Semiconductor Research Corporation

Mission: Driving Semiconductor Innovation

Vision: Neutral, Trusted, and Science-Driven

Values: Prosperity, People, and the Planet

JUMP 2.0 Virtual SAB – End of Year 1

December 5, 2023

Roman Caudillo
JUMP 2.0 and JUMP Director

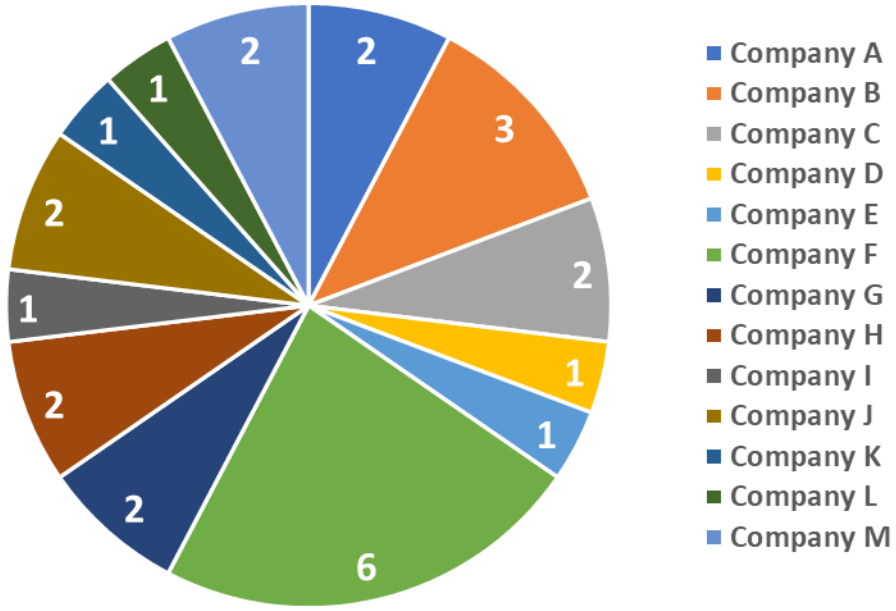
Adam Knapp
JUMP 2.0 and JUMP Program Manager

Tameka Bell
JUMP 2.0 Program Coordinator

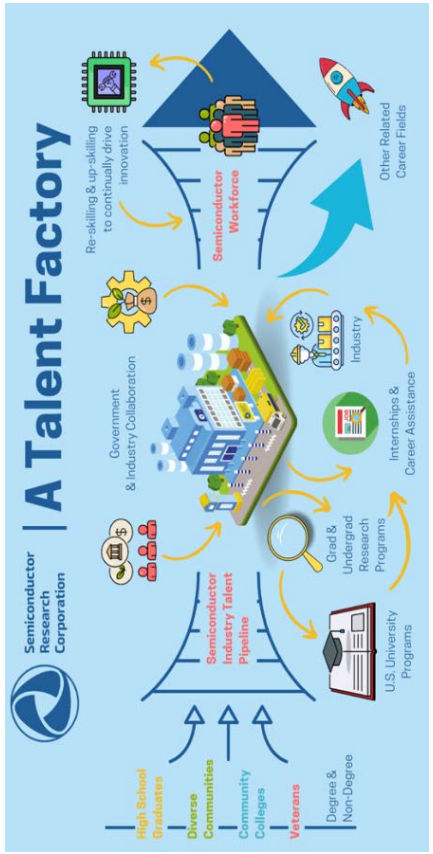
Holiday Greetings!

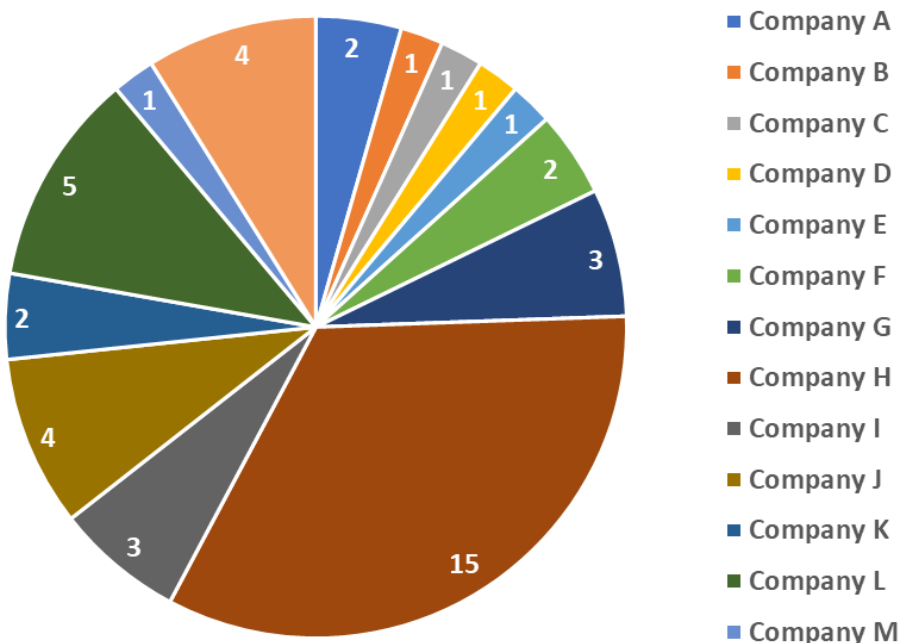
What has JUMP 2.0 delivered?

71 Full-time Hires and Interns for JUMP 2.0 for Year 1



26 FT Hires in 2023





45 Interns in 2023

SRC produces ~20% of the Semiconductor Industry's PhDs
We want to grow our number of hires by ~3x by 2030!

SRC's Broadening Participation Pledge

https://www.src.org/about/broadening-participation/

2030 Broadening Participation Pledge

"Throughout the decade, as SRC defines, selects, and manages its research and education programs, we will look to grow our student base, establish a balanced mix of bachelors, masters, and Ph.D.-level initiatives, and create a more diverse and inclusive community."

CoCoSys	CUbiC	CogniSense	ACE	PRISM	CHIMES	SUPREME
Yan Fang (Kenessaw State)	Michal Lipson (Columbia U.)	Tingyi Gu (U. Delaware)	Mohammad Alian (U Kansas)	Niema Moshiri (UCSD)	Zhiting Tian (Cornell)	Luqiao Liu (MIT)
	Alex Meng (Columbia U.)	Devon McLaurin (GIT), Lorry Chang (UD)	Charith Mendis (UIUC)		Rohit Sharma (Penn State)	Judy Cha (Cornell)
Pledge link	Pledge link	Pledge link	Pledge link	Pledge link	Pledge link	Pledge link

Each JUMP 2.0 Center has:

- Proposed Center commitments
- A public Center pledge
- Committed Center BP champion(s)
- The support of SRC, DARPA, and Industry

PRISM Broadening Participation Projects BP Champion: Prof Niema Moshiri, UCSD

- Students working on PRISM demos and projects at UCSD this summer
  - 20 undergraduates: 4 URM, 11 female students
    - **PRISM review has 7 posters and 5 demos with undergraduates**
  - 8 high school students: 5 URM, 4 female students
- Travel grants to PRISM annual review for undergraduates
- Symposia with SRC representatives
  - In-person during PRISM annual review (November)
  - Virtually via research symposium (Spring)
- Partnering with diversity-focused organizations at local ins
  - NSBE, SHPE, SWE, WIC, etc.; many universities already hav
- Working with mentorship programs at PRISM Universities,
  - Early Research Scholars Program (ERSP), ENLACE, McNair @UCSD;
  - Computer Science Summer Institute, AI Breakthrough Program @UCLA;
  - Scalable Asymmetric Lifecycle Engagement (SCALE) – a preeminent U.S. program for semiconductor workforce development in the defense sector @Gatech

Forge connections to new NSF REU Students in 2024

NSF and SRC invest in 6 sites for undergraduate research experiences in semiconductors

- **The first class of SRC-NSF REU (6 schools) each received 3-year awards ($2.2M total)**
  - Carnegie Mellon
  - U. Rochester
  - U. Dayton
  - U. Michigan
  - UT-Dallas (start 2024)
  - U. Penn (start 2024)
- **Summer of 2024 will support ~140 students**

Image Credit: Kelly Pena, Boston University Photonics Center

Ideas for Connecting:

- Invite students to attend virtual SRC Career Fair
- Virtually attend JUMP 2.0 eWorkshops during their REU
- Attend nearby annual reviews that are held in the summer (SUPREME, CUBIC, CHIMES)
- Present Posters at TECHCON

(1) Forge connections to 6 already announced sites from 2023
(2) Forge new connections to the ~8 new sites that will be awarded in 2024

250 JUMP 2.0 Technology Transfers for Year 1 (2023)

- **KPI collection & verification has *changed* for JUMP 2.0.**
- All KPI forms are uploaded to their JUMP 2.0 projects on Pillar and filled out by the academics before the center annual review.
- Liaisons then have 4 weeks after the annual review to object to their company's purported technology transfers and hires
- If ***no*** objection occurs, that task is considered approved.
- Example from CoCoSys →

Center	Tech Transfers (TTs)	Multi-company TTs	FT Hires	Interns
CoCoSys	21	15	3	7
CUbiC	33	19	0	7
SUPREME	46	33	11	11
CHIMES	76	51	7	6
ACE	32	11	1	8
CogniSense	7	3	2	2
PRISM	35*	15*	2*	4*
TOTAL	250	147	26	45

* Preliminary count

CoCoSys Projects	ID	PI	KPI Forms	TT Sponsors
Architectures for Neuro-symbolic-probabilistic Workloads	3131.005	Tushar Krishna	3131.005 KPI	TSMC, IBM, Intel
Full-stack Optimization and Software Frameworks for Cognitive Systems	3131.006	Yingyan Lin	3131.006 KPI	IBM, Intel
Technology and Integration-driven Cognitive Architectures	3131.007	Yu (Kevin) Cao	3131.007 KPI	Intel, IBM, Boeing, Raytheon, TSMC

SRC

Agenda

Time	Title	Presenter
6:00 - 6:10 pm	SRC Welcome and Overview	Roman Caudillo / SRC
610 - 6:20 pm	**CoCoSys** Highlights/Lowlights/Top Plans for Y2	Arjit Raychowdhury / GA Tech Anand Raghunathan / Purdue
6:20 - 6:30 pm	**CogniSense** Highlights/Lowlights/Top Plans for Y2	Saibal Mukhopadhyay / GA Tech James Buckwalter / UC Santa Barbara
6:30 - 6:40 pm	**ACE** Highlights/Lowlights/Top Plans for Y2	Josep Torrellas / UIUC Minlan Yu / Harvard
6:40 - 6:50 pm	**PRISM** Highlights/Lowlights/Top Plans for Y2	Tajana Rosing / UC San Diego Nam Sung Kim / UIUC
6:50 - 7:00 pm	**CHIMES** Highlights/Lowlights/Top Plans for Y2	Madhavan Swaminathan / Penn State Muhammad Bakir / GA Tech
7:00 - 7:10 pm	**SUPREME** Highlights/Lowlights/Top Plans for Y2	Grace Xing / Cornell Tomas Palacios / MIT
7:10 - 7:20 pm	**CUbiC** Highlights/Lowlights/Top Plans for Y2	Keren Bergman / Columbia Ali Niknejad / UC Berkeley
7:20 - 7:30 pm	Q&A and Discussion with SAB	

SRC

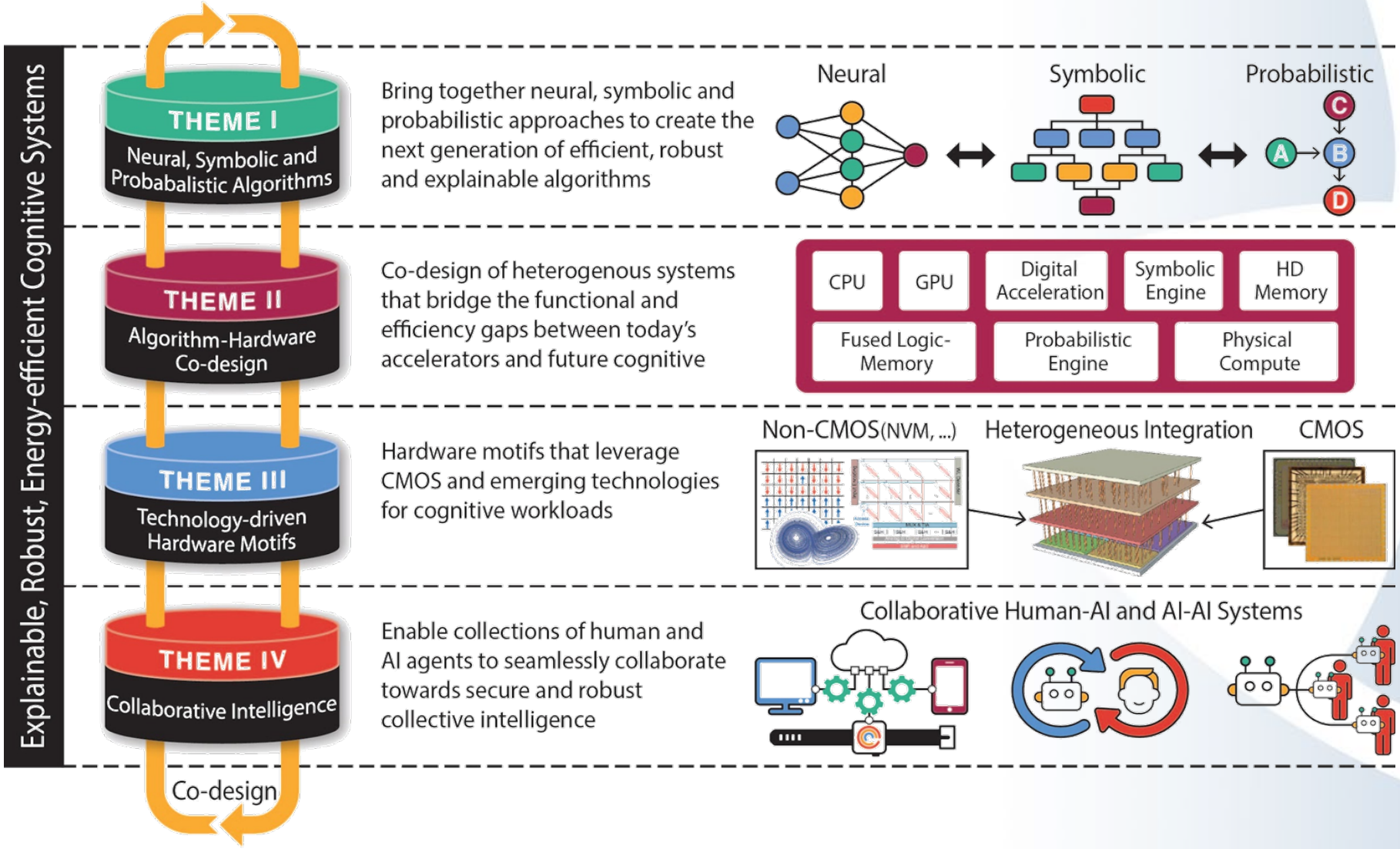
Semiconductor Research Corporation

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

CoCoSys Center SAB Update

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

Center Overview and Themes



SRC

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Year 1: Goals and Accomplishments

	Goals	Accomplishments
Technical	Conceive and demonstrate early examples of structured combinations of neural, symbolic and probabilistic methods to enable human-AI and AI-AI collaboration	Created early examples of neuro-symbolic and neuro-probabilistic algorithms for 3D scene and language understanding, including concept graphs, NeRF and conversational tables
Technical	Characterize emerging algorithms and workloads to identify needs and efficiency gaps in current cognitive hardware architectures	First study to characterize neuro-symbolic workloads on CPU and GPU architectures, identified bottlenecks for future HW development
Technical	Development of a technology (logic, memory and packaging) evaluation framework to project and benchmark technology capabilities for cognitive systems	Developed PDK with CMOS+ eNVM for circuit and architectural exploration, tools for workload-aware thermal analysis of heterogeneous integrated systems
Center Management	Team-building within center and across JUMP2 ecosystem and define the Center's grand challenge	Theme-level collaboration building meetings, early signs of cross-theme, cross-university and industry collaboration (joint projects, papers)
Center Management	Define broadening participation (BP) goals, identify BP champion and execution plan	Recruited Yan Fang (Kennesaw State) as BP champion, focus on undergraduate research and curriculum development

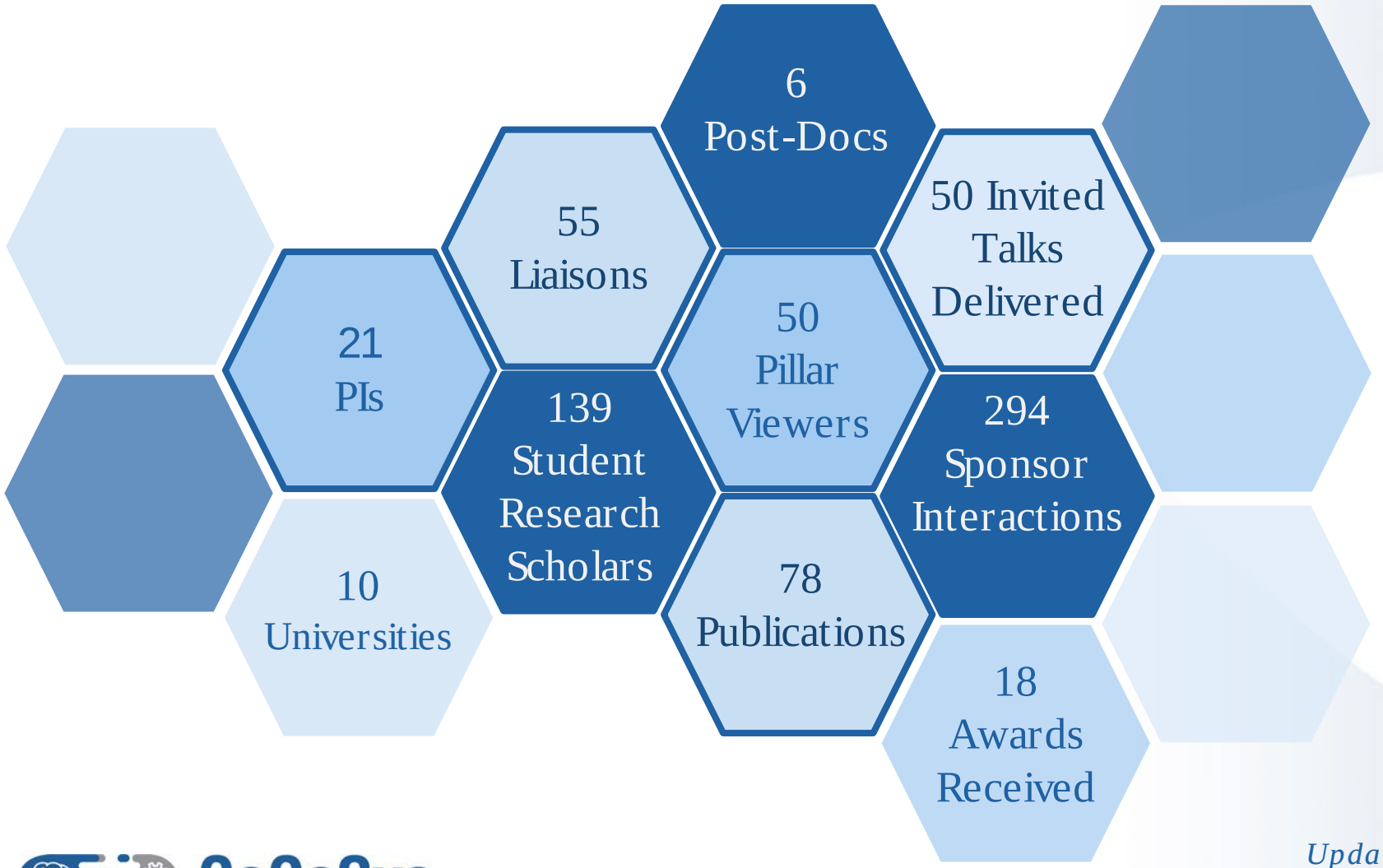
Year 1: Highlights

- Center hiring fully ramped up (139 student scholars, 6 post-docs)
- Firing on all cylinders
  - 78 publications (IEDM, ISSCC, NeurIPS, ISCA, ASPLOS, DAC, …)
  - 18 awards
  - 5 chip tapeouts, 2 open-source modeling and evaluation tools
- Annual review well-received despite being very early in the year
  - Great effort from PIs, students, theme leads
- Several industry visits and talks by PIs
- Early instances of technology graduating to CHIPS Act (ME Commons) and DARPA (OPTIMA) programs

Year 1: Challenges

- Initial lag in contracting, spending & invoicing
  - Expect to be back on track in Q4
- Sometimes thin attendance in weekly theme meetings
- No organized process to connect students to internship and full-time opportunities
  - Especially challenging for PIs who are new to SRC ecosystem
- Still working on defining grand challenge
  - Expect to finalize in Q1'24, present at the annual review

CoCoSys at a Glance



Updated December 2023

CoCoSys

CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS

Semiconductor Research Corporation

CogniSense
CENTER ON COGNITIVE MULTISPECTRAL SENSORS

CogniSense: Center on Cognitive Multispectral Sensors

Center Overview and Highlights

Director: Saibal Mukhopadhyay, Georgia Tech
Co-Director: James Buckwalter, University of California, Santa Barbara

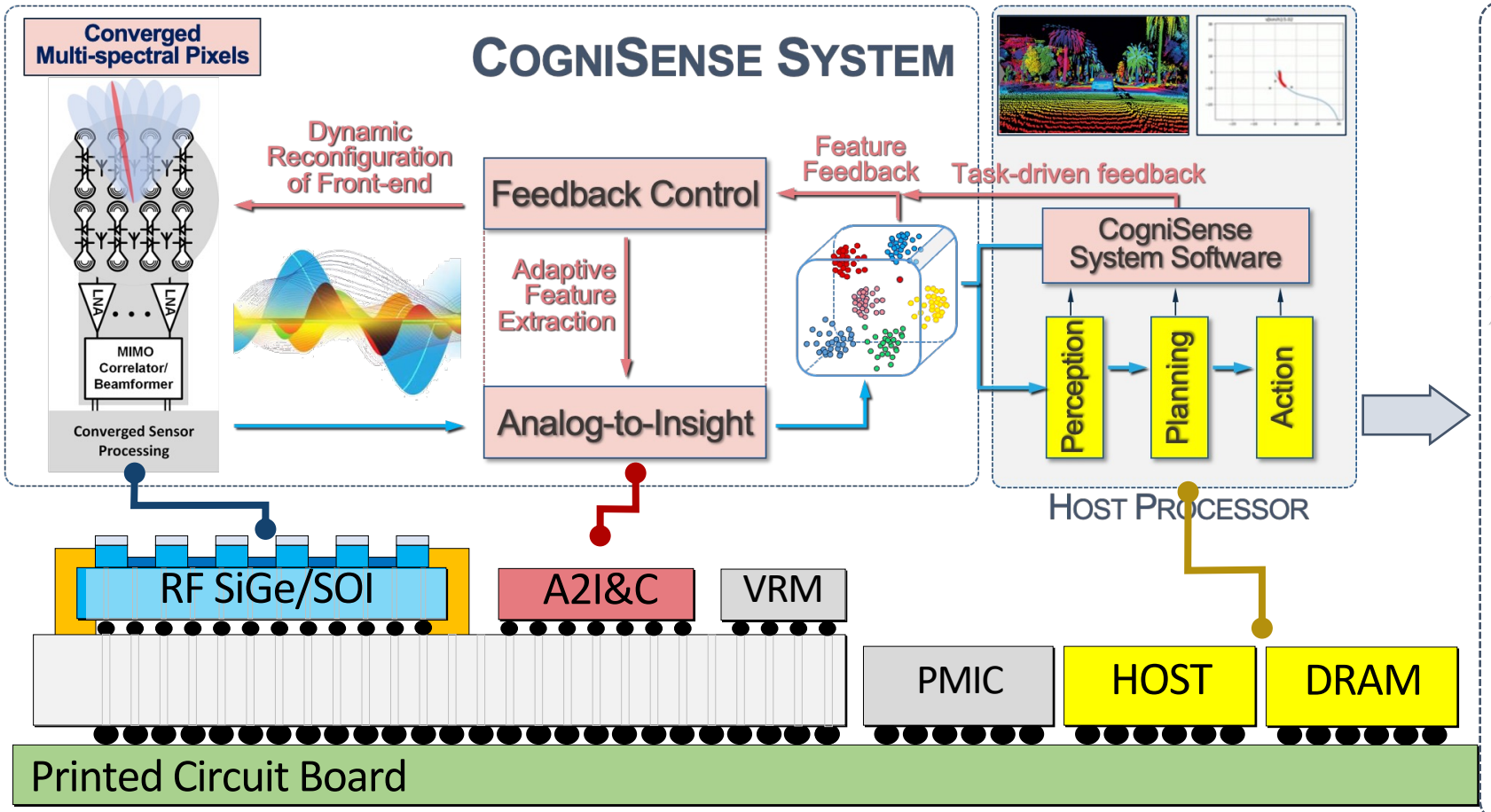
Georgia Institute of Technology, UC Santa Barbara, University of Michigan, University of Illinois Chicago, Columbia University, Purdue University, Cornell University, Massachusetts Institute of Technology, University of Maryland, University of Delaware, University of California, Davis, and Iowa State University,

12 Universities & **10** States

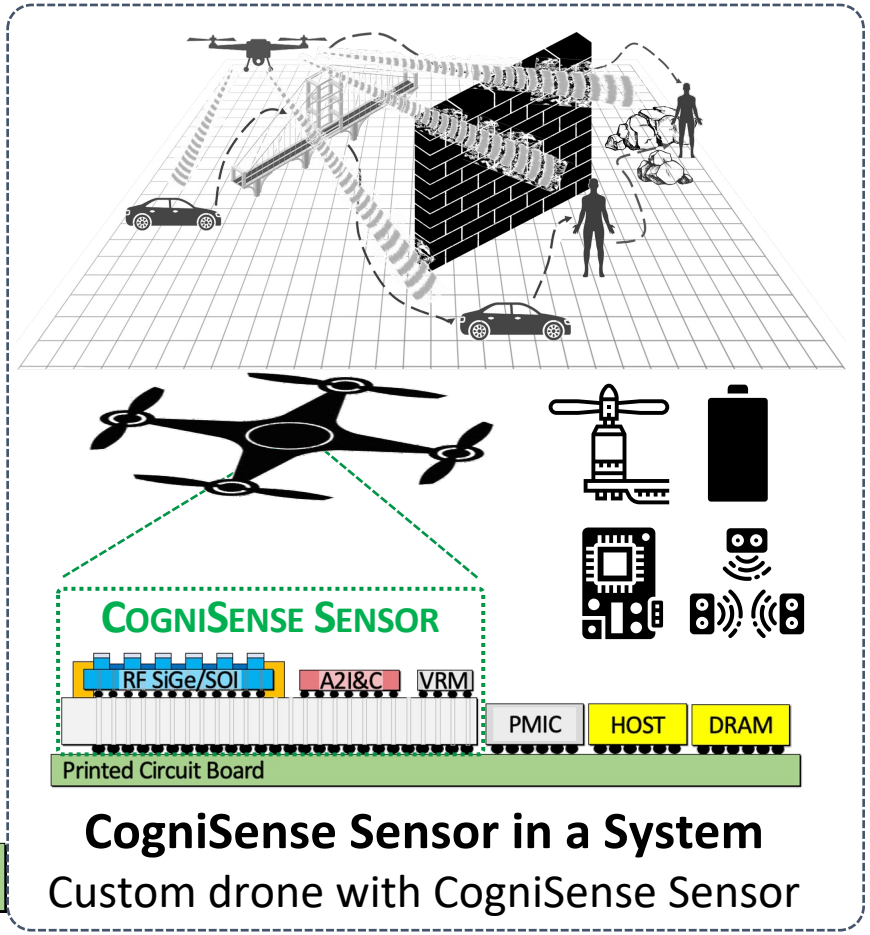
20 Principal Investigators

80+ Students + Post-docs

CogniSense Overview



COGNISENSE APPLICATIONS



CogniSense Sensor in a System

Custom drone with CogniSense Sensor

Cognitive multi-spectral sensors directly generate trustworthy insights from wideband multi-modal analog signals using closed-loop feedback control of the sensor hardware and feature extraction algorithms that enable energy-efficient sensing-to-action.

SRC

CogniSense
CENTER ON COGNITIVE MULTISPECTRAL SENSORS

Goals and Accomplishments

Goals	Accomplishments
1. Ramp-up the center with student recruitments and coordinated interaction among PIs, Students, and Liaison	**Accomplished** • Sub-contracting with most universities completed in Q1. • 80+ students and post-docs in the center • Regular cadence in liaison meetings - each thrust meets with liaison once in every six week.
2. Initiate research progress within and across thrusts	**Accomplished** • All thrusts have made major progress in research objectives • Cross-thrust activities have been demonstrated
3. Early start to center-wide prototyping activities to identify strengths & weaknesses of core concepts	**Accomplished** • Multiple hardware prototypes of CogniSense concepts and components have been developed across all thrusts • Students had “seven” live demonstrations of CogniSense concepts in the Annual Review.
4. Collaboration with other JUMP2.0 centers and plan cross-center activities	**Initial progress** • Initial interactions with PIs from CHIMES, CUBIC, and PRISM. • Identified topics for coordinated center-wide interactions.

Research Highlights from Year 1

Multi-spectral front-end & sensor integration

- Power-efficient 140GHz FMCW radar
- Beam-steering & switched beam Lidar
- High dynamic range passive imaging
- Front-end circuits for sensor convergence
- 3D HI (Cu-Cu bonding) optimized for electrical & thermal management in CogniSense sensor.

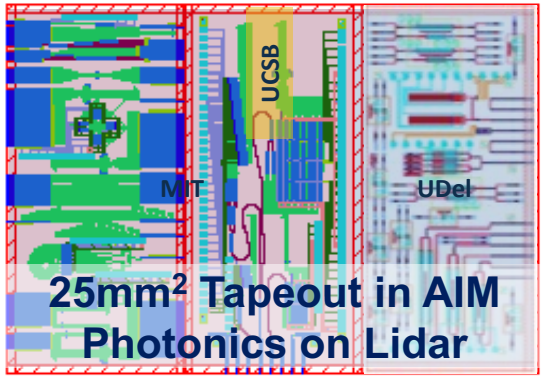
Algorithm & circuits for feature extraction to reduce analog data

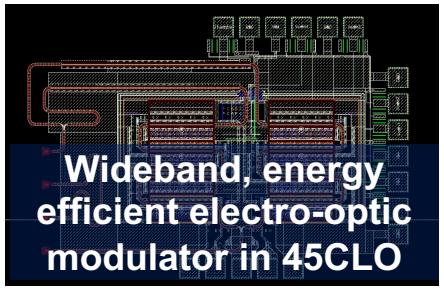
- Feature extraction from analog input with compute-in-memory circuits.
- Multi-modal digital feature processing with efficient neural network.
- Fundamental analysis of sensor trade-offs and ML-enabled sensor simulation framework

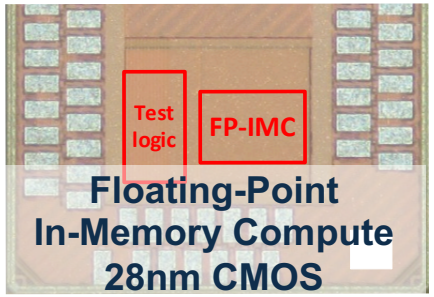
Cross-stack adaptation for power-quality trade-off in CogniSense sensing-to-action

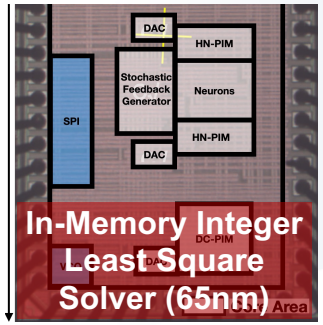
- Uncertainty-aware sensor processing with trust and reputation tracking
- Adaptive on-sensor feature extraction algorithms for analog & multi-modal inputs
- Algorithm for task-dependent sensor control & security architecture development
- Multi-modal custom drone platform to study CogniSense concepts

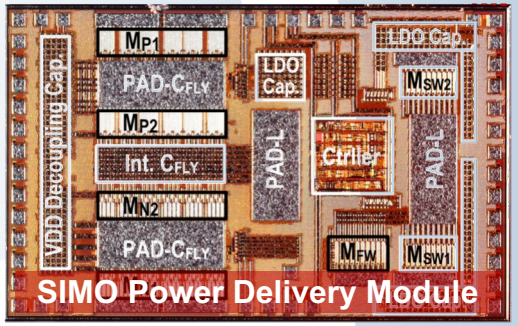
Research Highlights from Year 1

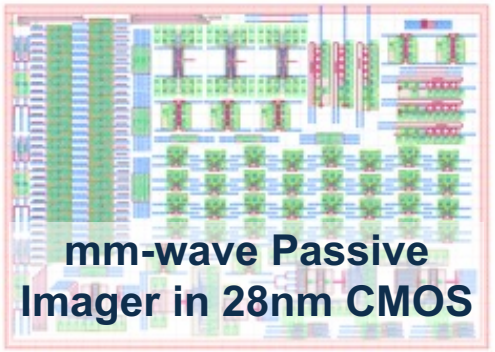


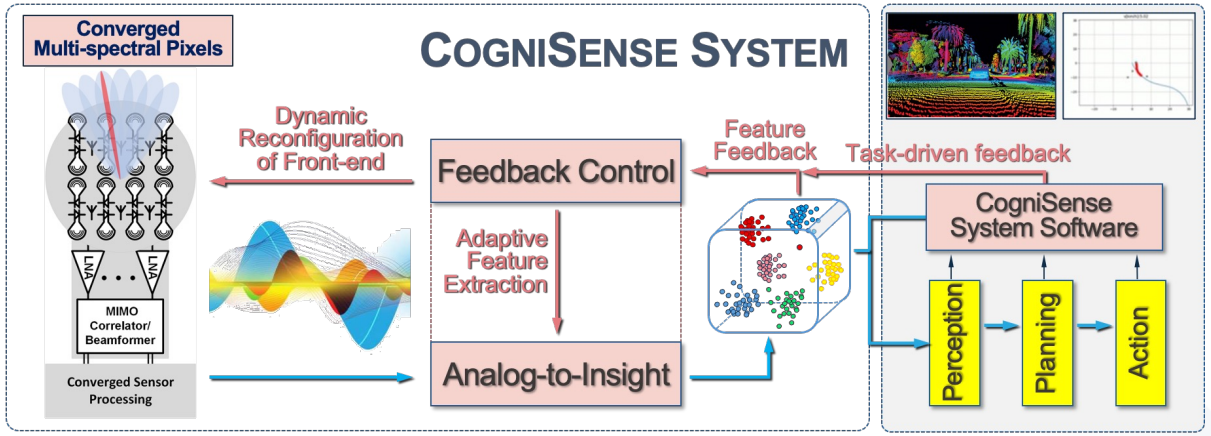


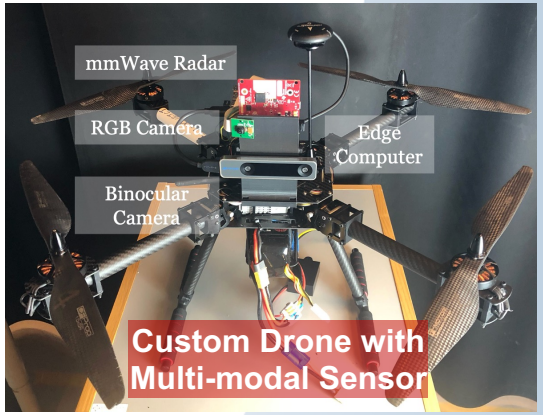


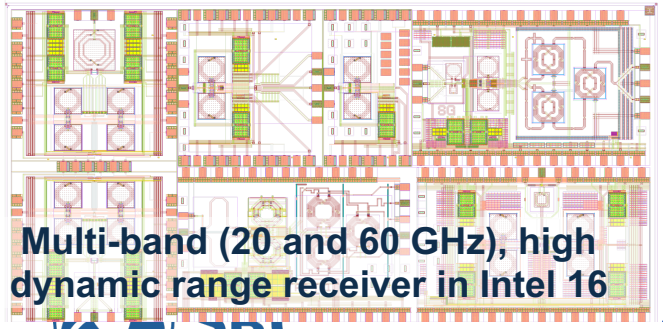


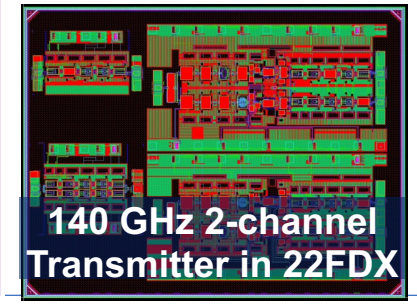


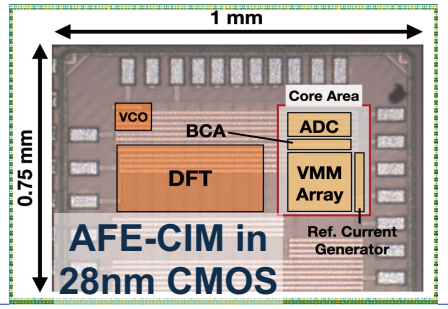


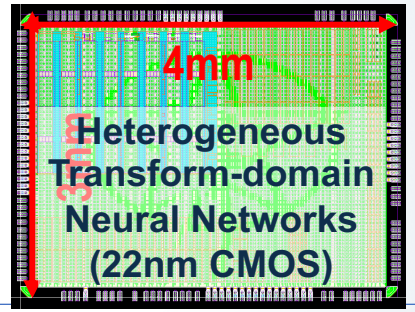


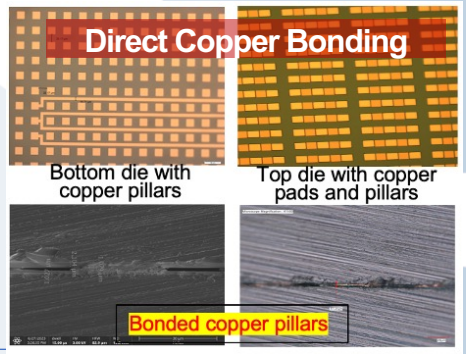












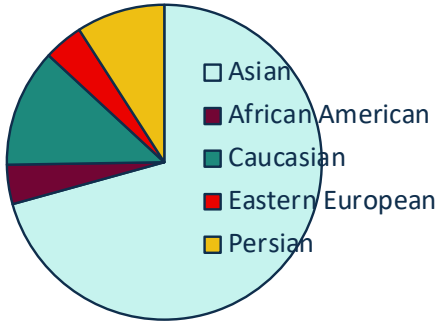
Needs Attention in Year 2

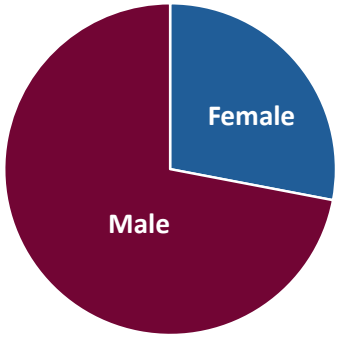
- **End-to-end simulation environment**
  - Radar and lidar-based input simulation
  - Integrated software pipeline coupling algorithmic innovations across center
  - Power/performance models to couple design innovations
- **System adaptation & sensing-to-action**
  - Define challenge application for case studies
  - Emphasize system & software challenges like security
- **Metrics & Benchmarking**
  - Establish thrust level metrics for sensing quality, analog data deluge, and adaptive power efficiency
  - Evaluate progress of each component, sub-system, and sensing-to-action against metrics.

Center	Potential Topic of Interactions
COCOSYS Georgia Tech	Leverage prior experiences of CBRIC and COCOSYS centers on drone platforms and new AI/ML models
CUbiC Columbia	Joint work on photonics, optimizing signal processing pipelines, and exploring joint communication and sensing
ACE UIUC	Security aspects associated with distributed computing and sensing that are relevant to CogniSense
PRISM UCSD	Impact of new memory technology in design of feature extraction algorithms and exploring memory solutions for streaming sensor data.
CHIMES Penn State	Coordination of 3D-HI activities with CHIMES and incorporate new heterogeneous integration and advanced cooling solutions explored in CHIMES into the CogniSense platform.

Student Participation

- **Student participation**
  - 82 students/post-docs: Graduate (71), Post-docs (9), and UG (2)
  - 50+ students in the Annual Review





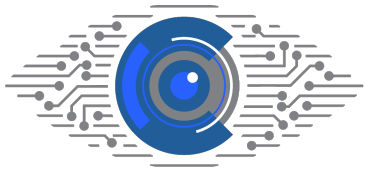
- **Student-led publications**
  - **11** students participated in TECHCON
  - **40+** Students papers
- **CogniSense student in Members**
  - **6+** Students joined **SRC Members**
  - **4+** student interned in **SRC Members**

Top Priorities for Year 2

- **Increase diversity of student body in Year 2**
  - URM & female students to > 40% of more
  - Number of UG students to 20 or more
  - Increase number of US Citizen students
- **Transition students to JUMP2.0 community**
  - Connect students to internships positions
  - Recommend students for full-time positions
  - Explore options for JUMP2.0 students in Govt. labs
  - Encourage students for faculty/post-doc in partner
- **Accelerate Broadening Participation Program**
  - High school & undergraduate participation
  - Leveraging IEEE/SPIE/Optica Centers
  - Female and minority engagement

Summary

- **What are we trying to do?**
  - Develop energy-efficient and trustworthy multi-spectral sensors for autonomy
- **How it's done today? What are the limitations?**
  - Sensing quality is improved via large and wideband pixel arrays; one array for each modality.
  - Continuous sensing and digitization of all pixels in a wideband pixel array leads to large volume of digitized data (Analog Data Deluge) and high sensing/digitization power.
  - Different sensors for each modality leads to higher sensing power and system cost.
- **What's new in our approach?**
  - Our ***cognitive*** multi-spectral sensors with pixel-level convergence will directly generate trustworthy insights from wideband multi-modal analog signals using closed-loop feedback control of the sensor hardware and feature extraction algorithms.
- **What if we are successful?**
  - Multi-modal sensing arrays that eliminate corner-case obstructions in machine perception and more efficiently use resources (data bandwidth and power).
  - New research direction in designing "adaptive sensors" that learns to traverse the quality versus resource space in evolving environment.



CogniSense

CENTER ON COGNITIVE MULTISPECTRAL SENSORS

What We Did and What is Coming

ACE Center for Evolvable Computing
SAB Meeting, December 2023

Director:
Josep Torrellas (UIUC)

Assistant Director:
Minlan Yu (Harvard)

https://acecenter.grainger.illinois.edu/

What ACE is About

Devise technologies for **distributed computing** to substantially improve the **energy efficiency** and the performance of applications

How to do it?

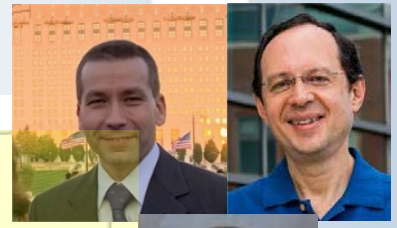
Leverage hardware accelerators and integration

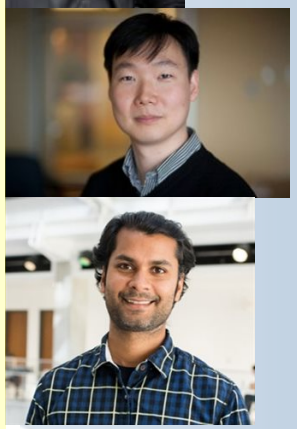
Minimize data movement

Co-design hardware and software innovations

Integrate security and correctness from the ground up

Tightly Coupled Organization





c) **Theme 3:** Fine-grained Communication & Coordination

Internet

Cloud

Edges

GPU

Switch

NIC

On-server, and in-network accelerators

DRAM

SSD

NVMe

a) **Theme 1:** Heterogenous Computing Platforms

b) **Theme 2:** Distributed Evolvable Memory and Storage

Security

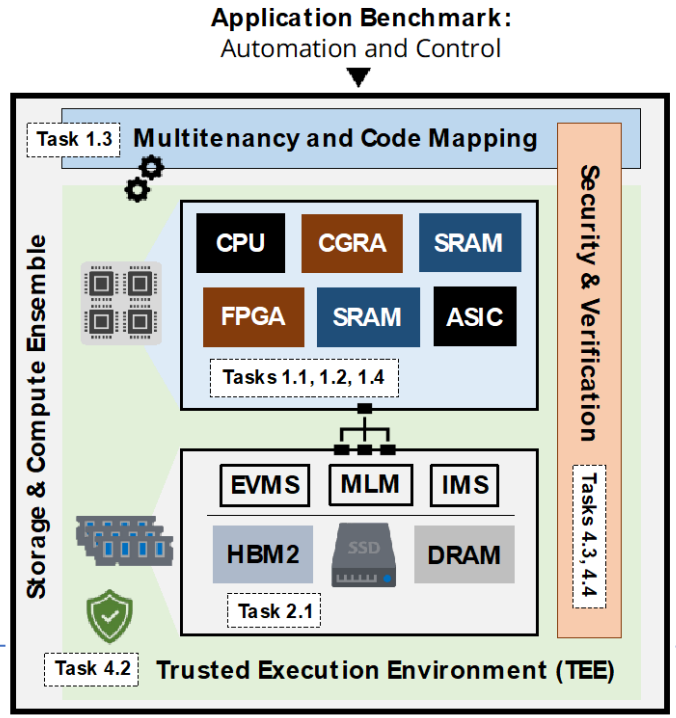
Privacy

Fidelity

d) **Theme 4:** Security, Privacy, Correctness

Demonstrator 1: A Reconfigurable Multi-accelerator Compute Ensemble

- Evolvable accelerators (FPGA, CGRA, and ASIC)
- Aggregated into a multichip module (MCM)
- Ensemble will be multitenant
- Compiler will generate code and map it to compute units
- Other software will be ported

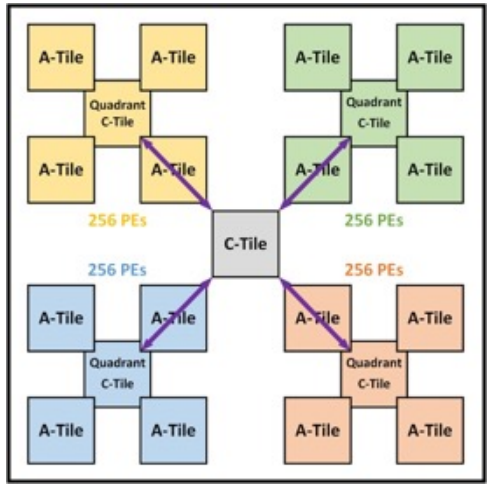


SRC ACE CENTER FOR EVOLVABLE COMPUTING

Configurable Chiplet for Composable Computing

- A configurable chiplet that can adapt to changing workloads (test chip tapeout planned for May 2024)
  - Made of a fine-grained mix of CPU and systolic array tiles for flexible workload mapping & adaptation
- A 16Gbps open UCIe interface for connecting chiplets (test chip tapeout planned for May 2024)
  - Make the interface synthesizable and develop an automatic interface generator →Public domain

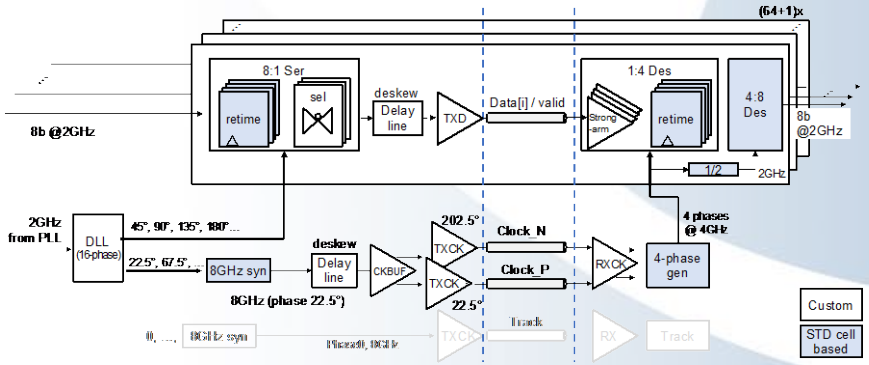
- PIs: Zhengya Zhang, Zhiru Zhang, Charith Mendis



Conceptual sketch of a configurable chiplet architecture

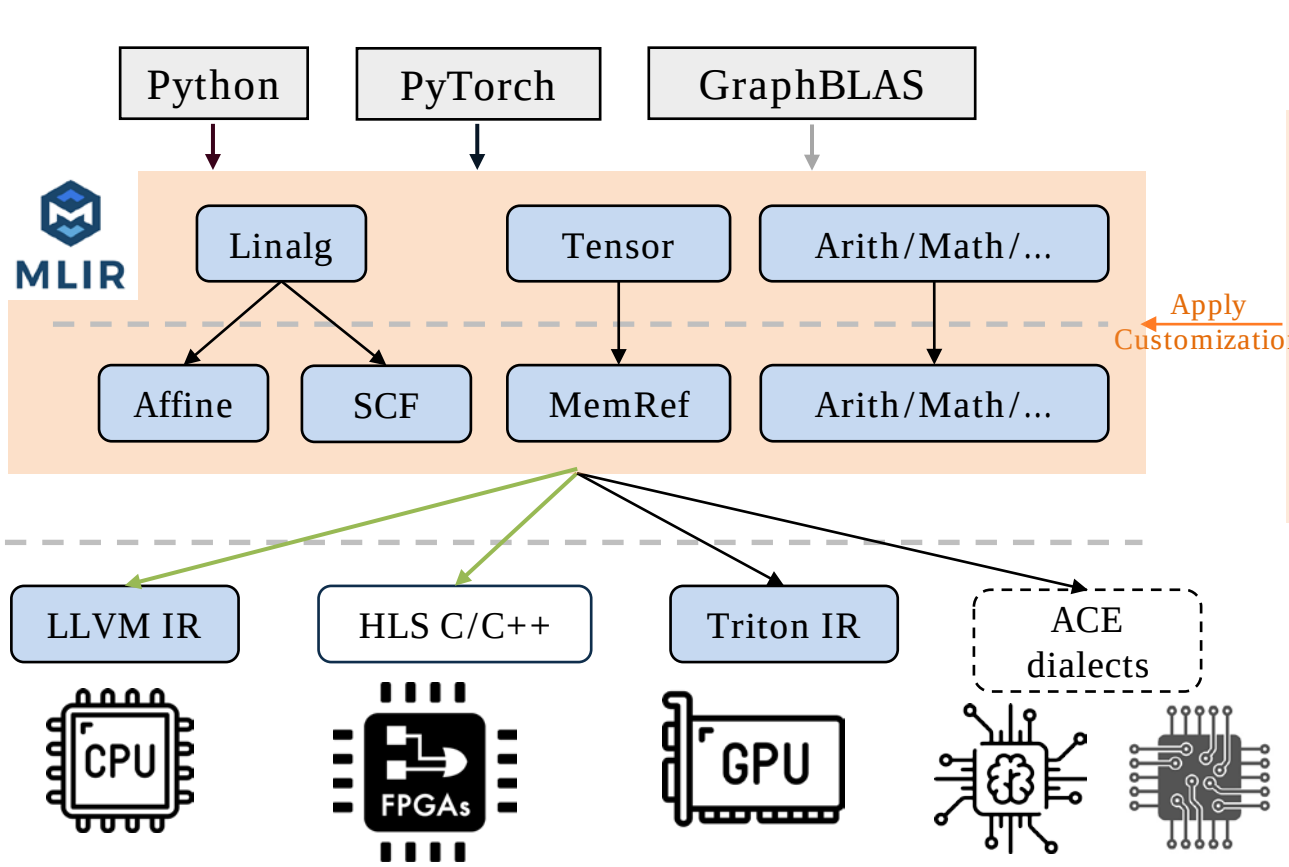
Goal: have a chiplet library of highly reusable chiplets. Show how they can be connected with modularity into bigger systems.

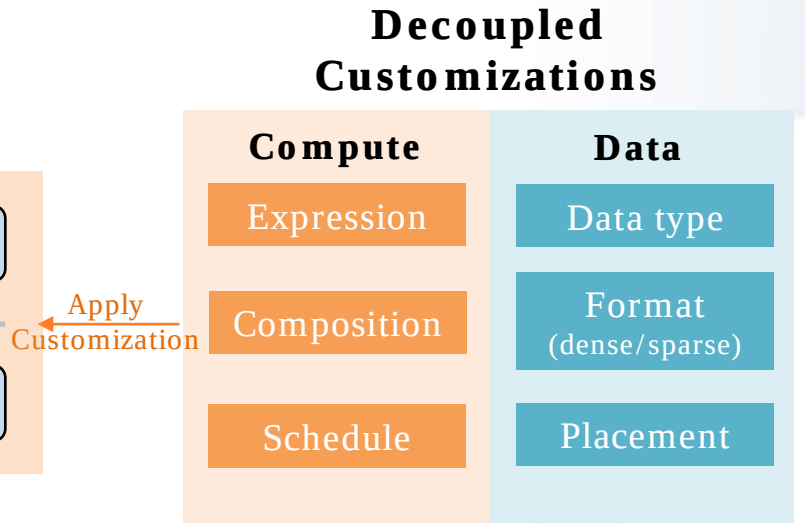
Baseline 16Gbps UCIe design



An Evolvable Intermediate Representation (IR) for Accelerator Programming

- PI: Zhiru Zhang, Charith Mendis



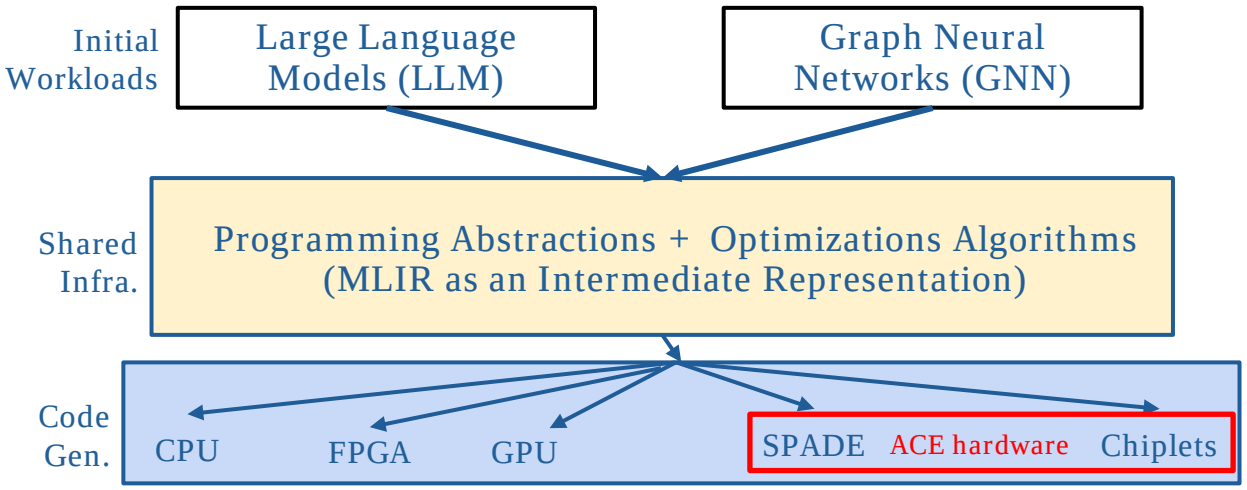


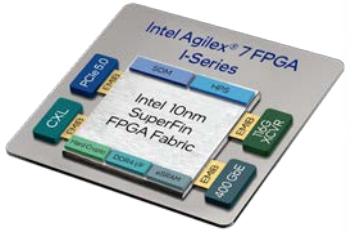
- Portable IR
- Evolvable optimizations with decoupled customizations in compute & data
  - Extensible support for data specialization
  - Formally verifiable code transformations
  - ML-based cost modeling + differentiable design space exploration

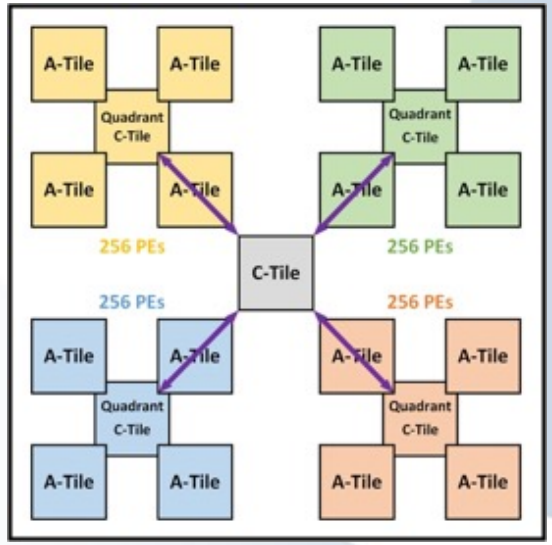
SRC ACE CENTER FOR EVOLVABLE COMPUTING

Next: Further Build-up Compiler + Hardware for Evolvable Accelerators

- A preliminary ACE compiler stack for evolvable accelerators
- Initial testbeds for prototyping distributed accelerator ensembles
- Tape out test chips for configurable chiplets







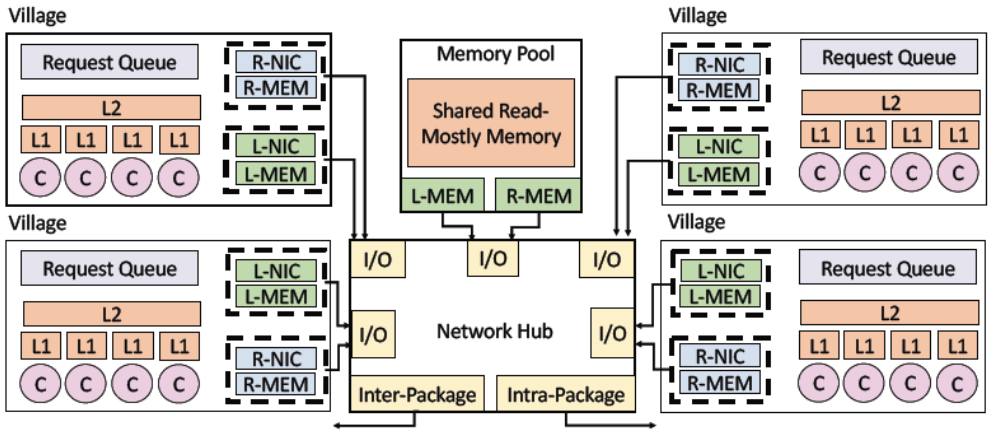
SRC ACE CENTER FOR EVOLVABLE COMPUTING

µManycore: A Cloud-Native Manycore for Tail at Scale

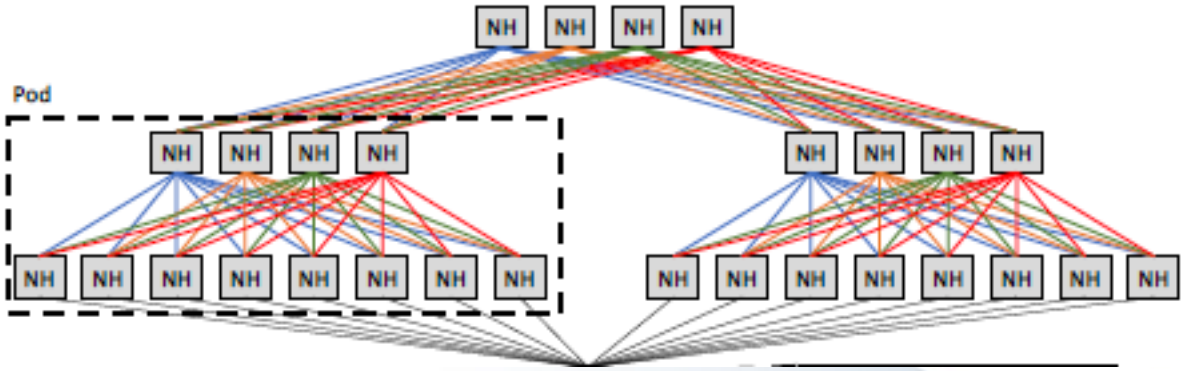
µManycore is a many-chiplet processor designed for cloud-native applications (microservices, etc)

- Main focus: limit **tail latency**
- Simple cores organized in coherent "villages"; no package-wide cache coherence
- Hardware-supported request queue and state save/restore at context switches
- Leaf-spine interconnect for many redundant, low-hop-count paths between clusters

PIs: Josep Torrellas and Muhammad Shahbaz



Processor organized in chiplets of simple cores



High-connectivity on-chip interconnection network

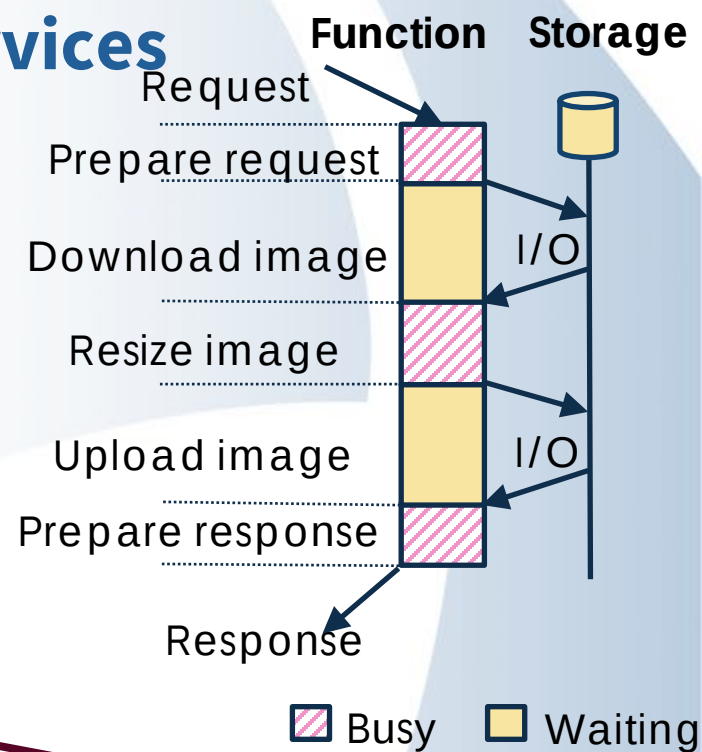
Next: A Programming Environment to Speedup Microservices

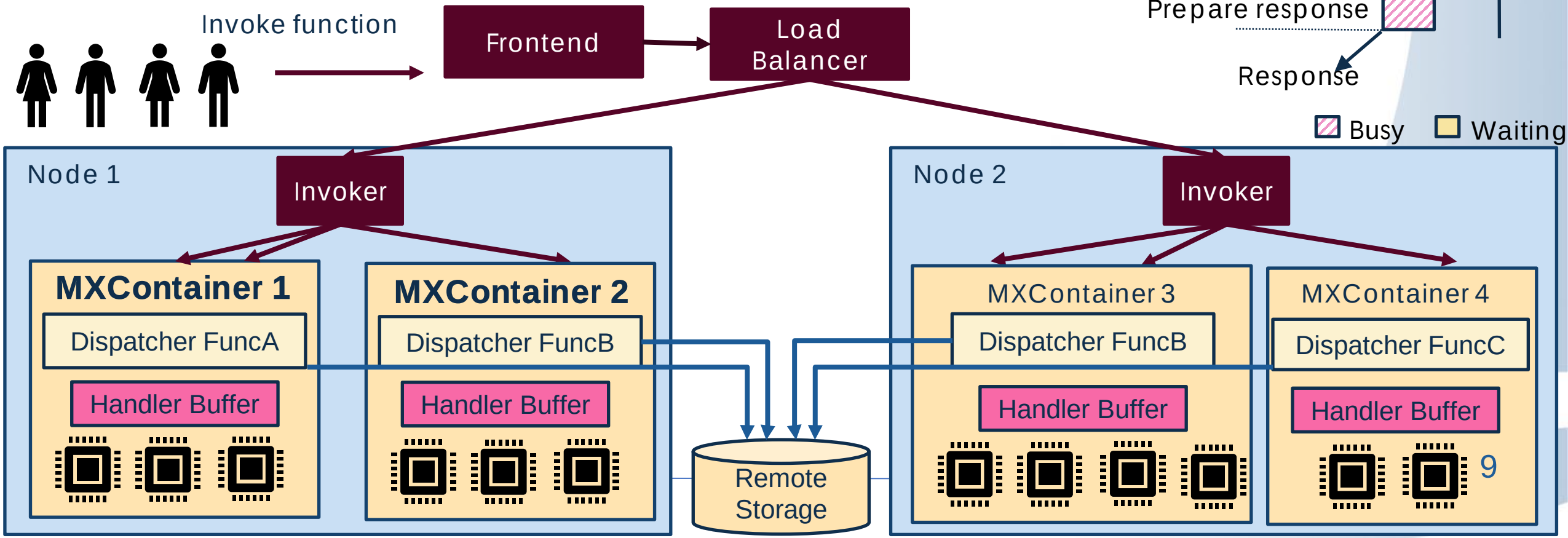
Current programming environment uses resources inefficiently:

→ Need re-use processor, memory state, and I/O

→ Need better data caching

PI: Josep Torrellas

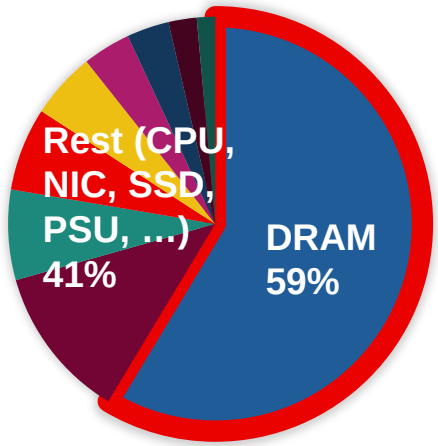




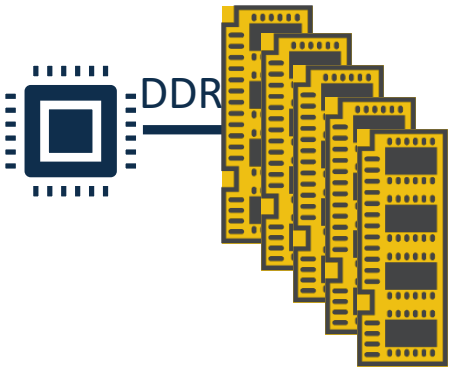
XFM: Near-Memory Acceleration of Software-Defined Far Memory

- Software-defined far memory (SFM)
  - Saves DRAM capacity at the expense of computation (i.e., compressing cold memory pages)
  - Hogs the memory channel
- Propose *XFM:* transparently offloads SFM operations on a Near-Memory Accelerator at the buffer device of the DIMM

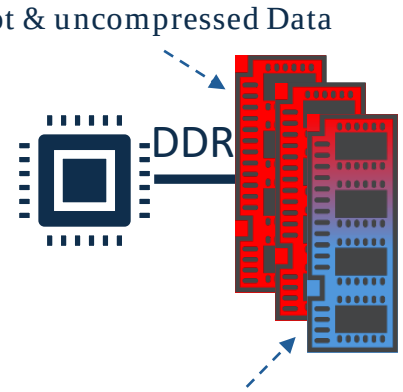
PI: Mohammad Alian



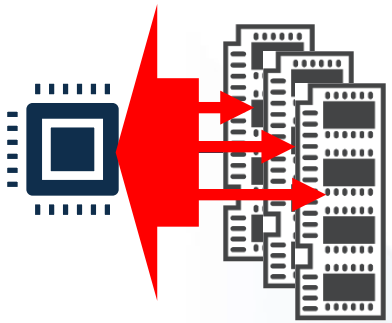
Emissions Breakdown of a A Dell R740 Server [Boavizta Project 2021]



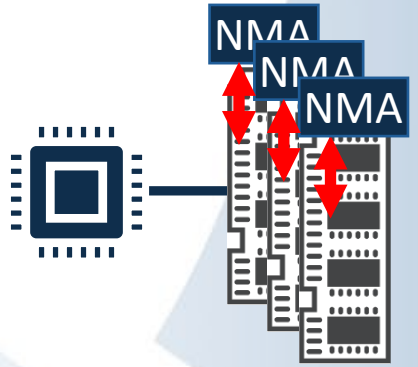
Conventional High Memory Capacity Server



Server Employing Software-Define Far Memory (SFM)



Baseline SFM

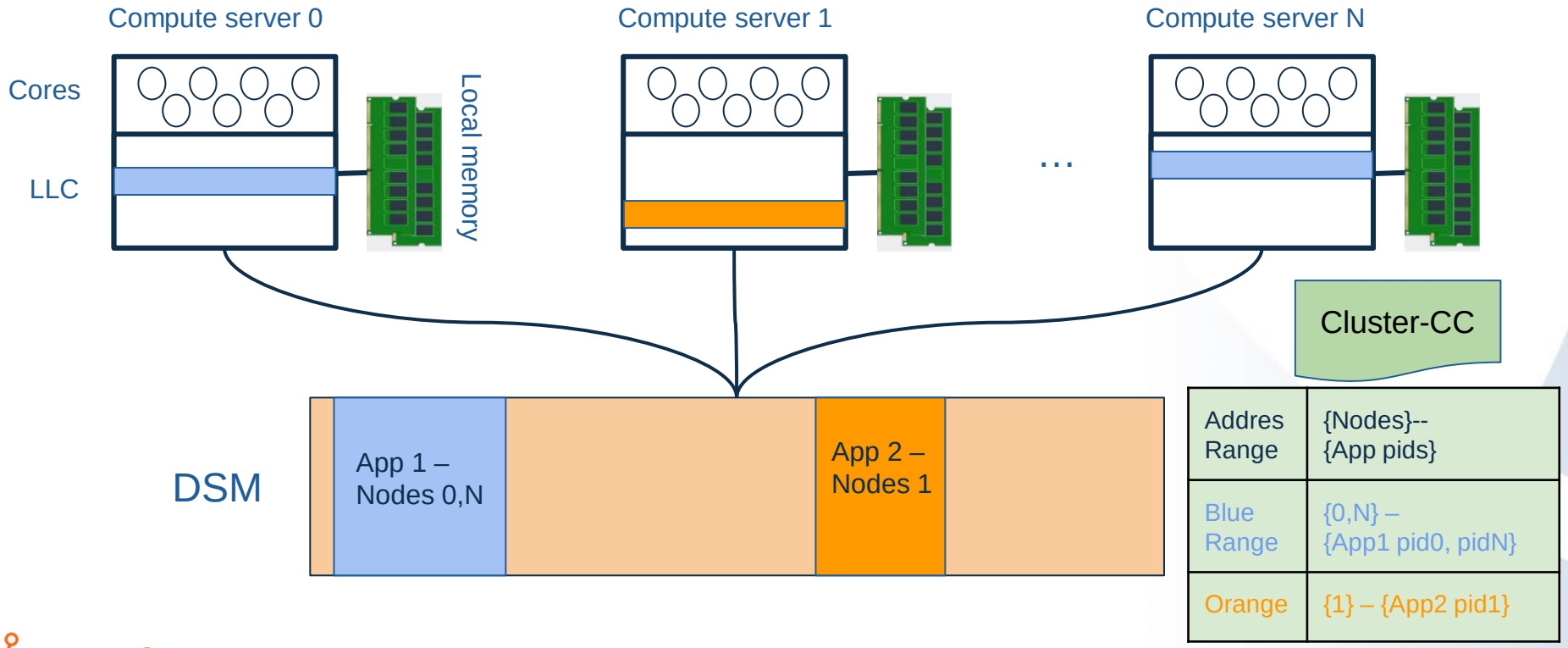


XFM

Next: Coherence at Scale with CXL 3.0

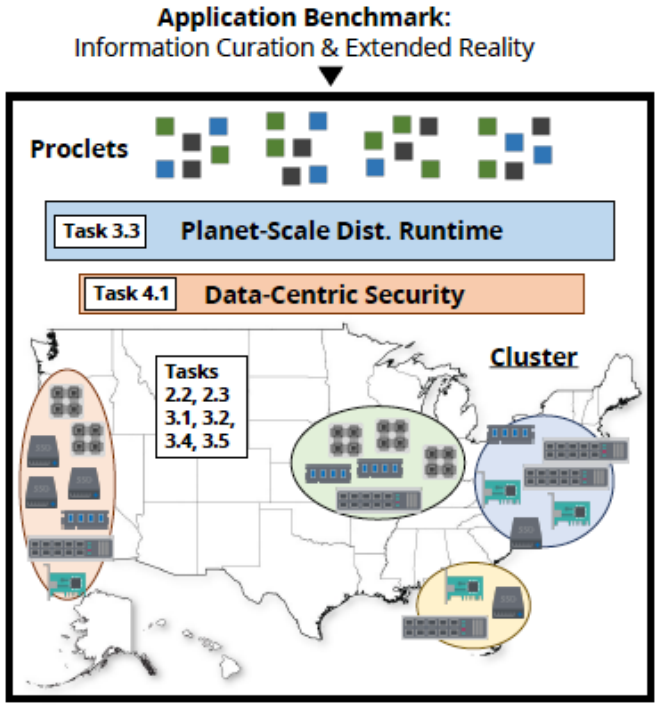
Distributed shared memory for CPUs and accelerators with CXL

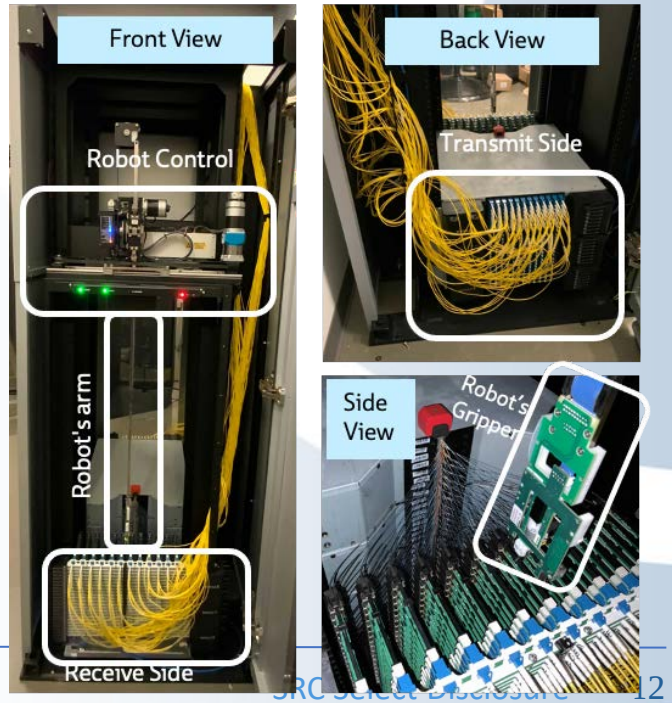
PIs: Steven Swanson, Mohammad Alian, Josep Torrellas



Demonstrator 2: Heterogeneous Large Cluster with Specialized Intelligence

- Large compute cluster composed of accelerators
- Runtime with fine-grained tasks & customized communication stack
- Include accelerators in switches and smartNICs
- Algorithms to apportion resources across apps
- Security mechanisms

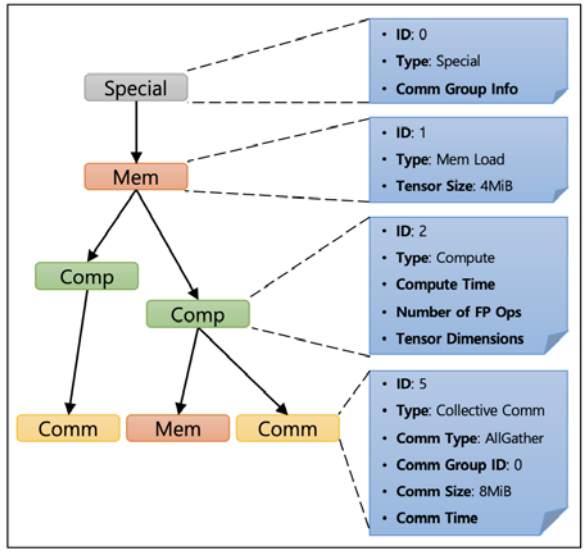


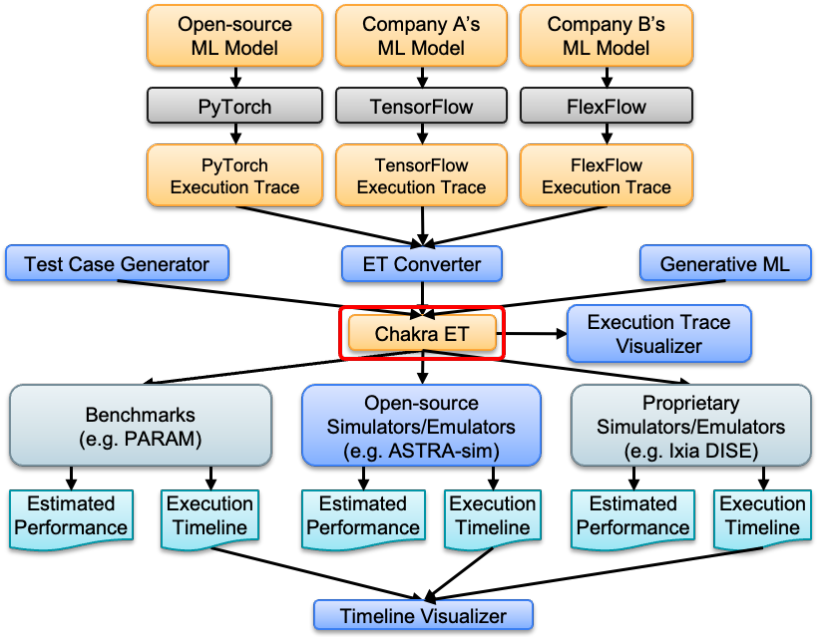


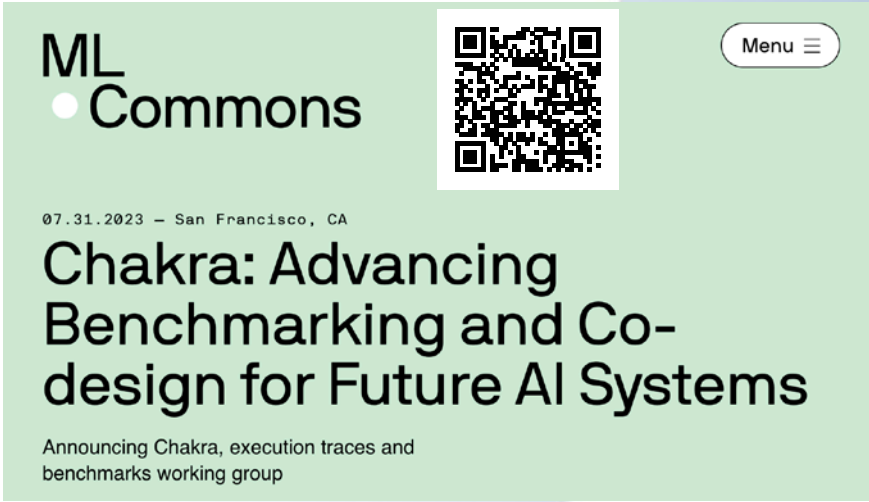
Benchmarking and Co-Design of Distributed AI Systems

- Developed *Chakra*, a framework for collecting execution traces of distributed AI workloads from production servers and replaying them through simulators for future system design.
- Chakra was formally adopted by MLCommons (maintainer of MLPerf) into a working group with several SRC member companies, cloud hyperscalers, and academics to standardize and release.

- PI: Tushar Krishna







SRC ACE CENTER FOR EVOLVABLE COMPUTING

CC-NIC: A Cache-Coherent Interface to the Network Interface Card (NIC)

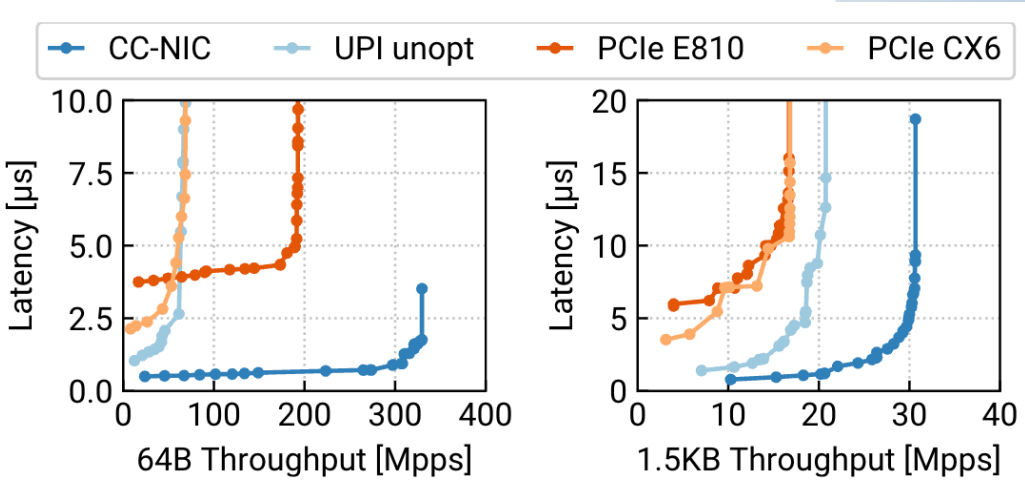
Today's PCIe NIC transmit-receive interfaces are inefficient: PCIe is >80% of intra-rack roundtrip latency

Emerging coherent interconnects (CXL, UPI, UCIe & others) offer a solution

- Devices participate in CPU cache coherence protocol

CC-NIC is a cache-coherent NIC design

- Coherence calls for new techniques in:
  - *Hardware host-device signaling*
  - *Cache-optimized memory layouts*
  - *Symmetric, shared buffer management*
- CC-NIC evaluated on Intel UPI: terabit throughput on SPR, significant latency reduction vs PCIe, up to 50% core savings for app workloads



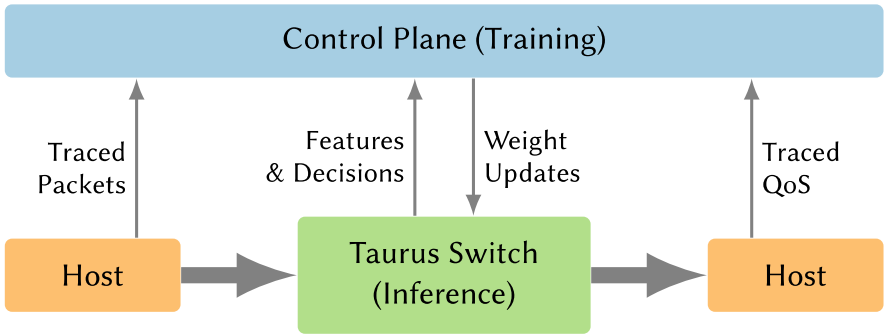
PI: Arvind Krishnamurthy

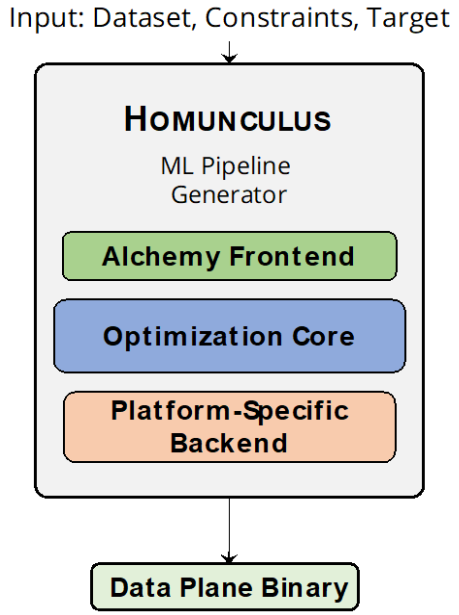
Generating Efficient Code for Datacenter Network Switches

ASPLOS 2023
Distinguished
Artifact

- Use machine learning (ML) to help ensure that networks are robust and secure
- *Homunculus* automatically generates efficient ML models to run on switches (e.g., check for malware)
- Operators only need to specify high-level directives, including datasets, objective, and backends
- Ongoing work: Improving accuracy of these models via online learning as application changes

- PI: Muhammad Shahbaz





App	Features	# NN Param	F1 Score	Accuracy increase
Base-AD	7	203	71.10	
Hom-AD	7	254	**83.10**	**+16.9%**
Base-TC	7	275	61.04	
Hom-TC	7	370	**68.75**	**+12.6%**
Base-BD	30	662	77.0	
Hom-BD	30	501	**79.8**	**+3.6%**

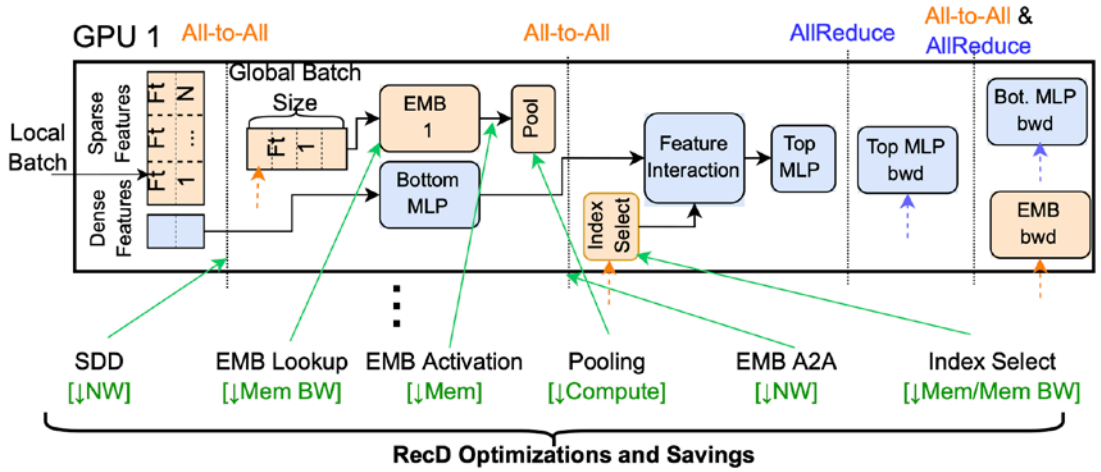
AD: Anomaly Detection
TC: Traffic Classification
BD: Botnet Detection

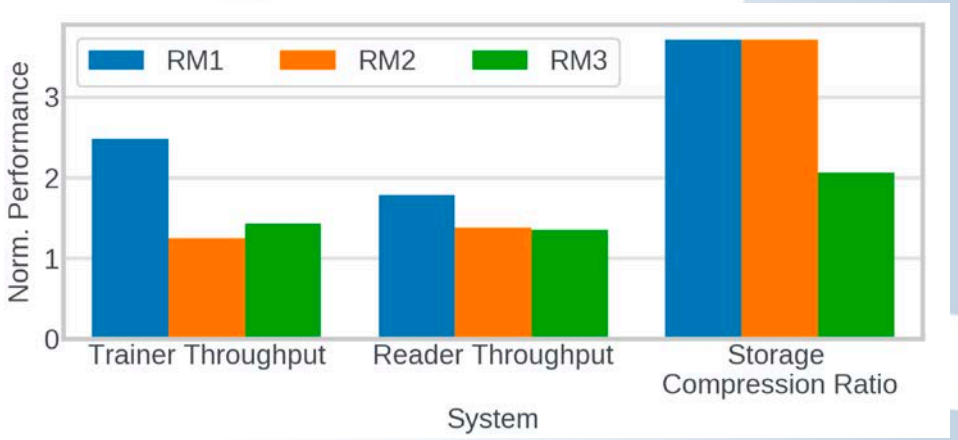
SRC ACE CENTER FOR EVOLVABLE COMPUTING

RecD: End-to-End Optimization of ML Training

- ML datasets exhibit high feature duplication, leading to massive inefficiencies in training
- RecD is a suite of optimizations for storage, preprocessing, and training
- RecD improves storage, preprocessing, and training efficiency by 3.71x, 1.79x, and 2.48x, respectively

- PI: Christos Kozyrakis



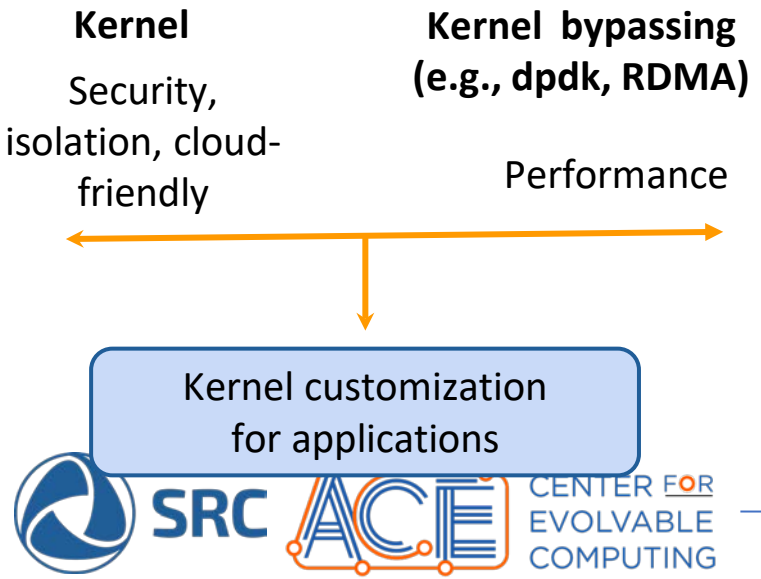


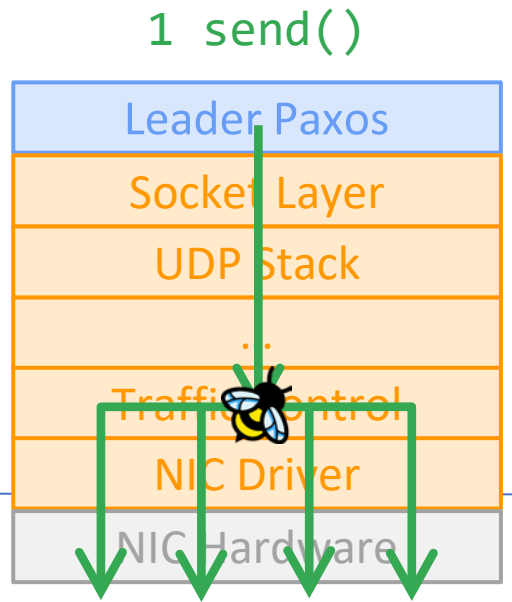
M. Zhao, et al., RecD: Deduplication for End-to-End Deep Learning Recommendation Model Training Infrastructure, MLSys'23

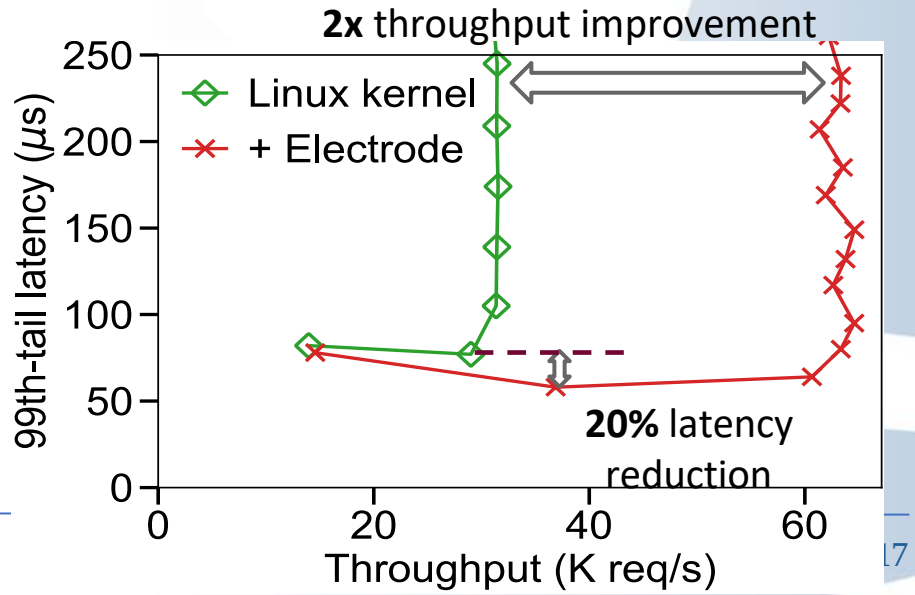
Accelerating Distributed Applications with eBPF

- Current communication stacks use the kernel: they are secure, isolated, and cloud-friendly but slow; bypassing the kernel (e.g., RDMA) is fast but insecure
- Propose intermediate solution: use eBPF to customize a solution to the particular use → fast & secure
- Demonstrated *Electrode*: acceleration of consensus protocols (e.g., Paxos)
  - Throughput increases by 2x; latency reduces by 20%
- Next: Acceleration of a full transaction system (lock server, key-value store)

- PI: Minlan Yu

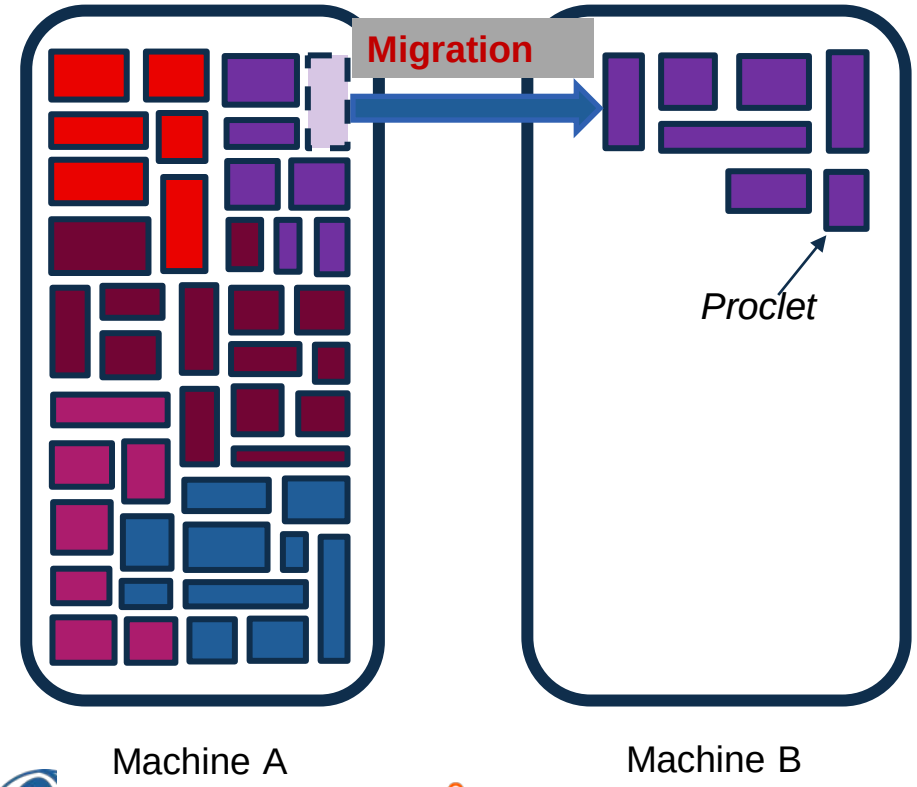




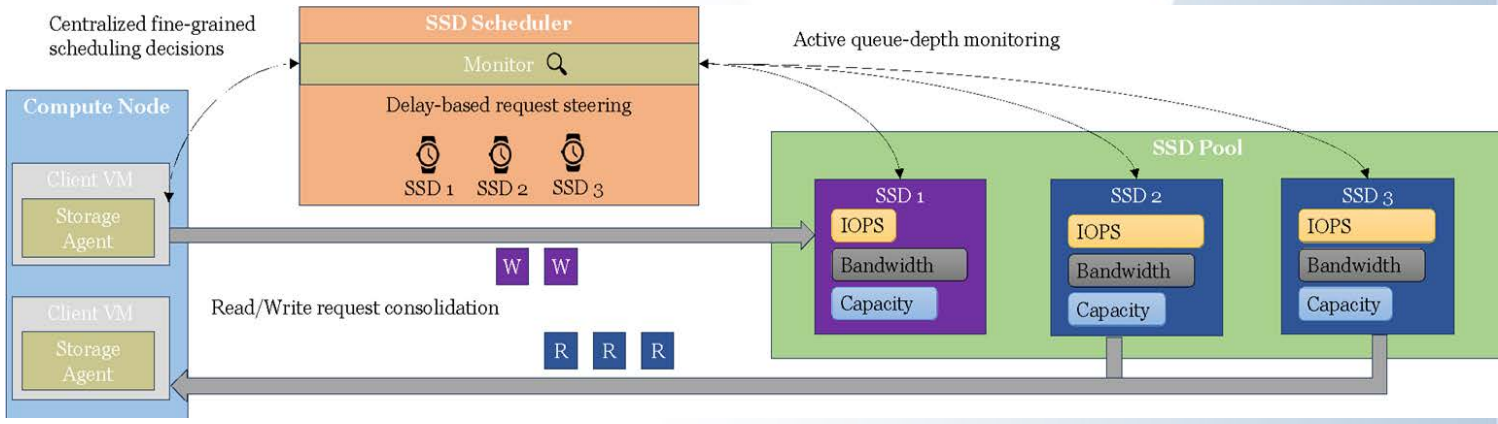


Hardware is Massively Underutilized in Cloud: Self Balancing Resources

- Fast migration to rebalance compute and memory across machines [NSDI 2023]
- PIs: Adam Belay, Minlan Yu



- Next: Make other resources self balancing → Flash SSDs
- Shard across pool of flash SSDs → poor performance
  - Hotspots
  - Heterogeneity
  - Read/Write interference
- *Sandook:* Rapid request steering to rebalance resource use



SpecVerilog: Adapting Information Flow Control for Secure Speculation

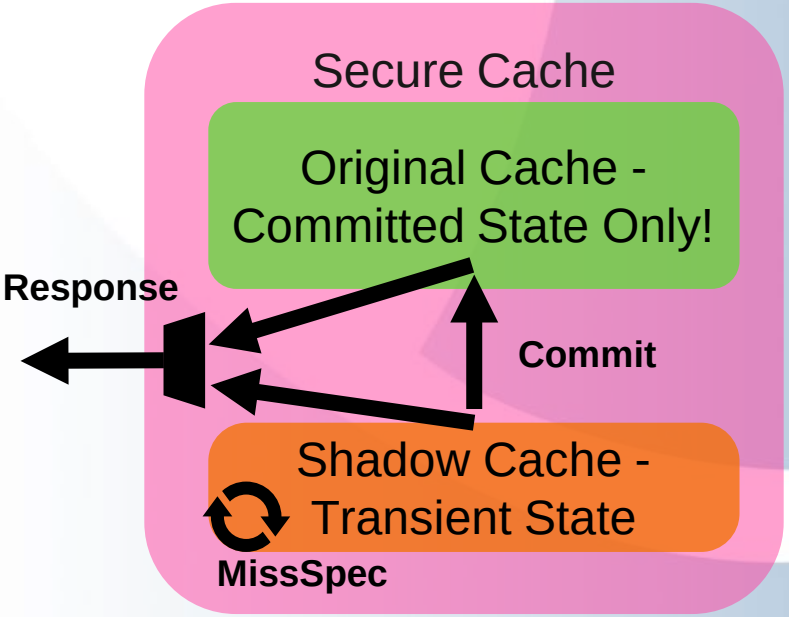
- First work to show that a security-typed hardware description language (HDL) can statically verify that a hardware design is free of transient execution vulnerabilities
  - Support transient non-interference: no information leakage from transient to non-transient
  - Key insight: erasure labels for the compiler to check that speculative state is removed on a misspeculation

- PI: G. Edward Suh

Timing-sensitive, RTL security analysis for **µArch**

Can verify key design patterns, e.g., **static partitioning, invalidation**, and **time-multiplexing**

```
input clk, miss {BOT};
reg specData {BOT ↑miss TOP};
always@(posedge clk) begin
  specData <= specData; //Rejected
  specData <= (miss) ?
              0 : specData; //Accepted
end
```



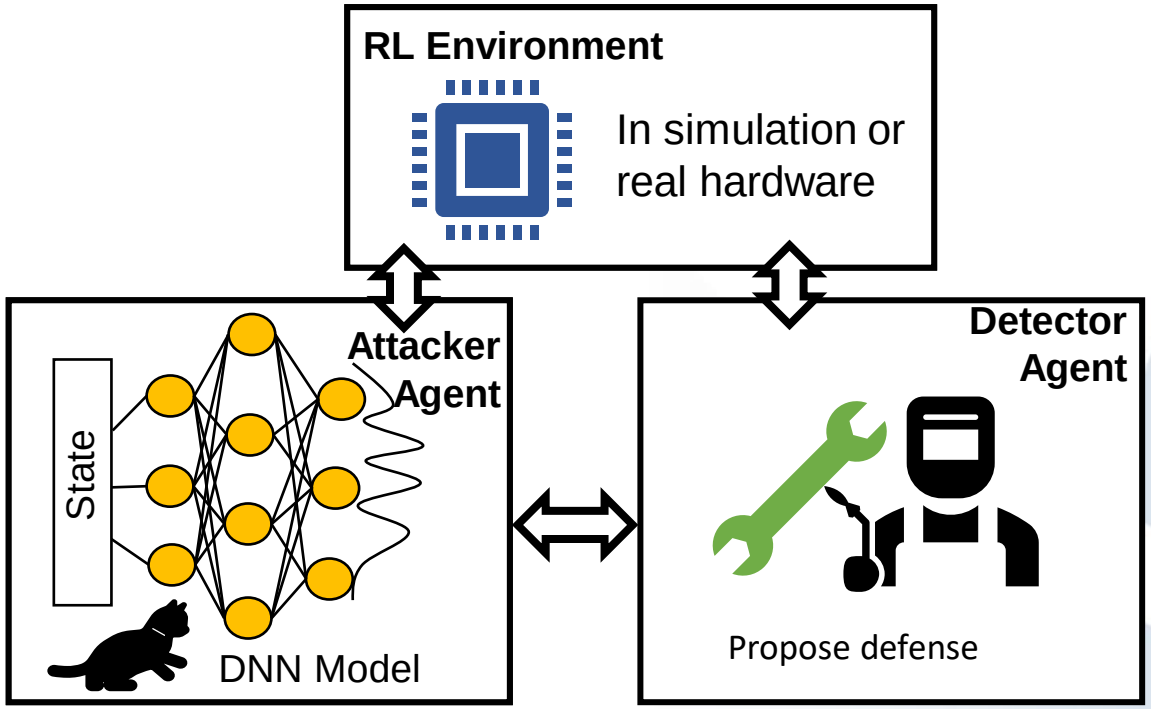
SRC ACE CENTER FOR EVOLVABLE COMPUTING

Now and Next: Reinforcement Learning for Automated Attack & Defense

Reinforcement Learning has been effective in game scenarios → Use it in security:

- RL for cache-timing attack exploration (*AutoCAT*): Finds new attack pattern on real hardware
- Multi-agent RL for automatically exploring cache-timing attacks and detection schemes together (*MACTA*):
  - The trained RL detector can generalize to unseen attacks

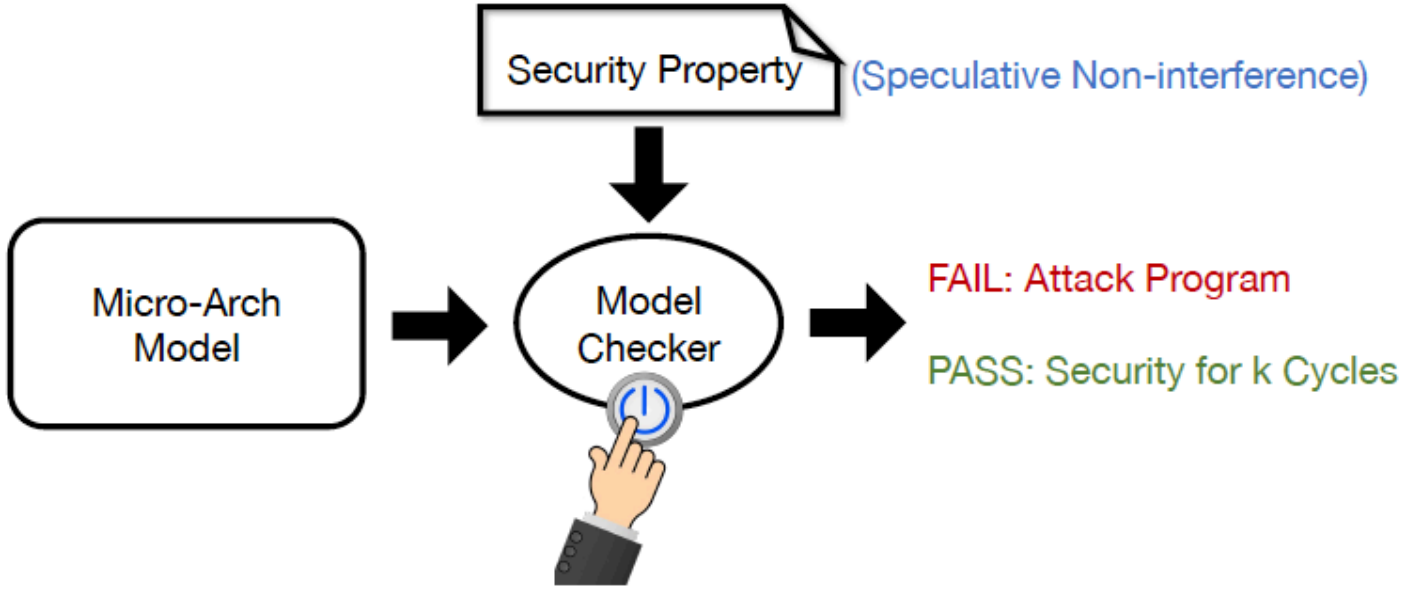
- PI: G. Edward Suh



Next: Assessing Security at Early Stage Microarchitectural Design

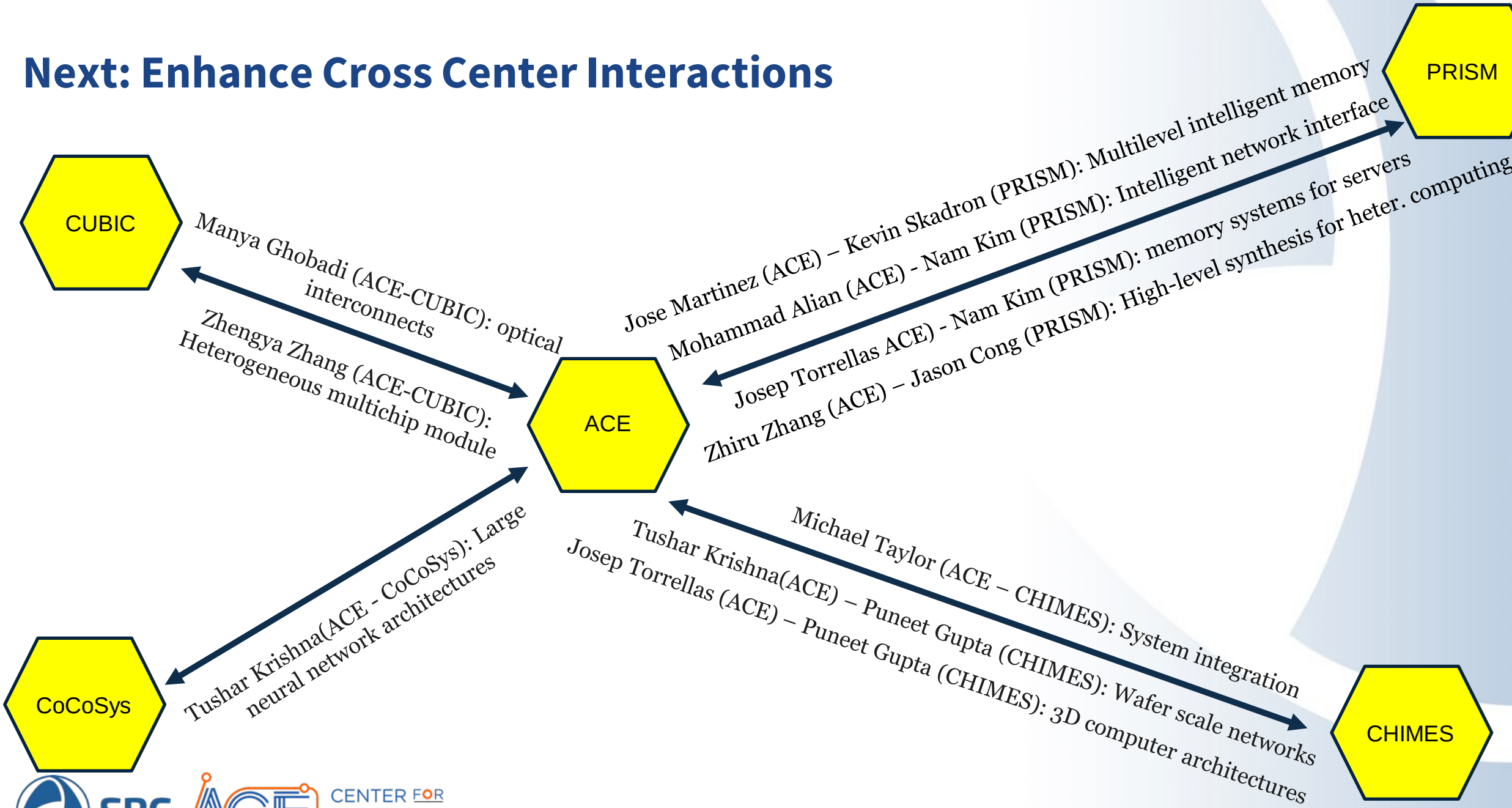
- Problem: A new mitigation mechanism is effective only until a new attack breaks it
- Goal: Model checking framework to automatically find possible vulnerabilities in early-stage microarchitectures
  - Leverages symbolic execution modeling discipline based on formal methods
  - Use a modular structure and handshaking understandable by computer designers [ISCA 2023]

- PI: Mengjia Yan



We will interface it with our existing work on Quick & Thorough Pre-silicon Verification for Evolving Designs

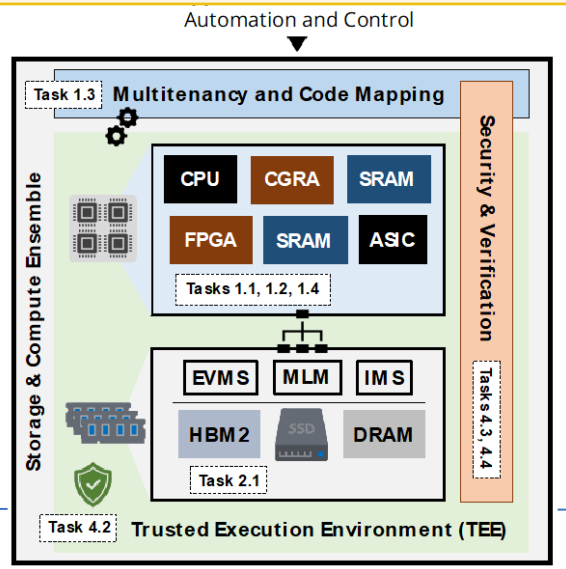
Next: Enhance Cross Center Interactions

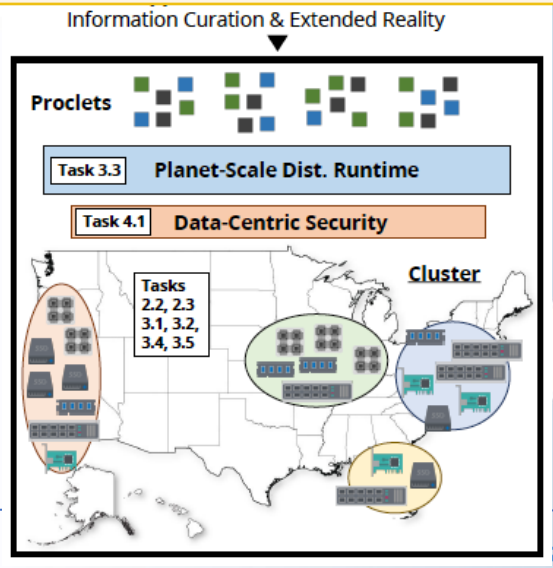


SRC ACE CENTER FOR EVOLVABLE COMPUTING

What will Success Be for Next Year

Work towards two demonstrators that integrate multiple innovations
Deepen existing research connections with JUMP 2.0 member companies and produce impactful results
Robust student engagement with JUMP 2.0 member companies in the form of mentorship, collaboration, and internships



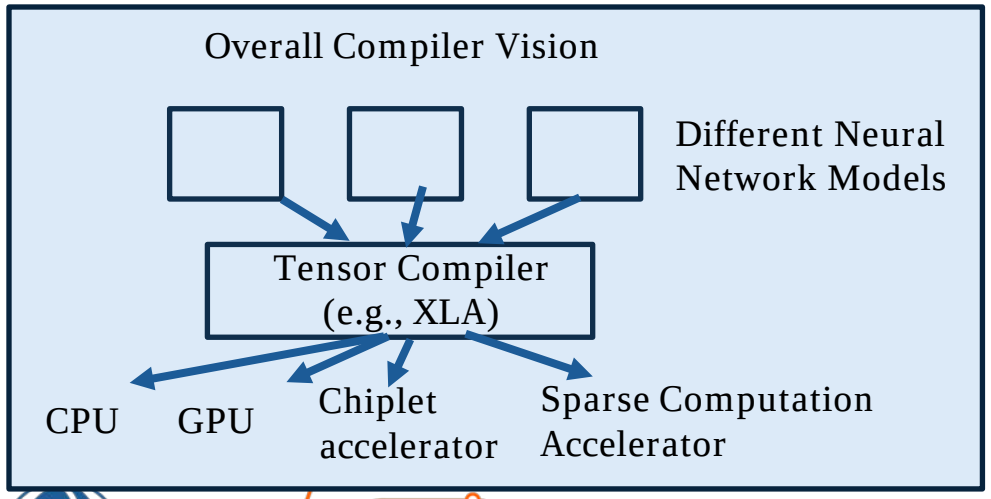


END

Optimizing, Compiling and Mapping Neural Networks to Accelerators

- We propose *TGOpt* optimizer: performs optimizations to eliminate redundancies in Temporal Graph Attention Networks
- Significant performance gains over many temporal graphs on both CPU and GPU
- We generalize TGOpt to a lightweight framework called *TGLite* with novel data abstractions for any temporal graph neural network (GNN)

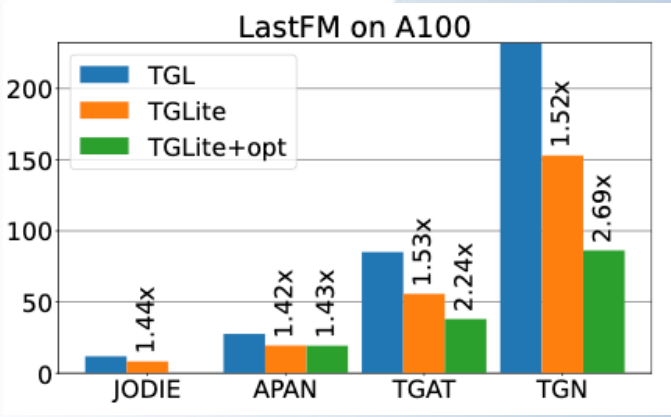
- PI: Charith Mendis; student: Yufeng Wang



Temporal GNN: GNN that changes over time

Use of temporal GNNs:
- Fraud detection: detecting anomalies in financial networks
- Social interaction analysis: analyzing how users interact in social networks

Performance comparison with TGL (state of the art system)

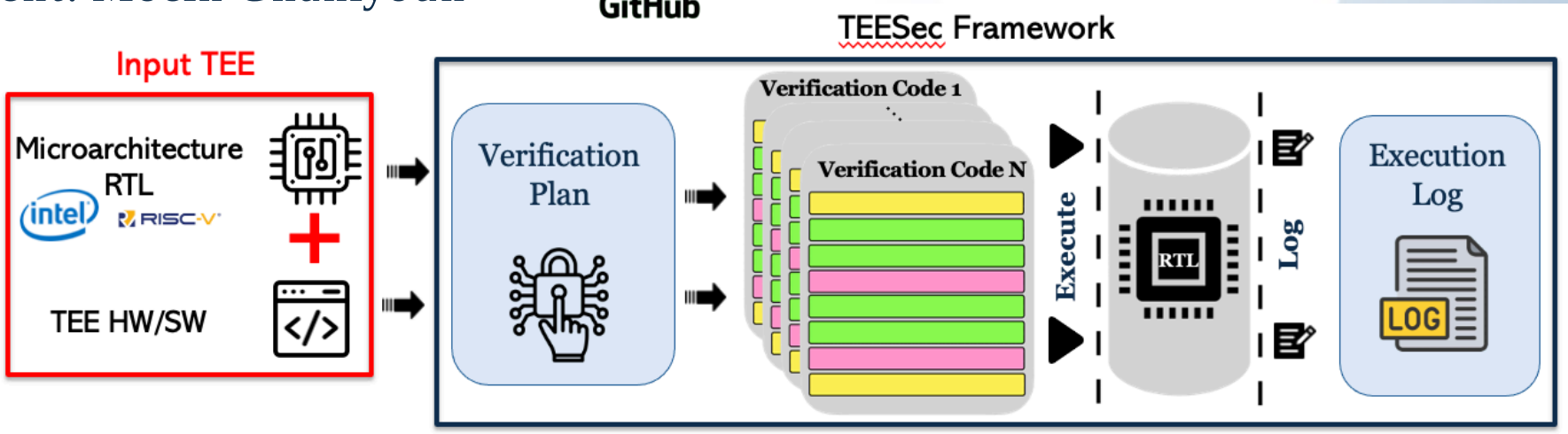


New Tool Released: Pre-Silicon System-Level Security Verification for Trusted Execution

- TEE vulnerabilities: interplay between microarchitecture/system/software
- Introduce *TEESec*: Joint verification of TEE HW/SW and microarchitectural implementation – pre-silicon, using detailed RTL simulation
- Uncovered 10 security violations in 3 open-source RISC V cores
- Open-sourced the tool in July 2023

PI: Radu Teodorescu; student: Moein Ghaniyoun

GitHub https://github.com/MoeinGhaniyoun/TEESec

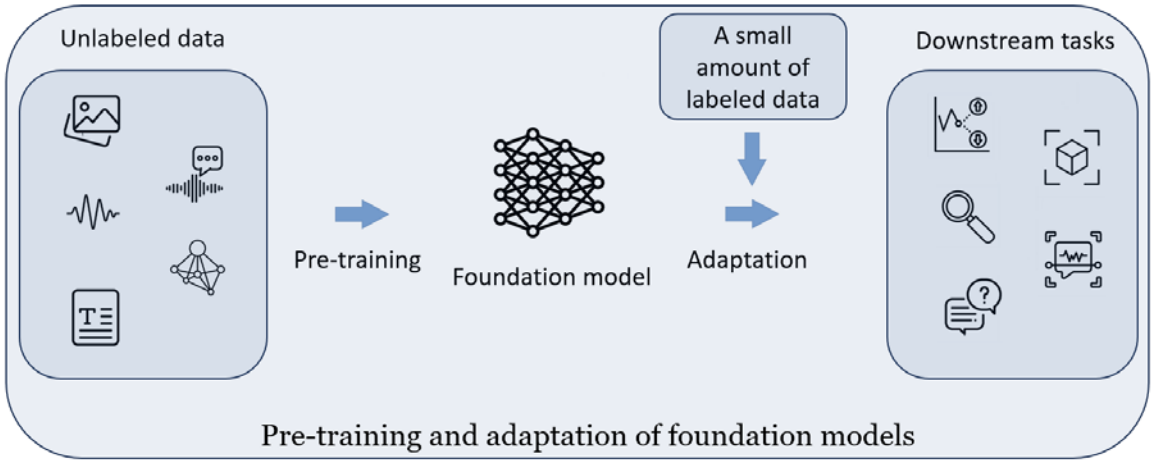


SRC ACE CENTER EVOLVABLE COMPUTING

*Ghaniyoun et al., TEESec: Pre-Silicon Vulnerability Discovery for Trusted Execution Environments, ISCA 2023

Towards Foundation Models for Internet of Things Applications

- Foundation models are an important self-supervised technique for training neural networks to accomplish domain-specific downstream tasks (with fine-tuning)
- We developed a simple foundation model for *vibrometry* (identifying events from their vibration signatures) to be used as application benchmark. It shows great improvements in downstream classification tasks when deployed in different environments than it was trained.
- PI: Tarek Abdelzaher; students: Yigong Hu



The Foundation Model Concept (top)

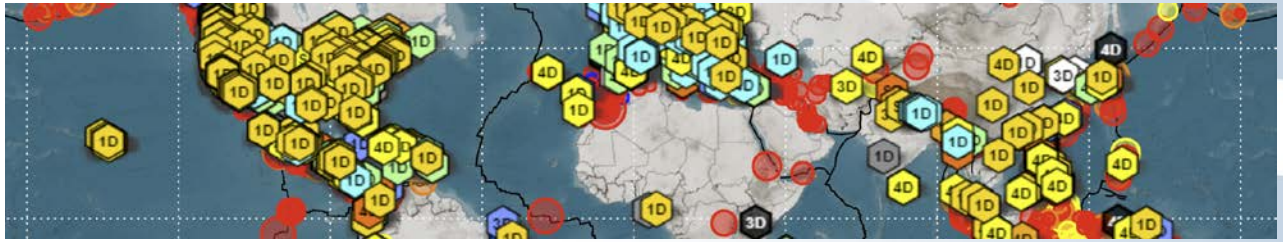
Pretraining: Data from deployed geophones (right figure) for earthquake measurements that happen to pick up numerous other nearby events (car traffic, pedestrian activity, etc)

Testing challenge: Vehicle classification from its vibration (geophone) signature, after fine-tuning with limited labeled data.

Model	Target Classification Accuracy (Same Environ. as Trained)	Target Classification Accuracy (Diff. Environ. as Trained)
Supervised (DeepSense)	83.8%	17.6%
Foundation Model	80.2%	40.3%

Computational bottleneck: Multi-head attention in transformer model:

- Complexity of multi-head attention: $O(n^2)$, where n is the input length.
- Larger n allows better performance but increases computational complexity



SRC ACE CENTER FOR EVOLVABLE COMPUTING

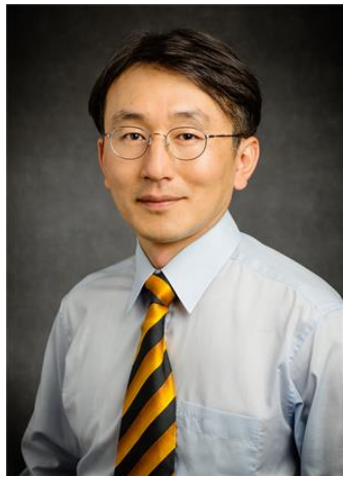
Semiconductor Research Corporation

PRISM

JUMP 2.0 PRISM Center Highlights & Plans

Center director: Tajana S Rosing, UCSD

Center co-director: Nam Sung Kim, UIUC



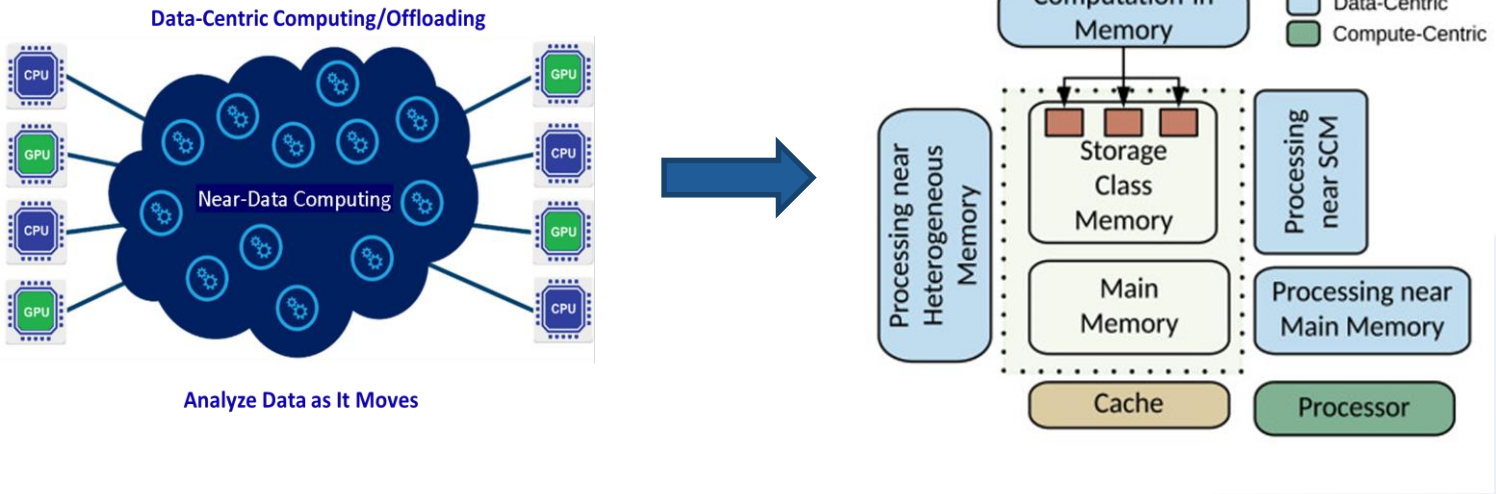
PRISM Goals

Solve fundamental IMS scale out and scale down challenges for 2030

By creating a novel IMS architecture that:

- Answers when, where and how to store and process which data
- Seamlessly integrates diversity of memory, storage, compute & software
- Holistic cross layer IMS optimization from devices to applications

Demonstrate 100x improvements using grand challenges

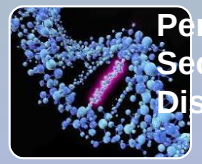


PRISM – Processing with Intelligent Storage and Memory

Director: Tajana Rosing, UCSD **Co-director: Nam Sung Kim, UIUC**

Theme 4: Grand Challenges

Leads: Vijay Narayanan
Yizhou Sun


Personalized & Secure Drug Discovery

Deep Insights

Theme 1: Systems & Software

Leads: Emmett Witchel & Ada Gavrilovska


Platform Abstractions


AIMS Controls


Scheduling & Placement

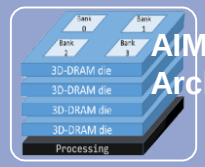

Programming & Compilers


Security & Privacy

Theme 2: Architecture

Leads: Nam Sung Kim & Jishen Zhao


Memory & Storage Architecture



AIMS Architecture


Controllers & Interfaces

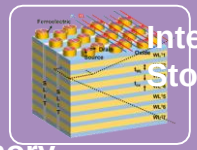

Disaggregated IMS Design


Hardware Security

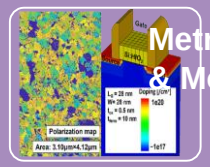
Theme 3: Devices & Circuits

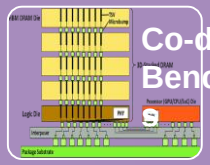
Leads: Suman Datta & Shimeng Yu


Intelligent Memory


Intelligent Storage


Emerging Memory


Metrology & Modeling


Co-design & Benchmark

Cross-cuts

Performance
Programmability
Energy Efficiency
Security
Scalability
Virtualization
Composability
Reliability
Resilience
Availability

23 PIs, 13 universities, 118 students

PRISM Highlights – KPI 35.3!

20 Awards

- **Kevin Skadron** - 2023 SIA-SRC University Researcher Award
- **Jason Cong** - EDAA Achievement Award 2023, Recipient of the "Global Industry Leader" Award from ChipEx'2023
- **Yizhou Sun** - IEEE AI's 10 to Watch
- **H.-S. Philip Wong** - Received the Test of Time Award of the Symposium on VLSI Technology and Circuits

41 Keynotes & Invited Talks

- **Suman Datta –** "A System Driven Approach to Semiconductor Innovation," EDTM Keynote
- **Jason Cong** - "Democratizing IC Designs -- with Deep Learning and Automated Microarchitecture Optimization" at University of Florida (DISTINGUISHED LECTURE)
- **Nam Sung Kim** - Joint tutorial on on-chip accelerators with Intel at ISCA'23
- **Yizhou Sun** – "Neural-Symbolic Reasoning on KGs" Keynote at The WebConf Knowledge Graph Special Day
- **Tajana Rosing** - "Accelerating Bioinformatics Workloads" Keynote at MPSoC 2023
- **Vikram Adve** – "Automating Retargetable Compiler Construction with Hydride," Keynote at Compiler Construction 2023
- **Fredrik Kjolstad** - "Portable Compilation of Sparse Computation," Keynote at PLDI DRAGSTERS
- **Sang-Woo Jun** - "Reconfigurable Hardware Acceleration Why and How" at University of Lisbon
- **Ada Gavrilovska** - "Simplifying Management of Complex Memory Fabrics" at Barcelona Supercomputing Center
- **Nam Sung Kim** jointly with Intel will do a CXL Type 2 use case demo at SC'23 offloading Linux kernel features to processing engines in a CXL device

30 News articles published

Publications: 270 datasets in Pillar Science

Students to sponsors: 17 internships, 9 full time

Industry liaisons & viewers: 95 total, 60 are liaisons

>300 interactions with member companies by PRISM team

PRISM Github: https://github.com/PRISM-T4-Grand-Challenges

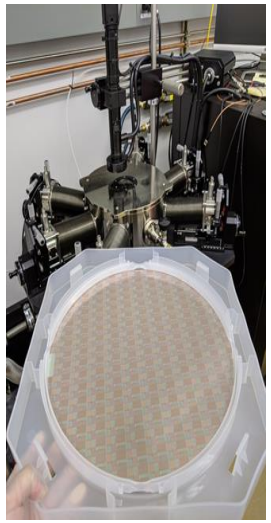
Broaden Participation: 20 undergrads & 8 high school students

86 Student Posters &
25 Demos at the Review!

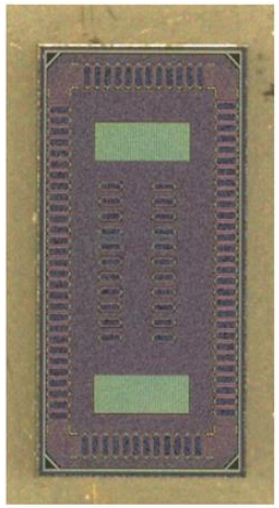
Collaboration with other JUMP 2.0 centers via **SEED projects**

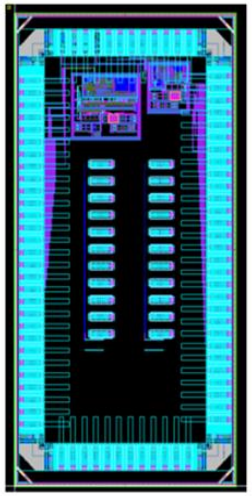
- Cognition - CoCoSys:
  - Patrick McDaniel: Intrusion detection in hosts and networks – Demo
  - Jason Cong: Near storage acceleration of clustering at scale - Poster
- Communication – CuBIC
  - Nam Sung Kim & Tajana Rosing: high bandwidth photonic connectivity to memory; deeply disaggregated connectivity architectures
- Distributed Sys & Arch – ACE:
  - Jishen Zhao: Tiering of Serverless Snapshots for Memory-Efficient Serverless Computing – Demo
  - Ada Gavrilovska: End-to-end In-Fabric Programming for Graph Analytics – Demo
  - Kevin Skadron: Integrating Heterogeneous PIM in Disaggregated Systems – Demo
  - Jason Cong: Improving high-level synthesis – Demo
- Integration – CHIMES:
  - Nam Sung Kim: Reconstituted Wafer-based Heterogeneous Integration Technology Tailored for Memory and Storage Devices - Poster
  - Shimeng Yu: 3D NAND acceleration of Mass Spectrometry – Demo
- Devices - SUPREME:
  - Vijay Narayanan: Embedding Security into FeFET NAND Array via Intrinsic Memory- Poster
  - Suman Datta: LCO based selector materials growth and device fabrication - Poster

PRISM Chips & Devices Gallery

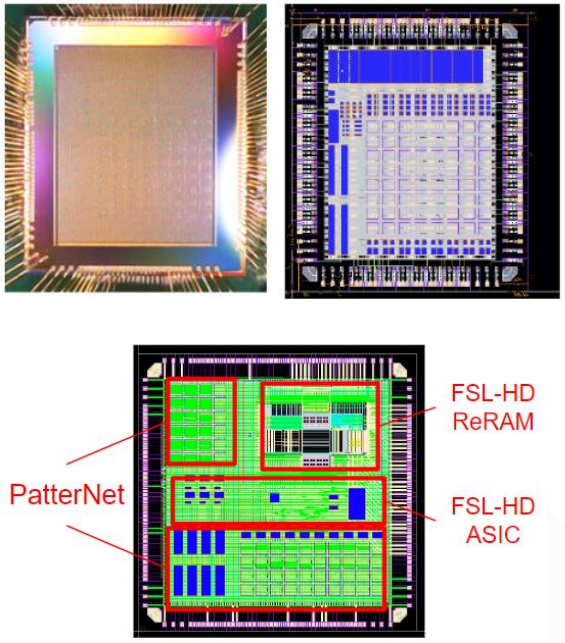


PI Shimeng Yu
Courtesy 300mm wafer from Global Foundries 28nm FeFET process for nvCap characterization
(with GF Dresden team)

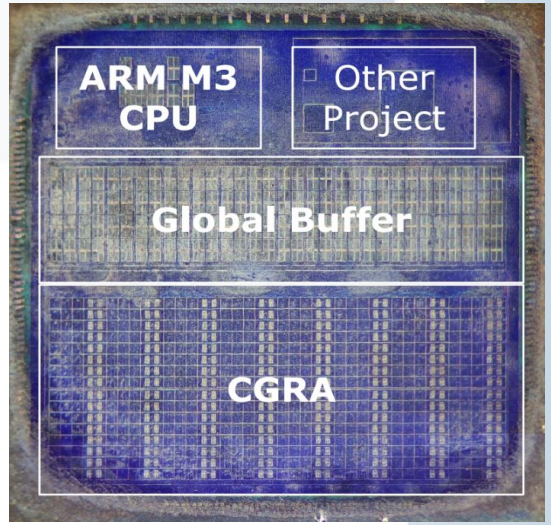




PI Shimeng Yu
1st gen FeFET PIM macro in GF 28SLPe shuttle via Fraunhofer IIS MPW shuttle w/ FeFET risk manufacturing

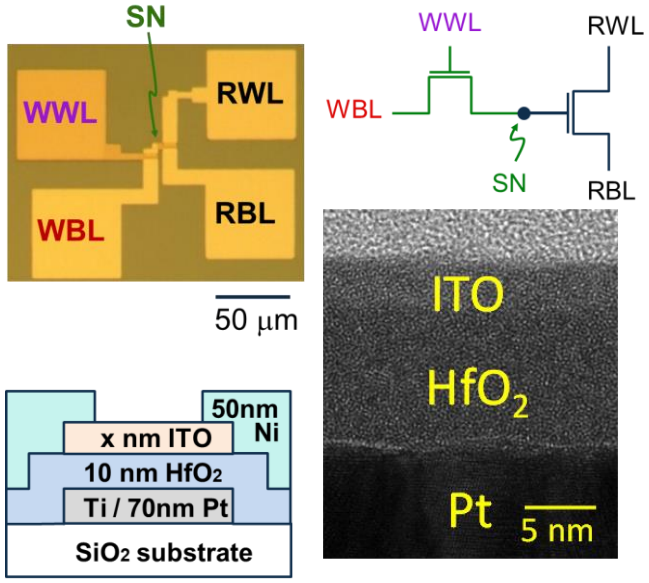


PI Tajana Rosing
40nm ASIC & ReRAM HDnn - Few shot learning for ImageNet size images
Collaboration with TSMC

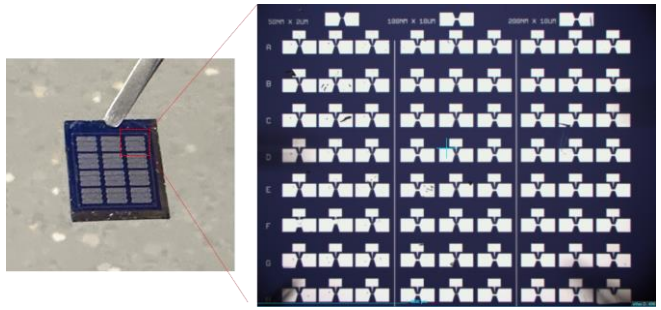


PI Priyanka Raina
First AIMS data processor for programmably accelerating both dense and sparse applications

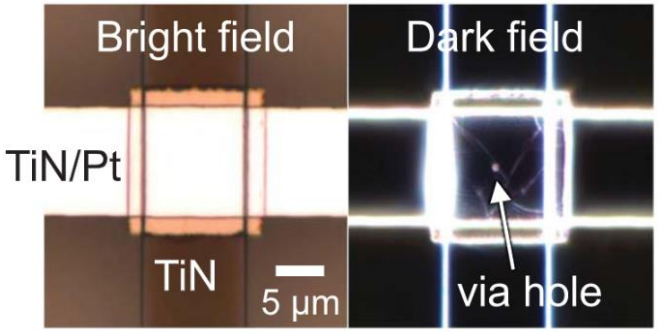
PRISM Chips & Devices Gallery (cont.)



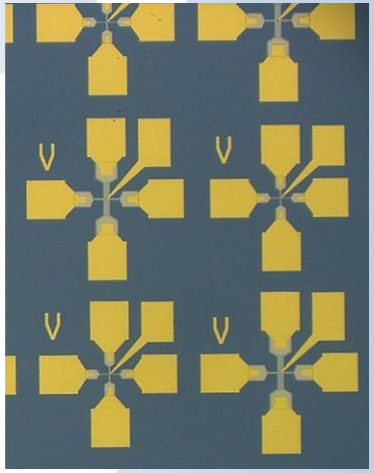
PI Philip Wong
Oxide Semiconductor Gain Cell fabricated at Stanford



PI Datta
Chip micrograph of asymmetric dual-gate (ADG) Ferroelectric memory cell array



PI Pop
Superlattice phase change memory fabricated at Stanford University



PI Salahuddin
Top view of 3D Interlayer-Exchange-Coupled Memory Devices

Theme 4: Drug Discovery Code & Data Repository

Dataset Description	Application	Description	Git Repository	PI
Proteomics Datasets (A, B: Kidney cells, 1.1M, 5.6-25GB; C: HeLa, 4.1M, 54GB; D: HEK293, 4.2M, 87GB; E: Human Proteome, 21.1M, 131GB)	Hyper-Spec	GPU enhanced computing to rapidly cluster spectra data in binary hyperdimensional space for improved quality and speed.	https://github.com/PRISM-T4-Grand-Challenges/Hyper-Spec	Rosing
Proteomics Datasets (Small: Yeast & Human HCD, 1.16M spectra, iPRG2012; Large: Human HCD, 2.99M, HEK293)	HyperOMS	GPU enhanced Open Modification Search for MS database search	https://github.com/PRISM-T4-Grand-Challenges/homs-tc	Rosing
Proteomics Datasets (A, B: Kidney cells, 1.1M, 5.6-25GB; C: HeLa, 4.1M, 54GB; D: HEK293, 4.2M, 87GB; E: Human Proteome, 21.1M, 131GB)	SpecHD	FPGA-based computing to rapidly cluster spectra data in binary hyperdimensional space for improved quality and speed.	https://github.com/PRISM-T4-Grand-Challenges/SpecHD	Rosing
Human reference genome, Viral and Bacterial datasets	GenoMix	FPGA Accelerated Simultaneous Analysis of Human Genomics, Microbiome Metagenomics, and Viral Sequences, published at BioCAS'23	https://github.com/PRISM-T4-Grand-Challenges/gemomix	Rosing

Theme 4: Deep Insights: Code & Data Repository

Dataset Description	Application	Description	Git Repository	PI
CUB-200-2011 (image dataset with 200 bird species)	Fine-Grained Image Classification	The process of distinguishing objects into precise subgroups within a larger category, requiring recognition of subtle visual distinctions.	https://github.com/PRISM-T4-Grand-Challenges/token-adaptive-vision-transformer	Narayanan
Stanford Cars (image dataset with 196 classes of cars)				Narayanan
Stanford Dogs (image dataset with 120 breeds of dogs)				Narayanan
NABirds (image dataset with 400 species of birds)				Narayanan
iNaturalist (image dataset with 5089 natural fine-grained categories)				Narayanan
NYU Depth Dataset v3 (RGB and Depth camera video sequences from a variety of indoor scenes)	Depth Estimation	The process of determining the distance or 3D spatial information of objects within an image or scene.	https://github.com/PRISM-T4-Grand-Challenges/depth-completion-gan	Narayanan
ShapeNet (image dataset with 3135 3D CAB models)			https://github.com/PRISM-T4-Grand-Challenges/robust-multimodal-fusion-gan	Narayanan
MVSEC (depth image dataset collected with Multi Vehicle Stereo Event Camera)	Multi-modal Depth Estimation		https://github.com/PRISM-T4-Grand-Challenges/ER-F2D	Narayanan
DENSE (depth image dataset with synthetic events)			https://github.com/PRISM-T4-Grand-Challenges/ER-F2D	Narayanan
Benchmark dataset (OKVQA: 14,055 open-ended questions); Knowledge-source (Knowledge Source Corpus Size Type of Text Avg. Text Length WIT [37] 5,233,186 Wikipedia Passage 258 CC12M [5] 10,009,901 Alt-Text Caption 37 VQA-V2 [12] 123,287 Question Answer 111 WikiData [40] 4,947,397 Linearlized Triplets 326)	REVEAL: Retrieval-Augmented Visual Language Model	an end-to-end Retrieval-Augmented Visual Language Model (REVEAL) that learns to encode world knowledge into a large-scale memory, and to retrieve from it to answer knowledge-intensive queries.	https://github.com/google-research/scenic/tree/main/scenic/projects/knowledge_visual_language	Sun
HLSYN (HLS design dataset with 42,000 labelled design over 42 unique kernels/programs)	ProgSG: HLS Design Automation	a multi-modality model with GNN and language model to predict the quality of HLS design to facilitate HLS design automation	https://github.com/PRISM-T4-Grand-Challenges/software-gnn	Sun

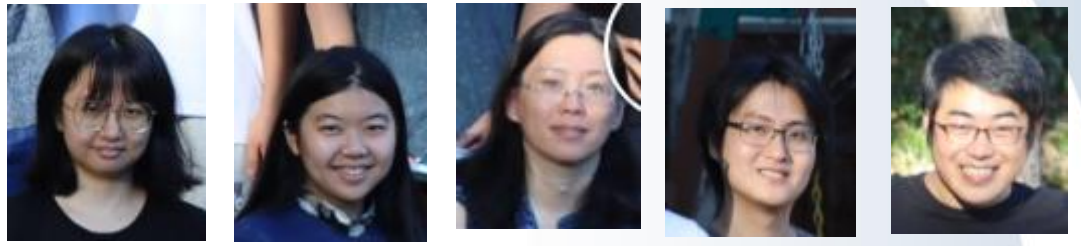
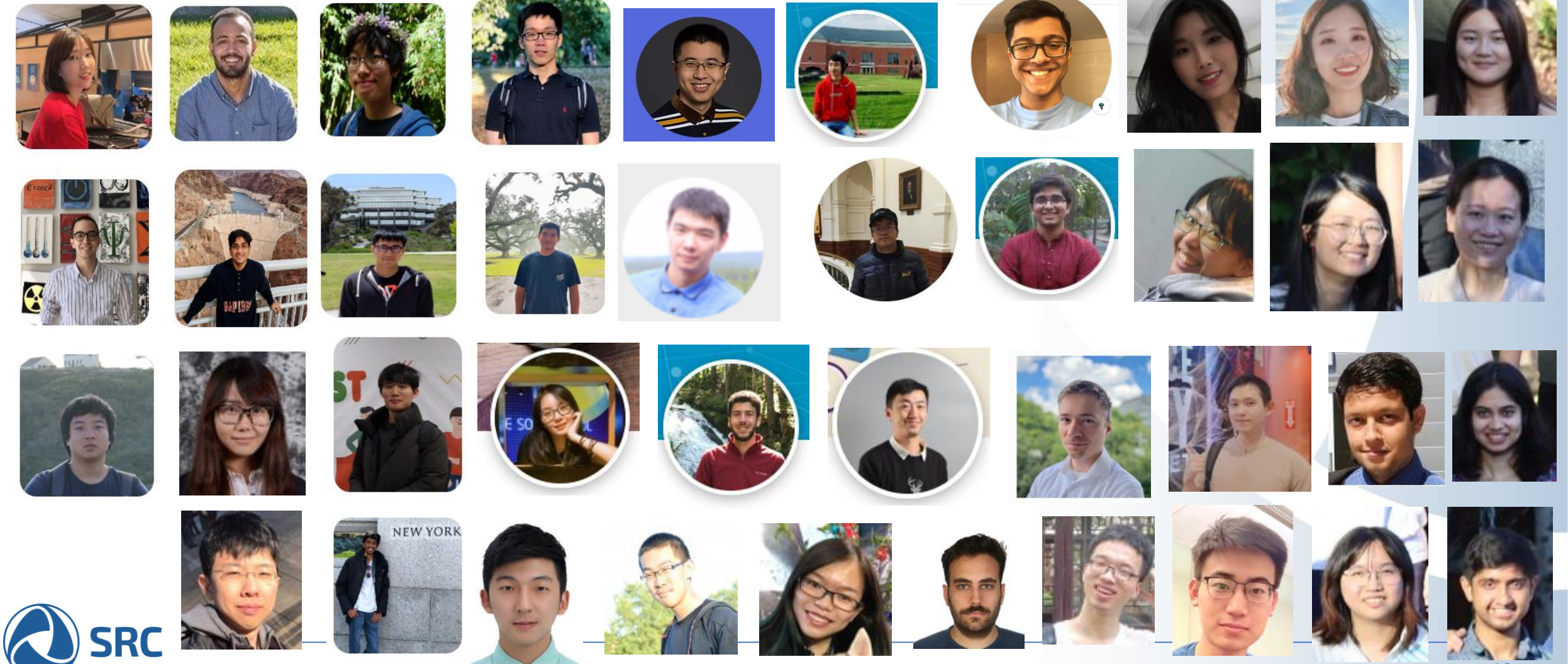
PRISM Year 1 Lowlights

- Center-wide CXL infrastructure slow to set up
  - Took much longer than expected → CXL server arrived a few months ago
  - NDA in progress to get CXL-enabled memory
- Liaison Theme Meetings
  - Not very efficient
- Time too short for demos and posters at the annual review

PRISM Center Year 2 Plans

- Center-wide CXL-enabled PRISM systems fully operational
  - Integration of accelerators into PRISM system via CXL, including software (compiler, workload placement/management) using large scale data
  - Demonstrate benefits of PRISM systems using grand challenge applications in demos that include contributions of all four themes
- Demonstrated results of expanded collaboration with other JUMP 2.0 centers, integrated into PRISM system
- Cross-center integration of security: systems, accelerators, CXL etc.
- Integration of PRISM systems with SRC member companies systems
  - Improve liaison theme/project meetings (2x per month)
- Growth of our broadening participation program
  - Undergraduate and high school research symposium in Spring'24; NSF REU

Many thanks from >118 students & 23 PIs from 13 universities!

Greetings from Penn State & Georgia Tech

Center for Heterogeneous Integration of Micro Electronic Systems
Highlights/Lowlights/Top Plans for Y2

Madhavan Swaminathan (Penn State), Center Director
Muhannad Bakir (Georgia Tech), Assistant Director
Rohit Sharma (Penn State), Managing Director

Center Year 1 Goals & Accomplishments – Year 1

Goal	Accomplishment
1. Revisit the four themes in the center, realign metrics and goals along with any changes in the plan and technical approach based on SAB Feedback.	✓ Accomplished. Modifications to themes and tasks proposed at the April ’23 SAB meeting. Approved by SRC and the information on research pillar has been updated.
2. Cluster the research tasks so that there is greater synergy between the tasks and PIs.	✓ Excellent Progress. Power Delivery (Theme I, II, III, IV); Photonics (Theme II, III); Wireless (Theme III, IV)
3. Ramp up the center through the recruitment of students and post docs.	✓ Accomplished. Our goal was 83-85 students. **We are at 85 students.**
4. Start discussions with other center directors for initiating inter-center collaborations. Also, set-up theme meetings to gather input from the industry liaisons on our roadmap and any changes required.	✓ Excellent Progress. Collaborations with ACE, CUbiC, PRISM, SUPREME, CogniSense established and growing.
5. Develop a Research Prototype Vehicle (RPV) that enables us to evaluate new technologies at an early stage, determine if these developments need to be continued, directions changed or should be sunset.	✓ One of the Top Plans for Y2. ▪ Discussed at PI meeting on May 22, 2023. ▪ Discussion w/ Theme Leaders on Nov 20, 2023

Year 1 Highlights

- Research Output
  - Papers: 58
  - Invited & Keynotes talks: 84
  - Awards: 24
  - Software releases: 3
- Participation in 3 Packaging Roadmaps (NIST)
  - MAPT (SRC), MAESTRO (iNEMI), MRHIEP (Semi/UCLA)
- Annual Review Sep 5-6, 2023
  - 160+ participants (phenomenal participation)
- Broadening Participation
  - Establishment of student council (7 members + Champion)
- SAB Feedback from Annual Review
  - Role model for other centers
- CHIPS Act
  - Several presentations to Industry Advisory Committee on CHIPS science and technology needs
  - Subu Iyer (UCLA), CHIMES PI becomes Director, NAPMP

Year 1 Highlights - Asia Trip

- Seminars and discussions at Samsung, SKHynix, and TSMC
- Led to several follow-up discussions with CHIMES PIs
  - E.g. Thermal Management & Thermal Interface Materials
- Experienced a 5.0 Earthquake while on the 14th floor in Taipei

Year 1 Lowlights

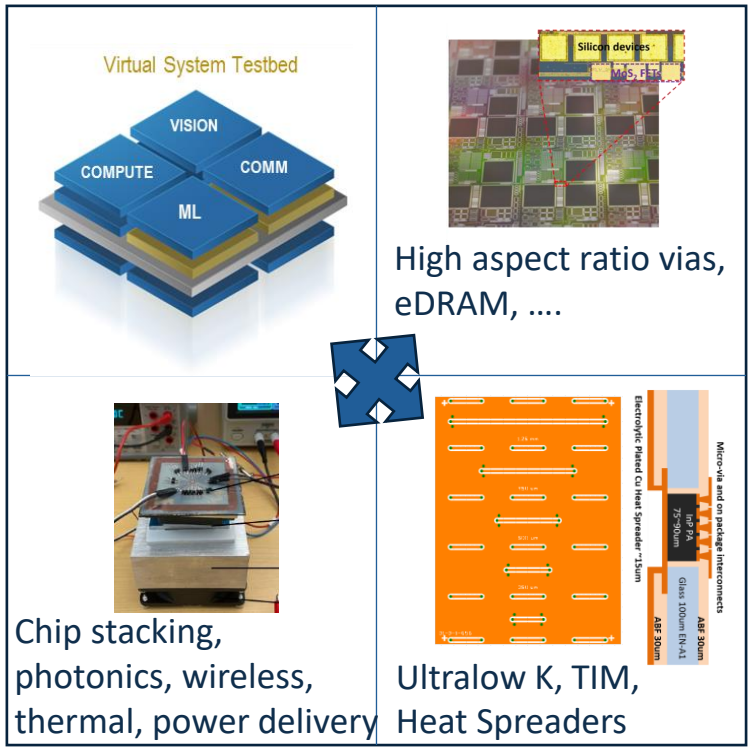
- Subu Iyer joining CHIPS – NAPMP (2-year absence)
  - Has been involved with CHIMES since White Paper
  - Helped shape CHIMES
- Limited cross theme interactions
  - Attributed to a late start due to students joining in Fall ’23
- Liaison Theme Meetings
  - Not very efficient
  - Not frequent enough

Year 2 – Top Plans

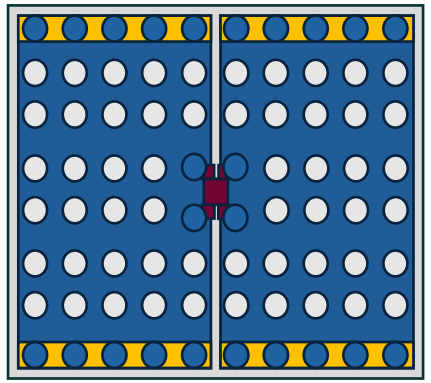
- ❑ Improve Liaison Theme Meetings
  - Increase frequency (Monthly meetings)
  - Theme Leader (15mins) + 3 Student Presentations (20mins each) + Q&A
  - Student presenters being finalized for Y2
- ❑ Intra and Inter Theme Collaborations
  - Better clustering of research tasks
  - Rapid Prototype Vehicle (RPV)
- ❑ Involvement of UG students
  - NSF proposal submitted for REU site
  - If funded, 8 UG Students + 4 G Mentors + 4 Faculty summer program
  - Hope to hear from NSF early next year
- ❑ Broadening Participation Student Council
  - Involvement in CHIMES activities (roadmaps, UG mentoring, ….)

Y2 Top Plan - Rapid Prototype Vehicle (RPV)

- ❑ Bulk Material or Component properties are easier to achieve.
- ❑ It is the interfaces and integration that are the challenge in Heterogeneous Integration.
- ❑ Purpose of RPV is to evaluate new technologies early in the cycle and only pursue ones that are promising.
- ❑ Foster Intra and Inter theme collaboration
- ❑ Enable Inter center collaboration

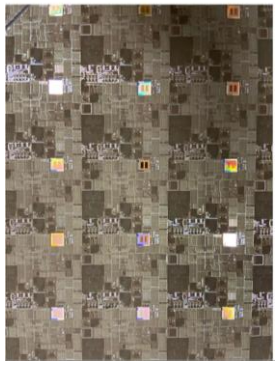


Access to CHIPLETS from other JUMP Centers

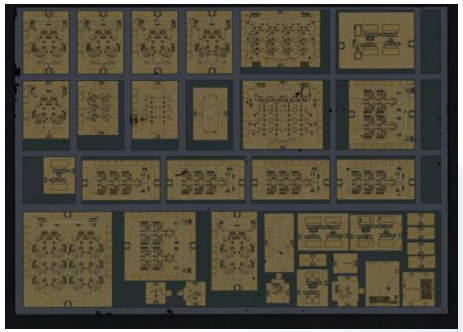


Thermal Test Chip
Goal: 2K/cm^2
Commercial: 300W/cm^2
In-house fabrication

TSMC 40nm



InP (140GHz) - CUbiC



CMOS Dies: 140GHz
Inhouse fabrication

Thank you
SRC, DARPA, Industry

www.chimes.psu.edu

Semiconductor Research Corporation

SUPeRior Energy-efficient Materials and dEvices

Grace Huili Xing (Cornell), Center Director

Tomás Palacios (MIT), Co-Director

Thomas Dienel (Cornell) Managing Director

Cornell University · MIT · The University of Texas at Austin · Rensselaer · Yale · UCSB · Stanford University · Georgia Tech · Boise State University · Northwestern University · RIT → Notre Dame · NC State

SUPREME

SUPeRior Energy-efficient Materials and dEvices

SUPREME Goals and Accomplishments—Year 1

Goal	Accomplishment
1. Center Management: Finalize all subawards for the grant and ensure that all partners meet the 95% spending goal, enable purchase of capital equipment to provide unique tools and capabilities for future success of the center, establish Pillar Science to disseminate all results, create a center website for outward facing communication, and effective communication both within and outside the Center on job opportunities, broadening participation etc.	✓ **Accomplished**
2. Team Building (PIs and students): Enable collaborative atmosphere for PIs and their groups to be aware of each other's work, broaden participation on all levels though center-wide initiatives, and to develop intracenter collaboration across thrusts and individual projects.	✓ **Accomplished**. We are at ~150 students and postdocs
3. Team Building (Liaisons, PIs, and students): Establish a culture that liaisons are extended researchers and mentors of the Center. Ensure that Liaisons and PIs are well-connected, including all team members (students, postdocs, etc.) to ensure a smooth exchange of ideas and facilitate outreach to other JUMP 2.0 centers for joint interest and expertise.	✓ **Excellent progress made; topic-focused meetings help to increase coherence and connections between materials and device projects**
4. The SUPREME Materials Palette: The materials palette is the unique identity of SUPREME, central to all aspects of the Center research thus most effective in connecting all projects, all researchers, all liaisons and research methods. Learn how to manage a dynamic materials palette, implement the cycle of materials discovery combining theoretical predictions, experimental realization, and thorough characterization to expand or down-select the materials palette. Update and present the updated Materials Palette at the annual center reviews, or other venues if applicable.	✓ **Excellent progress made** ✓ Top Goal for Year 2: Continuously updating the materials Palette
5. New research methodologies: Implement new ways of research including machine learning and artificial intelligence (ML/AI) approaches towards high-throughput methods and automated materials discovery, explore inverse design for device design whenever applicable.	✓ **Good progress made** ✓ Top Goal for Year 2: to organize topical workshops

Highlights

1. Growth of Center during year 1

- Students and Postdocs joined: 149
- Sponsor Liaisons at SUPREME projects: 98

2. Annual Review August 3-4, 2023

- 160+ on-site participants & additional 40 online
- Highlighted as a benchmark for efficient and productive event for other JUMP 2.0 Centers

3. Liaison Meetings weekly—exception apply

- Rotating through Thrusts of SUPREME
- Well-received by Liaisons
- At alternating times to balance time zones and availability
- Aim for higher student attendance and presentations

4. Broadening Participation

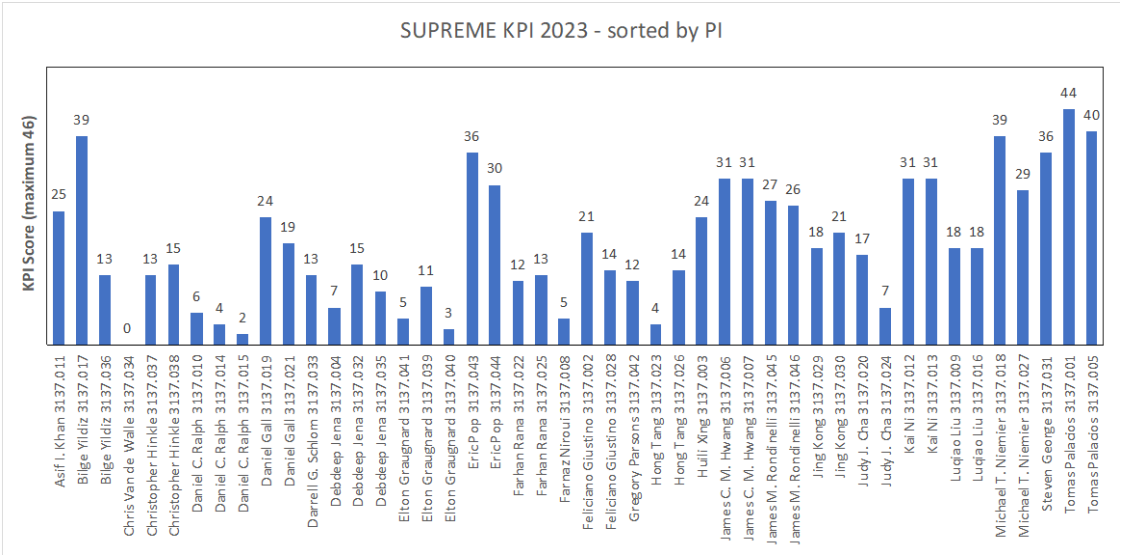
- Two BP Champions guide SUPREME's efforts
- Two awards established
  - SUPREME Undergraduate Microelectronics Fellows (5 awarded)
  - SUPREME Undergraduate Travel Grant (4 awarded)

5. **Publications**: 81+
6. **Keynotes**: 60+
7. **Awards**: 25+

Lowlights

1. Growth of Center during year 1

1. Slow process of completing subawards (last finalized October 2023)
2. Need to help improve PI's KPI, especially for those who are new to SRC



2. Varying efficiency of Liaison Meetings

1. Improve schedule to achieve larger Liaison attendance
2. Encourage student attendance
3. Continue combination of deep dive presentations and short updates
4. Move beyond thrust boundaries with focus on topical areas

3. Leveraging resources

1. PIs attempted to access the HPC systems via the DOD high-performance computing modernization program (HPCMP), but clearance seems to be a bottleneck

4. Meaningful connections with other JUMP 2.0 center

1. Need to be driven by PIs' genuine research needs
2. PIs need to have bandwidth

Year 2 - Top Plans

1. Generate concise technical documents for annual review
    1. Single document listing metrics for all materials candidates across the center
    2. Single document listing milestones for each task, & progress against those milestones
    3. Single document list of wafer-scale (8 or 12 inch) processing (in particular, deposition) capabilities
2. Highly abstracted presentations by the Center Directors and Thrust Leads
    1. Presentations with technical depth on progress, challenges and actions toward quantitative goals/deliverables
3. Improve further weekly Liaison meetings
    1. Encourage student attendance & presentations
    2. Continue engaging industry to seek out challenges to be addressed for near and midterm success
    3. Continue combination of deep dive presentations and short updates, always starting with an executive summary
    4. Continue topical areas beyond thrust boundaries
4. Generate a resume booklet of students who are looking for hire/internship
5. Further broadening participation by UGs
    1. NSF REU site proposal submitted, waiting for NSF decision in early 2024.
6. Foster connection and collaboration with other JUMP 2.0 centers

Year 2 - Top Plans

Accelerated Materials Discovery | **AI-assisted Synthesis & Manufacturing** | **Processing & Metrology**

1. **Improve fail-fast approaches**
2. **Identify target demonstrations to focus center achievements**

TMDs
$MoSe_2$, WS_2, WSe_2
MoS_2

Nitrides
AlScN, ScGaN, BN, InN
AlN, AlBN

Proton electrolyte in electrochemical RAM (ECRAM) - HfO_2, ZrO_2

Synthetic antiferromagnets
FeCoB/Pt/Ru/Pt/FeCoB

Oxide interconnect metals
$PdCoO_2$, $PdRhO_2$
$PtCoO_2$

Conventional semiconductors
Si, SiGe, Ge, GaN, SiC

2-D Materials
InN, ZnO, SnSe
h-Bi, GeSe, GeS, SnTe

Ferroelectrics
SnS, AlScN, $BaTiO_3$
$SrCaCo_2O_5$, $LaA_2Fe_3O_8$ (ordered O vacancy)

Proton channnel in ECRAM - MoO_3, WO_3, V_2O_5

Intermetallic interconnect metals
VPt_2, $MoNi_2$, $NiIr_3$
CoPt

Anisotropic spin sources
~~PtO, $Sr_3Ru_2O_7$~~
$PdCrO_2$, $Bi_2Ir_2O_7$, and IrO_2

Proton reservoir in ECRAM - Metal hydrides, PdH_x

Conductive boride-based interconnect metals - YCo_3B_2, NbFeB, Mn_2B
ReB_3

p-channel oxides
~~Ta_2SnO_6~~, ~~Ba_2BiTaO_6~~
K-doped SnO, $SrPd_3O_4$, a-PdO
PdO, ~~Ba_2PbWO_6~~ (not BEOL)

Weyl semi-metals, Heusler alloys - Co_2MnGa, Mn_3Sn

2D magnetic & topological insulator (demo only – low T):
$Cr_2Ge_2Te_6$ and $BiSbTeSe_2$

Dielectrics
SiO_2, SiN_x, SiCO, HfO_2, $HfZrO_2$.
$BaHf_{0.6}Ti_{0.4}O_3$ (BHTO) (k~150)

Topological interconnect
GaPt, CoSn, NbAs, ~~$CeCoIn_5$~~? (oxidation)

Proton channel:
Nb_2O_5, ~~Ta_2O_5~~ (low proton conductivity)

N-channel oxides
In_2O_3 with W, Mg, Sn, Ga, Ti

Molecules
$Ru(pap)_3(PF_6)_2$, $Ru(bpy)_3(PF_6)_2$

Optical interconnect
III-Vs on Si, graphene

SRC | SUPREME SUPeRior Energy-efficient Materials and dEvices

Semiconductor Research Corporation

SUPeRior Energy-efficient Materials and dEvices

Grace Huili Xing (Cornell), Center Director

Tomás Palacios (MIT), Co-Director

Thomas Dienel (Cornell) Managing Director

Cornell University · MIT · The University of Texas at Austin · Rensselaer · Yale · UCSB · Stanford University · Georgia Tech · Boise State University · Northwestern University · RIT → Notre Dame · NC State

SUPREME

SUPeRior Energy-efficient Materials and dEvices

Semiconductor Research Corporation

CUbiC

Center for Ubiquitous Connectivity

SAB Center Meeting

December 5, 2023

Director: Keren Bergman, Columbia University
Co-Director: Ali Niknejad, UC Berkeley

Cloud to Edge Connectivity Challenges

Explosive Growth in Data Communication Demands

Cloud Connectivity Challenges:

- Orders of magnitude gap between on-chip/off-chip BW
- Strong distance-dependent communication energy
- Scalability limited by energy and bandwidth tapering
- Massive heterogeneity – compute/memory/accelerator

Edge Connectivity Challenges:

- Driving mm-Wave capacity to meet data demand with robustness, reliability, mobility, and low cost
- Massive densification, power, loss, thermal cooling
- Long-range links - back-haul, long range front-haul, airborne links - limited by output power

System Connectivity Challenges:

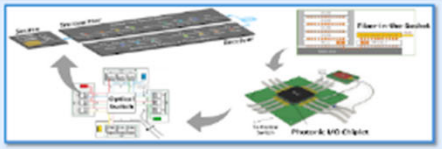
- Seamless connectivity between edge and cloud for optimized cross-layer performance
- Reconfigurable, adaptable connectivity to accelerate heterogeneous applications
- Secure and resilient connectivity across edge and cloud

SRC

CUbiC

Vertically Integrated Research Organization

Theme 2: Wireline and Lightwave Interconnects

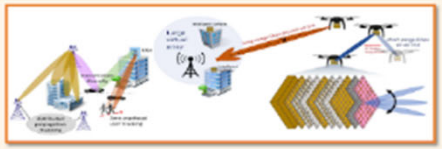


- 2.1 Systems & Algorithms for Connectivity
- 2.2 Circuits & Architectures for Links
- 2.3 Circuits & Architectures for Switches
- 2.4 Photonic Devices for Connectivity

Theme 1: Connectivity Networks and Systems

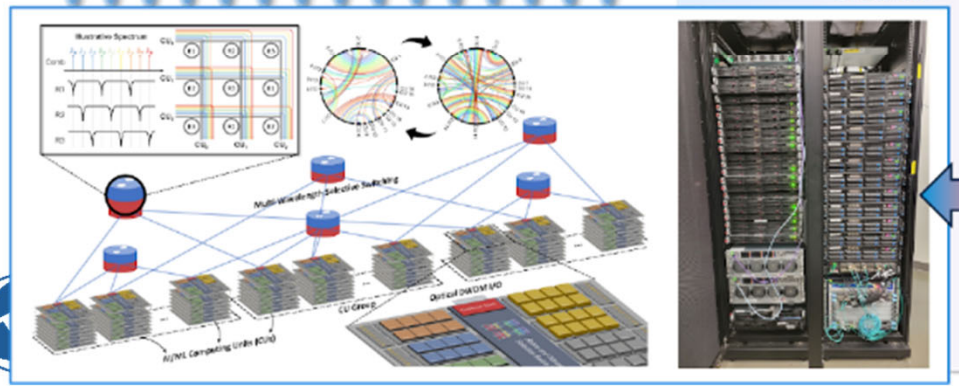
- 1.1 Terabit/s PHY Systems
- 1.2 Cross-layer Design of Terabit/s Networks
- 1.3 Security and Resiliency
- 1.4 System Connectivity Platforms

Theme 3: Wireless Circuits and Technology

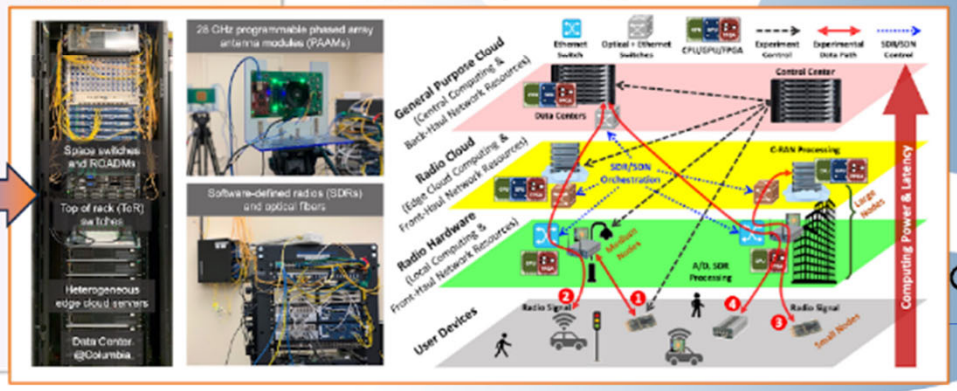


- 3.1 Large-Scale Millimeter-wave Arrays
- 3.2 Wideband HDR Analog Signal Processing Interfaces
- 3.3 High-Dimensional DSP for Emerging Wireless
- 3.4 Heterogeneous Technologies for Next Generation Wireless

SoSFab

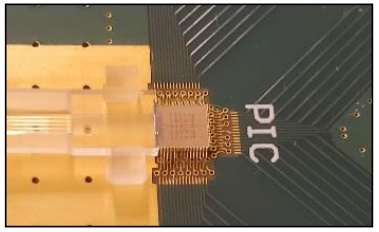


ReACT

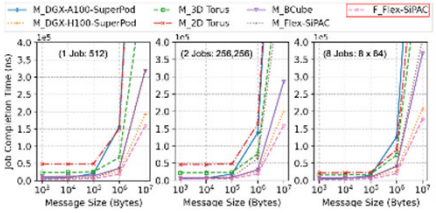


Theme 1 – Systems - Technical Highlights

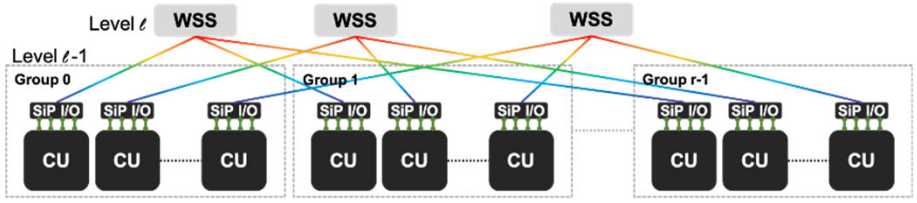
Photonic Networks



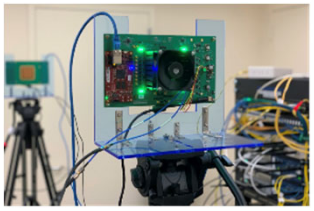
- **FlexPAC:** Packaged spatial-wavelength photonic switch for testbed experiments. Demonstrated improved performance over SOTA accelerator clusters



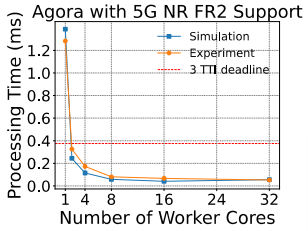
- **ALfrEd**: Develop the algorithmic and systems foundations for building application-aware clusters



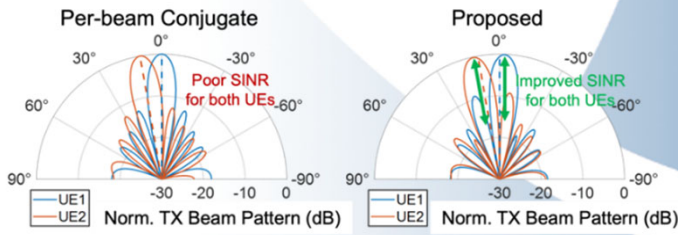
Wireless Networks



- **ReACT**: prototyping advanced system concepts with SOTA hardware; Insert T3 hardware; Focus on fundamental bottlenecks in 5G mmWave: mobility, blockage, interference



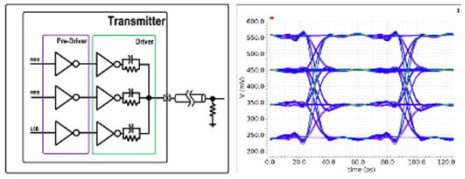
- **Architecture/algorithm** co-design guiding T3 hardware development; Tiled array architectures for scaling number of antennas and bandwidth



Theme 2 - Electrical Connectivity Highlights

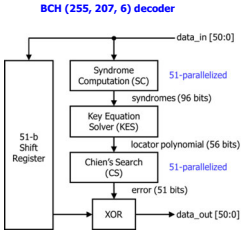
Integrated Massively Parallel Electrical Links

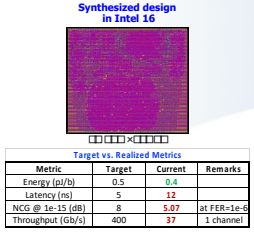
PAM4 Analysis with UCIe Channel (32 Gb/s)



Highlight: Developed a 0.4V inverter-based N-over-N voltage-mode driver achieving a 70mV PAM-4 eye-opening at 70 fJ/b energy efficiency, operating at 32 Gb/s in Intel 16nm (Tape-out - Feb'24).

Programmable Energy-efficient DSP Architectures



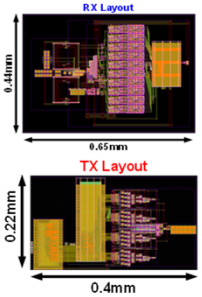


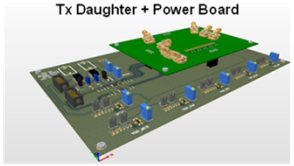
Target vs. Realized Metrics

Metric	Target	Current	Remarks
Energy (pJ/b)	0.5	0.4	
Latency (ns)	5	12	
NCG @ 1e-15 (dB)	8	5.07	at FER=1e-6
Throughput (Gb/s)	400	37	1 channel

Highlight: designed a hard-decision decoder achieving energy-efficiency of < 0.5 pJ/b at a single channel throughput of 37 Gb/s and a coding gain of 5dB in Intel 16 (Tape-out - December'23)

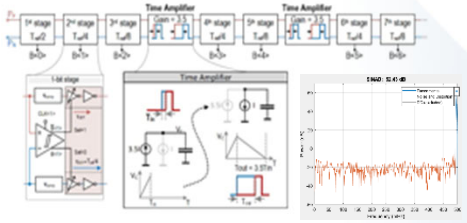
Machine Learning-inspired High-Speed Links





Highlight: designed and taped-out a PAM-4 50Gb/s transmitter and two versions of receiver in Intel 16nm targeting a FOM of 90 fJ/b/dB

Adaptive Low-Cost High-Speed ADC



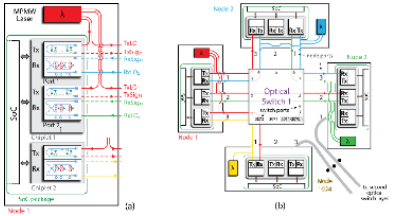
Highlight: designed a single channel ADC achieves a 56dB SNDR at ~5GHz input frequency, and 52dB SNDR at ~10GHz input frequency, with a power consumption of 27mW.

SRC

CUbiC

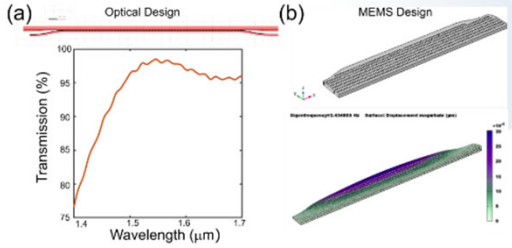
Theme 2 - Lightwave Connectivity Highlights

SuperFabric IO



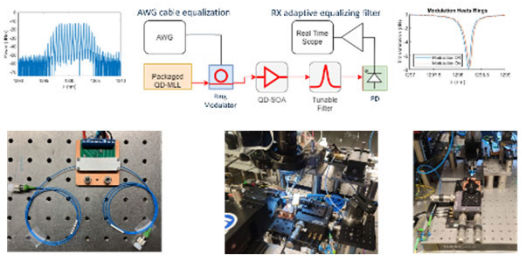
Highlight: developed laser-clock-forwarded coherent DWDM link, a pod architecture with a 1024-port optical switch, and a fast-lock simulation framework for DWDM receivers at 256Gb/s-to-1 Tb/s.

SuperSwitch



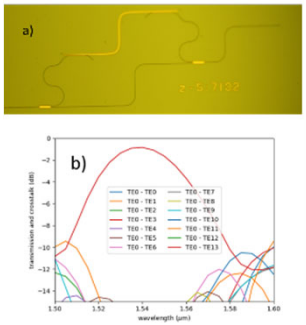
Highlight: developed a new silicon photonic MEMS design that is compatible with AIM Photonics process stack, and taped-out an 8x8 switch prototype

Fiber-In-The-Socket



Highlight: demonstrated modulation link experiment with a packaged comb source and 1 Tbps DWDM-compatible silicon photonics components to achieve BER=2.9e-5

High-bandwidth via Mode Coupling

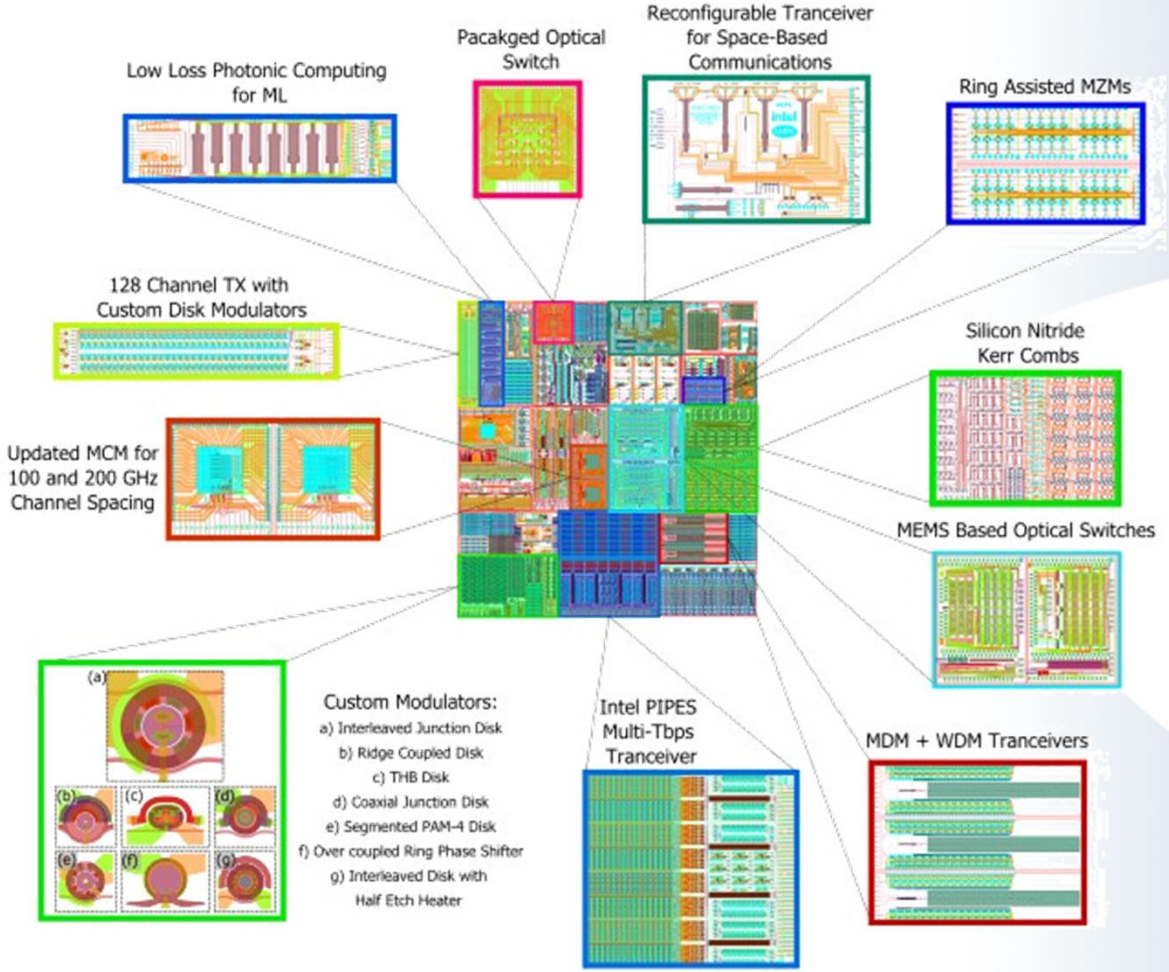


Highlight: designed compact, broadband, and robust, multimode bends with insertion loss < 0.4 dB, and crosstalk < -12 dB for all 5 supported modes over a BW of > 200 nm; and a new photonic mode converter to converted fromTE0 to TE13

Willow: CUbiC Full Photonic 300mm Wafer Tapeout! (Nov 2023)

Stojanovic

Bergman



Lipson

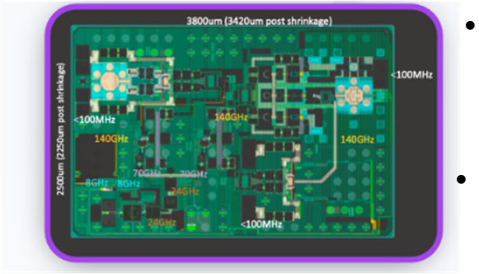
Wu

SRC

CUbiC

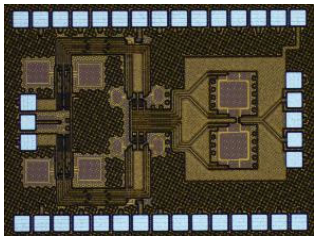
Theme 3 – Wireless Circuits and Technology Highlights

Large Scale Millimeter-Wave Arrays



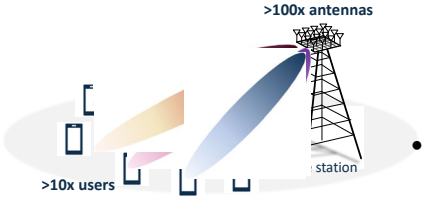
- **LAIR**: 4 Rx and Tx channels, are realized in small footprint of 2.1mm by 2.1mm
- **BIRDA**: Radar testing at 140 GHz to demonstrate the ability to cancel nearby reflections

Wideband High-Dynamic-Range Analog Signal Processing Interfaces



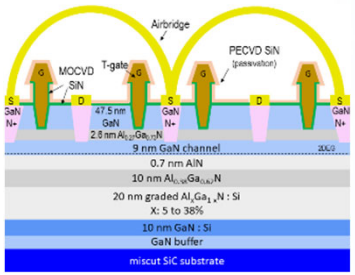
- **HI-SPAR: Co**-design of Linear Time Varying circuits with passive, low loss networks
- **T-SPAR:** Asymmetric Poly-Phase LO RX with Nonuniform Multi-level Time Approximation Filter

High-Dimensional DSP for Emerging Wireless



- **ADAPT:** Large antenna array and low load factor provide opportunity for efficient processing
- **PAYGO:** Develop a FEC decoder that provides differentiated treatment of errors.

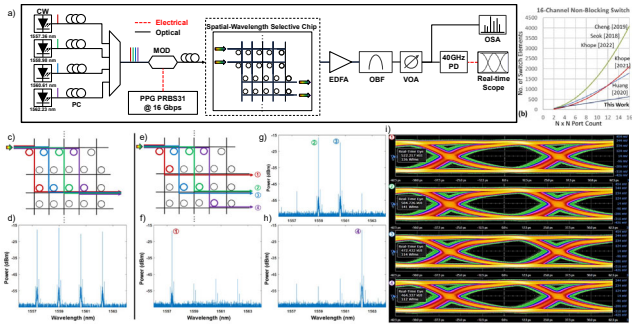
Heterogeneous Technologies for Next-Gen Wireless

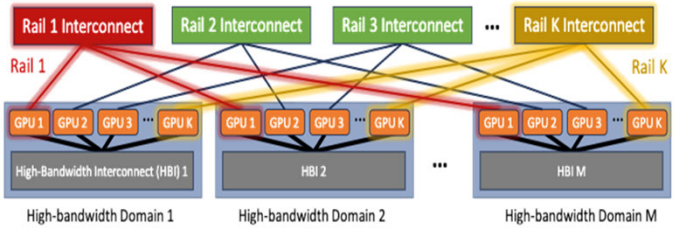


- **EVO:** Record single transistor output power at W-band
- **GOG:** A record high electron mobility of 2000 cm^2/Vs has been achieved near pinch off (charge density of $2x10^{12}$ cm^{-2})

Intra-CUbiC collaboration via SoSFab testbed – YR 2

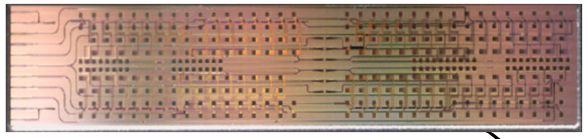
- The PIs worked towards framing the technical challenges of concept systems and applications embodying seamless, on-demand, Terabit/s connectivity.
- Significant progress towards reshaping the landscape of artificial intelligence, data center efficiency, and network optimization

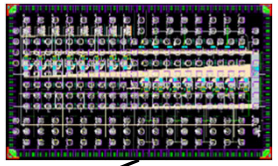




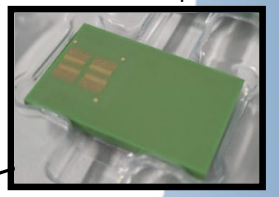
SoSFab platform: Optical Network Interface

PIC – 2 x 16-Channel

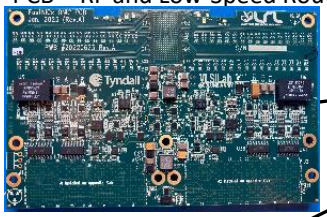


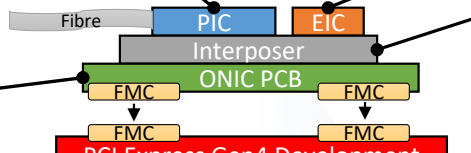


Ceramic Interposer

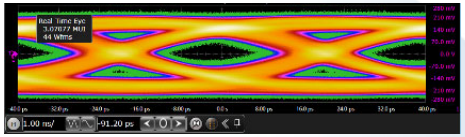


PCB – RF and Low-Speed Routing

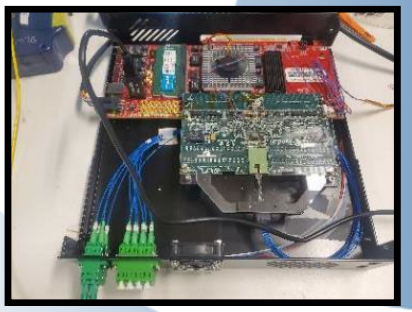




HTG 930 – PCIE to Host Server



Receiver BER < 10^{-12} up to 25Gbps/λ



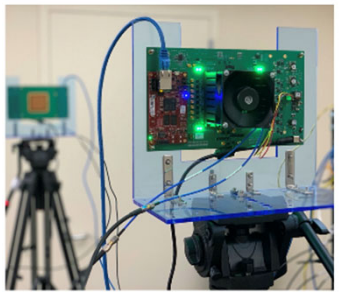
First optically-packaged ONIC, connected to HTG 930

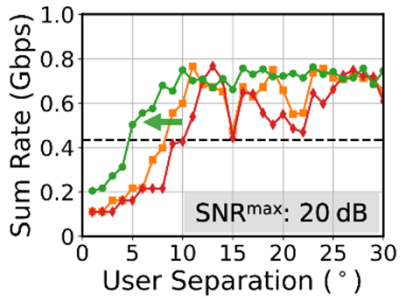
SRC

CUbiC

Intra-CUbiC collaboration via ReACT testbed – YR 2

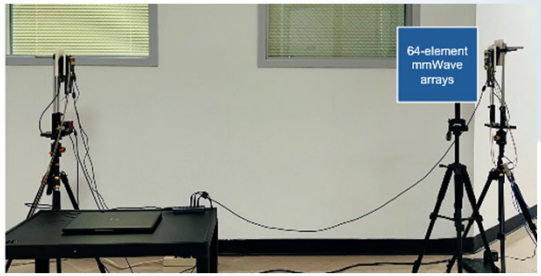
Interference-free MU-MIMO with closely spaced users

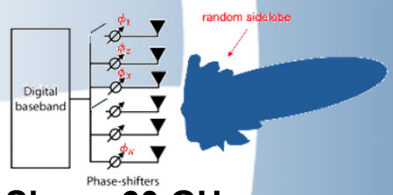




IBM 28-GHz PAAM connected to USRP SDRs

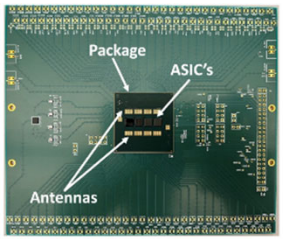
Array calibration and compressive tracking



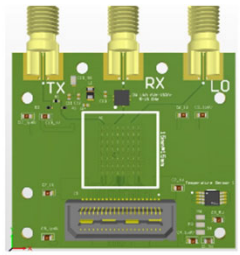


Sivers 60 GHz arrays

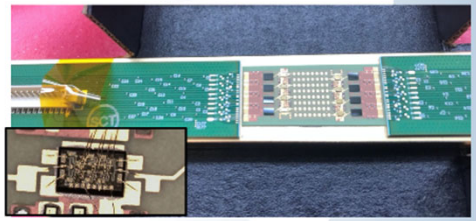
T1/T3 collaboration for Phase 2 Testbed:
Integrating CUbiC hardware



All-CMOS 140GHz MIMO Hub Module (Niknejad)

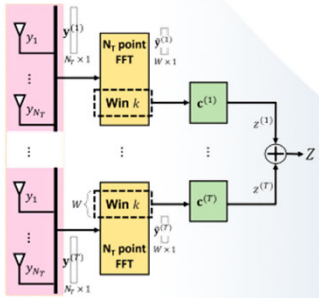


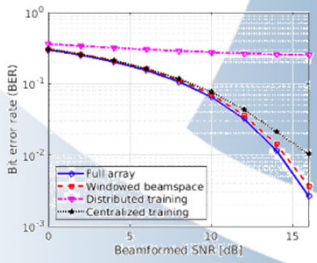
8x8 140GHz TRX Module with 50dBm EIRP (Rebeiz)



Gen-II 140GHz MIMO Hub Module (Rodwell)

T1/T3 collaboration between SPATS/ADAPT tasks
Beamspace DSP for tiled massive MU-MIMO





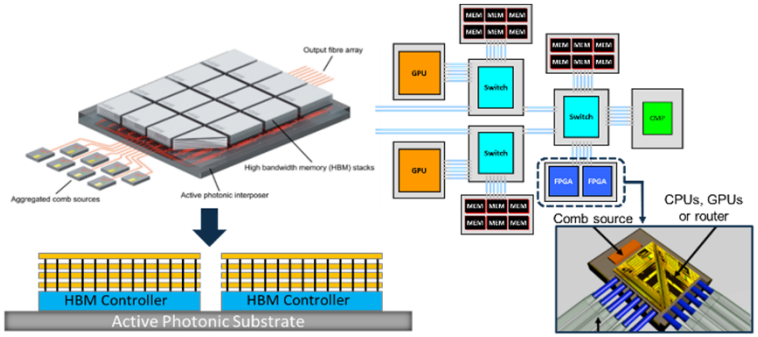
SRC

CUbiC

Cross-Center Collaborations

PRISM – CUbiC
Photonic Connected Memory

- Partner with PRISM (PIs: Zhao and Kim) for high-bandwidth photonic memory connectivity.
- Explore deeply disaggregated connectivity architectures



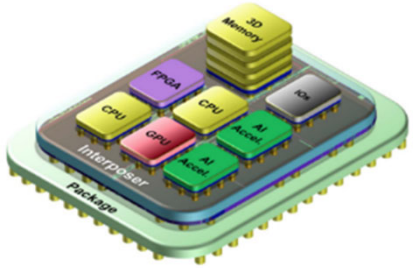

Bergman

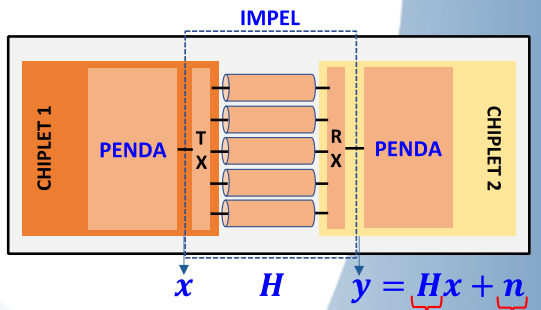

Zhao


Kim

CHIMES – CUbiC
Enabling Dense Die-to-Die Interconnect

- Collaborate with CHIMES (PI: Bakir), Samsung, ADI, and Intel to develop 2.5D/3D package models.
- Explore advanced packaging options in partnership with Intel and A*Star to enhance packaging technologies.






Shanbhag


Kumar


Bakir

SRC

CUbiC

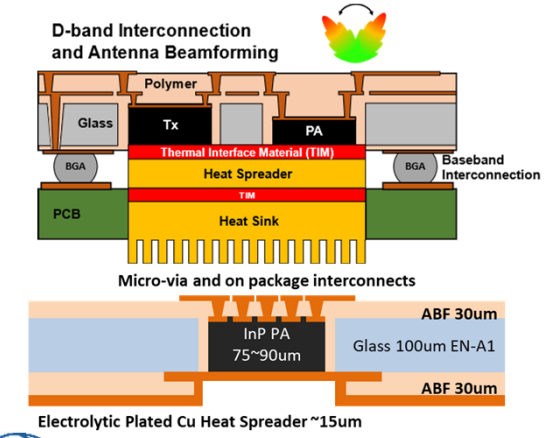
Cross-Center Collaborations

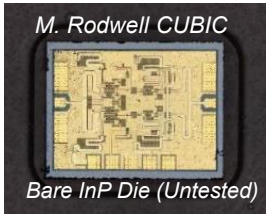
CHIMES – CUbiC
Packaging for high frequency wireless

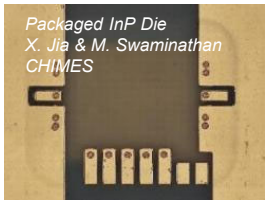
- CUbiC-CHIMES: Collaboration continuing to demonstrate functional module with beamforming.
- First demonstration of functional packaged PA @ 140GHz using embedded InP dies in glass.

Swaminathan

Rodwell



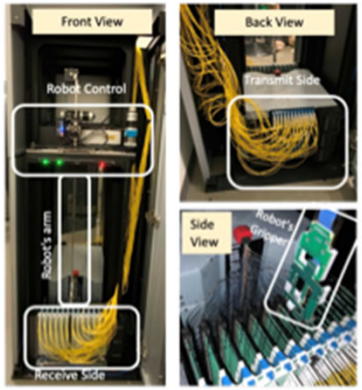


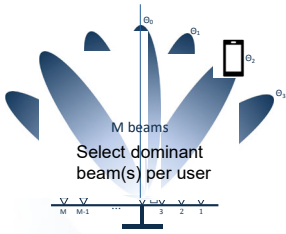


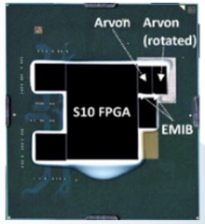
ACE – CUbiC
Accelerated AI/ML/HPC applications & Intel/Arvon co-packaged compute platform

- CUbiC-ACE collaboration drives reconfigurable datacenter architectures to accelerate AI/ML/HPC applications
- ACE task produces heterogeneously integrated compute platforms for the CUbiC wireless DSP workloads

CUbiC SoSFab







Intel FPGA + Arvon DSP co-packaged compute platform

Ghobadi

Krishna

Zhang

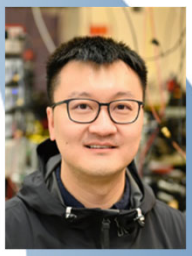
Belay

SRC

CUbiC

CUbiC Broadening Participation Initiatives


Michal Lipson


Alex Meng

Broadening Participation Pledge

At CUbiC, our goal is to enrich and uphold the involvement of a diverse student population in research and educational endeavors. The CUbiC Principal Investigators are committed to cultivating an inclusive environment for students, regardless of their race, gender, place of birth, or academic level.

- NSF REU Proposal - aiming to expand meaningful CUbiC research experiences to undergraduates. (submitted)
- Student Workshops - platforms for collaboration, diversifying participants, and promoting broader research understanding.
- CUbiC Travel Fellowship - recognizes academic excellence in minority students, offering opportunities for further education and professional growth.
- Joint Mentorship for Minority HS Students - Partnership with Science Honors Program, Summer Academic Program, and Young Women's Leadership Programs.

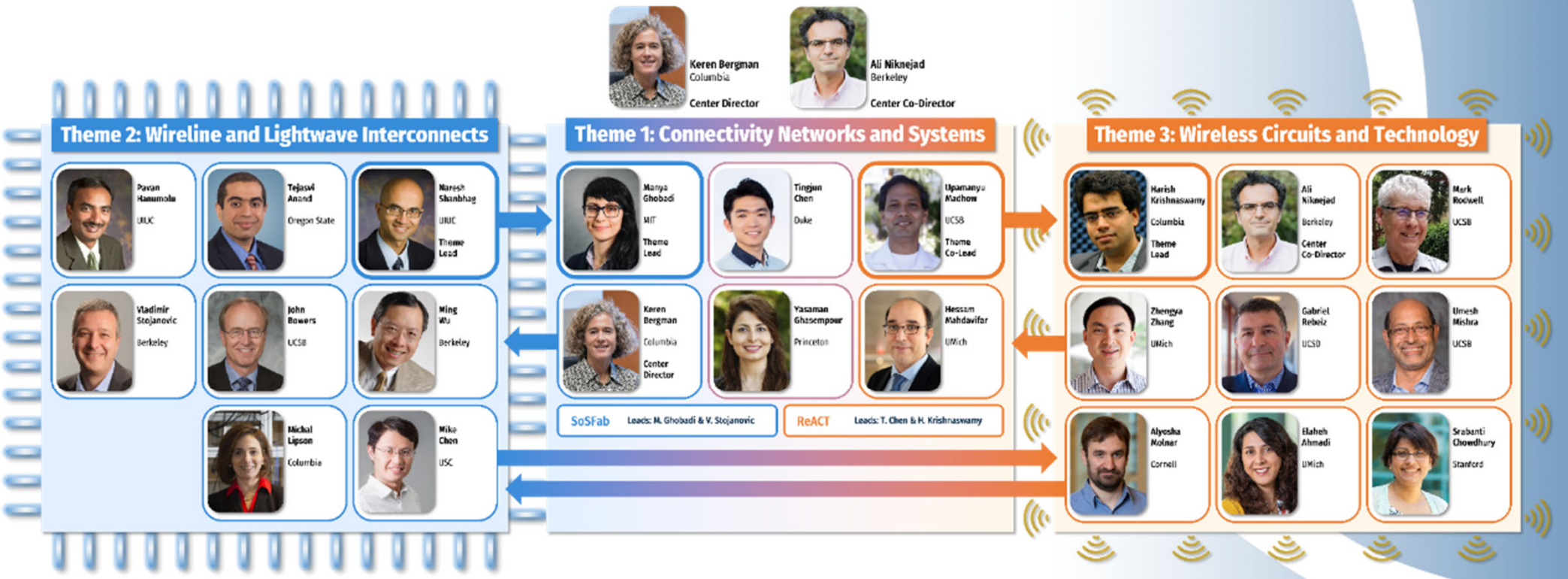
SRC

CUbiC

Enhanced Industry Partner Engagement

- CUbiC has three themes, each holds Liaison Meetings every 6 weeks
- Expanding the engagement with Liaisons and Industry Partners:
  - Monthly Student-Led Workshops: Collaborative sub-theme meetings that invite liaisons and viewers for deeper engagement. (eight sub-themes in total)
  - Student and Postdoc Leadership: CUbiC students and postdocs will lead (with PI participation) ensuring a student-driven focus.
  - Improve Collaboration: Monthly workshops enhance engagement with liaisons and collaborations among research groups on specific tasks.
- SRC Scholars Resume Book; Detailed Spreadsheet

CUbiC's Integrated Team



SRC

CUbiC

Summary

- CUbiC will strive to flatten the computation-communication gap, delivering seamless Edge-to-Cloud connectivity with transformational reductions in the global system energy consumption.

- Vertically integrated research agenda cross-cutting 3 technical themes
- Outstanding team of 23 PIs from 15 Universities
- 37 Research Tasks
- Currently >100 SRC graduate students scholars and 13 postdoctoral fellows


arm
IBM
ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™
SK hynix
Micron
BOEING
EMD ELECTRONICS
HRL LABORATORIES
tsmc
qorvo
NSF
intel
DARPA
GlobalFoundries™
Raytheon Technologies
SAMSUNG
MEDIATEK

SRC

CUbiC


CUbiC


ACADEMIA • INDUSTRY • GOVERNMENT