What is a cache (what does it provide?)

- Stores data we hope will be used again
  - Temporal locality
- Latency reduction
- Prefetch useful data
  - Spatial locality
- Bandwidth amplification
  - Buffer nearby writes to multiple addresses
  - Applications have a working set, focus memory system on that

Why does a cache work?

- Memory speed proportional to size and technology
- Applications have locality
  - Temporal
  - Spatial

What is locality?

- Information about the next address accessed is contained in the history of previous addresses accessed
  - Correlation
  - Predictability

Why is there a cache hierarchy?

Why are there separate I/D caches?

- Avoid pollution
- No correlation in the locality of the I and D streams
  - Can predict addresses from one to the other
- I cache is very closely wedded to the I fetch stage
- D cache is wedded to the Load/Store queue
-

Why prefetch?

- Reduce latency
- Idle anyway
- Spatial locality potentially exceeds a cache line
- Reduce cold misses

How to prefetch?

- Look at cache lines for address-like things
- Stride-prefetch (up or down)
- High-semantic instructions
- Use branch predictor state to prefetch I cache
  - fetch both sides of a branch
- prefetch instruction
- prefetch threads

When to prefetch?

- When the memory bus is otherwise idle
  - don't want to slow down accesses you definitely need
- Avoid polluting cache with prefetches that are not used
  - stream buffers
- Not too early and not too late
- Not too much and not too little
  - burns energy to prefetch too much

Why not to prefetch?