COLBY COLLEGE
Department of Mathematics
and Computer Science

CS357
Spring 1998

Allen B. Downey

Assignment 1 Solutions

Question 2.1

I don't remember why I thought this was a good question, but now it seems to me that I have no idea how to compare prefetching and spooling, since they are completely different things. But I can define each of them, so that's what I'll do.

Prefetching is a general technique, used up and down the memory hierarchy, for bringing data items into cache (any cache) before they are asked for, on the assumption that they will be used soon.

Having a block size larger than one is a form of prefetching, specifically tailored to take advantage of spatial locality.

Spooling means doing the I/O of one job at the same time as the computation of another job. It can refer to any kind of I/O, including disk accesses, printing, etc.. Spooling can improve the utilization of both the CPU and the I/O systems. The term is not used much any more — it went the way of multiprogramming systems.

Question 2.9

Caches are useful if you have multiple storage mechanisms with different speeds, sizes and costs. In general, you use a small, fast, expensive thing as a cache for a large, slow, cheap thing. If things go well, you get the illusion of a large, fast thing.

Caching only works if your access pattern has the property of locality, either temporal or spatial.

The problems associated with caching are reliability and consistency. The relibility issue is especially relevant when you use memory as a cache for disk. If you write something on disk, you expect it to be permanent, but if it is actually cached in memory, it is volatile. The consistency issue is that more than one person, or process, or processor, or whatever, might cache the same data item. In that case, changes to one of the cached copies would not be reflected properly in the other.