11 BONUS: Data Prefetching [50 points]

You and your colleague are tasked with designing the prefetcher of a machine your company is designing. The machine has a single processor attached to a main memory (DRAM) system.

You need to examine different prefetcher designs and analyze the trade-offs involved. For all parts of this question, you need to compute the *coverage* or *overhead* of the prefetcher in its **steady state**.

You run an application that has the following memory access pattern (note that these are cache block addresses). **Assume this memory access pattern repeats for a long time.**

$$A,\ A+1,\ A+9,\ A+10,\ A+18,\ A+19,\ A+27,\ A+28,\ A+36,\ A+37, ...$$

(a) [10 points] You first design a stride prefetcher $Pref_X$ that observes the last three cache block requests. If there is a constant stride S between the last three requests, $Pref_X$ issues a prefetch to the next cache block using the stride S. In absence of a constant stride, $Pref_X$ refrains from prefetching. What is the coverage of $Pref_X$ for the application? Show your work. Please recall, prefetcher coverage is defined as:

$$\frac{Total\ number\ of\ prefetch\ requests\ used\ by\ the\ program}{Total\ number\ of\ main\ memory\ requests\ without\ the\ prefetcher}$$

0%

Explanation: Since the stride in the address pattern is changing between +1 and +8, the stride prefetcher $Pref_X$ cannot learn any constant stride to issue prefetch requests.

(b) [10 points] You then design a next-N-block prefetcher $Pref_Y$. For every memory access to cacheline address A, the $Pref_Y$ prefetches addresses $A+1, A+2, ..., A+N$. What is the coverage of $Pref_Y$ if you set $N = 2$?

50%

Explanation: $Pref_Y$ will prefetch $A+1$ by seeing A, $A+9$ by seeing $A+8$, and so on. Hence every alternate memory requests will be sucessfully prefetched.

(c) [10 points] A prefetcher also incurs bandwidth overhead to the system. We define a prefetcher's bandwidth overhead to the the system as:

$$\frac{Total\ number\ of\ main\ memory\ requests\ with\ the\ prefetcher}{Total\ number\ of\ main\ memory\ requests\ without\ the\ prefetcher}$$

Please note that, if multiple prefetch requests are generated for one memory address, *only one* request goes to the DRAM.

What is the bandwidth overhead of $Pref_Y$ when $N = 2$? Show your work.

3/2

Explanation:
For $Pref_Y$:

- A will prefetch addresses $A + 1, A + 2$
- $A + 1$ will prefetch addresses $A + 2, A + 3$

So, for every 2 unique cache block requests without the prefetcher, there are 3 unique cache block requests with the prefetcher $Pref_Y$. Hence the bandwidth overhead is $\frac{3}{2}$.

(d) [10 points] What is the minimum value of N required to achieve a 100% prefetch coverage for $Pref_Y$? Show your work. Remember that you should consider the prefetcher's coverage in its steady state.

8

Explanation: At $N = 8$, $A + 1$ can prefetch for $A + 9$, thus acheiving 100% coverage.

(e) [10 points] What is the bandwidth overhead of $Pref_Y$ at the value of N you find for part (d)? Show your work.

9/2

Explanation:
For $Pref_Y$ at $N = 8$:

- A will prefetch addresses $A + 1, A + 2, A + 3, A + 4, A + 5, A + 6, A + 7, A + 8$
- $A + 1$ will prefetch addresses $A + 2, A + 3, A + 4, A + 5, A + 6, A + 7, A + 8, A + 9$

So, for every 2 unique cache block requests without the prefetcher, there are 9 unique cache block requests with the prefetcher $Pref_Y$. Hence the bandwidth overhead is $\frac{9}{2}$.