tinyML® Summit

Miniature dreams can come true...

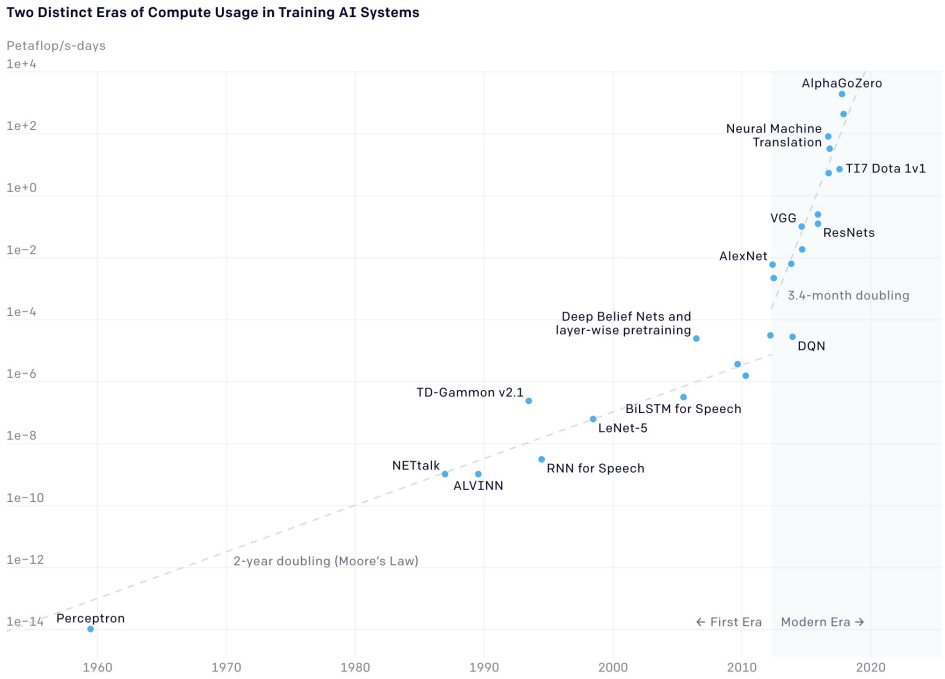
March 28-30, 2022 | San Francisco Bay Area

www.tinyML.org

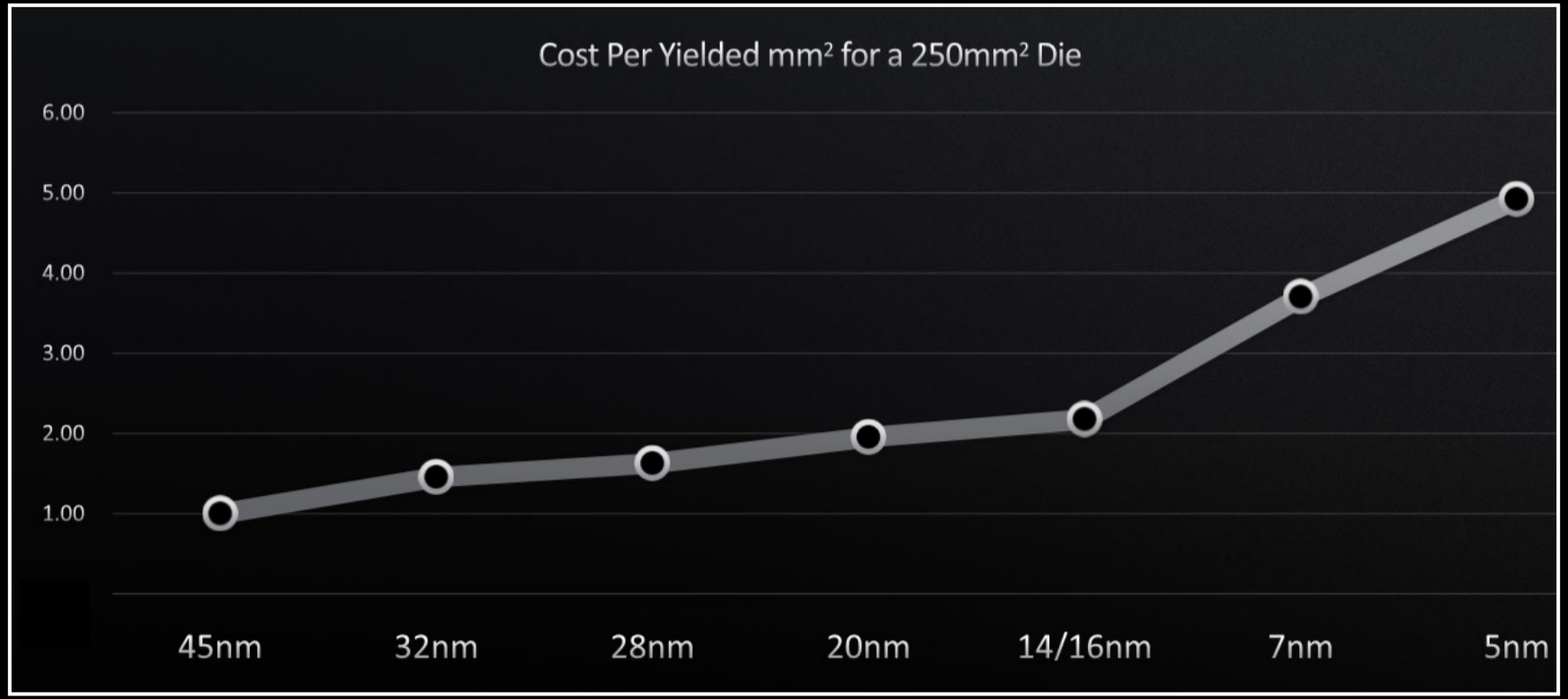
Next-Generation Deep-Learning Accelerators: From Hardware to System

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

Growing Demand in Computing



OpenAI



Slowing Supply in Computing

AMD, HotChips, 2019

Growing Demand in Computing

Slowing Supply in Computing

Growing Demand in Computing

+

Slowing Supply in Computing

=

Domain-Specific Accelerators

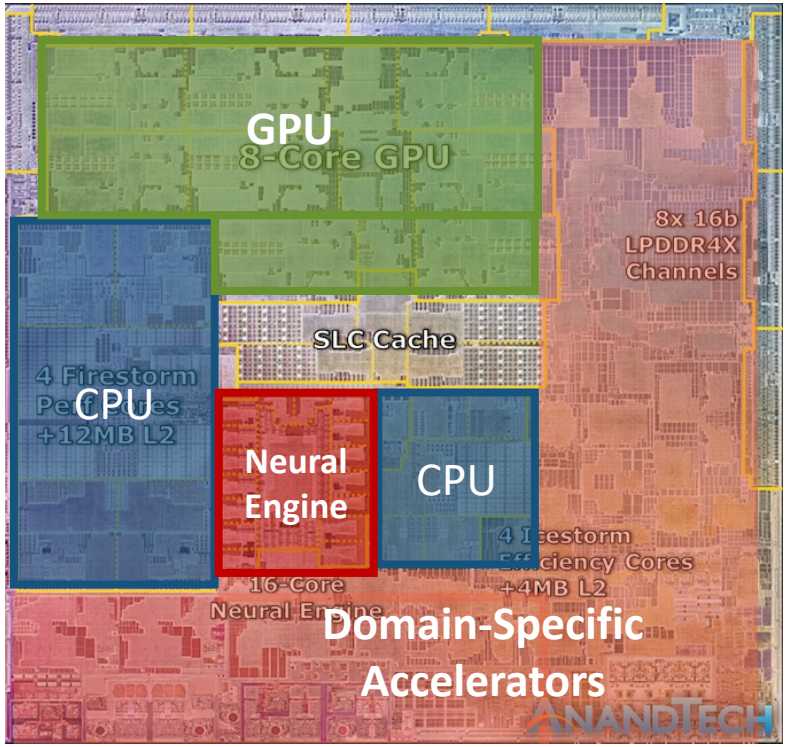
Growing Demand in Computing

Slowing Supply in Computing

Domain-Specific Accelerators

- Customized hardware designed for a domain of applications.

Apple M1 Chip
2020



* AnandTech

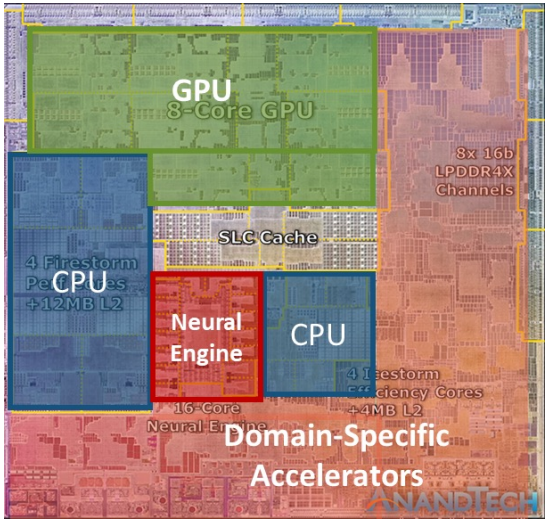
Full-Stack Optimization for DL Accelerators

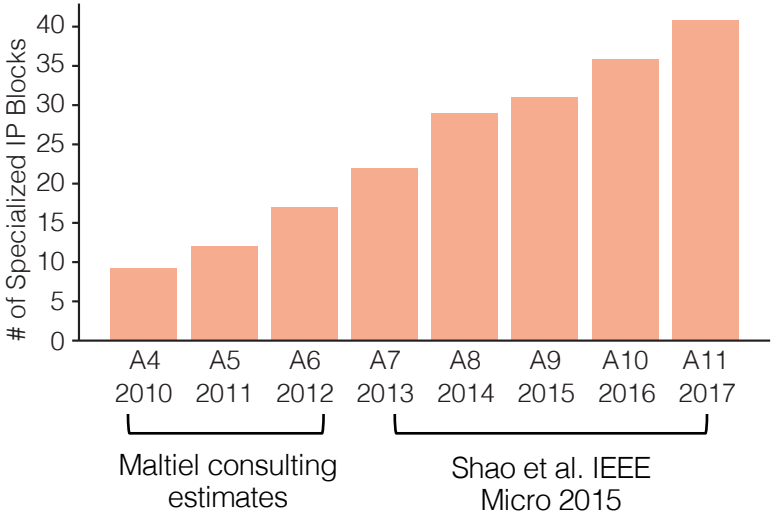
Design of Accelerators	• MAGNet [ICCAD'2019] • Simba [MICRO'2019 **Best Paper Award**]
Integration of Accelerators	• Chipyard [IEEE Micro'2020] • Gemmini [DAC'2021 **Best Paper Award**]
Scheduling of Accelerators	• CoSA [ISCA'2021]

Full-Stack Optimization for DL Accelerators

Design of Accelerators	• MAGNet [ICCAD'2019] • Simba [MICRO'2019 **Best Paper Award**]
Integration of Accelerators	• Chipyard [IEEE Micro'2020] • **Gemmini [DAC'2021 Best Paper Award]**
Scheduling of Accelerators	• CoSA [ISCA'2021]

Accelerators don’t exist in isolation.

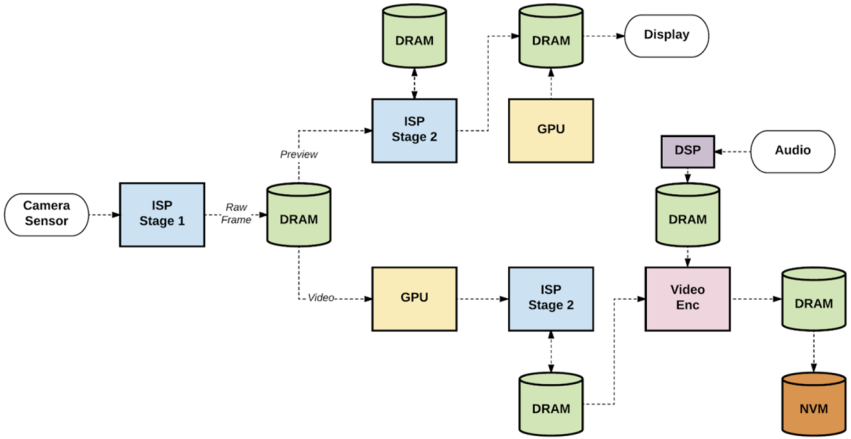




http://vlsiarch.eecs.harvard.edu/research/accelerators/die-photo-analysis/

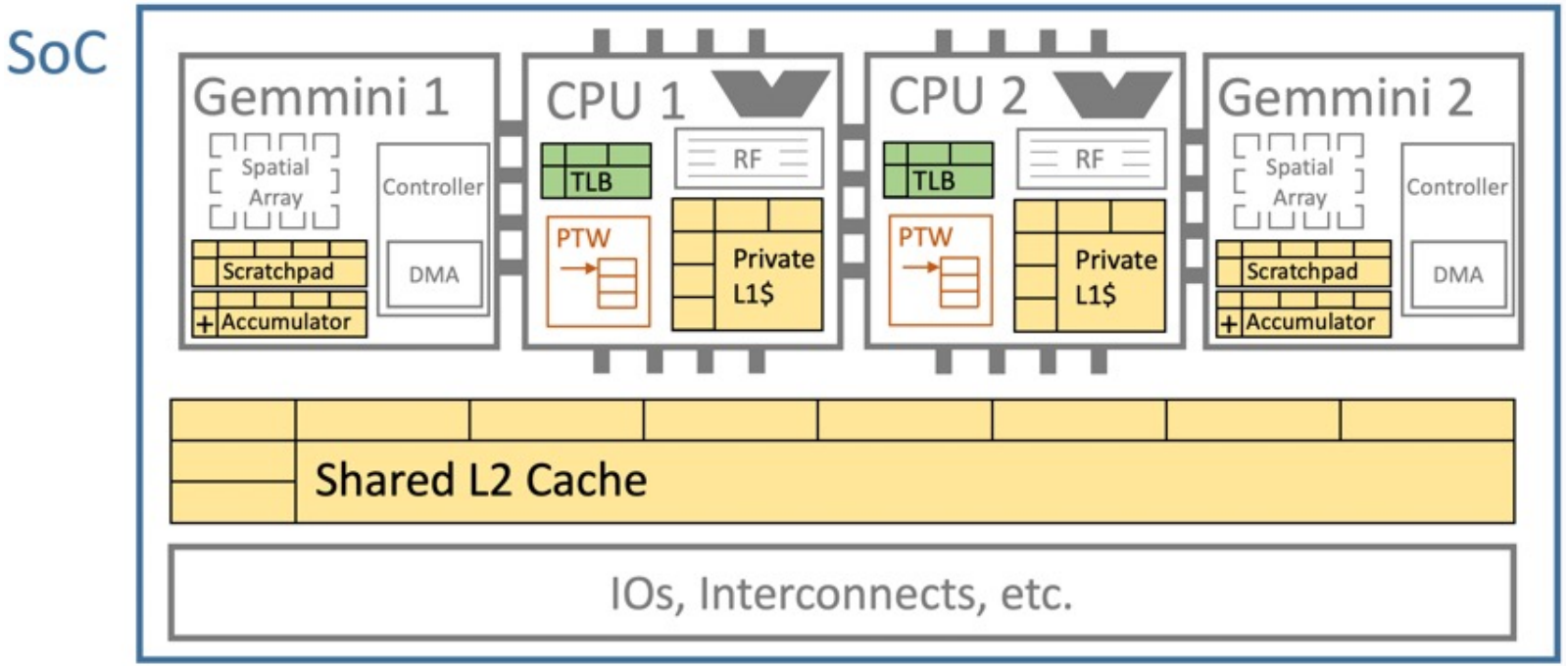
Mobile SoC Usecase

- 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

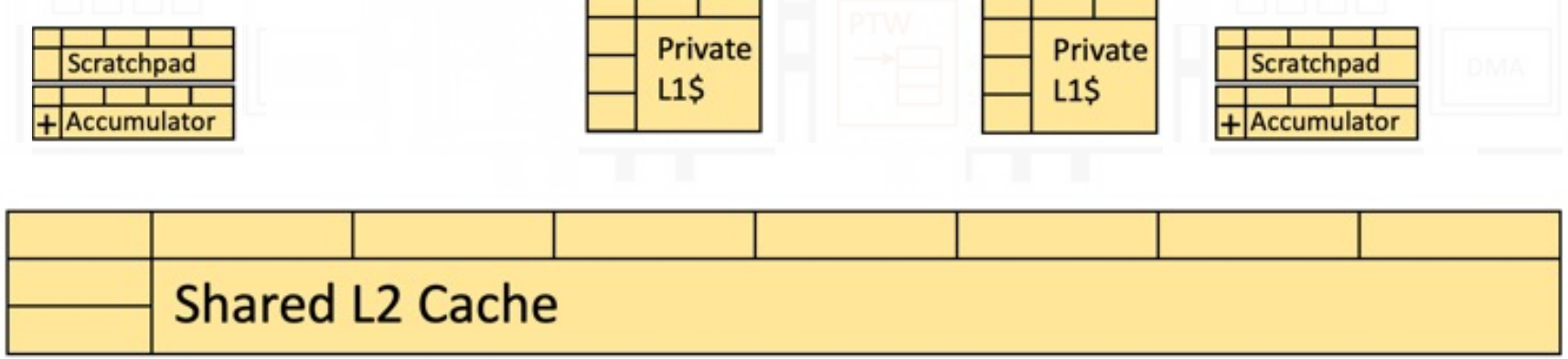


Two Billion Devices and Counting: An Industry Perspective on the State of Mobile Computer Architecture, IEEE Micro’2018

Full-System Visibility for DL Accelerators



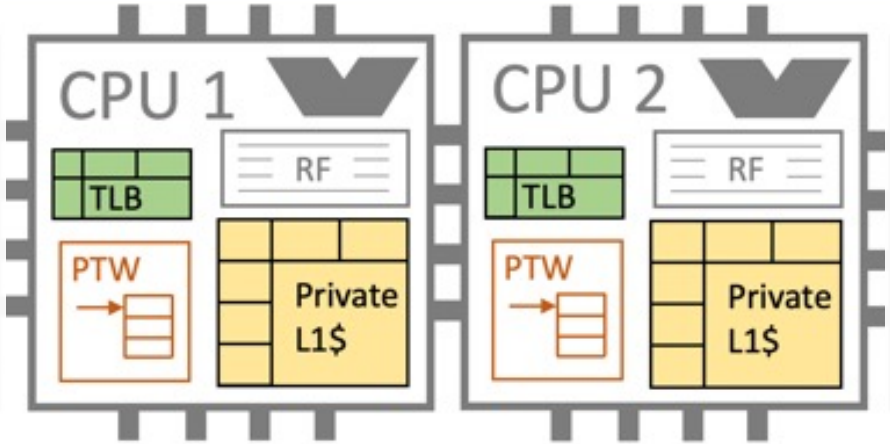
Full-System Visibility: Memory Hierarchy



Performance Impacts

Resource contention, cache coherence, etc.

Full-System Visibility: Host CPUs

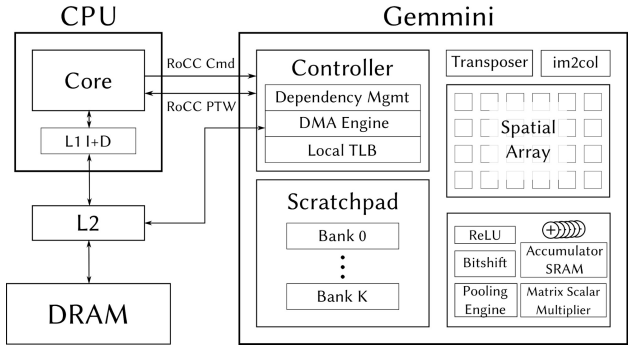


Performance Impacts

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

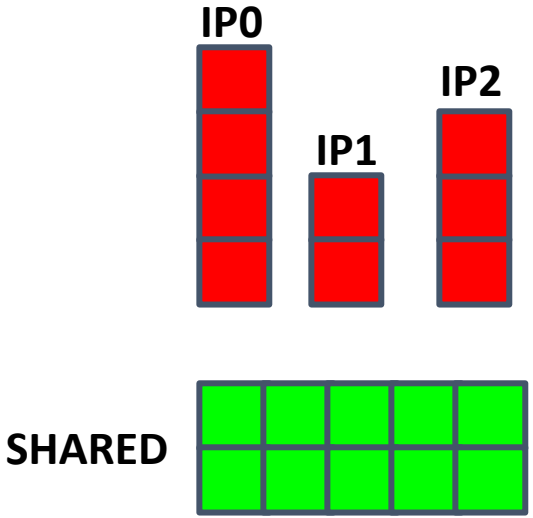


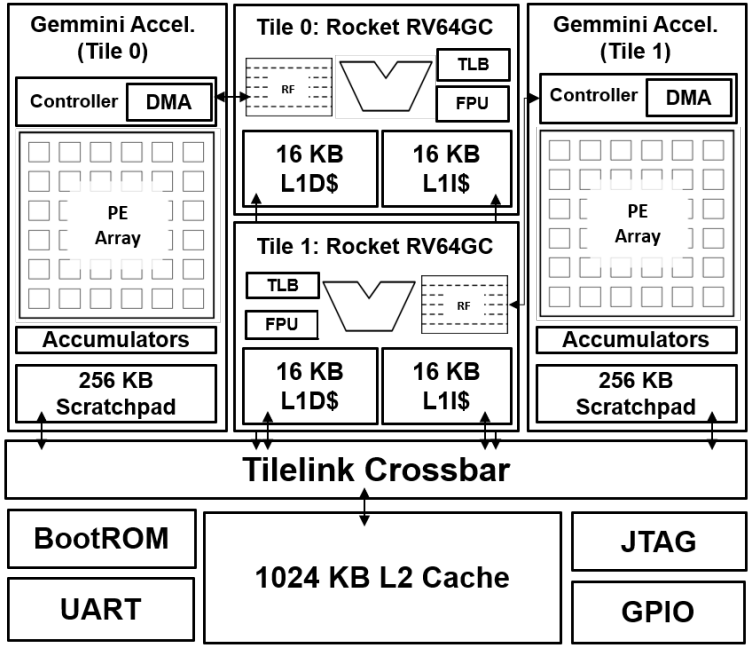
	Property	NVDLA	VTA	PolySA	DNNBuilder	MAGNet	DNNWeaver	MAERI	Gemmini
Hardware Architecture Template	Multiple Datatypes	Int/Float	Int	Int	Int	Int	Int	Int	Int/Float
	Multiple Dataflows	✗	✗	✓	✓	✓	✓	✓	✓
	Spatial Array	vector	vector	systolic	systolic	vector	vector	vector	vector/systolic
	Direct convolution	✓	✗	✗	✓	✓	✓	✓	✓
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



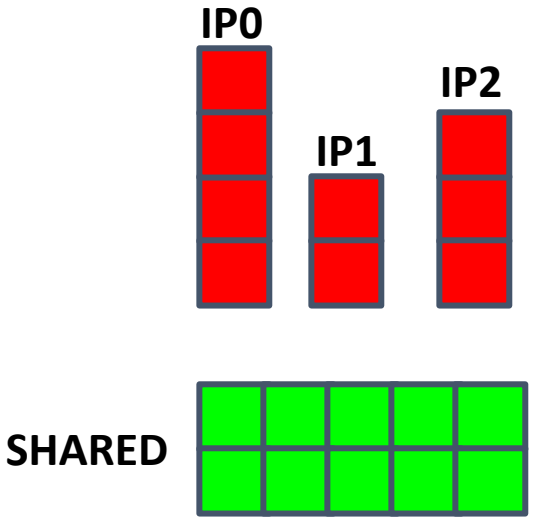


- **Where to allocated SRAM?**
  - Private within each IP
  - Shared

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

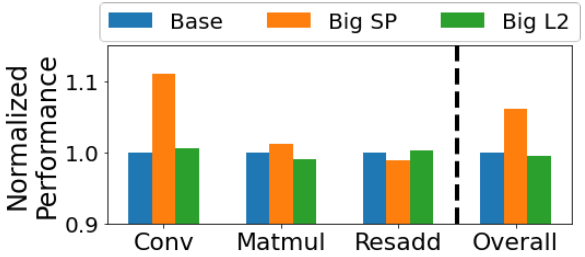
[DAC'2021 Best Paper Award]

Gemmini Case Study: Allocating on-chip SRAM



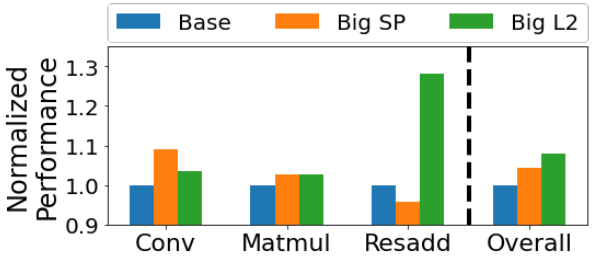
- **Where to allocated SRAM?**
  - Private within each IP
  - Shared

- **Application dependent.**



Single-Core SoC

- **SoC configuration dependent.**



Dual-Core SoC

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

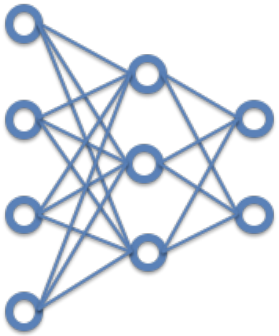
[DAC'2021 Best Paper Award]

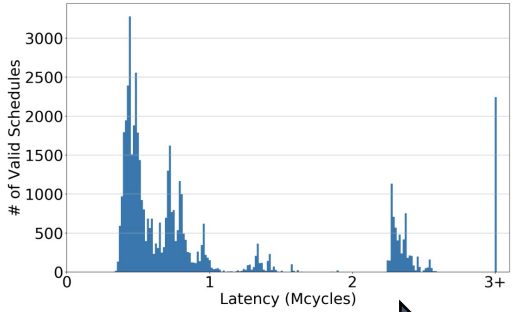
Full-Stack Optimization for DL Accelerators

Design of Accelerators	• MAGNet [ICCAD'2019] • Simba [MICRO'2019 **Best Paper Award**]
Integration of Accelerators	• Chipyard [IEEE Micro'2020] • Gemmini [DAC'2021 **Best Paper Award**]
Scheduling of Accelerators	• **CoSA [ISCA'2021]**

Large Space of Mapping Algorithms to ML Hardware

Algorithm

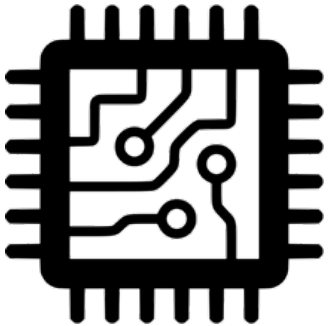




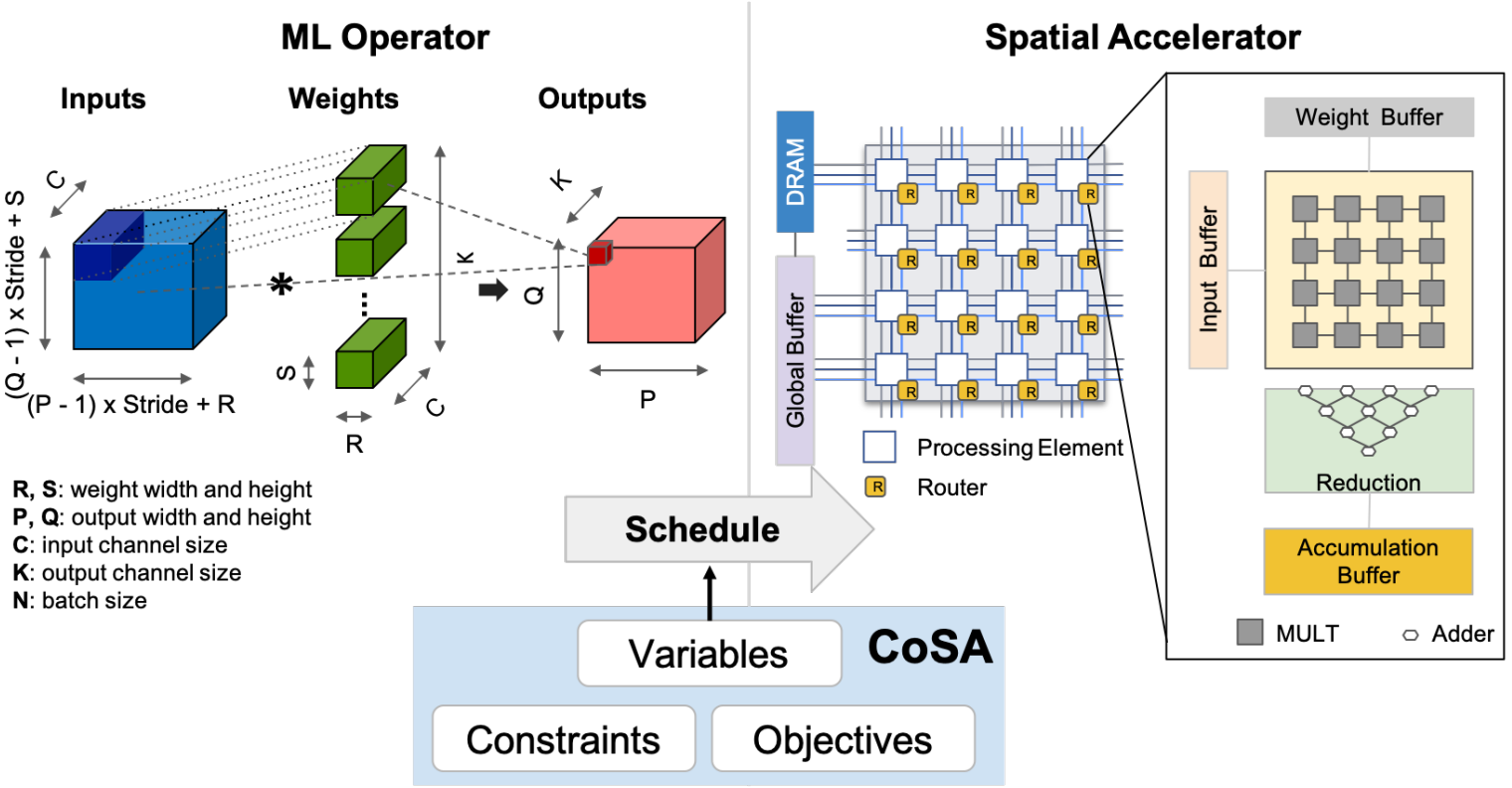
Scheduling

Scheduler	Search Algorithm
Brute-force approcahes:	
Timeloop	Brute-force & Random
dMazeRunner	Brute-force
Interstellar	Brute-force
Marvel	Decoupled Brute-force
Feedback-based Approaches:	
AutoTVM	ML-based Iteration
Halide	Beamsearch, OpenTuner
FlexFlow	MCMC
Constrained Optimization Approaches:	
CoSA	**Mixed Integer Programming (MIP)**

Hardware

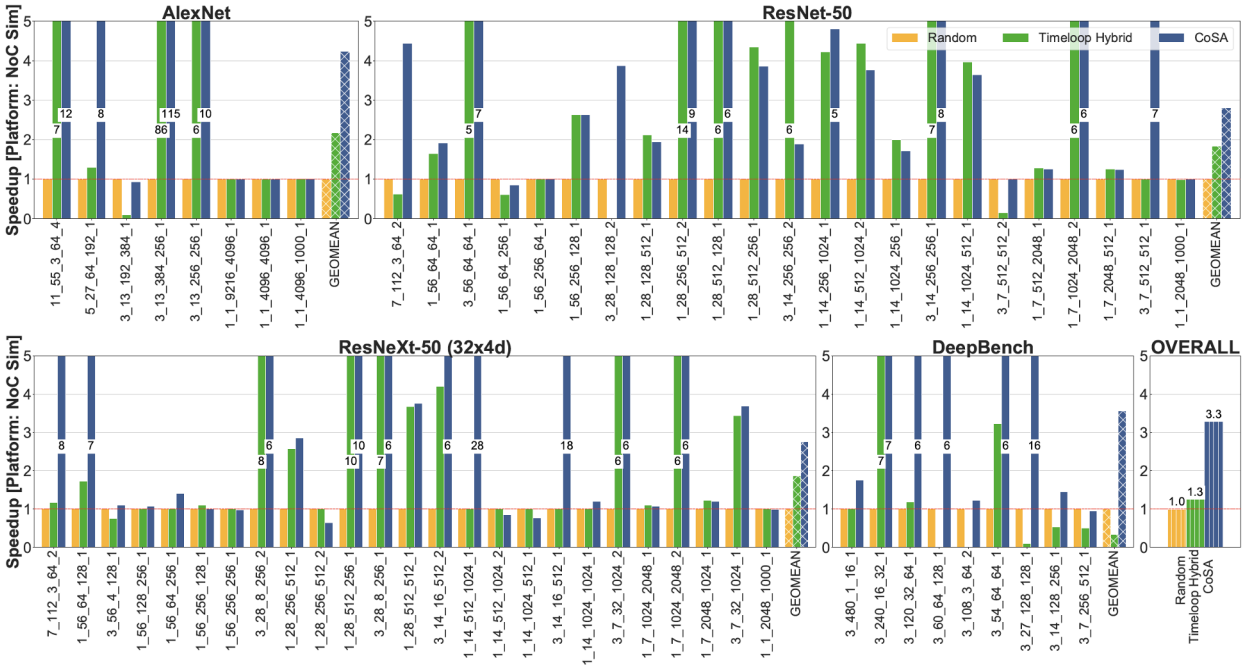


CoSA: Constrained-Optimization for Spatial Architecture



[ISCA'2021]

Results

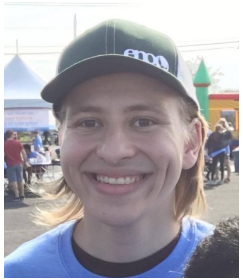


	CoSA	Random (5×)	Timeloop Hybrid
Avg. Runtime / Layer	4.2s	4.6s	379.9s
Avg. Samples / Layer	1	20K	67M
Avg. Evaluations / Layer	1	5	16K

2.5x speedup compared to SoTA with 90x faster time-to-solution.

[ISCA'2021]

Acknowledgement



Hasan Genc

Jenny Huang

Seah Kim

- Sponsored by DARPA and ADEPT/SLICE Lab industry sponsors!

Full-Stack Optimization for DL Accelerators

Design of Accelerators	• MAGNet [ICCAD'2019] • Simba [MICRO'2019 **Best Paper Award**]
Integration of Accelerators	• Chipyard [IEEE Micro'2020] • Gemmini [DAC'2021 **Best Paper Award**]
Scheduling of Accelerators	• CoSA [ISCA'2021]

tinyML Summit 2022 Sponsors

Gold:

Silver:

Copyright Notice