

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

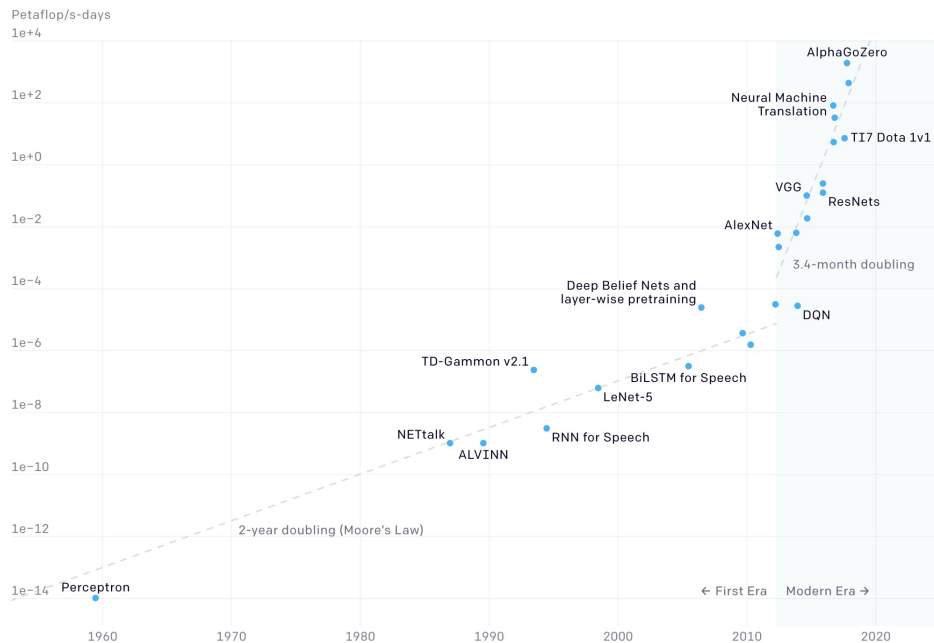
ysshao@berkeley.edu

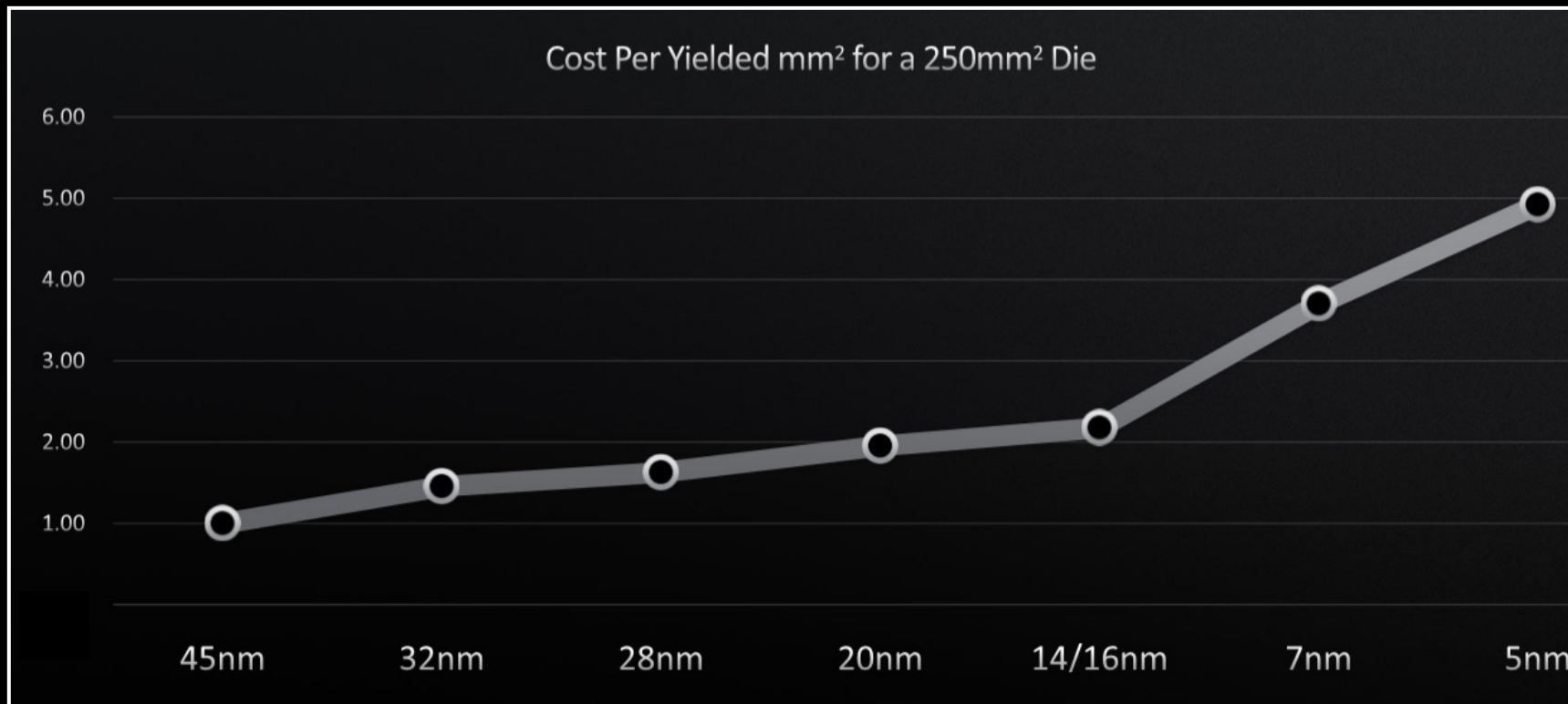
Electrical Engineering and Computer Sciences



Growing Demand in Computing

Two Distinct Eras of Compute Usage in Training AI Systems





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



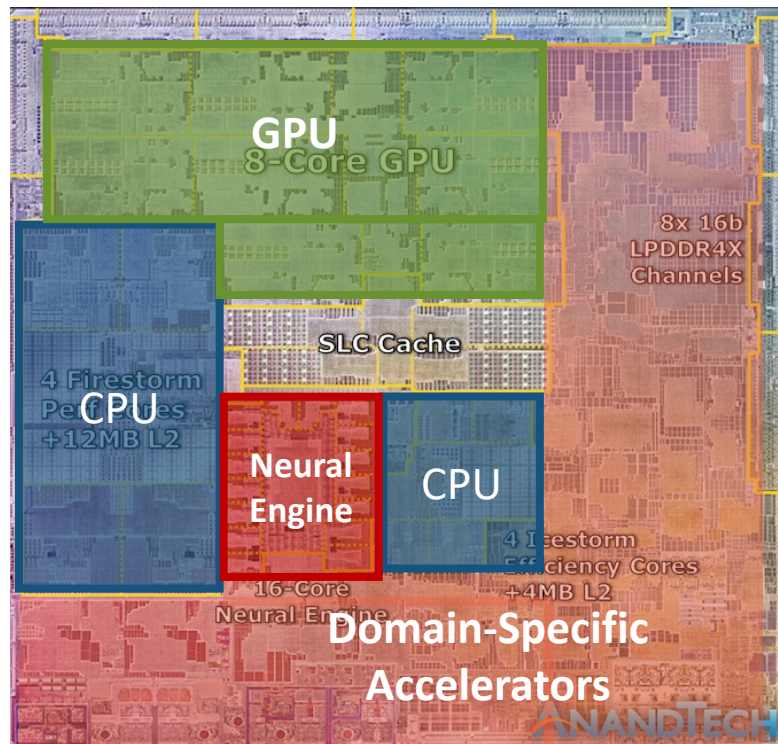
Slowing
Supply in
Computing

Domain-Specific Accelerators

- Customized hardware designed for a domain of applications.



Apple M1 Chip
2020



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**]

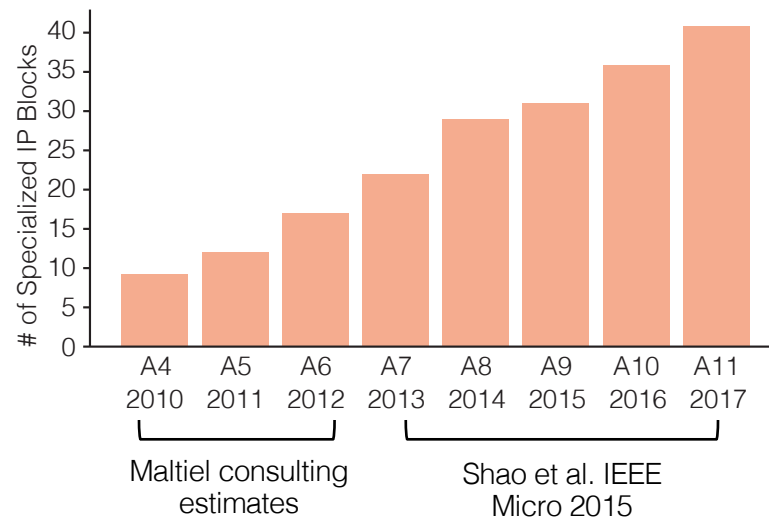
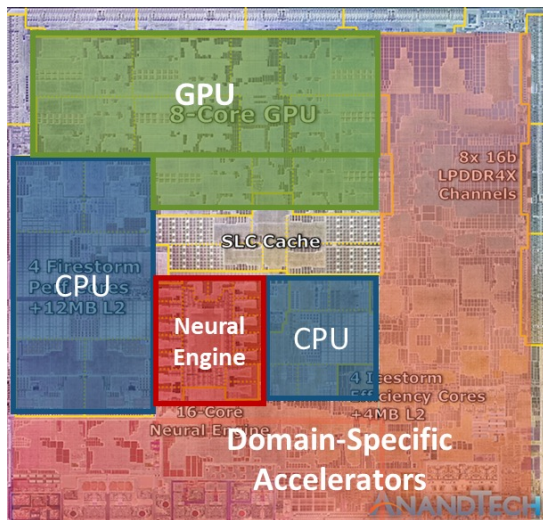
Integration of Accelerators

- Chipyard [IEEE Micro'2020]
- **Gemmini [DAC'2021 **Best Paper Award**]**

Scheduling of Accelerators

- CoSA [ISCA'2021]

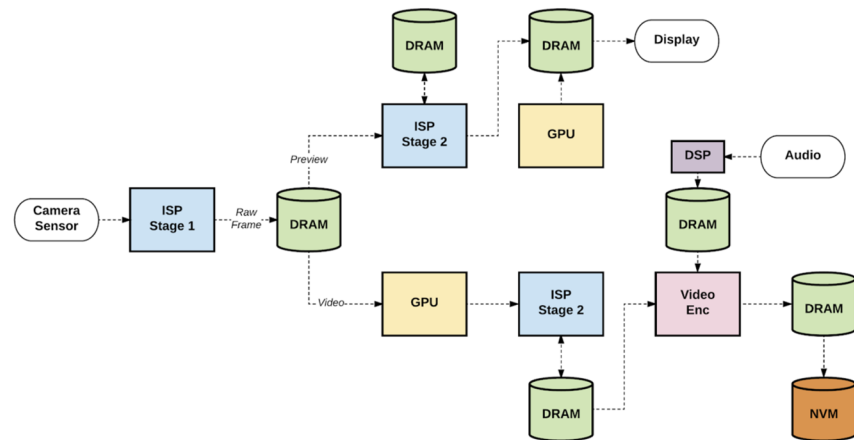
Accelerators don't exist in isolation.



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

Mobile SoC Usecase

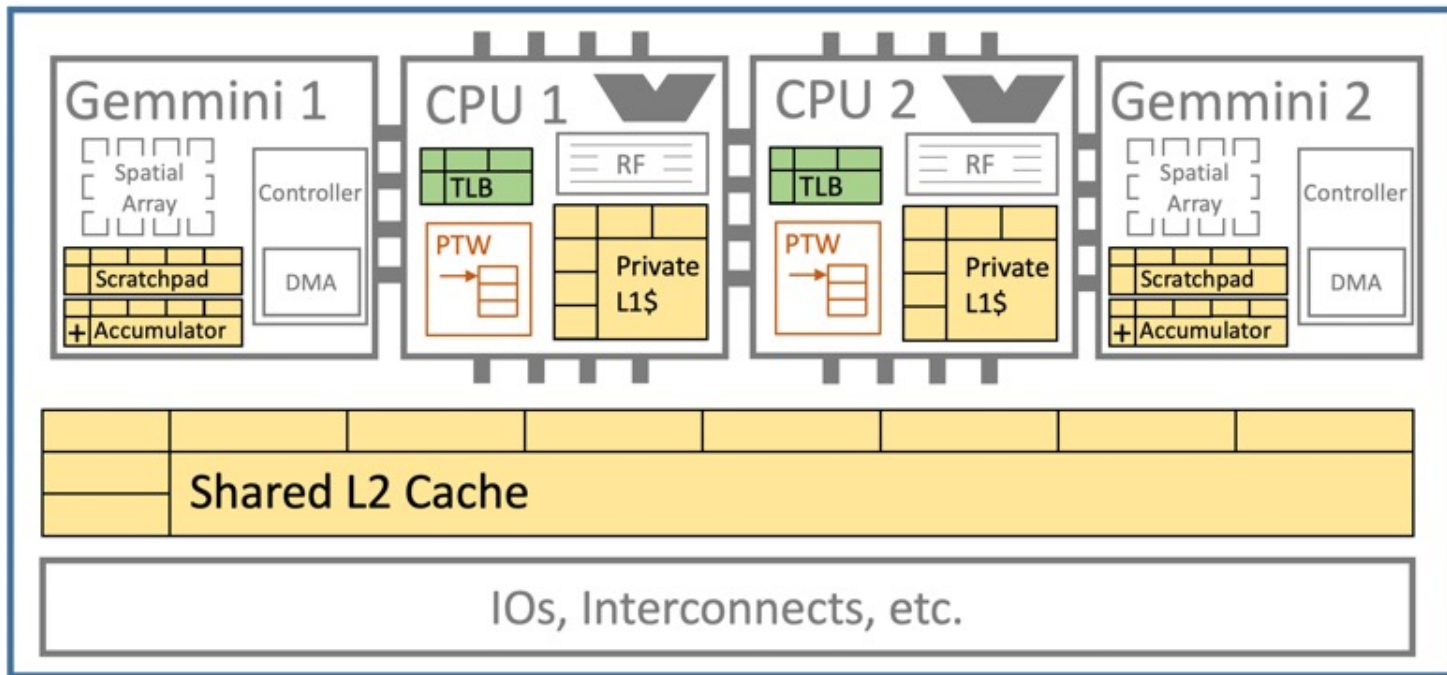
- Mainstream architecture has long focused on general-purpose CPUs and GPUs.
- In an SoC, multiple IP blocks are active at the same time and communicate frequently with each other.
- Example:
 - Recording a 4K video
 - Camera -> ISP
 - “Preview stream” for display
 - “Video stream” for storage
 - DRAM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



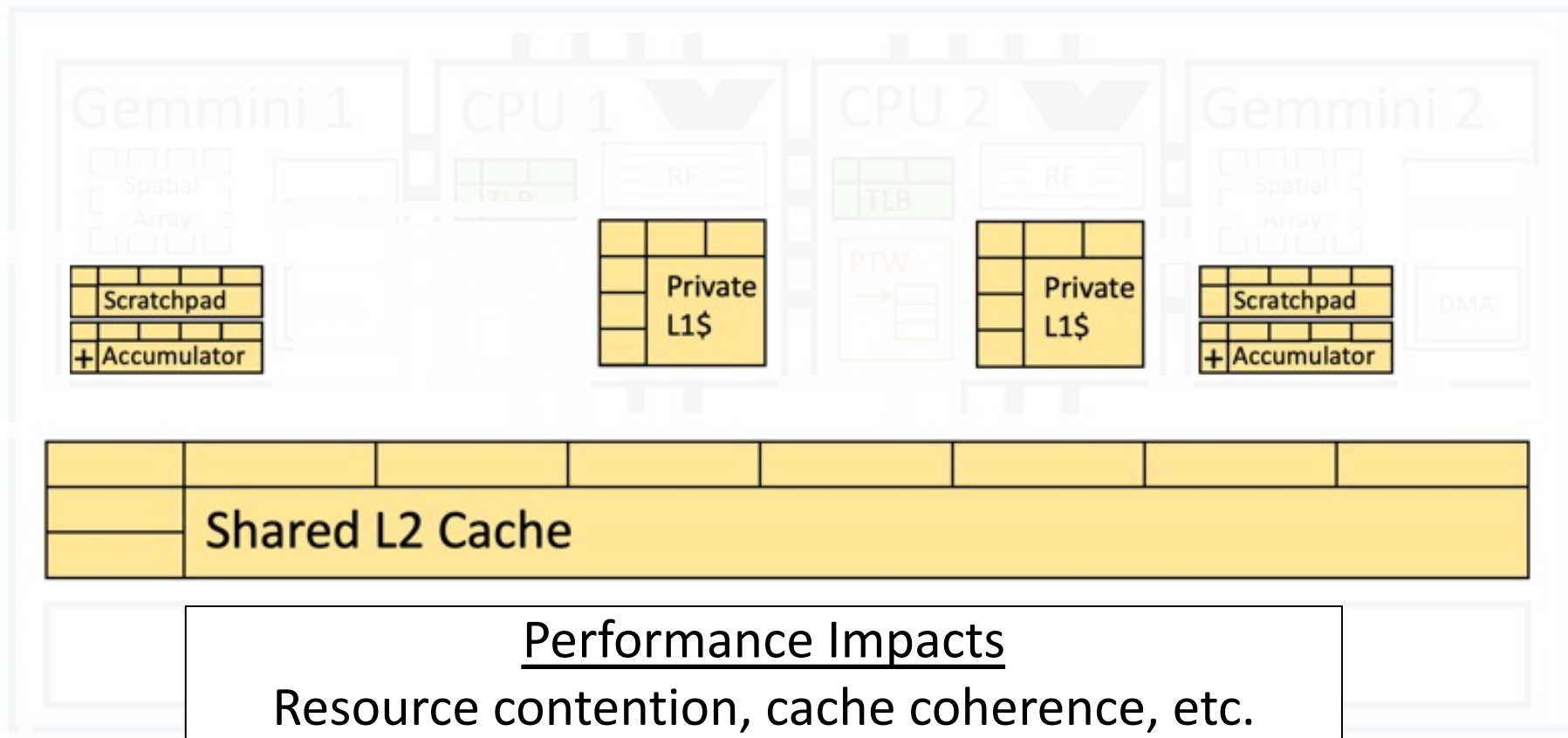
Two Billion Devices and Counting: An Industry Perspective on the State of Mobile Computer Architecture, IEEE Micro'2018

Full-System Visibility for DL Accelerators

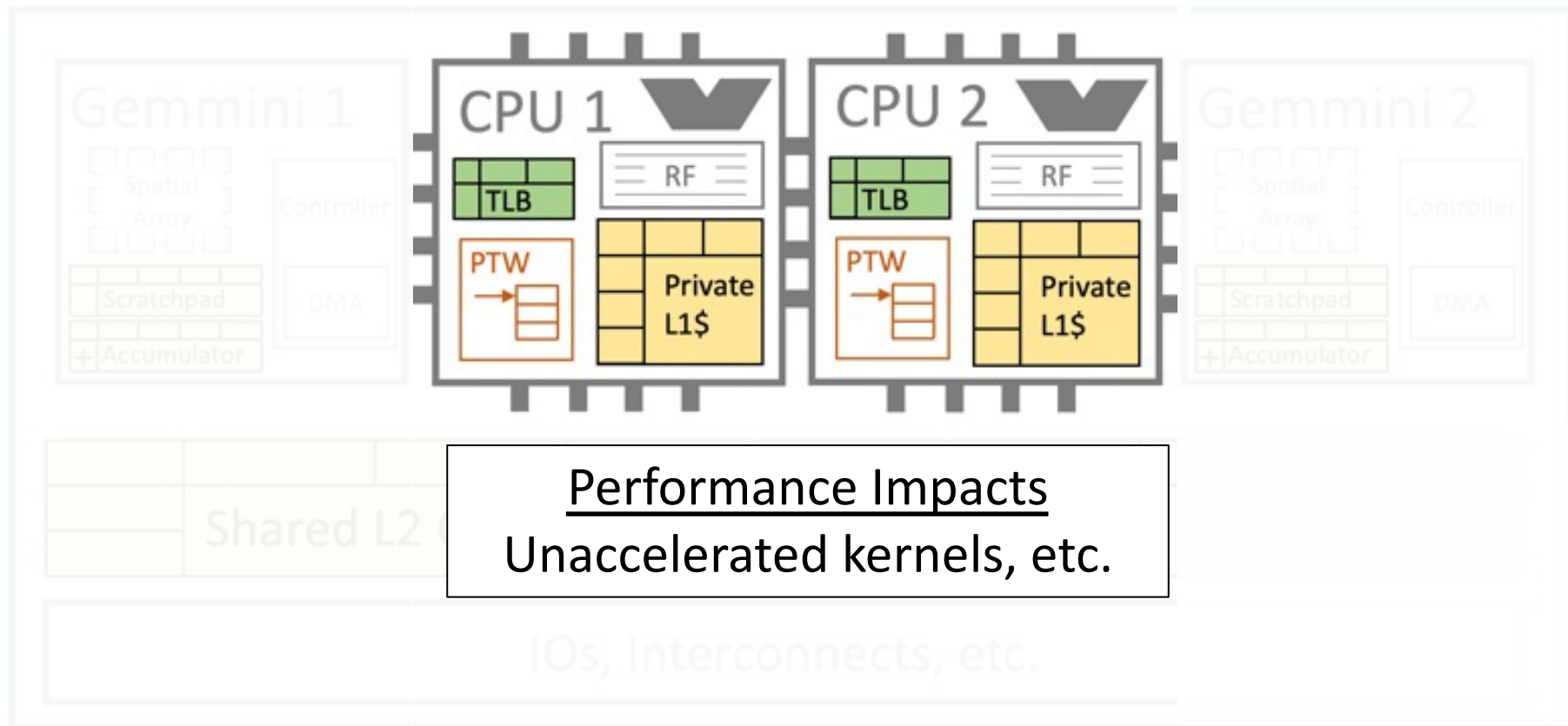
SoC



Full-System Visibility: Memory Hierarchy



Full-System Visibility: Host CPUs



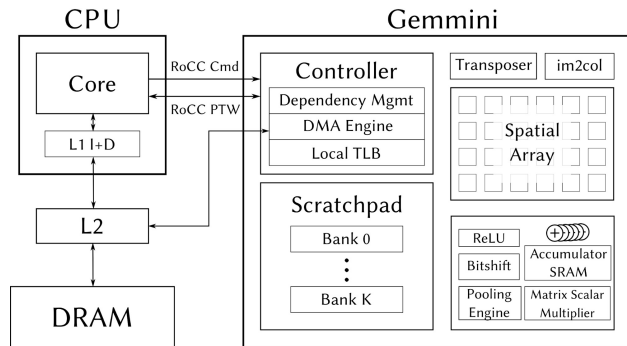
Gemmini: Full-System Co-Design of Hardware Accelerators

- **Full-stack**

- Includes OS
- End-to-end workloads
- “Multi-level” API

- **Full-SoC**

- Host CPUs
- Shared memory hierarchies
- Virtual address translation

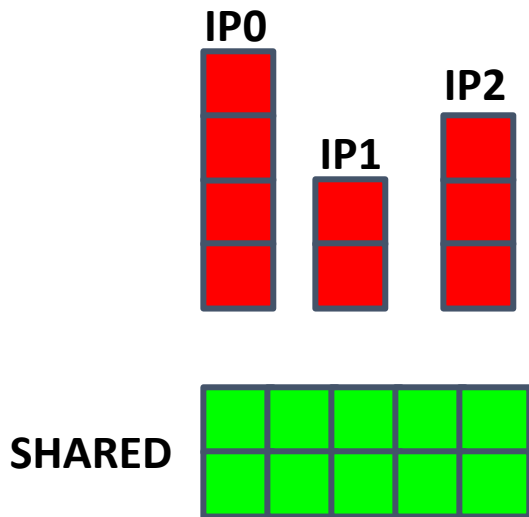


	Property	NVDLA	VTA	PolySA	DNNBuilder	MAGNet	DNNWeaver	MAERI	Gemmini
Hardware Architecture Template	Multiple Datatypes	Int/Float	Int	Int	Int	Int	Int	Int	Int/Float
	Multiple Dataflows	✗	✗	✓	✓	✓	✓	✓	✓
	Spatial Array	vector	vector	systolic	systolic	vector	vector	vector	vector/systolic
	Direct convolution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

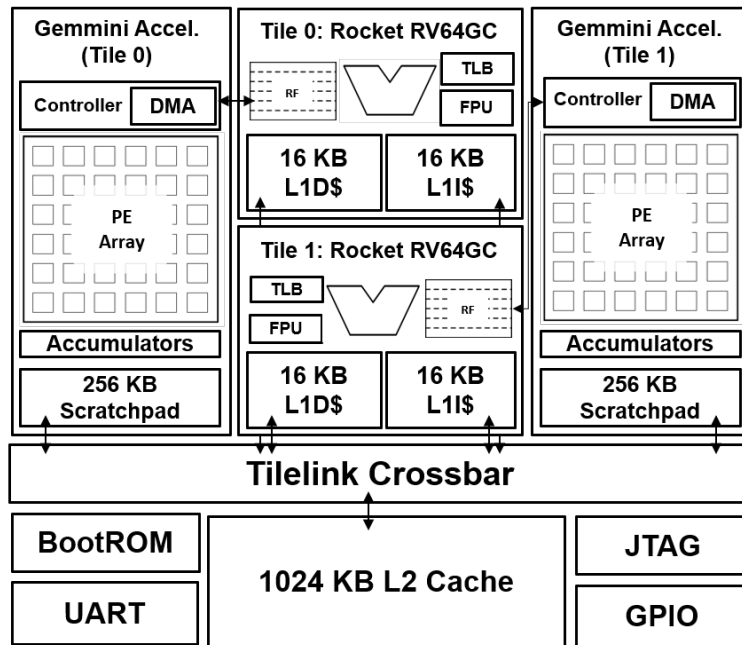
[DAC'2021 Best Paper Award]

Gemmini Case Study: Allocating on-chip SRAM



- **Where to allocated SRAM?**

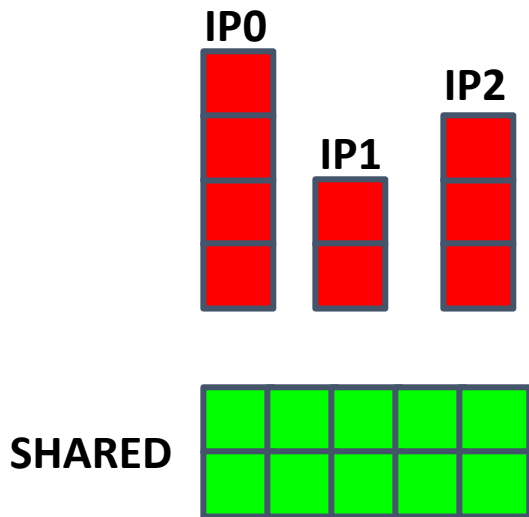
- Private within each IP
- Shared



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

[DAC'2021 Best Paper Award]

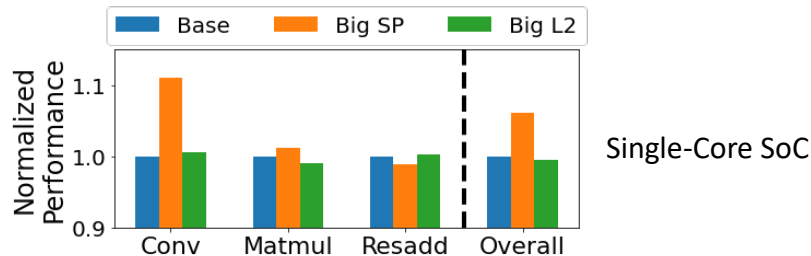
Gemmini Case Study: Allocating on-chip SRAM



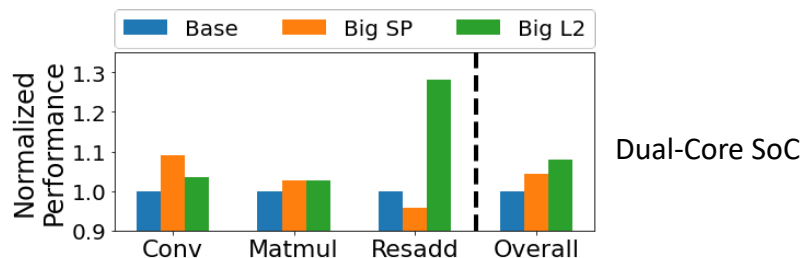
- **Where to allocated SRAM?**

- Private within each IP
- Shared

- **Application dependent.**



- **SoC configuration dependent.**



<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

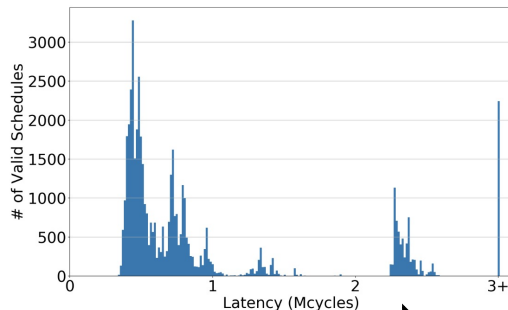
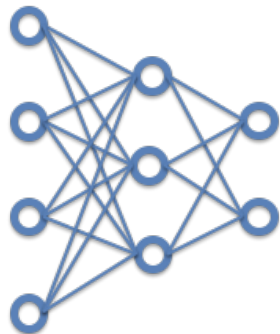
- Chipyard [IEEE Micro'2020]
- Gemmini [DAC'2021 **Best Paper Award**]

Scheduling of Accelerators

- CoSA [ISCA'2021]

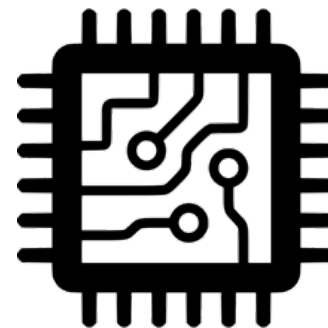
Large Space of Mapping Algorithms to ML Hardware

Algorithm



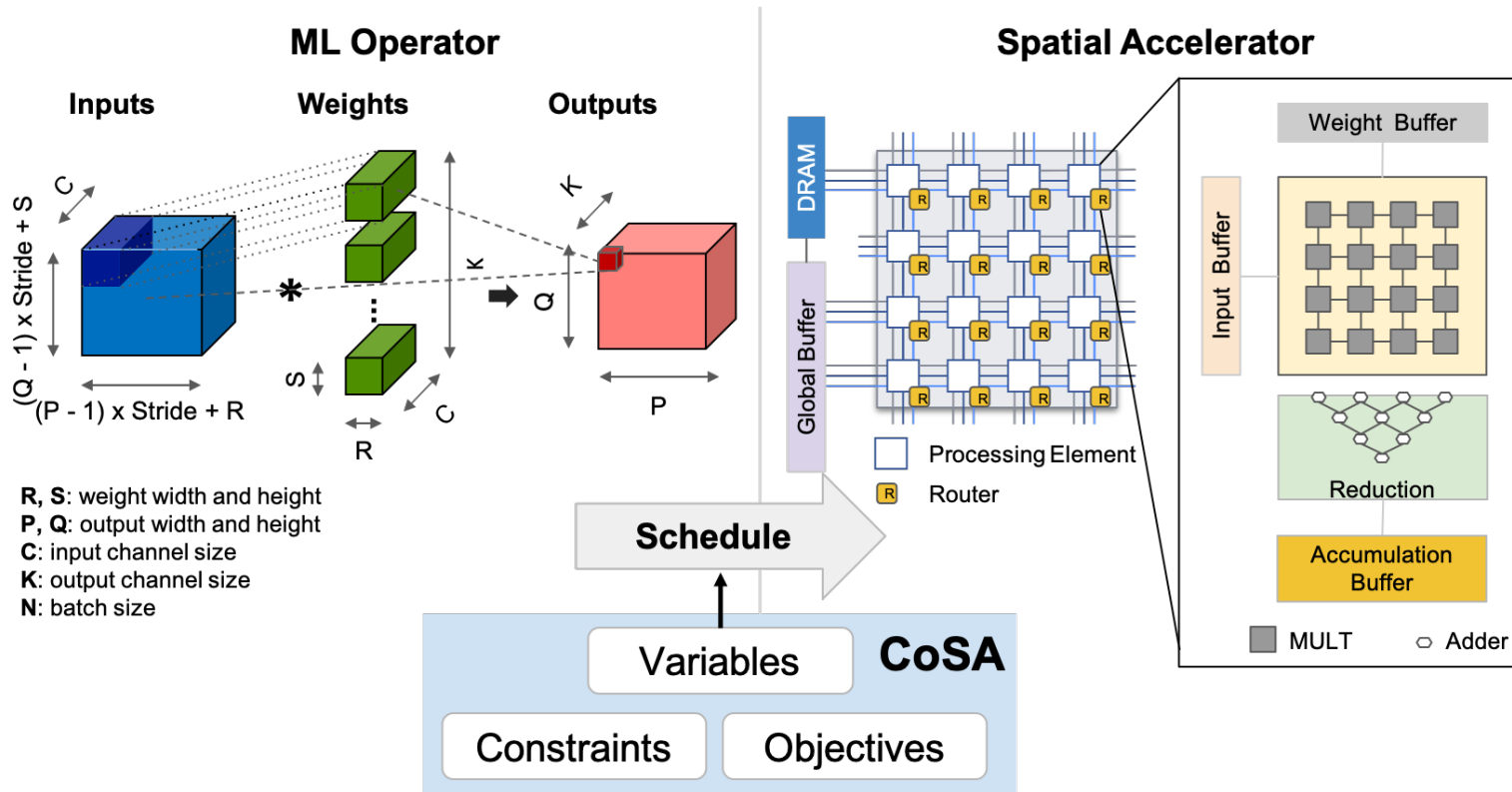
Scheduling

Hardware

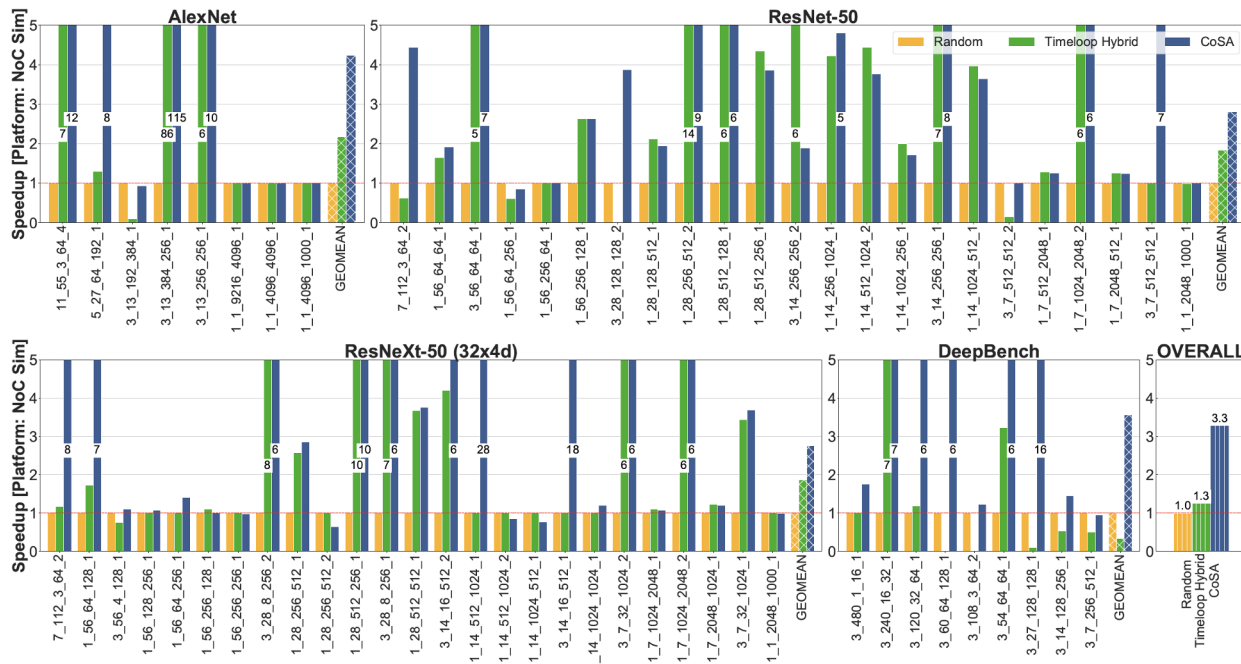


Scheduler	Search Algorithm
<i>Brute-force approaches:</i>	
Timeloop	Brute-force & Random
dMazeRunner	Brute-force
Interstellar	Brute-force
Marvel	Decoupled Brute-force
<i>Feedback-based Approaches:</i>	
AutoTVM	ML-based Iteration
Halide	Beamsearch OpenTuner
FlexFlow	MCMC
<i>Constrained Optimization Approaches: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

CoSA: Constrained-Optimization for Spatial Architecture



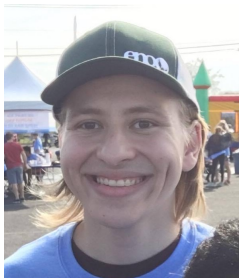
Results



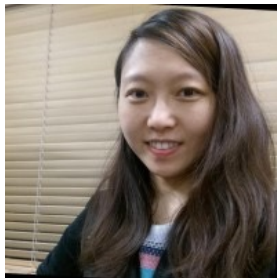
2.5x speedup
compared to SoTA
with 90x faster
time-to-solution.

	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

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:



FotaHub

Micro.ai



PROPHESÉE



seeed
The IoT Hardware Enabler



life.augmented



SynSense

XMOS

Silver:

AONdevices



ASPINITY

CEVA®



emza
visual sense

GREENWAVES
TECHNOLOGIES

Groovety Inc.



imagimob



itemis

LATTICE
SEMICONDUCTOR

nota



omniML



Plumerai



Rackner

REXEN
technology



STREAM ANALYZE



Tensil.ai



TensorFlow



Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® Summit 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at the tinyML Summit. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org