18-742: Research in
Parallel Computer Architecture
Intro and Logistics

Prof. Onur Mutlu

Carnegie Mellon University

Fall 2014

August 26, 2014

Agenda

- Syllabus
    - Course logistics, info, requirements

- Introduction

- Background Videos and Lectures to Study

Further Agenda

- Grading and Policies

- Details on Course Project

- Details on Paper Reviews

- Initial Assignments and Homeworks

Course Info: Who Are We?

- Onur Mutlu
    - Associate Professor @ Carnegie Mellon University ECE/CS
    - PhD from UT-Austin 2006, worked at Microsoft Research, Intel, AMD
    - http://www.ece.cmu.edu/~omutlu
    - omutlu@gmail.com (Best way to reach me)
    - http://users.ece.cmu.edu/~omutlu/projects.htm

- Research, Teaching, Consulting in
    - Computer architecture and systems, hardware/software interaction
    - Memory and storage systems, emerging technologies
    - Many-core systems, heterogeneous systems, core design
    - Interconnects

Interested in developing efficient, high-performance, and scalable systems; solving difficult architectural problems at low cost & complexity

    - …

Course Info: Who Are We?

- Teaching Assistant

  - Yixin Luo
    - camellyx@gmail.com

Where to Get Up-to-Date Course Info?

- Course Website: http://www.ece.cmu.edu/~ece742

- Your email

- Email to us
  - Please email both myself and Yixin

Course Location

- Most of the time here: CIC 4th Floor Panther Hollow Room

- Not always…
- Check your email

Related Videos and Course Materials

- **Computer Architecture Lecture Videos on Youtube**
  - https://www.youtube.com/playlist?list=PL5PHm2jkkXmidJOd59R Eog9jDnPDTG6IJ

- **Computer Architecture Course Materials**
  - http://www.ece.cmu.edu/~ece447/s13/doku.php?id=schedule

- **Advanced Computer Architecture Course Materials**
  - http://www.ece.cmu.edu/~ece740/f13/doku.php?id=schedule

- **Advanced Computer Architecture Lecture Videos on Youtube**
  - https://www.youtube.com/playlist?list=PL5PHm2jkkXmgDN1PLw OY_tGtUlynnyV6D

Background Required

- You MUST have taken at least one of:
    - 18-447
    - 18-740

- 18-447
    - http://www.ece.cmu.edu/~ece447/s13/doku.php?id=schedule

- 18-740
    - http://www.ece.cmu.edu/~ece740/f13/doku.php?id=schedule

- You MUST have the interest and enthusiasm to do or at least explore research

Course Overview and Components

- Focus is on research

- The course is designed to provide a strong experience in research in computer architecture via a rigorous, self-driven semester-long research project, critical research article reviews, in-class research article presentations and discussions, a literature survey and other assignments that focus on developing critical thinking, idea development, and clear explanation and presentation skills.

- Since the main focus is on research, the students are expected to be highly motivated and self driven in all aspects of this course.
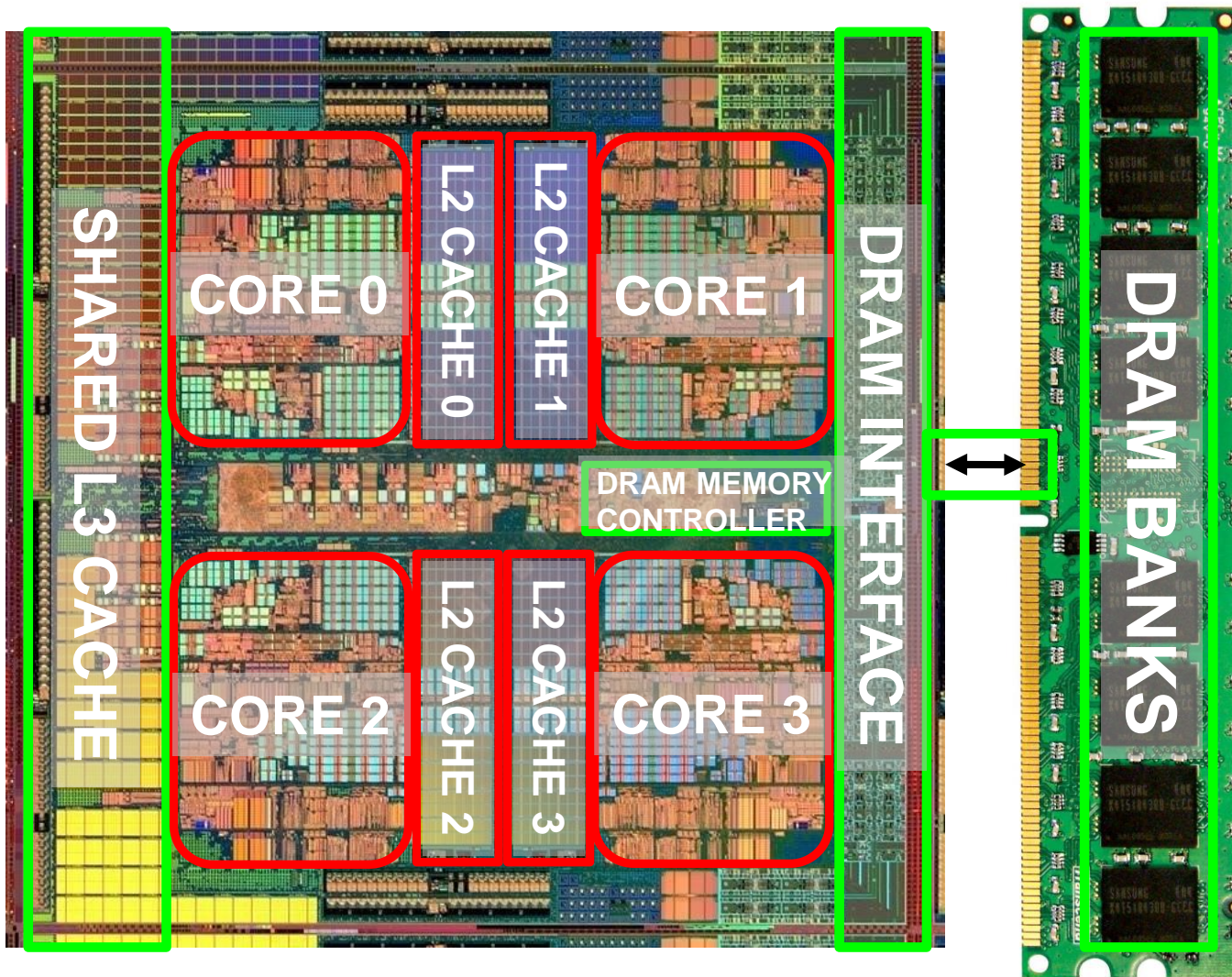
Major Course Component 1

- ***1. Critical analysis, review and in-class presentation and discussion of research papers and research seminars in computer architecture:*** Students will be expected to read and present recent as well as established research papers, critically review and analyze them in writing, and discuss them in class. Most of the class time will be dedicated to such presentations and discussions. I expect to assign at least two presentations per week. There will be minimal number of lectures, if any at all. In addition, students will be expected to attend selected computer architecture seminars and provide reviews of the seminars as well as the papers presented during the seminars.

Major Course Component 2

- **2. Semester-long research project:** Students will be expected to propose and carry out a clearly-defined research project in computer architecture. The goal of the research project will be to substantially advance the state of the art in computer architecture. The project is open-ended and has to be approved by the instructor. Project deliverables will consist of a project proposal writeup, milestone presentations, a literature survey report and presentation, and a final project report, poster and presentation. Students are expected to conduct research projects that are likely publishable.
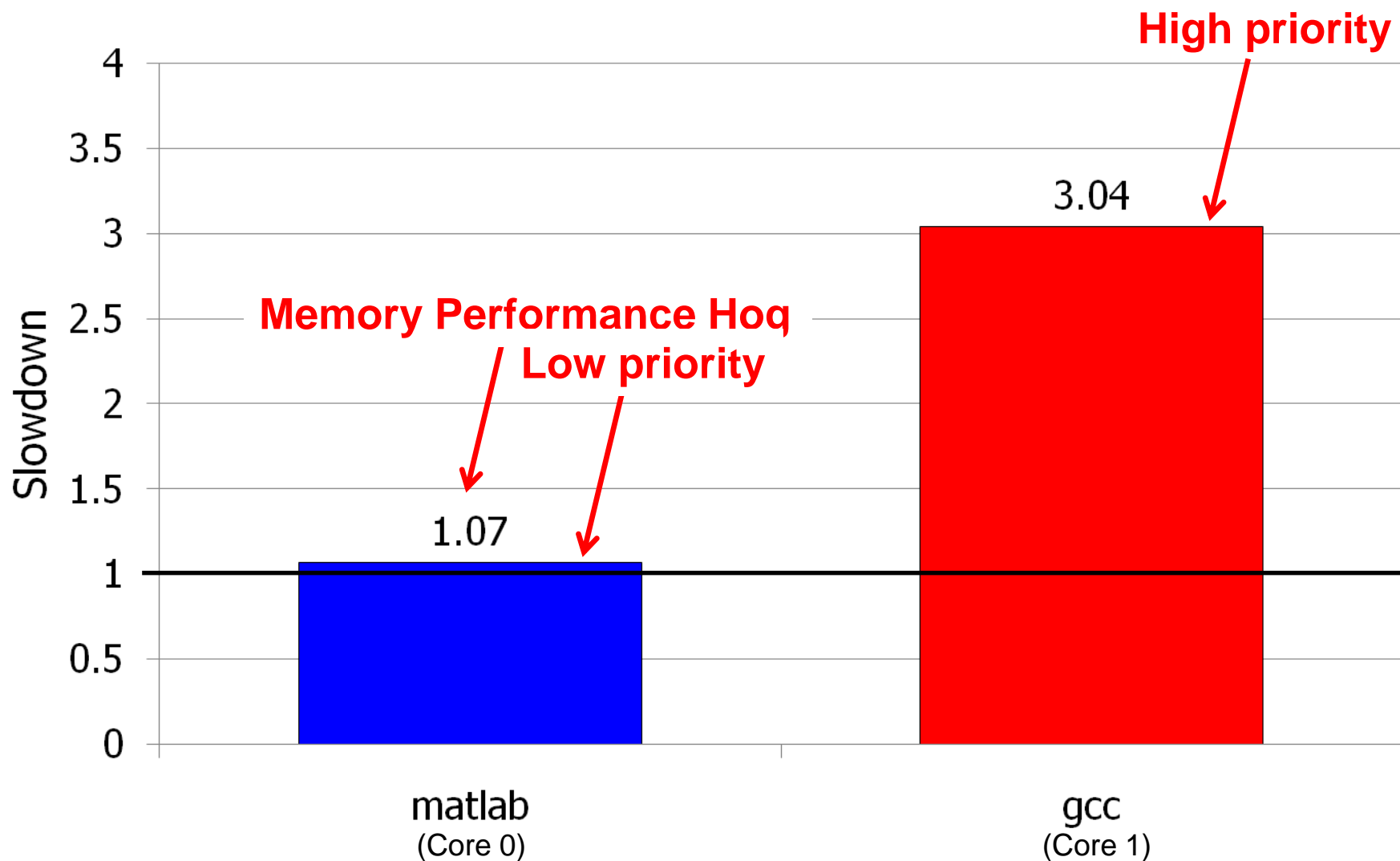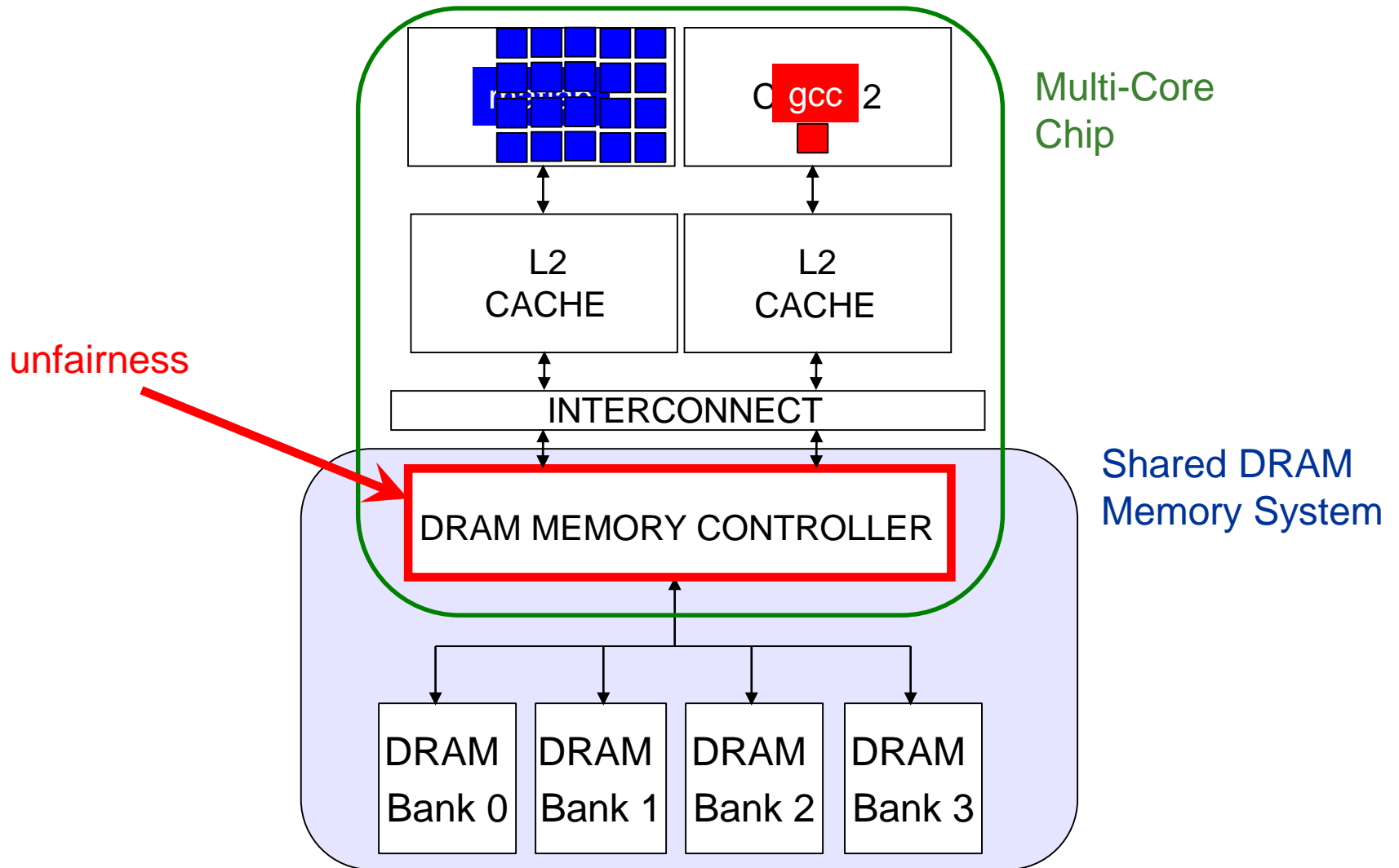
An Example Project Problem

Multi-Core
Chip

SHARED L3 CACHE

CORE 0

L2 CACHE 0

L2 CACHE 1

CORE 1

DRAM INTERFACE

DRAM MEMORY CONTROLLER

CORE 2

L2 CACHE 2

L2 CACHE 3

CORE 3

DRAM BANKS

*Die photo credit: AMD Barcelona

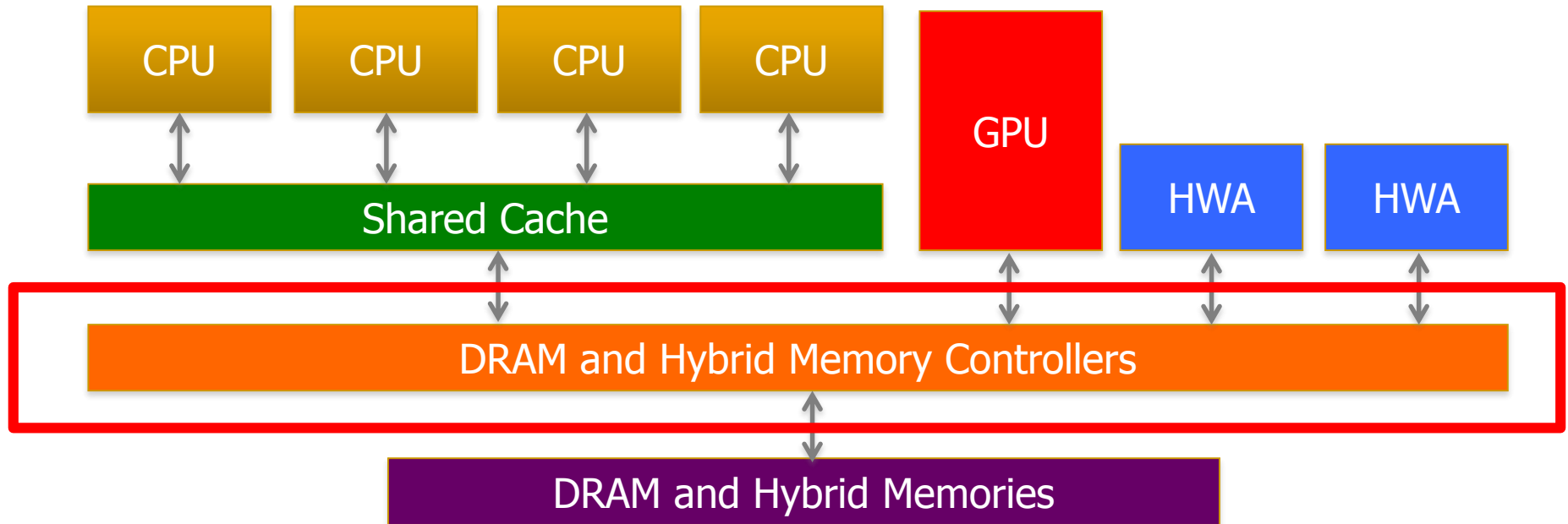# Unexpected Slowdowns in Multi-Core

Why the Disparity in Slowdowns?

For More Information, Read

- Mutlu and Moscibroda, "Memory Performance Attacks: Denial of Memory Service in Multi-Core Systems", USENIX Security 2007.

Predictable Performance in Complex Systems

- Heterogeneous agents: CPUs, GPUs, and HWAs
- Main memory interference between CPUs, GPUs, HWAs

How to allocate resources to heterogeneous agents
to mitigate interference and provide predictable performance?

Strong Memory Service Guarantees

- Goal: Satisfy performance/SLA requirements in the presence of shared main memory, prefetchers, heterogeneous agents, and hybrid memory/storage

- Approach:
  - Develop techniques/models to accurately estimate the performance of an application/agent in the presence of resource sharing
  - Develop mechanisms (hardware and software) to enable the resource partitioning/prioritization needed to achieve the required performance levels for all applications
  - All the while providing high system performance

- Example work: Subramanian et al., "MISE: Providing Performance Predictability and Improving Fairness in Shared Main Memory Systems," HPCA 2013.
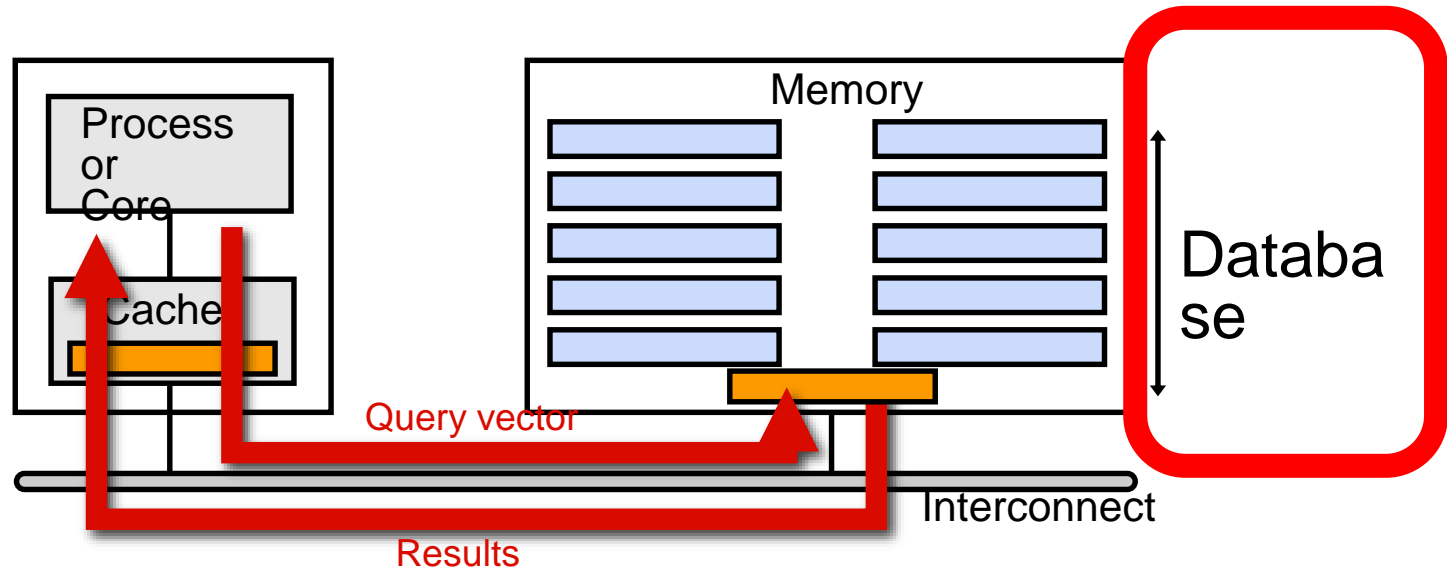
Many Project Ideas in the Previous Slide

- Another example, in the next two slides…

- Building upon
  - **"RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization"**, MICRO 2013.

Goal: Ultra-Efficient Processing By Data
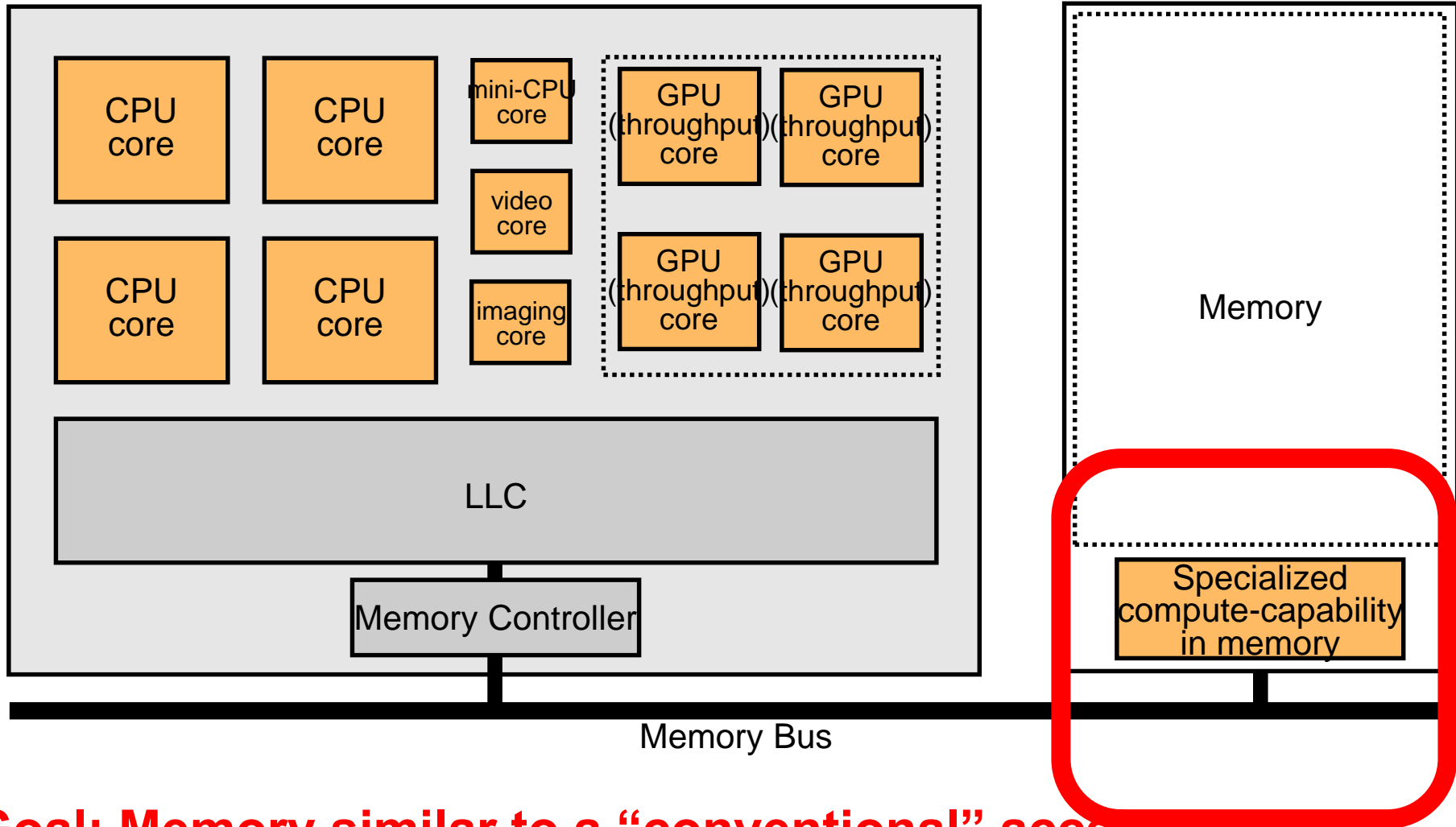
Goal: Memory similar to a "conventional" accelerator
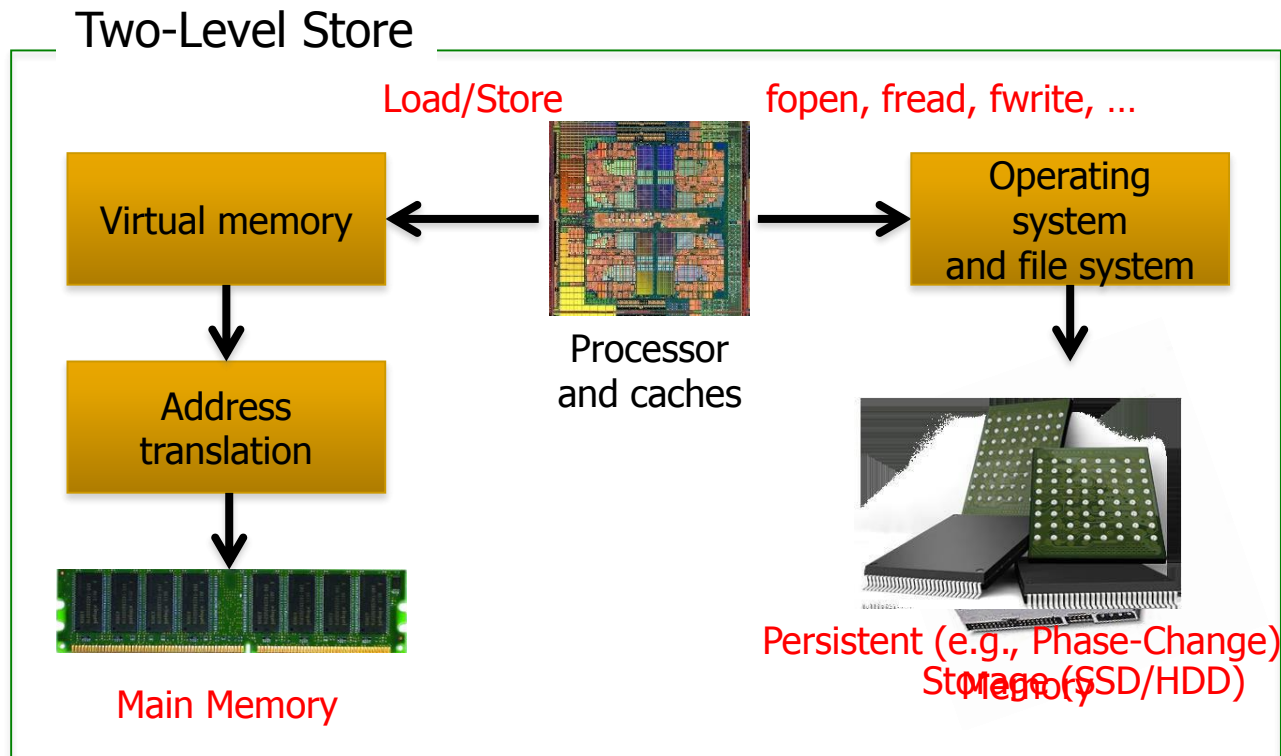
Enabling Ultra-Efficient Search

- What is the right partitioning of computation capability?

- What is the right low-cost memory substrate?

- What memory technologies are the best enablers?

- How do we rethink/ease (visual) search

Other Project Topics: Emerging Memory Tech.

- **Merging of memory and storage**
  - e.g., a single interface to manage all data

- **New applications**
  - e.g., ultra-fast checkpoint and restore

- **More robust system design**
  - e.g., reducing data loss

- **Processing tightly-coupled with memory**
  - e.g., enabling efficient search and filtering

SAFARI

Coordinated Memory and Storage with NVM (I)

- **The traditional two-level storage model is a bottleneck with NVM**
  - **Volatile** data in memory → a **load/store** interface
  - **Persistent** data in storage → a **file system** interface
  - Problem: Operating system (OS) and file system (FS) code to locate, translate, buffer data become performance and energy bottlenecks with fast NVM stores

Two-Level Store

Load/Store   fopen, fread, fwrite, …

Virtual memory

Address translation

Processor and caches

Operating system and file system

Main Memory

Persistent (e.g., Phase-Change Memory)
Storage (SSD/HDD)

Coordinated Memory and Storage with NVM (II)

- Goal: Unify memory and storage management in a single unit to eliminate wasted work to locate, transfer, and translate data
  - Improves both energy and performance
  - Simplifies programming model as well

Unified Memory/Storage

Persistent Memory Manager

Processor and caches

Load/Store   Feedback

Persistent (e.g., Phase-Change) Memory

Meza+, "A Case for Efficient Hardware-Software Cooperative Management of Storage and Memory," WEED 2013.
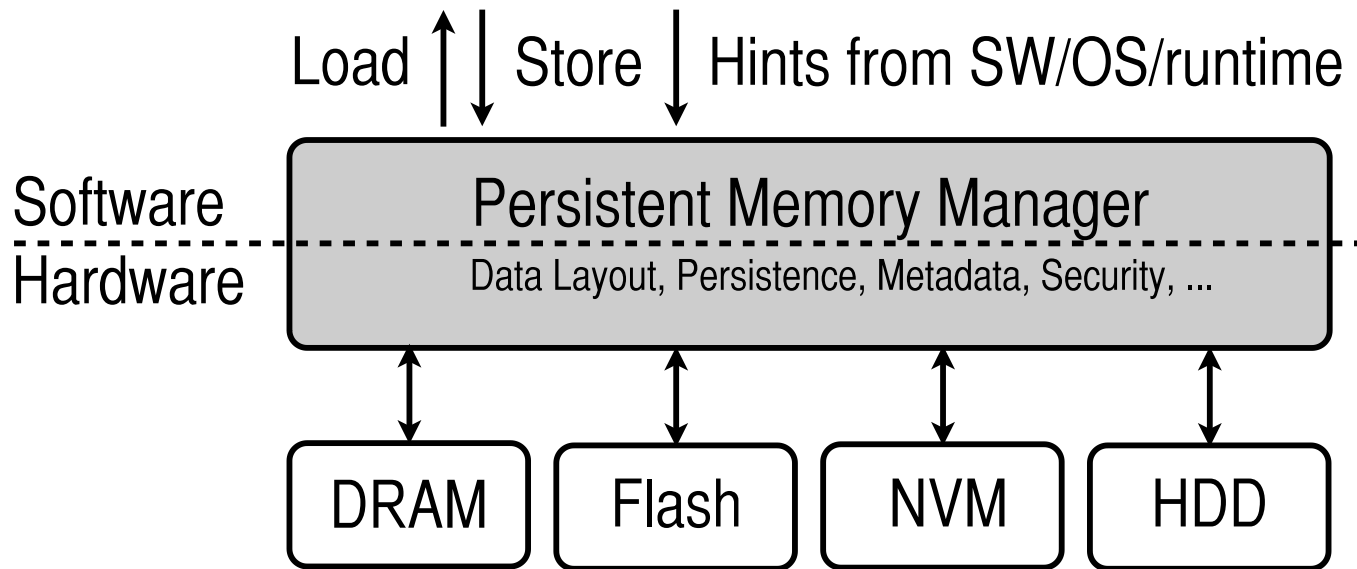
SAFARI
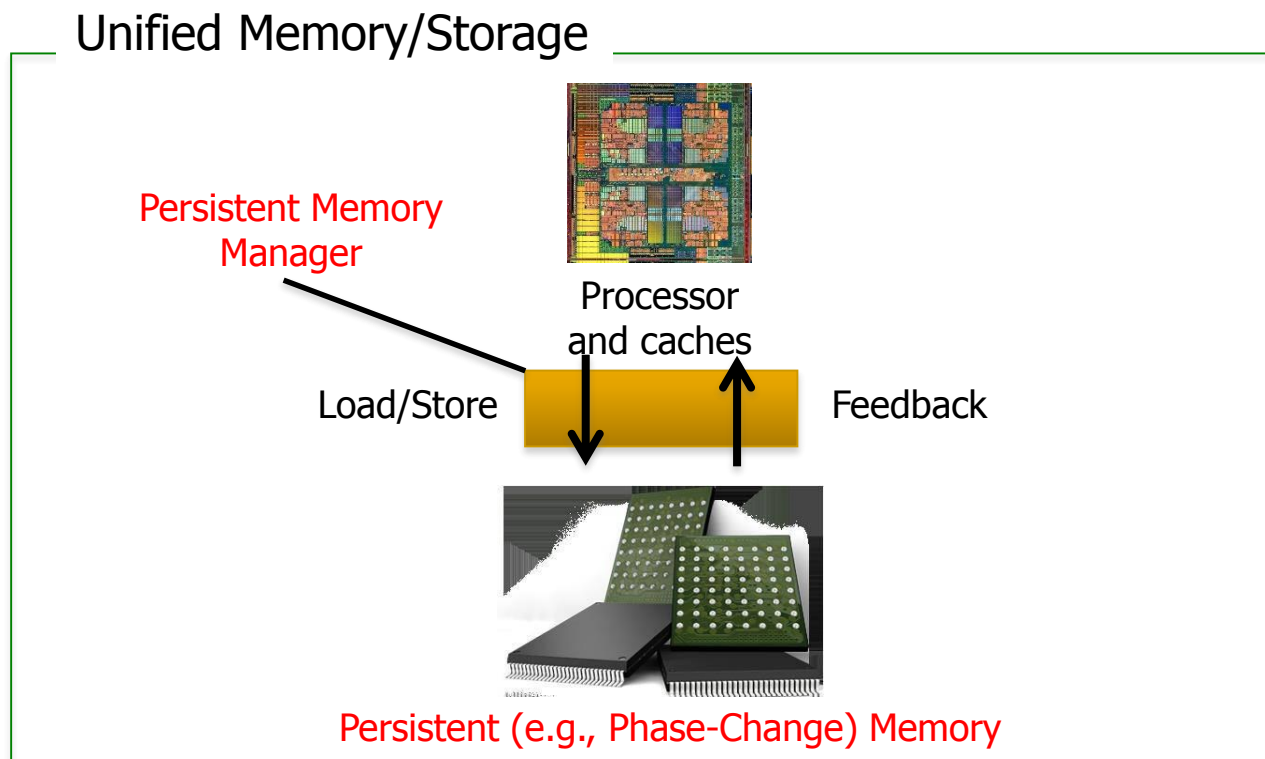
The Persistent Memory Manager (PMM)

- **Exposes a load/store interface to access persistent data**
  - Applications can directly access persistent memory → no conversion, translation, location overhead for persistent data

- **Manages data placement, location, persistence, security**
  - To get the best of multiple forms of storage

- **Manages metadata storage and retrieval**
  - This can lead to overheads that need to be managed

- **Exposes hooks and interfaces for system software**
  - To enable better data placement and management decisions

- Meza+, "A Case for Efficient Hardware-Software Cooperative Management of Storage and Memory," WEED 2013.

The Persistent Memory Manager (PMM)

```
1  int main(void) {
2    // data in file.dat is persistent
3    FILE myData = "file.dat";        Persistent objects
4    myData = new int[64];
5  }
6  void updateValue(int n, int value) {
7    FILE myData = "file.dat";
8    myData[n] = value; // value is persistent
9  }
```

Load ↕ Store ↓ Hints from SW/OS/runtime

Software | **Persistent Memory Manager**
Hardware | Data Layout, Persistence, Metadata, Security, ...
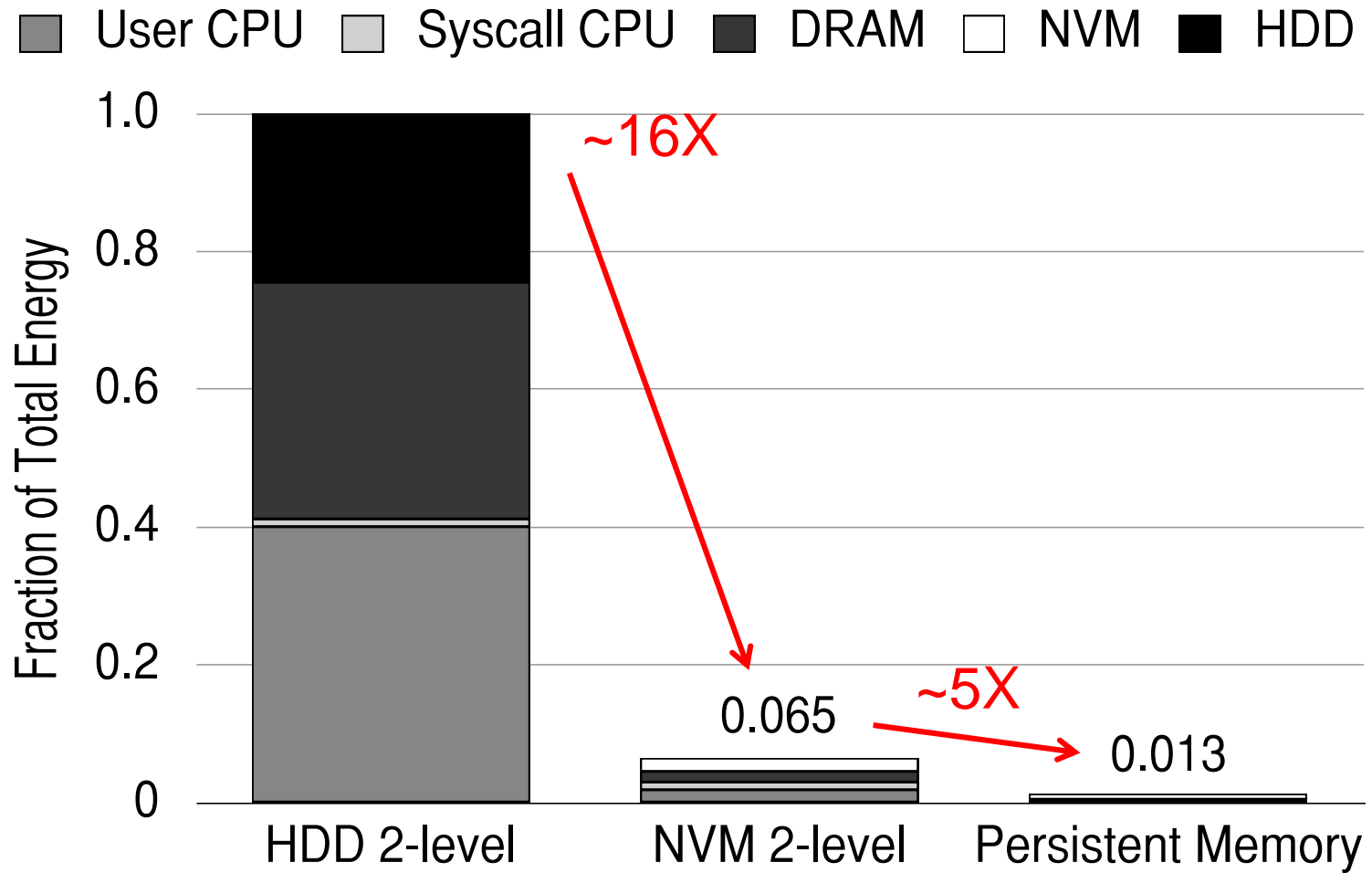
DRAM | Flash | NVM | HDD

PMM uses access and hint information to allocate, locate, migrate and access data in the heterogeneous array of devices

Performance Benefits of a Single-Level Store

Results for PostMark
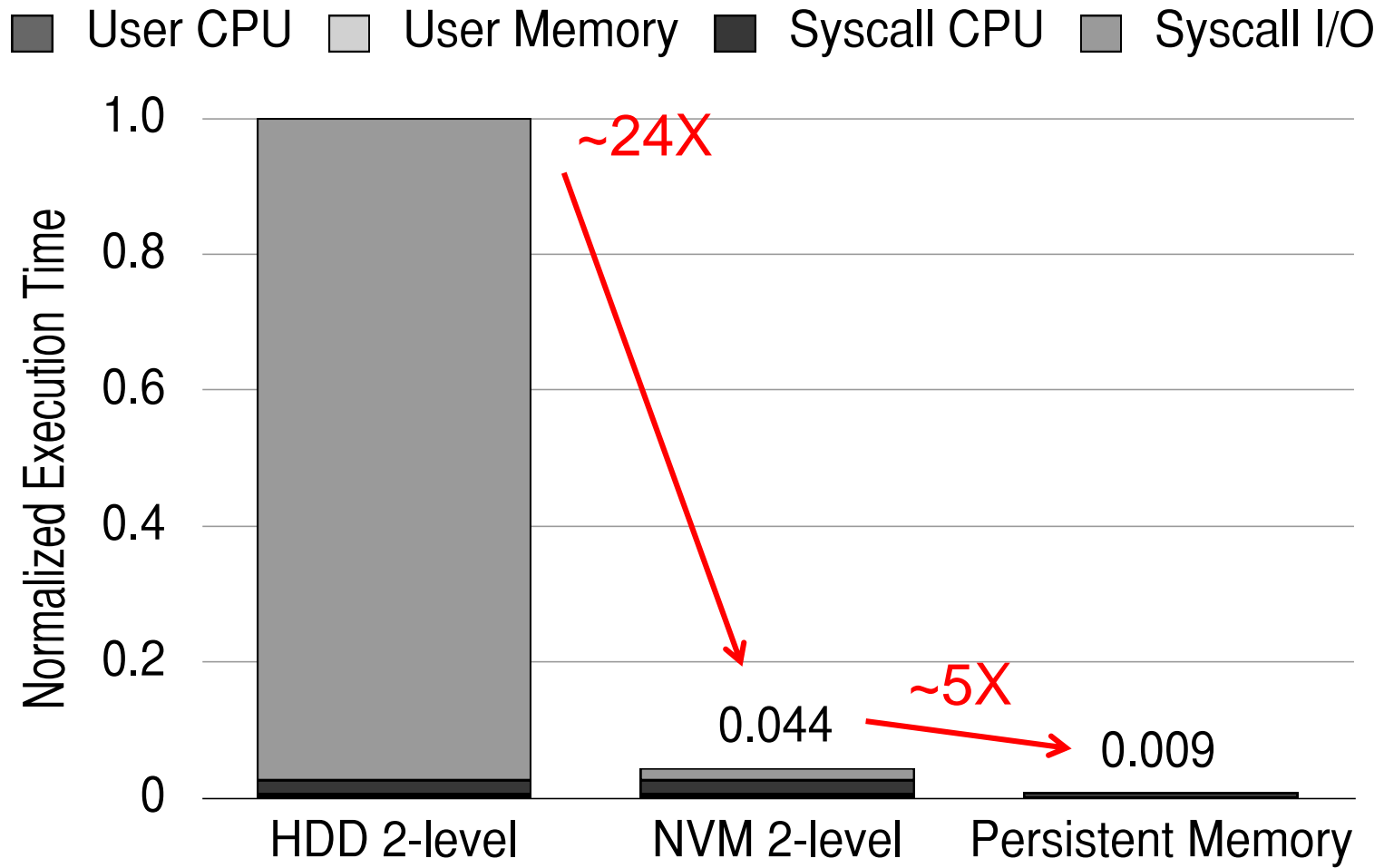
Energy Benefits of a Single-Level Store

SAFARI Results for PostMark

Many Exciting Project Topic Areas

- ***Rethinking Memory System Design** for Data-Intensive Computing*
  - *All aspects of DRAM, Flash Memory, Emerging Technologies*

- ***Single-Level Stores**: Merging Memory and Storage with Fast NVM*

- ***GPUs as First-Class Computing Engines***

- ***In-memory Computing:** Enabling Near-Data Processing*

- ***Predictable Systems**: QoS Everywhere in the System*

- ***Secure and Easy-to-Program/Manage Memories:** DRAM, Flash, NVM*

- ***Heterogeneous Systems:** Architecting and Exploiting Asymmetry*

- ***Efficient and Scalable Interconnects***

- ***Genome Sequence Analysis & Assembly:** Algorithms and Architectures*

SAFARI

29

Sample Past Projects

- **"ATLAS: A Scalable and High-Performance Scheduling Algorithm for Multiple Memory Controllers"** , HPCA 2010 Best Paper Session.

- **"Next Generation On-Chip Networks: What Kind of Congestion Control Do We Need?"** , HotNets 2010.

- **"Thread Cluster Memory Scheduling: Exploiting Differences in Memory Access Behavior"** , MICRO 2010, IEEE Micro Top Picks 2011.

- **"Reducing Memory Interference in Multicore Systems via Application-Aware Memory Channel Partitioning"**, MICRO 2011.

- **"RAIDR: Retention-Aware Intelligent DRAM Refresh"**, ISCA 2012.

- **"On-Chip Networks from a Networking Perspective: Congestion and Scalability in Many-core Interconnects"**, SIGCOMM 2012.

- **"Row Buffer Locality Aware Caching Policies for Hybrid Memories"**, ICCD 2012 Best Paper Award.

- **"HAT: Heterogeneous Adaptive Throttling for On-Chip Networks"**, SBAC-PAD 2012.

- **"Asymmetry-Aware Execution Placement on Manycore Chips"**, SFMA 2013.

- **"Exploiting Compressed Block Size as an Indicator of Future Reuse"**, SAFARI Technical Report 2013.

What Do I Expect From You?

- Be enthusiastic to learn, invent, question, create

- Work hard and diligently

- Ask questions, take notes, participate in discussion

- Do great research = preparation + opportunism + luck

- Start the research project early and focus on it

How Will You Be Evaluated?

- You should not be taking this course to get a grade.

- Learning how to do great research and think critically is the goal. A strong outcome is to have done research that is published.

- Participation is vey important.

- See syllabus for more details.

That Said …

- If you are really curious

40% - Research Project (detailed breakdown to follow, but will
milestone writeups/presentations, literature surveys, final prese
40% - Paper Reviews and Presentations
10% - Class Participation, Quizzes, Ad-Hoc Assignments
10% - Exam (Could potentially be an oral or take-home exam)
5%   - The teaching team's evaluation of your performance

Guidelines on Paper Reviews

Prof. Onur Mutlu

Carnegie Mellon University

Fall 2014

How to Do the Paper/Talk Reviews

- 1: Brief summary
  - What is the problem the paper is trying to solve?
  - What are the key ideas of the paper? Key insights?
  - What is the key contribution to literature at the time it was written?
  - What are the most important things you take out from it?
- 2: Strengths (most important ones)
  - Does the paper solve the problem well?
- 3: Weaknesses (most important ones)
  - This is where you should think critically. Every paper/idea has a weakness. This does not mean the paper is necessarily bad. It means there is room for improvement and future research can accomplish this.
- 4: Can you do (much) better? Present your thoughts/ideas.
- 5: What have you learned/enjoyed/disliked in the paper? Why?

- Review should be short and concise (~a page or shorter)

Advice on Paper/Talk Reviews

- When doing the reviews, be very critical

- Always think about better ways of solving the problem or related problems

- Do background reading
  - Reviewing a paper/talk is the best way of learning about a research problem/topic

- Think about forming a literature survey topic or a research proposal

Reading(s) on Refereeing CS Papers

- Smith, "The Task of the Referee," IEEE Computer 1990.
  - Provides an idea of the publication process
  - Provides guidance on how to perform technical reviews

- Also see:
  - Hill and McKinley, "Notes on Constructive and Positive Reviewing"
    - http://www.cs.utexas.edu/users/mckinley/notes/reviewing.html
    - Levin and Redell, "How (and how not) to write a good systems paper," OSR 1983.
    - Jones, "How to Write a Great Research Paper"

Literature Survey

- More information to come… In the meantime:

- Read a lot of papers; find focused problem areas to survey papers on and do your project on. It is best if your survey is related to your course project.

- We will provide a list of project ideas and papers associated with them

- A good way of finding topics to survey or do projects on is:
  - Examining the provided project ideas and papers
  - Reading assigned papers in lectures
  - Examining papers from recent conferences (ISCA, MICRO, HPCA, ASPLOS, …)

Project Proposal and Topics

Prof. Onur Mutlu

Carnegie Mellon University

Fall 2014

Research Project

- Your chance to explore in depth a computer architecture topic that interests you
- Hopefully publish your innovation in a top computer architecture conference.

- Start thinking about your project topic from now!
- Interact with me and the TAs
- Constantly think good project topics and ideas

- Groups of 2 students
- Proposal due: within ~3 weeks

Research Project

- Goal: Develop (new) insight
  - Solve a problem in a new way or evaluate/analyze systems/ideas
  - Type 1:
    - Develop new ideas to solve an important problem
    - Rigorously evaluate the benefits and limitations of the ideas
  - Type 2:
    - Derive insight from rigorous analysis and understanding of existing systems or previously proposed ideas
    - Propose potential new solutions based on the new insight

- The problem and ideas need to be concrete

- Problem and goals need to be very clear

Research Proposal Outline: Type 1

- **The Problem:** What is the problem you are trying to solve
  - Define very clearly. Explain why it is important.
- **Novelty:** Why has previous research not solved this problem? What are its shortcomings?
  - Describe/cite all relevant works you know of and describe why these works are inadequate to solve the problem. This will be your literature survey.
- **Idea:** What is your initial idea/insight? What new solution are you proposing to the problem? Why does it make sense? How does/could it solve the problem better?
- **Hypothesis:** What is the main hypothesis you will test?
- **Methodology:** How will you test the hypothesis/ideas? Describe what simulator or model you will use and what initial experiments you will do.
- **Plan:** Describe the steps you will take. What will you accomplish by Milestone 1, 2, 3, and Final Report? Give 75%, 100%, 125% and moonshot goals.

All research projects can be (and should be) described in this fashion.

Research Proposal Outline: Type 2

- **The Problem:** What is the problem/phenomenon you are trying to evaluate?
  - ❑ Define very clearly. Explain why it is important.
- **Novelty:** How has previous research evaluated this? What are its shortcomings?
  - ❑ Describe/cite all relevant works you know of and describe why these works are inadequate to solve the problem. This will be your literature survey.
- **Evaluation method:** How will you evaluate the phenomenon/idea? What experimental infrastructure will you design? How would that experimental infrastructure enable you to reliably evaluate the phenomenon/idea?
- **Hypotheses:** What hypotheses will you test?
- **Methodology:** What are your plans for evaluation? What experiments will you run? How will you do the data analysis?
- **Plan:** Describe the steps you will take. What will you accomplish by Milestone 1, 2, 3, and Final Report? Give 75%, 100%, 125% and moonshot goals.

All research projects can be (and should be) described in this fashion.

Heilmeier's Catechism (version 1)

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What's new in your approach and why do you think it will be successful?
- Who cares?
- If you're successful, what difference will it make?
- What are the risks and the payoffs?
- How much will it cost?
- How long will it take?
- What are the midterm and final "exams" to check for success?

Heilmeier's Catechism (version 2)

- What is the problem?

- Why is it hard?

- How is it solved today?

- What is the new technical idea?

- Why can we succeed now?

- What is the impact if successful?

- http://en.wikipedia.org/wiki/George_H._Heilmeier

Supplementary Readings on Research, Writing, Reviews

- Hamming, "You and Your Research," Bell Communications Research Colloquium Seminar, 7 March 1986.
  - http://www.cs.virginia.edu/~robins/YouAndYourResearch.html

- Levin and Redell, "How (and how not) to write a good systems paper," OSR 1983.

- Smith, "The Task of the Referee," IEEE Computer 1990.
  - Read this to get an idea of the publication process

- SP Jones, "How to Write a Great Research Paper"

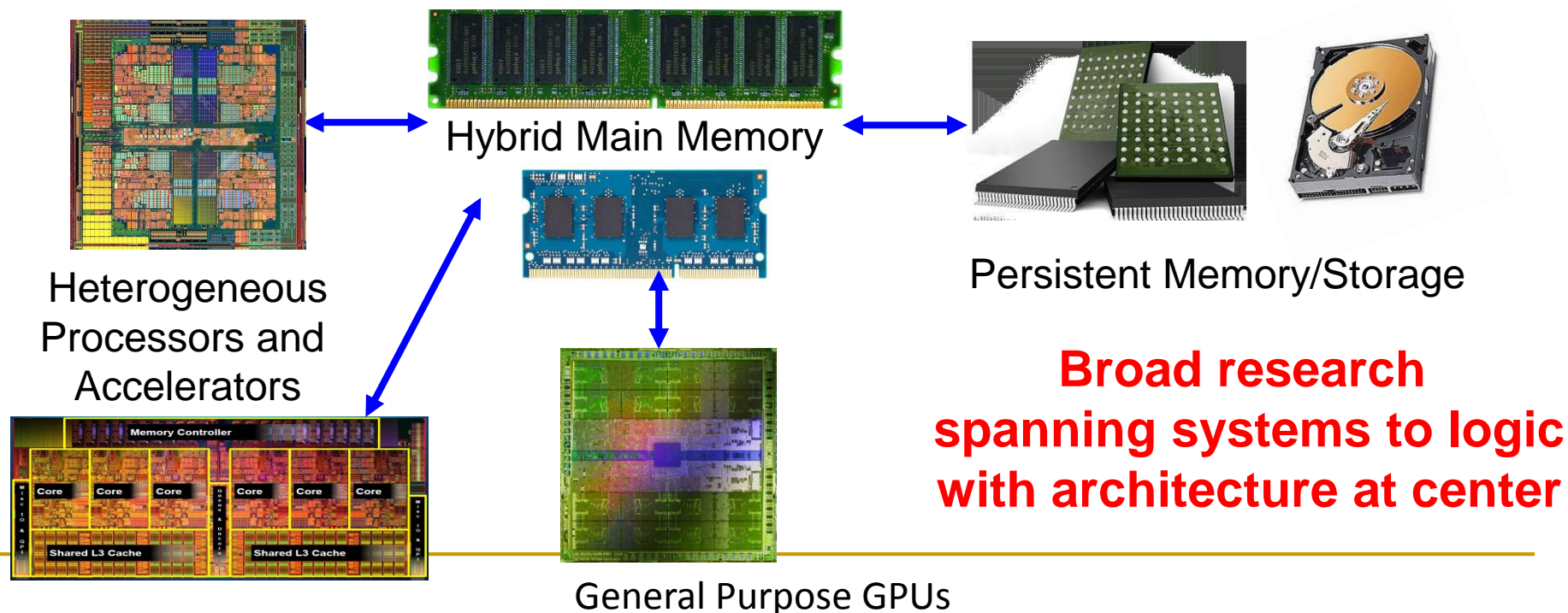- Fong, "How to Write a CS Research Paper: A Bibliography"

Where to Get Project Topics/Ideas From

- Project topics handout

- Assigned readings
  - e.g., today's book chapter on Main Memory Scaling

- Talk with me and my students

- Recent conference proceedings
  - ISCA: http://www.informatik.uni-trier.de/~ley/db/conf/isca/
  - MICRO: http://www.informatik.uni-trier.de/~ley/db/conf/micro/
  - HPCA: http://www.informatik.uni-trier.de/~ley/db/conf/hpca/
  - ASPLOS: http://www.informatik.uni-trier.de/~ley/db/conf/asplos/

Research Focus: *Computer architecture, HW/SW, bioinformatics*

- *Memory, memory, memory, storage, interconnects*
- *Parallel architectures, heterogeneous architectures, GP-GPUs*
- *System/architecture interaction, new execution models*
- *Energy efficiency, fault tolerance, hardware security*
- *Genome sequence analysis and assembly*

Hybrid Main Memory

Persistent Memory/Storage

Heterogeneous
Processors and
Accelerators

**Broad research
spanning systems to logic
with architecture at center**

General Purpose GPUs

Some Current Directions and Projects

- *Rethinking Memory System Design for Data-Intensive Computing*
  - *All aspects of DRAM, Flash Memory, Emerging Technologies*

- *Single-Level Stores: Merging Memory and Storage with Fast NVM*

- *GPUs as First-Class Computing Engines*

- *In-memory Computing: Enabling Near-Data Processing*

- *Predictable Systems: QoS Everywhere in the System*

- *Secure and Easy-to-Program/Manage Memories: DRAM, Flash, NVM*

- *Heterogeneous Systems: Architecting and Exploiting Asymmetry*

- *Efficient and Scalable Interconnects*

- *Genome Sequence Analysis & Assembly: Algorithms and Architectures*

Example Infrastructure for Future Memory Designs
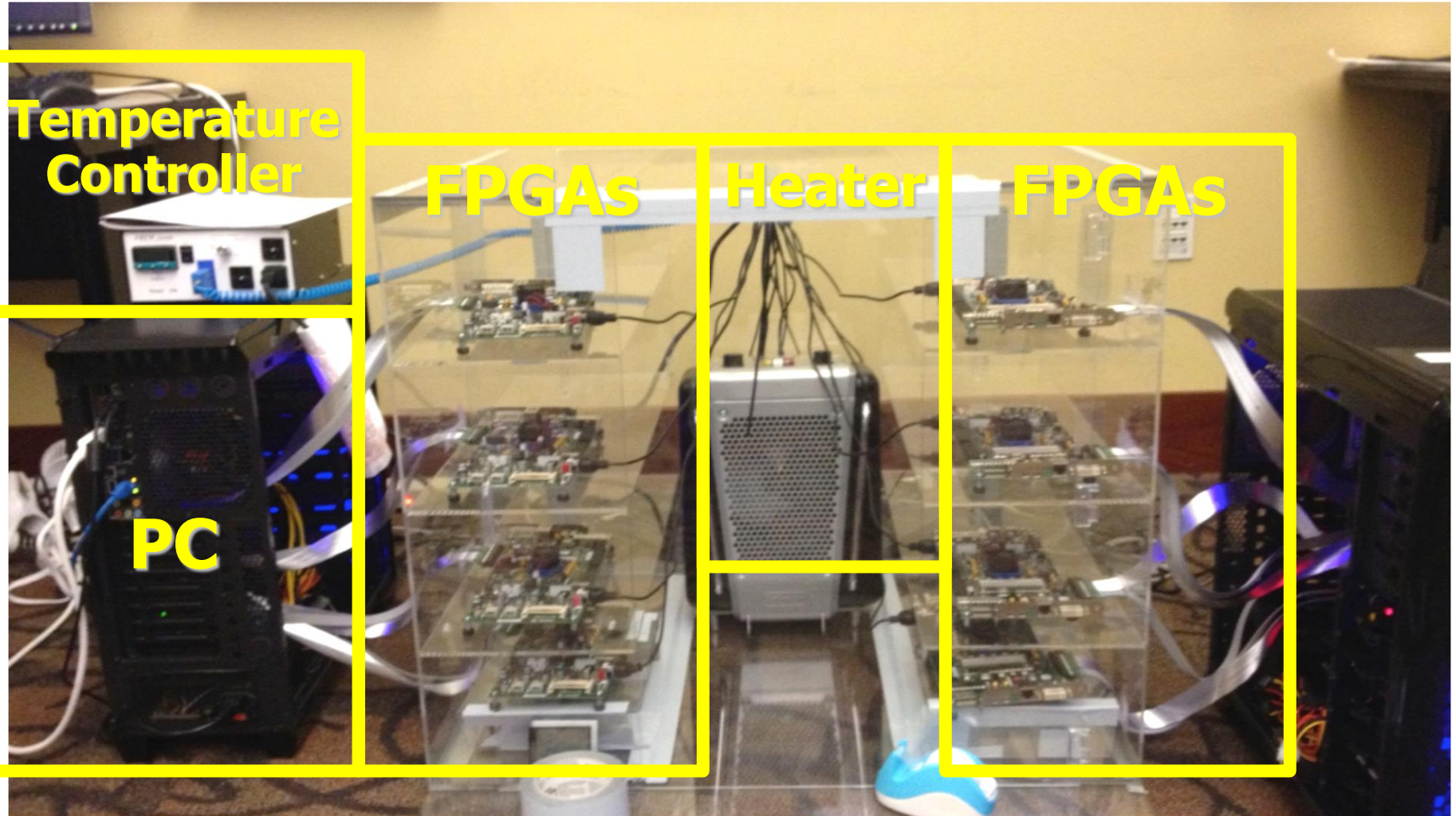
More Places to Get Ideas From

- **Computer Architecture Lecture Videos on Youtube**
  - https://www.youtube.com/playlist?list=PL5PHm2jkkXmidJOd59REog9jDnPDTG6IJ

- **Computer Architecture Course Materials**
  - http://www.ece.cmu.edu/~ece447/s13/doku.php?id=schedule

- **Advanced Computer Architecture Course Materials**
  - http://www.ece.cmu.edu/~ece740/f13/doku.php?id=schedule

- **Advanced Computer Architecture Lecture Videos on Youtube**
  - https://www.youtube.com/playlist?list=PL5PHm2jkkXmgDN1PLwOY_tGtUlynnyV6D

More Places to Get Ideas From

- Research Papers

  http://users.ece.cmu.edu/~omutlu/projects.htm

  http://scholar.google.com/citations?user=7XyGUGkAAAAJ&hl=en

Assignments for Next Week

Prof. Onur Mutlu

Carnegie Mellon University

Fall 2014

Assignments for Next Week

- Background brush-up assignments

- Homework 0: student information sheet (due September 2)

- Paper Review Set 1 (due September 2)

- Readings to prepare you for Research and Report Writing

Required Videos (for Background)

- Lecture 1: Basics of Computer Architecture (447 Spring 2013)
- http://www.youtube.com/watch?v=BJ87rZCGWU0&list=PL5PHm2jkkXmidJOd59REog9jDnPDTG6IJ&index=1

More Videos and Background

- **Computer Architecture Lecture Videos on Youtube**
  - https://www.youtube.com/playlist?list=PL5PHm2jkkXmidJOd59REog9jDnPDTG6IJ

- **Computer Architecture Course Materials**
  - http://www.ece.cmu.edu/~ece447/s13/doku.php?id=schedule

- **Advanced Computer Architecture Course Materials**
  - http://www.ece.cmu.edu/~ece740/f13/doku.php?id=schedule

- **Advanced Computer Architecture Lecture Videos on Youtube**
  - https://www.youtube.com/playlist?list=PL5PHm2jkkXmgDN1PLwOY_tGtUlynnyV6D

Assignments for Next Week

- Background brush-up assignments

- Quiz 0: student information sheet (due September 2)

- Paper Review Set 1 (due September 2)

- Readings to prepare you for Research and Report Writing

Homework 0: Information Sheet

- Please send me and Yixin an introduction email with at least the following (all are mandatory)
    - Your full name and picture
    - Your advisor's name (if you have one)
    - Your main research topic (if you have one)
    - Degree you are enrolled for and when you expect to finish
    - When did you take 447 and with which professor?
    - When did you take 740 and with which professor?
    - Why are you taking this course?
    - What topics in computer architecture excite you most?
    - What are your future career plans?
    - What keeps you up at night?
    - Any other comments?
- Due September 2, 2014

Assignments for Next Week

- Background brush-up assignments

- Homework 0: student information sheet (due September 2)

- Paper Review Set 1 (due September 2)

- Readings to prepare you for Research and Report Writing

Review Set 1: Due September 2-4

- Mutlu, "Main Memory Scaling: Challenges and Solution Directions," Preprint Book Chapter 2014.
    - Be ready to discuss and present the paper next week
    - Write a critical review online
    - **Due September 2**

- Pick 3 papers referenced by the above paper that pique your interest
    - Read each very carefully
    - Write a critical review for each online
    - Be ready to discuss and present each paper next week
    - **Due September 4**

While Reviewing …

- Think project ideas

- Think critically

- Exercise your creativity to solve the described problems in different ways

- Read more, dig deeper as you become curious

Review Site

- http://safari2.ece.cmu.edu/18742-f14/submissions/

- Send the 3 papers you select to Yixin and me for the next assignment

Assignments for Next Week

- Background brush-up assignments

- Homework 0: student information sheet (due September 2)

- Paper Review Set 1 (due September 2)

- Readings to prepare you for Research and Report Writing

Required Reading: **Due September 6**

- Hamming, "You and Your Research," Bell Communications Research Colloquium Seminar, 7 March 1986.

  - http://www.cs.virginia.edu/~robins/YouAndYourResearch.html

- Be ready to discuss and present the paper next week
- Write a critical review online
- **Due September 6**

Supplementary Readings on Research, Writing, Reviews

❑ Levin and Redell, "How (and how not) to write a good systems paper," OSR 1983.

▪ Jones, "How to Write a Great Research Paper"

▪ Smith, "The Task of the Referee," IEEE Computer 1990.
   ❑ Read this to get an idea of the publication process

18-742: Research in
Parallel Computer Architecture
Intro and Logistics

Prof. Onur Mutlu

Carnegie Mellon University

Fall 2014

August 26, 2014