Matthew Tran
CS152
MT1

- ISA - contract between hardware / software
  - often designed w/ microarchitecture in mind
  - ex. hardwire, pipeline, microcode
- Microprogramming
  - good for CISC and when ROM cheaper/faster than RAM
  - split datapath from control
  - state machine, reads from ROM microinstructions
  - [illegible], we use single-bus for RISC-V
  - split instruction into steps
    - instruction CPI > 1
  - Reduce ROM size?
    - [illegible] states, logic → ↓ height
    - encode control signals, [illegible] → ↓ width
  - horizontal microcode - wider, multiple ops per
  - vertical microcode - usually one datapath op per instr
  - nanocoding - combine both [illegible]
    - extra ROM indexed by microcode ROM
  - writable control store (WCS)
    - use RAM not ROM
    - [illegible] failed
  - today most instr hardwired but some complex ones are microcoded
    - [illegible] bad patch/bug fix for microcode
- Pipelining
  - split datapath into steps
  - for RISC-V, 5 stage classic
    - IF, ID, EX, MEM, WB
  - Iron Law of Processor Performance

$$\frac{\text{time}}{\text{program}} = \frac{\text{instructions}}{\text{program}} \cdot \frac{\text{cycles}}{\text{instruction}} \cdot \frac{\text{time}}{\text{cycle}}$$

    ↑ [illegible] ISA, arch | ↑ [illegible] | ↑ [illegible], technology
    ↑ pipeline hopefully reduces!
  - Hazards
    - structural - 2 instr use same resource
    - data - need prev instr data value
    - control - need prev instr addr (ex. branch)
  - Resolve structural hazard
    - stalling
    - [illegible] available w/ more hardware
    - normal 5 stage RISC-V has none
  - data hazard
    - types: RAW, WAR, WAW
    - interlock - hold in issue, stall
    - bypass - send from later to earlier
    - speculate - guess and check
      - [illegible] (branch, stack ptr, memory, [illegible])
  - control hazard
    - bubbles, kill bit (insert NOP)
  - branch delay slot
    - always do instruction after (or NOP)
    - not used anymore
      - ↑ I cache miss from NOP
      - complex
      - use better branch prediction instead
  - why CPI > 1?
    - bypassing expensive
    - [illegible] 2 cycle latency
    - [illegible] bubble
- Traps and Interrupts
  - trap - force transfer control by exception or interrupt
  - exception - internal event [illegible]
  - interrupt - external event
  - precise interrupt - all instr before finish
  - save EPC before [illegible] interrupt [illegible]
  - status register for cause
  - ERET instr to return
  - synchronous trap
    - exception on instruction, needs restart
    - need to undo effect of partial instruction?
    - system call trap is done, ex. ECALL
  - in 5 stage pipeline, hold until commit stage (M)
  - speculating
    - prediction - [illegible]
    - recovery - [illegible] partial instruction before exception [illegible]
- Reduce hazards?
  - re-order code
  - ↓ dynamic instructions, ex. loop unroll
  - decoupling - separate control/memory addr from data [illegible]
    - ex. separate FPU pipeline
- More Pipelining
  - latency -
  - bandwidth
  - occupancy - time unit blocked on op (structural hazard)
  - multiple functional units, [illegible]?
    - stall shorter ones, bypass to keep CPI same
    - Superscalar - issue multiple instructions!
      - ex. one FPU, one integer, do both in [illegible]
- Memory (cache)
  - DRAM architecture
    - rows of word lines } 2D array
    - columns of bit lines
    - 4-8 banks per chip
    - operation 3 steps
      - RAS - [illegible]
      - CAS - [illegible]
      - precharge
    - overall ~15-20ns per step
  - CPU-memory bottleneck
    - latency
    - bandwidth } memory hierarchy! cache, SRAM
    - occupancy
  - Locality
    - temporal - use again soon → store recent
    - spatial - use nearby → group in blocks
  - Direct-Mapped Cache
    - each memory has only 1 location possible
  - n-way set associative
    - split into n sets, can use any one
  - Fully associative
    - no index, all tag, any can hold
  - Replacement Policy
    - random, LRU, FIFO, [illegible]
  - block size larger?
    - ↑ spatial locality
    - ↓ tag size
    - ↑ burst transfer DRAM can be bigger
    - ↑ conflict → waste bandwidth
  - AMAT = hit time + miss rate × miss penalty
    - for 5 stage, [illegible] cache ≥ < 1 cycle hit, ~8-32KB
  - Cache Miss
    - compulsory - first use
    - capacity - too small, happens w/ perfect replacement
    - conflict - miss bc line placement, (don't happen if fully associative)
  - Write Policy
    - hit
      - write through - simpler
      - write back - cache only, dirty bit; [illegible]
    - miss
      - no write allocate - [illegible]
      - write allocate - fetch into cache
  - ↓ write hit time?
    - [illegible] in one cycle, [illegible]
    - pipeline [illegible]
    - [illegible]
  - ↓ read miss penalty?
    - write buffer ([illegible])
      - [illegible]
      - [illegible]
  - cache hierarchy
    - [illegible] L1 in L2, [illegible]
    - [illegible] L1 → [illegible]
    - write through L1! write back L2
  - inclusion policy
    - inclusive - L1 ⊂ L2, [illegible]
    - exclusive - [illegible]
  - Victim cache - [illegible]
    - [illegible]
- Address translation
  - location independent programs - base
  - protection - bound
  - base bound translation - simplest
    - [illegible]
    - [illegible]
  - paged memory
    - addr → page number → offset } [illegible] memory!
    - possible VPN → PPN
    - OS manages free list
    - internal fragmentation still an issue (last [illegible] page)
    - [illegible] page table
  - limited RAM? (paging memory)
    - manual overlay - [illegible]
    - software [illegible] - [illegible]
    - demand paging - RAM is cache for storage
  - linear page table
    - Page Table Entry (PTE) has valid bit, PPN or DPN, [illegible]
    - [illegible]
    - [illegible] to big!
  - hierarchical page table
    - [illegible] memory access! [illegible]
  - TLB (translation lookaside buffer)
    - cache for translations, usually 32-128 entries, fully associative
    - TLB reach - size of VA mapped by TLB
    - miss handler
      - software - simpler, [illegible]
      - hardware - [illegible]
  - [illegible]
- Prefetching
  - hide memory latency??
    - cache
    - [illegible]
    - software prefetch - [illegible] timing
    - speculative execution
  - Little's Law
    - occupancy = latency × throughput
  - hardware prefetching?
    - ex. next line on miss, [illegible]
      - prefetch [illegible]
      - [illegible]
  - metrics
    - accuracy = useful / total prefetch
    - coverage = total / [illegible]
    - timeliness = [illegible] / total prefetch
  - instruction prefetching
    - [illegible]
    - [illegible]
    - [illegible]
    - ex. hybrid [illegible]
    - [illegible]
      - branch target buffer (BTB), branch predictor
  - data prefetching
    - ex. next line - [illegible]
      - stride prefetcher
        - instruction based
        - cache block address based
    - other correlation based
      - [illegible]
    - content based
      - [illegible]
      - [illegible]
  - feedback - [illegible] throttling
- More Virtual Memory
  - virtual cache index?
    - [illegible] TLB [illegible]
    - [illegible] ASID [illegible]
    - aliasing - [illegible] page
    - complex cache coherence
    - TLB [illegible]
  - parallel TLB cache access
    - VA [illegible] PA [illegible]
    - [illegible], multiple [illegible]
      - ex. L2 [illegible]
      - [illegible]
- Extra
  - VIPT - [illegible]
  - [illegible] - [illegible]
  - [illegible] - [illegible]
  - [illegible] - [illegible]
  - [illegible] - [illegible]
  - [illegible] - [illegible]
  - [illegible] - [illegible]

	compulsory	conflict	capacity	hit time	[illegible]	[illegible]
- ↑ line size =	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
- [illegible] =	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
- prefetch =	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
- [illegible] =						

  - improve [illegible] (ISA?)
    - [illegible], CISC, [illegible]
  - improve CPI?
    - pipeline, [illegible], bypassing, branch delay, [illegible], branch predict, prefetching
  - [illegible] (TC) improve?
    - [illegible]
  - [illegible]
  - Page Table Walk (PTW)