Semiconductor Research Corporation

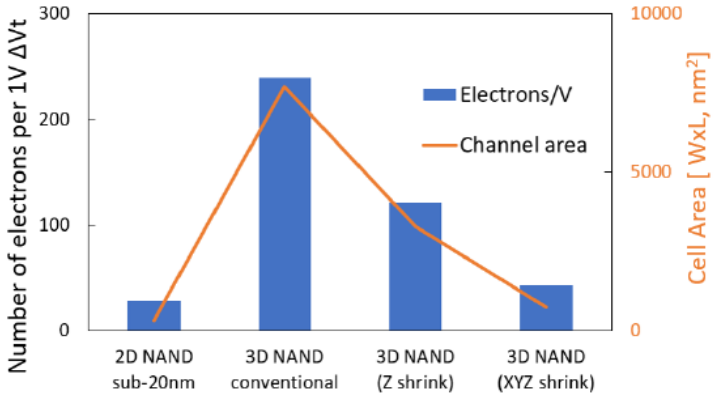
PRISM

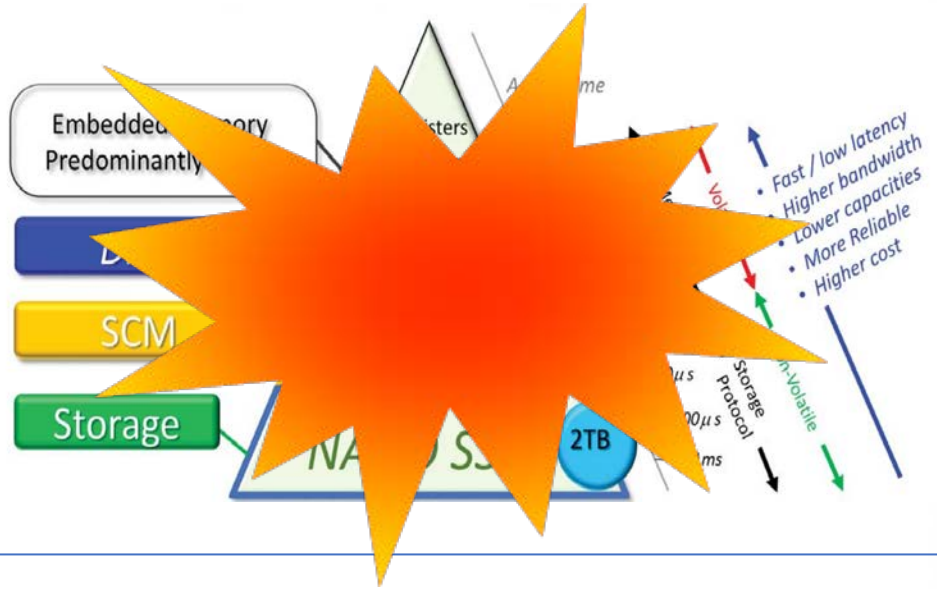
PRISM Highlights & Plan of Action

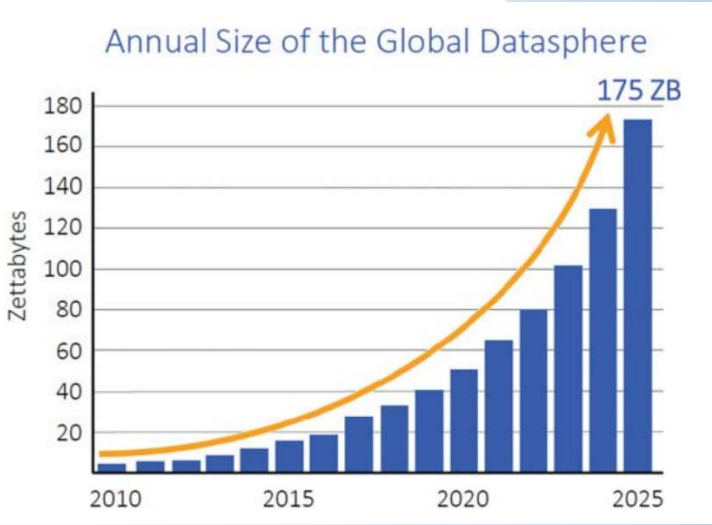
Center director: Tajana S. Rosing, UCSD
Center co-director: Nam Sung Kim, UIUC

PRISM Challenges

- Amount of data to be stored, moved & processed is rising exponentially
  - Global demand for memory/storage is growing rapidly, outpacing silicon production
  - Data movement is expensive → today most of the performance is lost due to data movement!
- Rising complexity makes programming and optimization harder
  - Heterogeneity of components and how they are integrated into systems
- Fundamental barriers to memory and storage technology scaling
  - Lower NAND string current, higher cell-to-cell interference, fewer electrons per stored state
  - Wordline disturbance, variable retention time, reduced sense margin







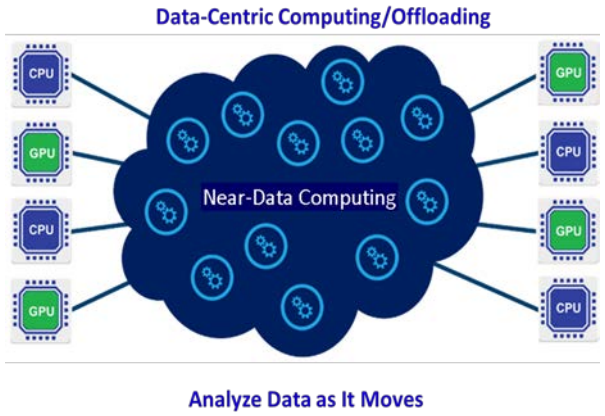
PRISM Goals

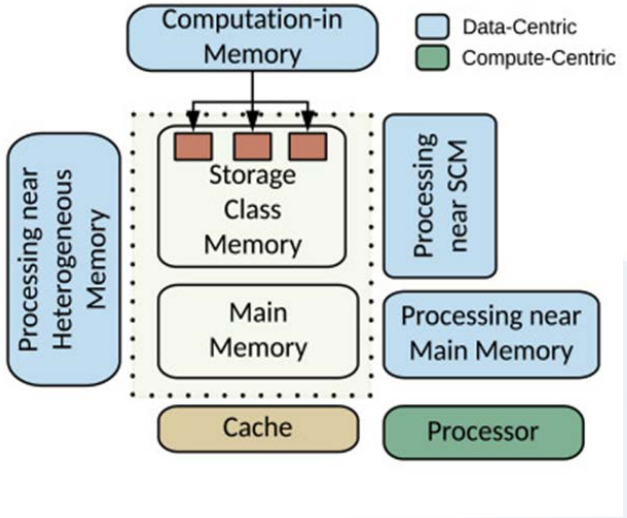
Solve fundamental IMS scale out and scale down challenges for 2030

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

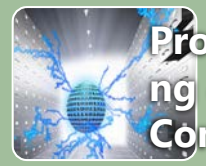
Theme 1: Systems & Software

Leads: Emmett Witchel & Ada Gavrilovska


Platform Abstractions


AIMS Controls


Scheduling & Placement

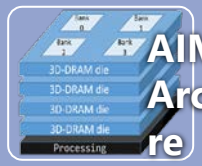

Programming & Compilers

Security & Privacy

Theme 2: Architecture

Leads: Nam Sung Kim & Jishen Zhao


Memory & Storage Architecture



AIMS Architecture


Controllers & Interfaces

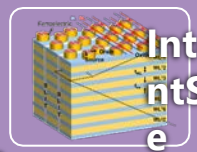

Disaggregated IMS Design

Hardware Security

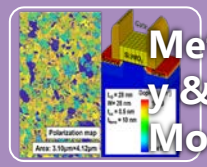
Theme 3: Devices & Circuits

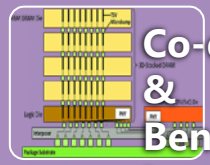
Leads: Suman Datta & Shimeng Yu


Intelligent Memory


Intelligent Storage


Emerging Memory


Metrology & Modeling


Co-design & Benchmarks

Cross-cuts

Performance
Programmability
Energy Efficiency
Security
Scalability
Virtualization
Composability
Reliability
Resilience
Availability

24 PIs, 13 universities, >120 graduate students

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

- Students working on PRISM demos and projects at UCSD to date:
  - **58 undergraduates**: 6 URM, 16 female students; **23 high school students**: 17 URM, 15 female students
- Presentations to the SRC representatives
  - PRISM annual review – best undergraduate poster award, with travel grants for undergraduate presenters
  - JUMP 2.0 Undergraduate Research Symposium
    - 26 presenters; three best undergraduate presentation and one graduate student/postdoc mentor award
  - **Goals:** help another JUMP 2.0 Center organize symposium, and bring undergraduates to Techcon
- Collaborate with mentorship programs at PRISM Universities
  - For example, undergraduates in UCSD ERSP, ENLACE, and McNair presented at Virtual Symposium
- High school outreach, with a focus on underserved communities
  - Ran 3 workshops at high schools in San Diego County
    - **Goal:** Continue in San Diego County; expand to local communities of other PRISM universities
  - PRISM High School Summer Research Program, 15 high school students for 2024 (12 URM, 11 female)
    - **Goal:** Continue in San Diego County; expand to high schools around other PRISM universities
    - **Long-Term Goal:** Create similar programs across all JUMP 2.0 Centers

PRISM Highlights for 2023

- **20 Awards**
  - **Vijay Narayanan** – 2023 American Association for the Advancement of Science Fellow
  - **Nam Sung Kim and Shimeng Yu** – Intel's 2023 Outstanding Researcher 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** - 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
  - **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
  - **Ada Gavrilovska** - "Simplifying Management of Complex Memory Fabrics" at Barcelona Supercomputing Center

86 Student Posters & 25 Demos at the

→ **Goal: longer time for industry-student interaction**

- **30 News articles published**
- **First JUMP 2.0 SRC patent application by PRISM PI Kevin Skadron, two more patents by PI Vijay Narayanan with SUPREME Center**
- **DARPA OPTIMA project initiated due to PRISM PIs work – PIs Suman Dutta, Shime**
- **Publications: 270 datasets in Pillar Science**

  → **Goal: increase uploads of papers, datasets and workloads; more proac**
- **PRISM Github:** https://github.com/PRISM-T4-Grand-Challenges

  → **Goal: provide more comprehensive datasets and workloads and create**
- **Students to sponsors:** 17 internships, 9 full time hires → **Goal: expand collaboration**
- **Industry liaisons & viewers:** 95 total
- **> 300 interactions with member companies by PRISM team → Goal: focus on deepening quality interaction with liaisons**

Members	Paper Collab	Interns	Tech Transfer	Full Time Hire
GlobalFoundries	✓		✓	
IBM	✓	✓	✓	✓
Intel	✓	✓	✓	✓
Micron	✓	✓	✓	✓
Samsung	✓	✓	✓	
SK Hynix	✓	✓	✓	
TSMC	✓	✓	✓	

PRISM's average KPI for the first year is 35.3 !!!

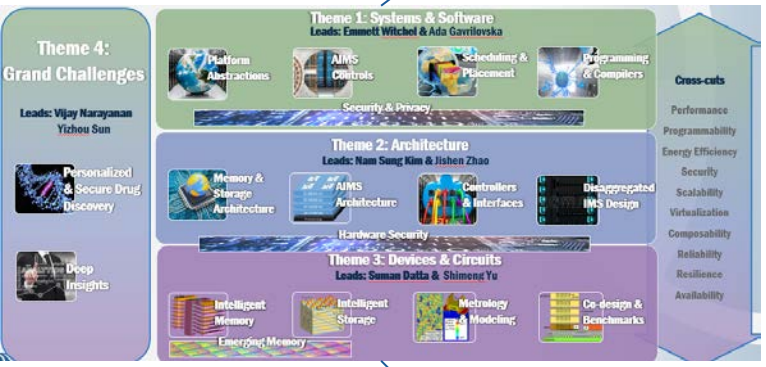
Collaboration with JUMP 2.0 centers via **SEED projects**

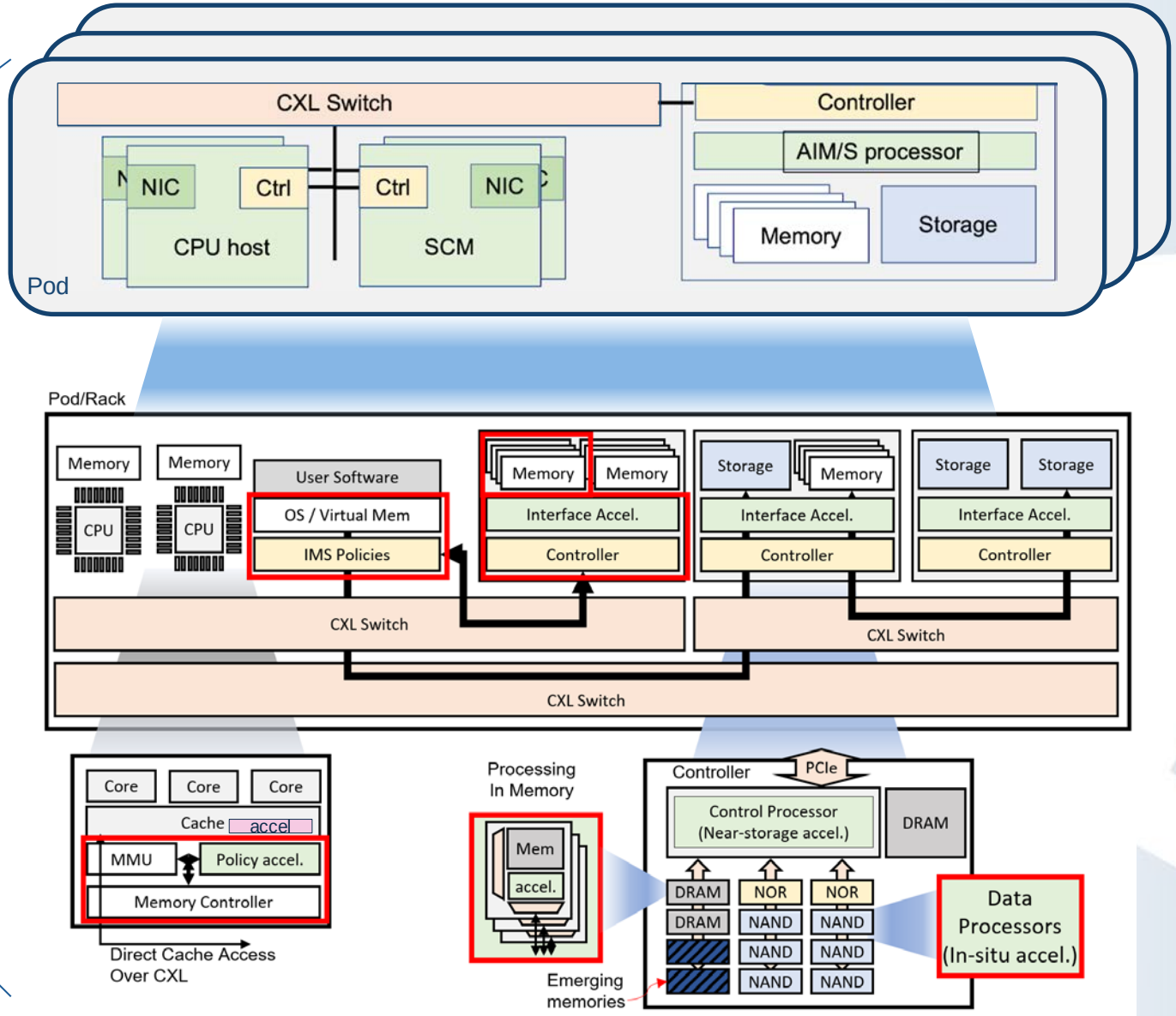
- **ACE**:
  - Jishen Zhao: Fixing Virtual Machine Memory Tiering Pathologies in the CXL Era
  - Ada Gavrilovska: End-to-end In-Fabric Programming for Graph Analytics
  - Baris Kasikci: Throughput-Optimized Datacenter Scale LLM
  - Yiying Zhang: SuperNICs
- **CHIMES**:
  - Nam Sung Kim: Development of SPICE Models for 2.5D Interconnects and Exploration of DRAM Design Space
  - Shimeng Yu, Nam Sung Kim, T. Rosing: 3D Stackable DRAM with CXL Interface for LLM Acceleration
  - K. Skadron, Shimeng Yu: Versatile BitSIMD AIMS Architecture For Processing In Embedded DRAM with Support for Bidirectional Data Access
  - Jason Cong, Shimeng Yu: Monolithic 3D FPGA Design and Synthesis with Back-End-of-Line Configuration Memories
  - Shimeng Yu: Security vulnerability of the eDRAM with oxide semiconductor transistors
- **CuBIC**:
  - Nam Sung Kim, Jishen Zhao & Karen Bergman: CXL and optics

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

- **CoCoSys**:
  - Vikram Adve, Eric Pop, T. Rosing: SW/HW optimization of Applications for MLC PCM
  - P. McDaniel, T. Rosing: HD computing for intrusion detection in hosts
  - Priya Panda & T. Rosing: Acceleration for Privacy-aware Federated Learning → **Goal: defense focus applications (SWaP-C)**
- **CogniSense**:
  - Hun Seok Kim, T. Rosing: Intelligent and Efficient On-Sensor Computing for LiDAR Segmentation
    → **Goal: more defense focus applications (SWaP-C)**
- **SUPREME → Goal: more collaboration with the SUPREME center**
  - Vijay Narayanan: Multi-Level Cell Sensing Inspired Robust Charge Domain In-Memory Computing with FeFET
  - Suman Datta, Asif Khan: Engineered FE Stack for 3D NAND
  - Suman Datta, Chris Hinkle: Phase transition for cross-point selectors
  - Shimeng Yu, Asif Khan: ML assisted modeling of FeFET variability and reliability
  - Asif Khan & Tajana Rosing: Computing near 3D MLC FeNAND Nonvolatile Memory for Database Search

PRISM Architecture

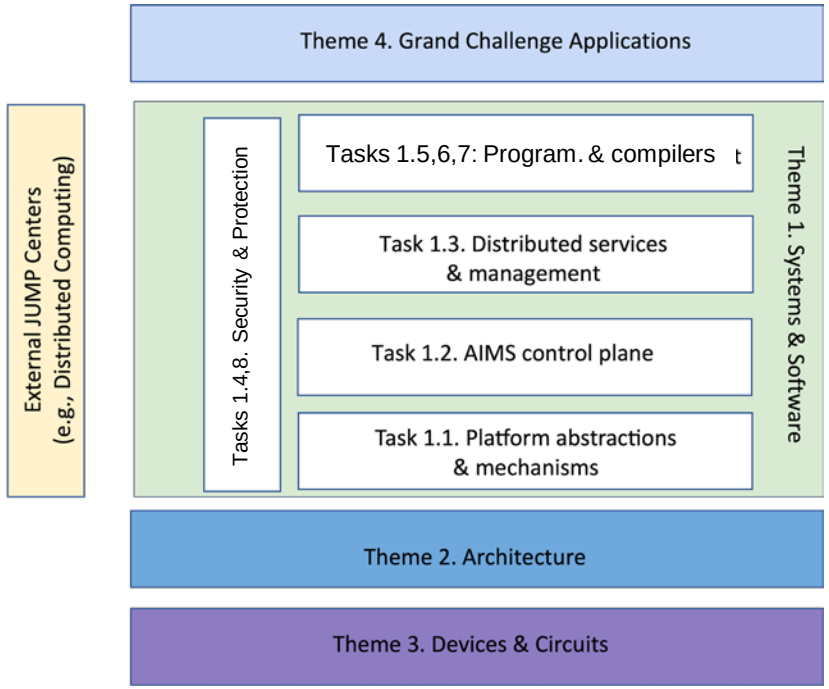


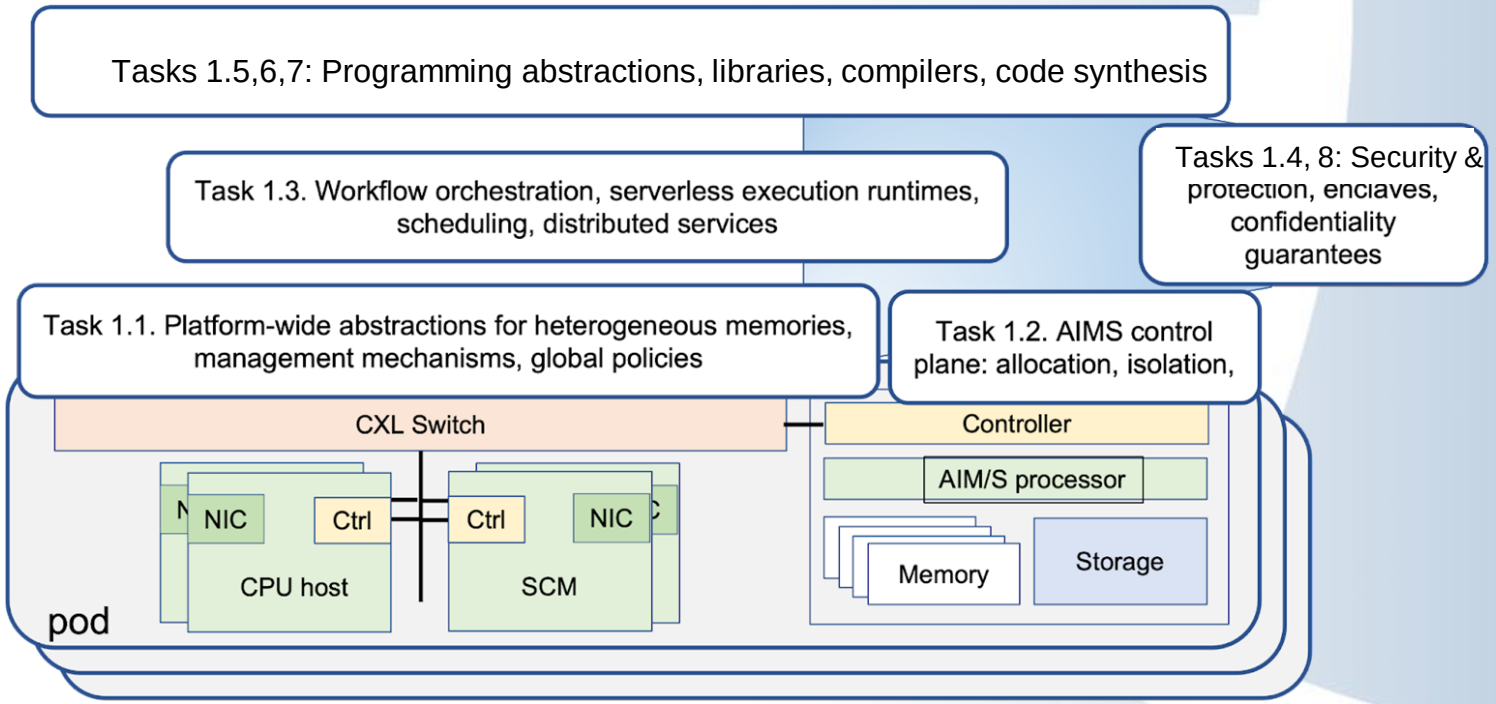


Theme 1: Systems and SW

Leads: Emmett & Ada

Goal: Seamless deployment of grand challenge applications in virtualized & distributed IMS systems with 100x improvement

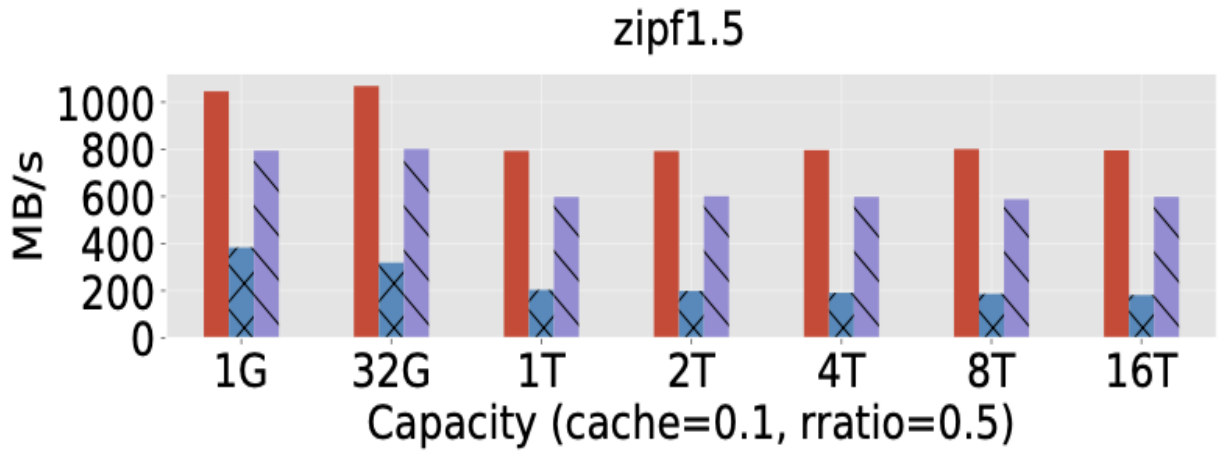




Theme 1: Efficient Integrity Protection for Trusted Cloud Storage

Patrick McDaniel (PI), Michael Swift (PI), Rachel King (Grad), Quinn Burke (Grad)

- **Problem:**
  - Integrity mechanisms for secure cloud storage (i.e., hash trees) and SoTA solutions are far from practical (>5X overhead)
- **Technical Approach:**
  - Our approach learns *workload* patterns and dynamically adjusts the hash tree on-the-fly to reduce costs
- **Key results and metrics vs. SoTA:**
  - We implement a custom block device on real Linux systems and run benchmarks against it with real-world datasets
  - Most overheads stem from writes, which can be optimized with log-structuring
  - Measured up to 3X speedups over the SoTA
- **Grand Challenge Application:** Deep insights & precision medicine: scalable and secure information retrieval



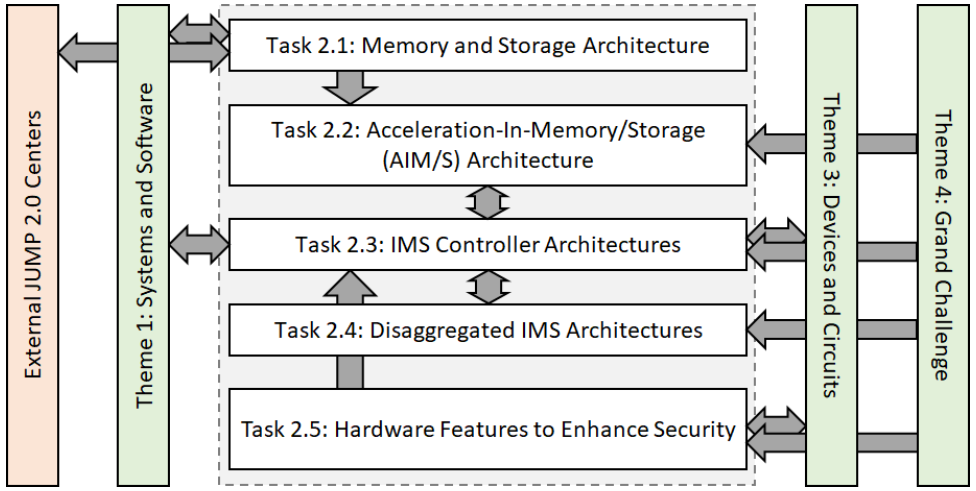
As storage capacity increases, **our approach (purple)** outperforms the **SoTA (blue)** and performs close to **optimal (red).**

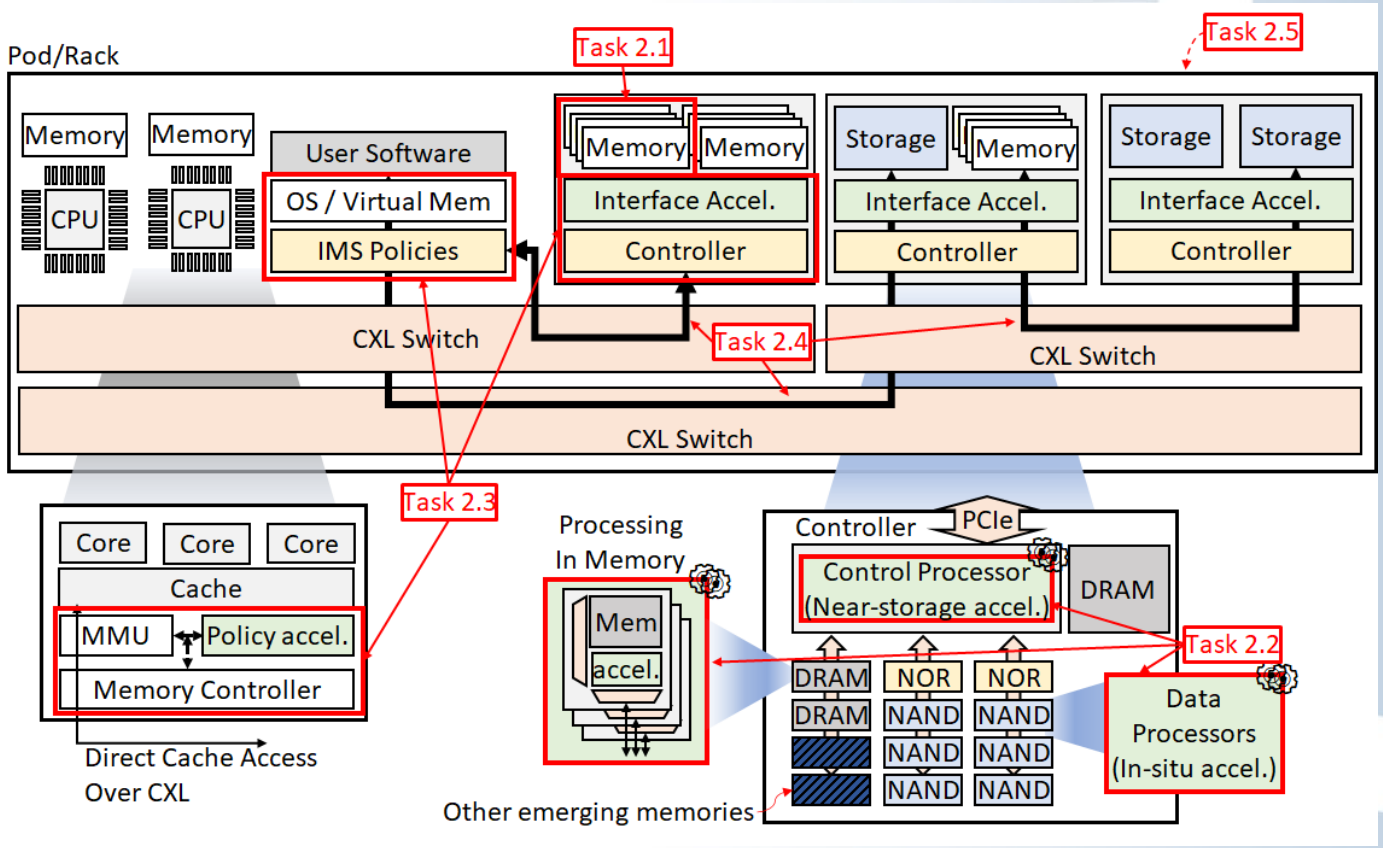
Quinn Burke, Ryan Sheatsley, Rachel King, Michael M. Swift and Patrick McDaniel, “Efficient Integrity Protection for Trusted Cloud Storage.” (in submission)

Theme 2: Architecture

Leads: Nam Sung & Jishen

Goal: Memory/storage architecture enabling 100x more powerful IMS computing capability at 10x larger capacity

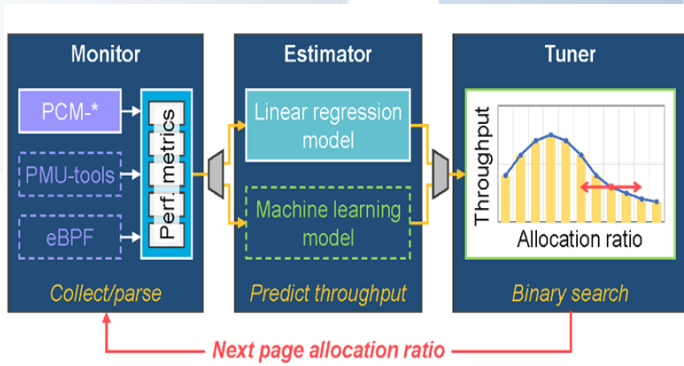


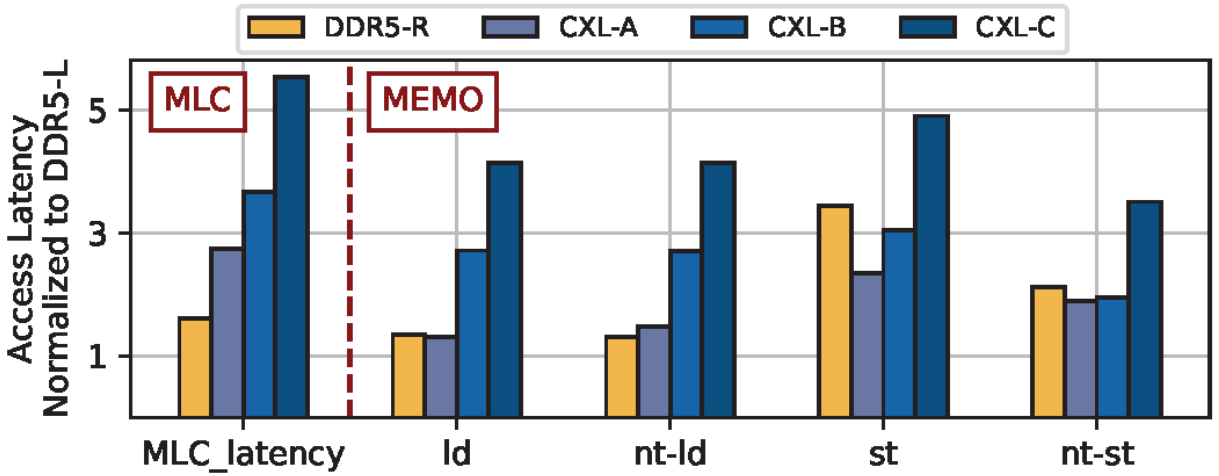


Themes 1 & 2: CXL system characterization & optimization

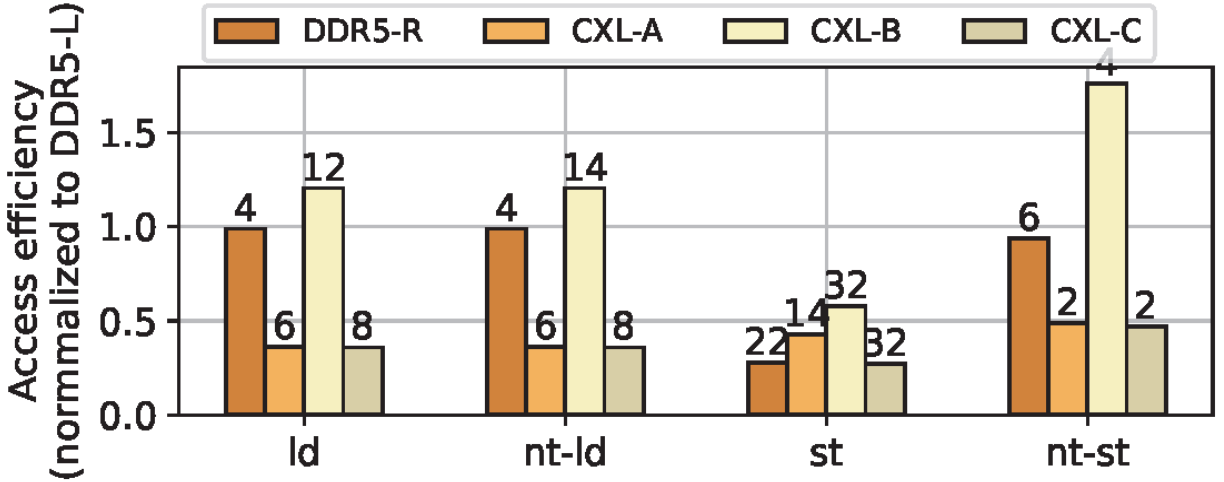
PIs NS Kim, M. Swift, B. Kasicki; Jointly with Y. Yuan & R. Wang Intel; and Samsung

- New insights in the performance impact of using true CXL-ready systems based on the latest 4th-generation Intel Xeon CPU with three CXL memory devices from different manufacturers
- PI Swift's team's new instrumentation of kernel locks provided hints for the level of data placement and sharing for this new data tiered system
- PI Kasicki is addressing microarchitectural inefficiencies from threads with small instruction counts between context switches by profile-guided fingerprinting of thread data accesses → guides scheduling and data placement
- Automatically determine % of pages to CXL.mem integrated w/ Linux
  - 8%--20% higher performance vs best DDR:CXL=50:50 & 100:0, allocating 29%--41% of pages to CXL.mem
  - Open-sourced and working with Samsung for tech transfer





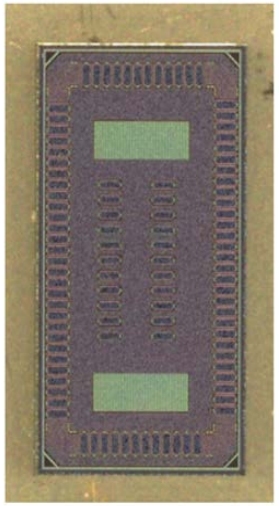
Random memory access latency



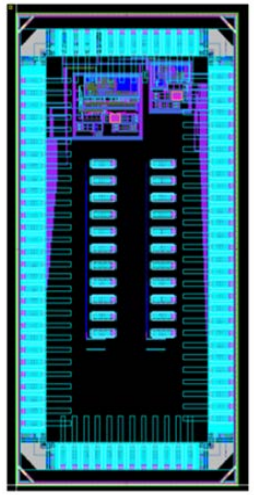
Efficiency of maximum sequential memory access bandwidth

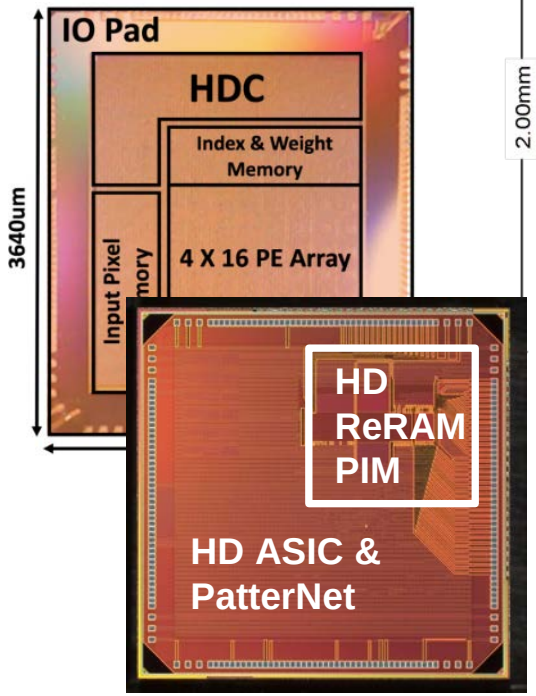
PRISM Chips Gallery

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

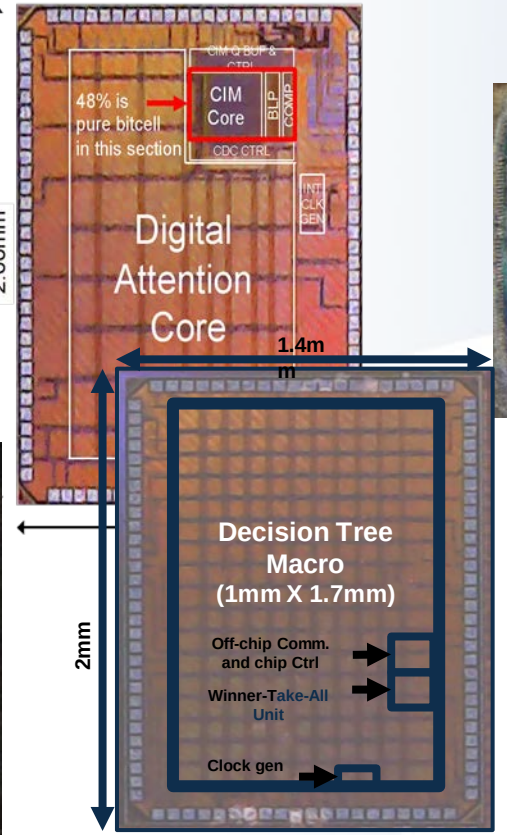


PI Shimeng Yu
1st gen FeFET PIM macro in GF 28SLPe shuttle via Fraunhofer IIS MPW shuttle w/ FeFET risk manufacturing

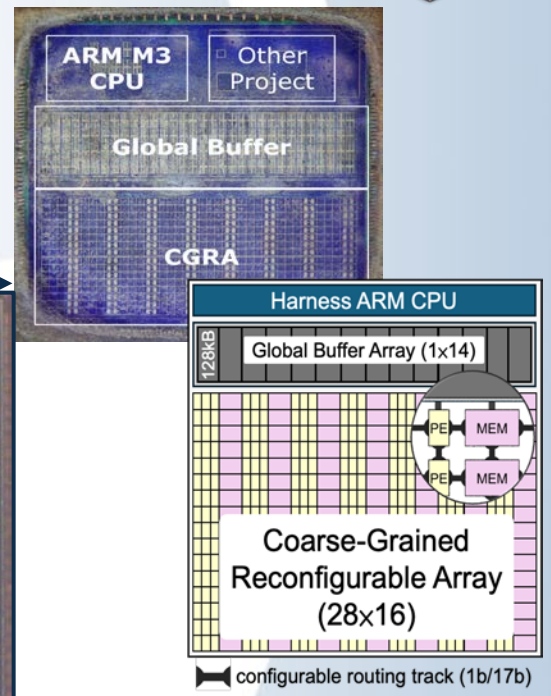




PIs Tajana Rosing & Mingu Kang
40nm ASIC & ReRAM HDnn - Few shot learning for ImageNet size images
Collaboration with TSMC



PIs Mingu Kang & Tajana Rosing
XGboost tree ensemble classifier [CICC'24] & SRAM-based attention core

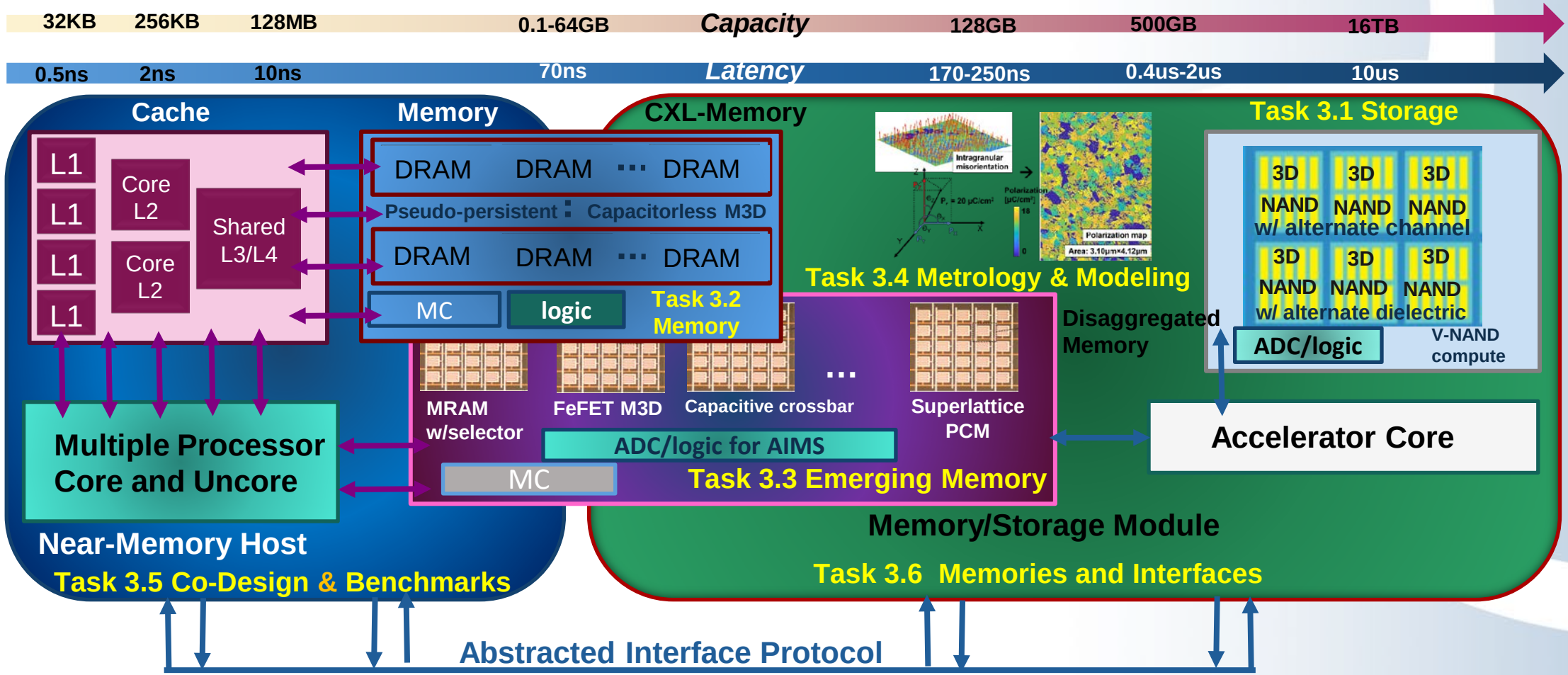


PI Priyanka Raina
Two AIMS data processors for programmably accelerating both dense and sparse applications

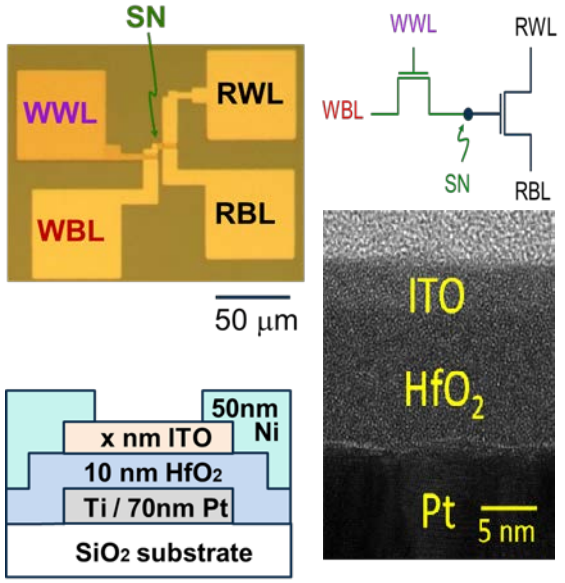
Theme 3: Devices & Circuits

Leads: Suman & Shimeng

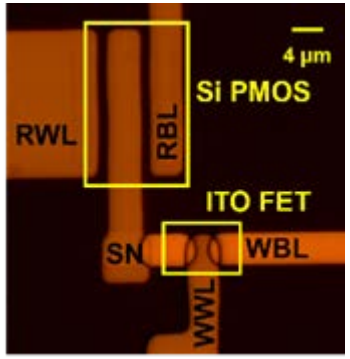
Goal: Fundamental advances in devices and their controls leading to 100x improvement in PPAC



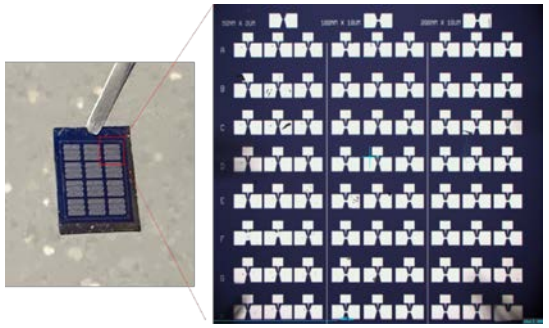
PRISM Devices Gallery



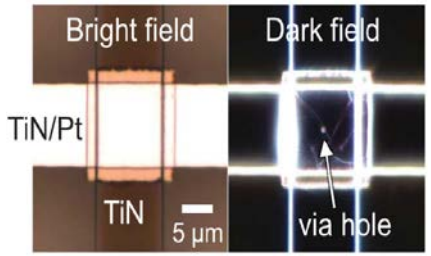
PI Philip Wong
Oxide Semiconductor Gain Cell fabricated at Stanford



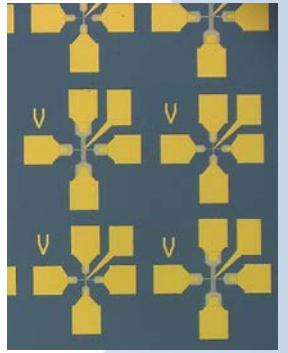
PI Philip Wong
Si PMOS / ITO-FET hybrid gain-cell memory [VLSI 2024]



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

Goals:

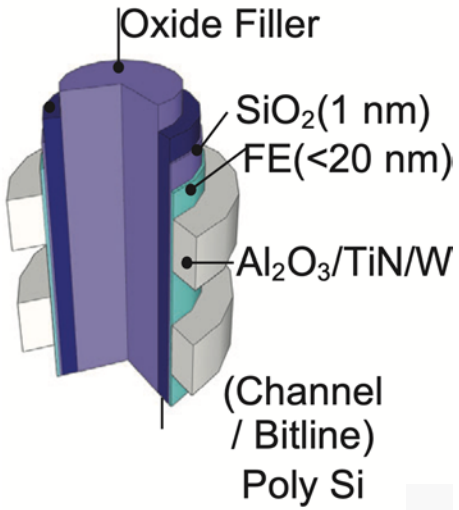
→ **Tighter collaboration between Theme 3: Devices & Circuits and Themes 1 & 2**

→ **Closer collaboration with the SUPREME Center**

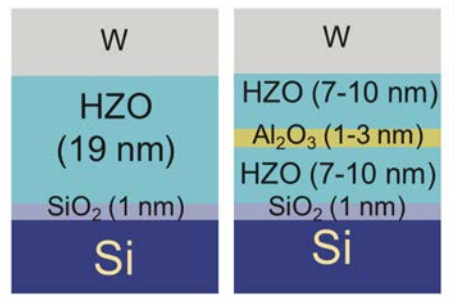
Theme 3: Disturb-Free 3D Ferroelectric NAND with Enhanced Memory Window

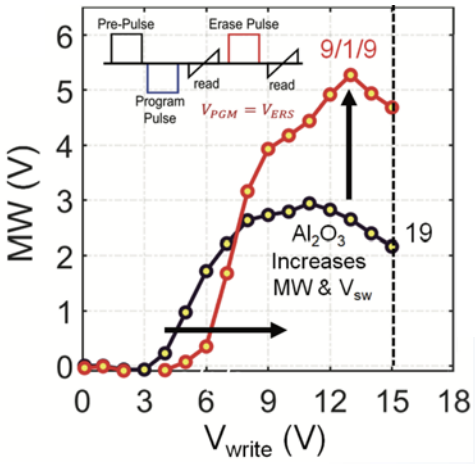
PI Suman Datta, collaboration with **SUPREME** PI Asif Khan

- **Problem**:
  - 3D NAND has high write voltage, cell-to-cell interference that affects z-scaling
- **Technical Approach:**
  - Oxide-semiconductor or poly-silicon channel with asymmetric dual-gate, one with HZO (FE), one with HfO2 (DE) (w/Asif Khan of SUPREME)
- **Results**:
  - Massive enhancement of memory window with dielectric layer insertion
  - Improved read disturb translated to higher accuracy for system-level applications as IMS for MS

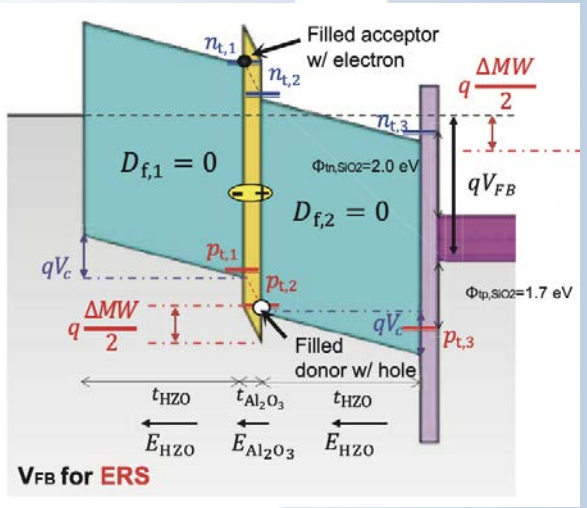


- TLC and QLC compatible operation.
- Low write voltage (<15V).
- Enables z-scaling.





Das et al. IEDM 2023

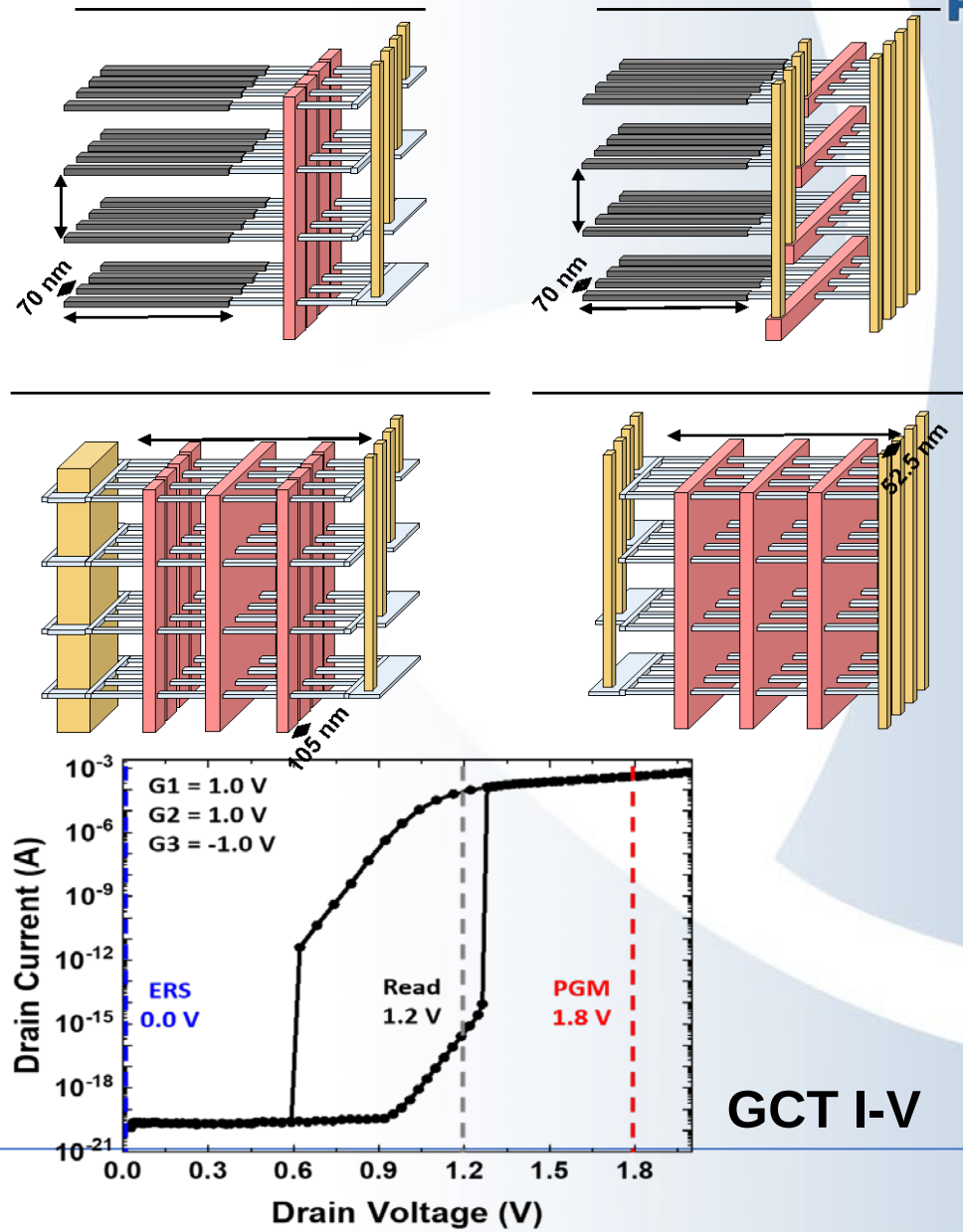


VFB for ERS

Understanding the role of tunnel dielectric and gate blocking layers and the interactions with poly-Si channel in ferroelectric NAND structures.

Themes 2,3,4: 3D DRAM acceleration of LLMs in CXL-based systems
PI Shimeng Yu (Theme 3), PI Nam Sung Kim & Tajana Rosing (Theme 2)

- Problem: 2D DRAM faces scaling challenges
  To address the scaling limit in 2D DRAM
  - ➢ To enable bit-cost effective scaling in the vertical direction
  - ➢ To improve the data retention by capacitorless design
  - ➢ To identify the new challenges in 3D DRAM design
- Technical approach:
  - Compare with industry concurrent effort in horizontal 1T1C vertical stacking
  - TCAD simulation of the capacitorless 3D DRAM with gate-controlled thyristor effect.
- Key results[1]:
  - Optimized GCT device shows the following metrics
    - 128 layer array provides 4x density improvement vs. SoTA 2D DRAM 1c node with similar latency & BW
- Collaboration with Theme 2 and Theme 3 for 3D DRAM's application in CXL enabled LLM acceleration (deep insights grand challenge)



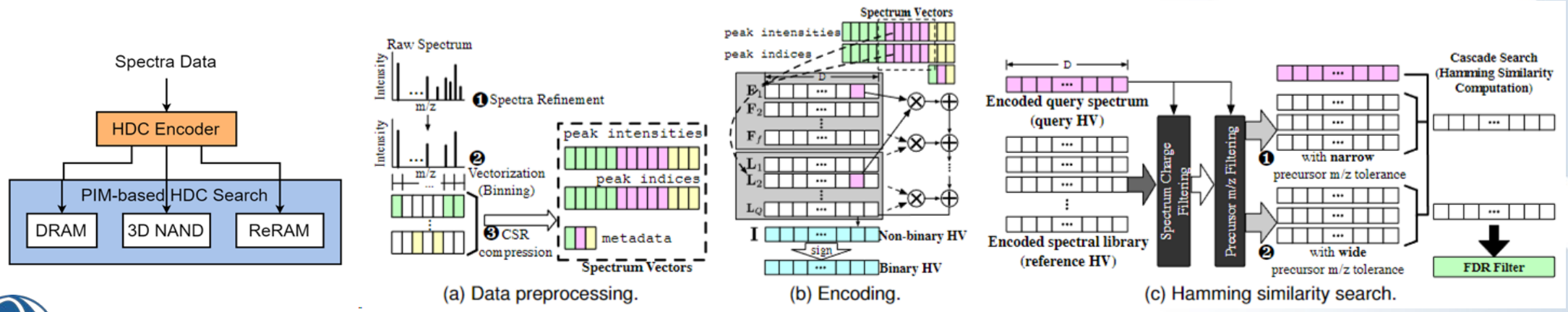
GCT I-V

Themes 1-4: Accelerating Search in Memory & Storage

Theme 1: Vikram Adve, Sang-Woo Jun, **Theme 2**: Jason Cong, Tajana Rosing,
Theme 3: Mingu Kang, Shimeng Yu, Philip Wong, Eric Pop; **SUPREME**: Asif Khan

- **Motivation:**
  - Protec
  - CPU/G
  - Novel
    autom
- **Technical**
  - Lever
    benef
  - Them
  - Use H
- **Grand cha**
  - **Datasets**: iPRG2012, HEK293 & human data from UCSD's MASSiVE database with comparable accuracy to SOTA AnnSoLo[JPR'23]

Architecture	SOTA-CPU	SOTA-GPU	GPU (HDC)	FPGA (HDC)	DRAM (HDC)	MLC ReRAM (HDC)	3D NAND (HDC)
Algorithm	AnnSoLo[JPR '23]	AnnSoLo	HOMS-TC	RapidOMS	HyperOMS-PIM	HyperOMS-PIM	HyperOMS-NAND
Tech Node (nm)	i7-8700K (14nm)	RTX 4090 (5nm)	RTX 4090 (5nm)	Versal Prime VM1802 (7nm)	22nm DRAM 28nm compute	130nm with 3M RRAM cell	ASIC 7/65nm 3D NAND 13nm
Speed	Baseline (1x)	3.09x	108x (35x)	56x	262x (84.76x)	76.7x	210x (24x)
Energy Efficiency	Baseline (1x)	1.41x	5.44x (3.85x)	71x	1,620x (2284x)	3,000x	5,400x



1. Kang J, Xu W, Bittremieux W, Moshiri N, Rosing T. Accelerating open modification spectral library searching on tensor core in high-dimensional space. Bioinformatics. 20231;39(7):btad404.
2. Kang J, Xu W, Bittremieux W, Moshiri N, Rosing T. DRAM-Based Acceleration of Open Modification Search in Hyperdimensional Space. IEEE TCAD. 2024.
3. Hsu PK, Xu W, Rosing T, Yu S. An In-Storage Processing Architecture with 3D NAND Heterogeneous Integration for Spectra Open Modification Search. MemSys 2023 (pp. 1-7).
4. Fan K,, Wong P., Rosing T. Efficient Open Modification Spectral Library Searching in High-Dimensional Space with Multi-Level-Cell Memory, DAC 2024.

PRISM Center Year 2 Plans

- Center-wide CXL-enabled PRISM systems fully operational & characterized running SOTA workloads
  - Closer collaboration and bi-directional technology transfer with member companies
- Demonstrate benefits of PRISM systems using grand challenge applications across all themes
  - Theme 3: design and compare various types of memory and storage devices
  - Theme 2: acceleration in and near memory and storage, CXL integration
  - Theme 1: systems, runtimes, and compilers
  - Use state of the art baselines for comparisons with clear metrics, project their evolution into the future
  - Goal: determine how **intelligence should and should NOT be added to memory and storage**
- Cross-center integration of security: systems, accelerators, CXL etc.
- Defense-oriented AI at the tactical edge where Size, Weight, and Power & Cost (SWaP-C) really matter
- Closer lab to fab collaboration with sponsors so new materials & processes are compatible with foundry
- Expand collaboration with other JUMP 2.0 centers; e.g. SUPREME
- Continued growth of our Broadening Participation program

Systems & Software

Emmett Witchel

Professor of Computer Science, University of Texas, Austin
witchel at cs dot utexas.edu

Ada Gavrilovska

Associate Professor, School of Computer Science, Georgia Tech ada@cc.gatech.edu | 404.894.0387

Yiying Zhang

Associate Professor, Computer Science and Engineering Department, University of California, San Diego
858.246.5216 | yiying@ucsd.edu

Yizhou Sun (孙怡舟)

Associate Professor, Department of Computer Science, University of California, Los Angeles
yzsun at cs dot ucla dot edu

Fredrik Kjolstad

Assistant Professor, Department of Computer Science, Stanford University
kjolstad@cs.stanford.edu

Franz Franchetti

Professor of Electrical & Computer Engineering, Carnegie Mellon University
412.268.8297 | franzf at ece.cmu.edu

Baris Kasikci

University of Washington's Paul G. Allen School of Computer Science and Engineering
baris@cs.washington.edu

Vikram S. Adve

Donald B. Gillies Professor, Computer Science Department, University of Illinois at Urbana-Champaign vadve@illinois.edu

Patrick McDaniel

Tsun-Ming Shih Professor of Computer Sciences in the School of Computer, Data & Information Sciences at the University of Wisconsin-Madison mcdaniel@cs.wisc.edu | (608) 263-1008

Next-Generation Architecture

Tajana Simunic Rosing

ACM & IEEE Fellow, Full Professor & Fratamico Endowed Chair, Dept of CSE and ECE, UCSD
tajana@eng.ucsd.edu

Nam Sung Kim

W.J. 'Jerry' Sanders III – Advanced Micro Devices, Inc. Endowed Chair in Electrical and Computer Engineering, University of Illinois
217.244.9169 | nskim@illinois.edu

Jishen Zhao

Associate Professor Computer Science and Engineering Department, Jacobs School of Engineering, University of California, San Diego
858.822.2449 | jzhao@ucsd.edu
(Photo credit: Darrell Long)

Kevin Skadron

Harry Douglas Forsyth Professor of Computer Science, Department of Computer Science School of Engineering and Applied Science, University of Virginia
434.982.2042 | skadron (ampersand) virginia.edu

Michael Swift

Professor Computer Sciences Department, College of Letters and Sciences, University of Wisconsin, Madison
608.890.0131 | swift at cs dot wisc dot edu

Jason Cong

Volgenau Chair for Engineering Excellence, Director, Center for Customizable Domain-Specific Computing Director, VLSI Architecture, Synthesis, and Technology (VAST) Laboratory (former VLSI CAD Laboratory)
310.206.2775 | cong@cs.ucla.edu

Sang-Woo Jun

Assistant Professor, Computer Science Department, Donald Bren School of Information and Computer Sciences, University of California, Irvine
swjun _AT_ ics.uci.edu

Vijaykrishnan Narayanan

A. Robert Noll Chair Professor, Computer Science & Engineering and Electrical Engineering, Pennsylvania State University vijay_at_cse.psu.edu

Priyanka Raina

Assistant Professor, Electrical Engineering, Stanford University
praina AT stanford DOT edu

Devices & Circuits

Shimeng Yu

Professor, School of Electrical and Computer Engineering, Georgia Institute of Technology
shimeng.yu@ece.gatech.edu

Suman Datta

Professor; Georgia Research Alliance Eminent Scholar; Joseph M. Pettit Chair in Advanced Computing ; Joint appointment with the School of Materials Science and Engineering (MSE)
sdatta68@gatech.edu

Sayeef Salahuddin

TSMC Distinguished Professor of Electrical Engineering and Computer Sciences, University of California Berkeley
510.642.4662 | sayeef at eecs.berkeley.edu

H.-S. Philip Wong

Professor of Electrical Engineering, Willard R. and Inez Kerr Bell Professor, School of Engineering, Stanford University
650.725.0982 | hspwong AT stanford DOT edu

Eric Pop

Professor of Electrical Engineering (EE) and Materials Science & Engineering, Stanford University
epop at stanford dot edu

Mingu Kang

Assistant Professor, Electrical and Computer Engineering, University of California San Diego
m7kang at ucsd dot edu

Center Directors

Tajana Simunic Rosing

ACM & IEEE Fellow, Distinguished Professor & Fratamico Endowed Chair, Dept of CSE and ECE, UCSD
tajana@eng.ucsd.edu

Nam Sung Kim

W.J. 'Jerry' Sanders III – Advanced Micro Devices, Inc. Endowed Chair in Electrical and Computer Engineering, University of Illinois
217.244.9169 | nskim@illinois.edu

Broadening Participation Champion

Niema Moshiri

Associate Teaching Professor, Computer Science & Engineering Department at the UC San Diego. a1moshiri at ucsd dot edu

Team PRISM

SRC

24 PIs, 13 Universities, 1 BPC