UC San Diego

CSE 291: Operating Systems in Datacenters

Amy Ousterhout

Nov. 17, 2022

UC San Diego

Agenda for Today

- Project presentations
- Memory
- Llama discussion

Project Presentations

Project Presentations: Logistics

- During class on 11/29 and 12/1
- Talk duration depends on group size
  - For 1/2/3 students you will have 12/15/17 minutes
- 3 minutes of questions after your presentation
  - Ask questions of your peers!
- Use slides
- Will post on Canvas:
  - Details about project presentations
  - Sign up slots

Project Presentations: Content

- The problem and motivation
  - Your research problem or question and why it’s important
- Background
  - Information needed to understand the rest of your presentation
- Solution
  - How did you solve your problem or try to?
  - What worked well and what didn’t?
- Evaluation
  - Experimental setup
  - Results and implications
- Future work
  - What steps would you take next to continue this work?

Tips for Giving a Good Talk

- Consider your audience. What do they know or not know?
  - In this case audience == your peers
- Motivate the problem. Why should your audience care?
- Explain **why**, not just **what**
  - What: LegoOS has an ExCache on each pComponent and the rest of the memory is on the remote mComponent
  - Why: Because of the high latency to access remote memory in the mComponent, LegoOS adds an extra cache, the ExCache, to pComponents
- For experiments, tell us what question you're trying to answer
  - For example: how does disaggregation impact the performance of applications?
- Practice!

Tips for Designing Good Slides

- Give your slides meaningful titles
  - “Background” vs. “Hardware Support for Disaggregation”
- Use diagrams and graphs to illustrate your ideas
  - Hint: you can re-use these for your write-up
- Use text sparingly

Memory

Memory in Datacenters

- Storage technology is not improving significantly
  - Capacity has increased (16667x from 1980s -> 2009)
  - Transfer rate has increased less (50x from 1980s -> 2009)
  - Latency has improved even less (2x from 1980s -> 2009)
- Rise of data-intensive applications
  - Machine learning
  - Analytics
  - Complex web applications
- Motivates storing more data in memory
  - Low latency, high-bandwidth access

The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM [SOSP ‘09]

Research Challenges

- How to avoid overloading the TLB?
  - Huge pages!
- But huge pages raise other challenges
  - Fragmentation
  - Learning-based Memory Allocation for C++ Server Workloads [ASPLOS ‘20]
- How to reduce TLB overheads?
  - Don't shoot down TLB shootdowns! [EuroSys ‘20]

UC San Diego

Llama Discussion