1998 Paper 8 Question 4

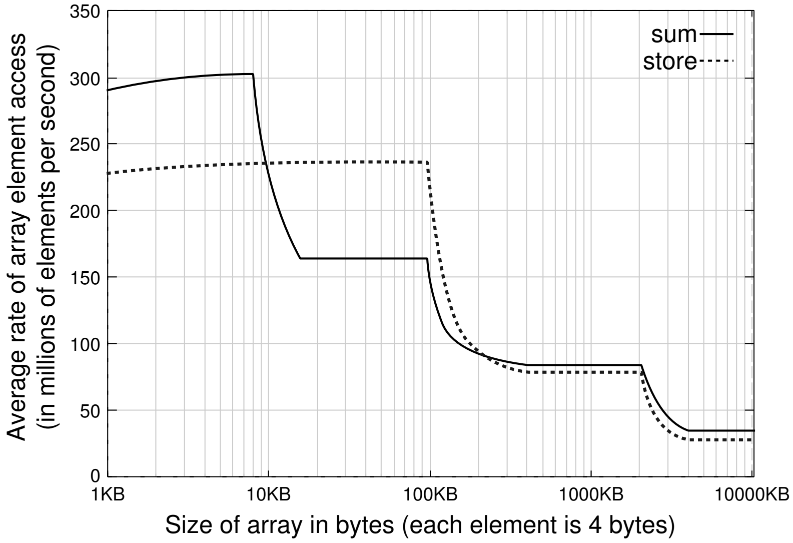
Comparative Architectures

Modern workstations typically have memory systems that incorporate two or three levels of caching. Explain why they are designed like this. [4 marks]

In order to investigate the performance of a system's memory hierarchy, a user writes a simple test program. The program, known as `sum`, accepts N as an input parameter, and allocates an array of N 32-bit integer elements in physically contiguous memory. `sum` contains an inner loop that scans sequentially over the array and computes the sum of all the elements. The program measures the total time taken to execute several thousand iterations of the scan, and uses this to compute the average rate at which the computation proceeded.

Write sample assembly code for a performance-optimised implementation of `sum`'s inner loop for a super-scalar RISC processor. You may assume that the array always contains a multiple of eight elements, and you are encouraged to demonstrate your knowledge of techniques such as loop unrolling and instruction scheduling. Indicate how your loop might execute on a processor capable of issuing two integer operations per cycle. [8 marks]

In addition to `sum`, the user writes a similar program called `store`. Instead of totalling the array, `store` simply writes to each element, setting its contents to zero. The user invokes the programs over different sizes of array in order to generate the following graph. The x axis indicates the size of the array (in bytes) that the program was operated over, while the y axis shows the average rate at which array elements were processed (in millions per second).



Ignoring startup effects, describe the behaviour of the memory system, and hence account for the graph. What can you deduce about the workstation's memory hierarchy? [8 marks]