Computer Design (2004) Paper 3 Question 1 (SWM)

Sketch Answer

(a) A RAM stores data in a contiguous block of addresses. In order to perform a lookup the address is decoded into a row select signal which specifies which row in memory the data is in. Typically several bytes of data are returned on the column bit lines, and the remaining part of the address selects between these bit lines. A CAM on the other hand is a sparsely addressed memory. Address, data pairs are stored. Data is accessed by presenting the appropriate address which is compared in parallel with the addresses stored in the CAM. If a match is found, the appropriate data value is returned. This is a form of fully associative memory.

(b) A fully associative mapping allows (address,data) pairs to be stored anywhere in the memory. A direct mapped lookup relies on taking a hash of the address which is then used as an address into a RAM. A set associative cache combines direct mapped and fully associative features. There are two forms of set associative cache. The first performs a hash function in order to determine which of several blocks of fully associative cache are to be used for the lookup. The second relies on a set of direct mapped caches, each of which are accessed in parallel with the expectation that just one (or none) will return valid data.

(c) A TLB is a small cache which stores recent address translations. Unlike a cache line in a data or instruction cache, each TLB entry stores a translation for an entire page of memory. Thus, just a few TLB entries covers quite a lot of memory. Consequently the TLB can be small (e.g. 64 entries) whilst still performing well.

(d) A small fully associative memory is used for the TLB. Since it is so small, any size overhead due to the associative lookup logic is not an issue. It might be possible to have a direct mapped or set associative cache but these caches have nasty pathological cases when access patterns cause the caches to continually spill and refill the cache.