Differentiated Access Memories

Philip Levis and Caroline Trippel
First MemoryDAX/DAM Winter Workshop
1/10/2025

In One Slide

- Memory is increasingly the bottleneck in computing systems
  - ML accelerators: SRAM, HBM bandwidth and capacity
  - Cloud and datacenter servers: DDR bandwidth and cost
- Further gains in performance require transformative changes to memory
- Understanding *which* changes requires understanding many layers
  - Devices and circuits: what memories are possible and what are their tradeoffs?
  - Architecture: how do we organize memories and maintain coherence?
  - Software: what do we need memory to do?
- We've just started a 5-year project to explore and help define the future of memory: 6 months in

Our Thesis

- Over the past 20 years, processors have increasingly relied on specialization to improve performance and efficiency
- Over the next 20 years, memory will too
- Computing system memory will be heterogeneous
- ***Differentiated access memories***
  - Differ in read/write properties (write once, write many)
  - Differ in optimized access pattern
  - Differ in retention/lifetime of data
  - Differ in write endurance
  - Differ in density/capacity

Many Kinds of Memory...

DRAM
SRAM
MRAM
RRAM
FRAM
PCM
Flash
GC
HGC
FeFet
OS-OS

	Energy/power (active)		Energy/ power (standby)	Access time, latency		endurance	retention	Density (capacity)	On-logic chip integration	
	read	Write		read	Write				One layer (possible)	Multiple layers for density
High	RRAM, MRAM, PCM, FeRAM,	RRAM, MRAM, PCM, Flash	DRAM	Flash	Flash	DRAM, SRAM, OS-OS GC, HGC	Flash, RRAM, MRAM, PCM, FeFET, FeRAM	Flash, FeFET	MRAM, PCM, RRAM, FeRAM,	FeFET, OS-OS GC
Medium	DRAM	DRAM, FeRAM	SRAM	RRAM, PCM, FeFET, FeRAM	RRAM, PCM, FeFET, FeRAM	FeRAM, MRAM	OS-OS GC, HGC	DRAM, FeRAM, OS-OS GC	DRAM	
Medium low	FeFET, OS-OS GC	FeFET	HGC, OS-OS GC	DRAM, MRAM, OS-OS GC	DRAM, OS-OS GC, HGC	PCM, RRAM	DRAM	HGC, MRAM, RRAM, PCM,		
low	SRAM, HGC	SRAM, HGC, OS-OS GC	RRAM, MRAM, PCM, FeFET, FeRAM, Flash	SRAM, HGC	SRAM	Flash, FeFET		SRAM	Flash	Flash, DRAM

Results in the Past 6 Months

- Classifying memory: broad groups, defined by software use cases
  - Long term memory (LtRAM), short term memory (StRAM)
- Tools and hardware to guide which memories to use, when
  - Dynamic software in datacenter servers
  - Designing and provisioning ML accelerators
  - Gain cells and their design
- Integrating heterogeneous memory models and memories
  - Packaging, integration, and thermal concerns
  - Memory consistency and correctness
- Architectural memory primitives and verification

Five Groups

Five Groups

Structure

Benefits

Drawbacks

Uses

Five Groups

	SRAM
Structure	6T
Benefits	Fast Easy to integrate Low static power
Drawbacks	Sparse
Uses	Fast read/write caches

Five Groups

	SRAM	DRAM
Structure	6T	1T1C
Benefits	Fast Easy to integrate Low static power	Dense
Drawbacks	Sparse	No logic High power
Uses	Fast read/write caches	Large, random-access RW data

Five Groups

	SRAM	DRAM	Block Flash
Structure	6T	1T1C	1G
Benefits	Fast Easy to integrate Low static power	Dense	HUGE Capacity
Drawbacks	Sparse	No logic High power	No logic Low endurance Expensive, slow erases Block access Low bandwidth
Uses	Fast read/write caches	Large, random-access RW data	Large, read-mostly data

Five Groups

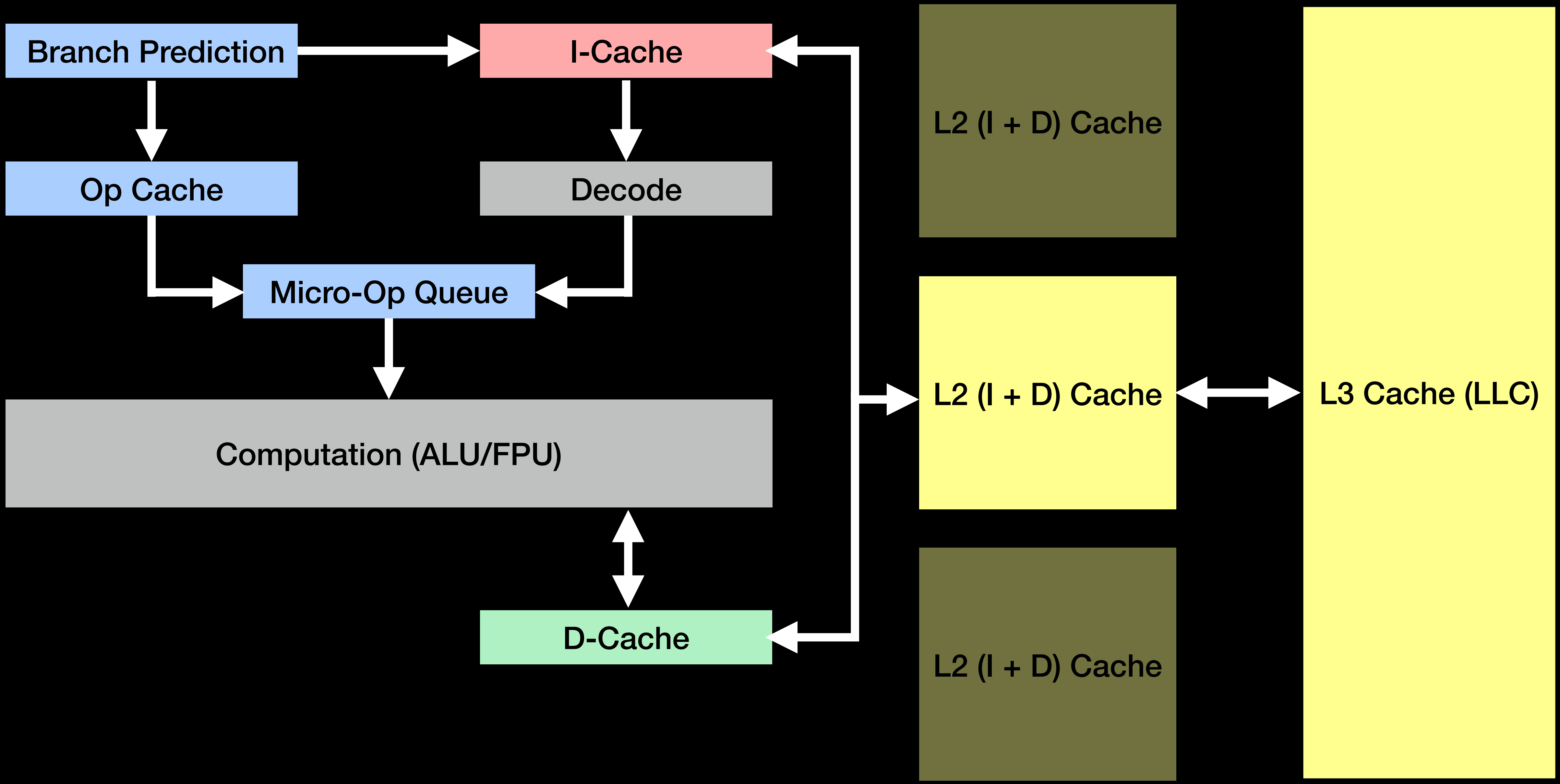
	SRAM	DRAM	Block Flash	LtRAM (long-term RAM)
Structure	6T	1T1C	1G	FeRAM, MRAM, RRAM, FRAM
Benefits	Fast Easy to integrate Low static power	Dense	HUGE Capacity	Dense Low Read Energy
Drawbacks	Sparse	No logic High power	No logic Low endurance Expensive, slow erases Block access Low bandwidth	Writes are slow and high energy Limited endurance
Uses	Fast read/write caches	Large, random-access RW data	Large, read-mostly data	Write rarely (static caches)

Five Groups

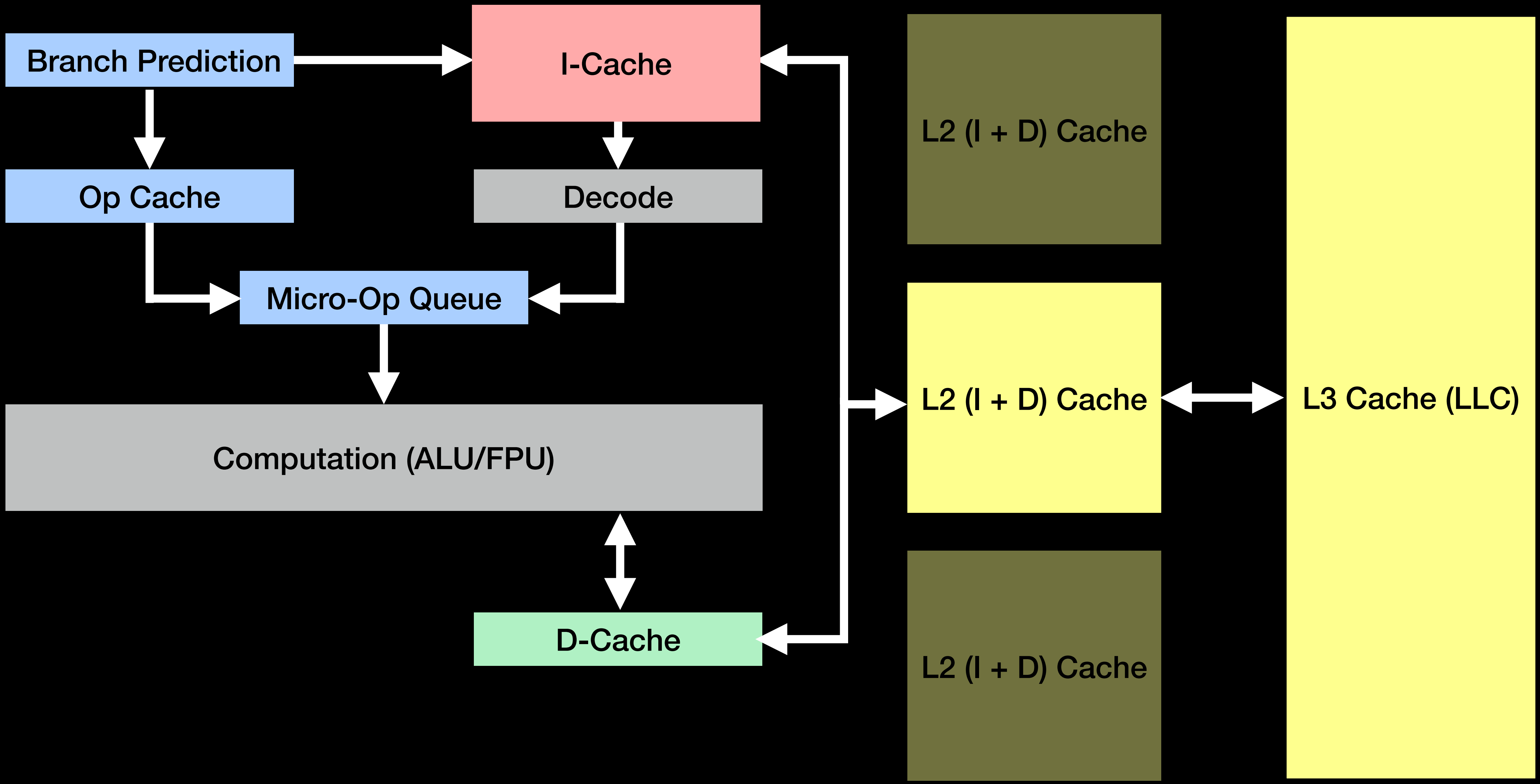
	SRAM	DRAM	Block Flash	LtRAM (long-term RAM)	StRAM (short-term RAM)
Structure	6T	1T1C	1G	FeRAM, MRAM, RRAM, FRAM	Gain Cells (2T, 3T)
Benefits	Fast Easy to integrate Low static power	Dense	HUGE Capacity	Dense Low Read Energy	Dense Low Energy
Drawbacks	Sparse	No logic High power	No logic Low endurance Expensive, slow erases Block access Low bandwidth	Writes are slow and high energy Limited endurance	Active research Refresh power
Uses	Fast read/write caches	Large, random-access RW data	Large, read-mostly data	Write rarely (static caches)	Write-and-read Write-and-read

Two Example Uses of StRAM and LtRAM: Datacenter Servers, ML Accelerators

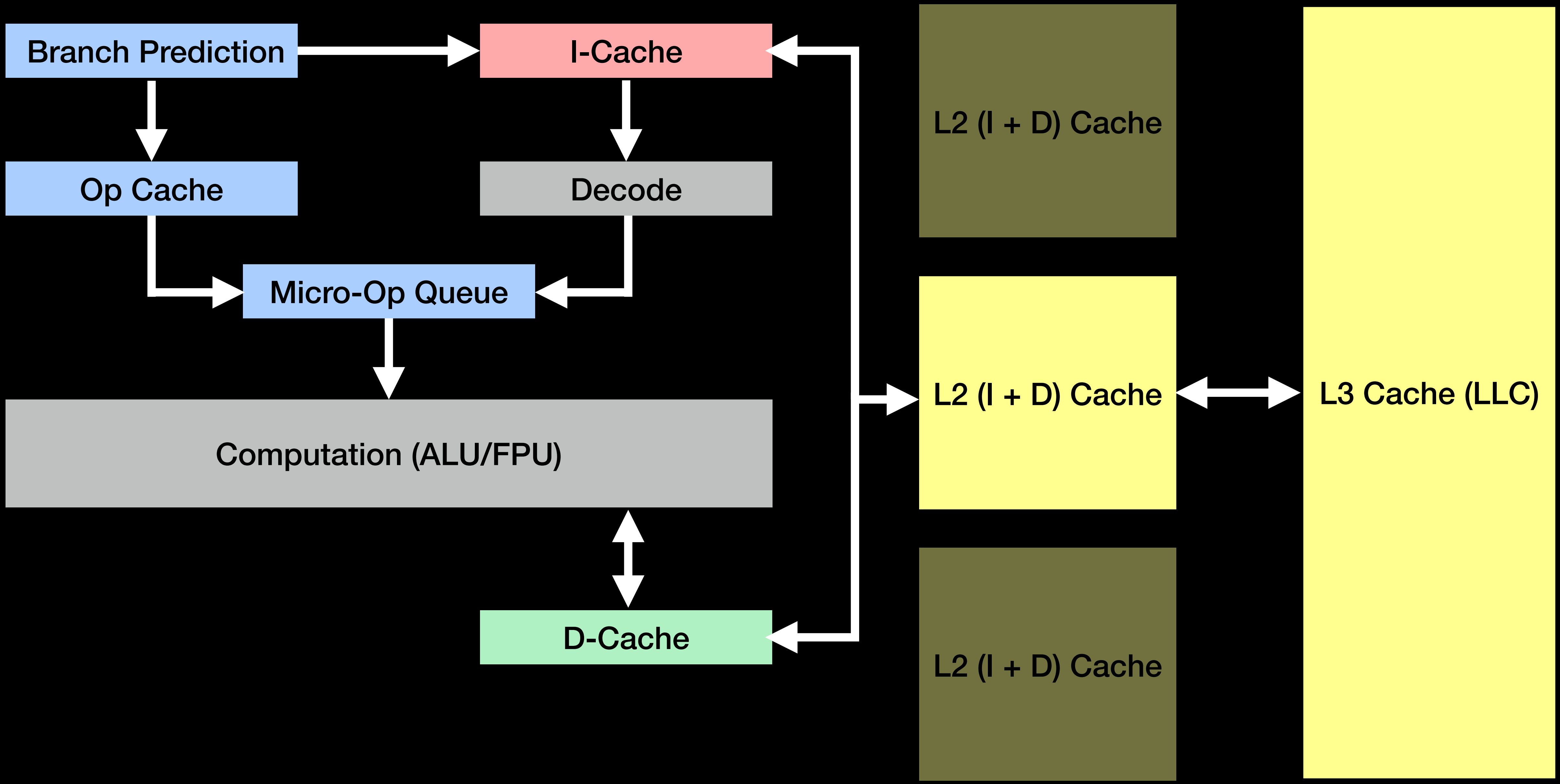
In Servers (x86)



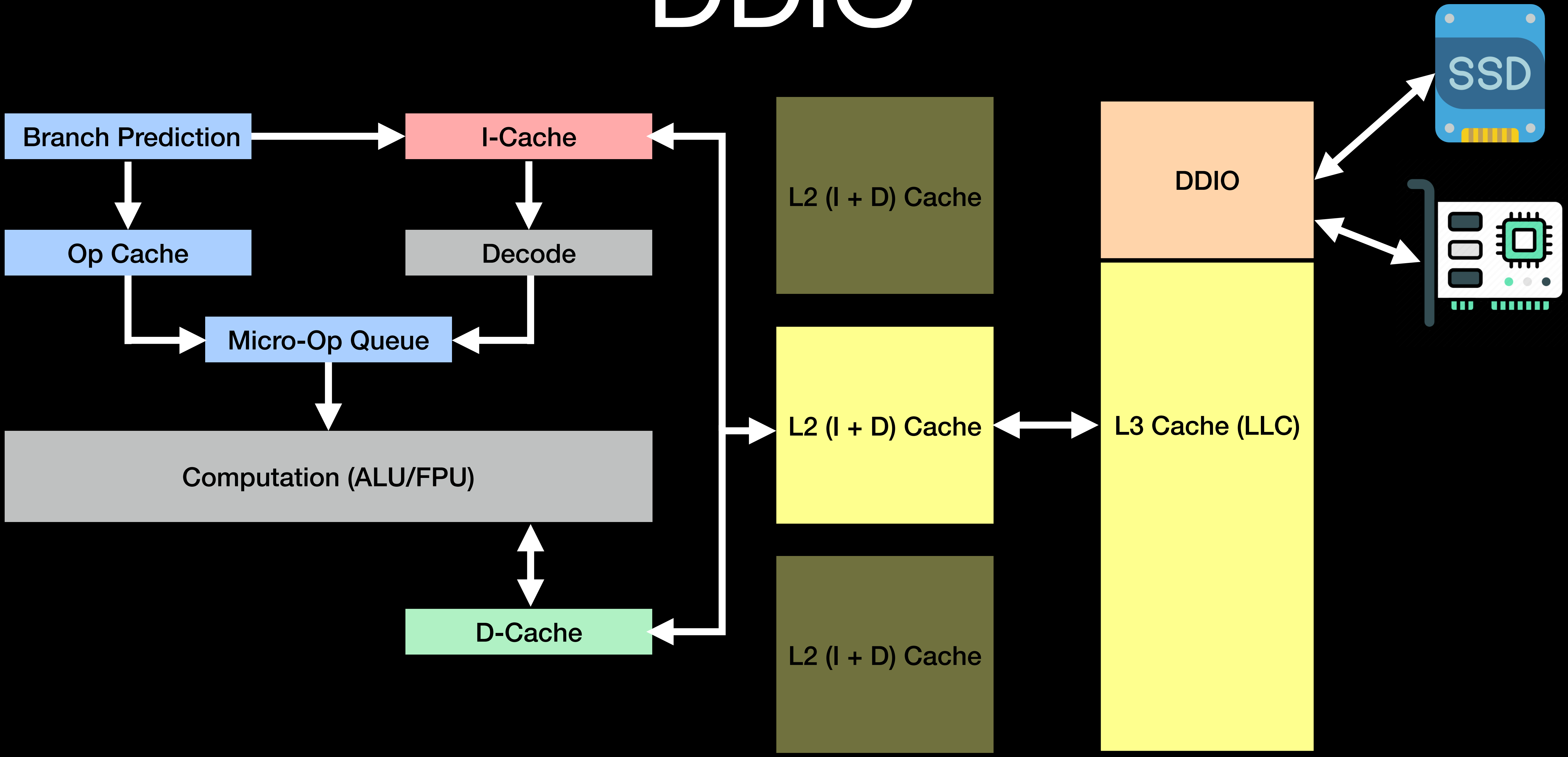
I-Caches (StRAM)



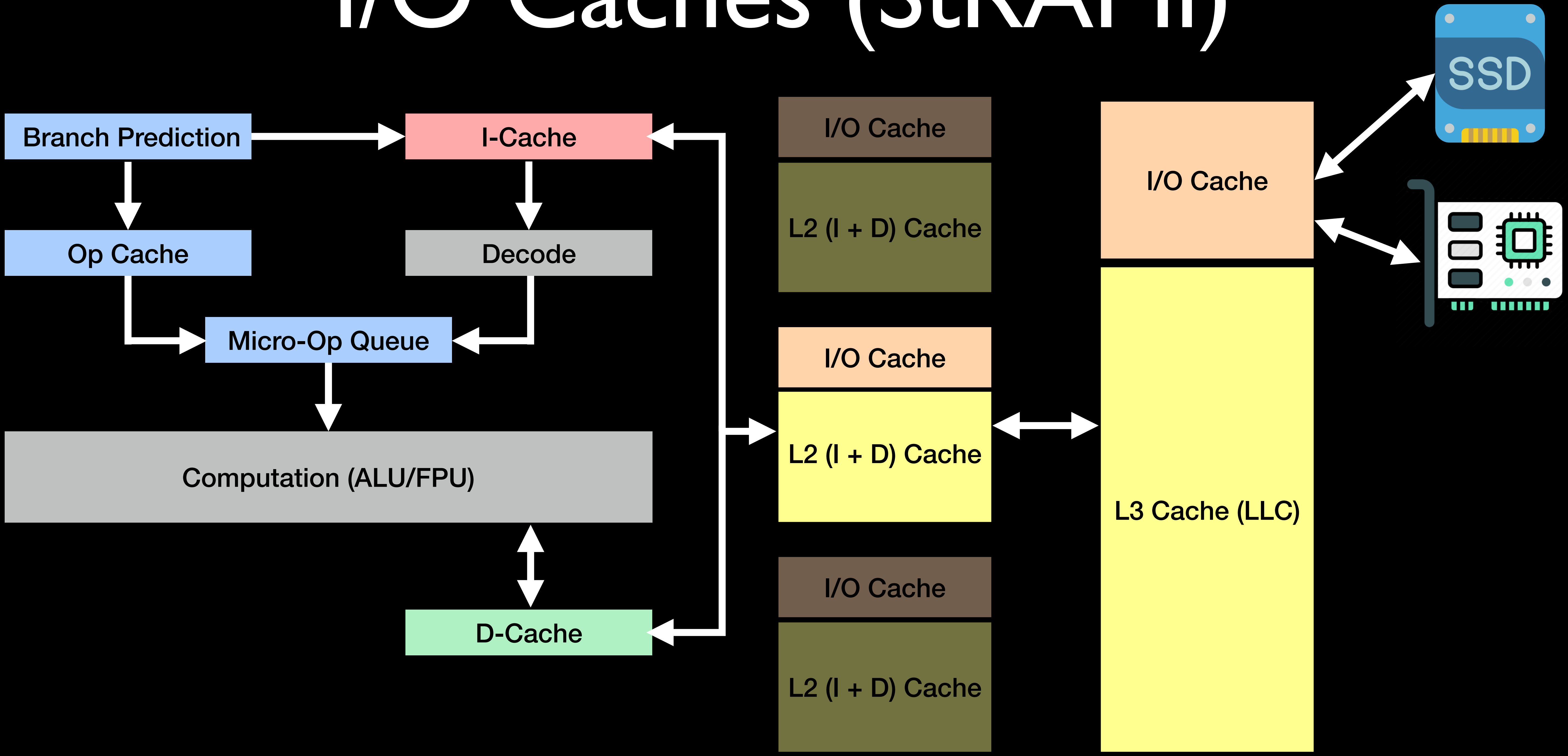
In Servers (x86)



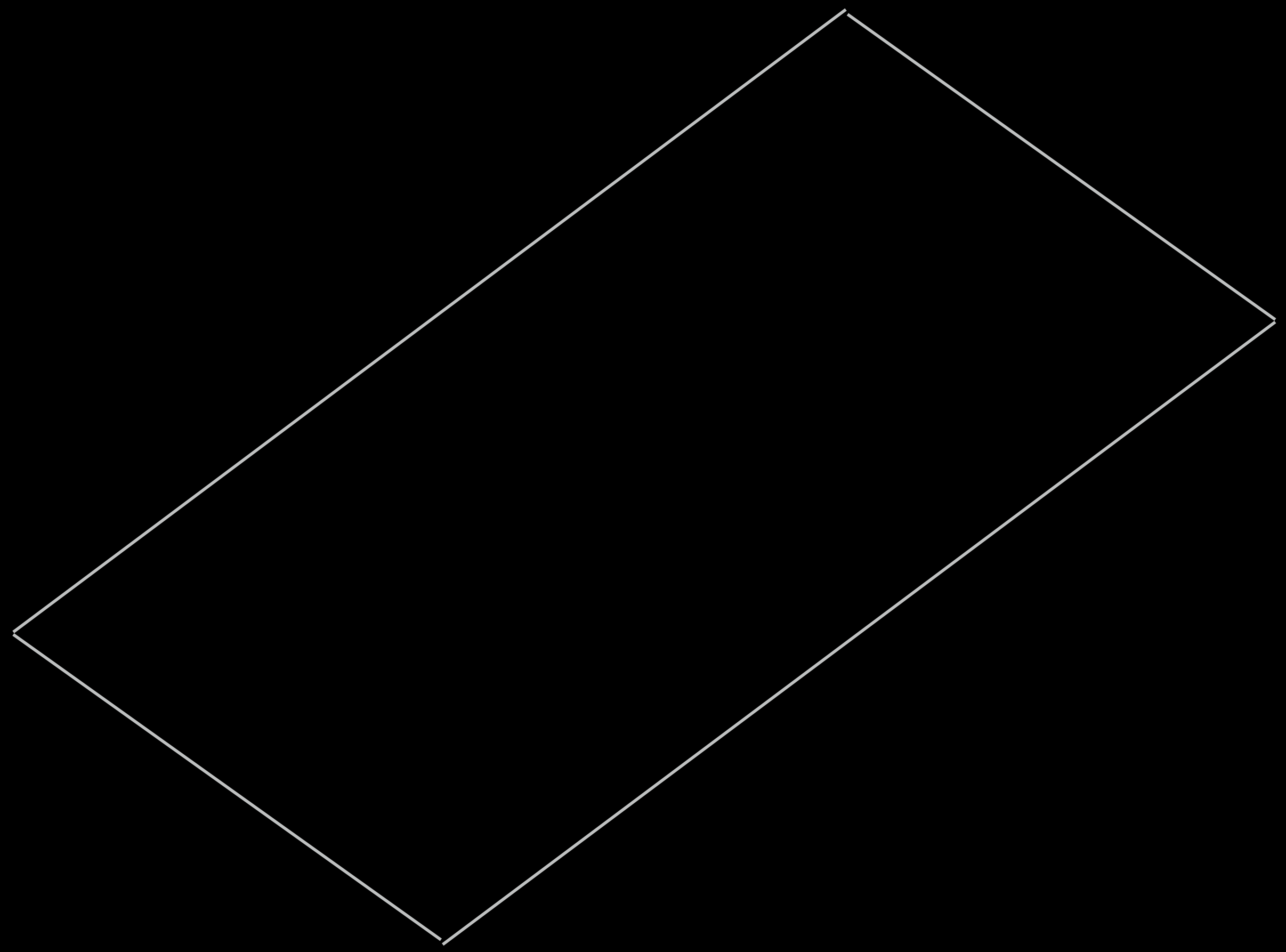
DDIO


Branch Prediction
I-Cache
Op Cache
Decode
Micro-Op Queue
Computation (ALU/FPU)
D-Cache
L2 (I + D) Cache
L2 (I + D) Cache
L2 (I + D) Cache
DDIO
L3 Cache (LLC)
SSD

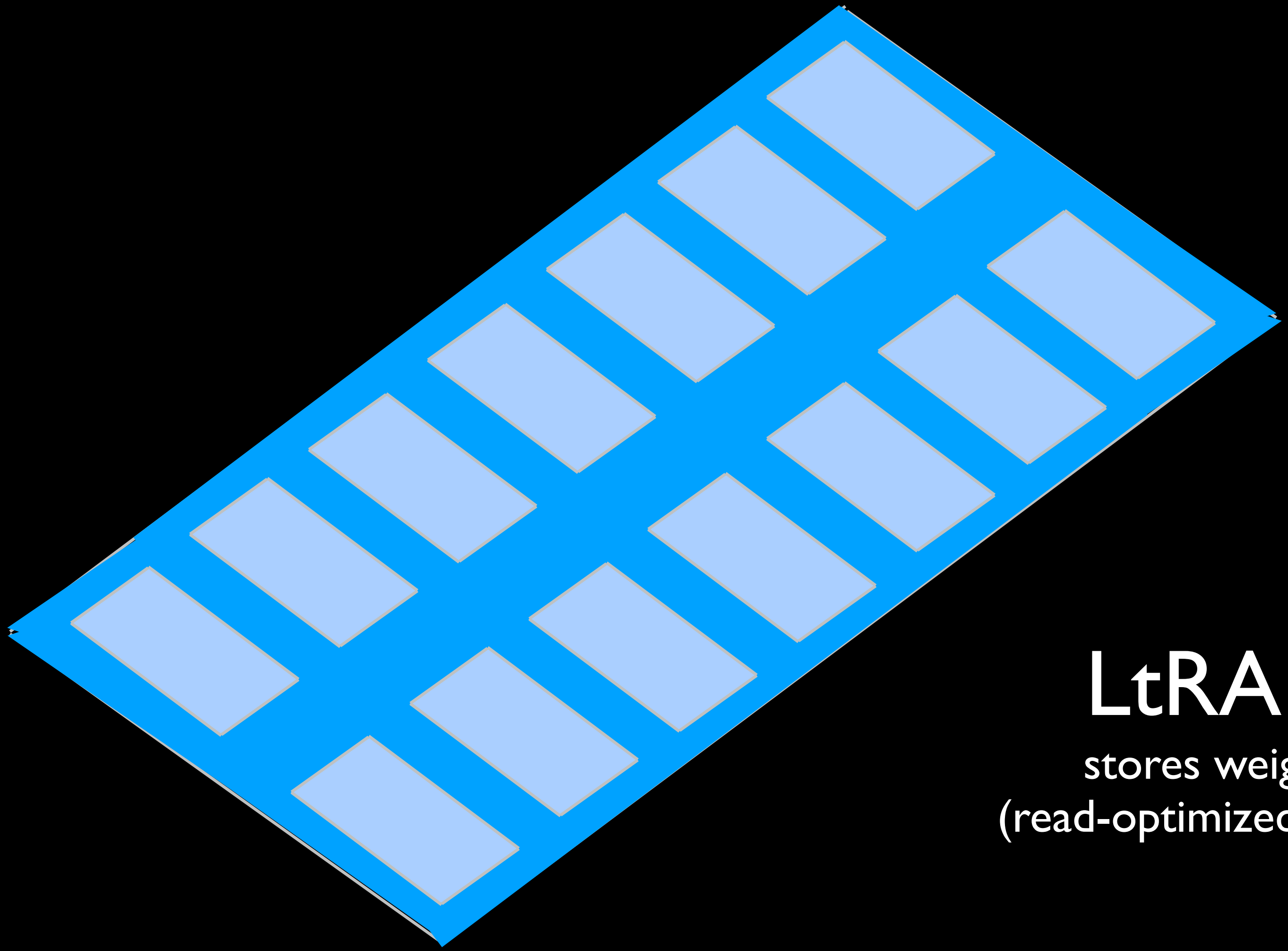
I/O Caches (StRAMI)



Inference Accelerator



Inference Accelerator

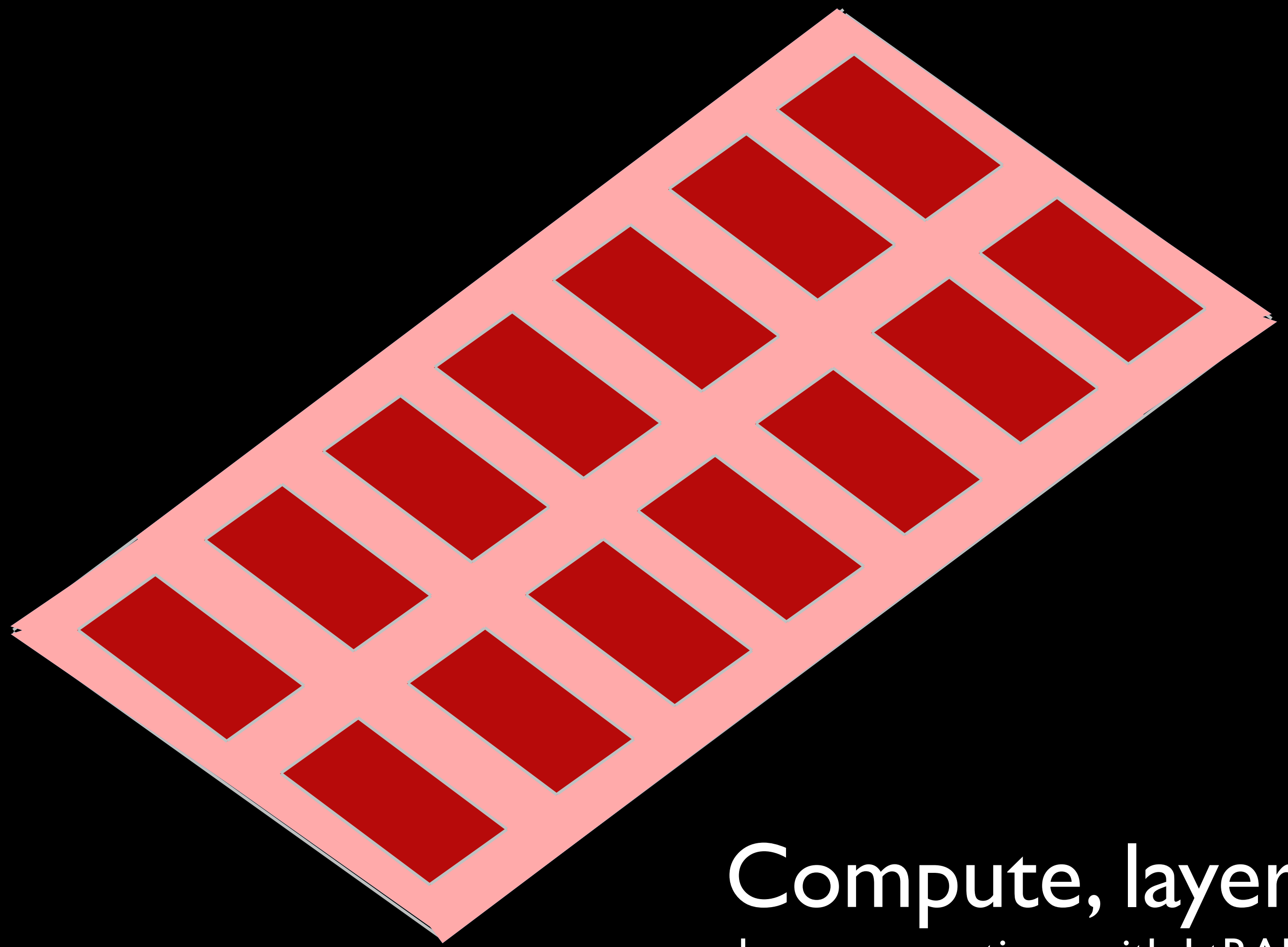


LtRAM

stores weights

(read-optimized, dense)

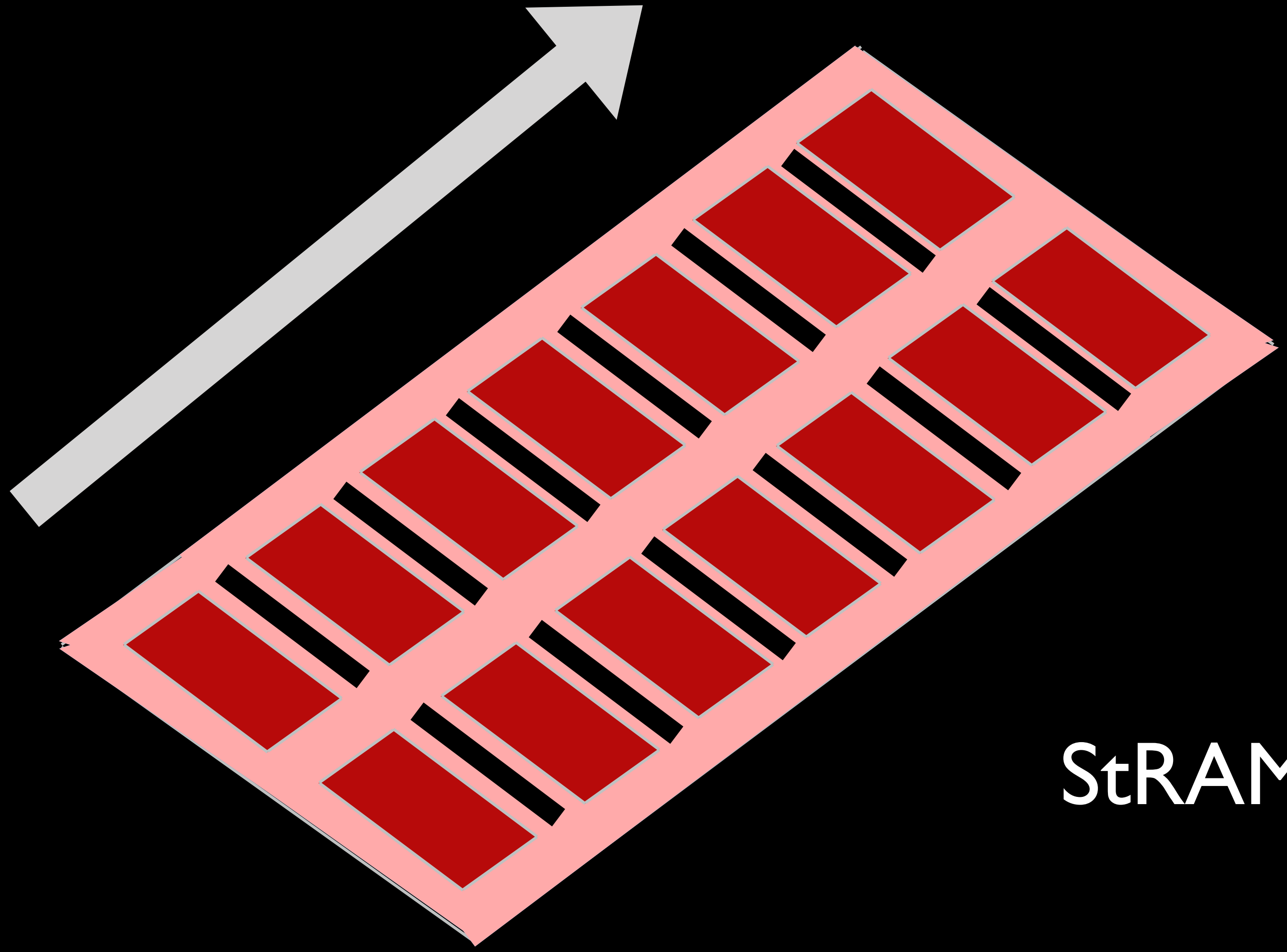
Inference Accelerator



Compute, layered on top

dense connections with LtRAM below

Inference Accelerator



StRAM for Activations

Faculty Team

Philip Levis

Caroline Trippel

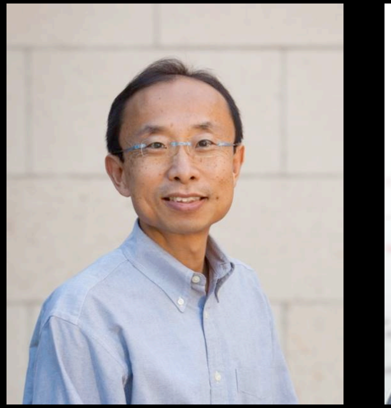
Chris Gregg

Mark Horowitz

Subhasish Mitra

Thierry Tambe

Keith Winstein



H.-S. Philip Wong

Mary Wootters

Today

9:00	Welcome and Project Overview
	Massive, Diverse, Tightly Integrated with Compute – from Device to Software
	Memory Access Pattern Classification
	Data Lifetime and Its Refresh Implications
10:45	Break
11:00	MemGlue for Heterogeneous Architectures
	Conserving Memory Bandwidth with Virtual Gather
12:00	Walk and Lunch
1:30	Integration: Performance, Power, and Thermal Constraints
	Gain Cell Compiler
	Synthesizing High Level Models from RTL
3:00	Break
3:15	Panel: Memory Has to Change
4:15	Differentiated Access Memories: What's Next
4:45	Closing