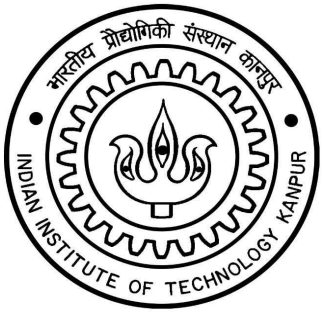
Sangam: A Multi-component Core Cache Prefetcher

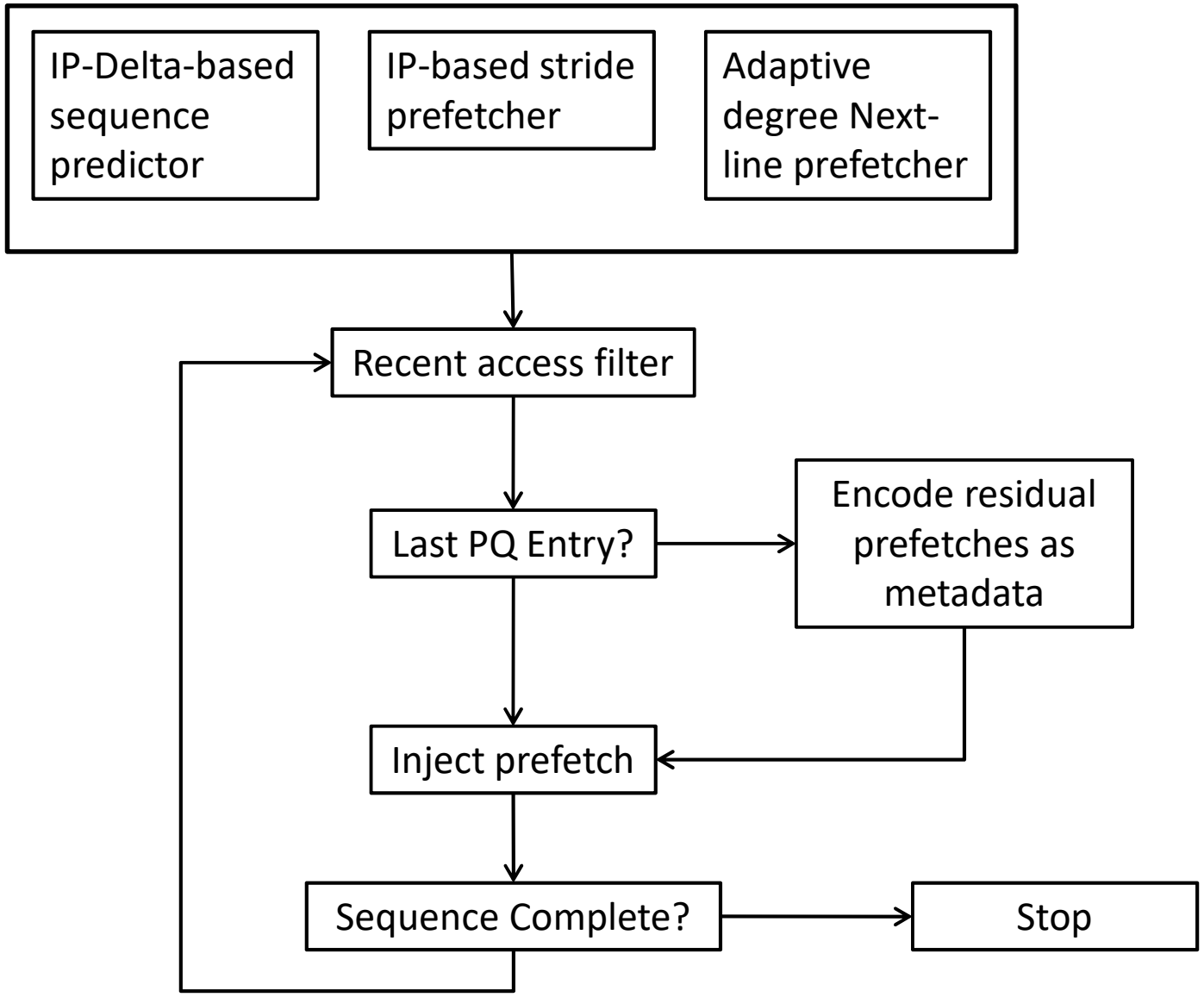
Mainak Chaudhuri, **Nayan Deshmukh**



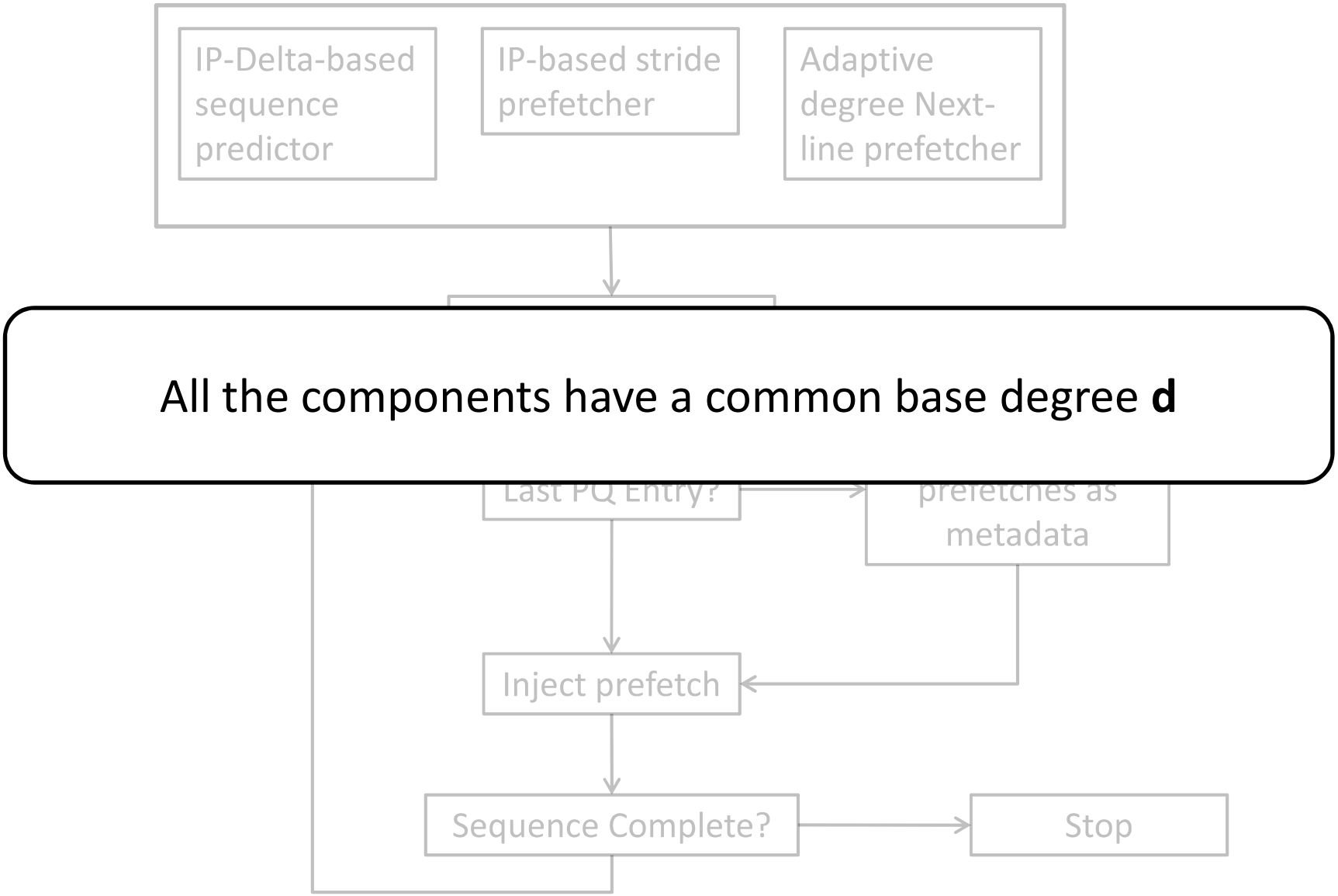
Introduction

- The word ‘Sangam’ refers to a confluence of 3 rivers which corresponds to 3 core components in our prefetcher
- We achieve 40.3% speedup over no prefetching for 46 single core workloads
- For 4 core we achieve 19.5% speedup over no prefetching for 100 multiprogramed workloads (45 homo, 55 hetro)

Sangam


IP-Delta-based sequence predictor
IP-based stride prefetcher
Adaptive degree Next-line prefetcher
Recent access filter
Last PQ Entry?
Encode residual prefetches as metadata
Inject prefetch
Sequence Complete?
Stop

Sangam


IP-Delta-based sequence predictor
IP-based stride prefetcher
Adaptive degree Next-line prefetcher
All the components have a common base degree **d**
Last PQ Entry?
prefetches as metadata
Inject prefetch
Sequence Complete?
Stop

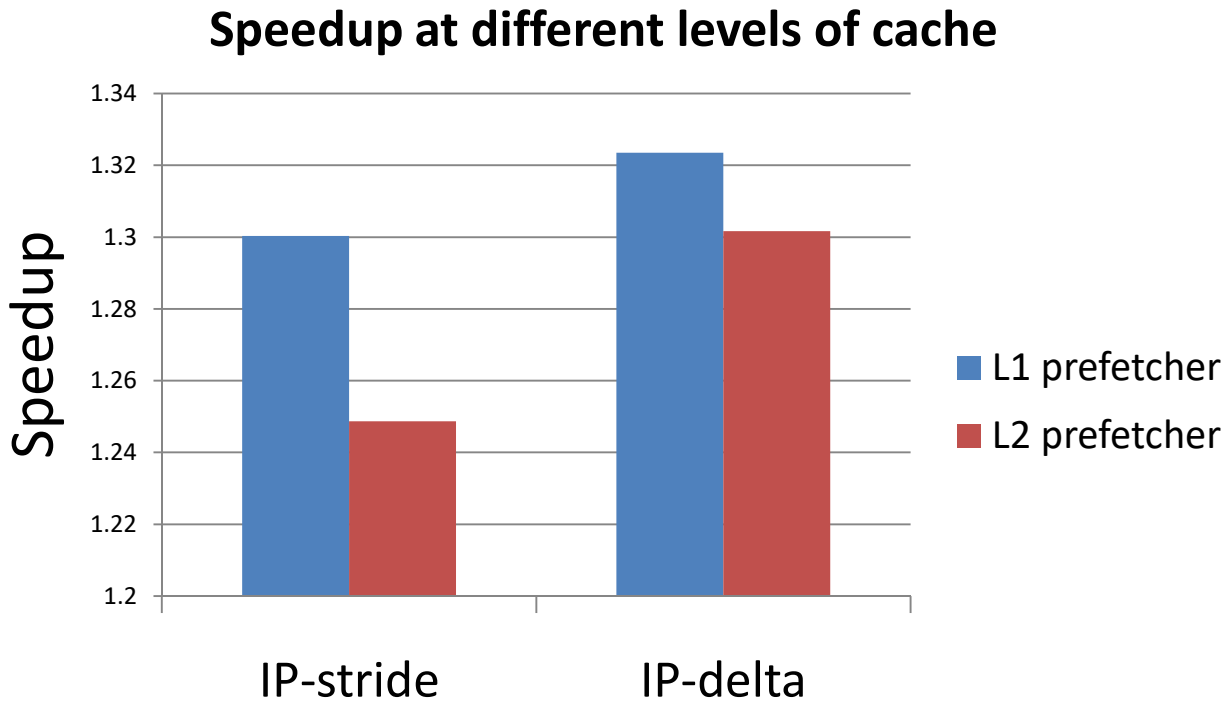
Where?

Where?

- Where to place the prefetcher
- L1 allows for better learning whereas L2, L3 allows for more hardware resources

Where?

- Where to place the prefetcher
- L1 allows for better learning whereas L2, L3 allows for more hardware resources



IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses

IP Table

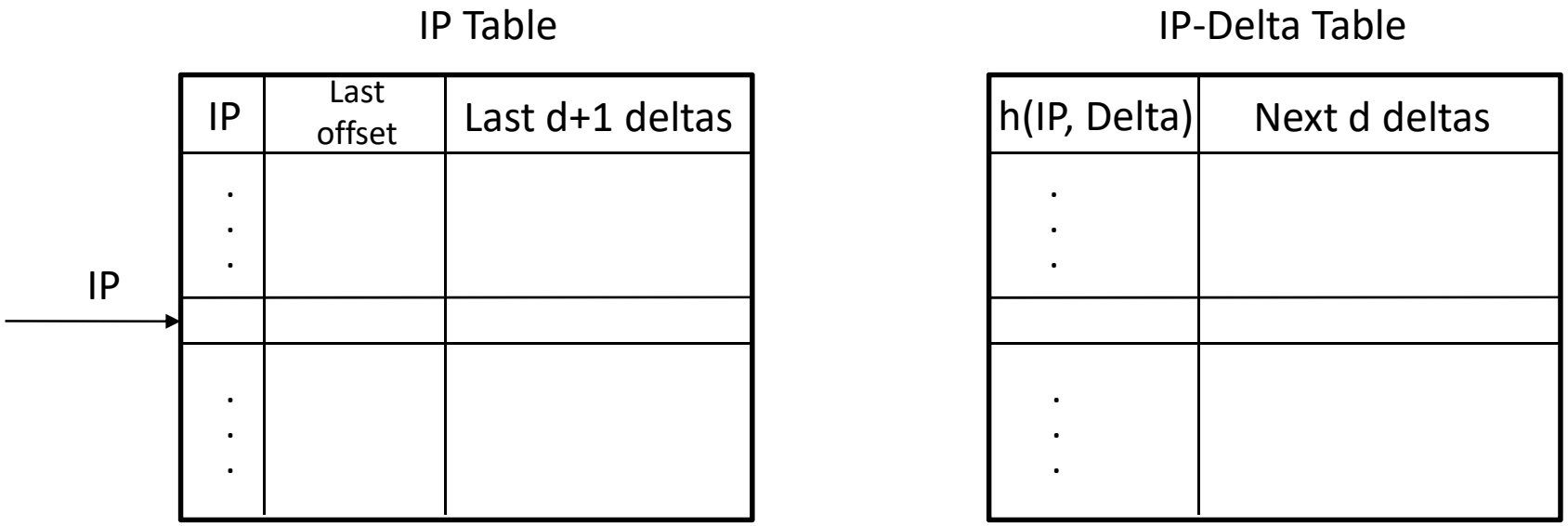
IP	Last offset	Last d+1 deltas
. . .		
. . .		

IP-Delta Table

h(IP, Delta)	Next d deltas
. . .	
. . .	

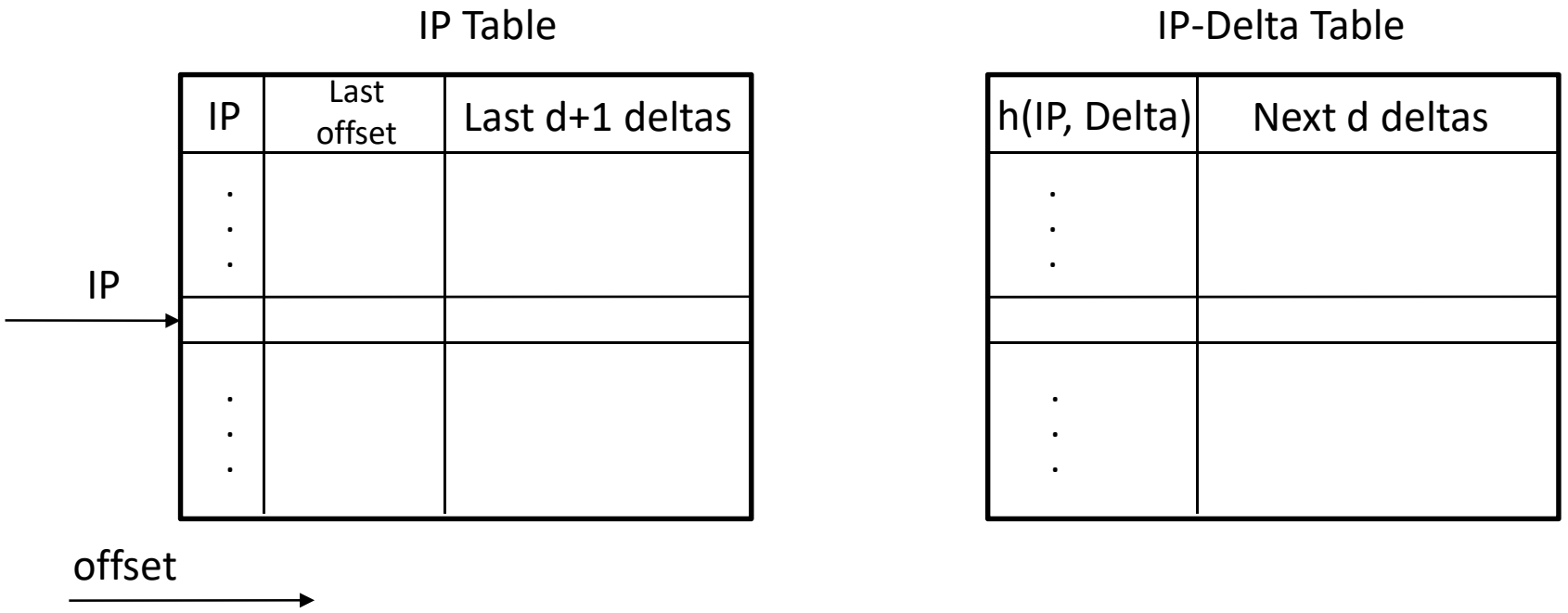
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



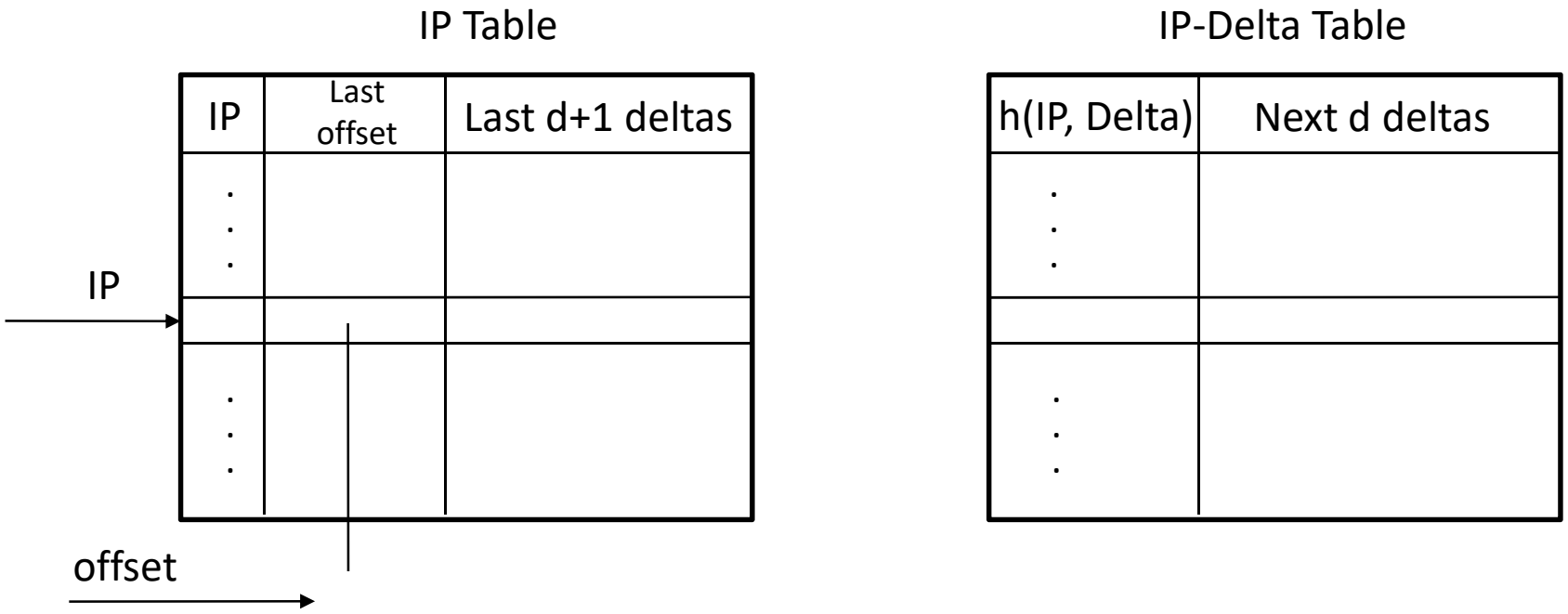
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



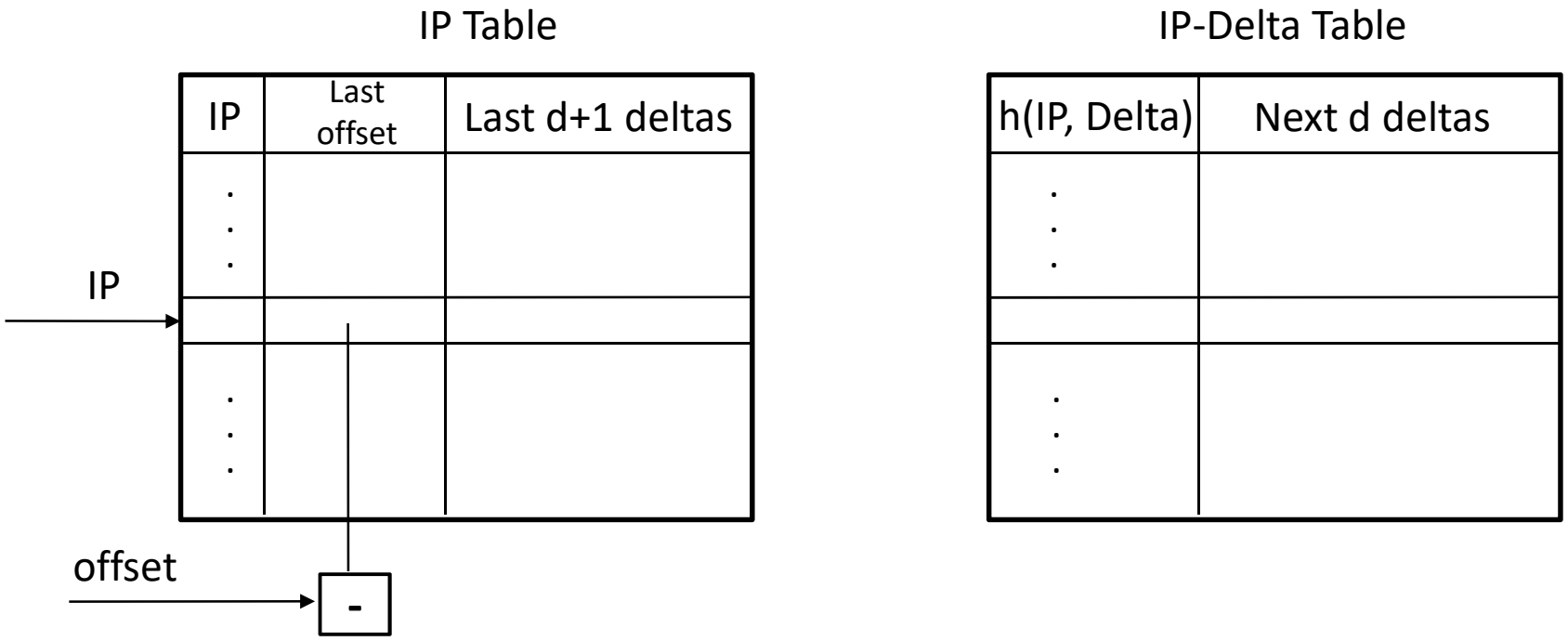
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



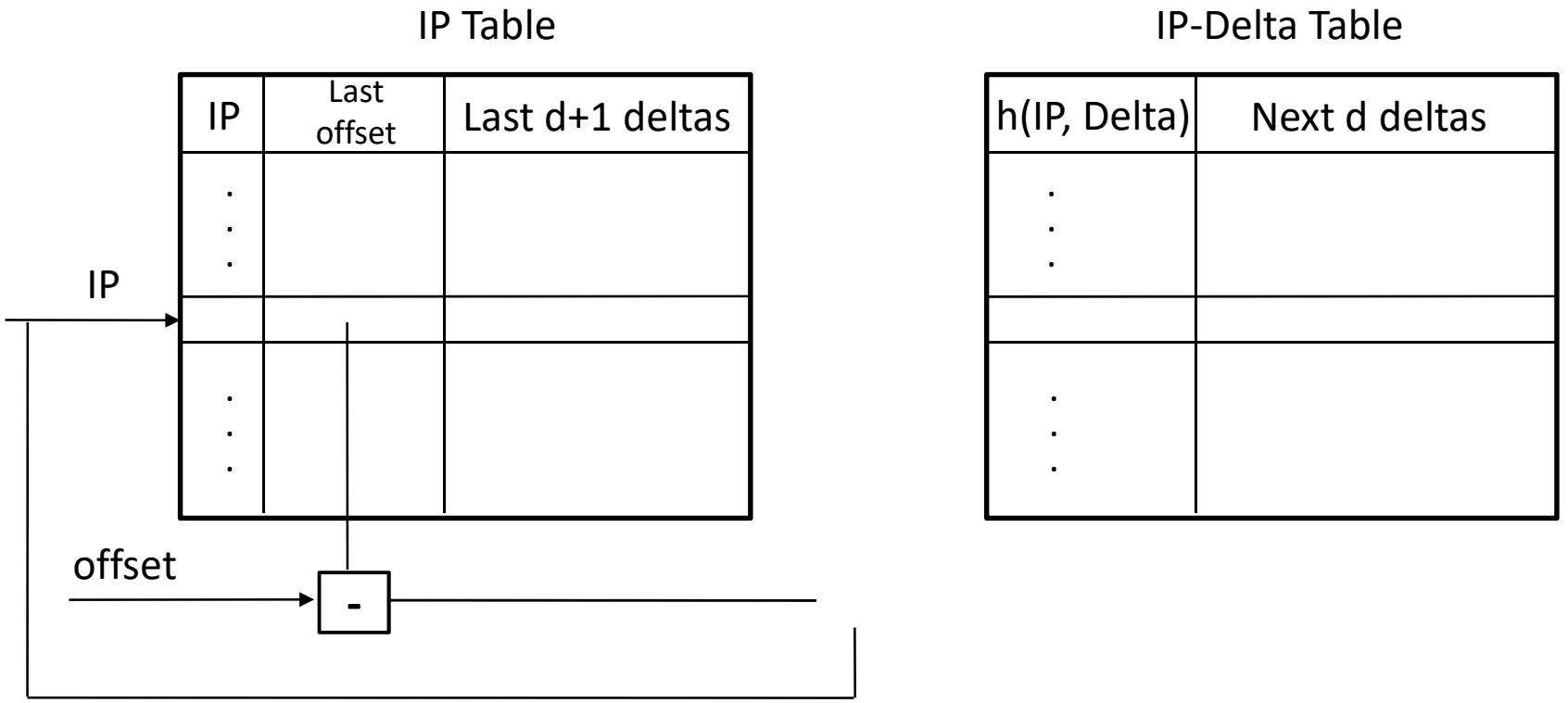
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



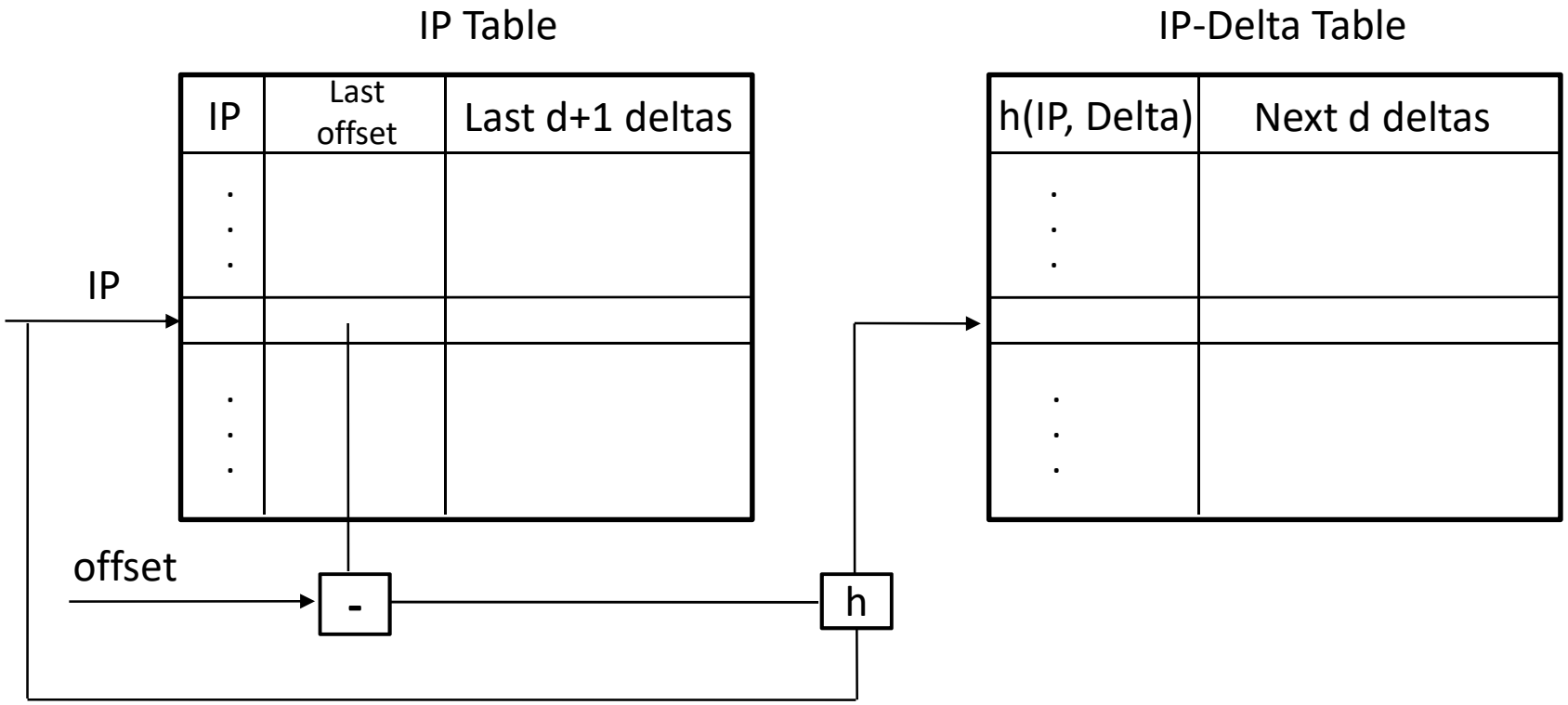
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



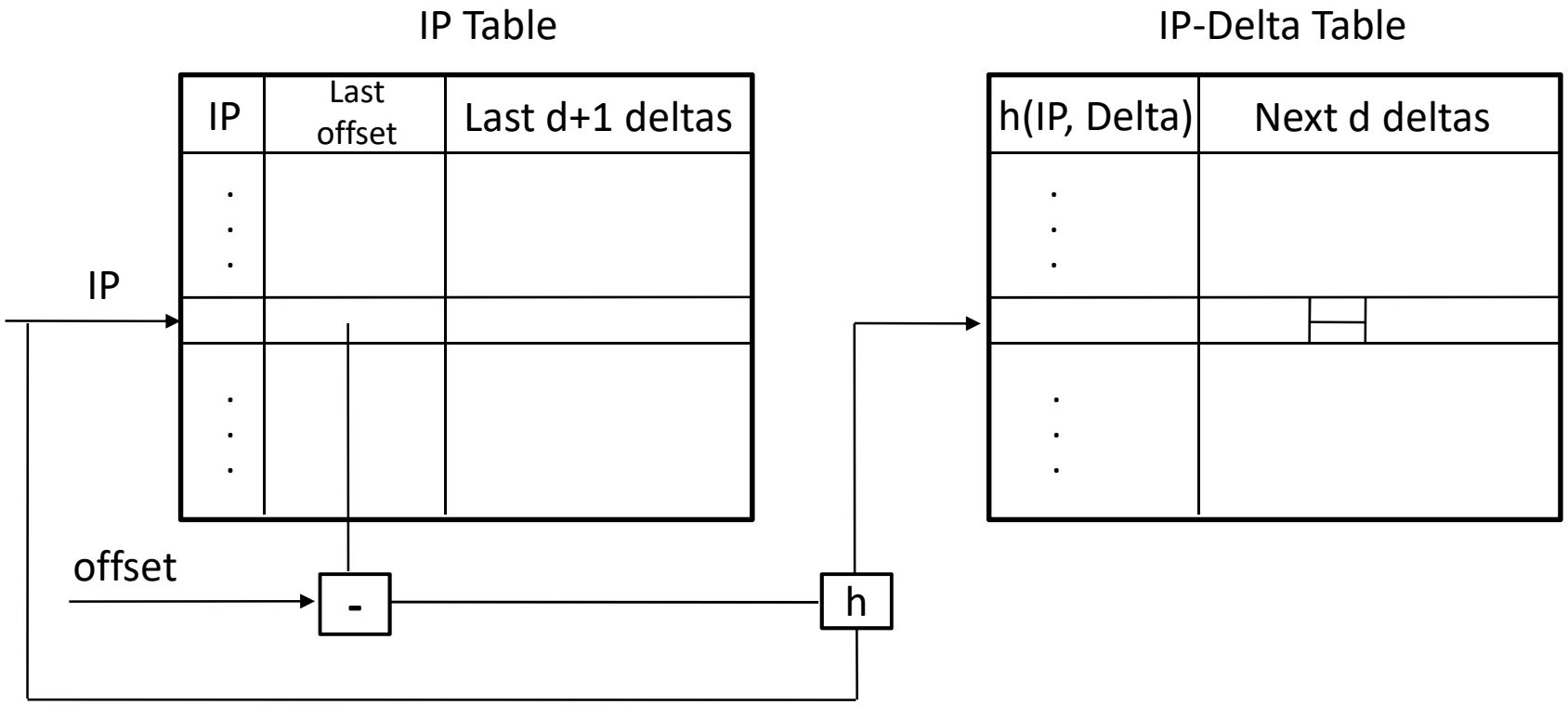
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



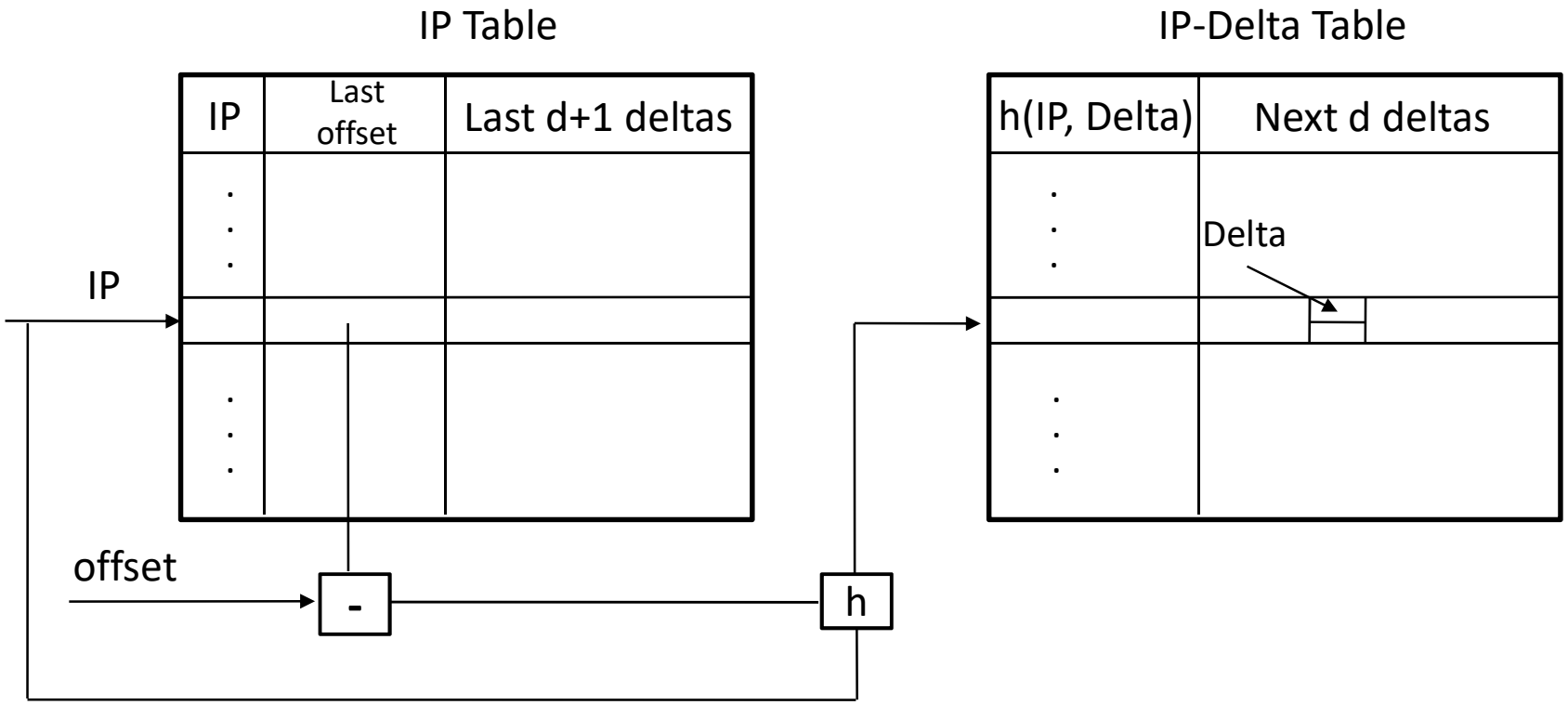
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



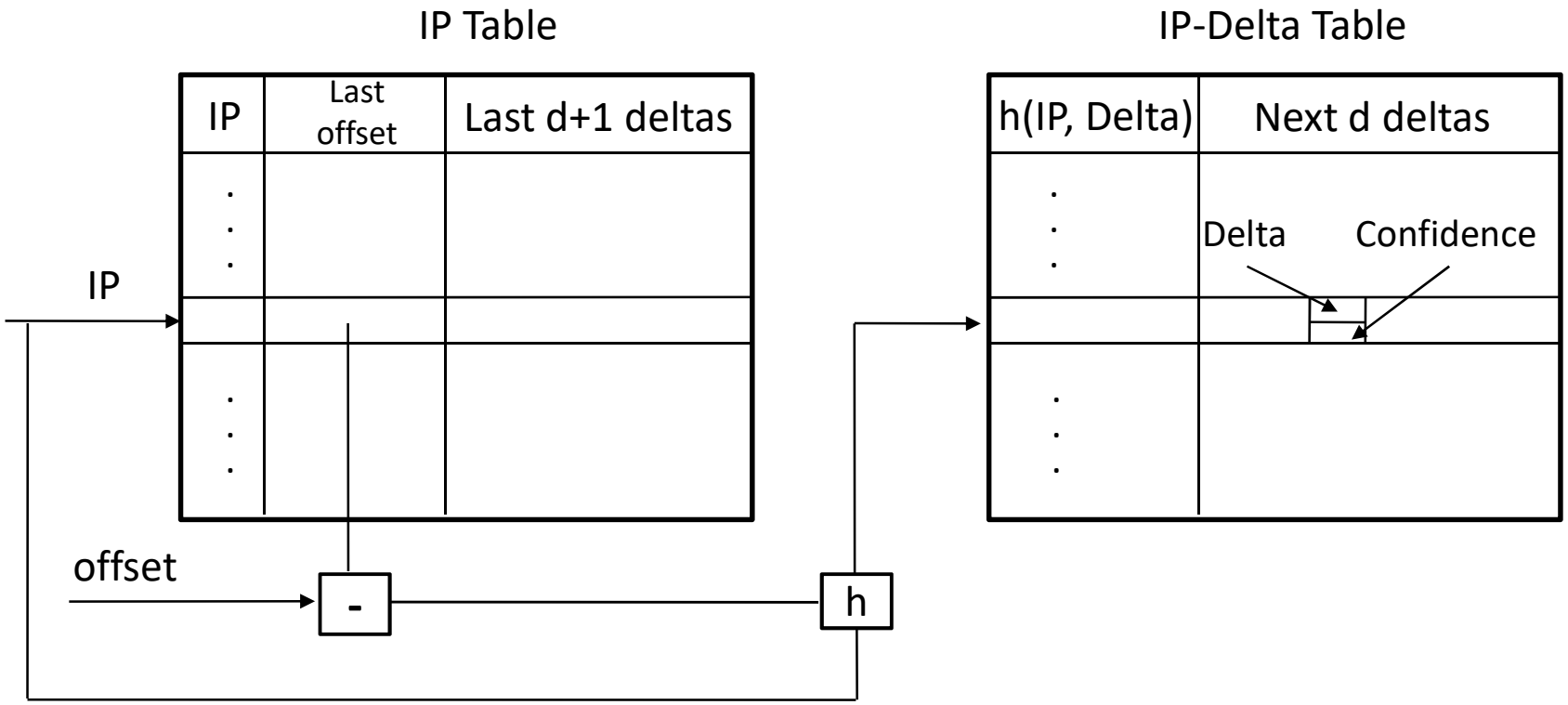
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



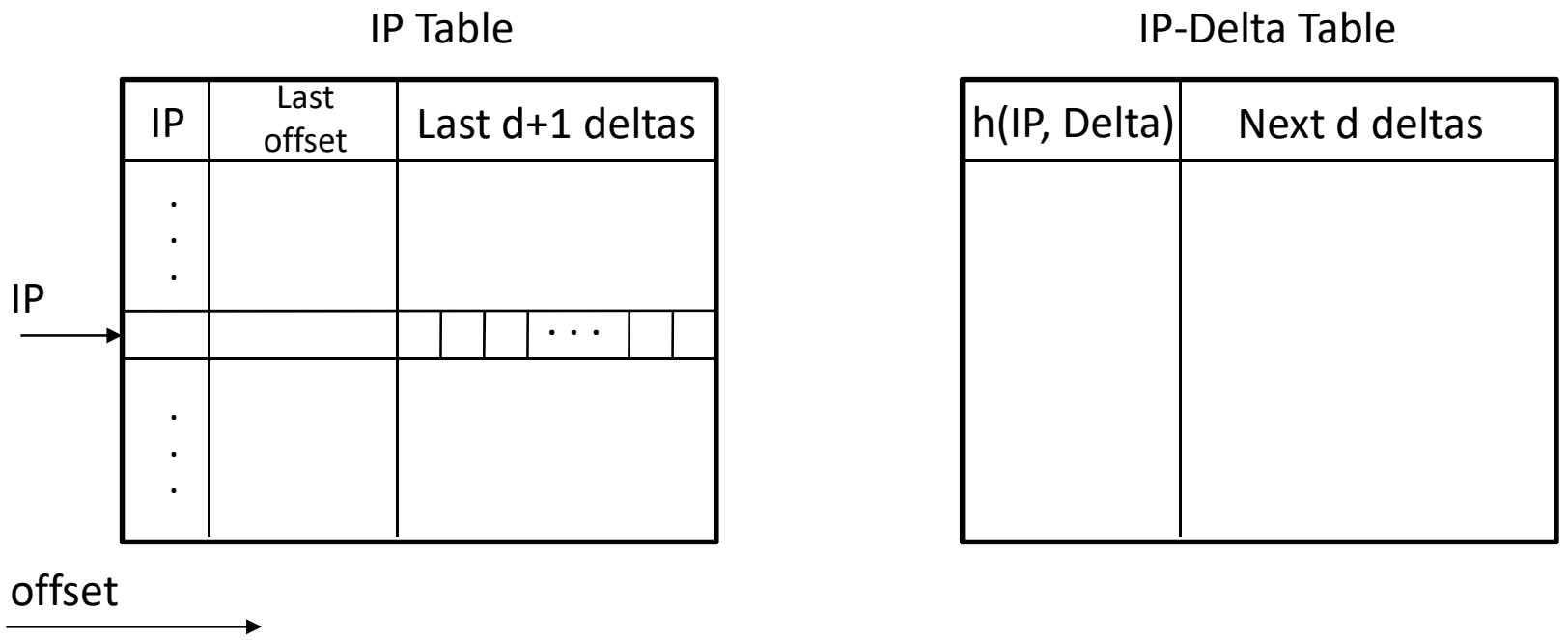
IP-Delta-based Sequence predictor

- Uses both control-flow and data-flow information to predict a sequence of accesses



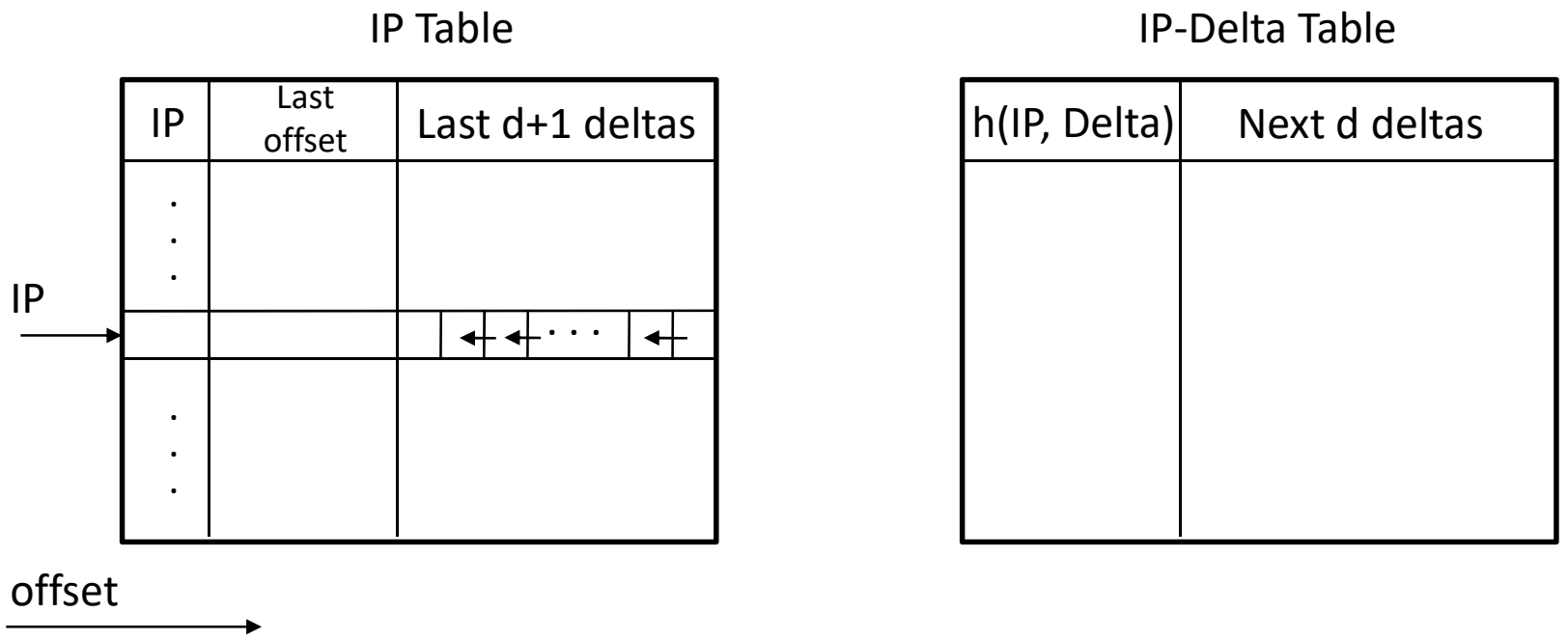
IP-Delta-based Sequence predictor

- Learning



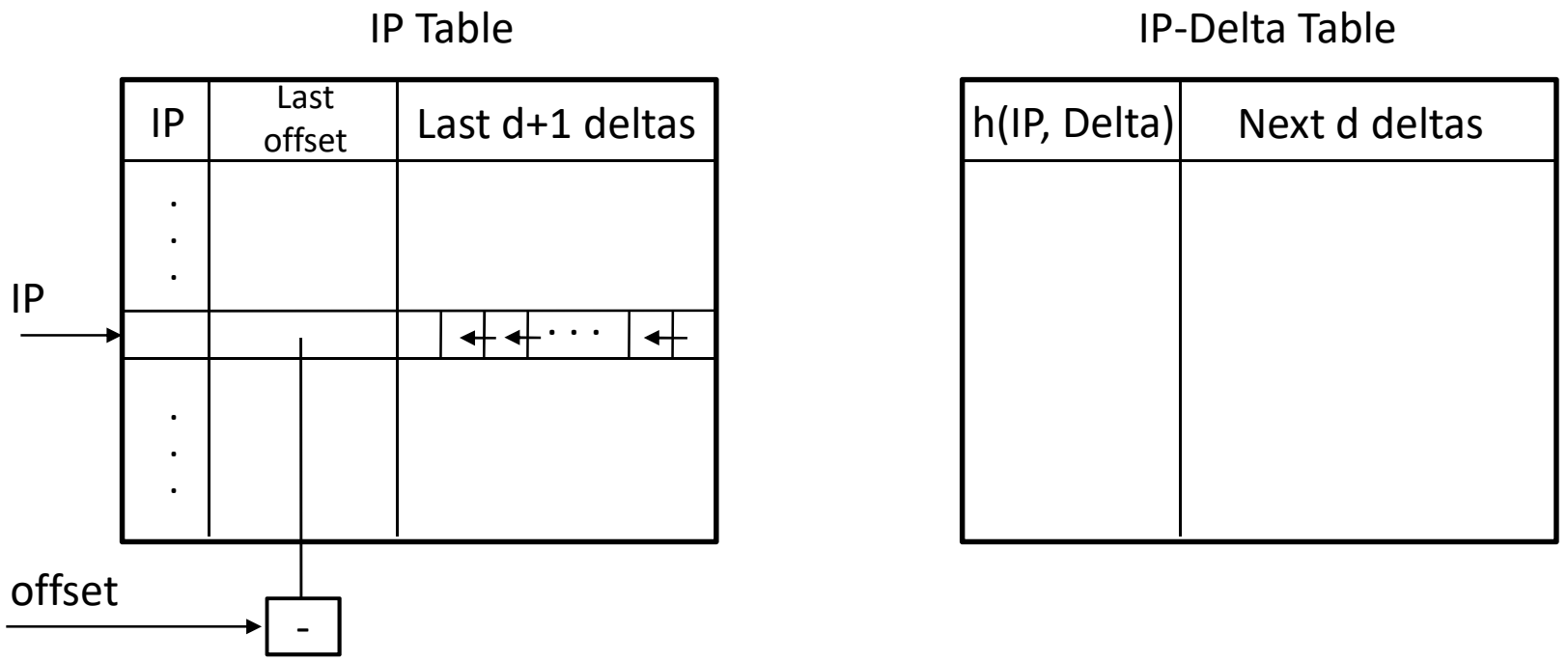
IP-Delta-based Sequence predictor

- Learning



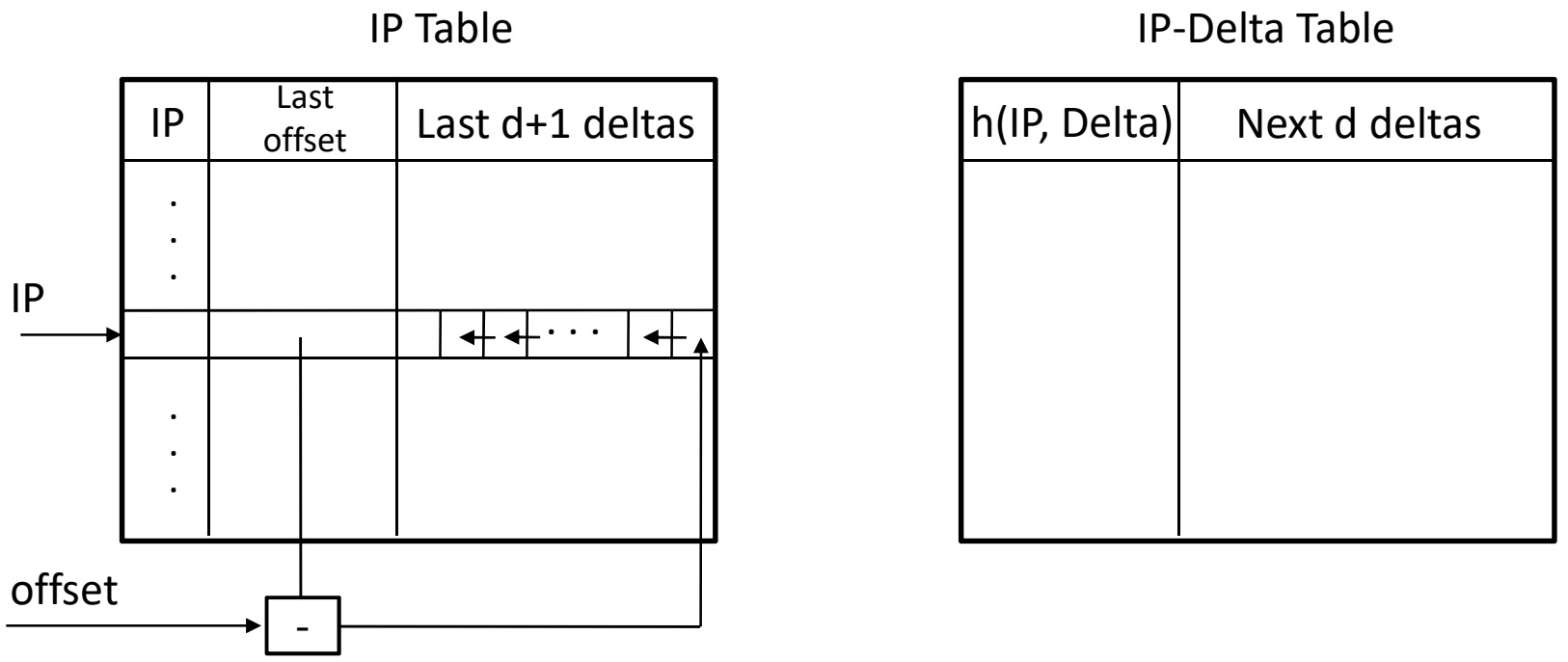
IP-Delta-based Sequence predictor

- Learning



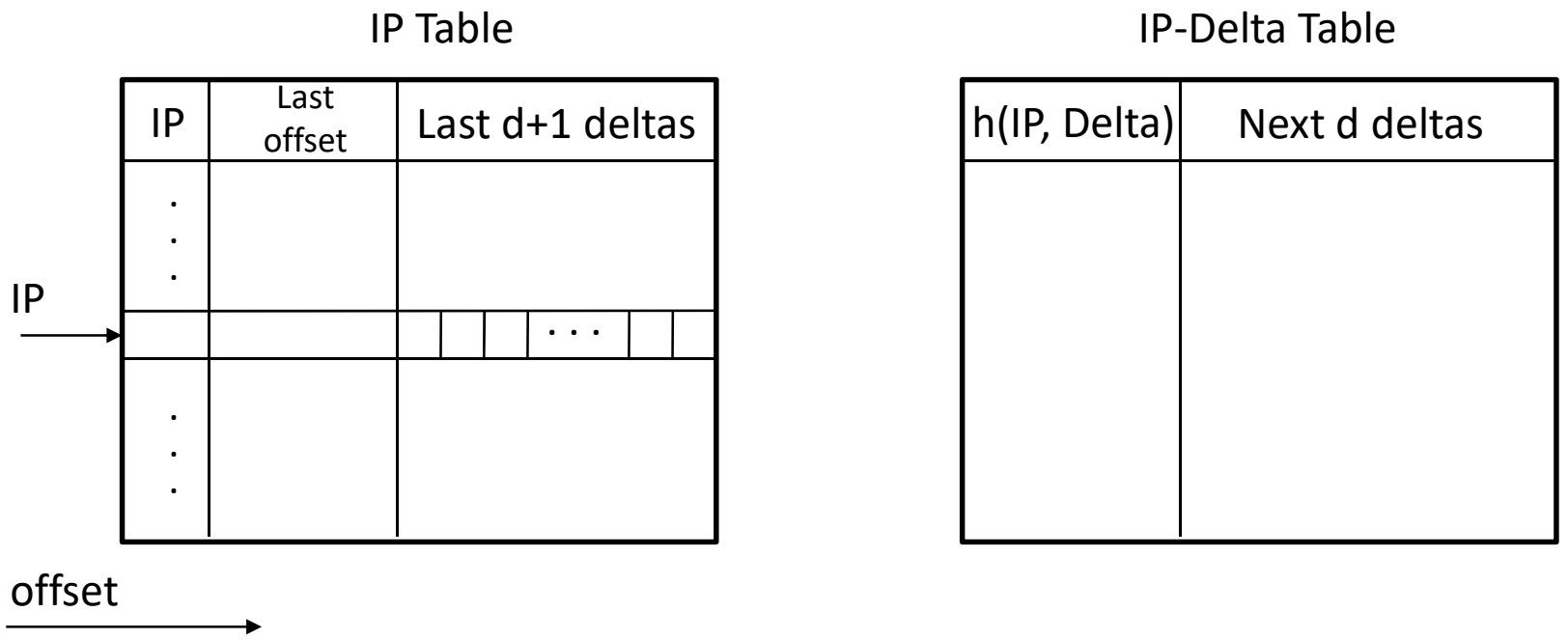
IP-Delta-based Sequence predictor

- Learning



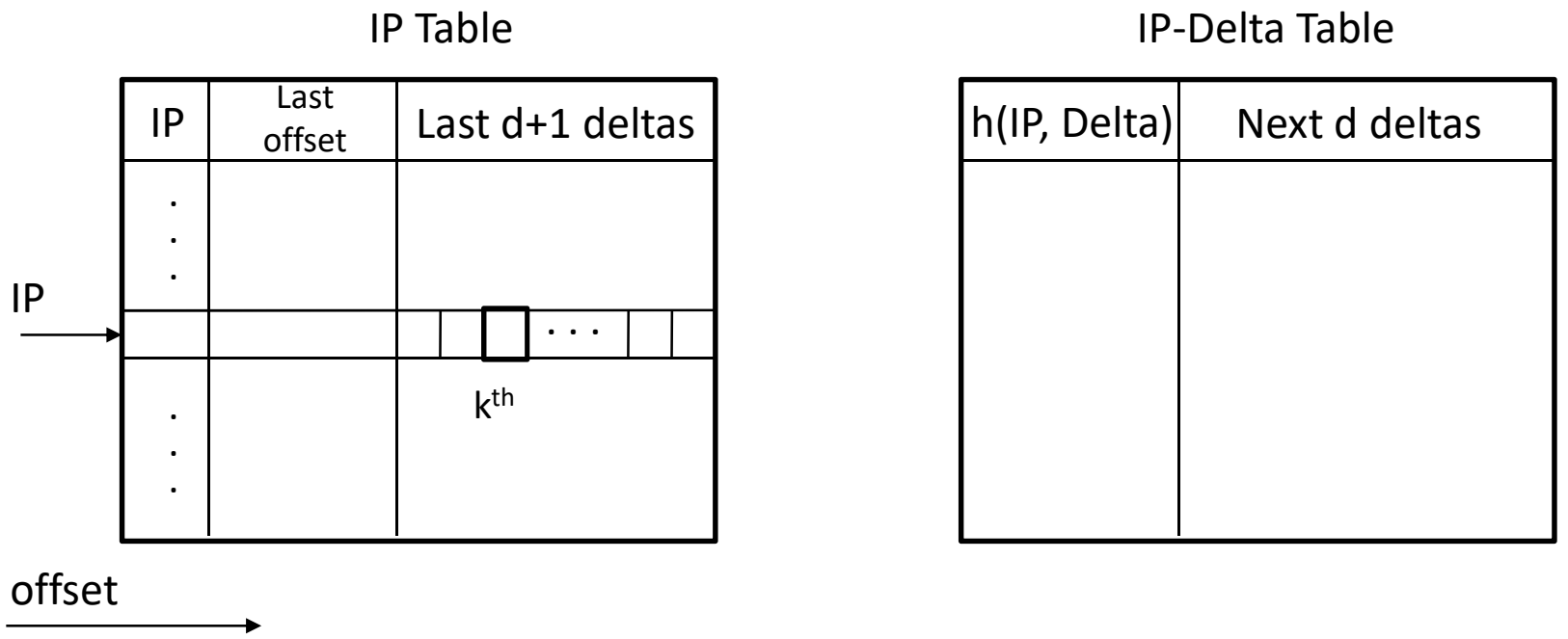
IP-Delta-based Sequence predictor

- Learning



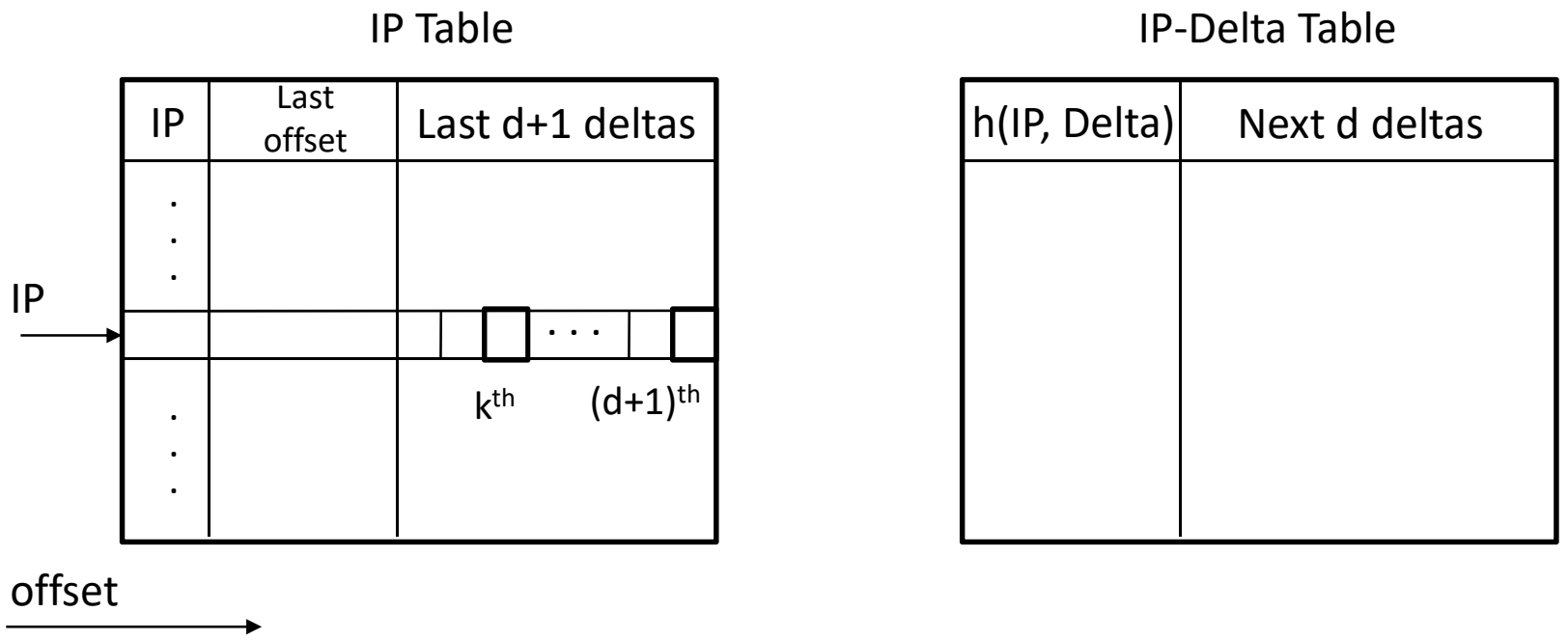
IP-Delta-based Sequence predictor

- Learning



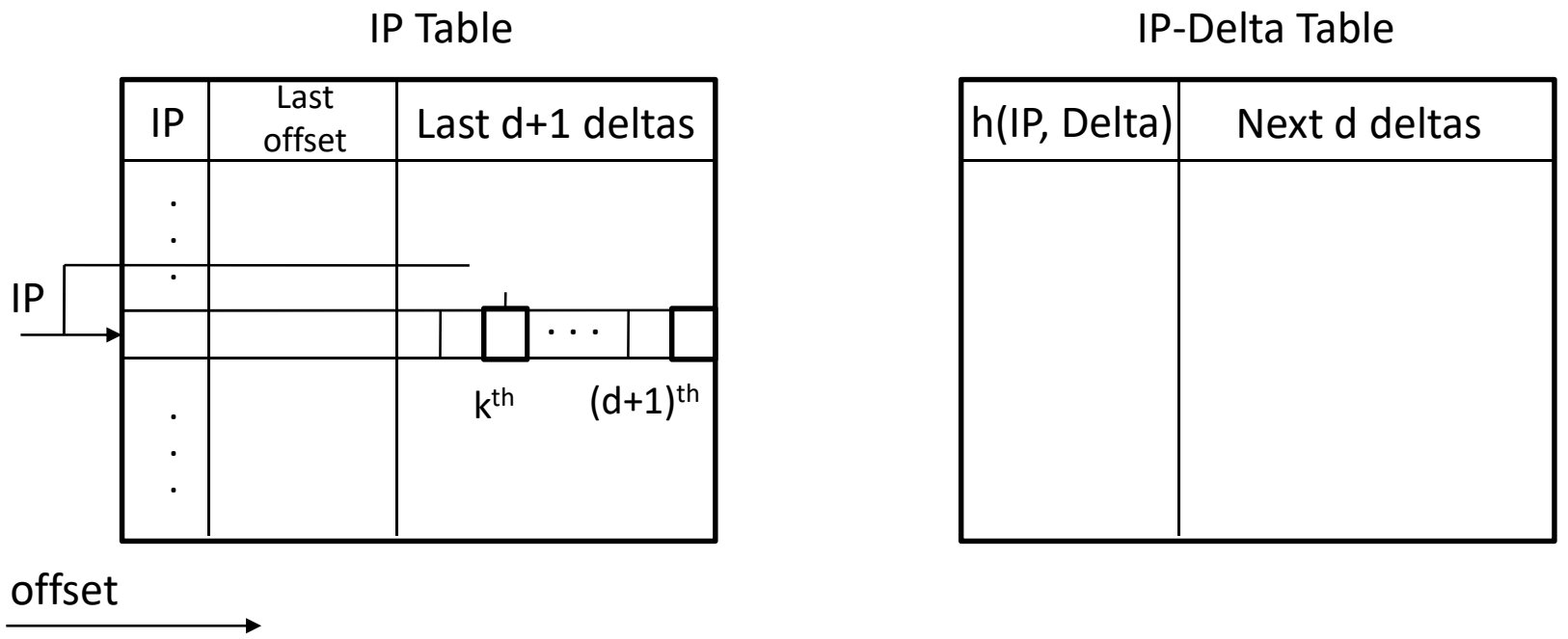
IP-Delta-based Sequence predictor

- Learning



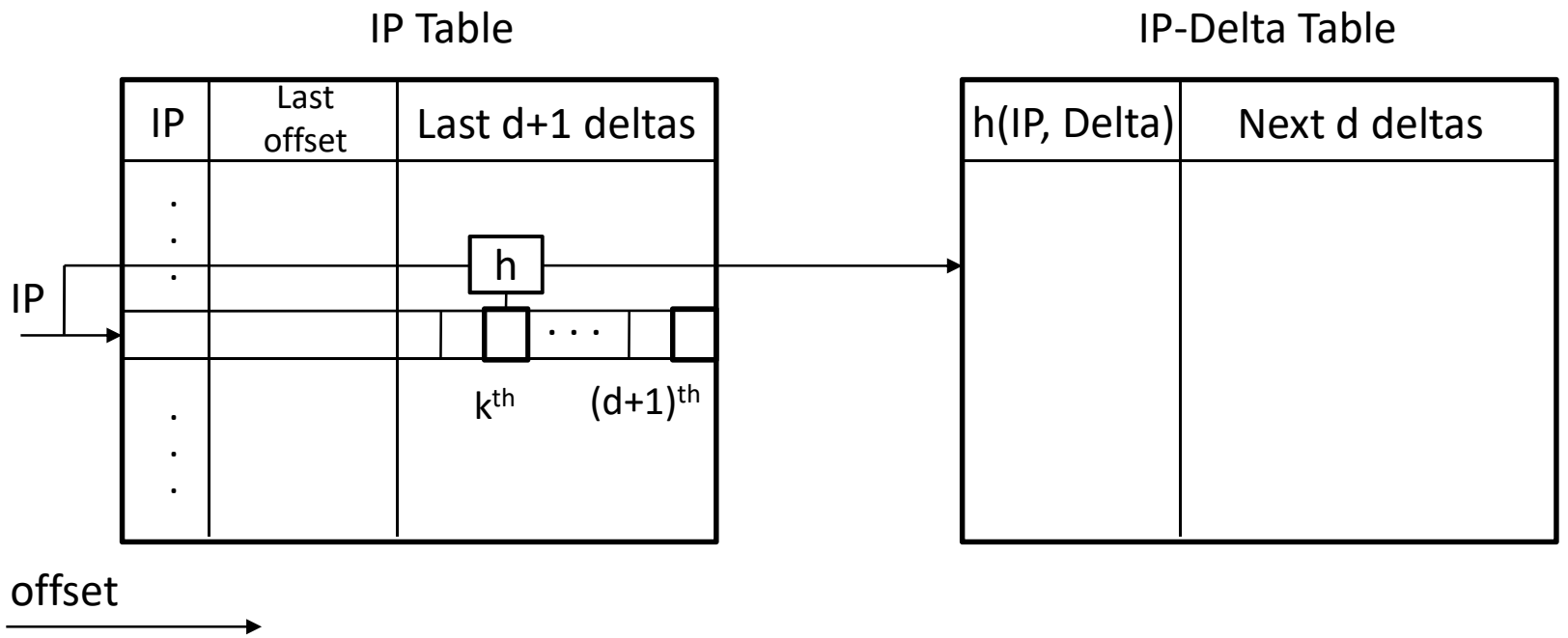
IP-Delta-based Sequence predictor

- Learning



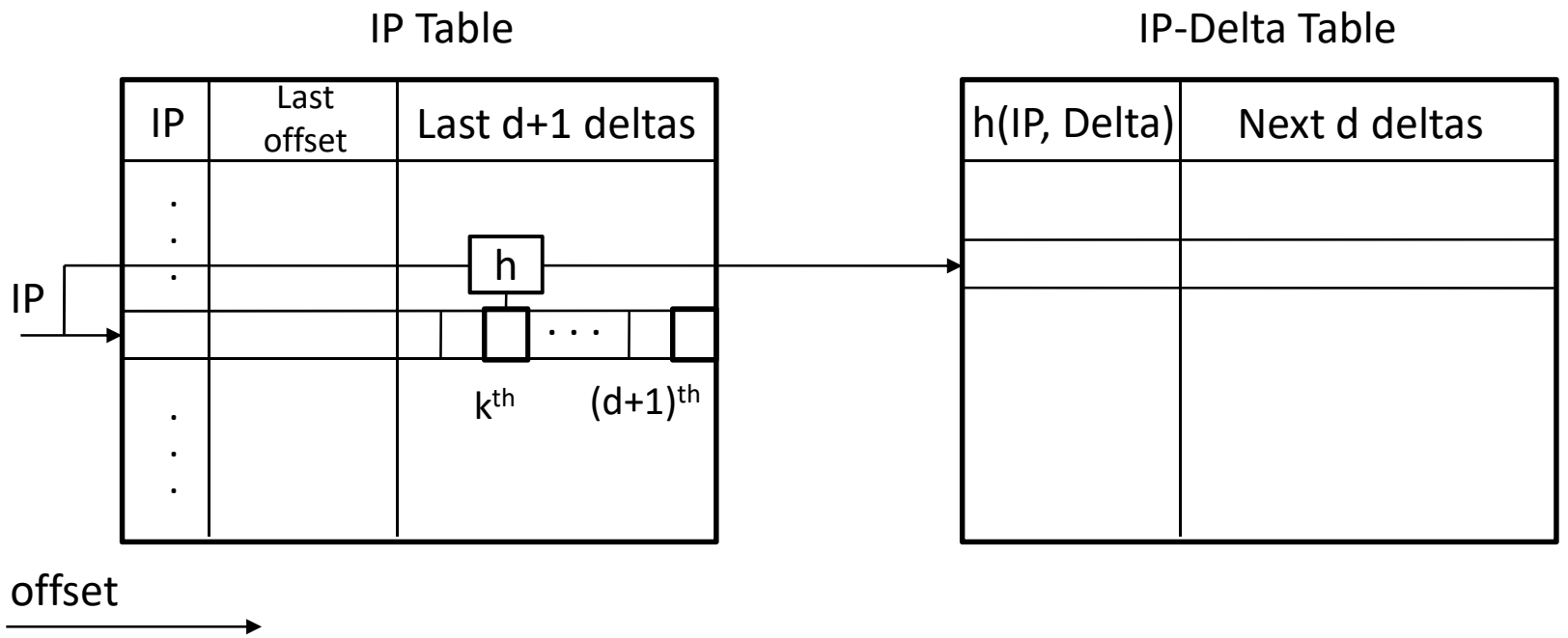
IP-Delta-based Sequence predictor

- Learning



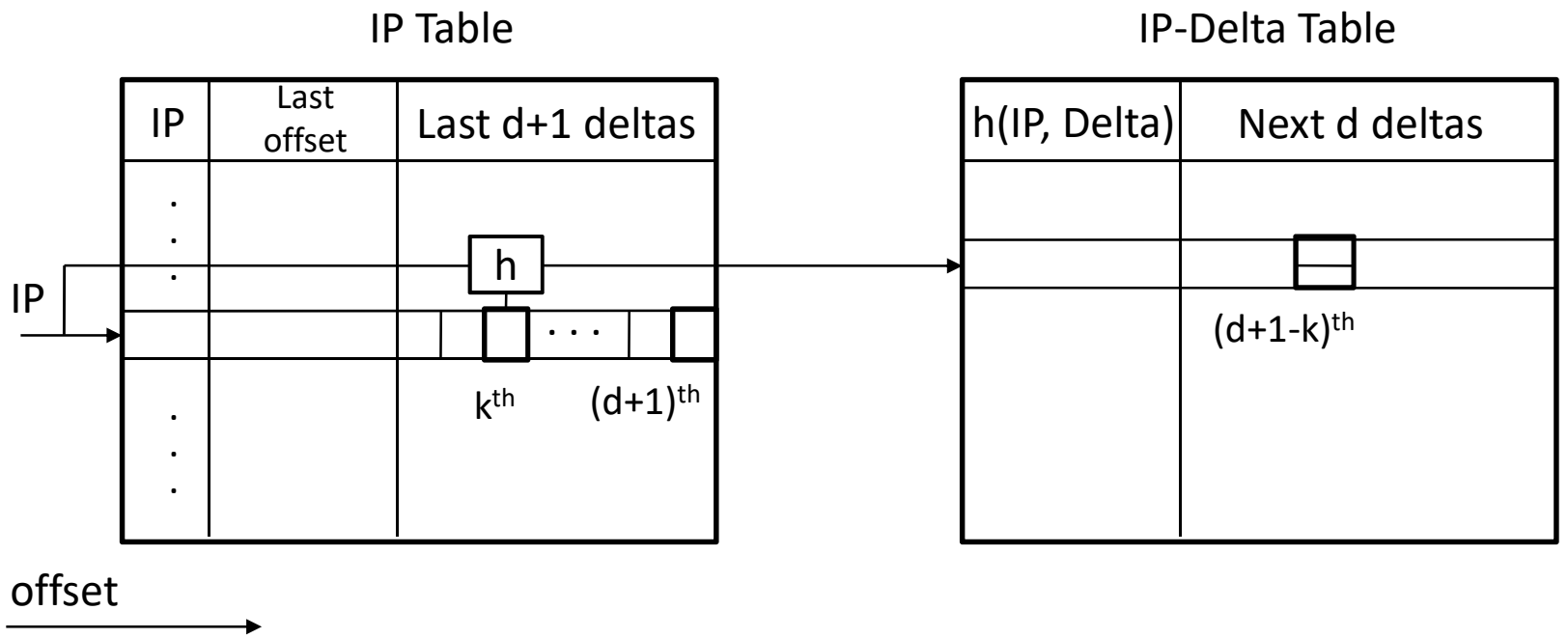
IP-Delta-based Sequence predictor

- Learning



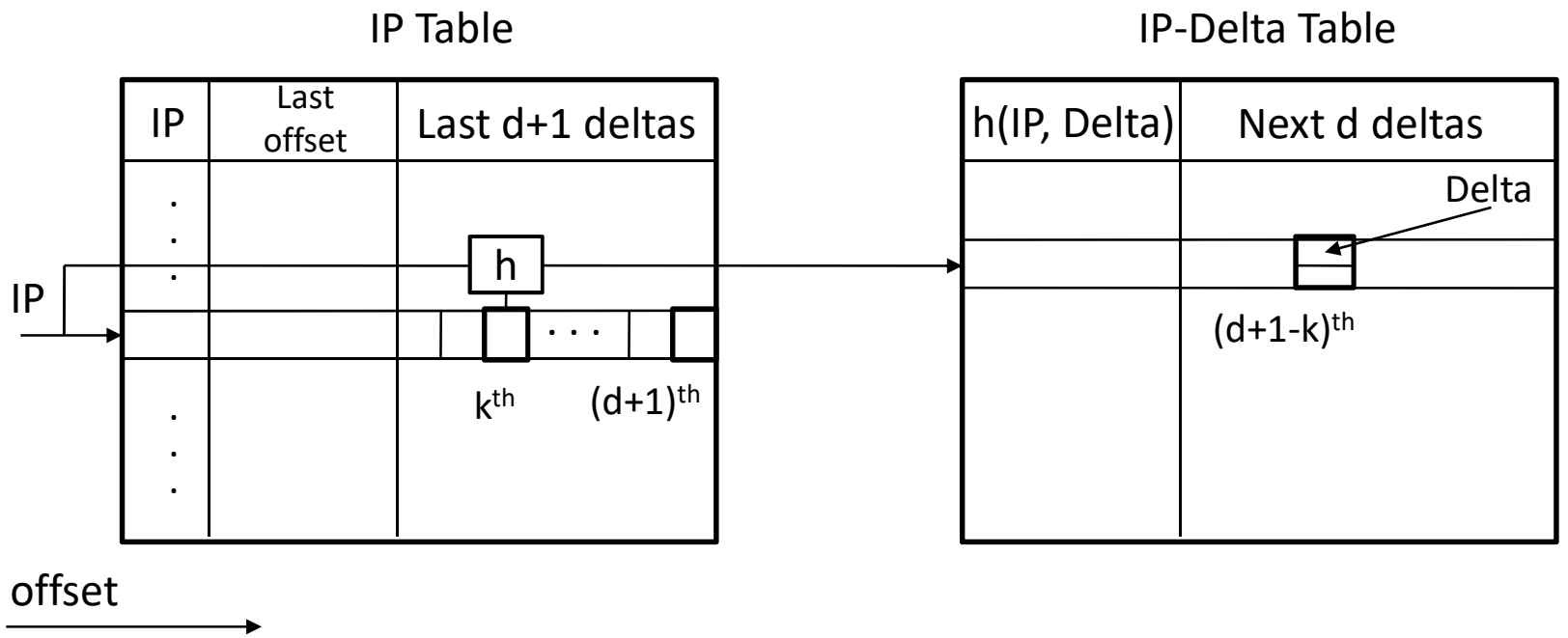
IP-Delta-based Sequence predictor

- Learning



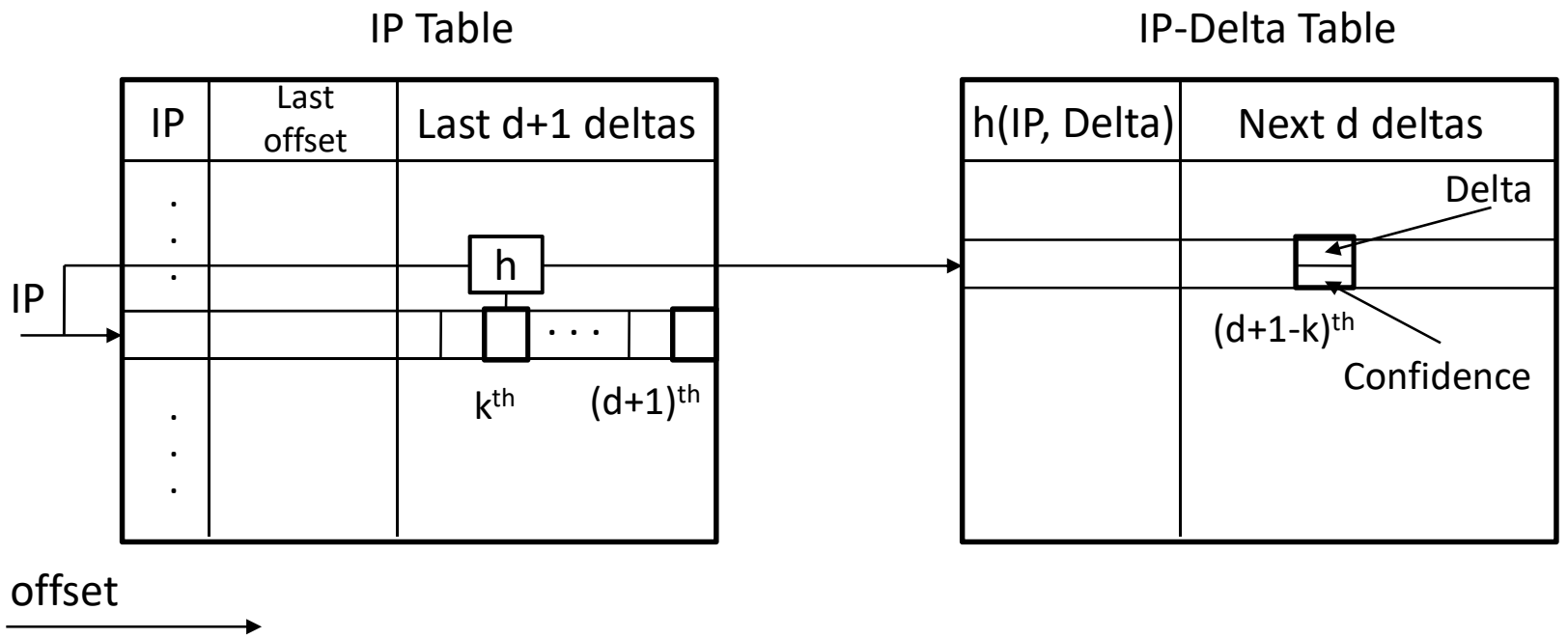
IP-Delta-based Sequence predictor

- Learning



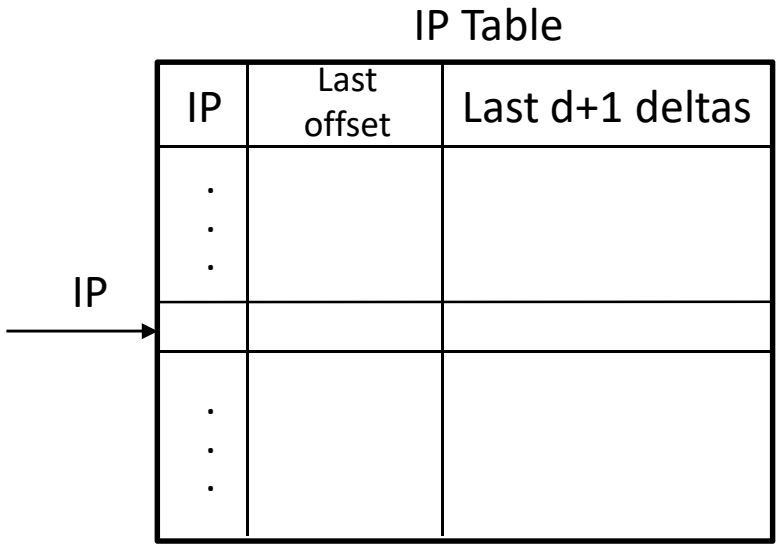
IP-Delta-based Sequence predictor

- Learning



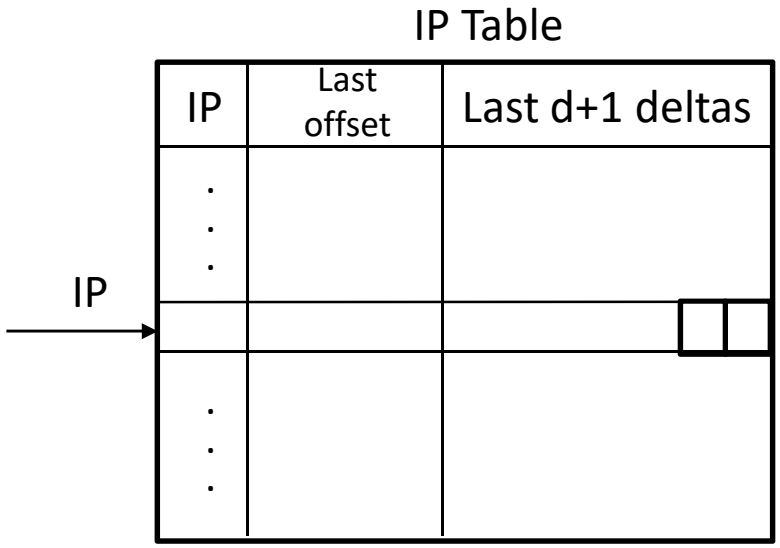
IP-based stride prefetcher

- We use IP based stride predictor when IP-delta predictor can no longer offer predictions
- This covers both cases when either the entry is missing from IP-delta table or the sequence is below confidence threshold



IP-based stride prefetcher

- We use IP based stride predictor when IP-delta predictor can no longer offer predictions
- This covers both cases when either the entry is missing from IP-delta table or the sequence is below confidence threshold



We use the IP stride predictor when the last two deltas seen for IP are equal

Next-line prefetcher

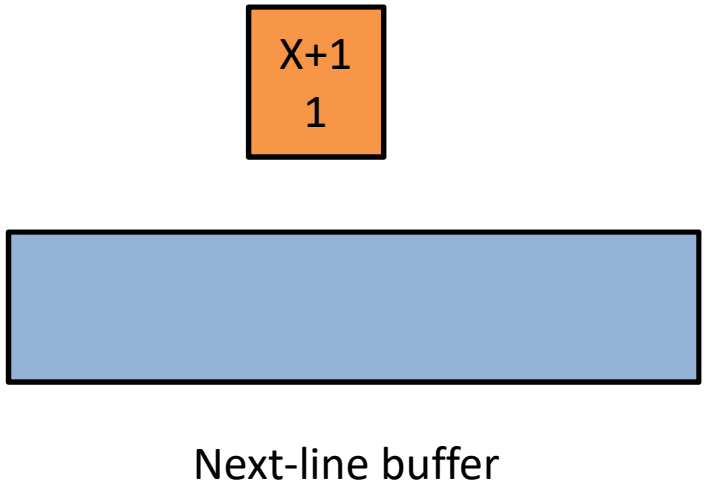
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

Next-line buffer

Degree	1	2	...	d
Hits	2	1	...	0
Insertions	4	4	...	4

Next-line prefetcher

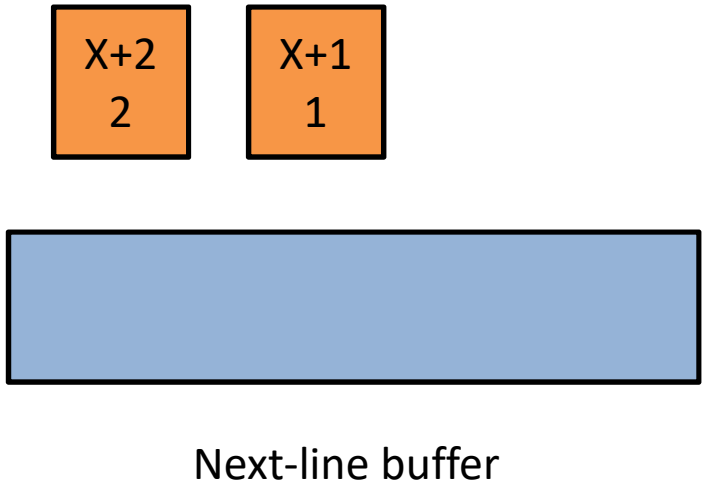
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	4	4	...	4

Next-line prefetcher

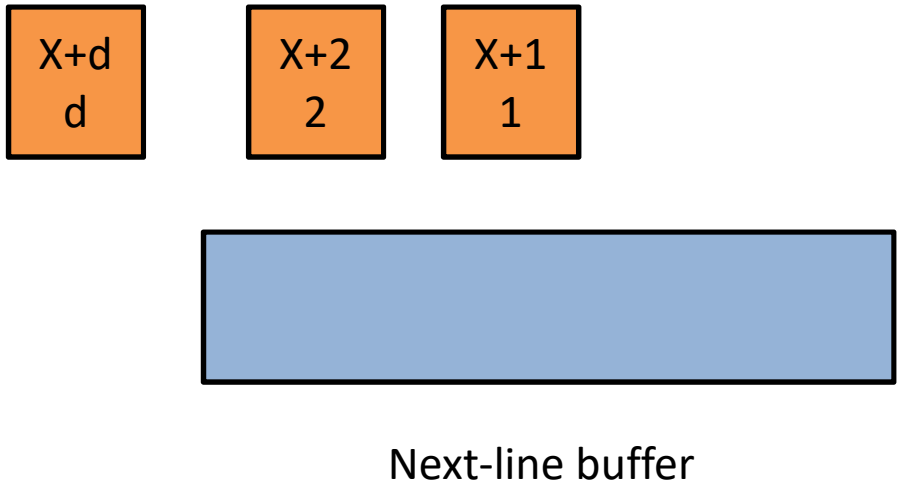
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	4	4	...	4

Next-line prefetcher

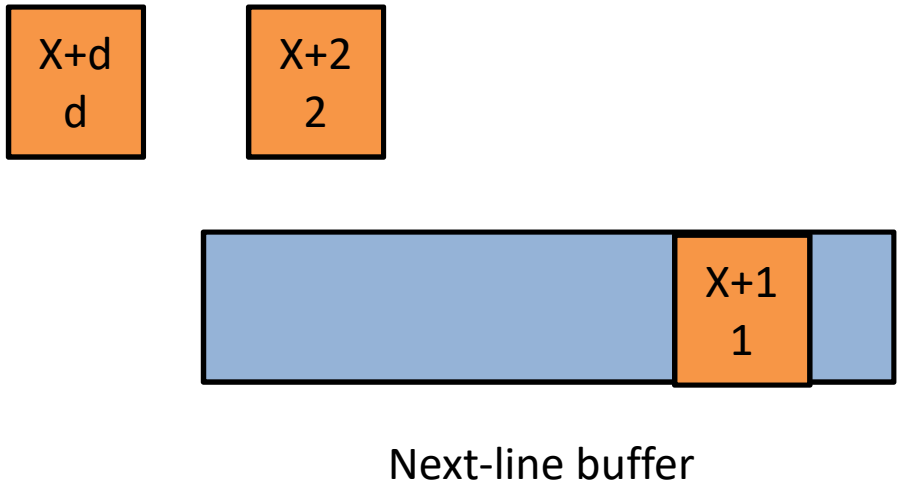
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	4	4	...	4

Next-line prefetcher

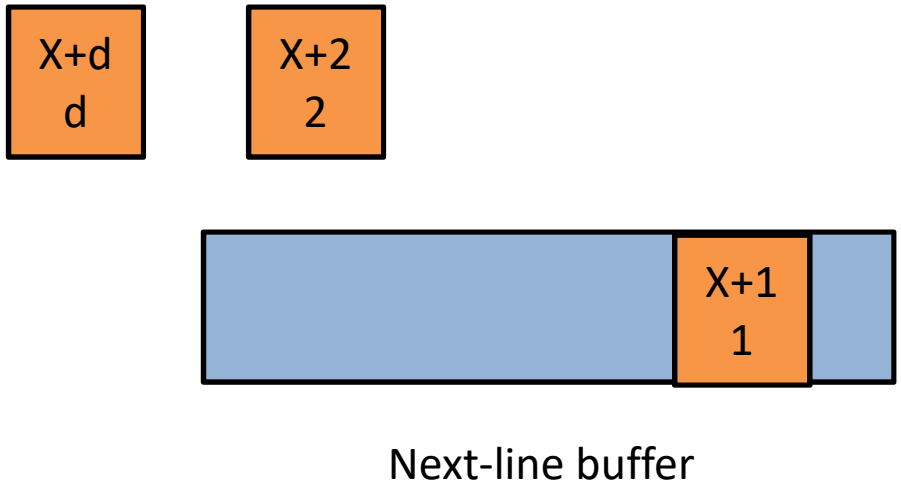
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	4	4	...	4

Next-line prefetcher

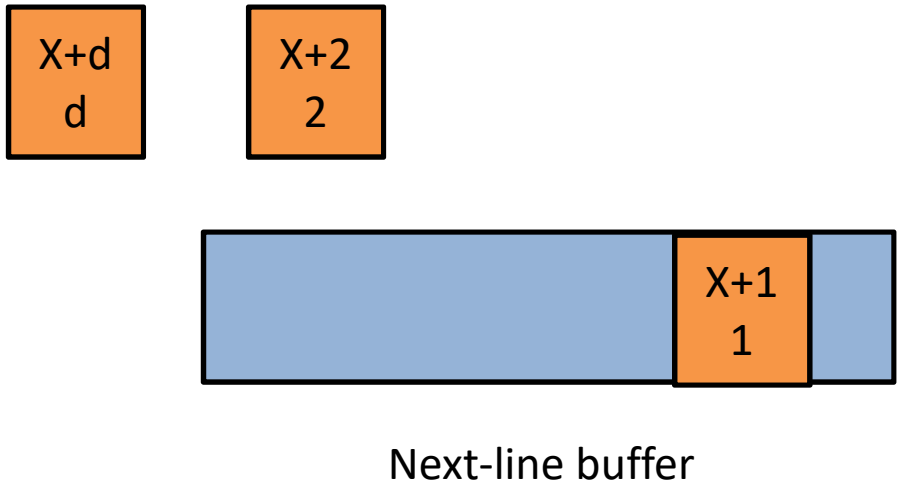
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	4	4	...	4

Next-line prefetcher

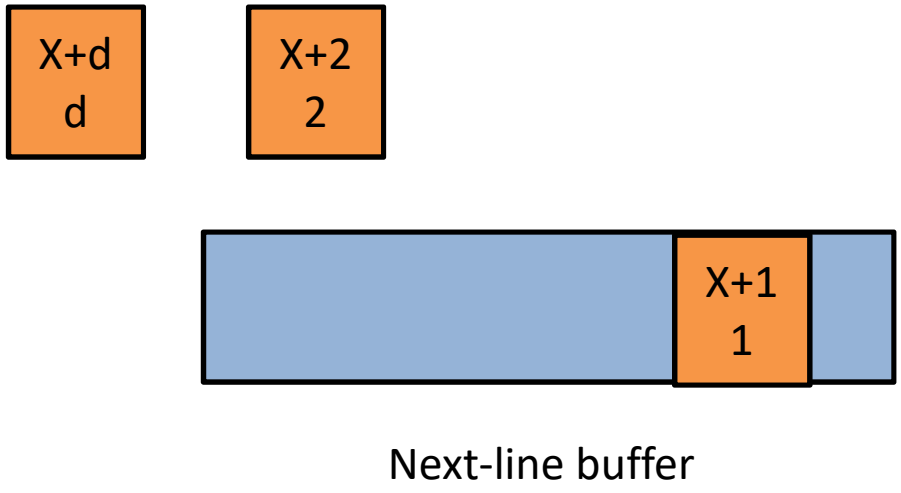
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	4	...	4

Next-line prefetcher

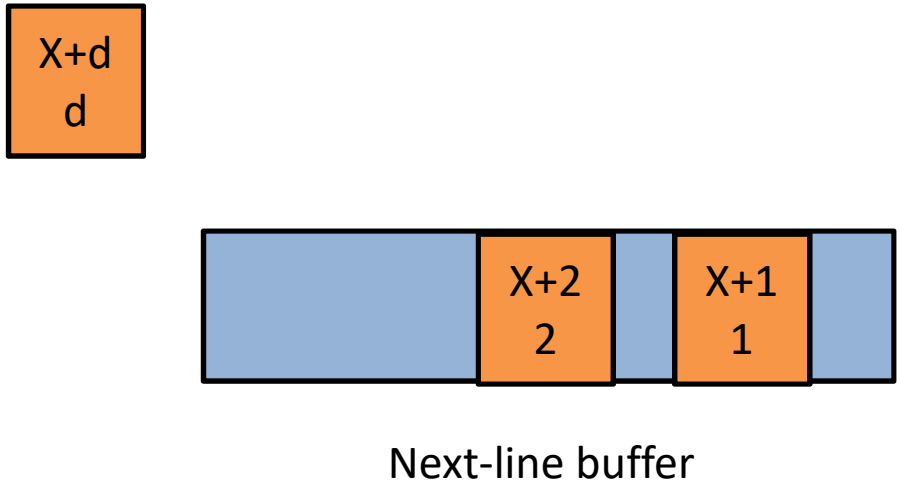
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	4	...	4

Next-line prefetcher

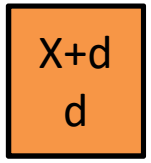
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

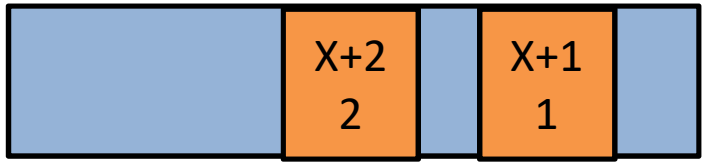


Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	4	...	4

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



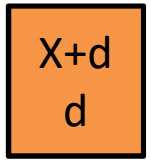


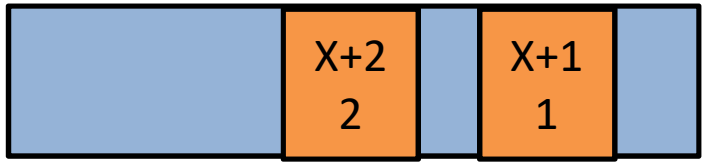
Next-line buffer

Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	4	...	4

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



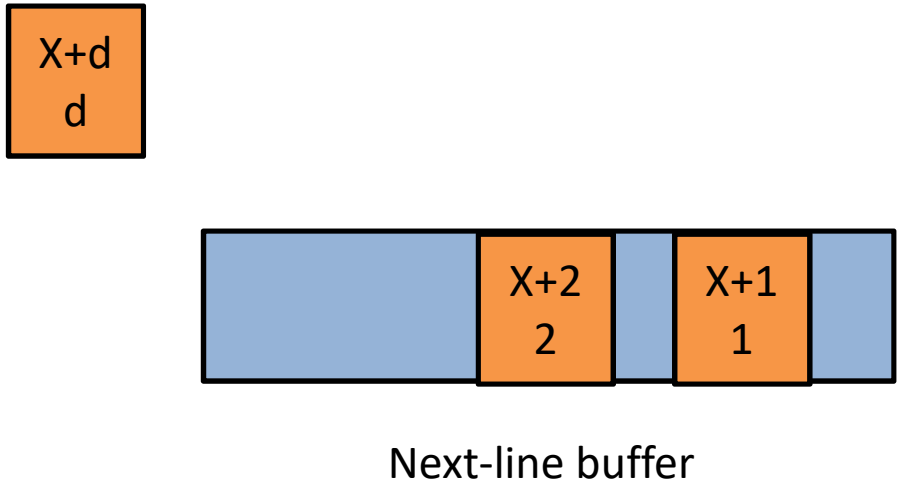


Next-line buffer

Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	4

Next-line prefetcher

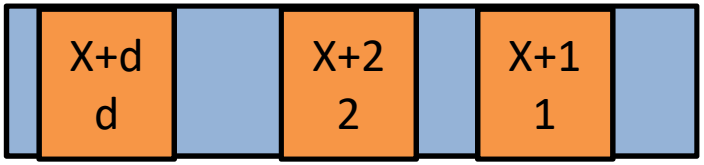
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	4

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

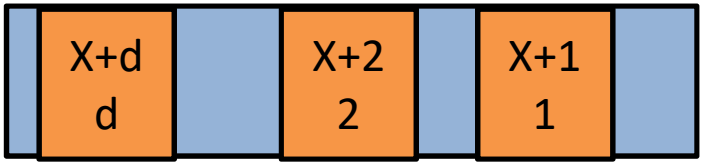


Next-line buffer

Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	4

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



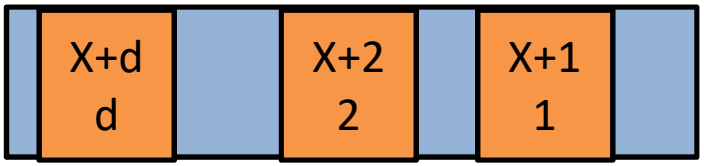
Next-line buffer

Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

Demand Access ⟹ X+1

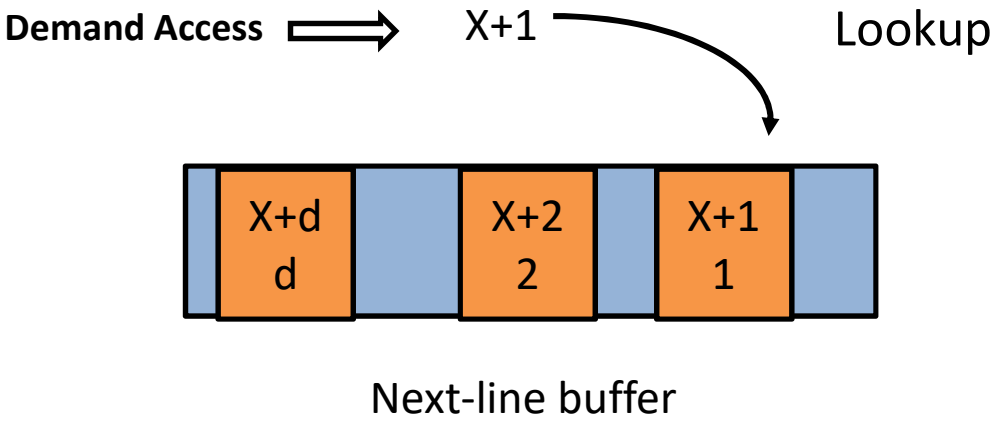


Next-line buffer

Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

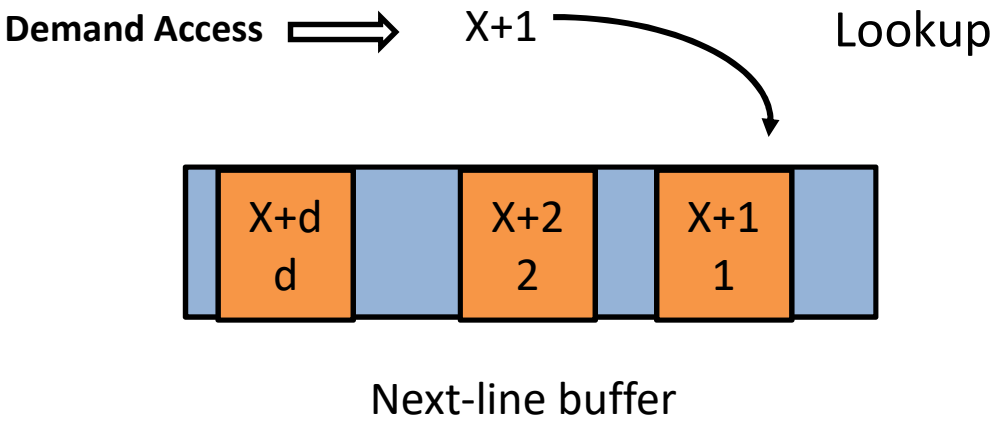
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

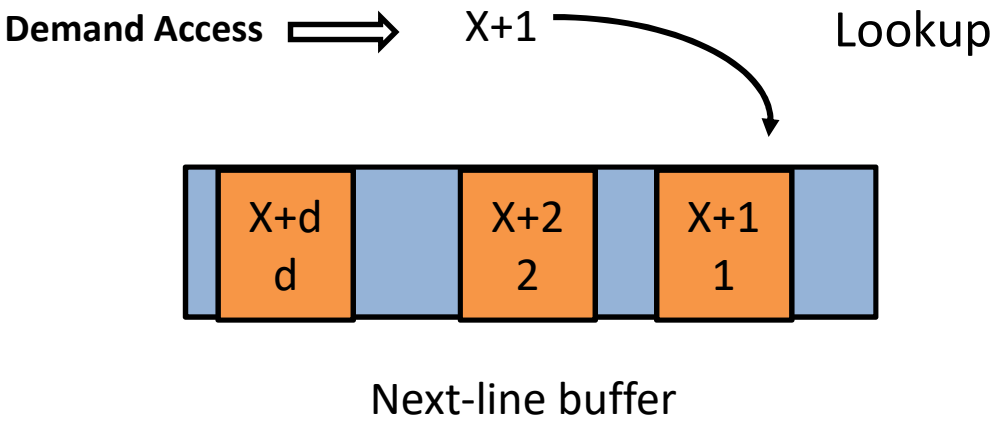
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	2	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

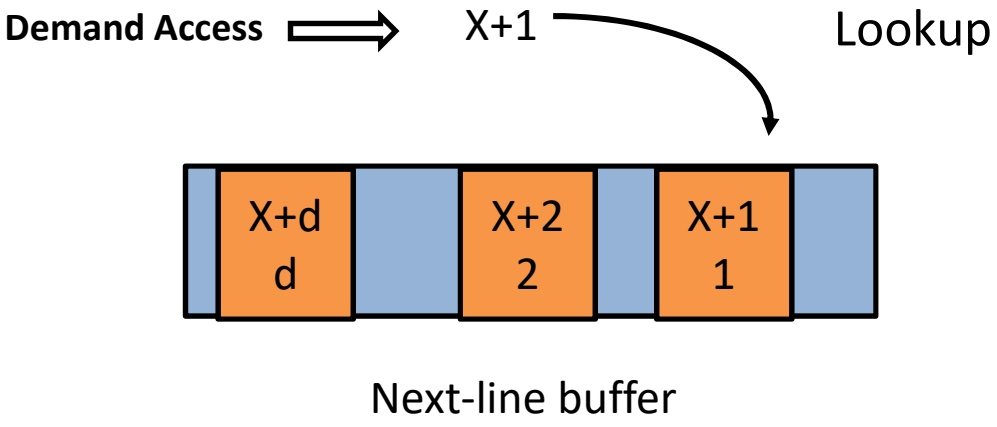
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Degree	1	2	...	d
Hits	3	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

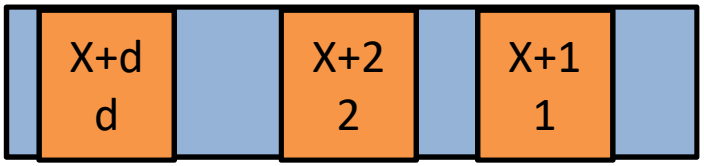


Degree	1	2	...	d
Hits	3	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

Demand Access ⟹ X+1

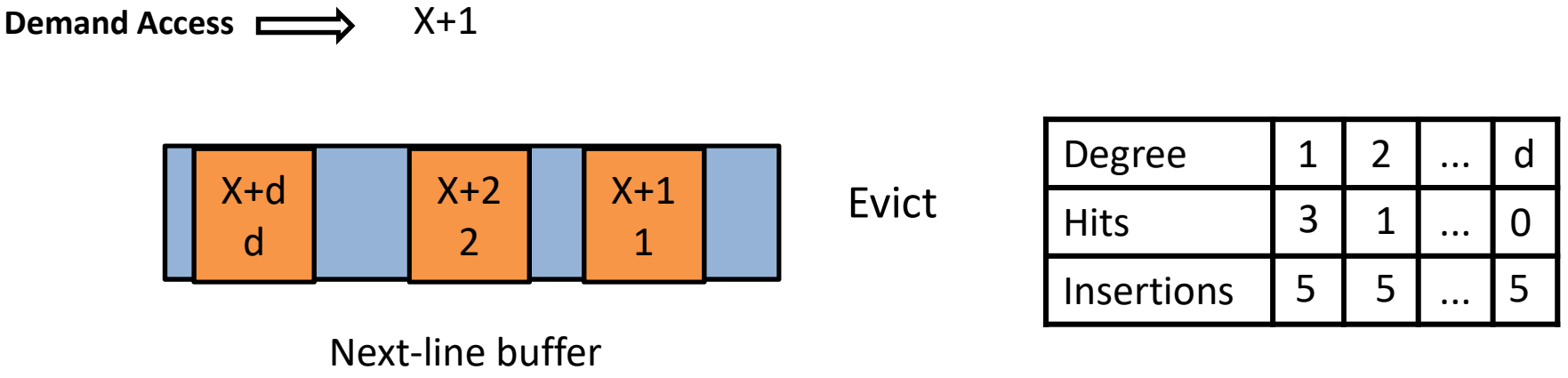


Next-line buffer

Degree	1	2	...	d
Hits	3	1	...	0
Insertions	5	5	...	5

Next-line prefetcher

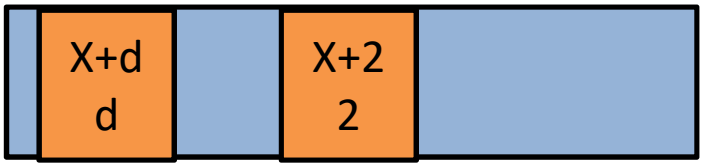
- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection



Next-line prefetcher

- Maintaining coverage at the cost of accuracy leads to overall better performance
- Used when both IP-delta and IP stride prefetcher cannot offer prediction
- Feedback directed degree selection

Demand Access ⟹ X+1



Next-line buffer

Evict

Degree	1	2	...	d
Hits	3	1	...	0
Insertions	5	5	...	5

Recent access filter

- Lot of overlapping prefetches due to 3 different components
- Could also be in a single component (e.g. Next-line)
- Should efficiently use Prefetch Queue
- Store the recent demand and prefetch accesses in a small fully associative buffer
- Only issue the prefetch requests if it misses in the recent access filter
- Small in size to avoid missing genuine requests

Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching

Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching

Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching

32 bits

Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching

32 bits

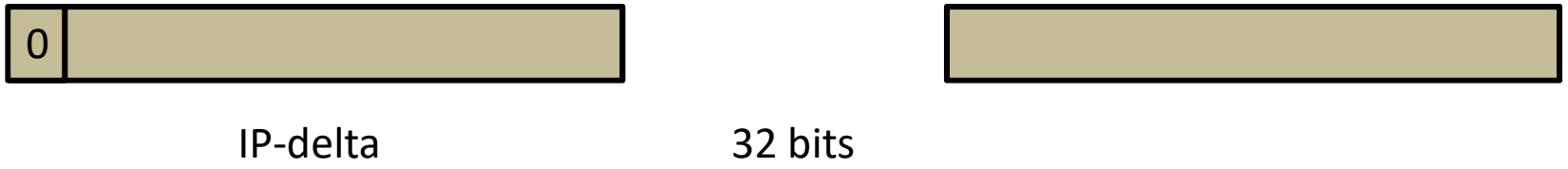
Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching

32 bits

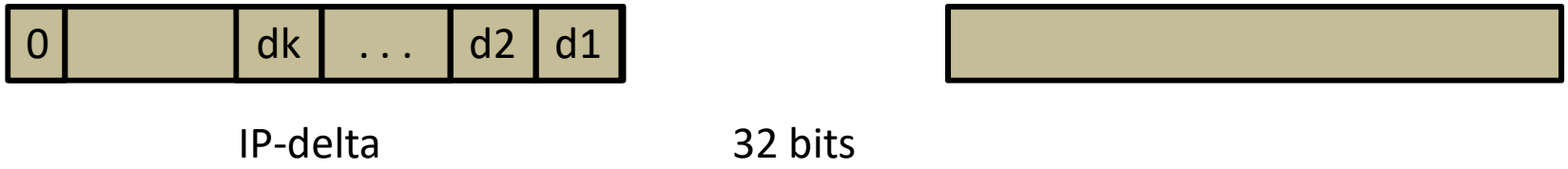
Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching



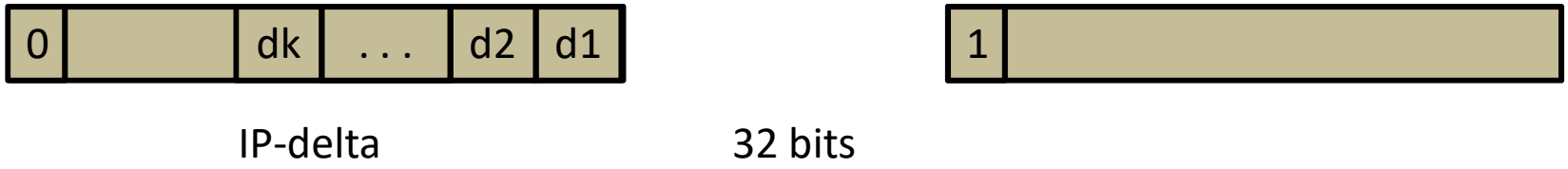
Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching



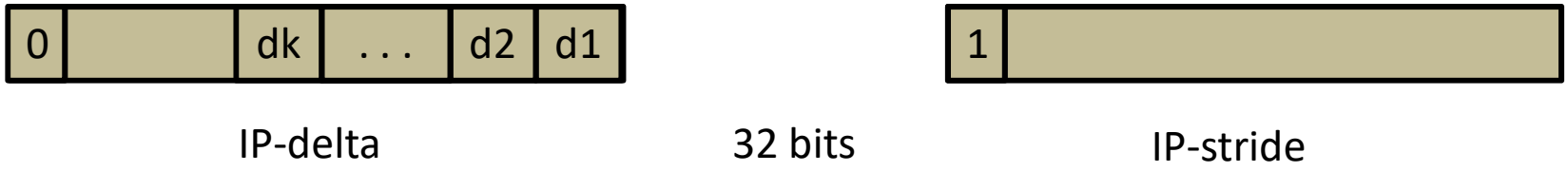
Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching



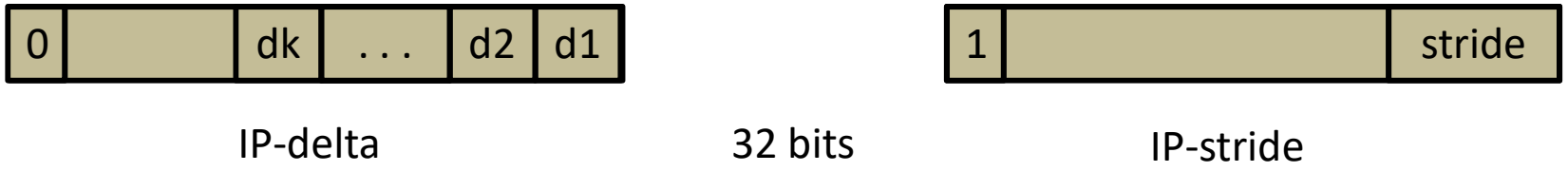
Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching

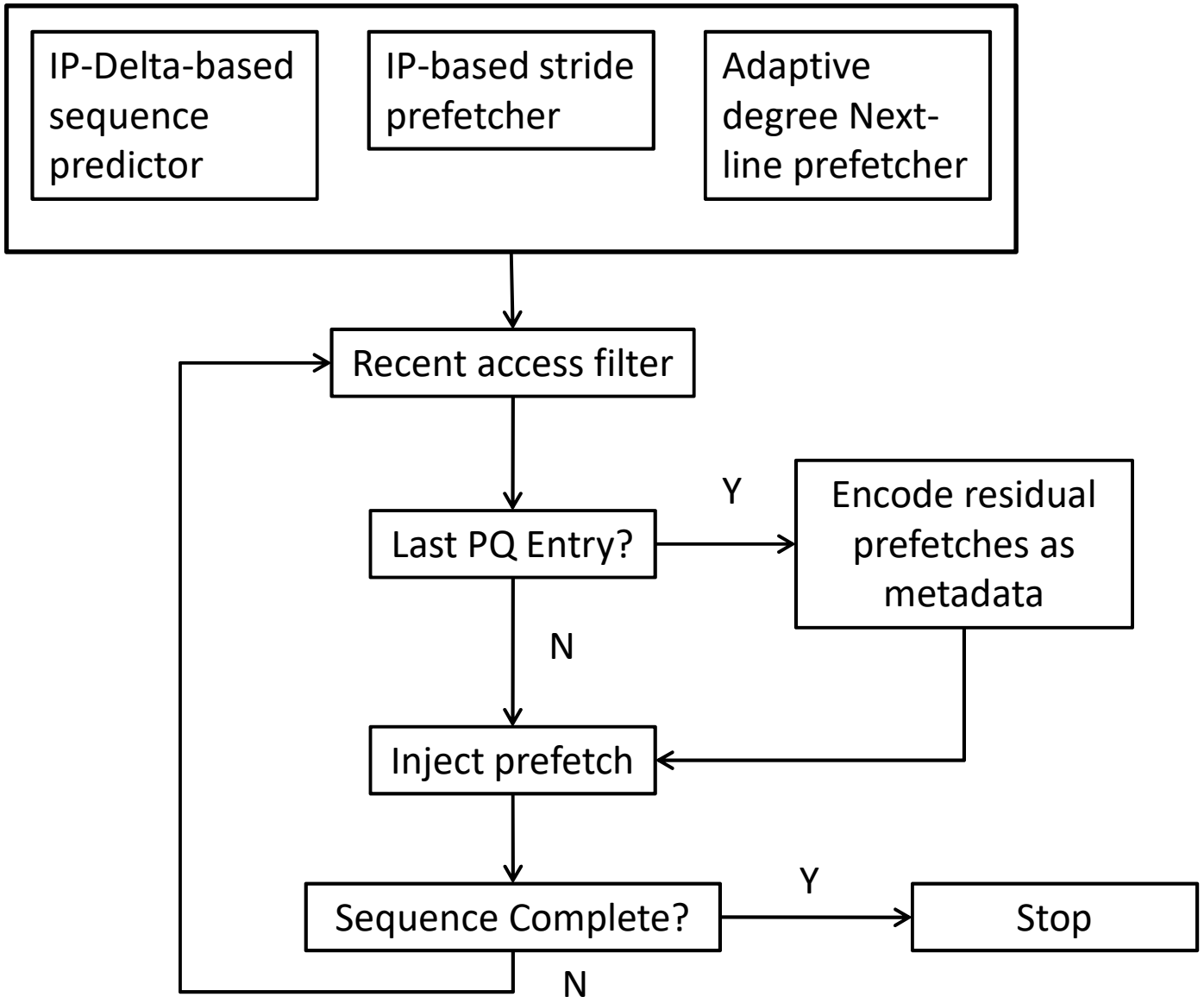


Handling resource shortage

- Short size of L1 prefetch queue restricts aggressive prefetching
- Leverage the communication b/w L1 and L2 prefetcher
- When the PQ has only entry left then we piggyback the remaining prefetch info with the last prefetch
- L2 cache uses this info to complete the prefetching



Sangam


IP-Delta-based sequence predictor
IP-based stride prefetcher
Adaptive degree Next-line prefetcher
Recent access filter
Last PQ Entry?
Y
Encode residual prefetches as metadata
N
Inject prefetch
Sequence Complete?
Y
Stop
N

Storage Overhead

- L1 prefetcher overhead

Structure		Storage (bits)	TOTAL
IP Table	128 sets, 15 ways	120960	259870 bits = **31.72 KB**
IP-Delta Table	256 sets, 8 ways	131072	
NL buffer	64 entries	4672	
Recent Access Filter	40 entries	2840	
Auxiliary Counters	316 bits		

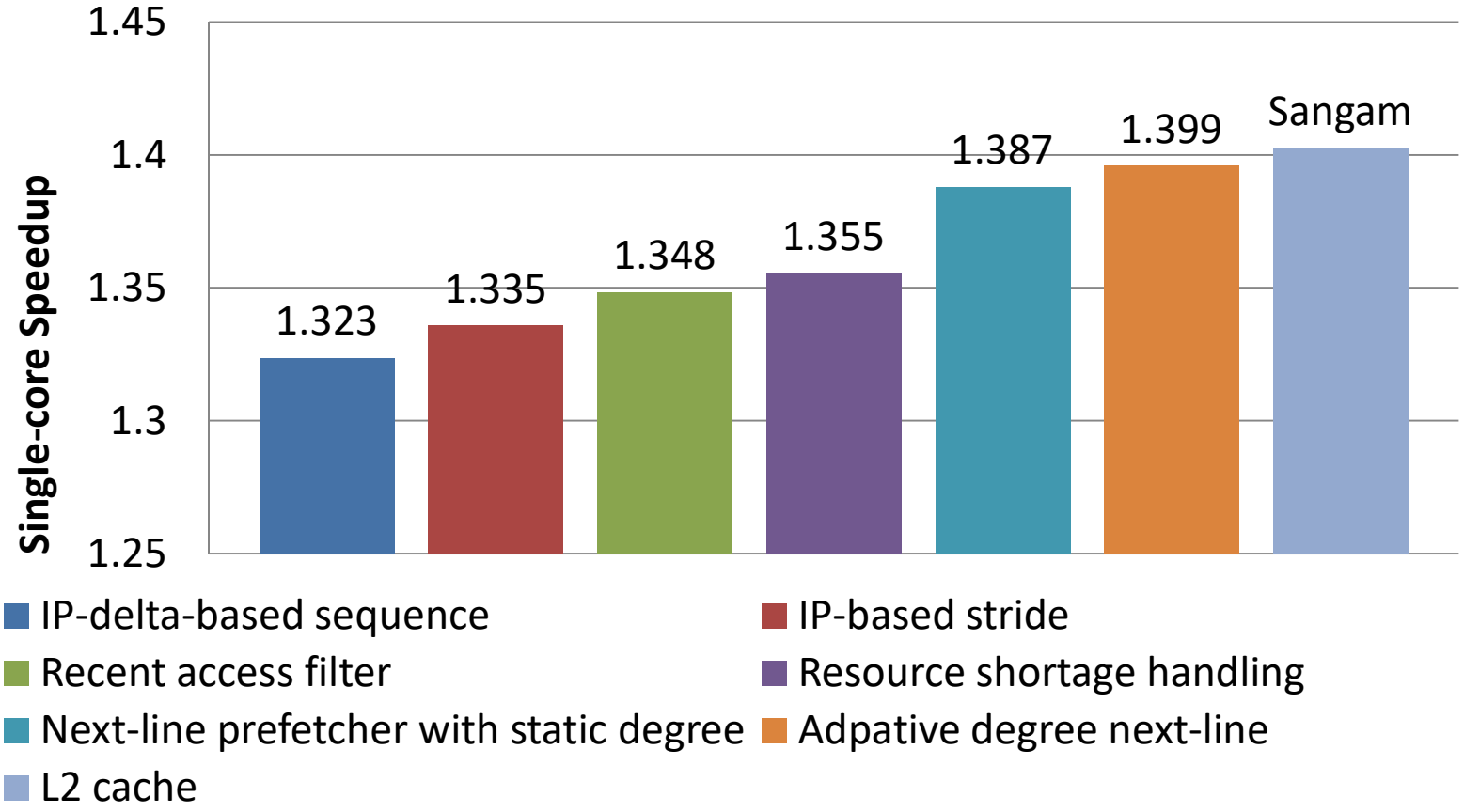
Storage Overhead

- L1 prefetcher overhead

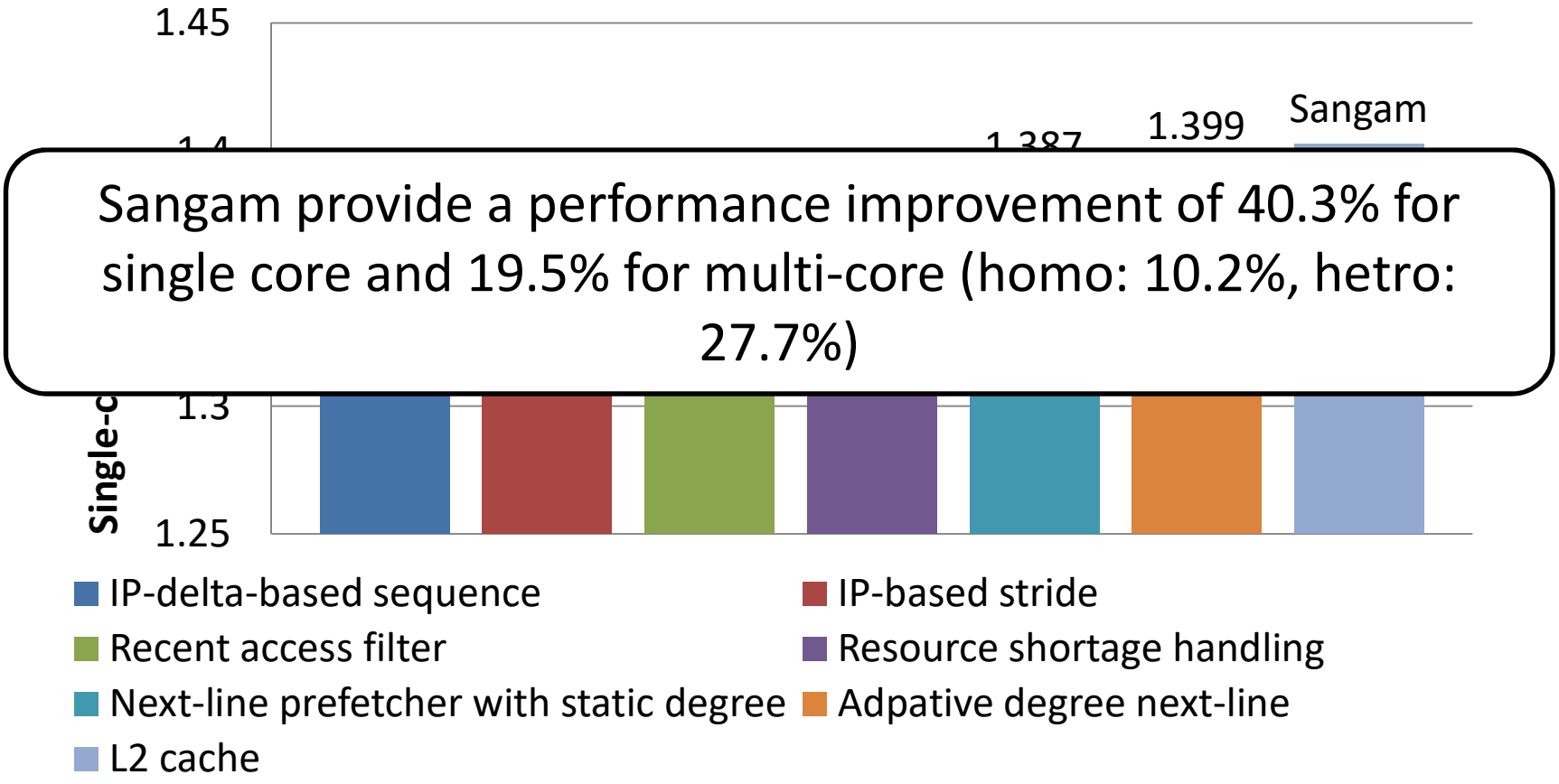
Structure		Storage (bits)	TOTAL
IP Table	128 sets, 15 ways	120960	259870 bits = **31.72 KB**
IP-Delta Table	256 sets, 8 ways	131072	
NL buffer	64 entries	4672	
Recent Access Filter	40 entries	2840	
Auxiliary Counters	316 bits		

- L2 prefetcher overhead = 31.36 KB

Performance distribution



Performance distribution



Sangam provide a performance improvement of 40.3% for single core and 19.5% for multi-core (homo: 10.2%, hetro: 27.7%)

Thank You

Questions?