CS698Y: Modern Memory Systems
Lecture-11 (Hardware Prefetching)

Biswabandan Panda

biswap@cse.iitk.ac.in

https://www.cse.iitk.ac.in/users/biswap/CS698Y.html

Flow of the Module

Data Prefetching Techniques

Metrics Related to Prefetching

Interaction with Cache Replacement

Instruction Prefetching

But, Why Prefetching? Remember Memory Wall: It is still hurting

Hardware Prefetching

What?

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

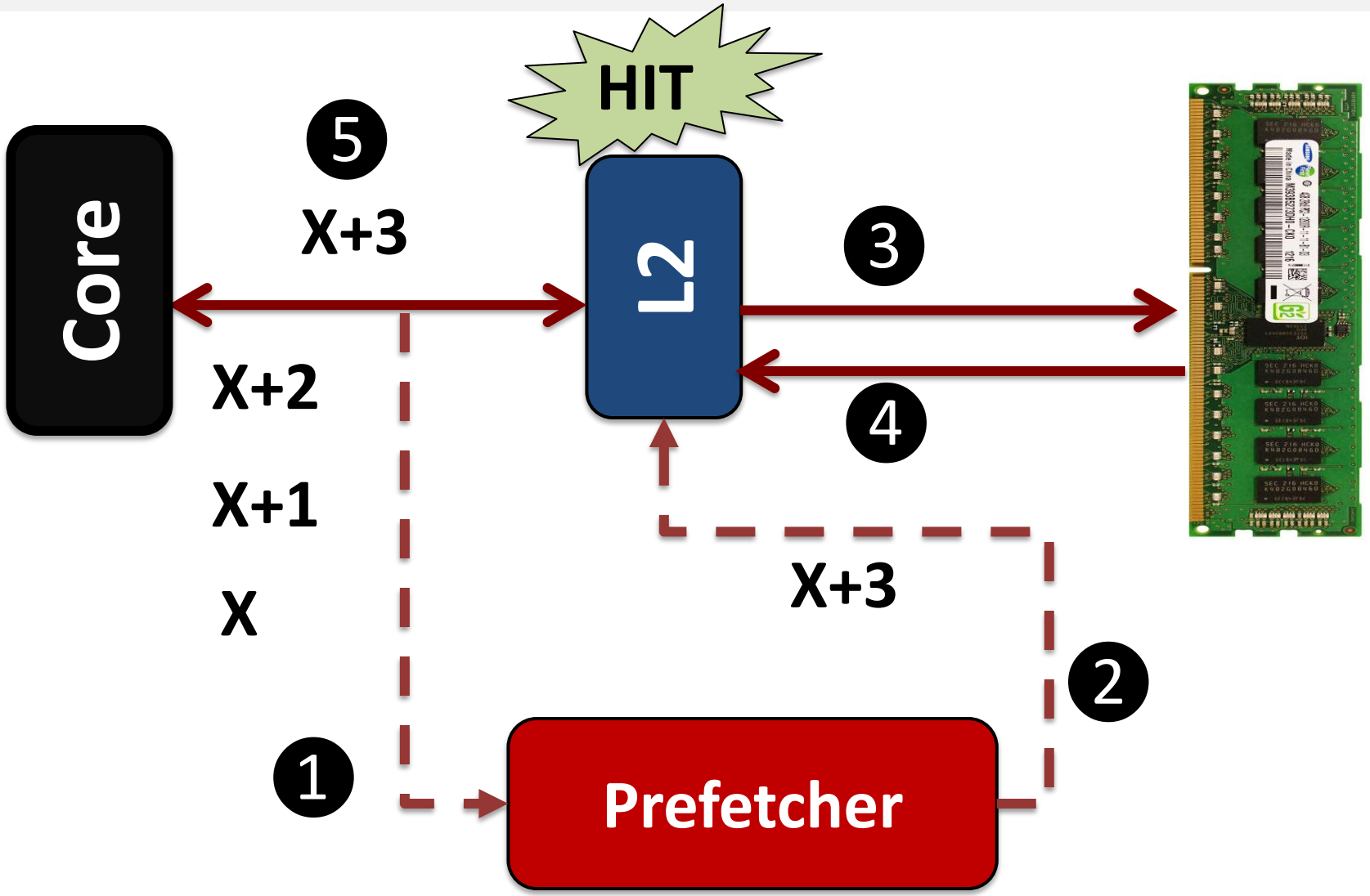
Why?

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

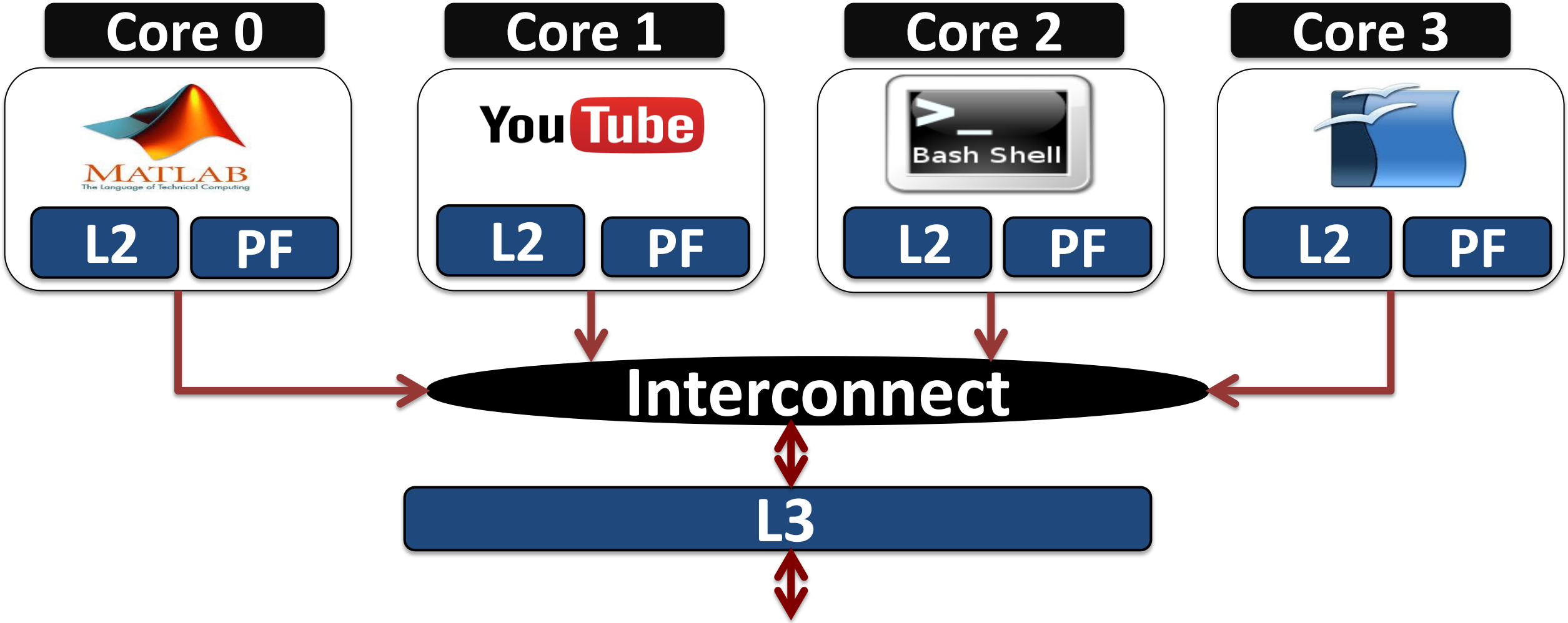
How?

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

Hardware Prefetch Engine

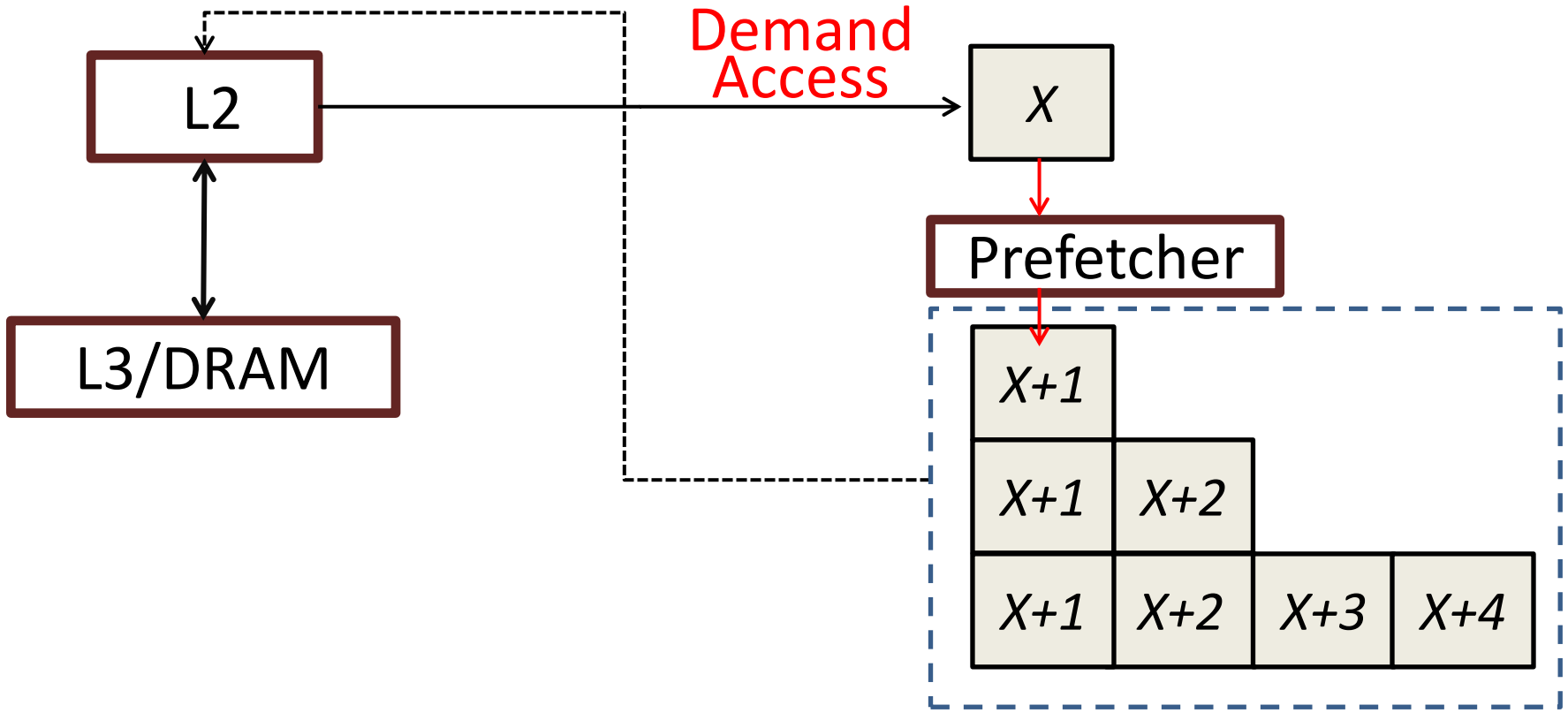


Prefetchers in Multicore - 101



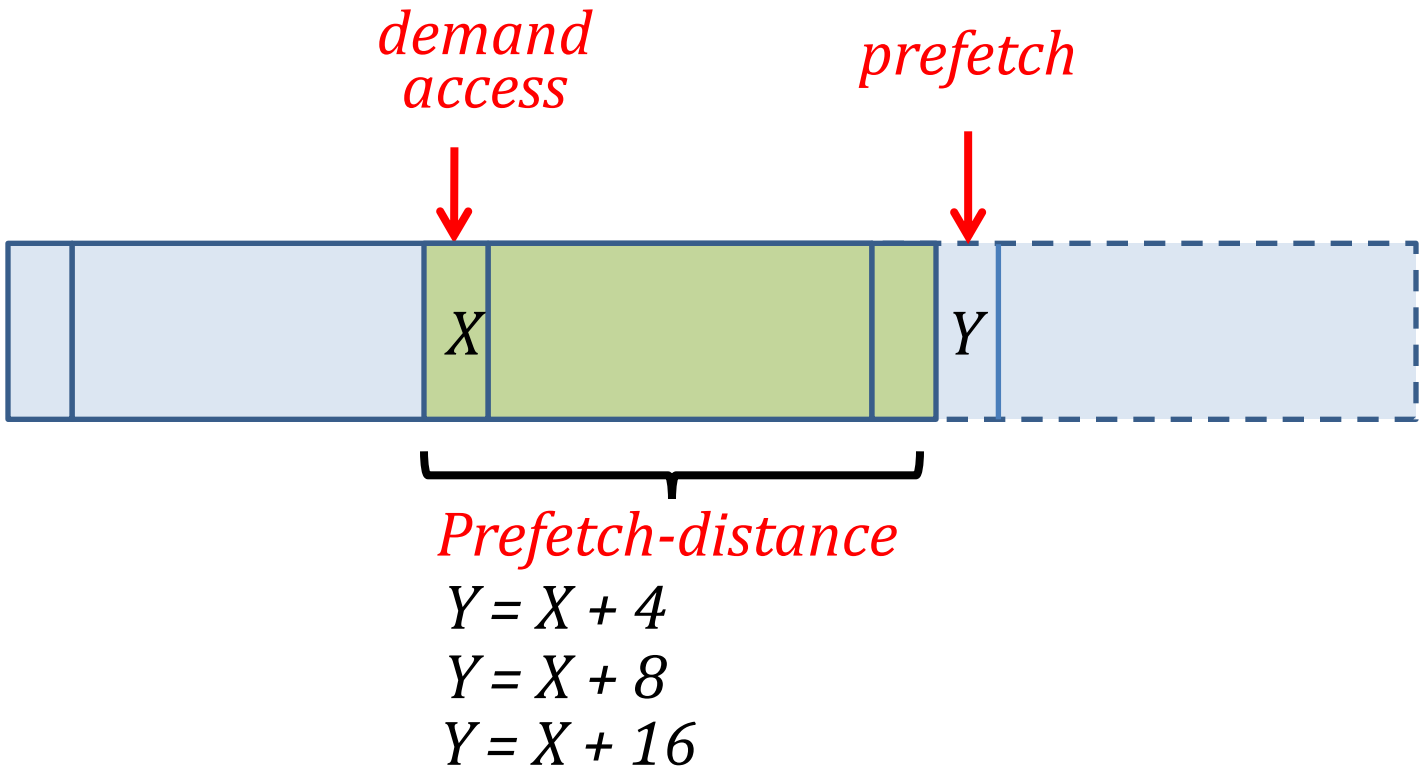
Prefetching Knobs

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



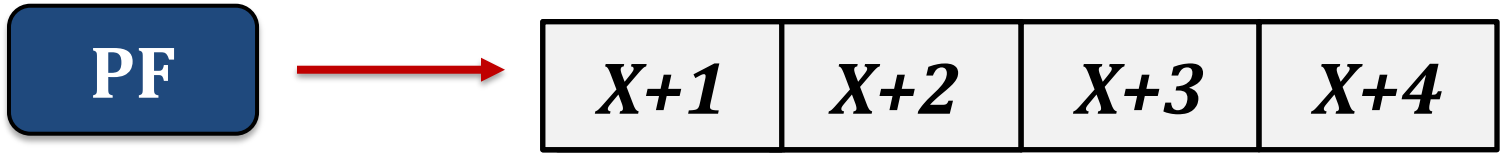
Prefetching Knobs

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

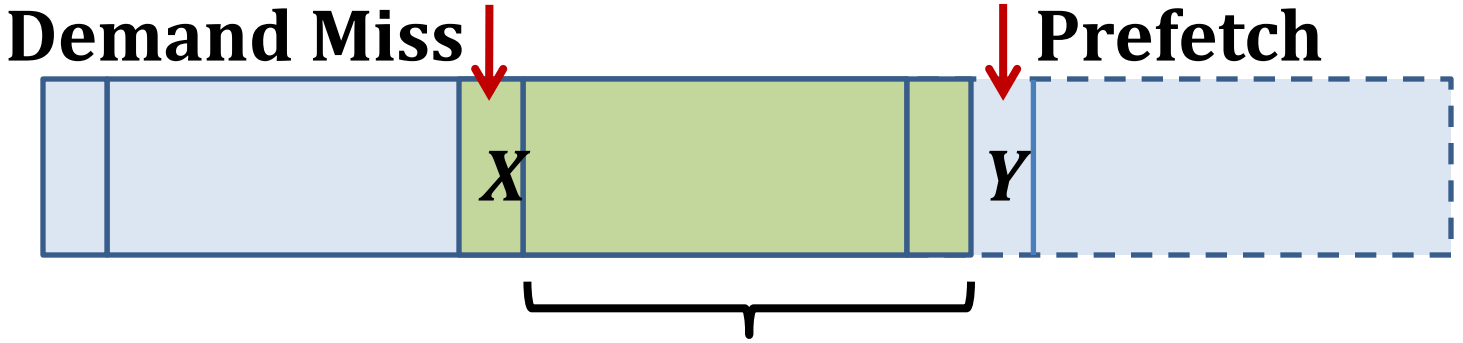


Aggressiveness [degree, distance]

Prefetch degree: #Prefetch requests issued on a miss



Prefetch distance: How far ahead (in terms of # blocks) of the demand miss ?



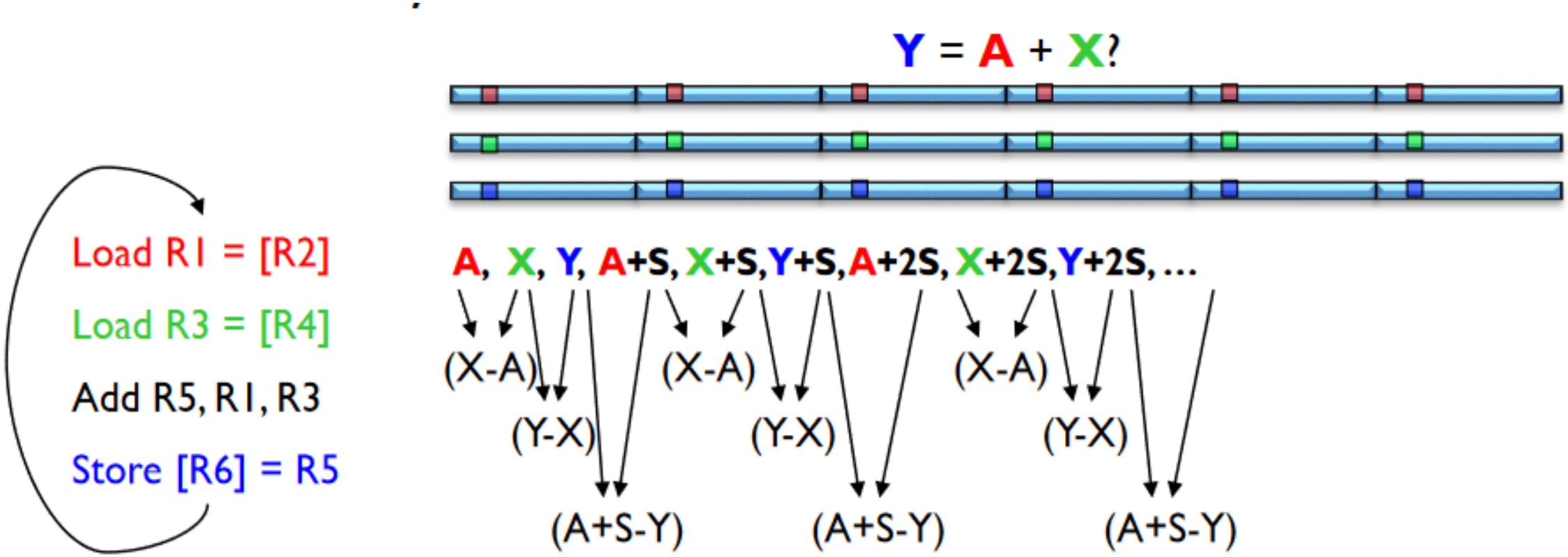
The Simplest Prefetcher

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

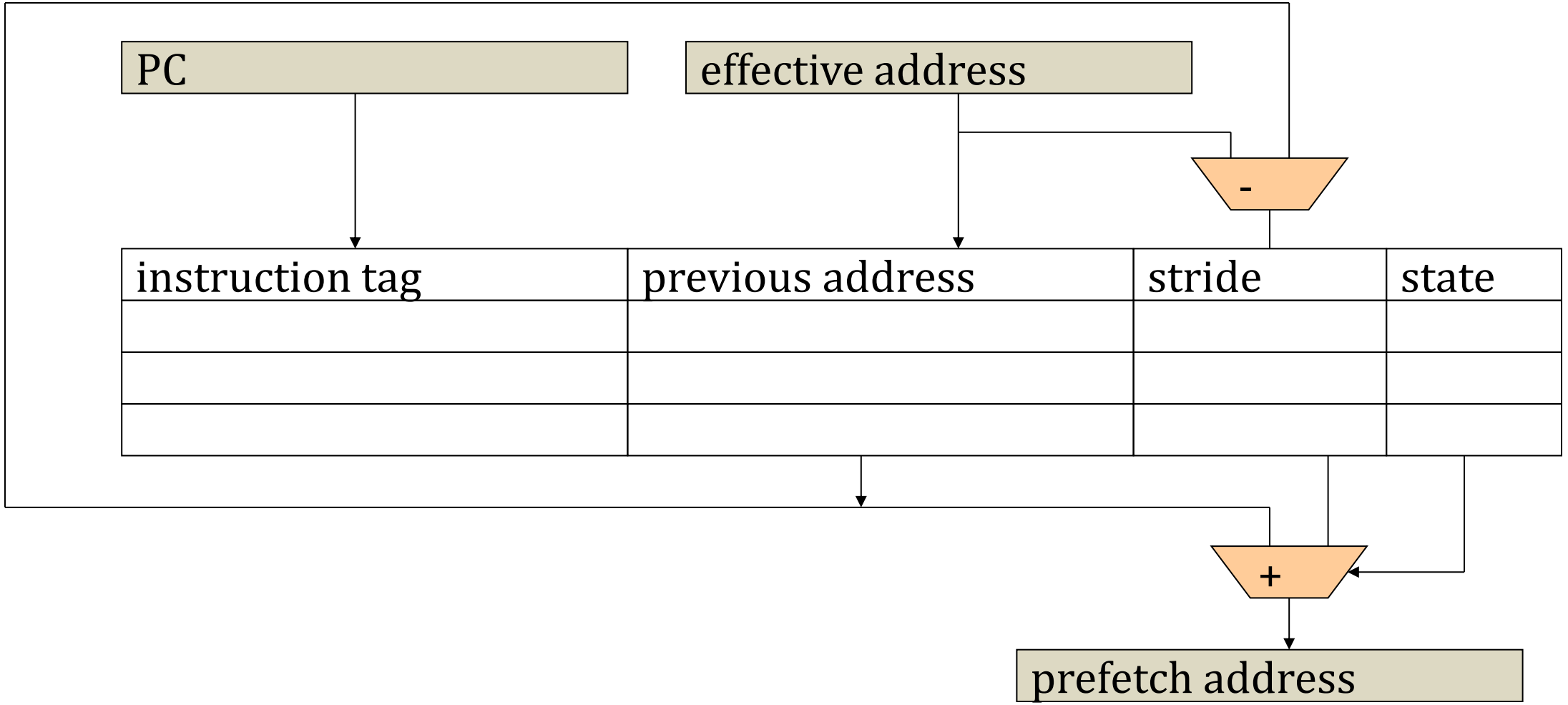
Works well for L1 Icache and L1 Dcache.

Next N Line: Miss to cache block X , prefetch X+1, X+2, ….. X+N, Degree=N, Distance=1

What about this?



Stride Prefetching

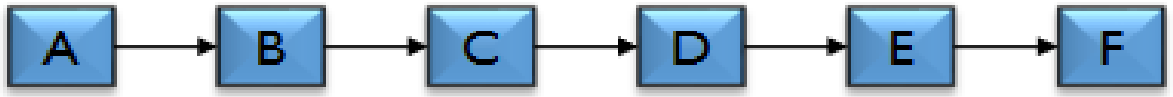


An Example

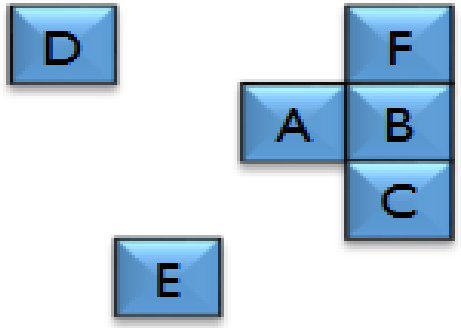
```
float a[100][100], b[100][100], c[100][100];
...
   for (i = 0; i < 100; i++)
    for (j = 0; j < 100; j++)
        for (k = 0; k < 100; k++)
            a[i][j] += b[i][k] * c[k][j];
```

instruction tag	previous address	stride	state
ld b[i][k]	50008	4	steady
ld c[k][j]	90800	800	steady
ld a[i][j]	10000	0	steady

Pointer Chasers



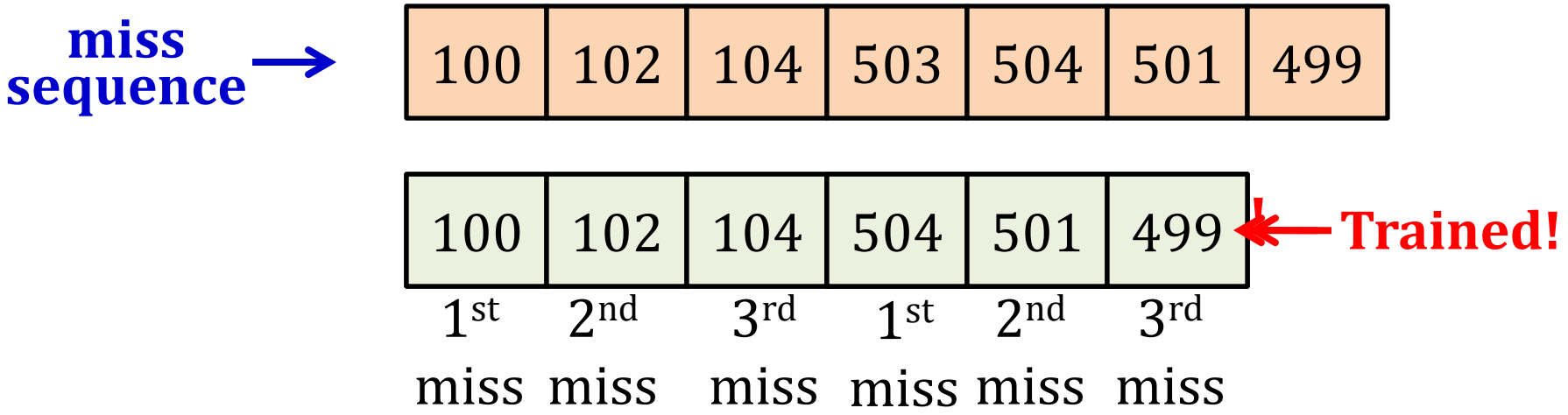
Linked-list traversal



Actual memory layout

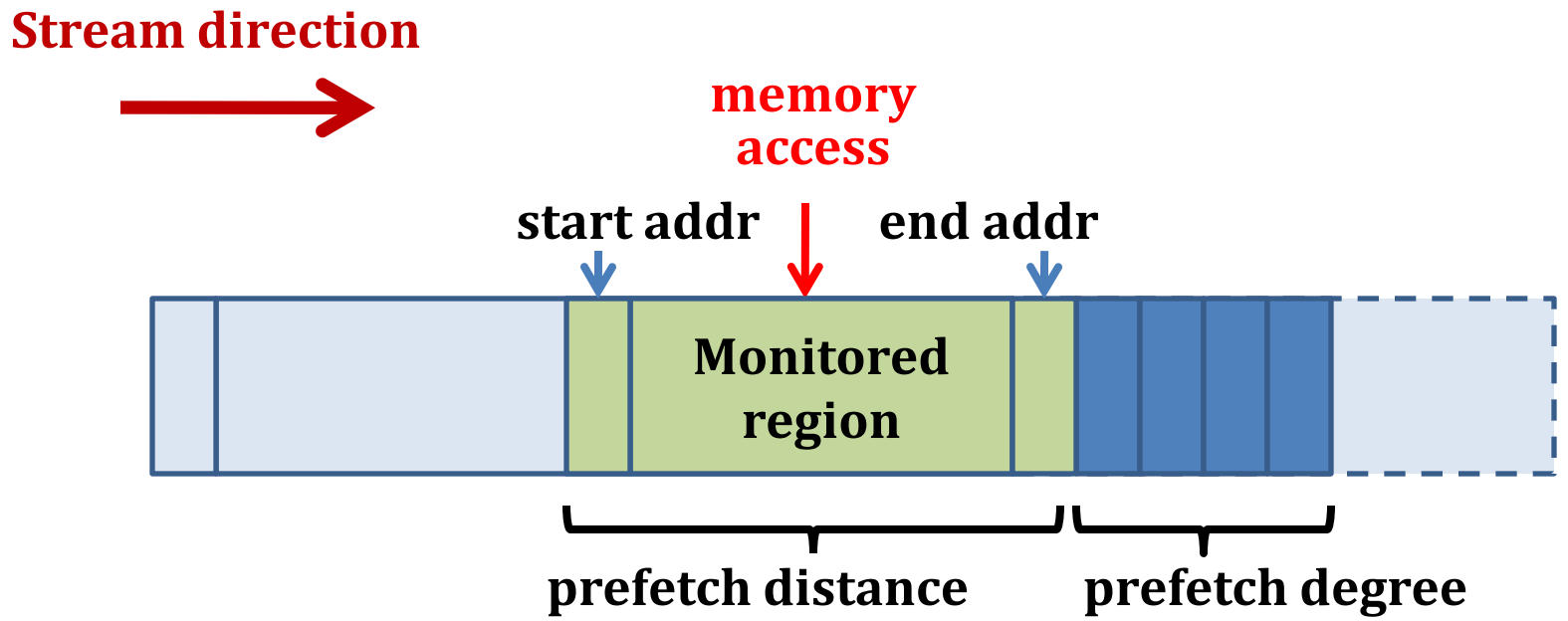
(no chance to detect a stride)

Stream Prefetching [DPC1]



Training: Consecutive misses in the same direction.

Stream Prefetcher in Action



Start address and end address will be shifted by blkSize*prefetch-degree after prefetch.

Stream + Stride