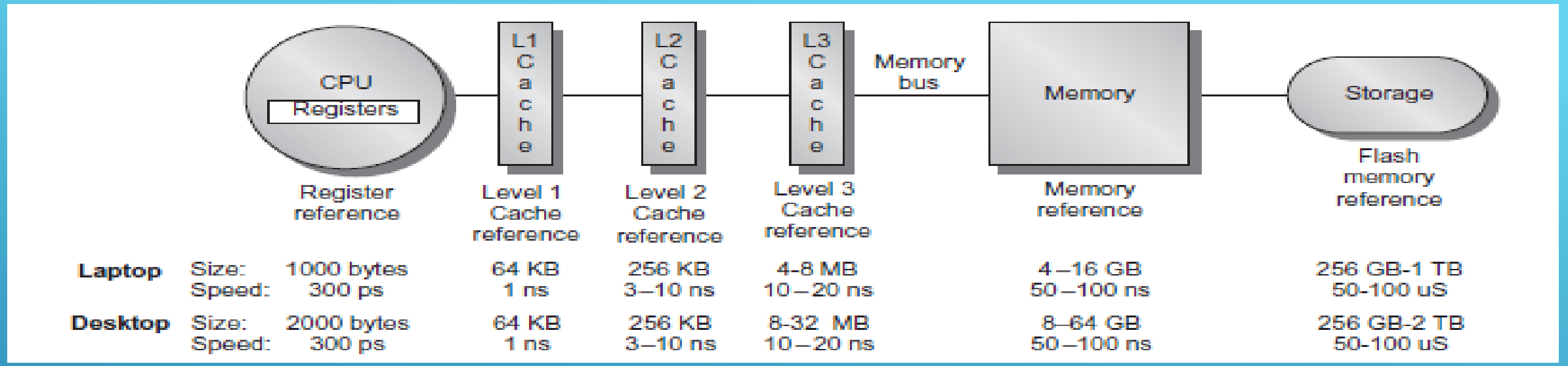
CACHE EVOLUTION

Current cache design

Cache basic fact

- FACTS
  - –BIG IS SLOW
  - –FAST IS SMALL
- INCREASE PERFORMANCE BY HAVING "HIERARCHY" OF MEMORY SUBSYSTEMS
- "TEMPORAL LOCALITY" AND "SPATIAL LOCALITY" ARE BIG IDEAS



- L1 caches are usually on chip, and hence are area-constrained. L2 and L3 caches are implemented off chip, and hence have lesser stringent area constraints.
- L1 caches target lowering the access time, whereas lower level caches target to reduce the miss rates, due to an orders of magnitude higher miss penalties. Note that after the last level cache takes a miss, the processor needs to go to main memory, whose access time is a few hundreds of clock cycles for today's processors that clock GHz frequencies.
- On a multicore system, which is roughly all systems you encounter these days, L1 caches are private to each core, whereas lower level caches are shared. Hence the size difference. L1 size is reported on a per core basis, whereas the reported sizes of L2 and L3 are those of the entire shared chunk of cache memory at the respective levels.

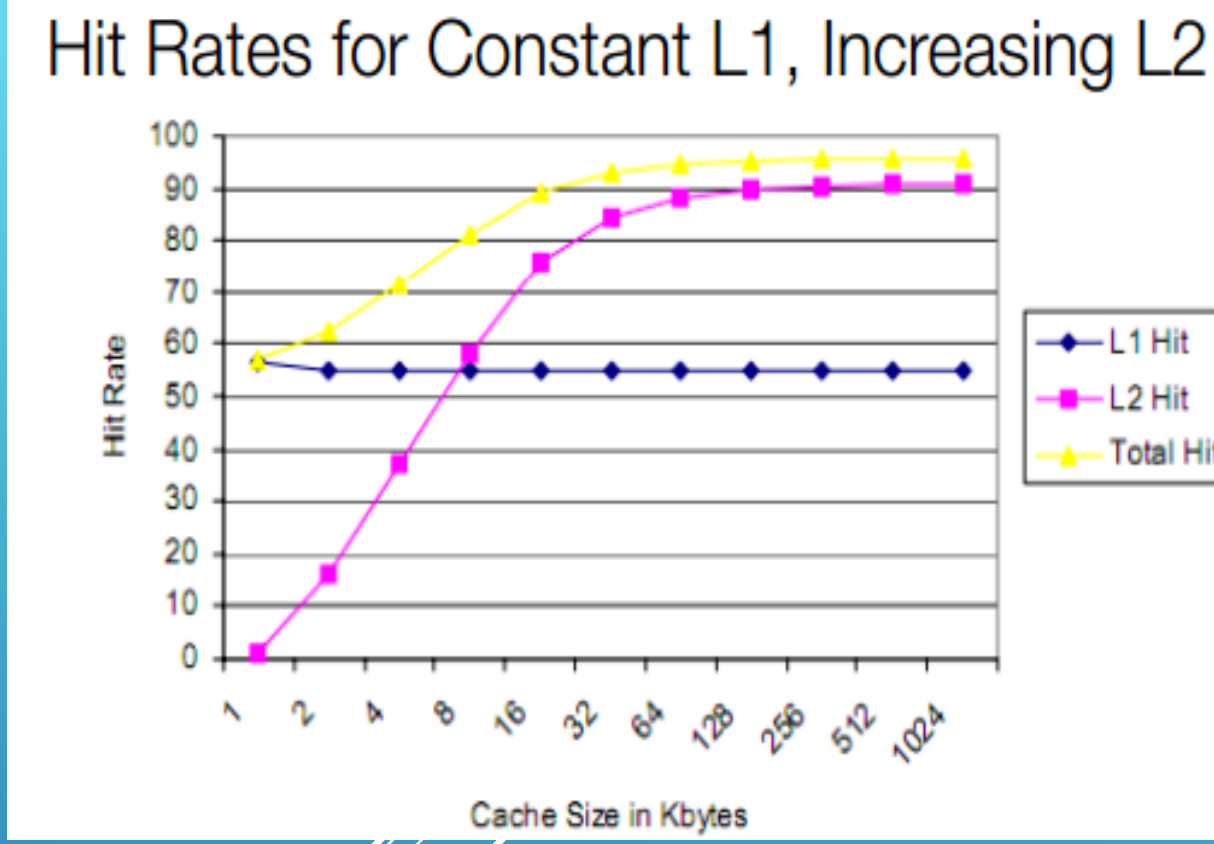
we want to be able to access it quickly (low access time) and we'd also like to hit in the cache as much as possible. Another thing we want is good bandwidth

Memory Size and Distance are certainly key factors.

- Size of L3 cache > Size of L2 cache > Size of L1 cache,

which implies, the time to search and get a cache hit is inversely proportional to the size of the cache. If T is the time to search for a cache line and bring it into the processor,

- T (L3) > T (L2) > T(L1)



THANK YOU

Rohit Kumar