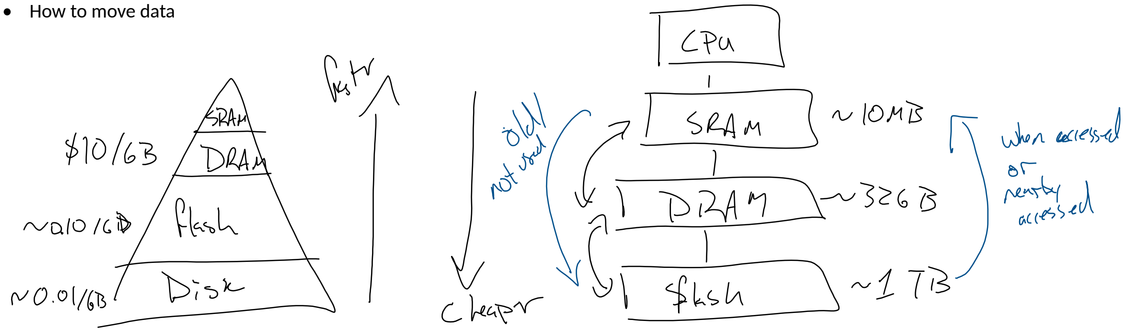
Discussion 7: Hierarchy

Thursday, February 21, 2019 2:34 PM

Outline

- Hierarchy
- Locality
    - Temporal
    - Spatial
- How to move data



how to decide when/what to move?

→ how often accessed

frequent things close

things close to eachother (sequentially)

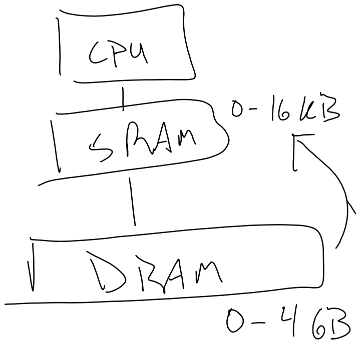
↳ pre load / move together

locality

temporal

spatial

↳ "prediction"



1) try to find data in SRAM

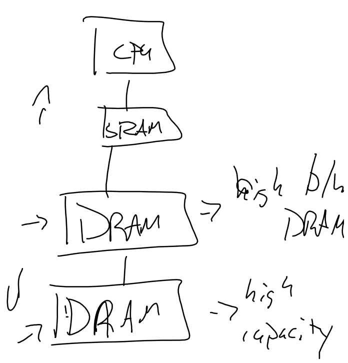
↳ hash table maps DRAM addr to SRAM addr

2) go to DRAM

3) put in SRAM

hardware cache

knights landing



two pools of memory

Let programmer move data from SRAM/DRAM and back

Scratchpad memory or software cache → GPU

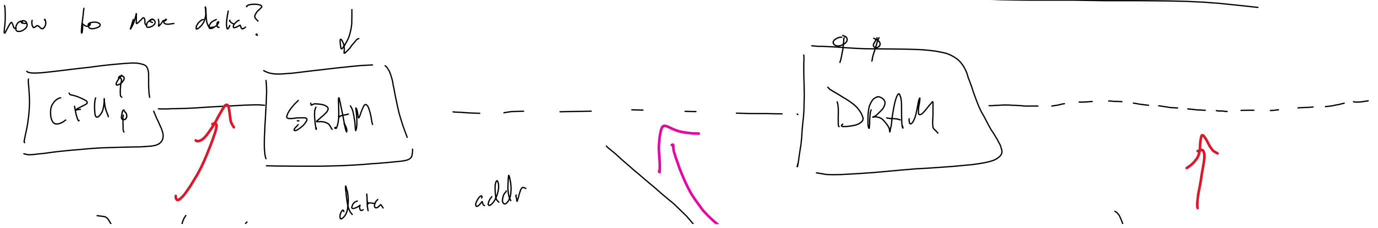
More complex?

hardware cache

↳ lots of metadata to track addresses present + where addresses' data is

↳ tag + valid + state

how to move data?



bus / wires ~64 bit + 32 bit

Parallel up to 512 bit

Short → very fast → @ clock speed

wide

on chip

low power → small wires

Off chip limit ↗

↳ through pins onto fiberglass ~~board~~ board

thick + long wires

↳ higher power

Clock is slower

narrower bus → pretty wide 64 bits →

On motherboard

↳ thick

Serial bus

↳ incr lat

share package pi

low bandwidth

Radio/ microwave interconnect

↳ low latency @ speed of light

↳ interference

↳ expensive ↳ power

↳ convert to e/m waves

Optical interconnect

↳ thicker waveguide

↳ low latency

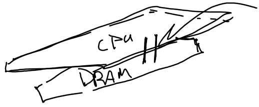
↳ high bandwidth per "pin"

↳ incompatible w/ logic

↳ convert electric signals to light

↳ expensive

3D die stacking



through silicon vias (TSVs)

copper wires like on chip wires

wide busses

high bandwidth

low latency

low power

→ heat

→ design packages

→ new VLSI tools

↳ very expensive