Memory & Caches I

CSE 351 Autumn 2023

Guest Instructor:

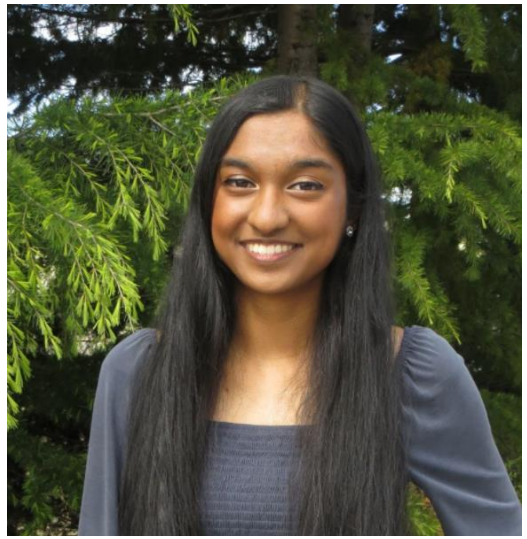
Nayha Auradkar

Teaching Assistants:

Afifah Kashif
Bhavik Soni
Cassandra Lam
Connie Chen
David Dai
Dawit Hailu
Ellis Haker
Eyoel Gebre
Joshua Tan
Malak Zaki
Naama Amiel
Nayha Auradkar
Nikolas McNamee
Pedro Amarante
Renee Ruan
Simran Bagaria
Will Robertson

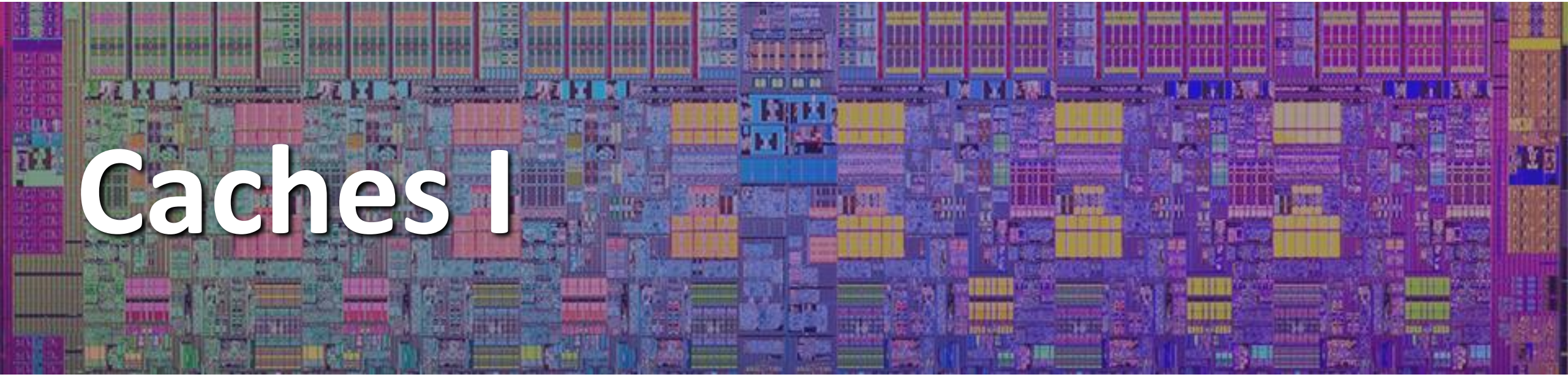
Introduction

- Guest Lecturer: Nayha Auradkar
  - CSE 351 TA
  - 5th year Master's student
  - Research focus on AI/ML and accessibility



Relevant Course Information

- Midterm starts tomorrow (11/2-11/4)
  - Only private posts on Ed Discussion
  - Staff will post clarifications and corrections as we go
- hw15 due Monday (11/6), hw16 due Wednesday (11/8)
- Lab 3 due next Friday (11/10)
  - Make sure to look at HW15 before starting
- Veteran's Day next Friday (11/10); no lecture



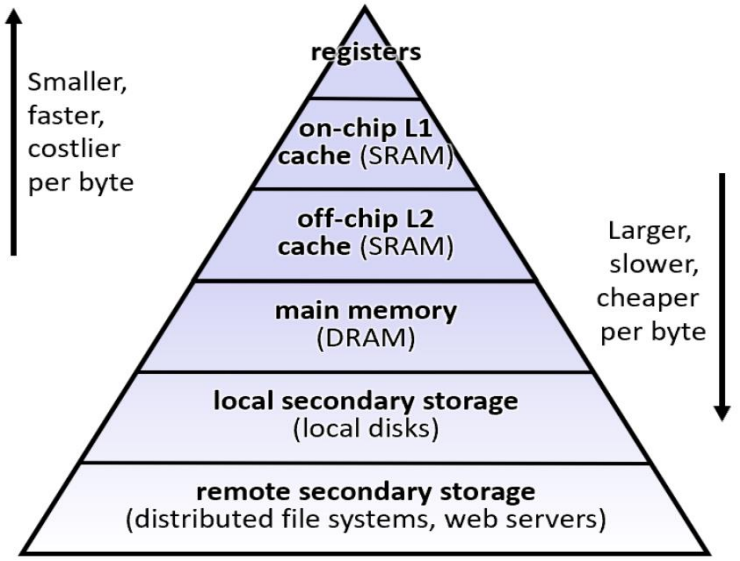
Caches I

Lesson Summary (1/2)

- Caches are intermediate storage levels used to optimize data transfers between any system elements with different characteristics
  - Exploits *temporal and spatial locality*

- Memory Hierarchy
  - Successively higher levels contain "most used" data from lower levels

- Cache Performance
  - Ideal case: found in cache (hit)
  - Bad case: not found in cache (miss), search in next level
  - Average Memory Access Time (AMAT) = HT + MR × MP
    - Hurt by Miss Rate and Miss Penalty

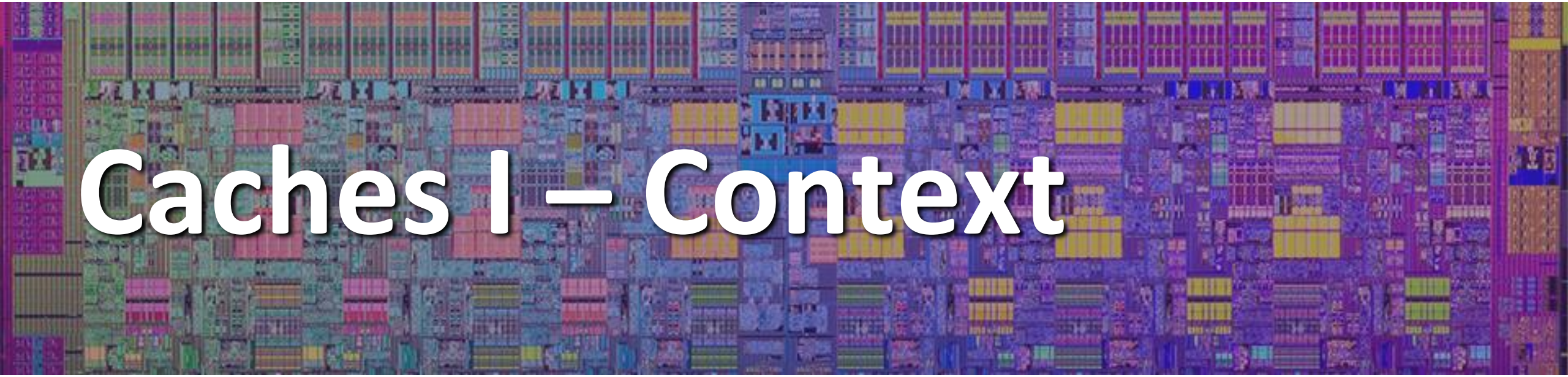


Lesson Summary (2/2)

- Terminology:
  - Caches: cache blocks, cache hit, cache miss
  - Principle of locality: temporal and spatial
  - Average memory access time (AMAT): hit time, miss penalty, hit rate, miss rate

- Learning Objectives:
  - Describe the memory hierarchy and explain the relationship between cost, size, and access speed of its layers.
  - Analyze how changes to [cache parameters and policies] affect performance metrics such as AMAT

- What lingering questions do you have from the lesson?



Caches I – Context

AMAT, Revisited

- *Average Memory Access Time* (AMAT): average time to access memory considering both hits and misses

  AMAT = Hit time + Miss rate × Miss penalty

  (abbreviated AMAT = HT + MR × MP)

- We called this a *cache performance* ***metric***
  - This isn't the only metric we could have used!

Metrics in Computing

- Generally, folks care most about performance
  - Energy-efficiency is more important now since the plateau in 2004/2005
  - This is why we have so many specialized chips nowadays

- Really, this is just **efficiency** – making efficient use of the resources that we have
  - Performance: cycles/instruction, seconds/program
  - Energy efficiency: performance/watt
  - Memory usage efficiency: bytes/program, bytes/data structure
  - Algorithm efficiency: Big-O Complexity analysis

Metrics

- What do we do with metrics?
  - We tend to optimize along them!
  - Especially when jobs/funding depend on better performance along some metric
    - See all of Intel under "Moore's Law"
- Sometimes, strange incentives emerge
  - "Minimize the number of bugs on our dashboard"
    - Does it count if we make the bugs invisible?
  - "Make this faster for our demo in a week"
    - Shortcuts might hurt performance at scale
  - "Minimize our average memory access time"
    - What if we add *more* memory accesses that we know will hit?

Metrics and Success

- Success is *defined along metrics*
  - This affects how we measure and optimize

- Let's say that we choose **performance/program** or **performance/program set** (*i.e.*, benchmarks) as our metric:
  1. Define what success means using this metric
  2. Measure existing performance
  3. Come up with optimizations that would improve performance
  4. Select some to build into the "next version"

Metrics and Success

- Success is *defined along metrics*
  - This affects how we measure and optimize

- Let's say that we choose **profit/year** or **stock price**:
  - Success means earning more profit than last year
  - Improvement or optimizations might include:
    - Reduce expenses, cut staff
    - Sell more things or fancier things (*e.g.*, in-app purchases)
    - Make people pay monthly for things they could get for free
    - Increase advertising revenue:

The New York Times

Whistle-Blower Says Facebook 'Chooses Profits Over Safety'

Frances Haugen, a Facebook product manager who left the company in May, revealed that she had provided internal documents to journalists and others.

Discussion Questions

- Discuss the following question(s) in groups of 3-4 students
  - I will call on a few groups afterwards so please be prepared to share out
  - Be respectful of others' opinions and experiences

- Suppose our metric is participation of minoritized folks in undergraduate computing education.
  - What does success mean?
  - How can we improve or optimize for this metric based on how we define success?

Design Considerations

- Regardless of what we build, the way that we define our metrics and success shapes the systems we build
  - Choose your metrics carefully
  - There's more to choose from than performance (*e.g.,* usability, access, simplicity, agency)

- Metrics are a "heading" (in the navigational sense)
  - Best to reevaluate from time to time in case you're off course or your destination changes

Caches I – Practice

Group Work Time

- During this time, you are encouraged to work on the following:
  1) If desired, continue your discussion
  2) Work on the lesson problems (solutions at the end of class)
  3) Work on the homework problems

- Resources:
  - You can revisit the lesson material
  - Work together in groups and help each other out
  - Course staff will circle around to provide support

Practice Questions (1/2)

SIZE PREFIXES (10^x for Disk, Communication; 2^x for Memory)

SI Size	Prefix	Symbol	IEC Size	Prefix	Symbol
10^3	Kilo-	K	2^{10}	Kibi-	Ki
10^6	Mega-	M	2^{20}	Mebi-	Mi
10^9	Giga-	G	2^{30}	Gibi-	Gi
10^{12}	Tera-	T	2^{40}	Tebi-	Ti
10^{15}	Peta-	P	2^{50}	Pebi-	Pi
10^{18}	Exa-	E	2^{60}	Exbi-	Ei
10^{21}	Zetta-	Z	2^{70}	Zebi-	Zi
10^{24}	Yotta-	Y	2^{80}	Yobi-	Yi

- Convert the following to or from IEC:
  - 512 Ki-books
  - 2^{27} caches
- Compute the average memory access time (AMAT) for the following system properties:
  - Hit time of 1 ns
  - Miss rate of 1%
  - Miss penalty of 100 ns

Practice Questions (2/2)

- **Processor specs:** 200 ps clock, MP of 50 clock cycles, MR of 0.02 misses/instruction, and HT of 1 clock cycle

  AMAT =

- Which improvement would be best?

  A. 190 ps clock

  B. Miss penalty of 40 clock cycles

  C. MR of 0.015 misses/instruction