What is pipelining?

- Break logic into different stages
- Minimize maximum gate delay to increase clock speed
- Not all operations are created equal
  - e.g. not want to clock machine at speed of FP MUL
- Carry out later stages of one operations while you carrying out earlier stages of another.
- Overlap the execution of multiple instructions

Pros and Cons?

- Pro:
  - Increase clock frequency
    - ideally while maintaining high IPC
    - Greater concurrency
- Con:
  - Increases complexity
  - Increased latency
  - In a microprocessors deeper pipeline equates to high misspeculation penalty

Limits?

- Mis-speculation delay pushes to shallower pipeline stages
- Latch delay limits FO4 depth
- Balancing power/performance limits FO4 ~ 20 in modern microprocessors

What is a branch prediction?

- The next instruction to fetch
- A prediction of the given outcome of a branch
- educated guess of a branch
- Branch predictor: a bit that indicates branch direction
- Branch target buffer: an address
- Return Address Stack: an address

When do we make a prediction?

What can we predict on?

- history: local and global branch history
- instruction type
- register state
- PC

What's the limit of branch prediction?

- finite space to learn with
- warm-up
- entropy