Next-Generation Domain-Specific Accelerators: From Hardware to System

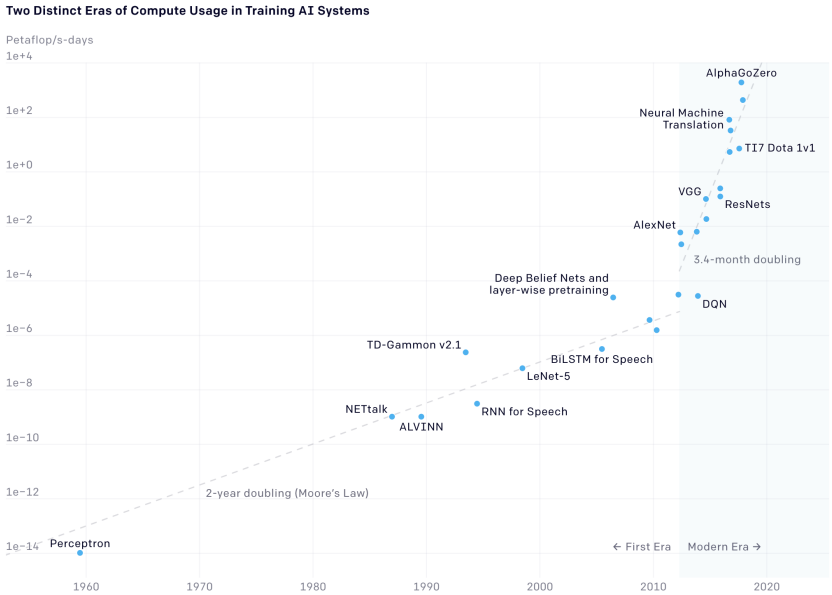
Sophia Shao
ysshao@berkeley.edu
Electrical Engineering and Computer Sciences
Berkeley
UNIVERSITY OF CALIFORNIA

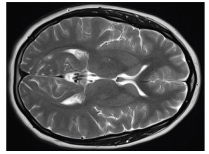
Growing Demand in Computing

Engadget

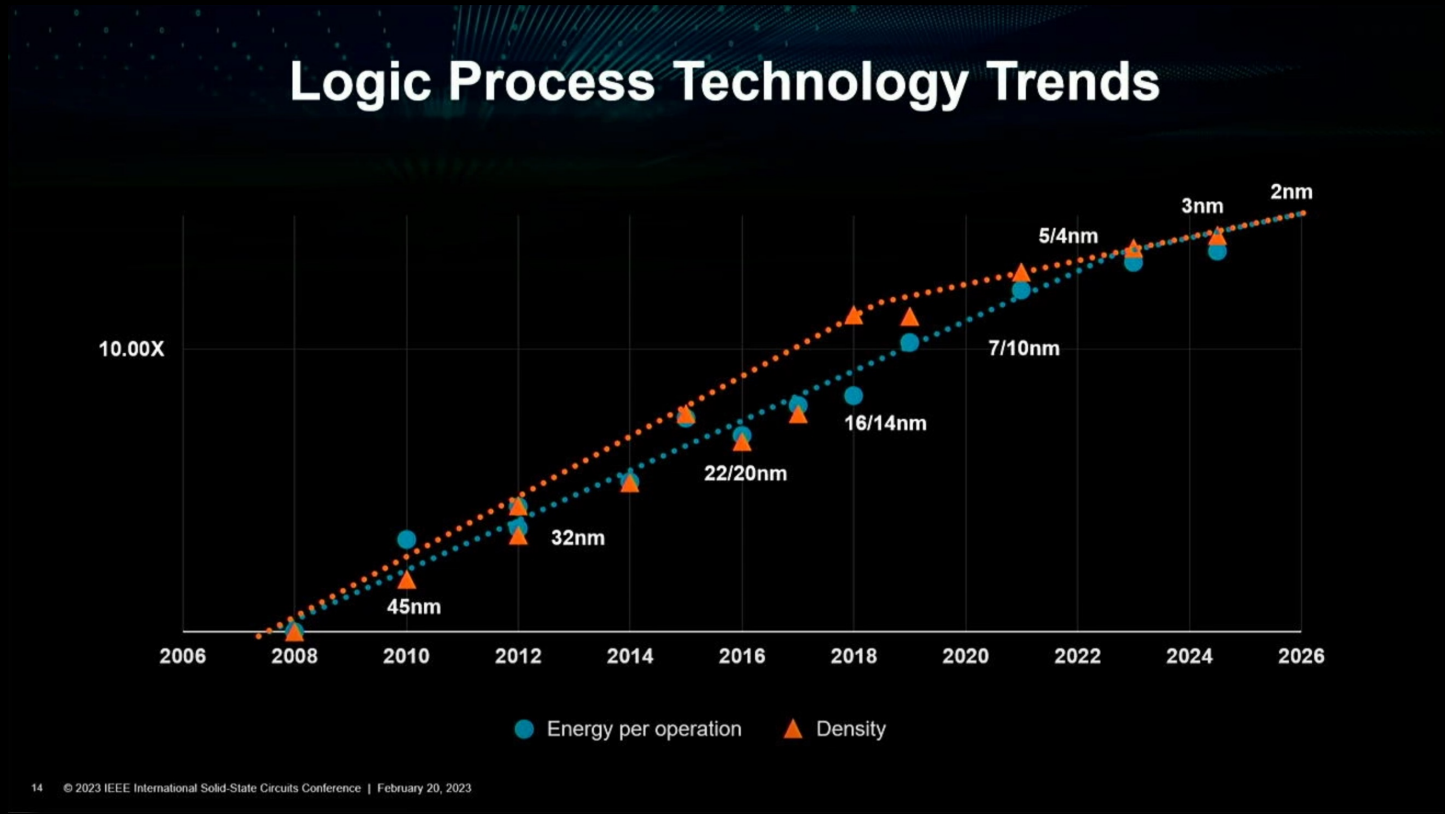
ChatGPT reportedly reached 100 million users in January

ChatGPT reportedly reached 100 million users in January ... According to a study by analytics firm UBS, it averaged 13 million unique visitors a...





OpenAI



Slowing Supply in Computing

AMD, ISSCC, 2023

“It was the best of times,
it was the worst of times.”

- *Dickens, A Tale of Two Cities, 1859*

Growing Demand in Computing

Slowing Supply in Computing

Growing Demand in Computing

Slowing Supply in Computing

Domain-Specific Accelerators

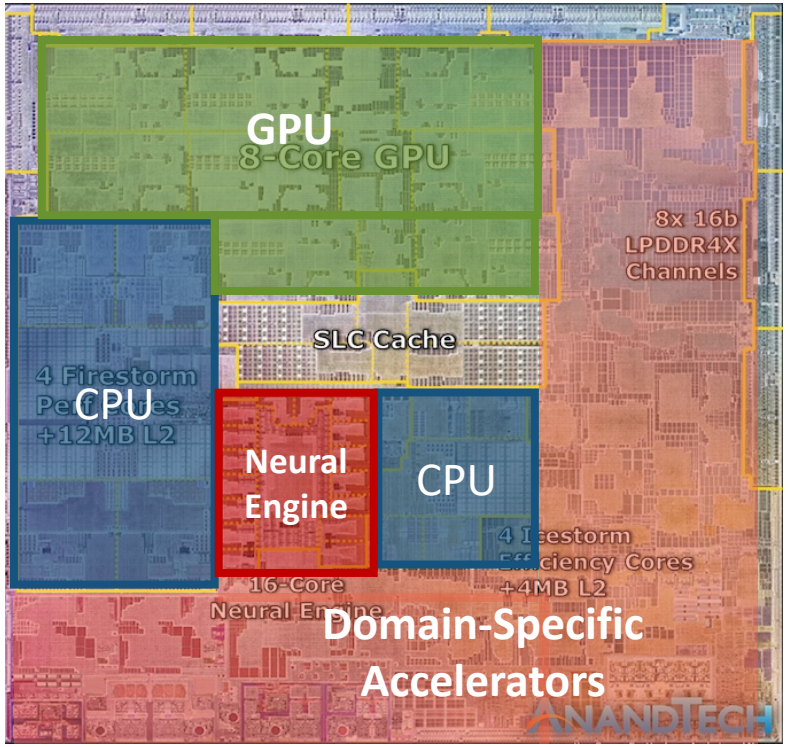
Growing Demand in Computing

Slowing Supply in Computing

Domain-Specific Accelerators

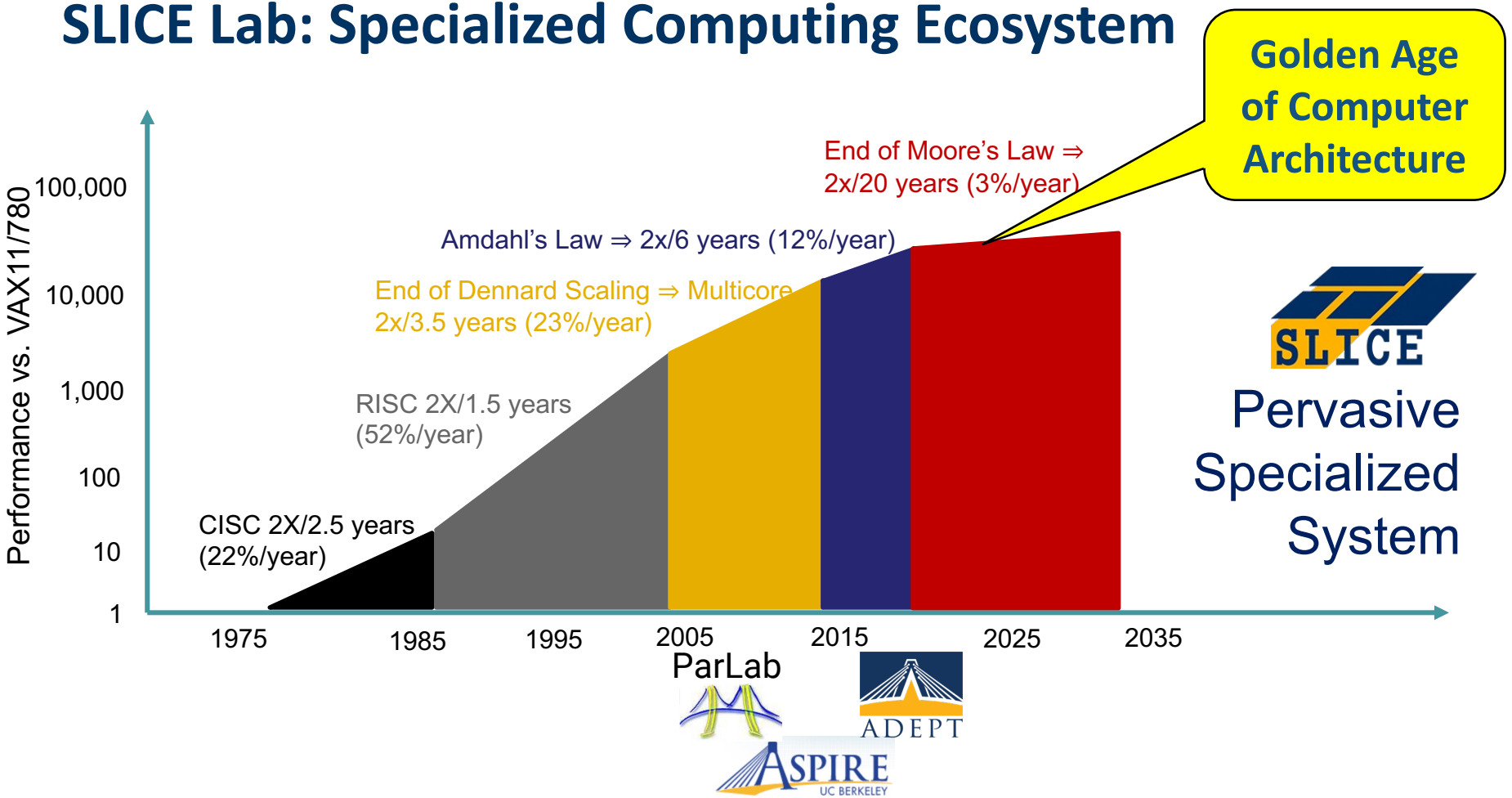
- Customized hardware designed for a domain of applications.

Apple M1 Chip
2020



* AnandTech

SLICE Lab: Specialized Computing Ecosystem



Full-Stack Optimization for Domain-Specific Systems

Design of Accelerators

- Simba [MICRO'19 **Best Paper Award**, CACM RH, VLSI'20, JSSC'20 **Best Paper Award**]

Integration of Accelerators

- Chipyard [IEEE Micro'20]
- Gemmini [DAC'21, **Best Paper Award**]

Scheduling of Accelerators

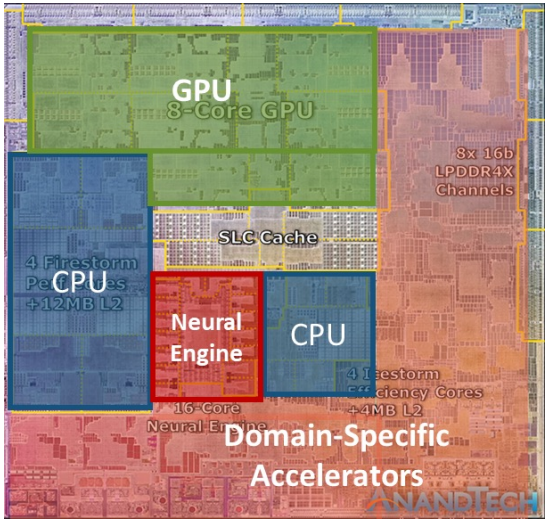
- CoSA [ISCA'21]
- MoCA [HPCA'23]

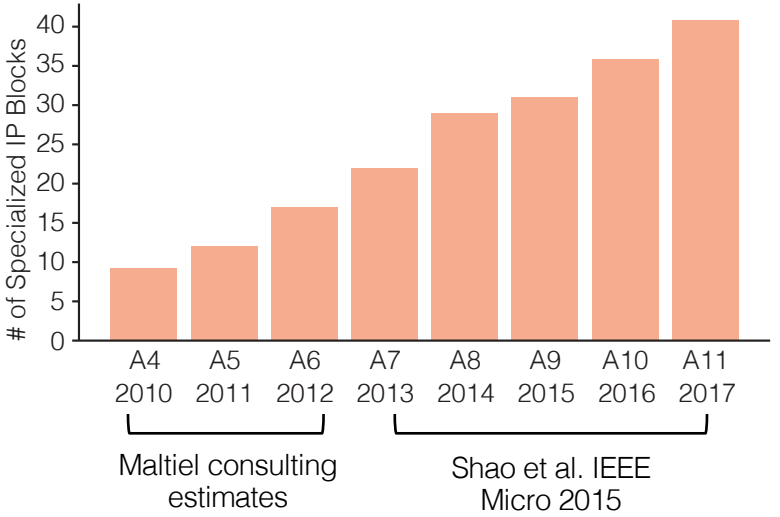
Closed-Loop Design Flow

RoSÉ [ISCA'2023]

Full-Stack Optimization for Domain-Specific Systems

Accelerators don’t exist in isolation.

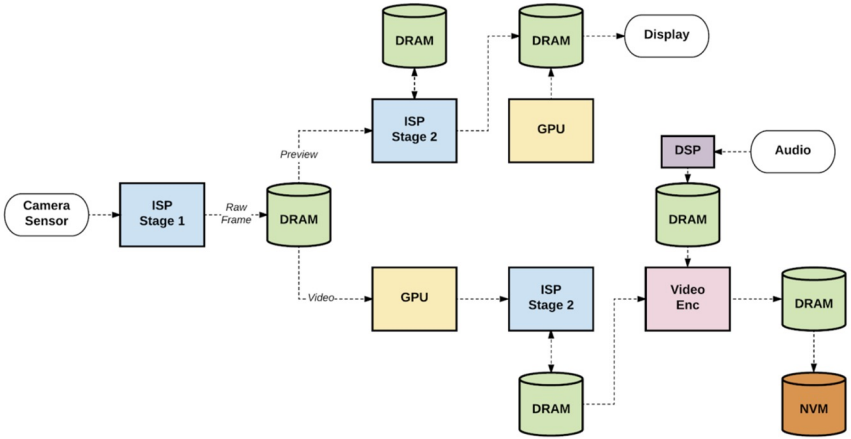




http://vlsiarch.eecs.harvard.edu/research/accelerators/die-photo-analysis/

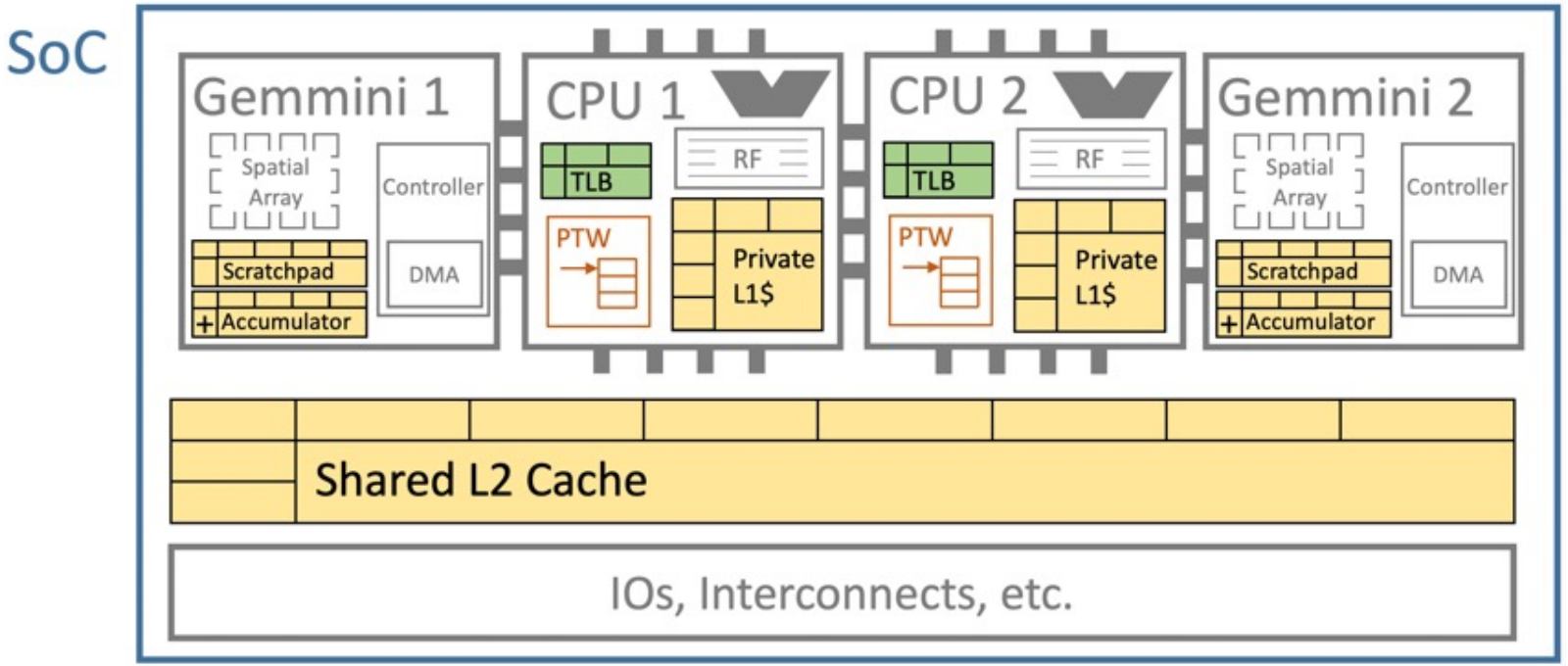
Mobile SoC Usecase

- Mainstream architecture has long focused on general-purpose CPUs and GPUs.
- In an SoC, multiple IP blocks are active at the same time and communicate frequently with each other.
- Example:
  - Recording a 4K video
  - Camera -> ISP
    - “Preview stream” for display
    - “Video stream” for storage
  - DRAM for data sharing

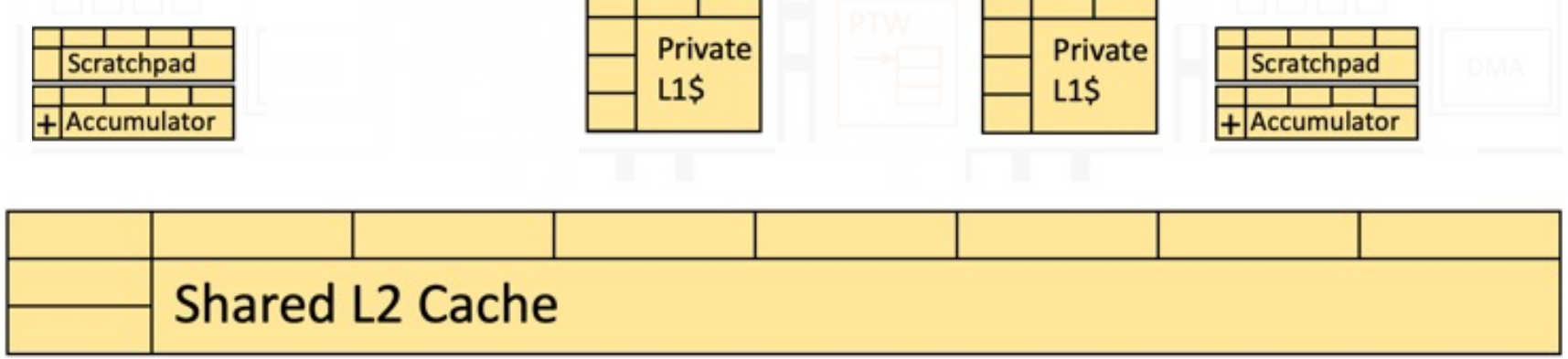


Two Billion Devices and Counting: An Industry Perspective on the State of Mobile Computer Architecture, IEEE Micro’2018

Full-System Visibility for DL Accelerators



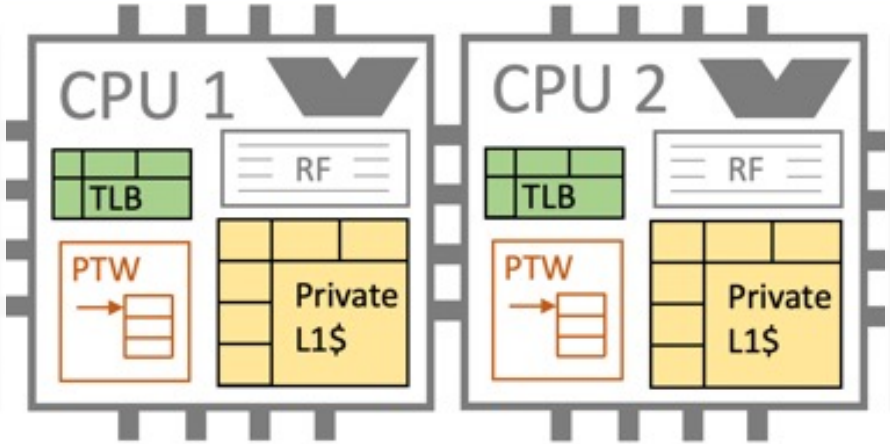
Full-System Visibility: Memory Hierarchy



Performance Impacts

Resource contention, cache coherence, etc.

Full-System Visibility: Host CPUs

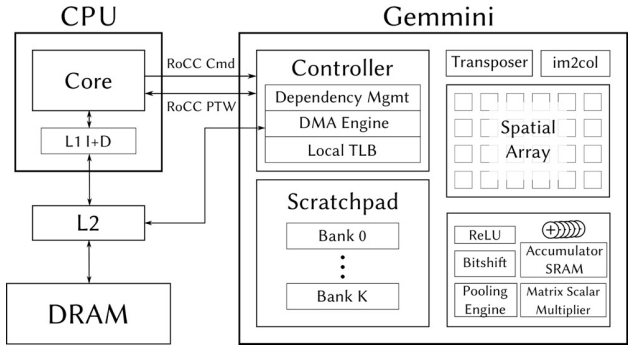


Performance Impacts

Unaccelerated kernels, etc.

Gemmini: Full-System Co-Design of Hardware Accelerators

- **Full-stack**
  - Includes OS
  - End-to-end workloads
  - “Multi-level” API
- **Full-SoC**
  - Host CPUs
  - Shared memory hierarchies
  - Virtual address translation

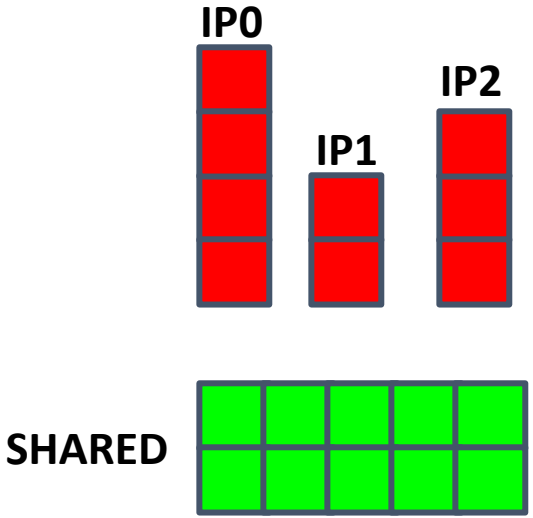


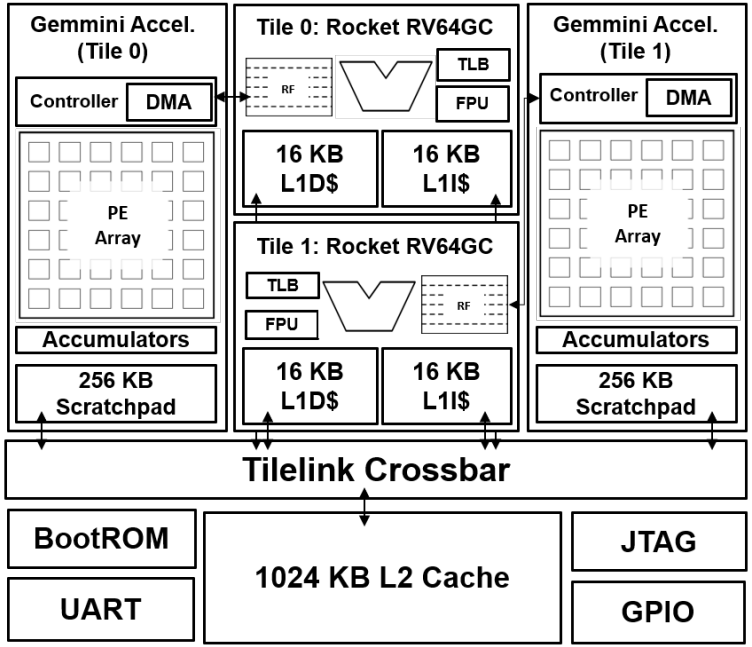
	Property	NVDLA	VTA	PolySA	DNNBuilder	MAGNet	DNNWeaver	MAERI	Gemmini
Hardware Architecture Template	Multiple Datatypes	Int/Float	Int	Int	Int	Int	Int	Int	Int/Float
	Multiple Dataflows	✗	✗	✓	✓	✓	✓	✓	✓
	Spatial Array	vector	vector	systolic	systolic	vector	vector	vector	vector/systolic
	Direct convolution	✓	✗	✗	✓	✓	✓	✓	✓
Programming Support	Software Ecosystem	Custom Compiler	TVM	Xilinx SDAccel	Caffe	C	Caffe	Custom Mapper	ONNX/C
	Hardware-Supported Virtual Memory	✗	✗	✗	✗	✗	✗	✗	✓
System Support	Full SoC	✗	✗	✗	✗	✗	✗	✗	✓
	OS Support	✓	✓	✗	✗	✗	✗	✗	✓

https://github.com/ucb-bar/gemmini

[DAC’2021 Best Paper Award]

Gemmini Case Study: Allocating on-chip SRAM



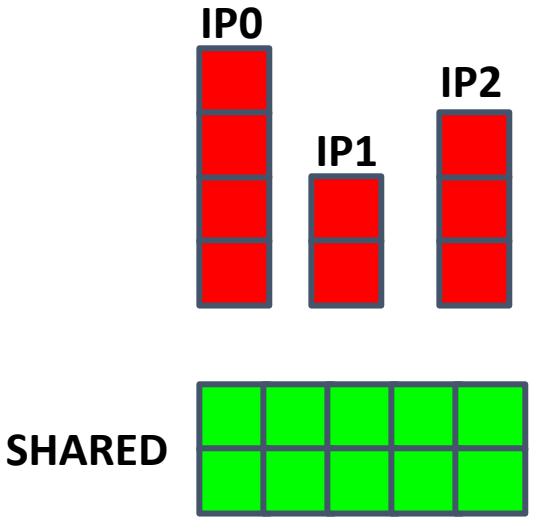


- **Where to allocated SRAM?**
  - Private within each IP
  - Shared

https://github.com/ucb-bar/gemmini

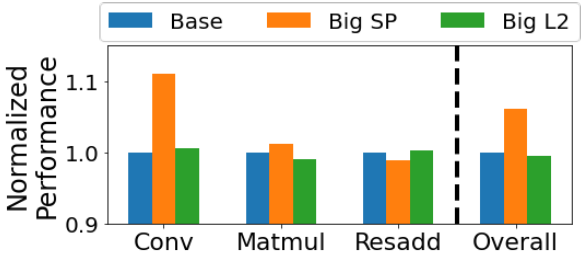
[DAC'2021 Best Paper Award]

Gemmini Case Study: Allocating on-chip SRAM



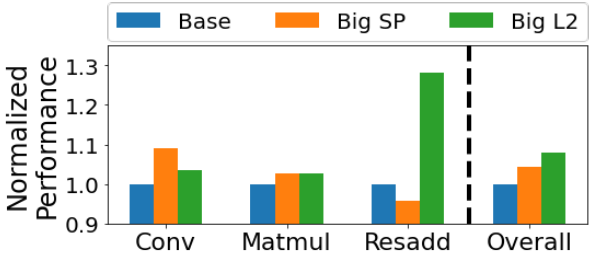
- **Where to allocated SRAM?**
  - Private within each IP
  - Shared

- **Application dependent.**



Single-Core SoC

- **SoC configuration dependent.**

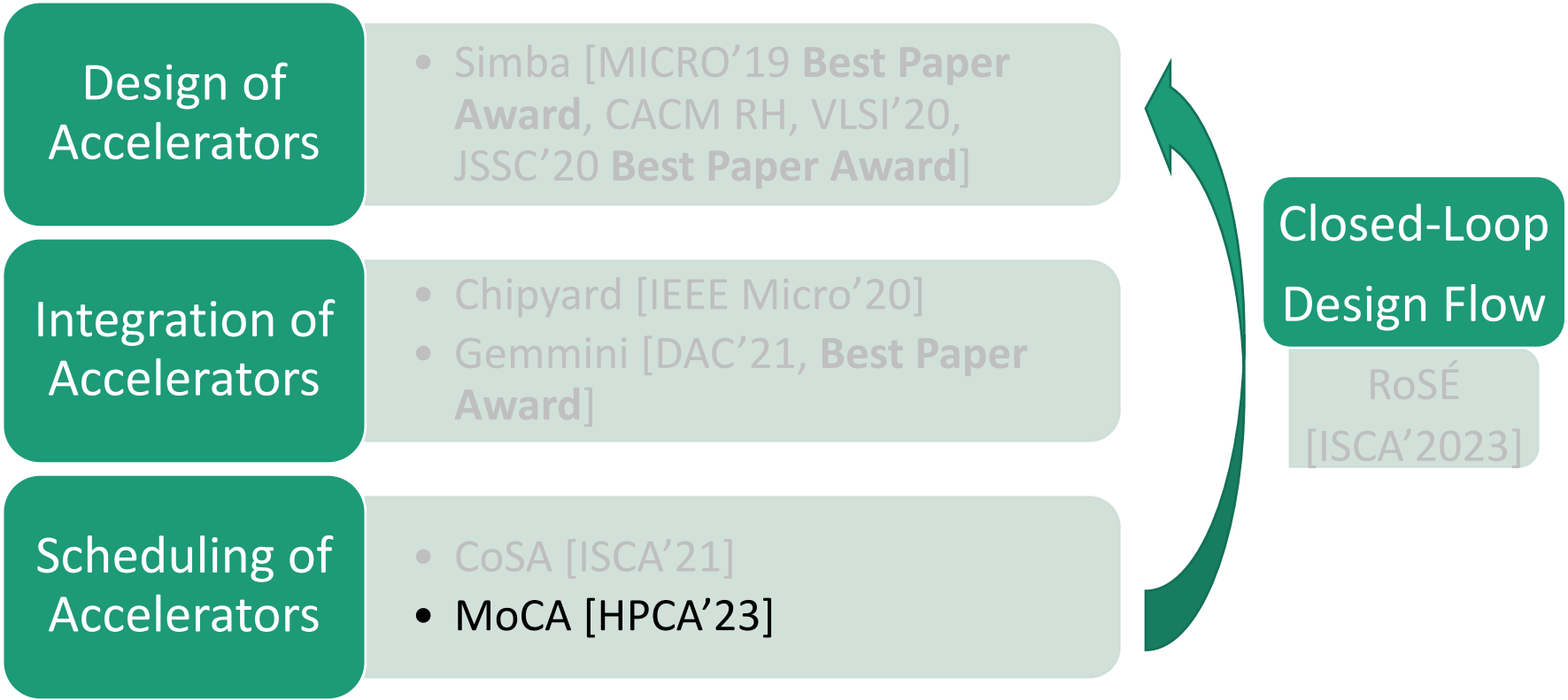


Dual-Core SoC

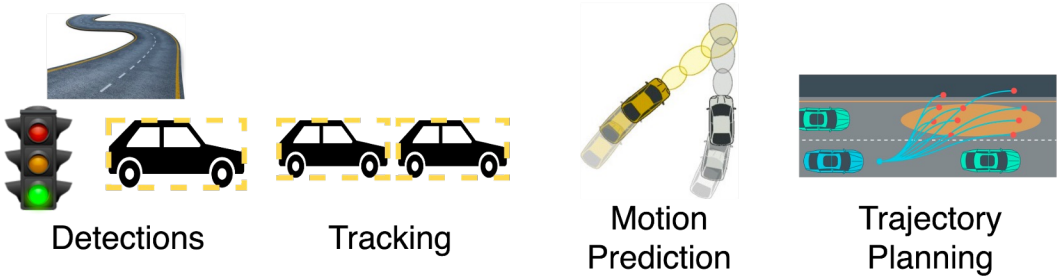
https://github.com/ucb-bar/gemmini

[DAC'2021 Best Paper Award]

Full-Stack Optimization for Domain-Specific Systems

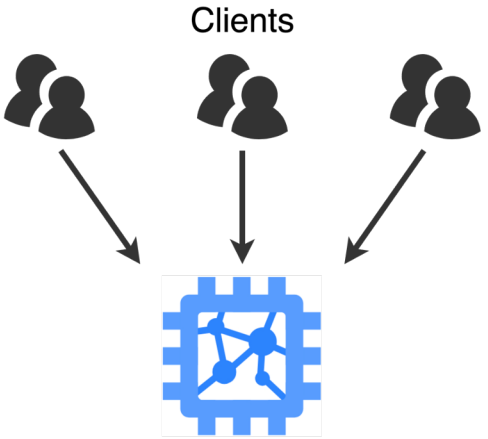


Tasks are not running in isolation



Autonomous Vehicle

@ Edge

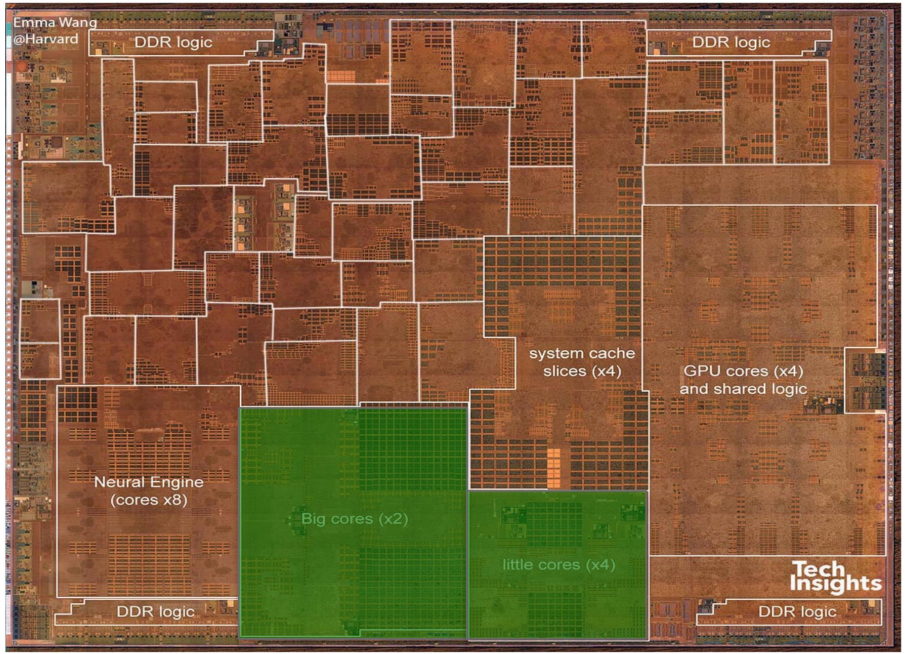


@ Cloud

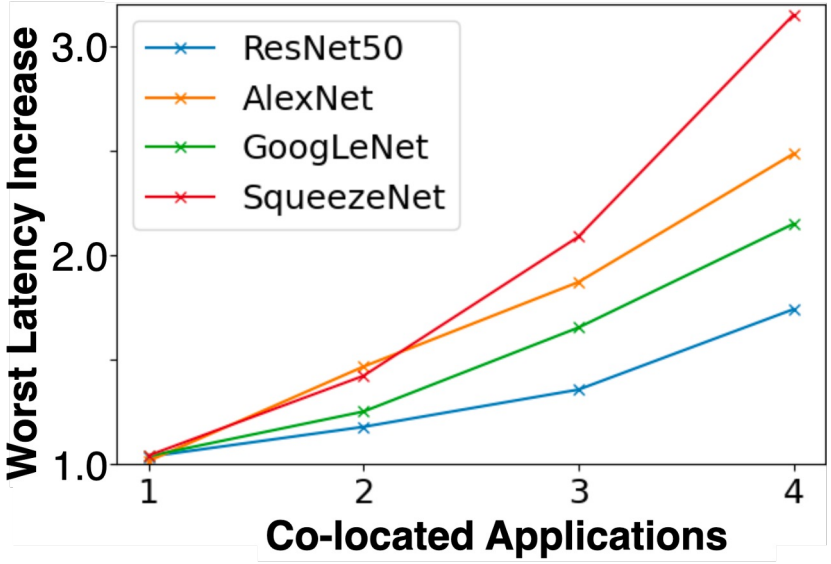
https://github.com/ucb-bar/MoCA

[HPCA'2023]

Accelerator Concurrency to Support Parallel Tasks

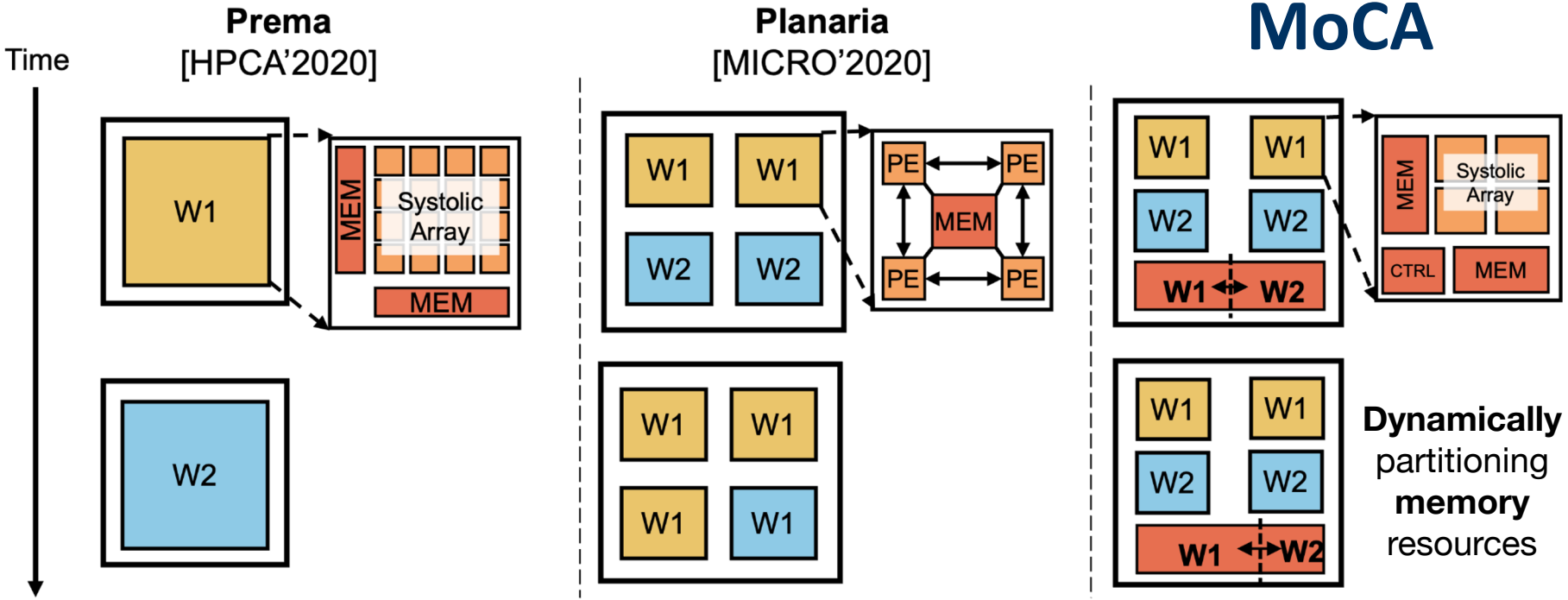


2019 Apple A12 w/ 42 accelerators



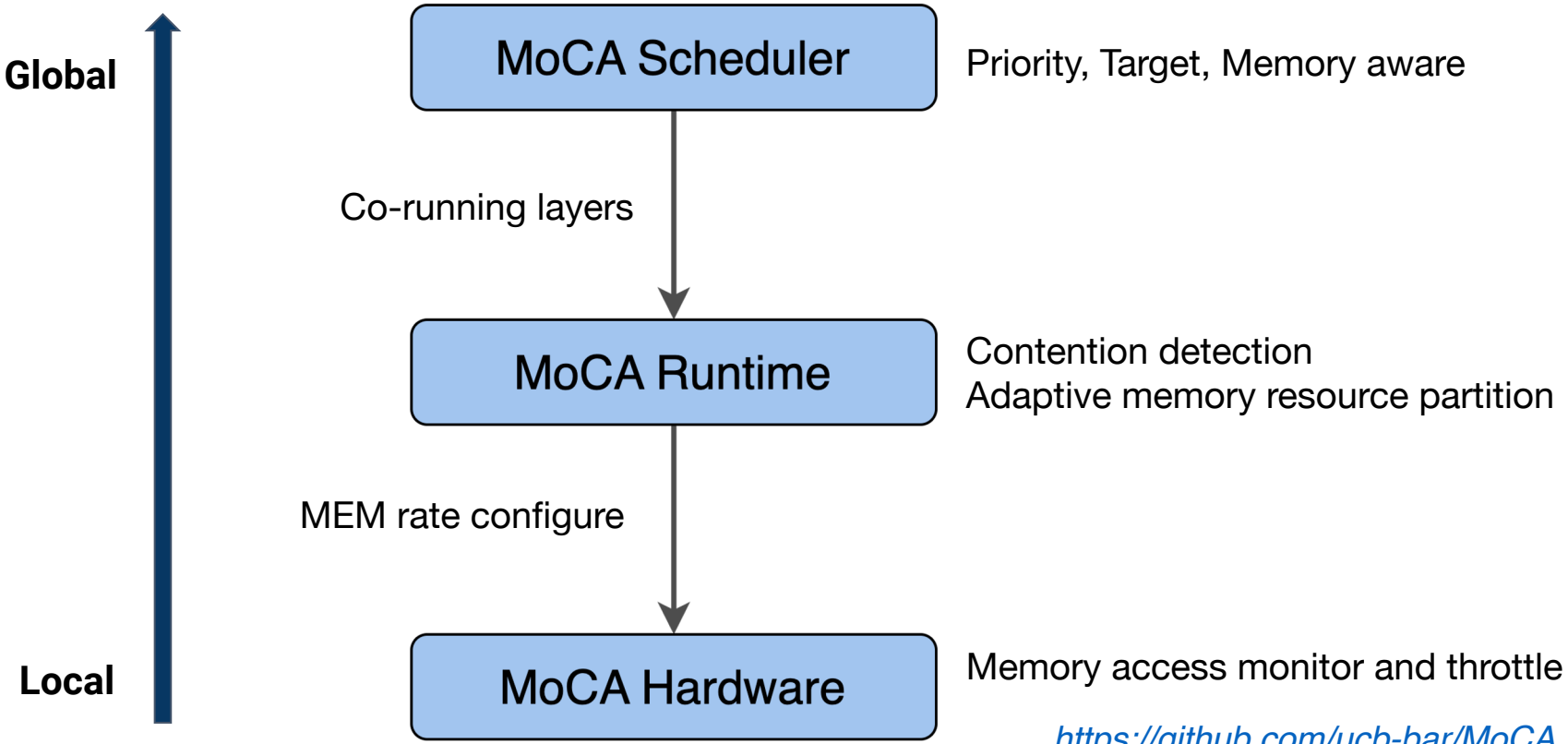
https://github.com/ucb-bar/MoCA
[HPCA'2023]

Need for Adaptive Resource Partition



https://github.com/ucb-bar/MoCA
[HPCA'2023]

MoCA's Full-Stack Approach

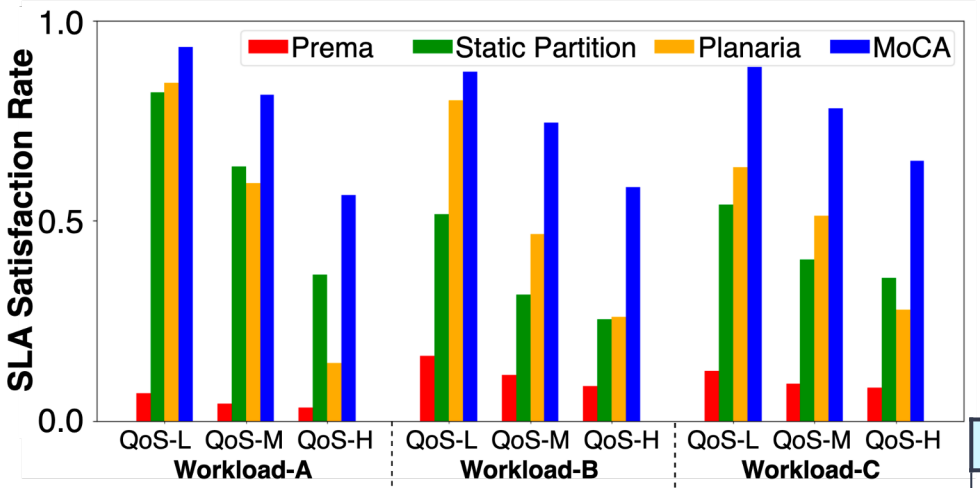


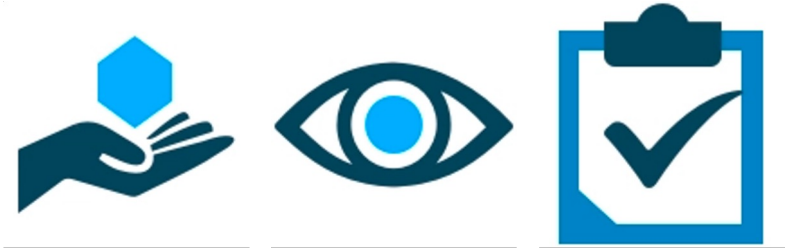
https://github.com/ucb-bar/MoCA

[HPCA'2023]

Results

- Improve the overall quality of service (QoS) by 2-8X on average.





Artifact evaluated & reproduced

Workload	Model Size	DNN Models
Workload set-A	Light	SqueezeNet, Yolo-LITE, KWS
Workload set-B	Heavy	GoogLeNet, AlexNet, ResNet50, YoloV2
Workload set-C	Mixed	All

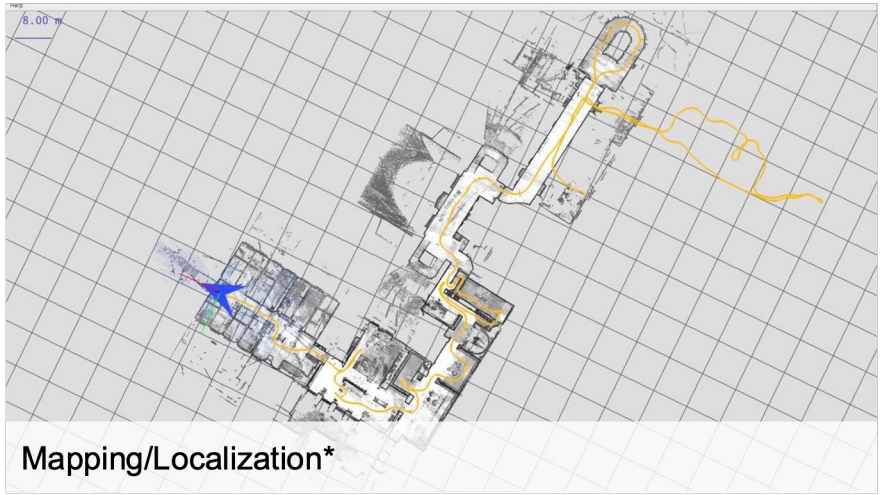
GEMMINI CHIPYARD FireSim

https://github.com/ucb-bar/MoCA
[HPCA'2023]

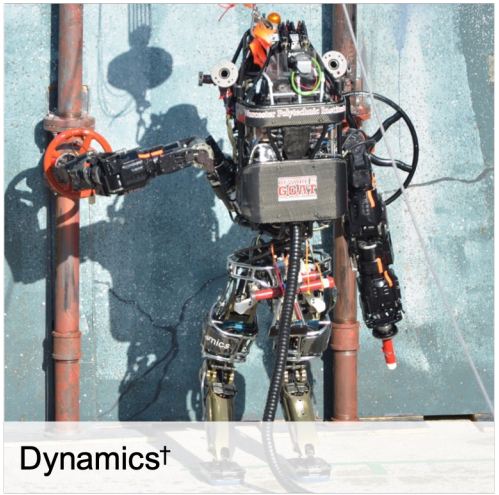
Full-Stack Optimization for Domain-Specific Systems

Why Robotics?

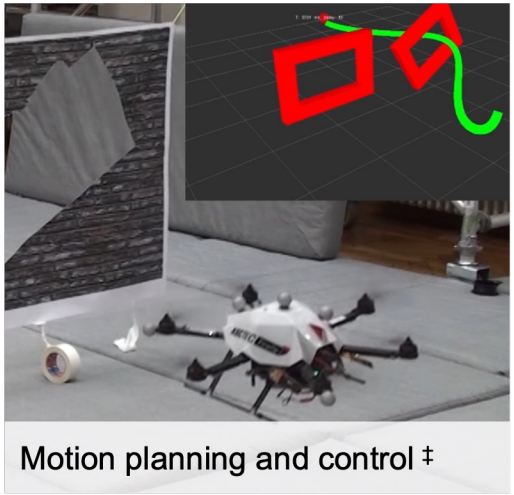
- Increasingly complex systems, tighter latency/energy constraints



Mapping/Localization*



Dynamics†



Motion planning and control ‡

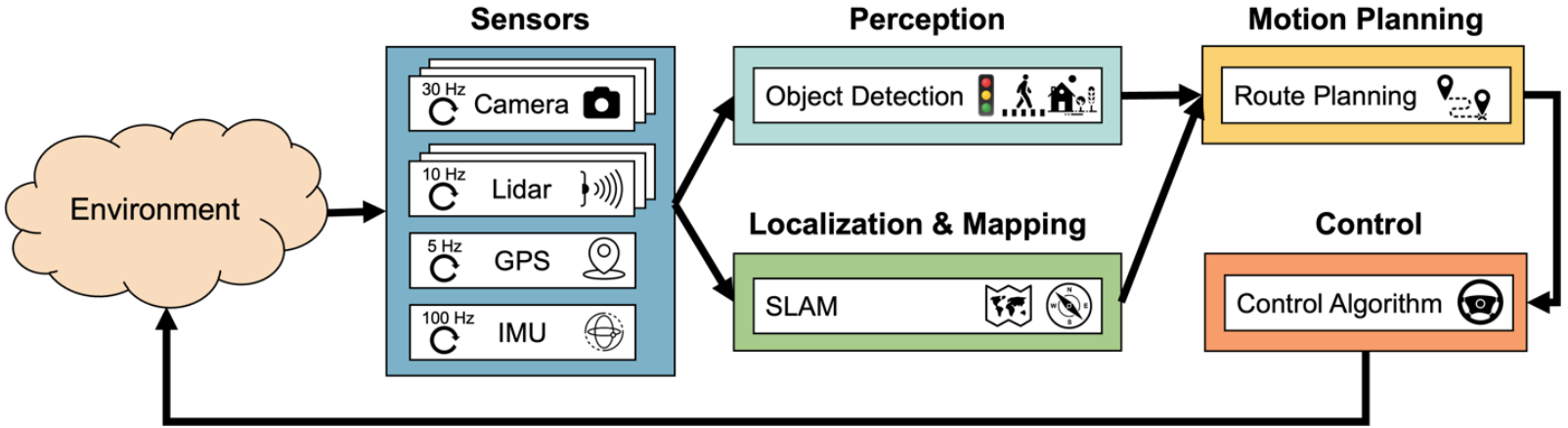
* https://google-cartographer.readthedocs.io/en/latest/

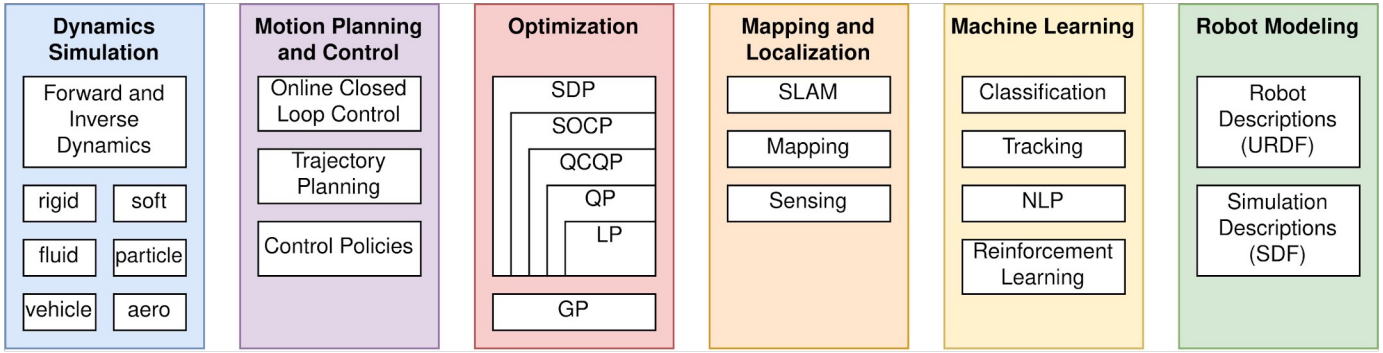
† S. Feng, E. Whitman, X. Xinjilefu and C. G. Atkeson, "Optimization based full body control for the atlas robot," 2014 IEEE-RAS International Conference on Humanoid Robots, Madrid, Spain, 2014, pp. 120-127, doi: 10.1109/HUMANOIDS.2014.7041347.

‡M. Neunert et al., "Fast nonlinear Model Predictive Control for unified trajectory optimization and tracking," 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 2016, pp. 1398-1404, doi: 10.1109/ICRA.2016.7487274.

https://github.com/ucb-bar/RoSE
[ISCA'2023]

Challenge: Diverse, Closed-loop Tasks

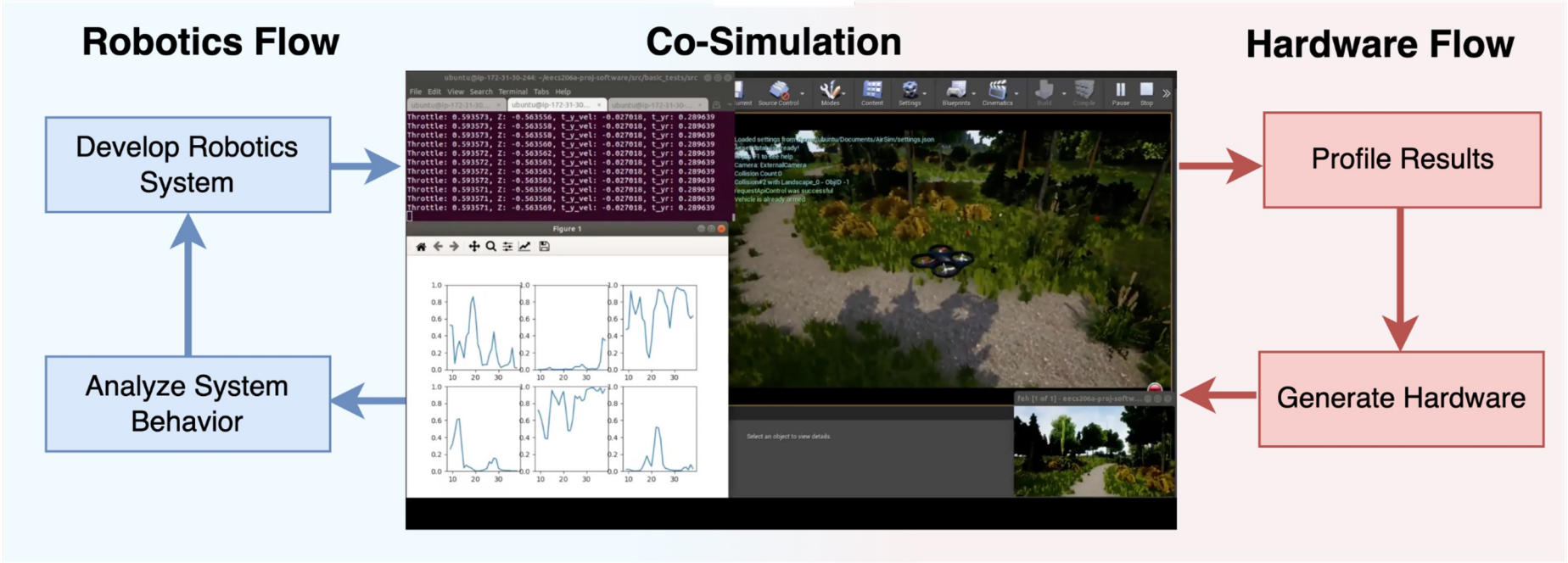




https://github.com/ucb-bar/RoSE

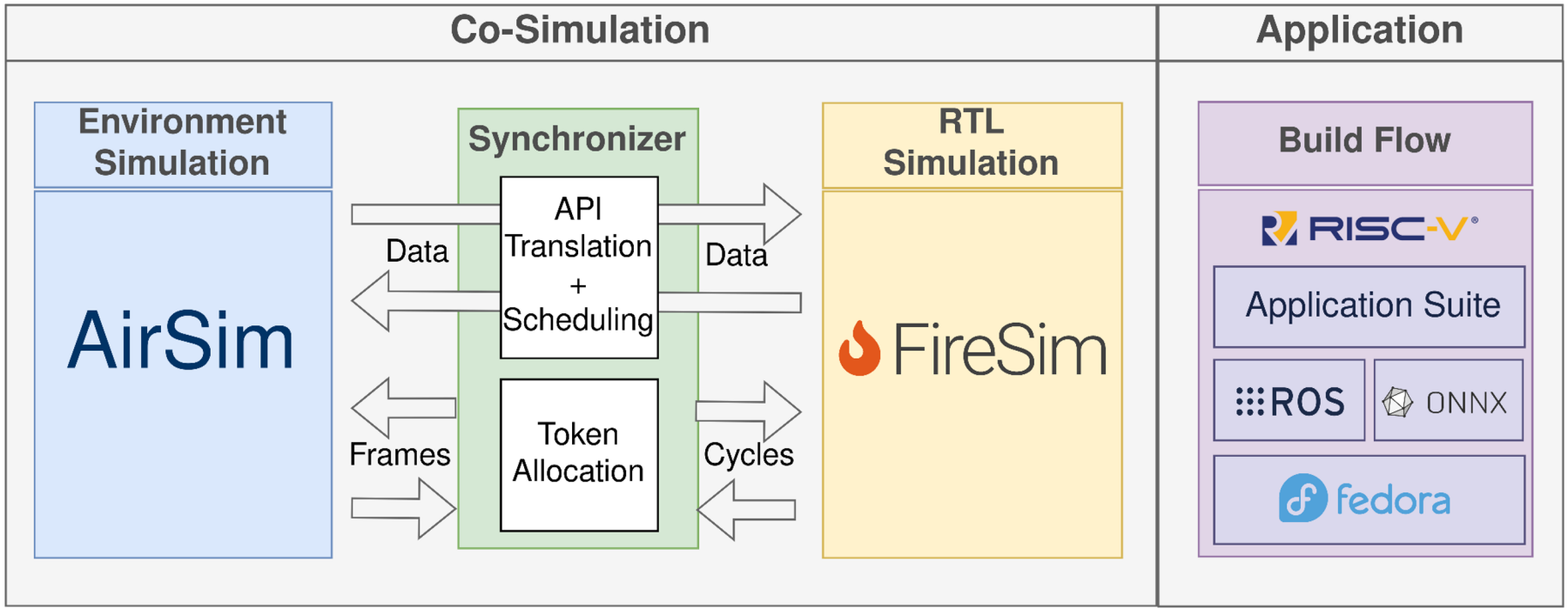
[ISCA'2023]

Capture Closed-Loop Effects w/ HW-SW Co-Simulation



https://github.com/ucb-bar/RoSE
[ISCA'2023]

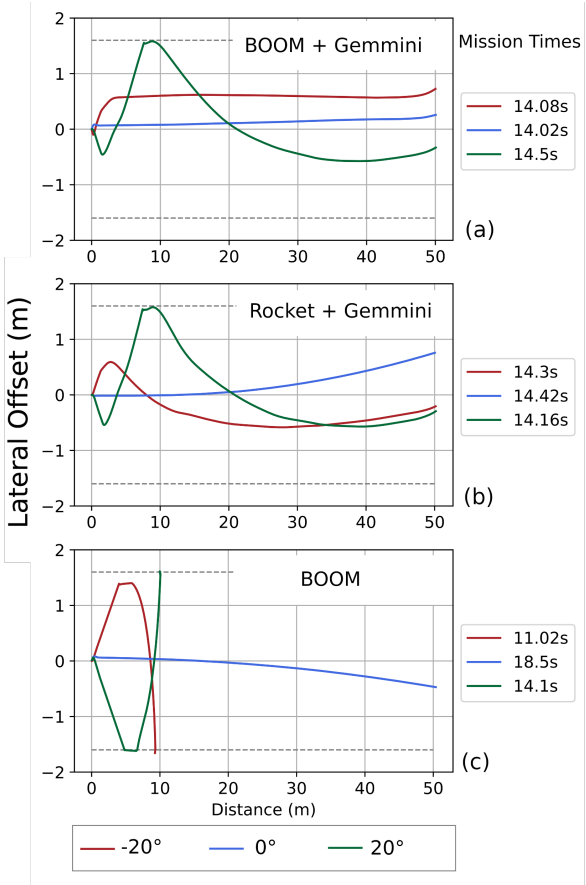
RoSÉ: Pre-Silicon Full-Stack Robotics Soc Evaluator



https://github.com/ucb-bar/RoSE

[ISCA'2023]

End-to-End Evaluation with RoSÉ



https://github.com/ucb-bar/RoSE

[ISCA'2023]

Co-Simulation in Realistic Environment

Video link: https://youtu.be/AqtMBSd9bbM

https://github.com/ucb-bar/RoSE
[ISCA'2023]

Full-Stack Optimization for Domain-Specific Systems

Design of Accelerators

- Simba [MICRO'19 **Best Paper Award**, CACM RH, VLSI'20, JSSC'20 **Best Paper Award**]

Integration of Accelerators

- Chipyard [IEEE Micro'20]
- Gemmini [DAC'21, **Best Paper Award**]

Scheduling of Accelerators

- CoSA [ISCA'21]
- MoCA [HPCA'23]

Closed-Loop Design Flow

RoSÉ [ISCA'2023]

Acknowledgement

- Sponsored by DARPA, NSF, an Intel Rising Star Faculty Award, a Google Research Award, and ADEPT/SLICE industry sponsors!


Kevin Anderson


Jun Sun Choi


Prashanth Ganesh


Charles Hong


Coleman Hooper


Roger Hsiao


Hansung Kim


Seah Kim


Vadim Nikiforov


Joonho Whangbo


Jingyi Xu


Hasan Genc

Full-Stack Optimization for Domain-Specific Systems

Design of Accelerators

- Simba [MICRO'19 **Best Paper Award**, CACM RH, VLSI'20, JSSC'20 **Best Paper Award**]

Integration of Accelerators

- Chipyard [IEEE Micro'20]
- Gemmini [DAC'21, **Best Paper Award**]

Scheduling of Accelerators

- CoSA [ISCA'21]
- MoCA [HPCA'23]

Closed-Loop Design Flow

RoSÉ [ISCA'2023]