Semiconductor Research Corporation

CoCoSys
CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS

CoCoSys Center Overview : Year 2 Annual SAB Meeting

Arijit Raychowdhury
Anand Raghunathan
Anca Dragan
Azad Naeemi
Bruno Olshausen
Jae-sun Seo
James DiCarlo
Jan Rabaey
Josh Tenenbaum
Kaushik Roy
Larry Heck
Michael Carbin
Naresh Shanbhag
Priya Panda
Priyanka Raina
Sumeet Gupta
Tajana Rosing
Tushar Krishna
Vijay Raghunathan
Yingyan (Celine) Lin
Yu (Kevin) Cao

AI Challenges

- Narrow (specific to task or input modality)
- Lack of explainability and transparency
- Lack of robustness
- Scaling depends on larger models and datasets
- Algorithms driven by today’s hardware (GPUs and digital accelerators)

Why CoCoSys?

Current AI Systems

- Black-box (not explainable or interpretable)
- Reliant on large datasets, networks and compute
- Mostly monolithic CMOS technology

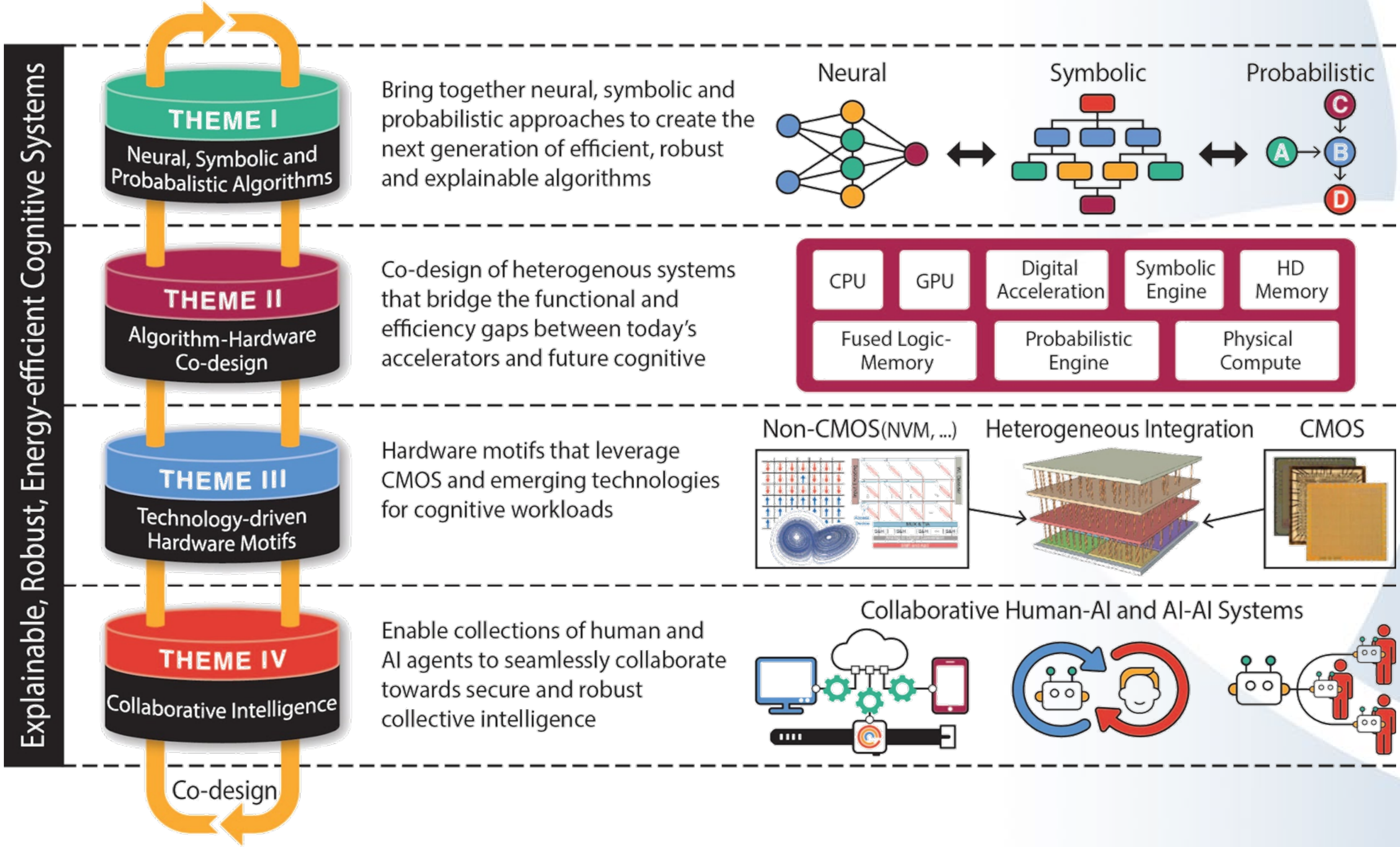
GRAND CHALLENGES

- Can we stem the unsustainable trends in compute requirements for AI?
- Can a fusion of neural, symbolic and probabilistic methods lead to more scalable, robust and explainable AI?
- Can cognitive algorithms perform the entire gamut of tasks involved in collaborative AI systems (perception, reasoning and decision making)?
- Can cross-layer design of cognitive algorithms and hardware improve energy efficiency by over 100X?

Future Collaborative AI Systems

- Seamless human-AI and AI-AI collaboration
- Explainable, robust and secure
- Hardware and algorithms co-designed to optimize energy efficiency, latency and throughput
- Leverage future logic, memory and integration technologies

Center Overview and Themes



Explainable, Robust, Energy-efficient Cognitive Systems

THEME I – Neural, Symbolic and Probabalistic Algorithms

Bring together neural, symbolic and probabilistic approaches to create the next generation of efficient, robust and explainable algorithms

THEME II – Algorithm-Hardware Co-design

Co-design of heterogenous systems that bridge the functional and efficiency gaps between today's accelerators and future cognitive

THEME III – Technology-driven Hardware Motifs

Hardware motifs that leverage CMOS and emerging technologies for cognitive workloads

THEME IV – Collaborative Intelligence

Enable collections of human and AI agents to seamlessly collaborate towards secure and robust collective intelligence

Neural + Vector Symbolic Architectures

Computational Models | Theory | Algorithms

Neural Network
Scalable, Flexible,
Handle inconsistency

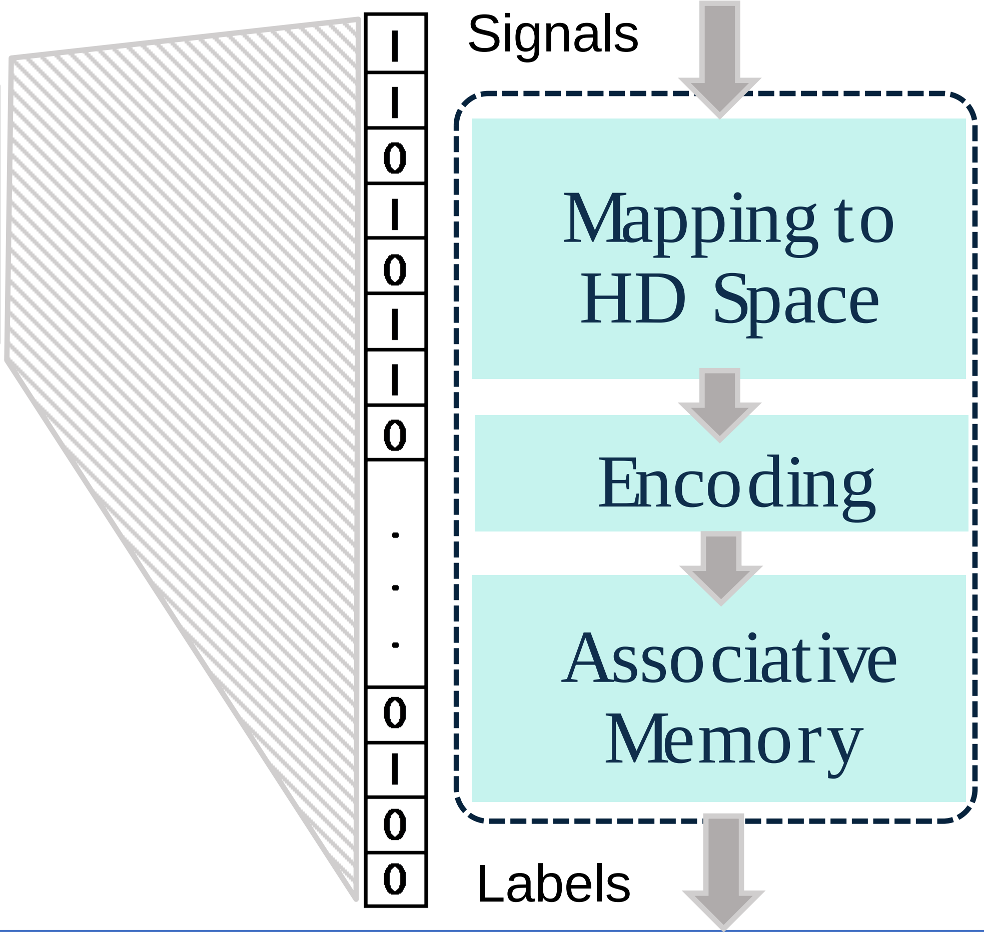
+

Symbolic
Interpretable, Explainable,
Data-efficient

+

Probabilistic
Robust to
uncertainty

- 3D scene perception
- LMMs
- Digital Assistants
- Conversational AI
- AI-AI and AI-Human Collaboration



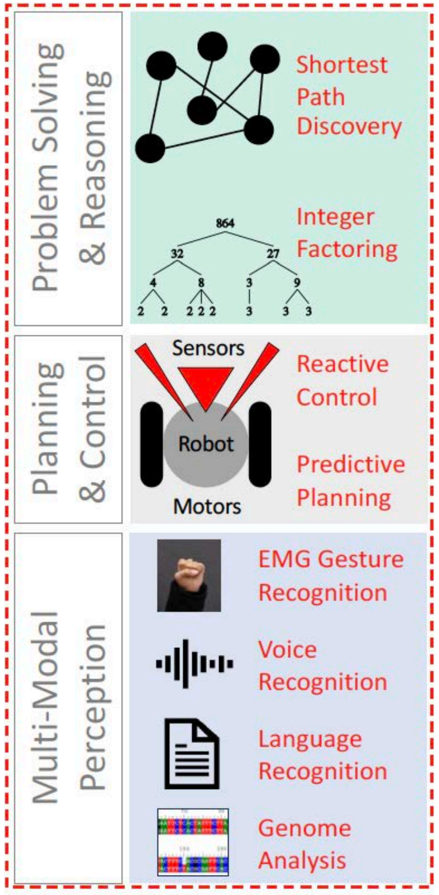
Signals

Mapping to HD Space

Encoding

Associative Memory

Labels



Problem Solving & Reasoning

Shortest Path Discovery

Integer Factoring

Planning & Control

Reactive Control

Predictive Planning

Multi-Modal Perception

EMG Gesture Recognition

Voice Recognition

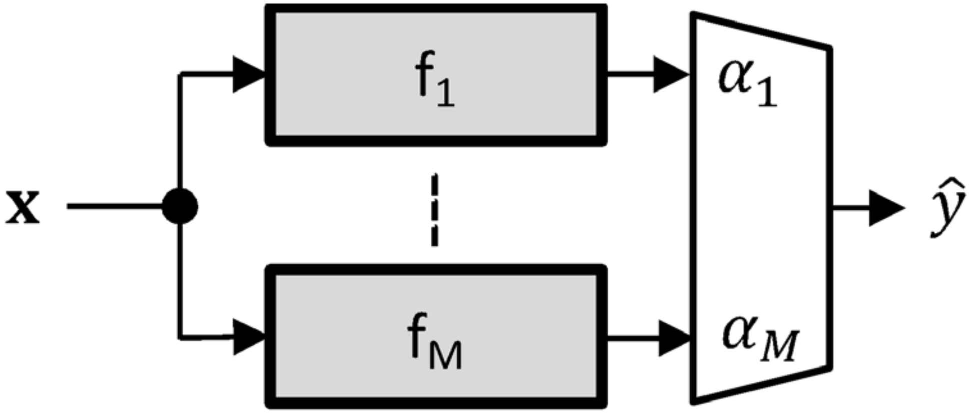
Language Recognition

Genome Analysis

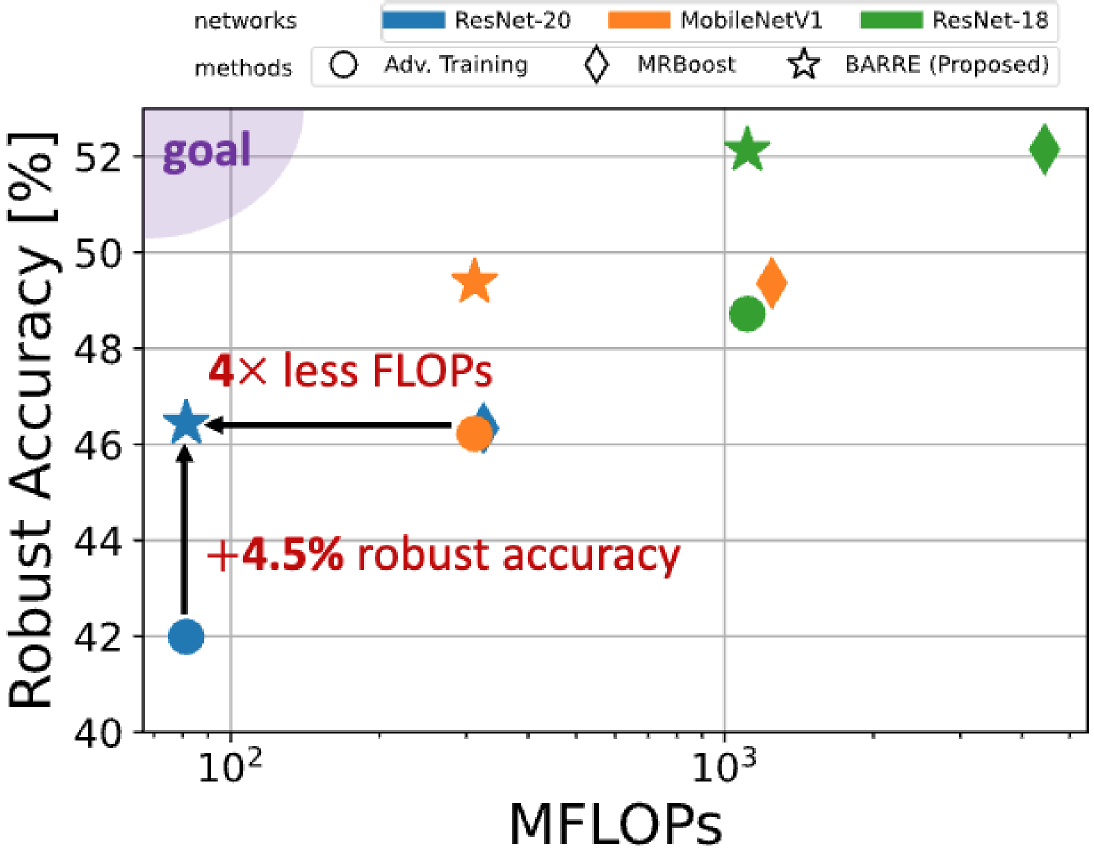
The Theory of Robustness and Beyond

Computational Models | **Theory** | Algorithms

Fundamental limits on accuracy-robustness-efficiency (ARE) trade-offs



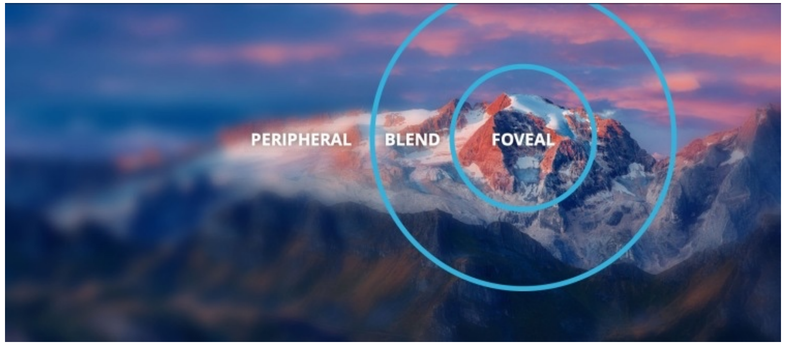
Randomized Ensemble of Classifiers (RECs)



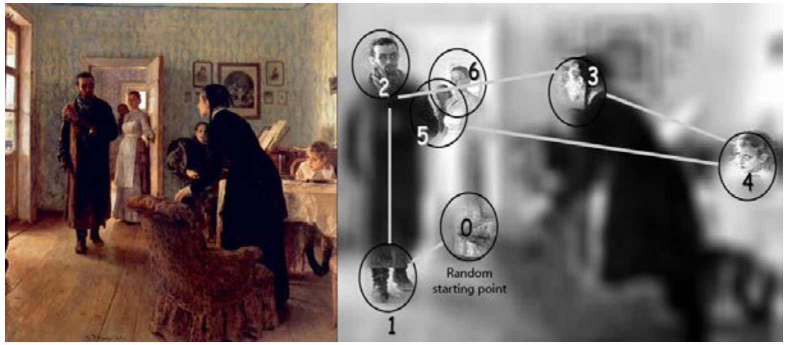
Improving Vision Systems through Foveation and Saccades

Computational Models

Foveation (variable resolution)

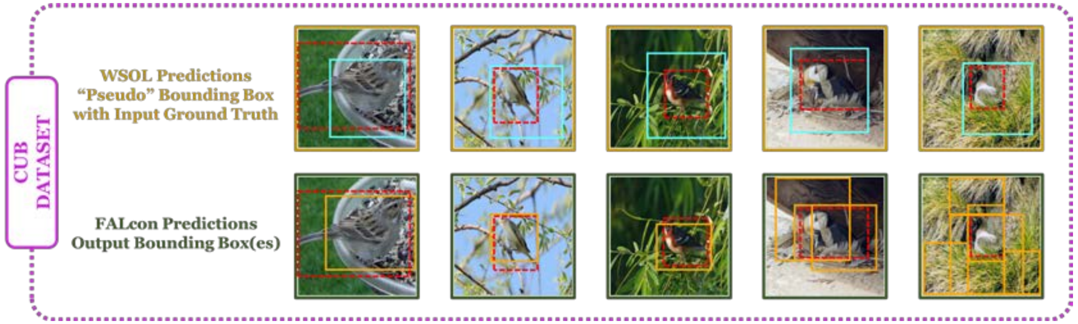


Saccade (quick eye movement)

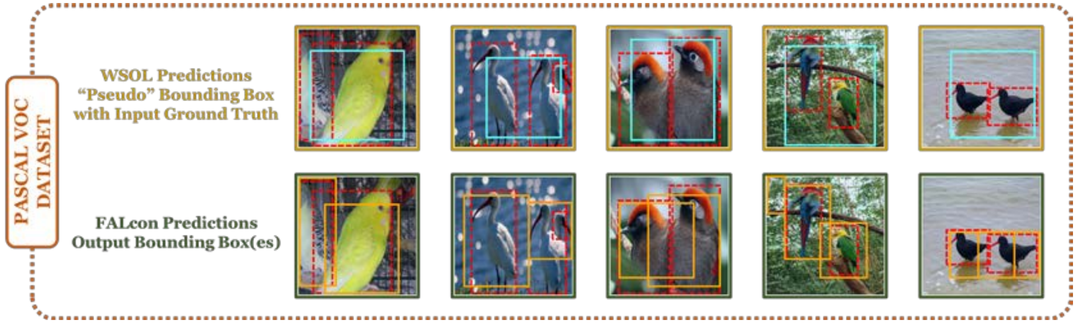


Theory

- **Advantage 1:** more accurate bounding boxes:

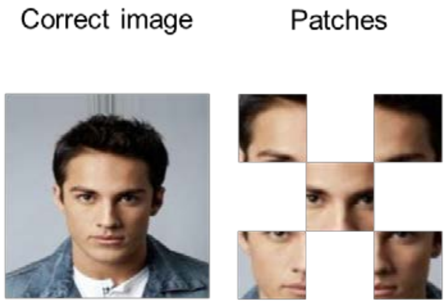

CUB DATASET
WSOL Predictions "Pseudo" Bounding Box with Input Ground Truth
FALcon Predictions Output Bounding Box(es)

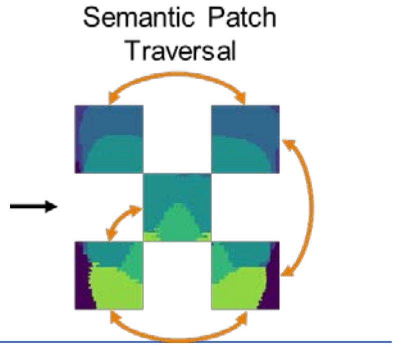
- **Advantage 2:** more resilient localization pipeline,


PASCAL VOC DATASET
WSOL Predictions "Pseudo" Bounding Box with Input Ground Truth
FALcon Predictions Output Bounding Box(es)

Algorithms

- **Advantage 3**: Can define Image grammer (semantics & syntax)

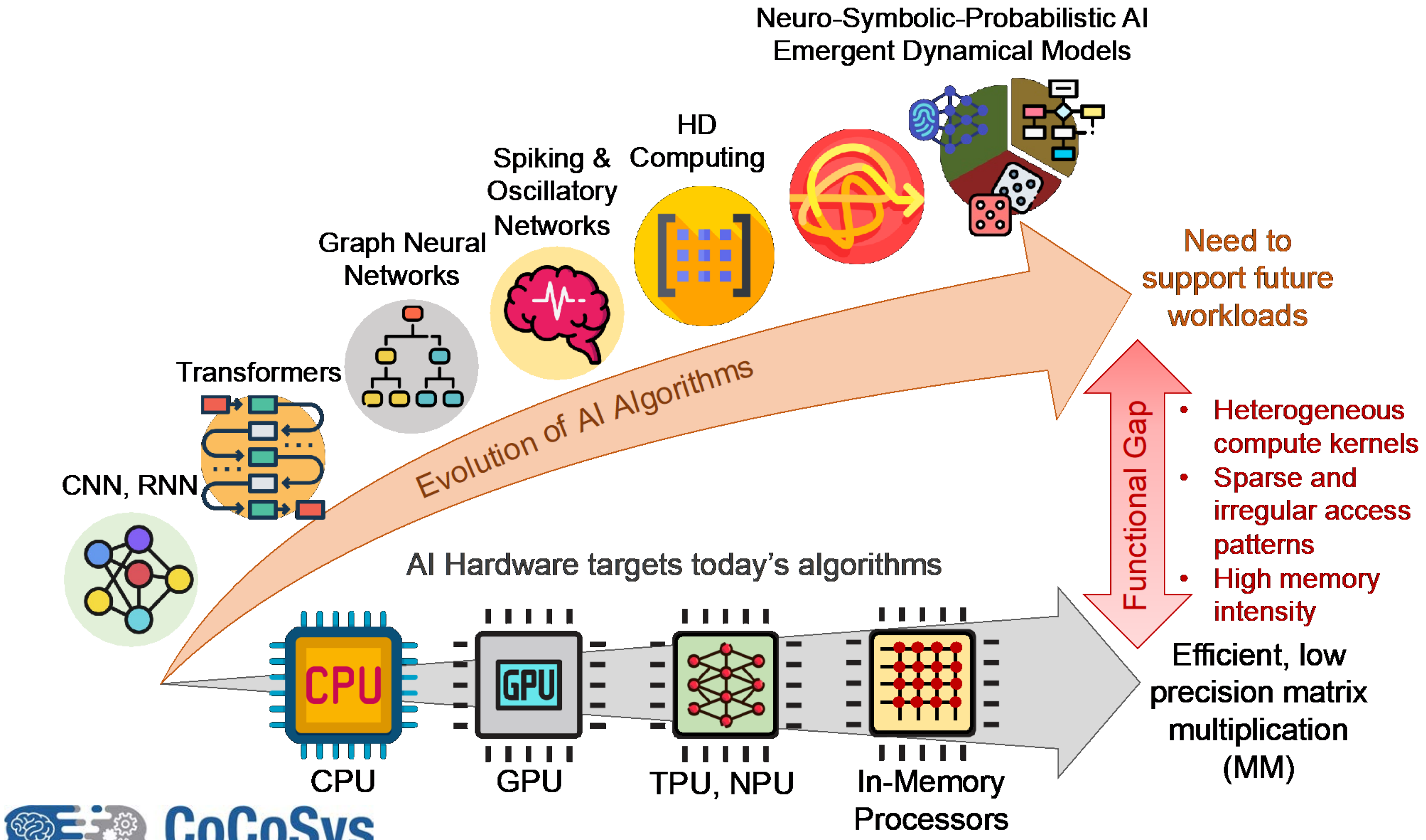




SRC

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Needs of Future Neuro-Symbolic-Probabilistic Workloads



Neuro-Symbolic-Probabilistic AI
Emergent Dynamical Models

HD Computing

Spiking & Oscillatory Networks

Graph Neural Networks

Transformers

CNN, RNN

Evolution of AI Algorithms

Need to support future workloads

Functional Gap

- Heterogeneous compute kernels
- Sparse and irregular access patterns
- High memory intensity

AI Hardware targets today's algorithms

CPU

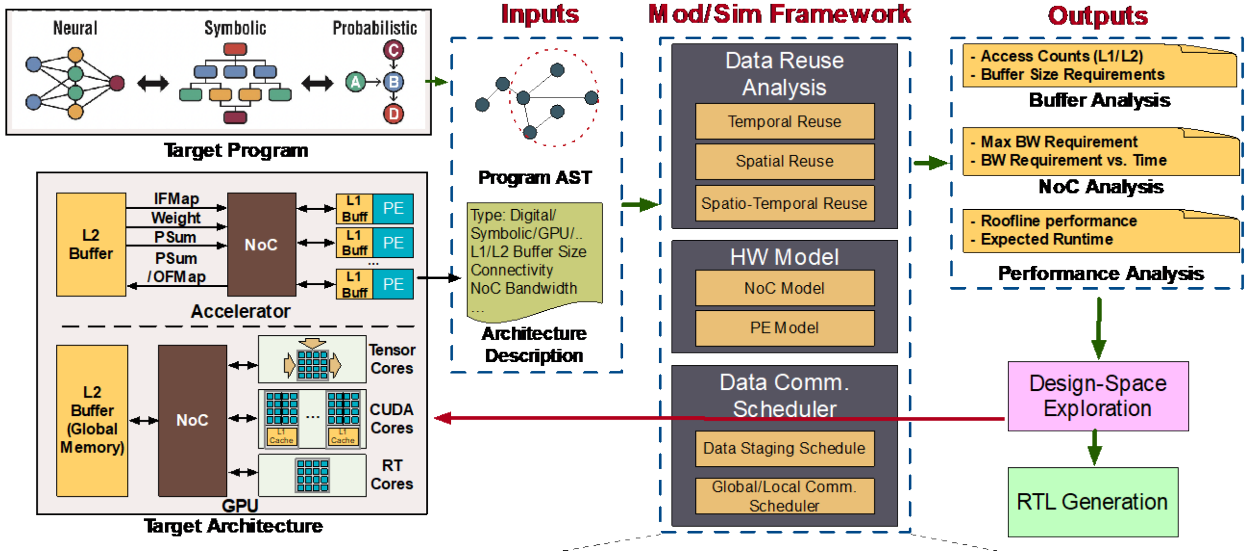
GPU

TPU, NPU

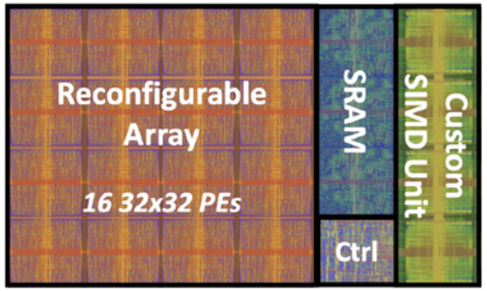
In-Memory Processors

Efficient, low precision matrix multiplication (MM)

Algorithm-Hardware Co-design



Layout of Proposed CogSys Accelerator



Accelerator Specs

Technology	28 nm	Frequency	800 MHz
#Reconfi-gurable PEs	16384	Voltage	1 V
#SIMD PEs	512	Power	1.18 W
SRAM	4.5 MB	Area	4.0 mm^2

- Open-source tools and tool-chains for system exploration within CoCoSys

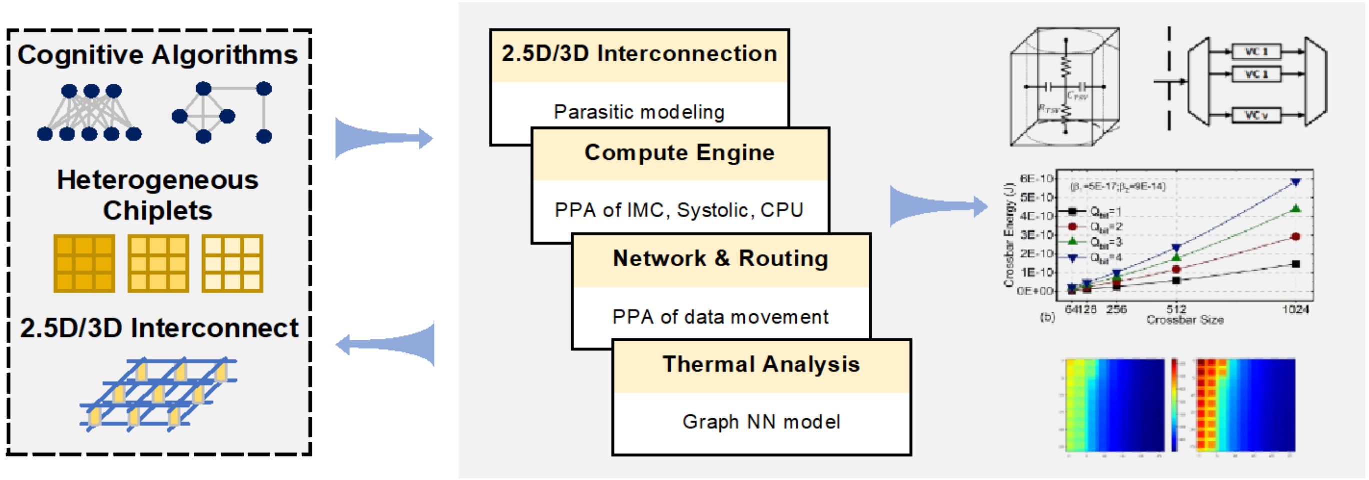
- Industry collaborations (joint papers, joint conference sessions)

- Hardware artifacts to quantify system benefits

- **Innovation:** First effort towards algorithms-to-hardware co-design of Neuro-symbolic-probabilistic AI systems
- **Key Result**: 2-3 orders of magnitude faster & more energy efficient than CPUs and GPUs

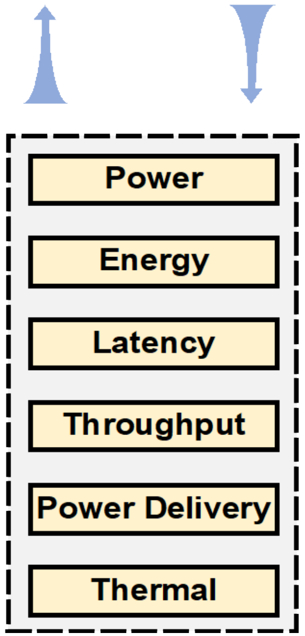
SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

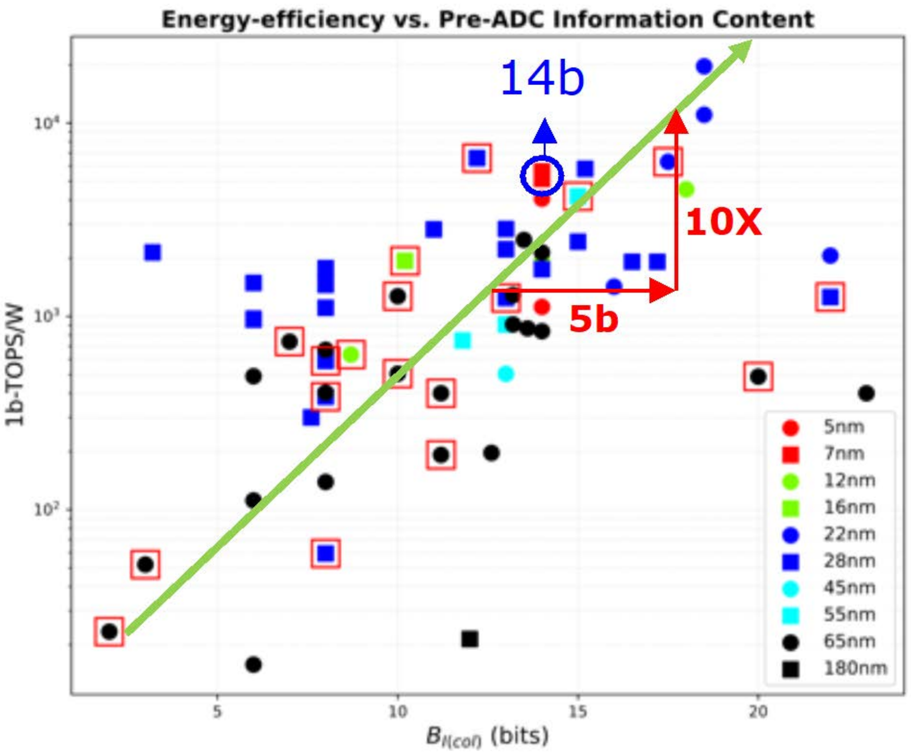
System Design with Advanced and Emerging Technologies



Heterogeneous Integration Simulator with Interconnect Modeling (**HISIM**)

- 2.5D/3D interconnection, in-memory computing chiplets, network-on-packaging, thermal
- <u>Analytical performance models</u> that are **10^4x-10^6x faster** than NeuroSim and other SOTAs

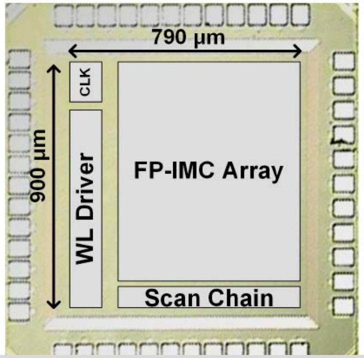


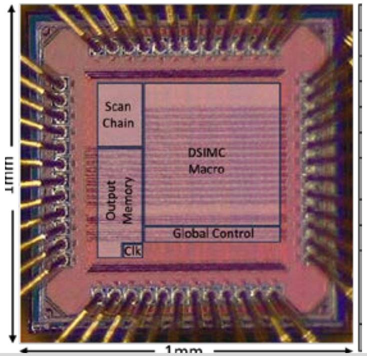


IMC benchmarking tool

Reducing Data-movement through In-X Computing

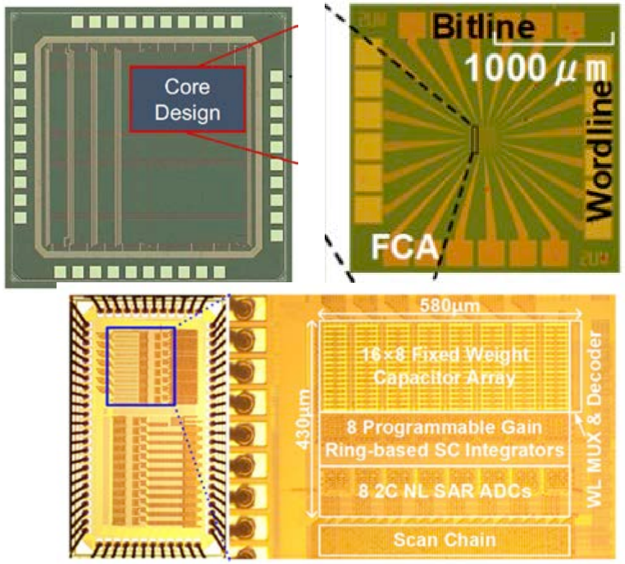
SRAM





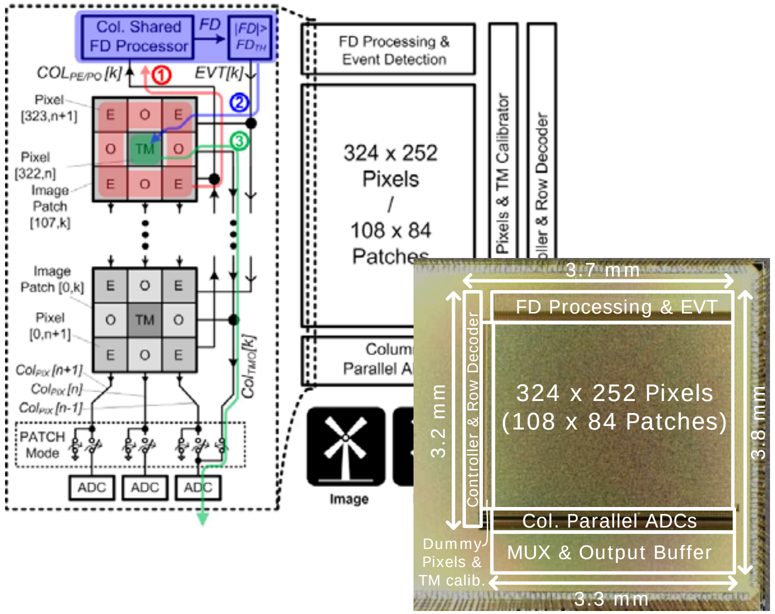
- FP-IMC: TSMC 28nm digital floating-point IMC macro chip (ESSCIRC'23)
- SP-IMC: TSMC 28nm digital sparsity-integrated IMC macro w/ compressed computing (CICC'24)
- IMC w/ delta-sigma modulator for variable input precision (CICC'23, SSCL'23) Multi-step cap.-coupling IMC SRAM Macro in 28nm (JSSC'24)
- Accurate/ Approximate CAM (TBP)

NVM



- 65nm RRAM for genome sequencing
- 180nm IMC macro chip for NVM ferroelectric capacitor array with PoT ADC (SSCL'23)
- 40nm RRAM VLIW processor for edge inference and robot manipulation

Sensor

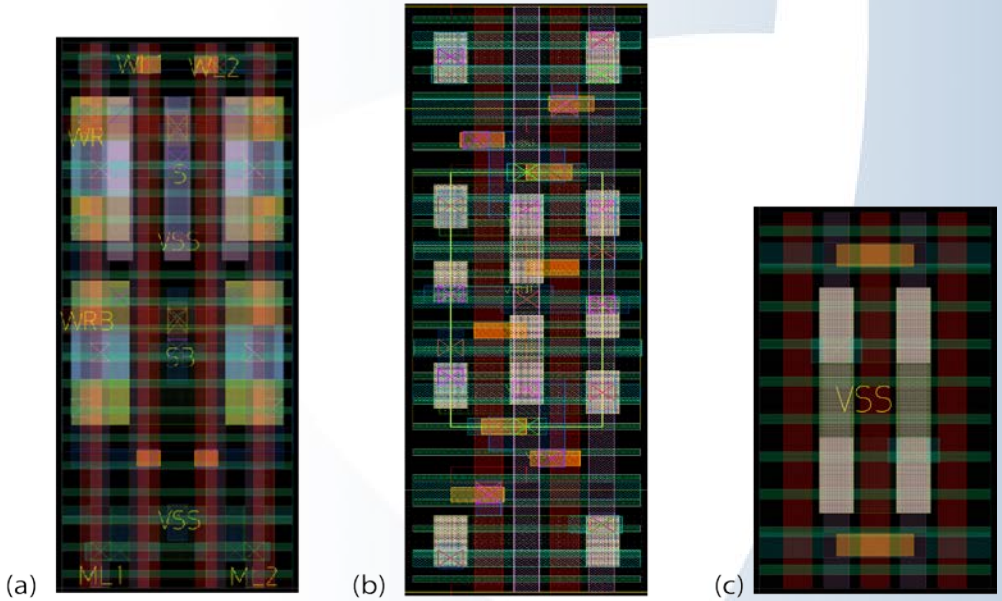


- Time-memory-based CMOS vision sensor w/ in-pixel temporal derivative comp. (ESSCIRC'23)
- Multi-mode: image sensor, event, temporal deriv.

Tool-chain for DTCO in the Context of AI Workloads

Cross-layer Modeling and Design of 7nm PDK for DTCO

- Circuit parasitics extracted from SOT, SRAM, and FeFET-CAM layouts based on ASAP7 PDK using state of the art EDA tools
- Parasitics used in SPICE simulations to extract ML discharge delays
- Interconnect parasitics – IR drop and RC delay degrade similarity search performance for larger array sizes
- Two solutions explored – using wider search lines (S2x) and matching clk delay (Clk match)



(a) SOT (b) SRAM and (c) FeFET CAM cell layout

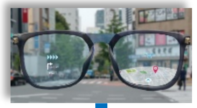
Human-AI Interactions: Conversational Agents

Conversational Expression

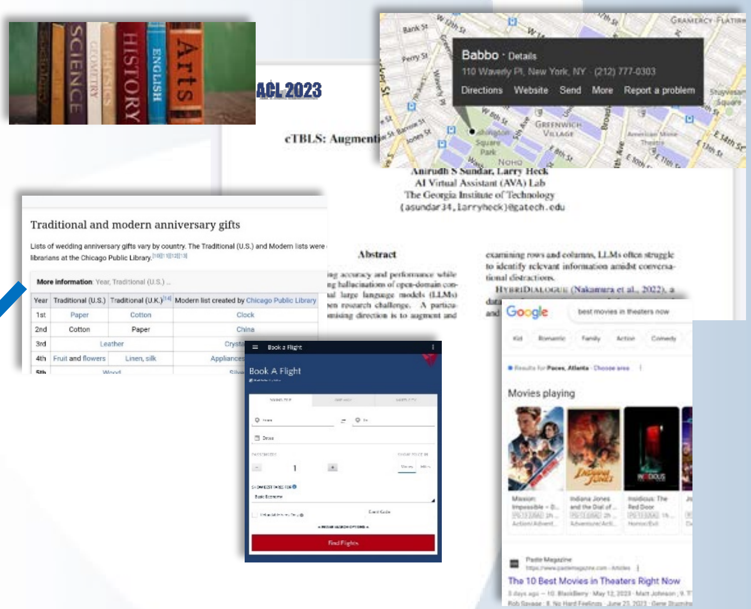
- Neural rendering
- Large Body Language Models (LBLM)
- Large Face Language Models (LFLM)

Conversational Vision

- Visual Dialogue
- Visual QA (VQA)
- AR/VR



Conversational Content



ACL 2023

cTBLS: Augmenti

Anirudh S Sundar, Larry Heck
AI Virtual Assistant (AVA) Lab
The Georgia Institute of Technology
{asundar34,larryheck}@gatech.edu

Abstract

ing accuracy and performance while ing hallucinations of open-domain con- al large language models (LLMs) en research challenge. A particu- mising direction is to augment and

examining rows and columns, LLMs often struggle to identify relevant information amidst conversational distractions.

HybriDialogue (Nakamura et al., 2022), a

Babbo · Details
110 Waverly Pl, New York, NY · (212) 777-0303
Directions Website Send More Report a problem

Traditional and modern anniversary gifts

Lists of wedding anniversary gifts vary by country. The Traditional (U.S.) and Modern lists were librarians at the Chicago Public Library.[10][11][12][13]

More information: Year, Traditional (U.S.) ...

Year	Traditional (U.S.)	Traditional (U.K.)[14]	Modern list created by Chicago Public Library
1st	Paper	Cotton	Clock
2nd	Cotton	Paper	China
3rd	Leather		Crysta
4th	Fruit and flowers	Linen, silk	Appliances
5th	Wood		Silve

Book A Flight

1

Google best movies in theaters now

Kid Romantic Family Action Comedy

Movies playing

Mission: Impossible – D...

Indiana Jones and the Dial of ...

Insidious: The Red Door

The 10 Best Movies in Theaters Right Now

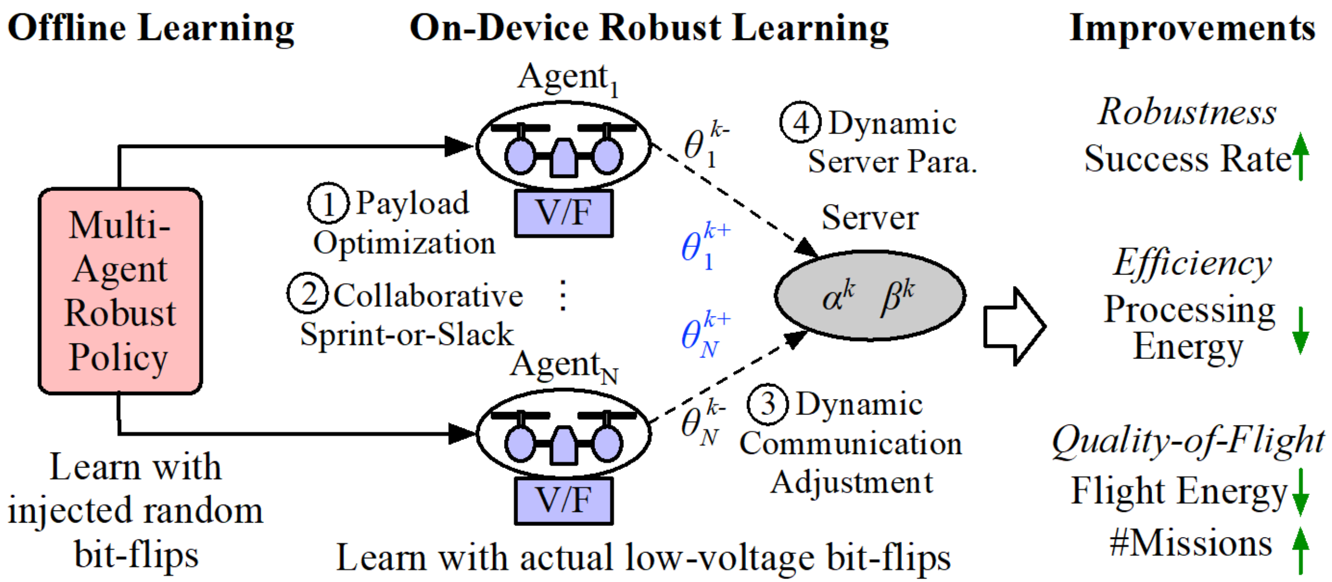
- GUI links
- Fields
- Lists
- Forms
- Tables
- Equations
- Plots
- Figures
- Maps
- Text (web, books, papers)
- etc…

Conversational Model

$$\arg\max_{r} \; p(r|m, u, c, s, d, \ldots)$$

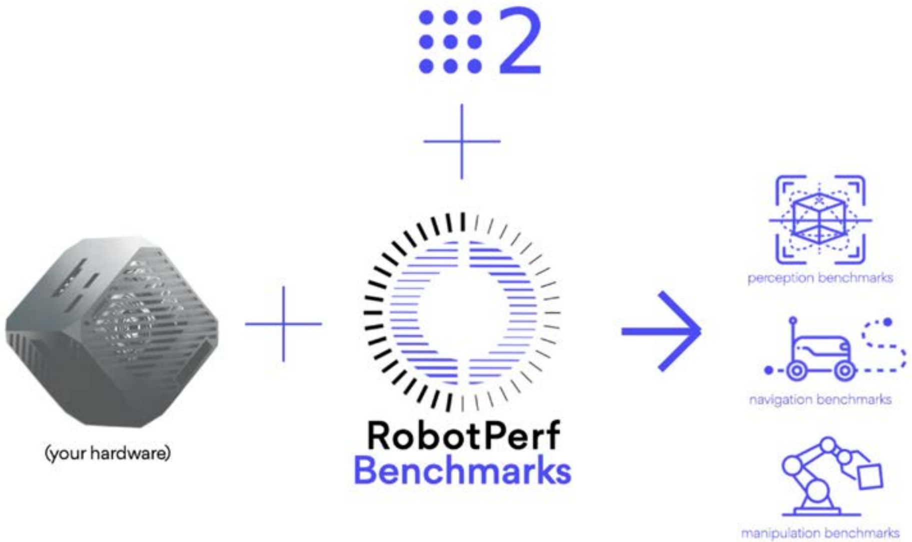
SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

AI-AI Interactions: Autonomous UAV Swarms



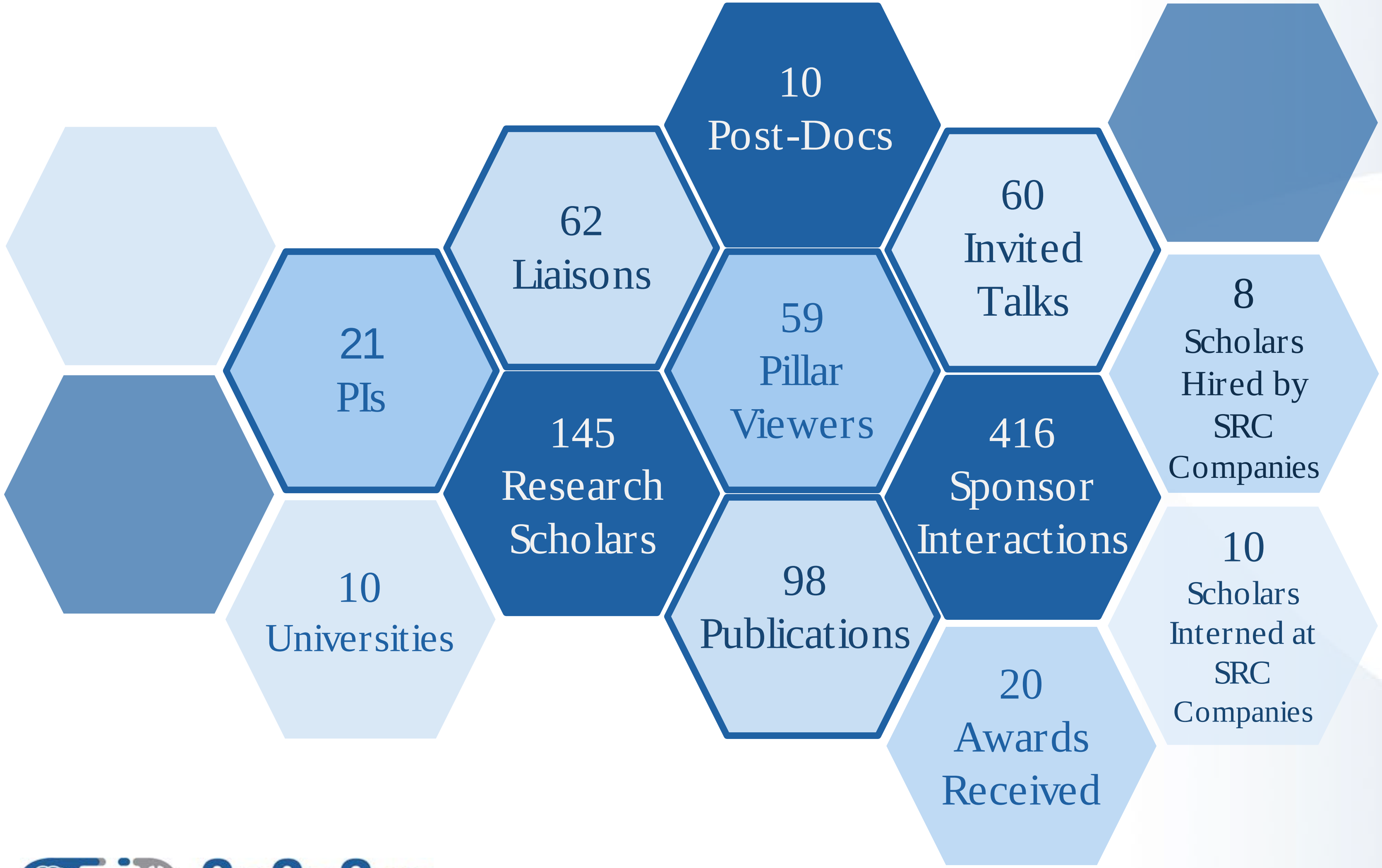
Joint work with IBM

- 18.9% reduction in UAV flight energy
- 22.1% increase in number of successful missions
- 4.07x reduction in processing energy
- Generalize across chips, voltages, UAV numbers, and autonomy policies



- First benchmark suite for evaluating robotic computing system performance.
- Usage across academia (Harvard, GT, CMU, Boston Univ, Columbia Univ, etc) and industry (Intel, Ford, AMD, etc). 123 GitHub stars.

CoCoSys at a Glance



21 PIs

10 Universities

62 Liaisons

145 Research Scholars

10 Post-Docs

59 Pillar Viewers

98 Publications

60 Invited Talks

416 Sponsor Interactions

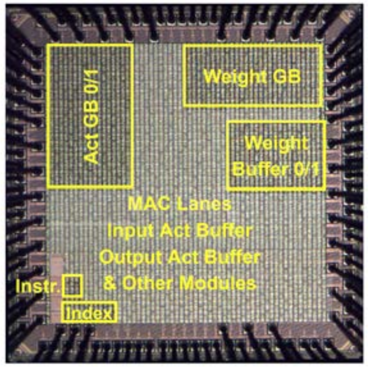
20 Awards Received

8 Scholars Hired by SRC Companies

10 Scholars Interned at SRC Companies

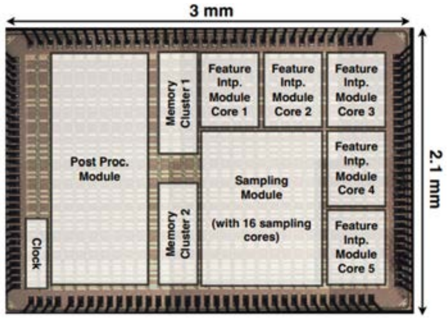
Updated March 2024

CoCoSys Hardware Gallery

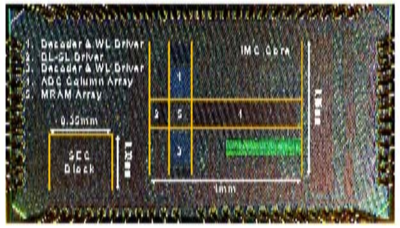


28nm eye tracking
PI: Celine Lin (ISCA'22)

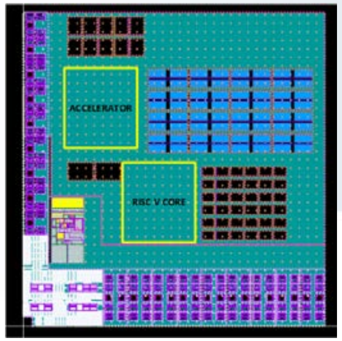
28nm 3D reconstruction
PI: Celine Lin (ISCA'23)


3 mm
2.1 mm
Post Proc. Module
Clock
Memory Cluster 1
Memory Cluster 2
Feature Intp. Module Core 1
Feature Intp. Module Core 2
Feature Intp. Module Core 3
Feature Intp. Module Core 4
Feature Intp. Module Core 5
Sampling Module
(with 16 sampling cores)

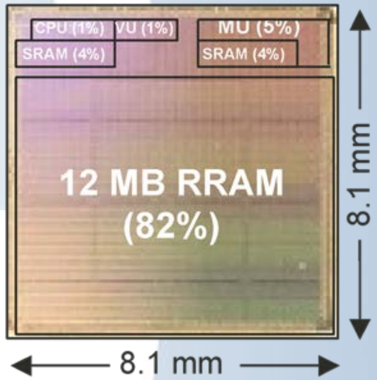
28nm NeRF acc.
PI: Celine Lin



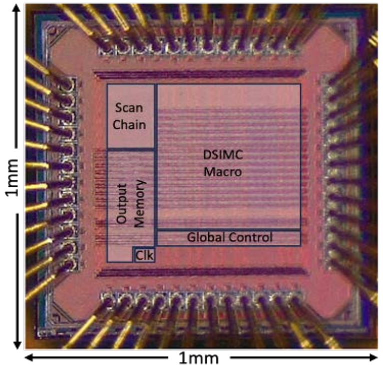
28nm MRAM IMC
PI: N. Shanbhag w/ Raytheon



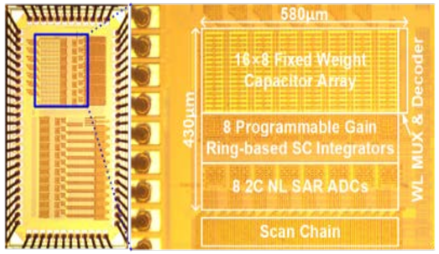
16nm adaptive GNN
PI: Yu Cao



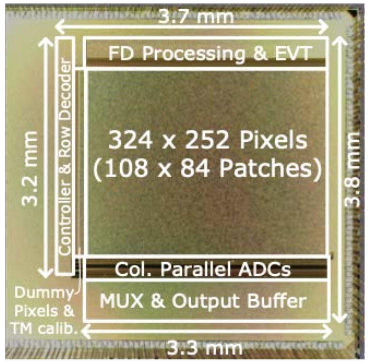
40nm Transformer
PI: Priyanka Raina



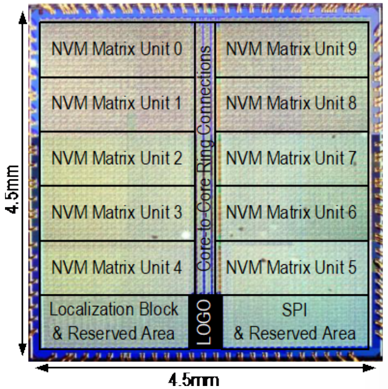
28nm sparsity IMC
PI: Jae-sun Seo (CICC'24)


580µm
430µm
16×8 Fixed Weight Capacitor Array
8 Programmable Gain Ring-based SC Integrators
8 2C NL SAR ADCs
Scan Chain
WL MUX & Decoder

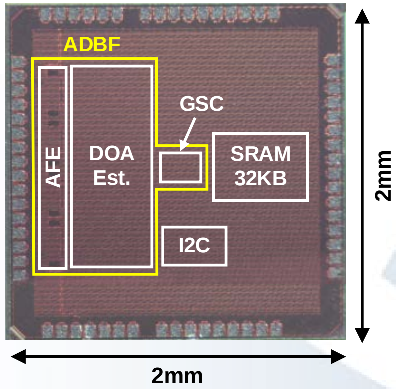
180nm Ferro-Cap IMC
PI: Jae-sun Seo (SSCL'24)
w/ CHIMES center



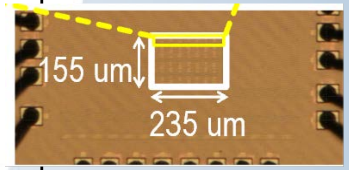
180nm vision sensor
PI: Jae-sun Seo & Yu Cao
(ESSCIRC'23)



40nm VLIW SoC
PI: Raychowdhury
w/ TSMC (ISSCC'24)

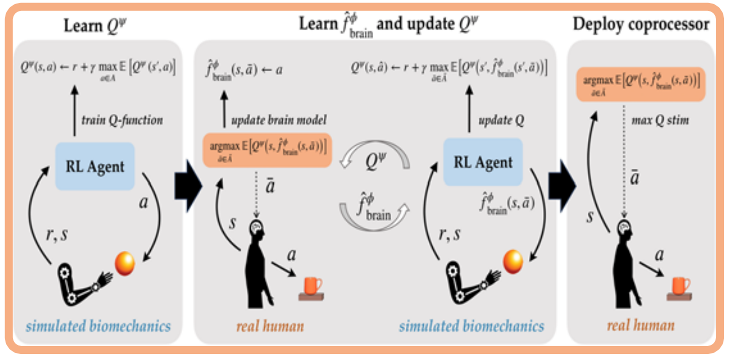


65nm beamformer
PI: Raychowdhury
(VLSI'23)

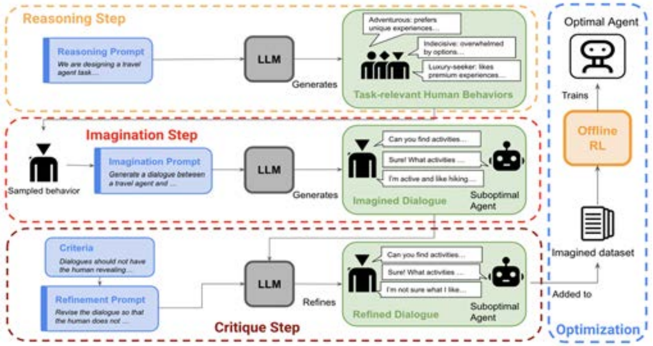


28nm SRAM
Complex CIM
PI: Raychowdhury
(Submitted)

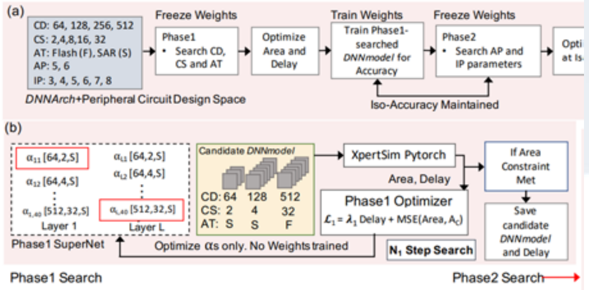
CoCoSys Software Artifact Gallery



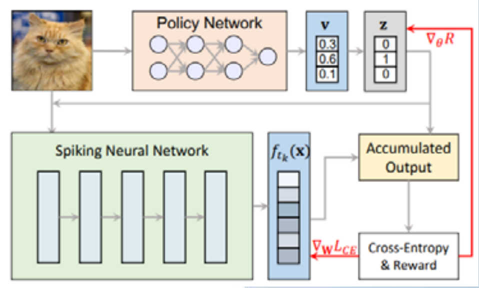
Model-based RL for brain simulation
PI: Anca Dragan



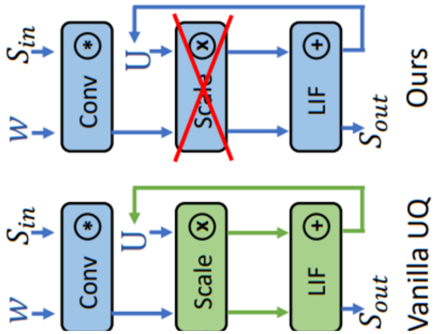
Offline RL for dialogue agent
PI: Anca Dragan



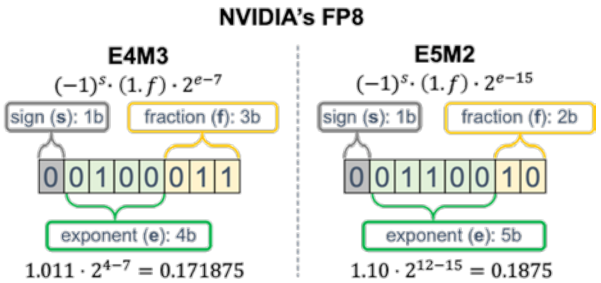
Xpert – network circuit co-search
PI: Priya Panda (DAC’23)



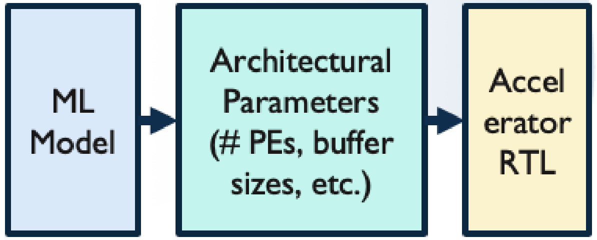
Dynamic timestep SNN
PI: Priya Panda
(DAC’23, NeurIPS’23)



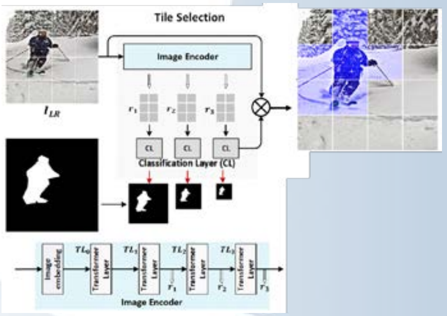
MINT – SNN quantization
PI: Priya Panda
(ASP-DAC’24)



Transformer quantization
PI: Priyanka Raina

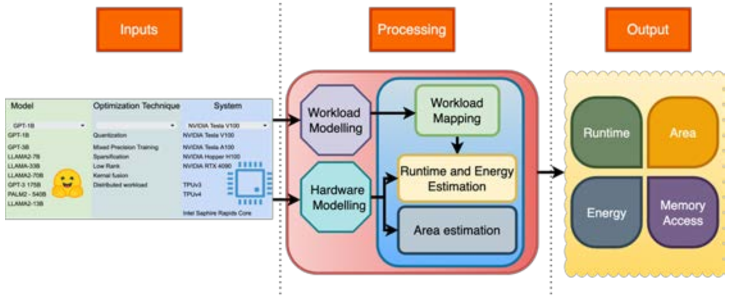


NN accelerator generator
PI: Priyanka Raina

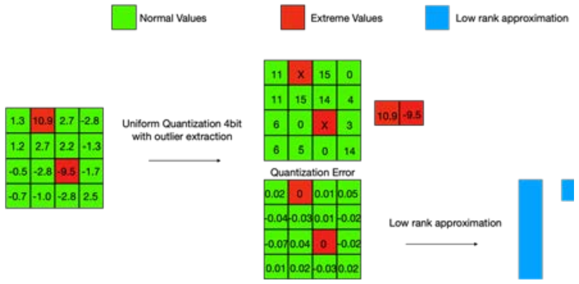


3D in-sensor computing
PI: Yu Cao
(WACV’24, AAAI’24)

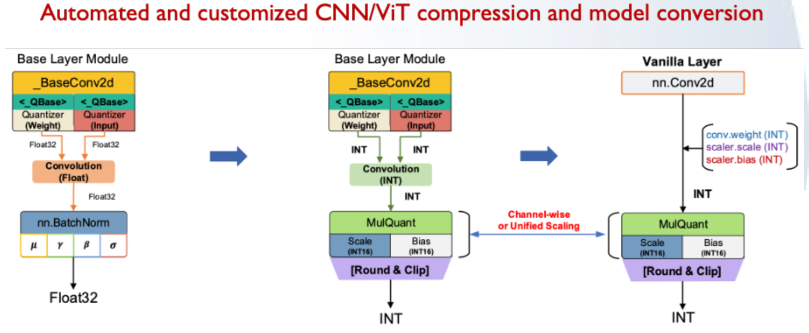
CoCoSys Software Artifact Gallery



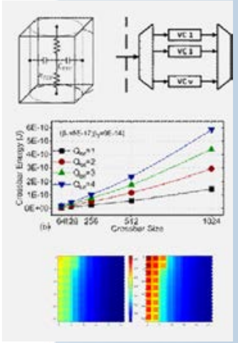
Rapid DSE for LLM
PI: Tushar Krishna



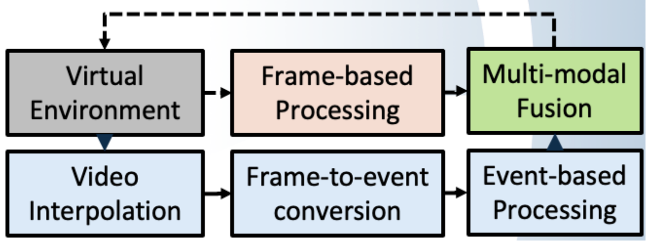
LLM KV Cache compression
PI: Tushar Krishna (w/ Intel)



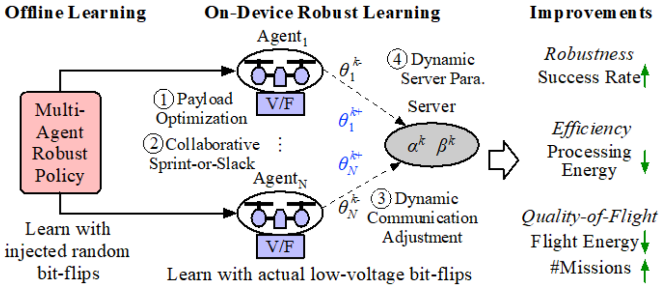
Automated CNN/ViT compression
PI: Jae-sun Seo



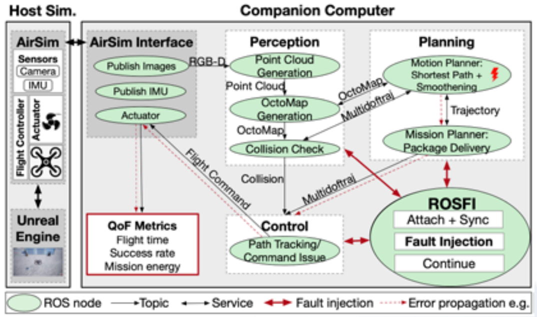
2.5D/3D benchmark
PI: Yu Cao
(ASP-DAC'24)



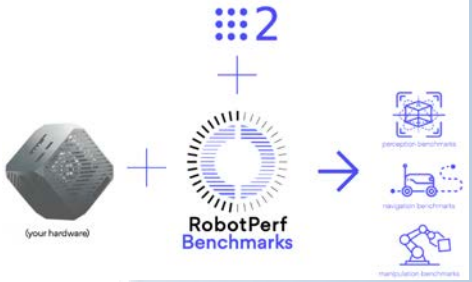
Event-based drone simulation
PI: Arijit Raychowdhury
(Frontiers in NeuroSci'24)



Swarm drone optimization
PI: Arijit Raychowdhury (w/ IBM)
(DAC'23, ASPLOS'24)



AutoSys reliability analysis
PI: Arijit Raychowdhury
(TCAD'23, Comm of ACM'24)



Robotic benchmarking
PI: Arijit Raychowdhury
(ICRA'24)

Software and Hardware Artifacts

https://drive.google.com/file/d/1uivIeDm1ClUjA2O4rZLmcW0YuYvqwgfK/view?usp=sharing

Software Artifacts

https://drive.google.com/file/d/17-wh_sf_Jf72Kc91GJcLMA6dir4REkhR/view?usp=share_link

Hardware Artifacts

ACADEMIA • INDUSTRY • GOVERNMENT

CoCoSys

CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS