Nov 9th

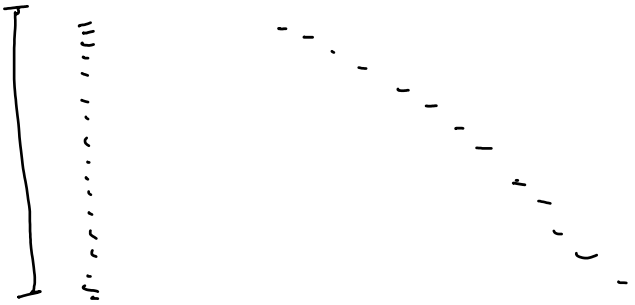
Wednesday, November 9, 2016 1:23 PM

Can we have really long latency L1 cache access?

latency short — less parallelism needed overall.

100 cycles/access

Two extremes: 2 threads to 100 threads



less latency → will require fewer threads to hide the latency.

32 KiB L1 cache. eg.

128B Cache Line size —