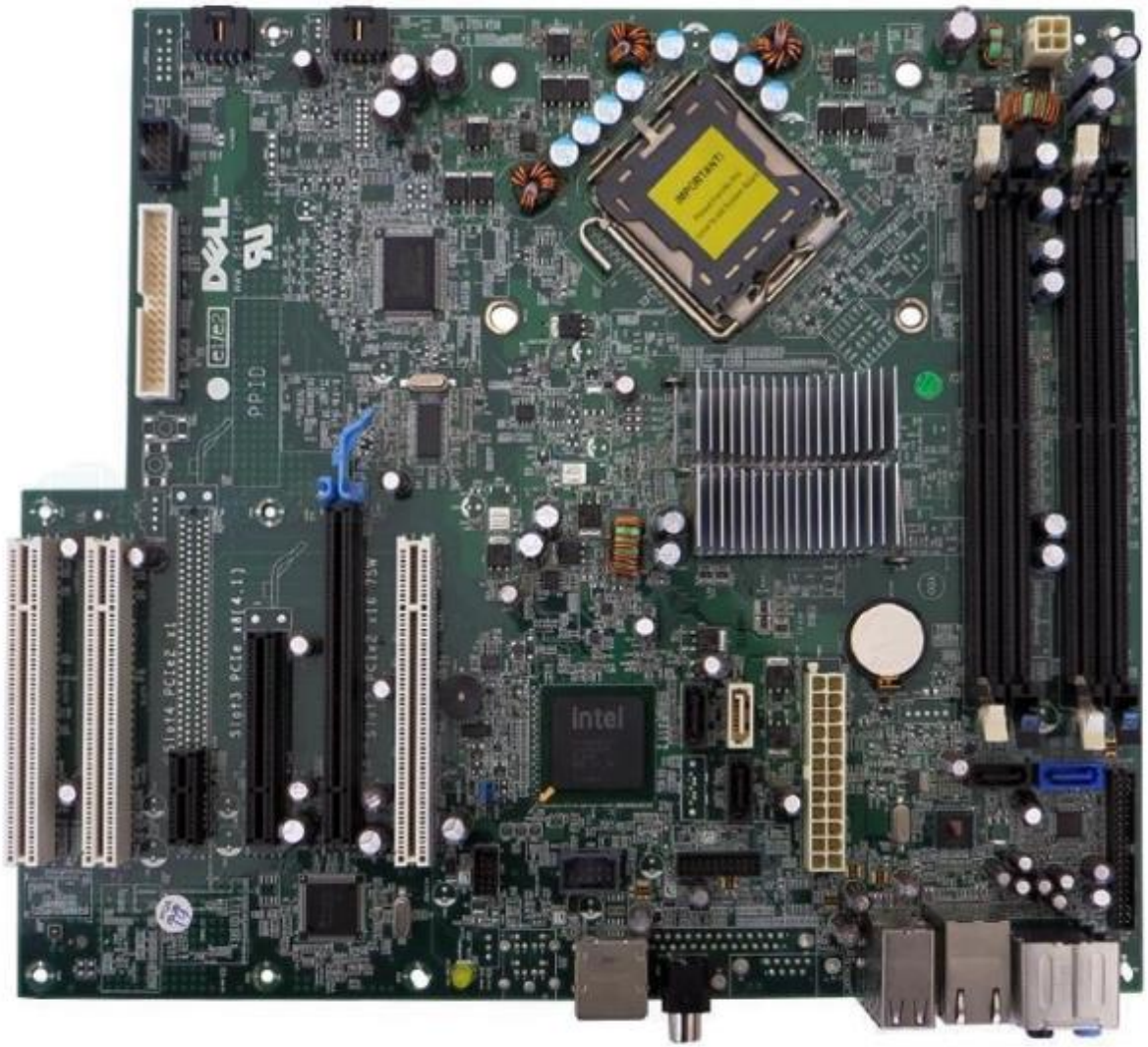
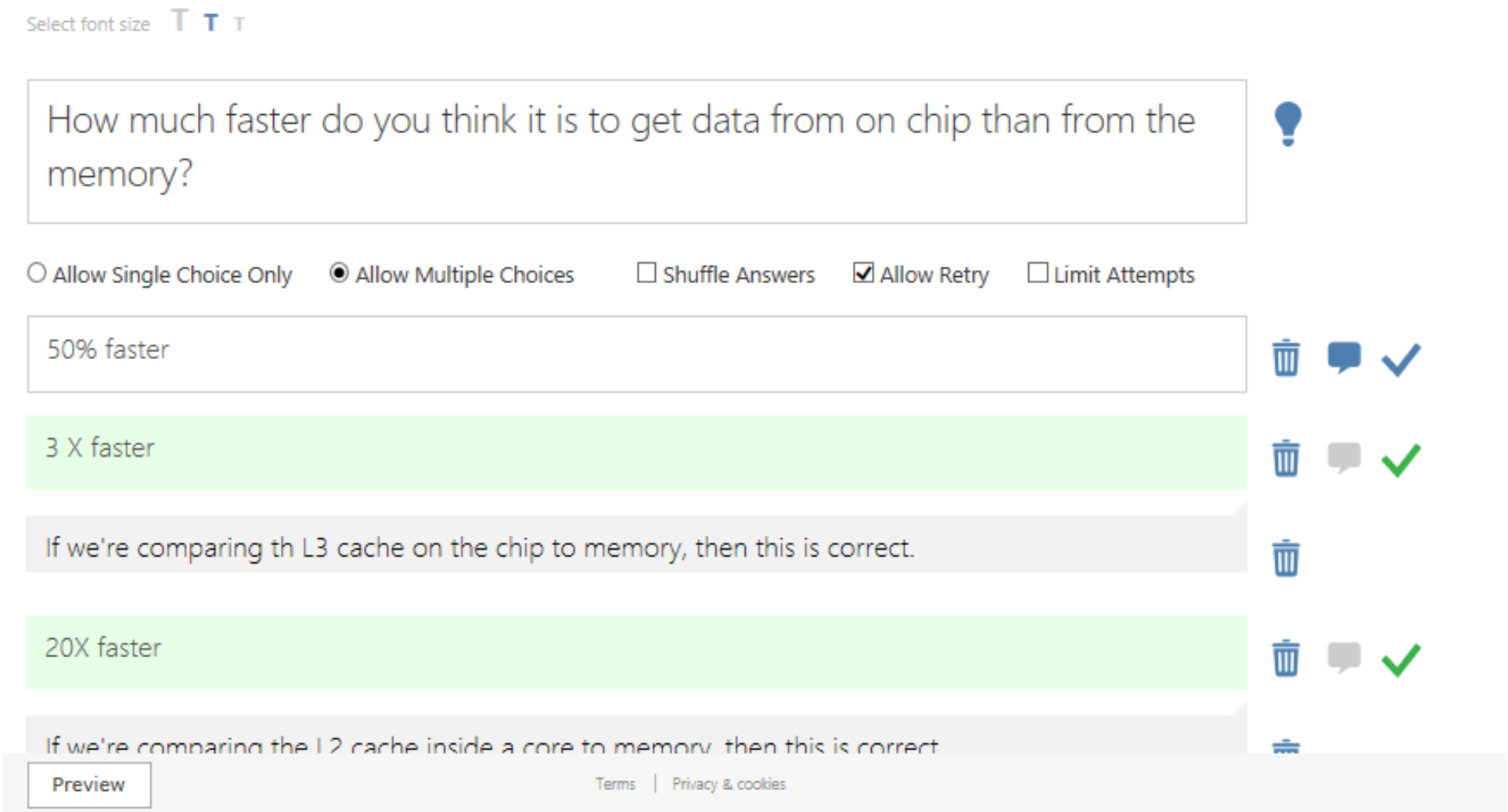
Unit 3: Architecture, the Memory Hierarchy, and Caching

- Learning Objectives (unit)
  - Leverage caching to overcome the differences in performance available at different levels of the memory hierarchy.
- Learning Objective (this video)
  - Define memory hierarchy.
  - Evaluate the performance differences found at the different layers of the hardware memory hierarchy.
  - Explain the different kinds of caching that processors and hardware systems perform to mitigate the performance differences between the levels of the memory hierarchy.

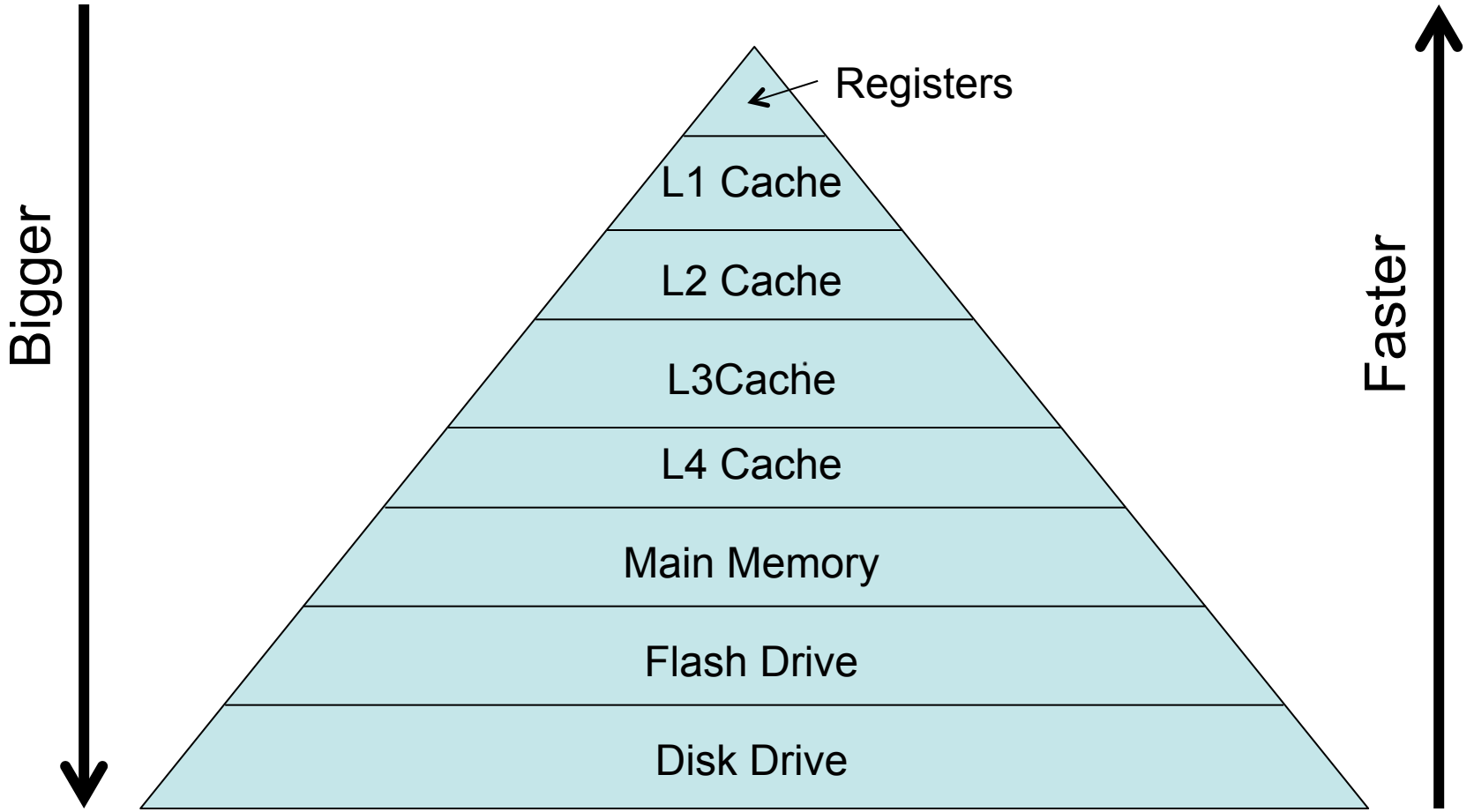

DELL
intel


CPU
DELL
intel

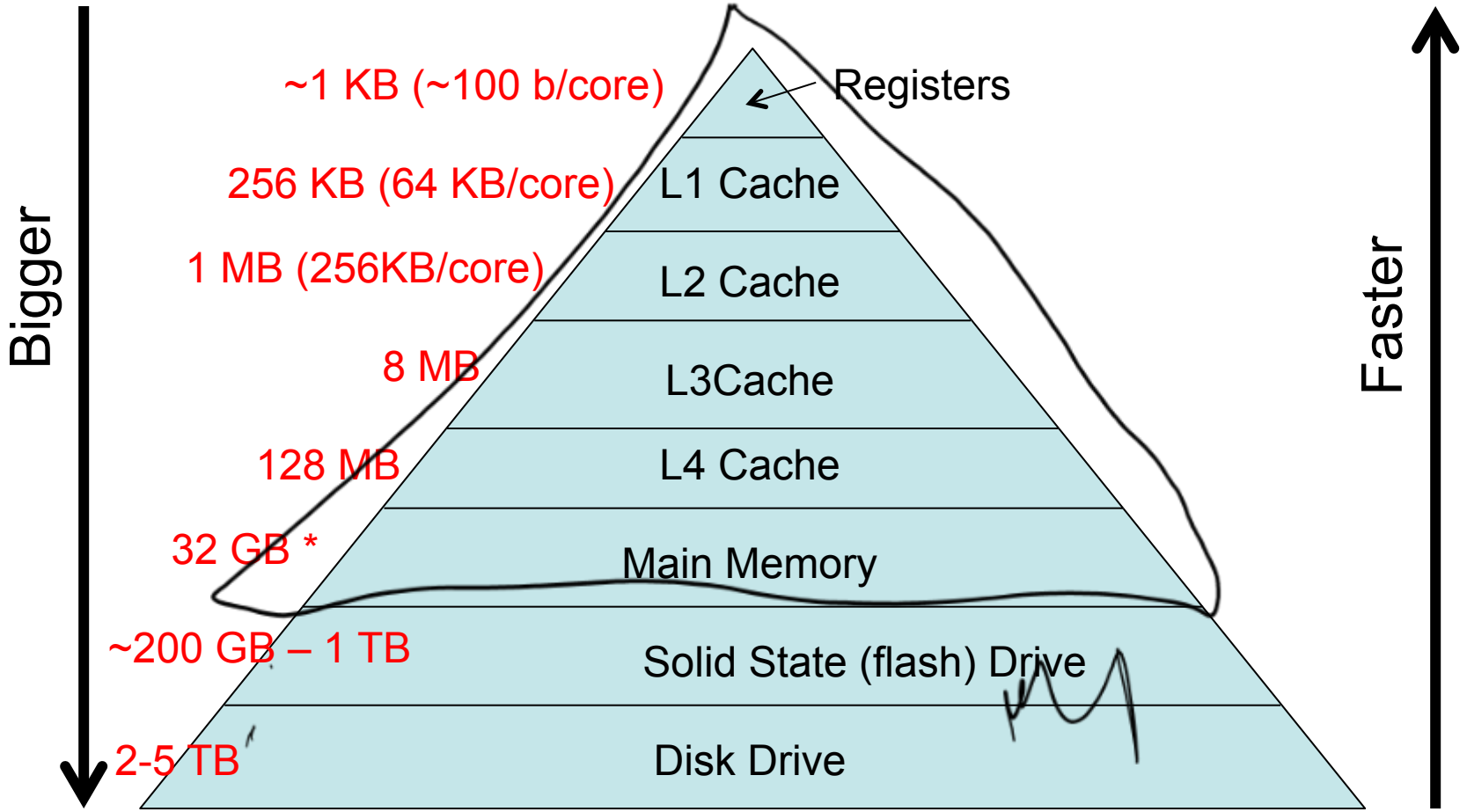

CPU
Memory Slots


Select font size T T T
How much faster do you think it is to get data from on chip than from the memory?
Allow Single Choice Only
Allow Multiple Choices
Shuffle Answers
Allow Retry
Limit Attempts
50% faster
3 X faster
If we're comparing th L3 cache on the chip to memory, then this is correct.
20X faster
If we're comparing the L2 cache inside a core to memory, then this is correct
Preview
Terms | Privacy & cookies

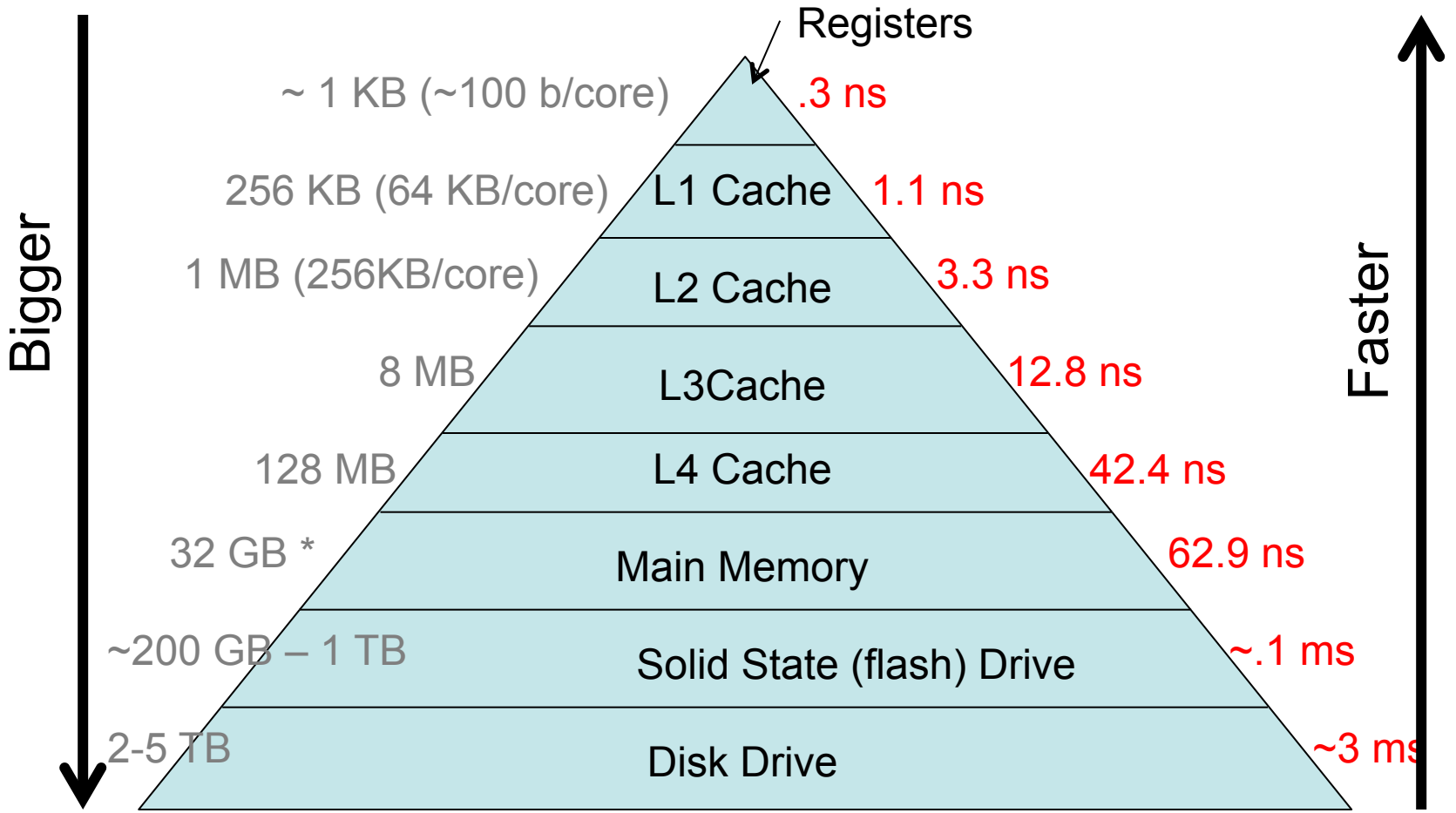
The Memory Hierarchy



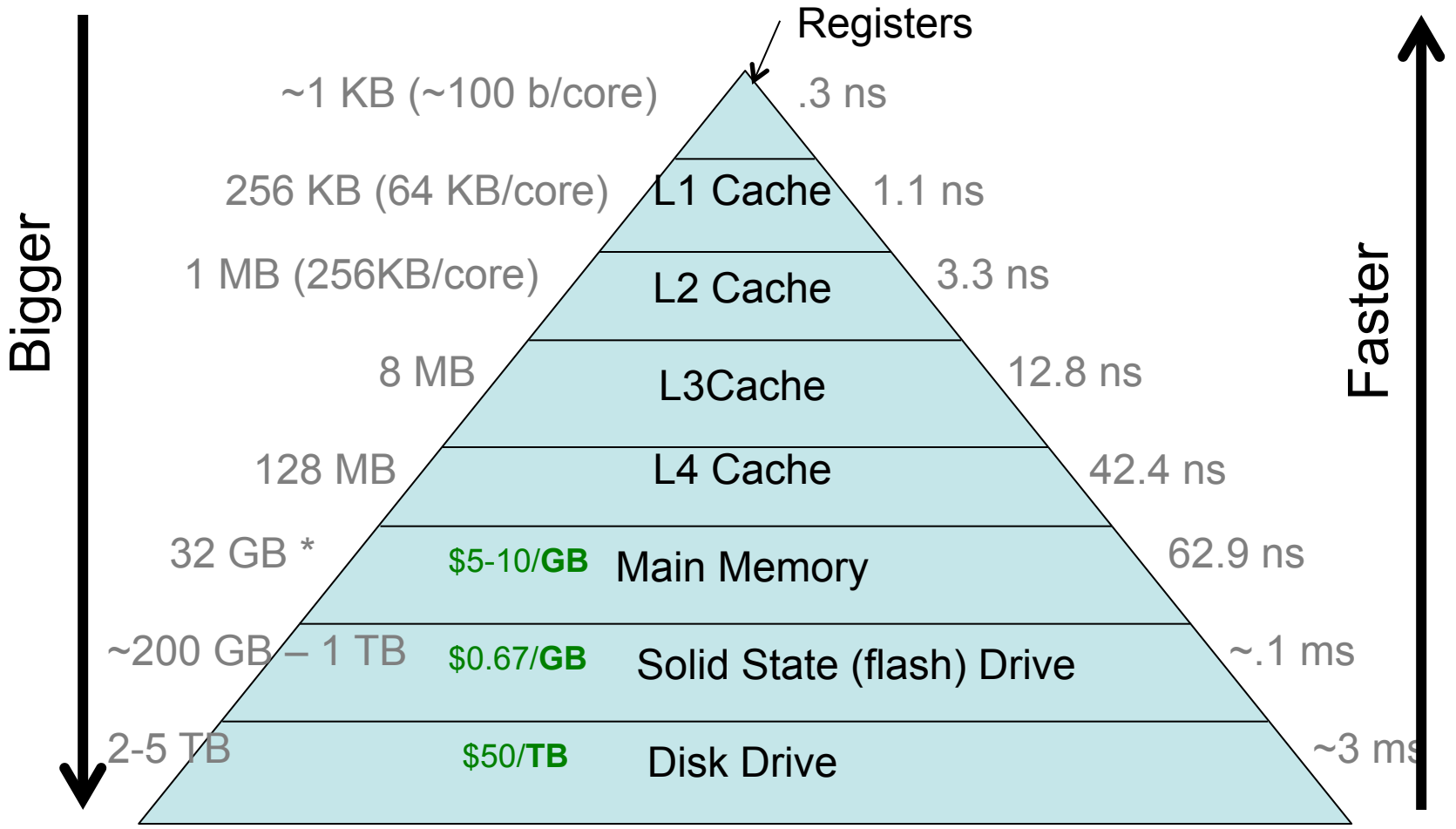
The Memory Hierarchy -- Size



The Memory Hierarchy -- Speed



The Memory Hierarchy -- Price



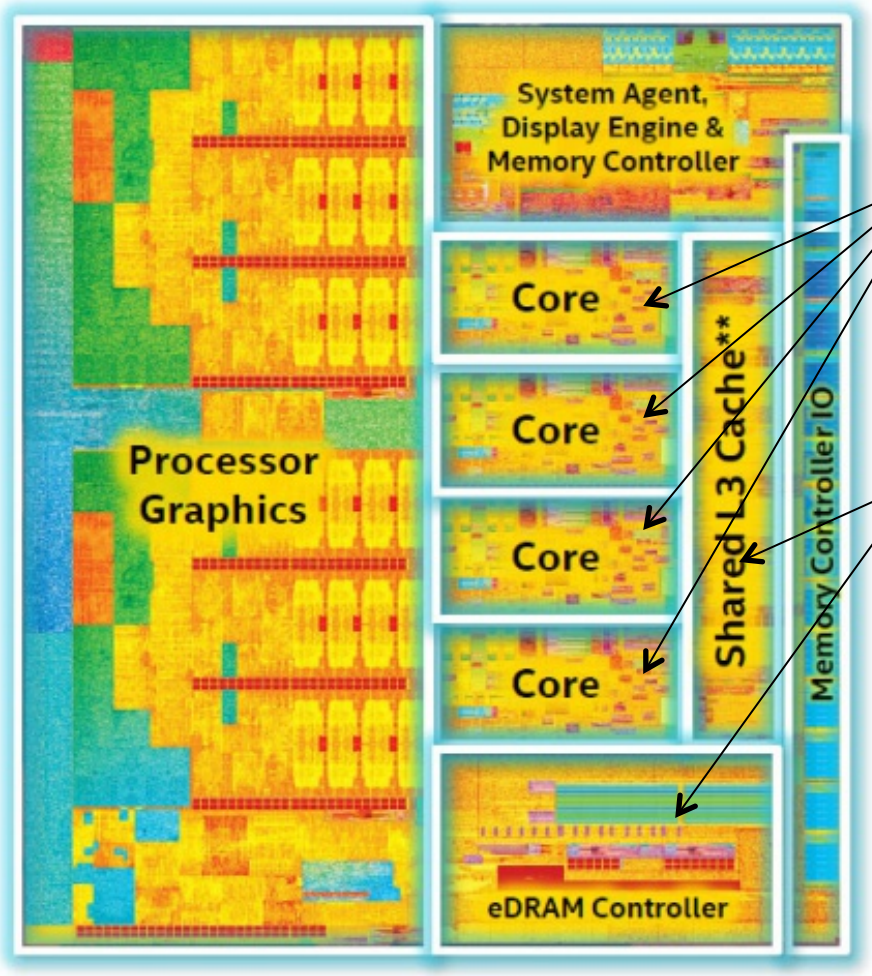
Implications

- We saw many factors of 10 or 100 in:
  - Size
  - Performance
  - Price
- "When you see a factor of 100, it's going to affect how you program." – E. Kohler
- As the ratios between different parts of the system change, so do our priorities.
  - 1956:
    - $/MB(mem) : $/MB(disk)=> $411M : $9200 => 44,673 X
  - 2015:
    - $/MB(mem): $/MB(disk) => $0.005 : 0.0000317=> 158 X
  - Cost of memory in 1956 versus 2015:
    - 1956 $ / MB : 2015 $ /MB => $411M/.005 => 82 trillion …

Caching

- Definition:
  - Colloquially: store away in hiding or for future use
  - Applied to computation:
    - Placing data somewhere it can be accessed more quickly.
- Examples:
  - In assembly language: we move data from memory to registers.
  - In the hardware: we move data from main memory into memory banks that live on the processor (more on this in a moment).
  - In software: we read things into our program's local buffers and manipulate them there.

Processor Caches



L1 and L2 are per core

L3 and L4 are shared
Across cores

Intel Core i7-5775C

Wrapping Up

- Caching is ubiquitous throughout our computing systems:
  - In the processor
  - In the operating system
  - In databases
  - In middleware
  - In applications
- Writing efficient and correct software requires a deep understanding of caching and its implications.