LOGISTICS AND INTRODUCTION

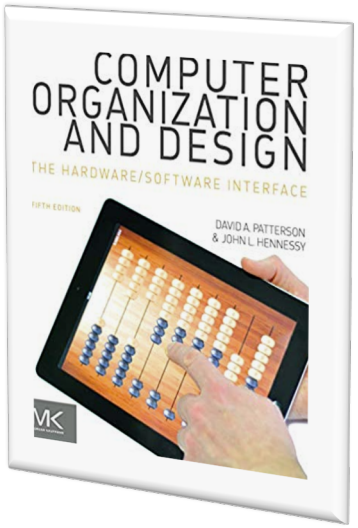
Mahdi Nazm Bojnordi

Assistant Professor

School of Computing

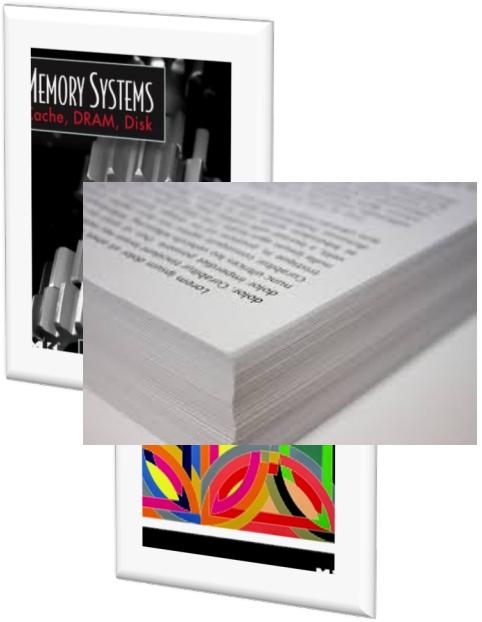
University of Utah

Advanced Computer Architecture



Basics of Computer Systems: CPU, Memory, Storage, IO, etc.

Processor/Memory Performance Optimization: ILP, TLP, AMAT, etc.



Today/Future Concerns: Power Wall, Energy-efficiency, Security, etc.

Logistics

Course organization and rules

Instructor

- Mahdi Nazm Bojnordi
  - Assistant Professor, School of Computing
  - PhD degree in Electrical Engineering (2016)
  - Worked in industry for four years (before PhD)
- Research in Computer Architecture
  - Energy-efficient computing
  - Emerging memory technologies
- Office Hours
  - Please email me for appointment
  - MEB 3418

This Course

- Prerequisite
  - CS/ECE 6810: Computer Architecture
- Advanced topics in computer architecture
  - cache energy innovations
  - memory system optimizations
  - interconnection networks
  - cache coherence protocols
  - emerging computation models

Resources

- Recommended books and references
  - "Memory Systems: Cache, DRAM, Disk", Jacob et al
  - "Principles and Practices of Interconnection Networks", Dally and Towles
  - "Parallel Computer Architecture", Culler, Singh, Gupta
  - "Synthesis Lectures on Computer Architecture", Morgan & Claypool Publishers

- Class webpage
  - http://www.cs.utah.edu/~bojnordi/classes/7810/s21/

Class Webpage

- Please visit online!

CS/ECE 7810: Advanced Computer Architecture

Course Information

- Time: Mon/Wed 11:50-01:10PM
- Location: CANVAS
- Instructor: Mahdi Nazm Bojnordi, email: lastname@cs.utah.edu, office hours: email me for appointment, MEB 3418
- Pre-Requisite: CS/ECE 6810
- Textbook: "Memory Systems: Cache, DRAM, Disk", Jacob et al.
- Textbook: "Principles and Practices of Interconnection Networks", Dally and Towles.
- Textbook: "Parallel Computer Architecture", Culler, Singh, Gupta.
- "Synthesis Lectures on Computer Architecture", Morgan & Claypool Publishers.
- Canvas is the main venue for class announcements, homework assignements, and discussions.
- Description: This course is based on advanced topics in computer architecture, including cache energy innovations, memory system optimizations, interconnection networks, cache coherence protocols, and emerging computation models.
- Expectation: In addition to homework assignment and final exam, students are expected to present a conference paper related to their course project in April. A project presentation is expected for each group of students in the last two classes and a final project reports is due in May. Important dates are listed below.

Important Policies

Please refer to the College of Engineering Guidelines for disabilities, add, drop, appeals, etc. Notice that we have zero tolerance for cheating; as a result, please read the Policy Statement on Academic Misconduct, carefully. Also, you should be aware of the SoC Policies and Guidelines.

Class rosters are provided to the instructor with the student's legal name as well as "Preferred first name" (if previously entered by you in the Student

Course Expectation

- Use Canvas for all of your submissions
  - No scanned handwritten documents please!
- Grading

	Fraction	Notes
Project	50%	One simulation-based project
Homework	20%	One homework assignment
Paper presentation	10%	One in class paper presentation
Final	20%	

Course Project

- A creative, simulation-based project on
  - Memory system optimization (SRAM, DRAM, RRAM, etc.)
  - Data movement optimizations (Off/On–chip interfaces)
  - Hardware accelerators (GPU, FPGA, ASIC)
  - …
- Form a group of 2 people by Feb. 2
- Choose your topic by Feb. 10
- Prepare for an in-class presentation in April
- Prepare a conference-style report by end of May

Paper Presentation and Assignment

- Every student presents a paper in class
  - A related work on your course project is recommended
  - Three main components must be included
    - The goal and key idea
    - Strengths and weaknesses
    - Future work
  - Email me your paper by Mar. 29
    - Conferences such as ISCA, MICRO, ASPLOS, HPCA
- A homework assignment will be posted on Feb. 24
  - Due on Mar. 4 (11:59PM)

Academic Integrity

- Do NOT cheat!!
  - Disciplinary hearings are no fun
  - Please read the Policy Statement on Academic Misconduct, carefully.
  - We have no tolerance for cheating
- Also, read the College of Engineering Guidelines for disabilities, add, drop, appeals, etc.

Important Policies

Please refer to the College of Engineering Guidelines for disabilities, add, drop, appeals, etc. Notice that we have zero tolerance for cheating; as a result, please read the Policy Statement on Academic Misconduct, carefully. Also, you should be aware of the SoC Policies and Guidelines.

Class rosters are provided to the instructor with the student's legal name as well as "Preferred first name" (if previously entered by you in the Student Profile section of your CIS account). While CIS refers to this as merely a preference, I will honor you by referring to you with the name and pronoun that feels best for you in class, on papers, exams, group projects, etc. Please advise me of any name or pronoun changes (and please update CIS) so I can help create a learning environment in which you, your name, and your pronoun will be respected.

About You …

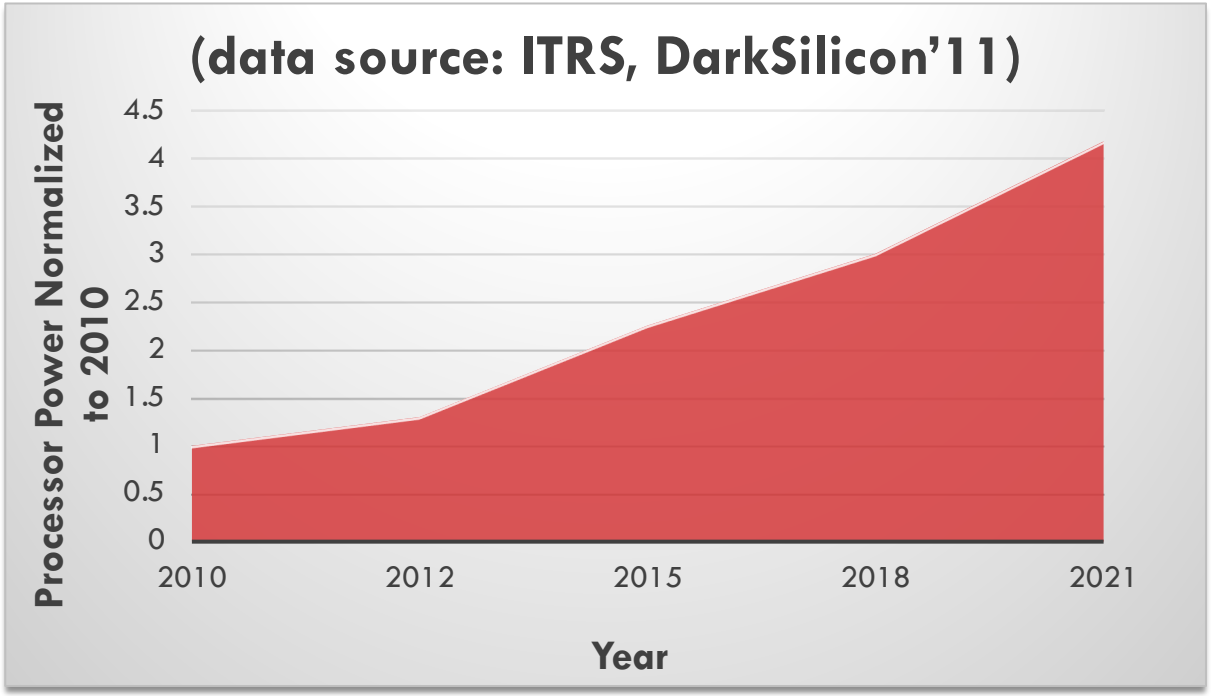
- Are you working in a research area?
- Do you know programming languages?
  - C/C++, etc.
- Do you know any hardware description languages?
  - Verilog
- Are you familiar with simulators?

Energy-efficient Computing

The importance of energy efficient computing

Energy and Power Trends

- Power consumption is increasing significantly



CPU Power Consumption

- Major power consumption issues

Peak Power/Power Density

- Heat
  - Packaging, cooling, component spacing
- Switching noise
  - Decoupling capacitors

Average Power

- Battery life
  - Bulkier battery
- Utility costs
  - Probability, cannot run your business!

New Challenges

- Excessive energy consumption
  - More energy-efficient architectures are needed

200M wearable devices will be sold in 2019 *(source: IDC forecast)*

New Challenges

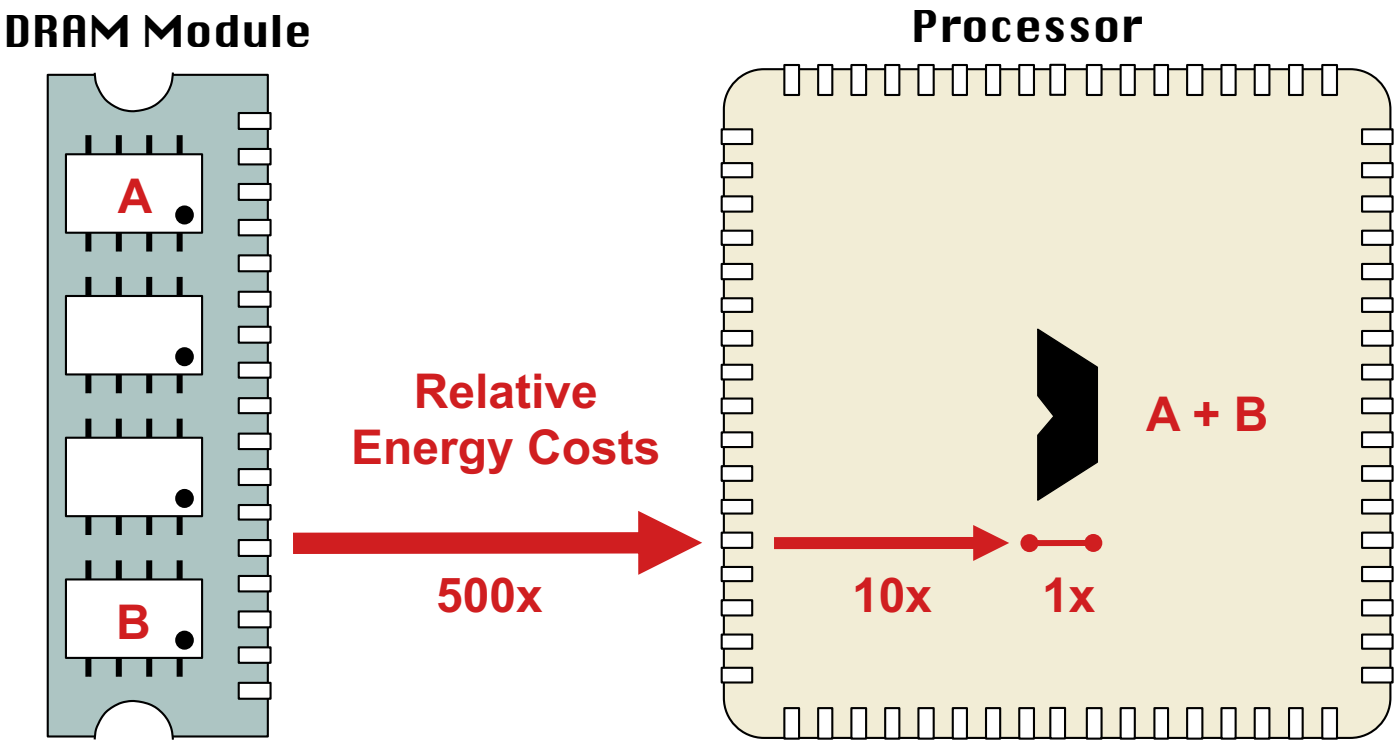
- Power delivery and cooling systems
  - More energy-efficient architectures are required

Facebook datacenter at edge of the Arctic circle *(source: CNET, 2013)*

Microsoft underwater datacenter *(source: NYTimes, 2016)*

The High Cost of Data Movement

- Data movement is the primary contributor to energy dissipation in nanometer ICs.

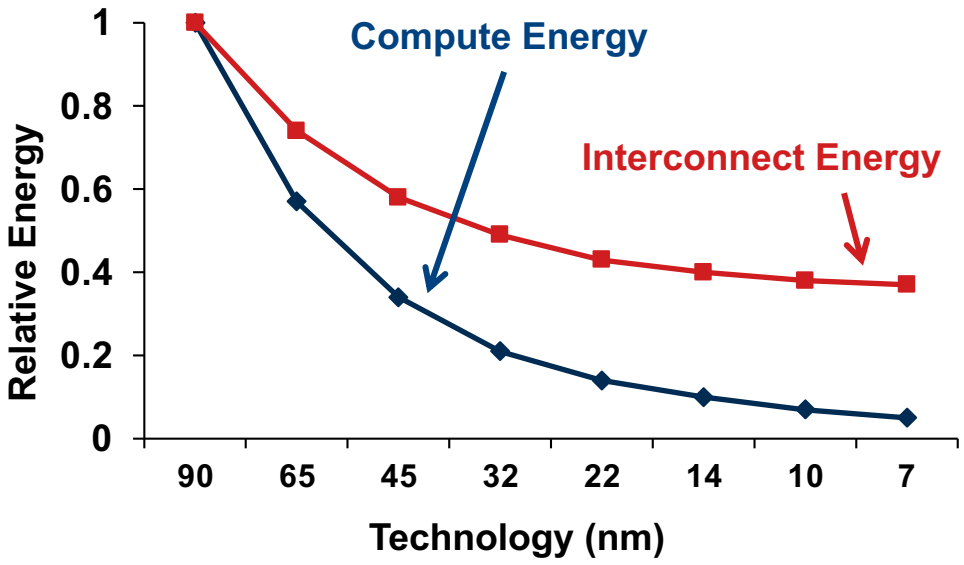


Source: NVidia

Data Movement Energy Increasing

- By 2020, the energy cost of moving data across the memory hierarchy will be orders of magnitude higher than the cost of performing a floating-point operation.

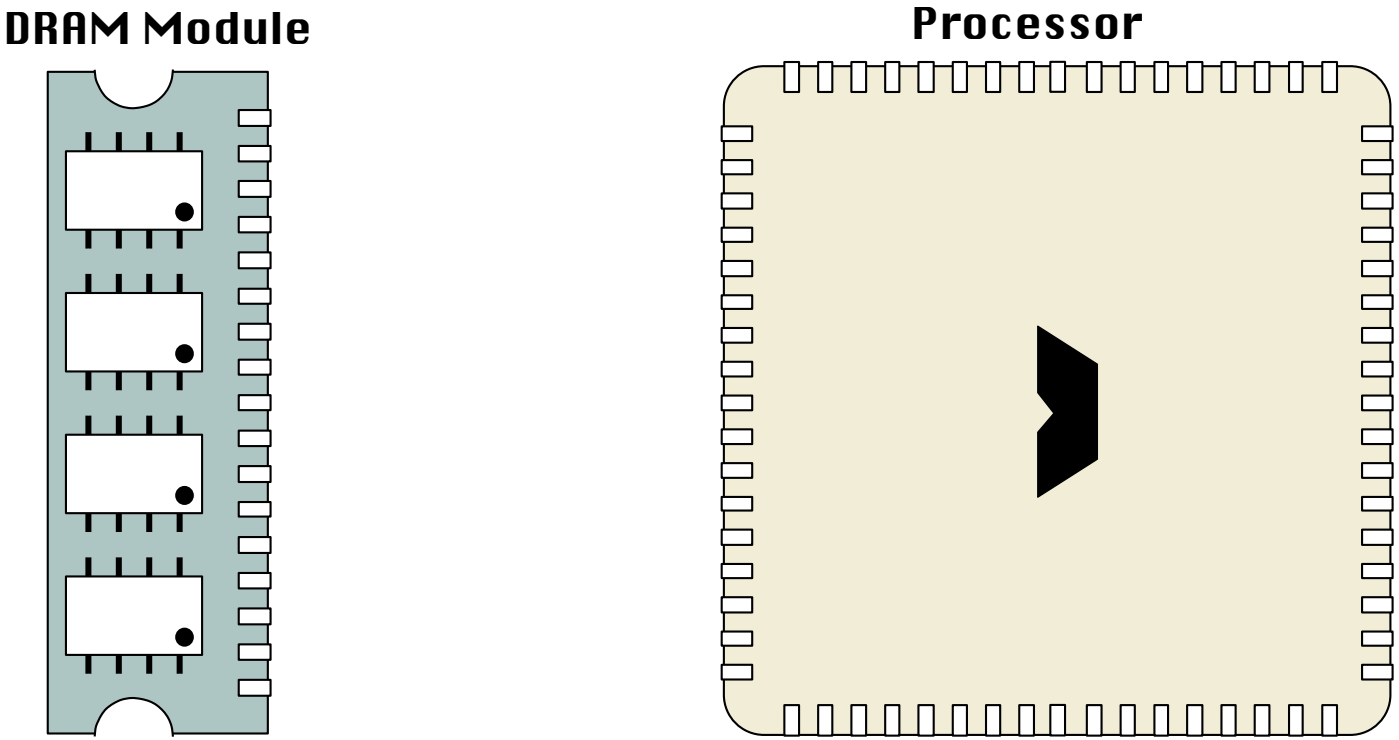
-- U.S. Department of Energy, 2014



Shekhar Borkar, Journal of Lightwave Technology, 2013

Possible Solutions

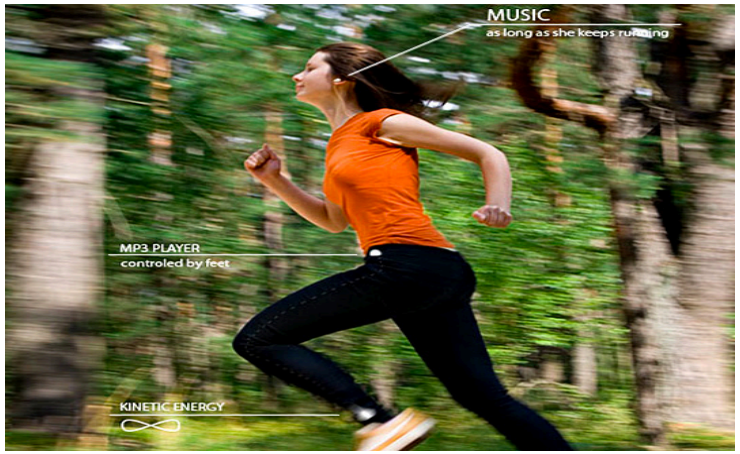
- How to minimize data movement energy?



Problem: Energy Efficiency

- Unconventional solutions are needed!
  - Hardware
  - Software

Solar powered dresses
(source: www.ecochunk.com)



Harvesting motion energy
(source: www.ecouterre.com)

Hardware Architecture

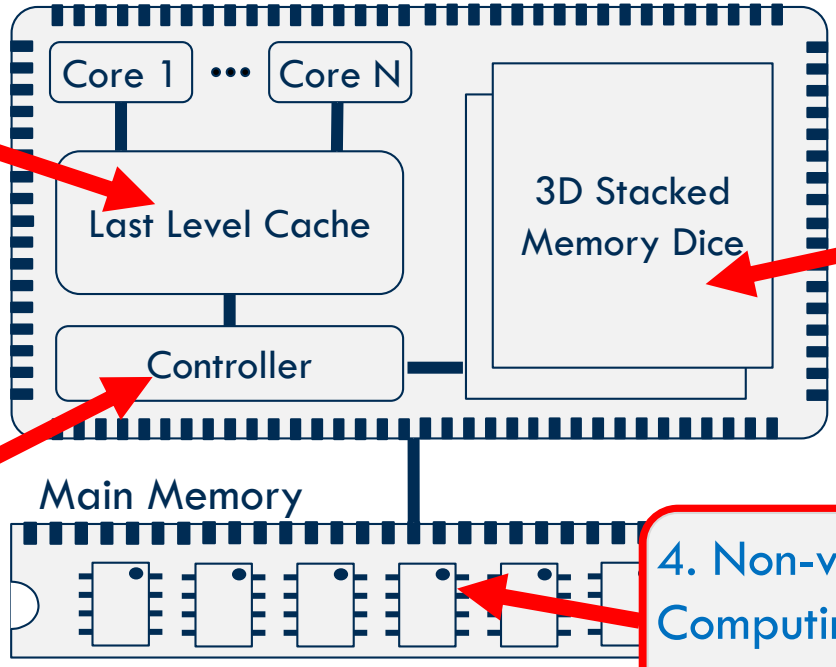
"People who are really serious about software should make their own hardware."

— Alan Kay

Research Examples

- Goal: enable energy and bandwidth efficient data movement between memory and the processor cores.

1. Energy efficient data encoding for large on-die cache



3. Efficient In-Package Memory Systems

2. Bandwidth and Energy Efficient Interface

4. Non-von Neumann Computing In Memory Modules with Emerging Technologies

Memory Bandwidth Demand

- Currently, unfathomable amounts of data generated in various domains

petabytes per hour

US Walmart Supermarkets [dezyre.com]

hundreds of petabytes

Large Synoptic Survey Telescope [lsst.org]

exabytes

Large Hadron Collider [HSF-CWP-2017-01]

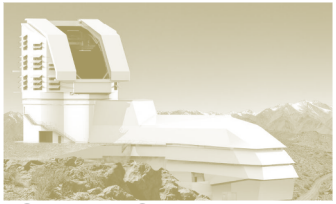
Memory Bandwidth Demand

- Currently, unfathomable amounts of data generated in various domains

petabytes per hour

US Walmart Supermarkets [dezyre.com]

hundreds of petabytes



Large Synoptic Survey Telescope [lsst.org]

exabytes

Large Hadron Collider [HSF-CWP-2017-01]

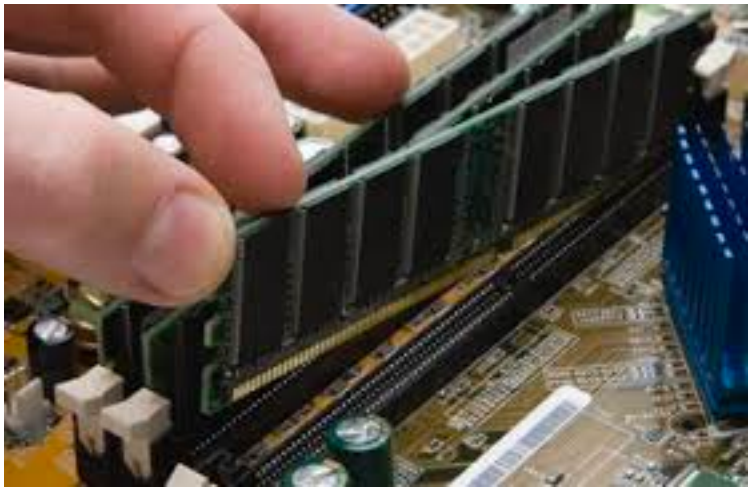
- By 2030, the required bandwidth for fully autonomous self-driving car is expected to reach near 1TB per second.

-- Micron 2018

Emerging Technologies

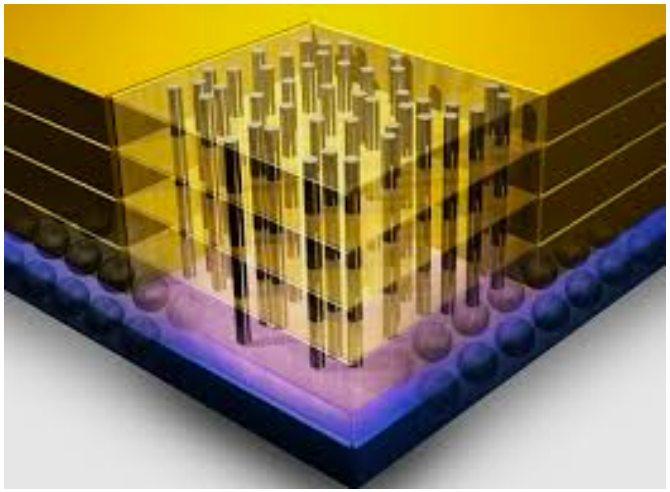
- High bandwidth memory

Off-chip Memory



Lower Bandwidth
Lower Costs

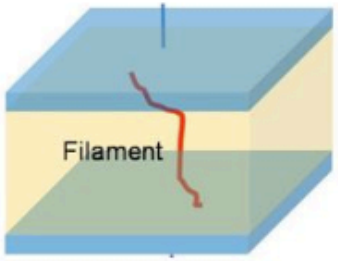
3D Stacked Memory



Higher Bandwidth
Higher Costs

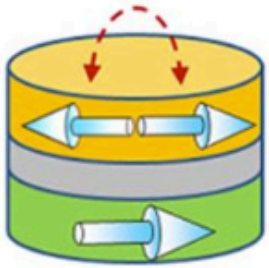
Emerging Non-volatile Memories

- Use resistive states to represent info.
  - Can we build non-von Neumann machines?
    - In-Memory and In-situ computers



RRAM

PCM



STT MRAM

CBRAM