Lecture-5
(10K feet view: Hardware Prefetching)
CS422-Spring 2019

Biswa@CSE-IITK

Hardware Prefetching

What?

Latency-hiding technique - Fetches data before the core demands.

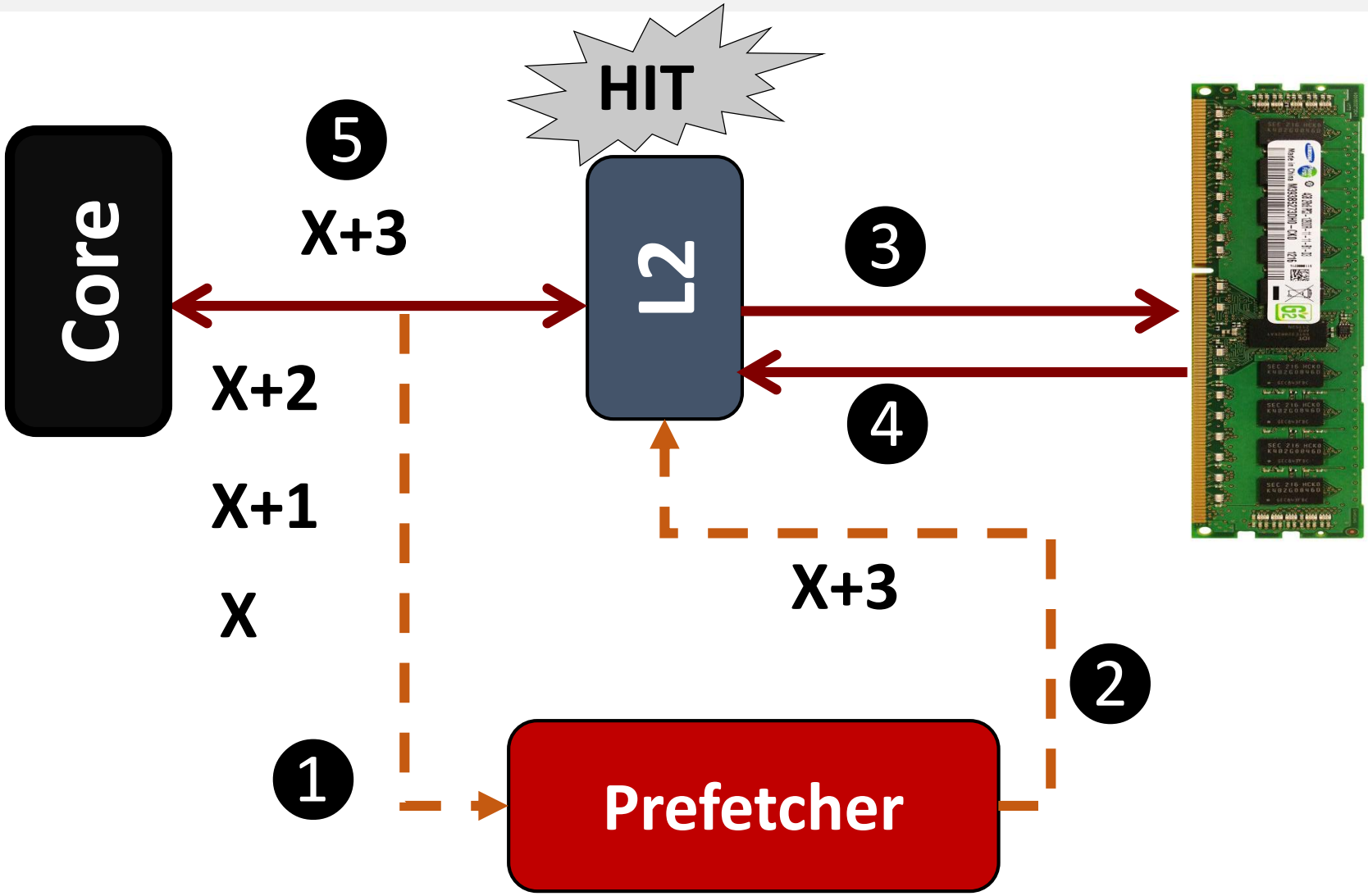
Why?

Off-chip DRAM latency has grown up to 400 to 800 cycles.

How?

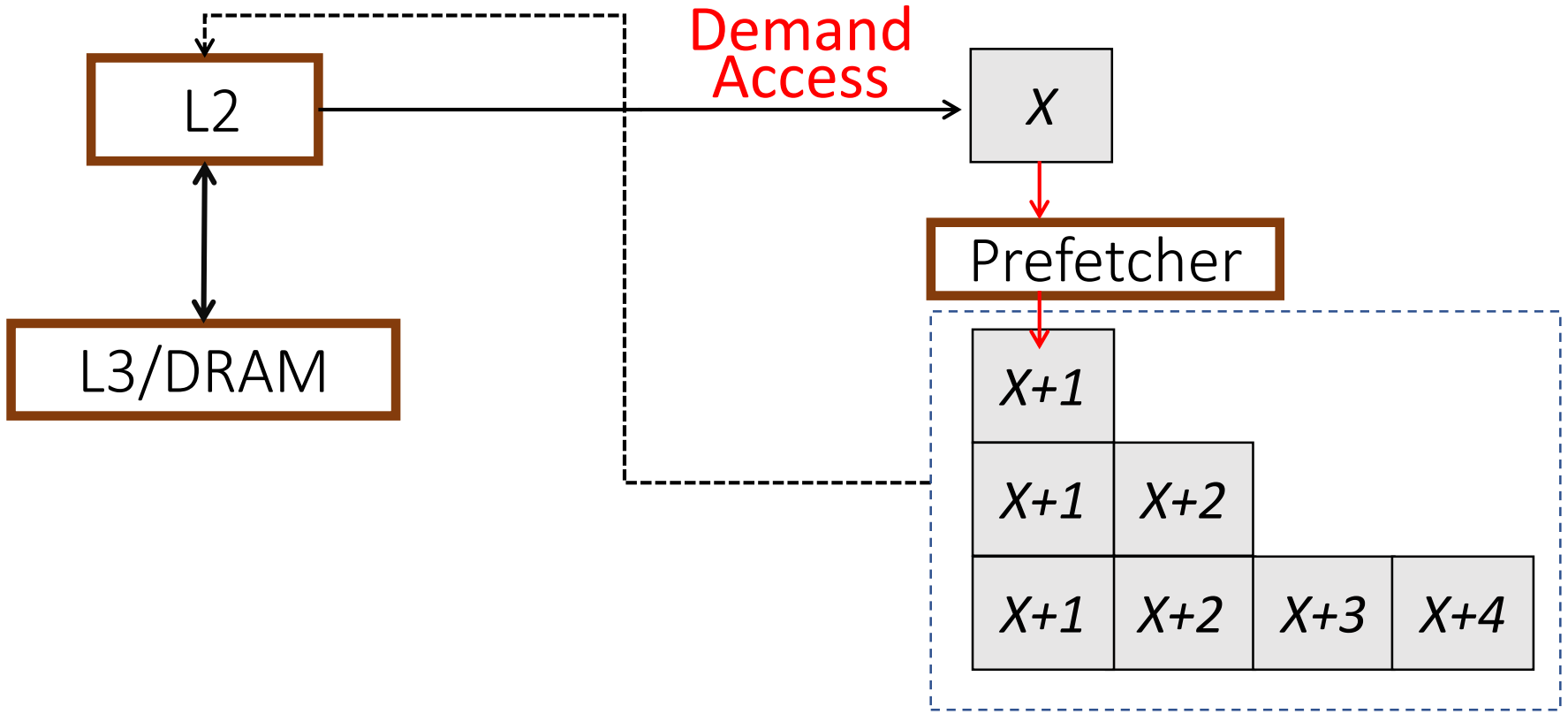
By observing/predicting the demand access (LOAD/STORE) patterns.

Prefetch Engine



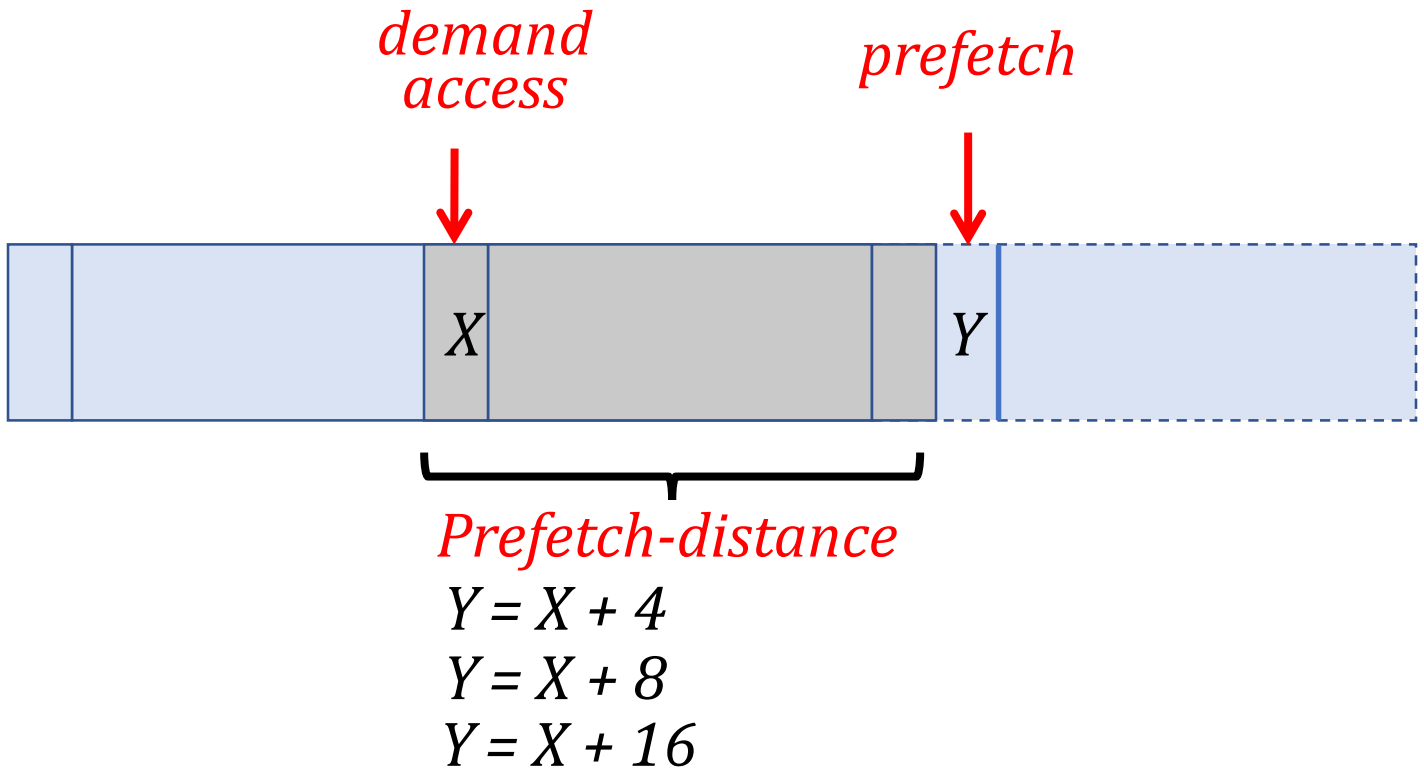
Prefetch Degree

Prefetch Degree: Number of prefetch requests to issue at a given time.



Prefetch Distance

Prefetch Distance: How far ahead of the demand access stream are the prefetch requests issued?

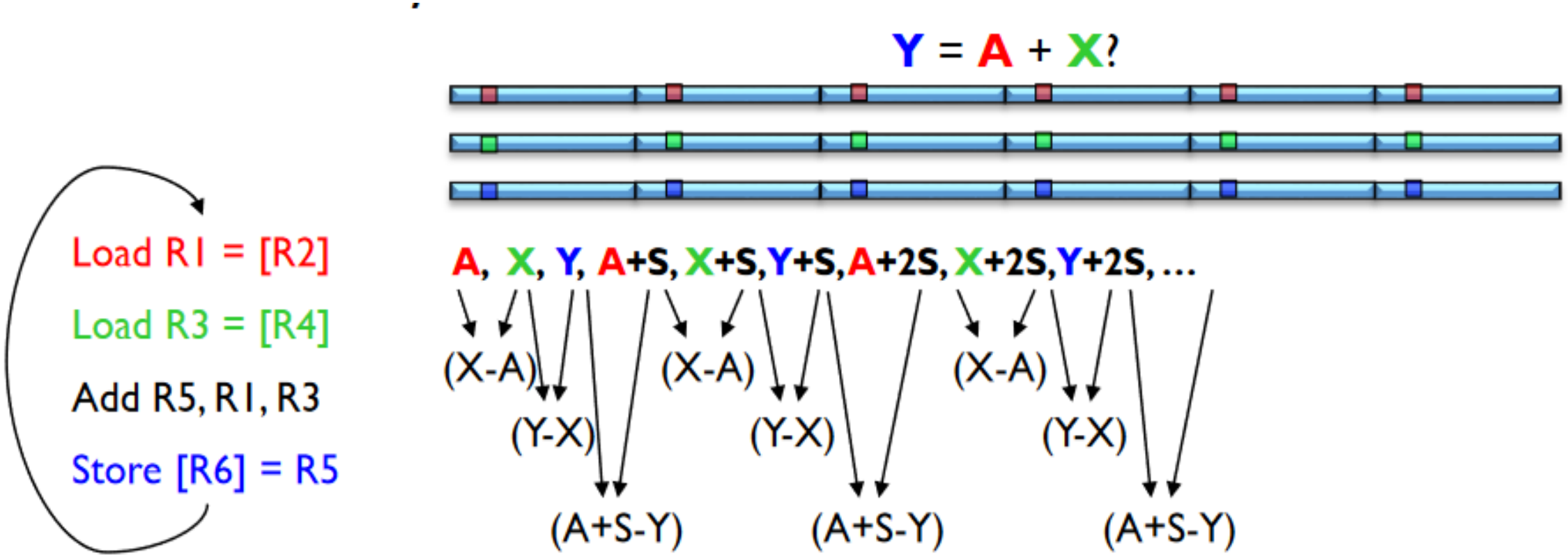


Next-line Prefetcher

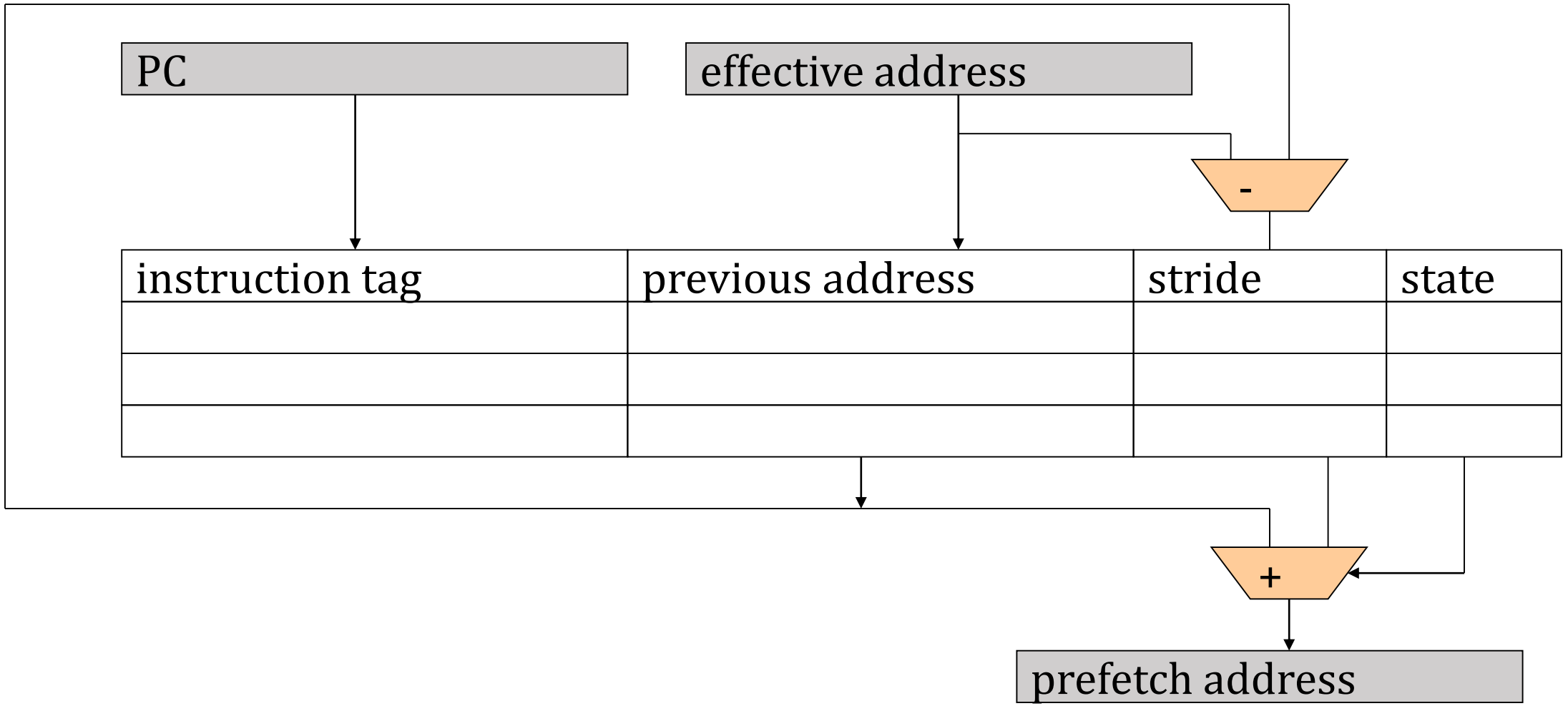
Next Line: Miss to cache block X , prefetch X+1. Degree=1, Distance=1

Works well for L1 Icache and L1 Dcache.

What About This?



Stride Prefetching



Quantifying Prefetchers

$$\text{PrefetchAccuracy}(i) = \frac{\text{Prefetch}_{\text{hits}}(i)}{\text{Prefetch}_{\text{issued}}(i)}$$

Prefetched Block in the Cache.

$$\text{Lateness}(i) = \frac{\text{Prefetch}_{\text{late}}(i)}{\text{Prefetch}_{\text{hits}}(i)}$$

Prefetched Block Still on its way

$$\text{Pollution}(i) = \frac{\text{LLC Poll}(i)}{\text{Demand}_{\text{misses}}(i)}$$

Prefetched Block evicted a demand block that will be reused

$$\text{Coverage}(i) = \frac{\text{Prefetch Hits}(i)}{\text{Prefetch Hits}\,(i) + \text{Demand}_{\text{misses}}(i)}$$

Fraction of misses avoided