Cache Lab Q&A

we use 8-bit addresses
e.g. 01101010

spatial locality means programs will access addresses that differ mainly in their low-order bits

break the address up into 3 parts, starting at bit zero (the right side):
- line offset
- index
- tag

valid bit: does not come out of the address or the data.

It just records whether our cache entry holds cached data or not.
How are threads executed?

1. Run all threads on a single processor by switching between them at some regular interval. Useful for latency hiding, not for performance.

2. On a machine with multiple processors (or cores), run threads across all processors. Use threads to keep multiple cores busy.

3. Use simultaneous multithreading to run more than one thread on a single core. Other names: hardware multithreading, hyperthreading. Use threads to keep all or most functional units in multiple-issue datapath busy.
Cache Coherence

multiple cores with separate caches

Core 0

Core 1

cache

x[2]

x[2]

snoop

invalidate

True Sharing

main memory

| x[0] |

| y[0] |

core 0:
1. read x
2. write x
5. read x

core 1:
3. read x
4. write x
Cache Coherence

multiple cores with separate caches

MESI: modified exclusive shared invalid

Core 0

Core 1

Cache

\[ x \square \ y \square \]

\[ x \square \ y \square \]

\(\odot\) snoop

\(\oplus\) invalidate

\(\odot\) snoop

False Sharing

Cache line

Main memory

\[ x \square \]

\[ y \square \]

core 0:

\(\odot\) read \(x\)

\(\odot\) write \(x\)

\(\odot\) read \(x\)

core 1:

\(\odot\) read \(y\)

\(\odot\) write \(y\)