Computer Systems Organisation (CS2.201)

Summer 2021, IIIT Hyderabad

14 July, Wednesday (Lecture 23) – Memory Hierarchies

Taught by Ziaul Choudhury

Caches (contd.)

Structure of the Cache

The cache consists of a storage unit and a controller. The controller decides what is to be stored in the storage unit.

Fundamental parameters	
Parameter	Description
$S = 2^s$	Number of sets
E	Number of lines per set
$B = 2^b$	Block size (bytes)
$m = \log_2(M)$	Number of physical (main memory) address bits

Derived quantities	
Parameter	Description
$M = 2^m$	Maximum number of unique memory addresses
$s = \log_2(S)$	Number of *set index bits*
$b = \log_2(B)$	Number of *block offset bits*
$t = m - (s + b)$	Number of *tag bits*
$C = B \times E \times S$	Cache size (bytes) not including overhead such as the valid and tag bits

Figure 1: Cache Parameters

When memory has to be updated, there are two general principles – write-back (where the block updated in the cache until it is removed, at which point it is updated in the main memory) and write-through (where the updates happen simultaneously in cache and main memory).

There are two types of caches – direct map caches (where the set bits of an address dictate the position of the block in the cache) and fully associative caches (which can store blocks in any empty positions in the cache).