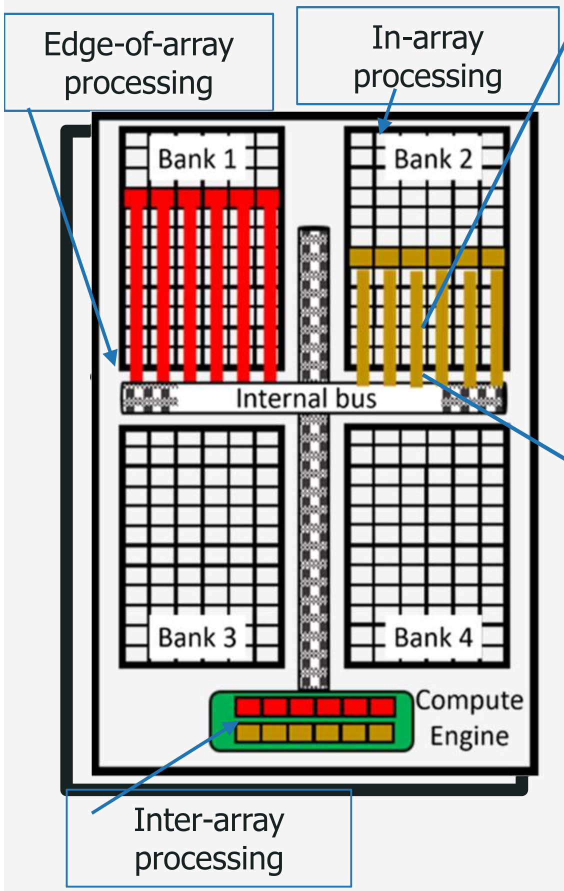

CRISP – Center for Research on Intelligent Storage and Processing-in-Memory
New Software Ecosystems and Application
Video Analytics
Precision Medicine
Cognitive Computing
Big Data Analytics
Theme 3: Scaling Applications and Making the Programmer's Life Easy
Theme 2: System Support for Massively Parallel Heterogeneity
Theme 1: Hardware Support For Massively Parallel, Hierarchical Processing in Memory and Storage
Modeling
Metric-centric engineering
Hardware prototyping
HMC
3D X-point
3D NAND
PCRAM
STT-MRAM
ReRAM
FeRAM
DRAM
Example Mature and Emerging Memory Devices

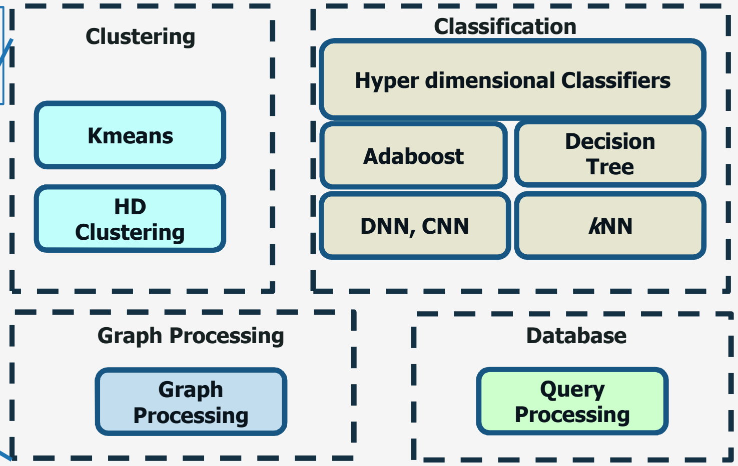
THEME 1 - HARDWARE

Reconcile underlying technology with effective architectural abstractions to operate as close as possible to parallelism available in the memory and storage

- Task 1.1: Organization and Hierarchy
  - Computing in the array, at the edge of the array, at the chip interface, etc.
  - ISA abstractions
- Task 1.2: Role of emerging semiconductor technology
  - Role of emerging devices
  - Integration options, from PIM to 3D to interposer
- Task 1.3: Node organization and HW/SW interface
  - Rethink node organization when processing coupled to a quantum of data
  - Introspection—HW for performance transparency and tradeoff management
- Task 1.4: Thermal/power
- Task 1.5: Simulation and prototyping

PROCESSING IN MEMORY / NEAR DATA (TASK 1.1)





PIM HD classification vs:

- DNN [ISCA'16] : 1,417× more efficient, 100x faster
- AMD R390 GPU: **12,000× higher energy efficiency & 1000× faster**

- Invert the relationship between computing & mem.
- Computing elements tightly integrated within *every possible level* of the memory/storage hierarchy
  - Progressively smaller processing units as we go farther into the mem/storage hierarchy

THEME 2 – SYSTEM SUPPORT

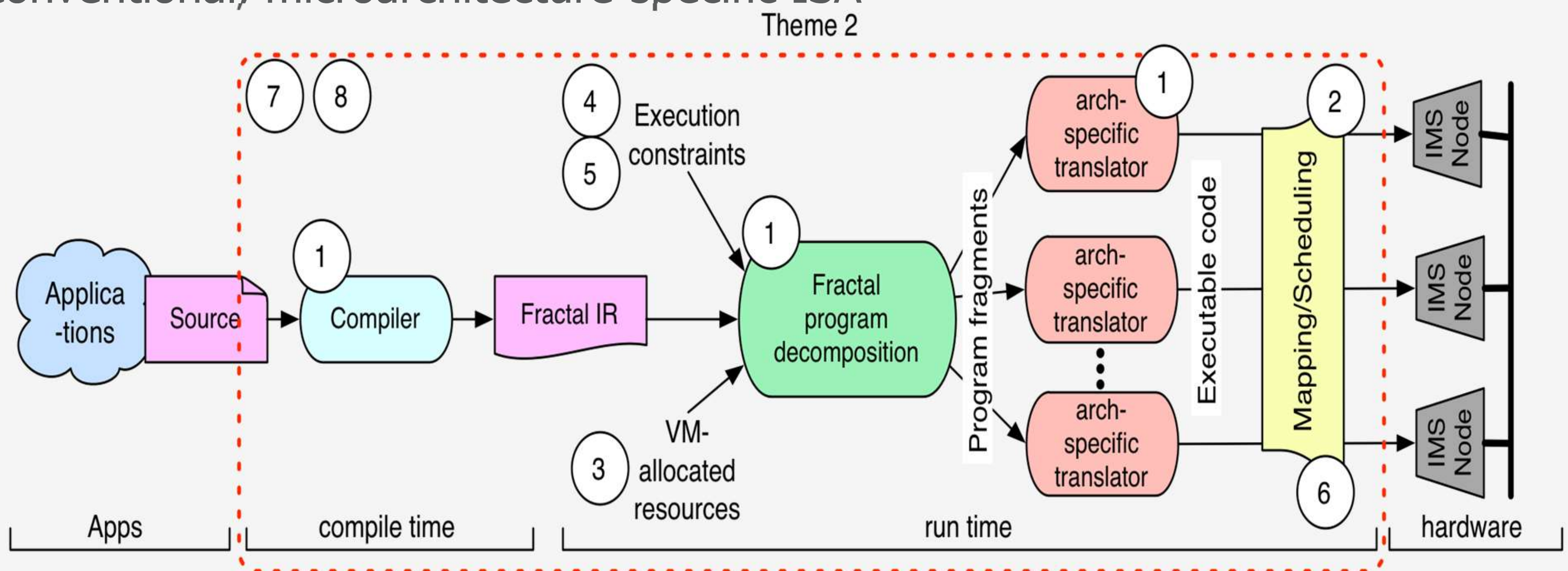
Connect programmer objectives to hardware, while providing dependability, tunability, etc.

- Task 2.1: Compiler
- Task 2.2: Runtime Middleware and Scheduling
- Task 2.3: Operating System and Virtualization
- Task 2.4: Persistence
- Task 2.5: Resilience
- Task 2.6: Scalability
- Task 2.7: Security
- Task 2.8: Introspective Execution

FRACTALS FOR SYSTEM SUPPORT

High-level tasks broken down into fractals & then mapped onto appropriate compute elements

Fractal ISA combines properties of compiler intermediate representation with more conventional, microarchitecture-specific ISA



(numbers = tasks in Theme 2)

THEME 3 – MAKING THE PROGRAMMER'S LIFE EASY

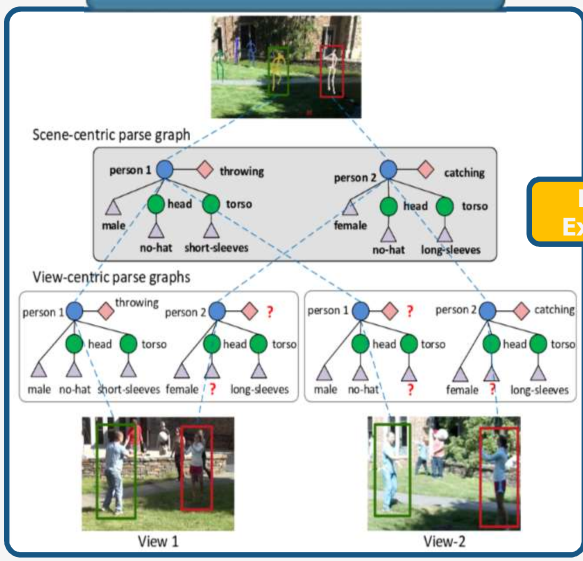
Relay **improvements in hardware** (Theme 1) & the **systems abstractions over hardware** (Theme 2) into benefits that can be **realized by "everyday" programmers in end-user applications**

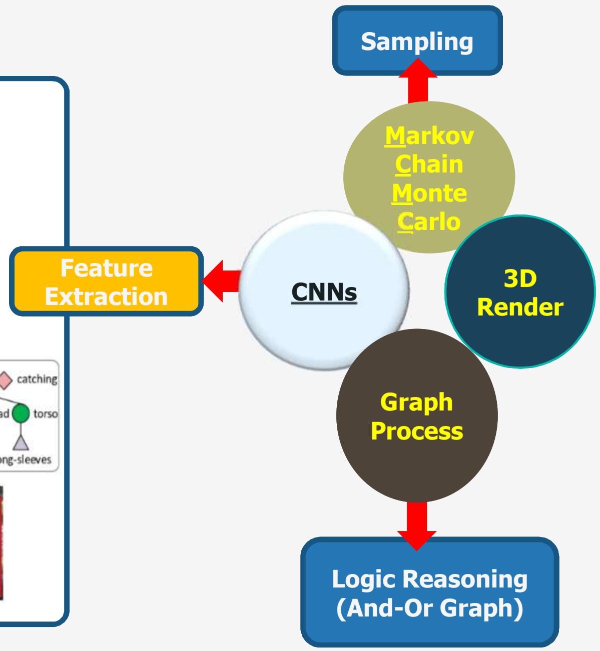
- Task 3.1: Programming framework
- Task 3.2: Big-data analytics
- Task 3.3: Video analytics
- Task 3.4: Medical Imaging
- Task 3.5: Genome, Microbiome, & Cognitive Wellness
- Task 3.6: Cognitive Architectures
- Task 3.7: Benchmarking

PARSING OF CROSS-VIEW VIDEOS & 3D RECONSTRUCTION AND QUESTION ANSWERING (TASK 3.3)

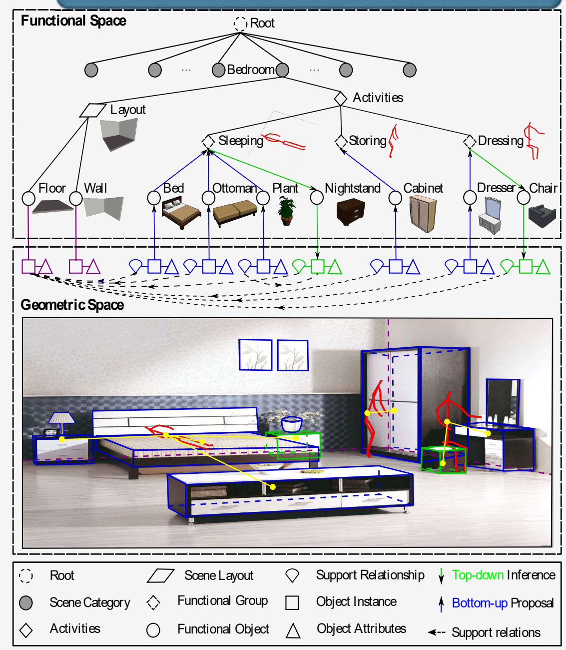
- Goal: Reconstruct, parse, and interpret 3D scenes
- Massive datasets, complex graph traversals
- Identify data bottlenecks and optimize

Video Analytic Framework





3D Reconstruction Framework



THE TEAM: 22 FACULTY, 10 UNIVERSITIES, $28M/5YRS

- Univ. of Virginia
  - Samira Khan
  - Kevin Skadron (Director)
  - Mircea Stan
- Cornell
  - José Martínez (Theme 1 Lead)
  - Zhiru Zhang
- Penn State
  - Vijay Narayanan
  - Anand Sivasubramaniam
- Wisconsin
  - Jing Li
  - Jignesh Patel (Theme 3 Lead)
  - Kevin Eliceiri
- UIUC
  - Wen-Mei Hwu
- Stanford
  - Christos Kozyrakis
- UC Santa Barbara
  - Dmitri Strukov
  - Yuan Xie (Associate Director)
- UC San Diego
  - Rob Knight
  - Tajana Rosing
  - Steve Swanson (Theme 2 Lead)
  - Jishen Zhao
  - Yuanyuan Zhou
- UCLA
  - Jason Cong
  - Song-Chun Zhu
- Univ. of Washington
  - Luis Ceze

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**