SOLUTION NOTES

Computer Design 2001 Paper 6 Question 2 (SWM)

(a) What is a data cache and what properties of data access does it exploit? [5 marks]

Ans: A data cache is a small fast local memory which stores recently used results (temporal locality) and results which are neighbor of recently used results which may be used soon (spatial locality).

(b) What is a direct mapped cache and under what conditions will it exhibit poor performance? [5 marks]

Ans: A direct mapped cache determines where to store data using a simple hash function (e.g. some subset of the bits of an address). Thus, data corresponding to a particular address can only be stored in one location. This cache will exhibit poor performance when two streams of data being processed map to the same area of the cache which results in data being continually spilled and refilled. For example, consider a simple memory copy routine:

```
  for(i=0; i<1024; i++) x[i]=y[i];
```

If beginning of arrays x and y map to the same place in a direct mapped cache then accessing writing the first item to x will result in y being displaced from the cache. Thus the cache will have to be spilled and refilled to fetch the second item of y.

(c) Under what circumstances might a word of data in main memory be simultaneously held in two separate first level cache lines? [5 marks]

Ans: If the cache works with virtual (rather than physical) addresses then shared data mapped to the same physical address but with more than one virtual address may be incorrectly mapped to two separate spaces in the cache.

(d) A translation look aside buffer is a specialised cache. What does it typically store and why is it often a factor of 1000 smaller than a data cache? [5 marks]

Ans: A TLB caches recent virtual to physical address translations.

Given that address translation is page based, each entry in the TLB covers quite a large area of memory (e.g. 8kbytes). Consequently, few TLB entries are required to cover the working set of an application.