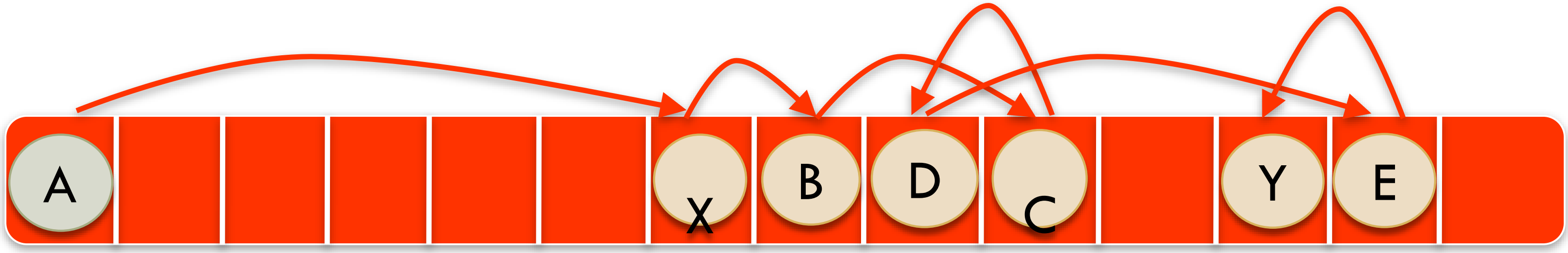
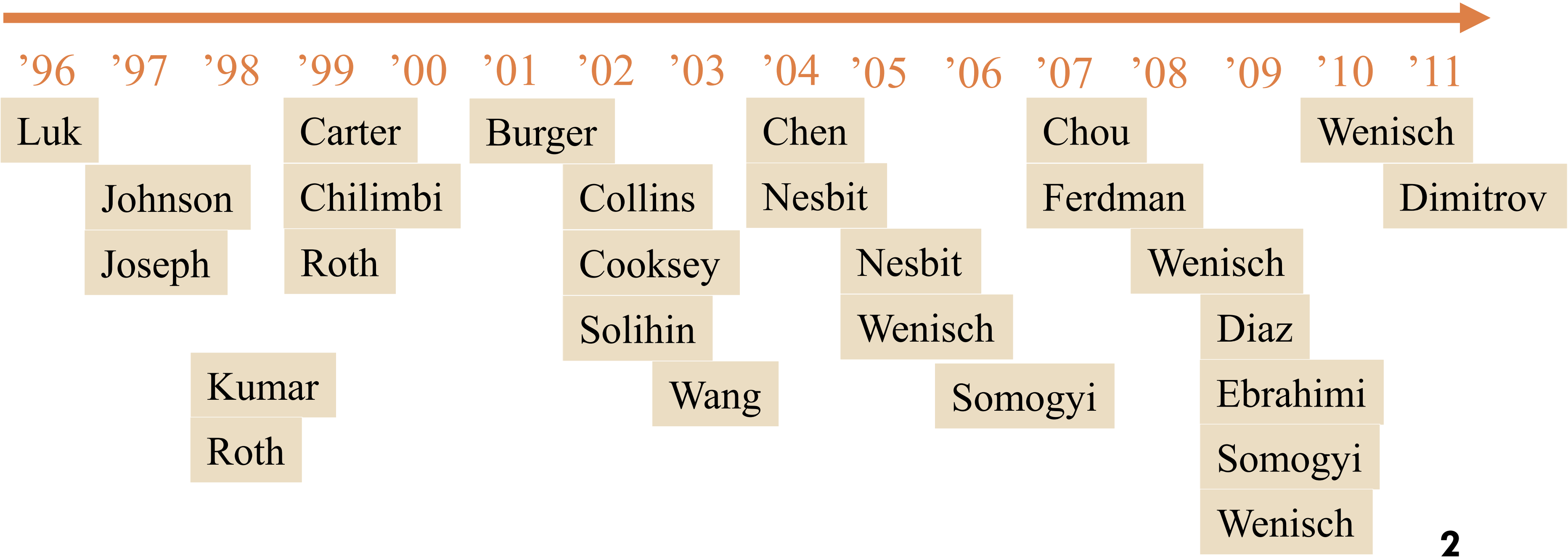
Linearizing Irregular Memory Accesses for Improved Correlated Prefetching

Akanksha Jain, Calvin Lin

University of Texas at Austin

The Problem : Irregular Prefetching





'96 '97 '98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11

Luk

Johnson

Joseph

Kumar

Roth

Carter

Chilimbi

Roth

Burger

Collins

Cooksey

Solihin

Wang

Chen

Nesbit

Nesbit

Wenisch

Somogyi

Chou

Ferdman

Wenisch

Diaz

Ebrahimi

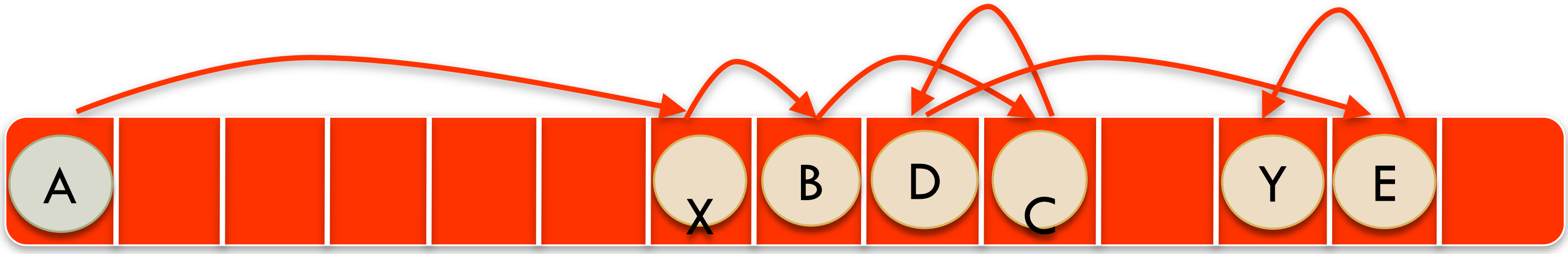
Somogyi

Wenisch

Wenisch

Dimitrov

The Problem : Irregular Prefetching



'96 '97 '98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11

Luk

Johnson

Joseph

Kumar

Roth

Carter

Chilimbi

Roth

Burger

Collins

Cooksey

Solihin

Wang

Chen

Nesbit

Nesbit

Wenisch

Somogyi

Chou

Ferdman

Wenisch

Diaz

Ebrahimi

Somogyi

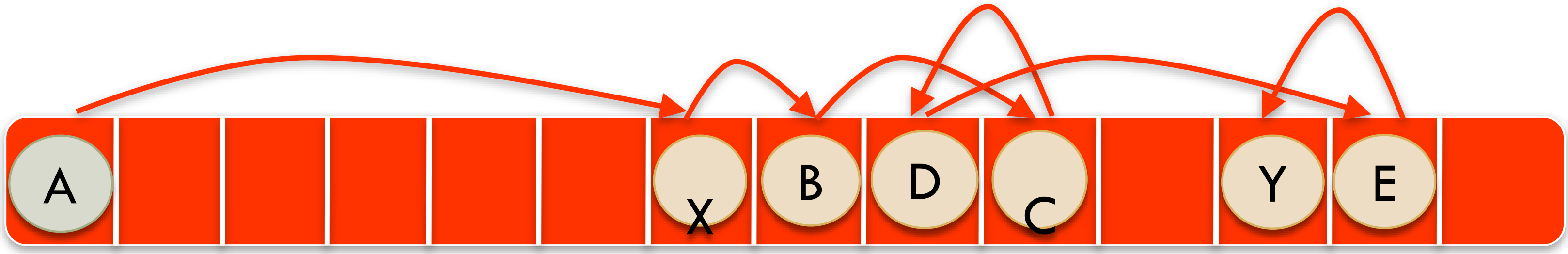
Wenisch

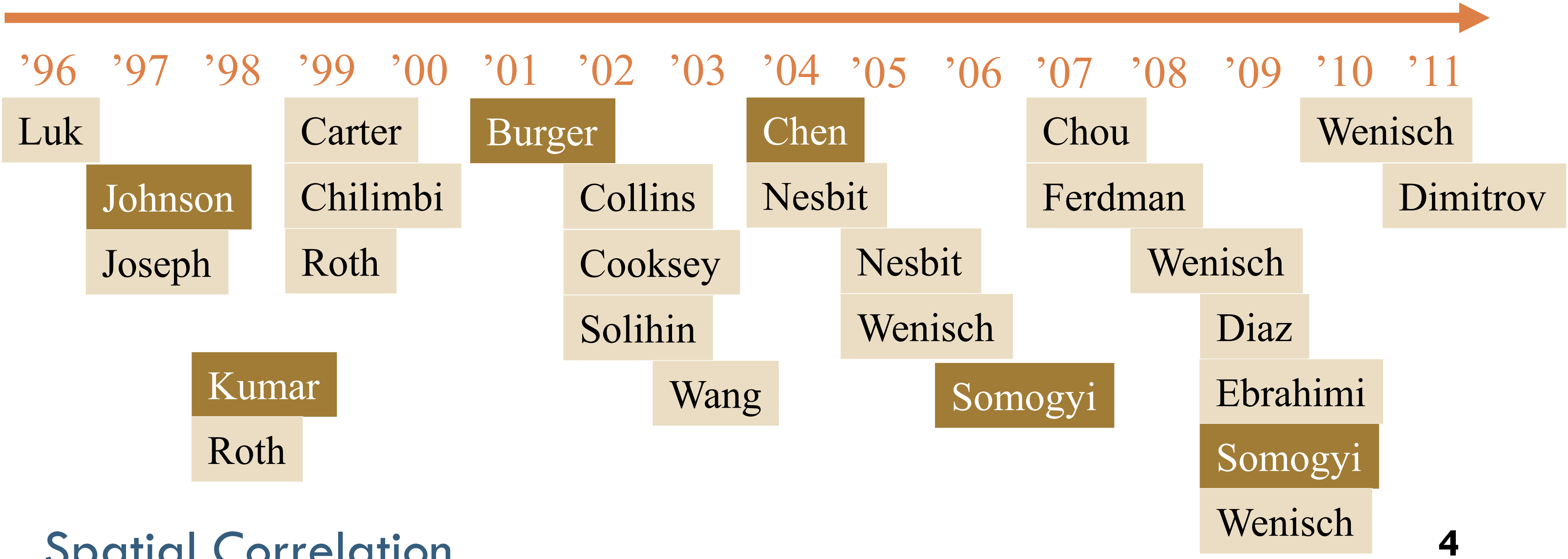
Wenisch

Dimitrov

Pointer-based

The Problem : Irregular Prefetching





'96 '97 '98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11

Luk

Johnson

Joseph

Kumar

Roth

Carter

Chilimbi

Roth

Burger

Collins

Cooksey

Solihin

Wang

Chen

Nesbit

Nesbit

Wenisch

Somogyi

Chou

Ferdman

Wenisch

Diaz

Ebrahimi

Somogyi

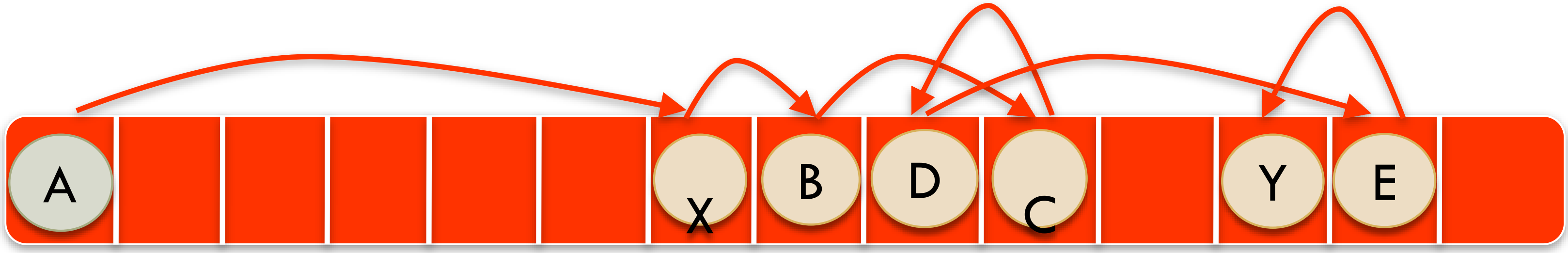
Wenisch

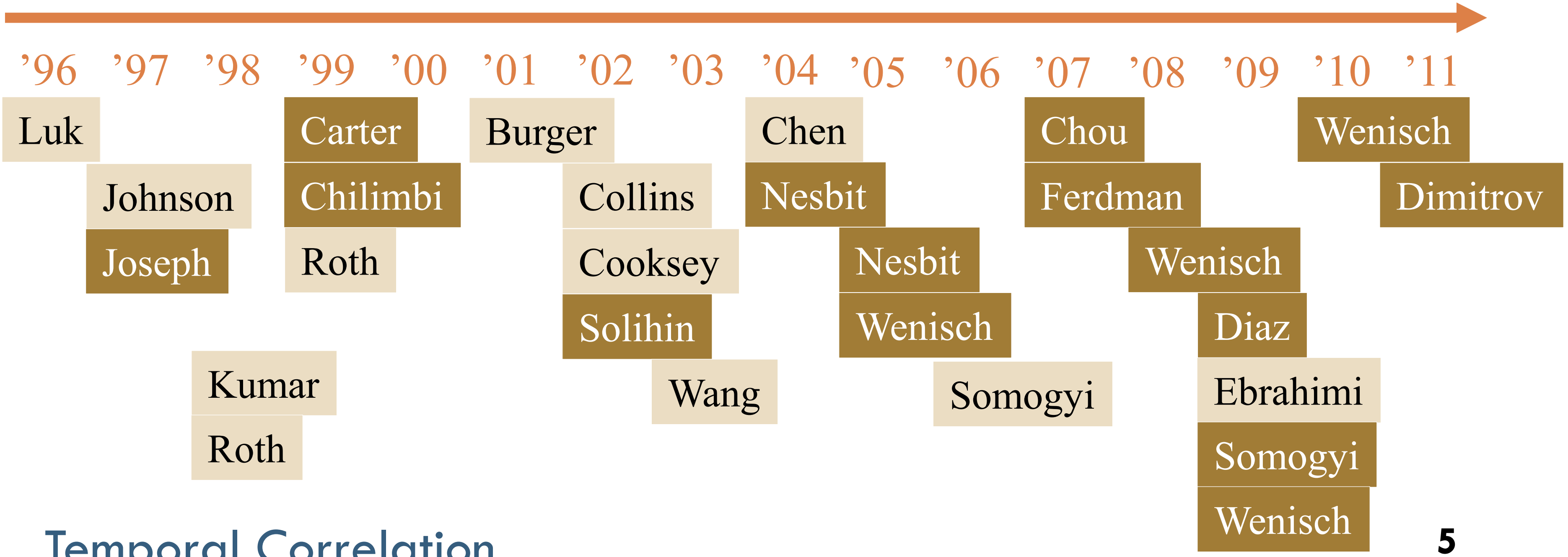
Wenisch

Dimitrov

Spatial Correlation

The Problem : Irregular Prefetching




'96 '97 '98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11

Luk

Johnson

Joseph

Kumar

Roth

Carter

Chilimbi

Roth

Burger

Collins

Cooksey

Solihin

Wang

Chen

Nesbit

Nesbit

Wenisch

Somogyi

Chou

Ferdman

Wenisch

Diaz

Ebrahimi

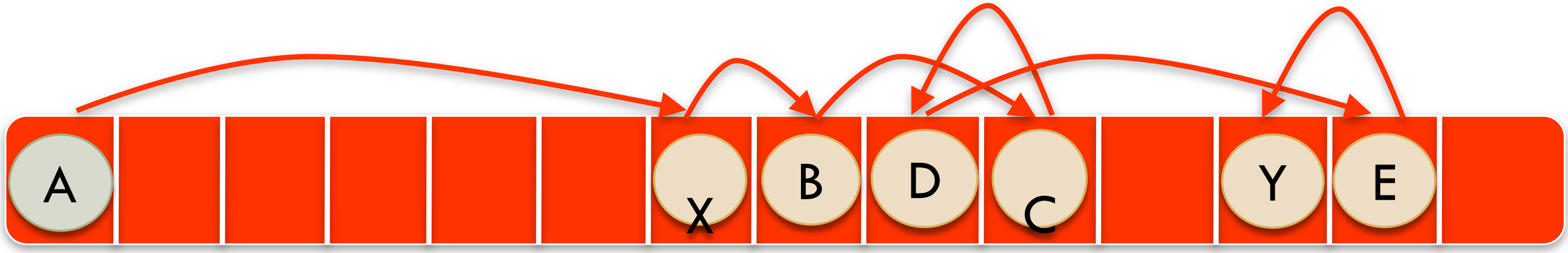
Somogyi

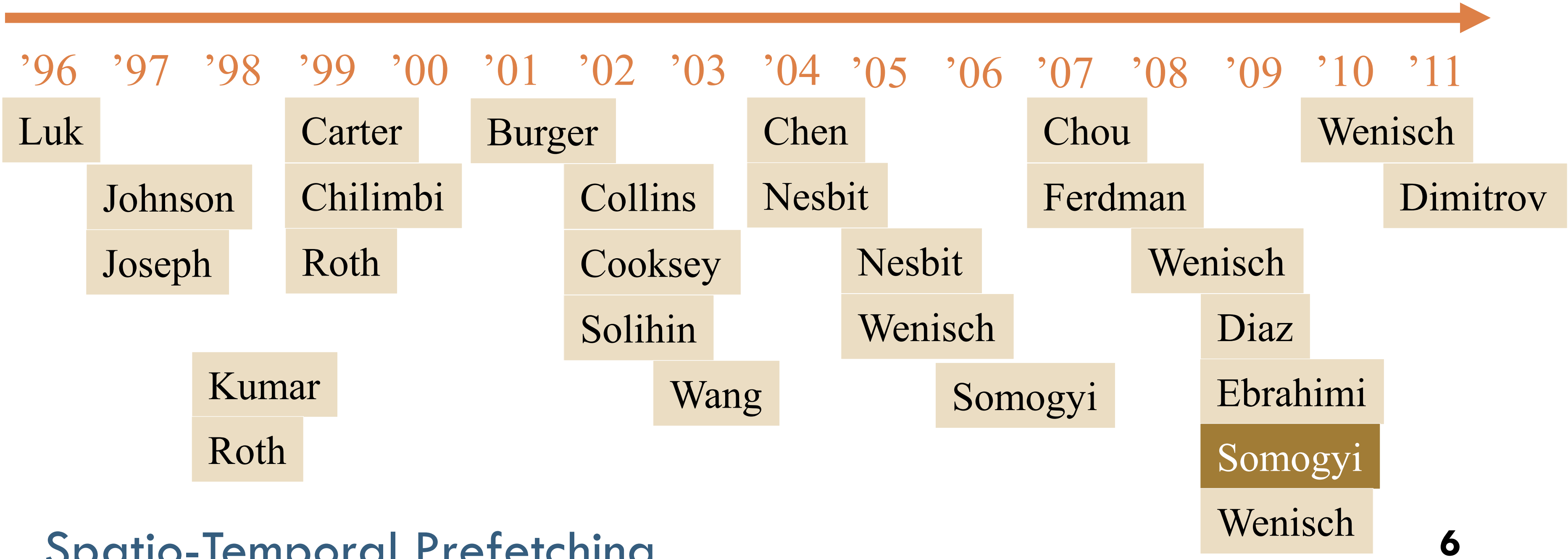
Wenisch

Wenisch

Dimitrov

The Problem : Irregular Prefetching





’96 ’97 ’98 ’99 ’00 ’01 ’02 ’03 ’04 ’05 ’06 ’07 ’08 ’09 ’10 ’11

Luk

Johnson
Joseph

Kumar
Roth

Carter
Chilimbi
Roth

Burger

Collins
Cooksey
Solihin

Wang

Chen
Nesbit

Nesbit
Wenisch

Somogyi

Chou
Ferdman

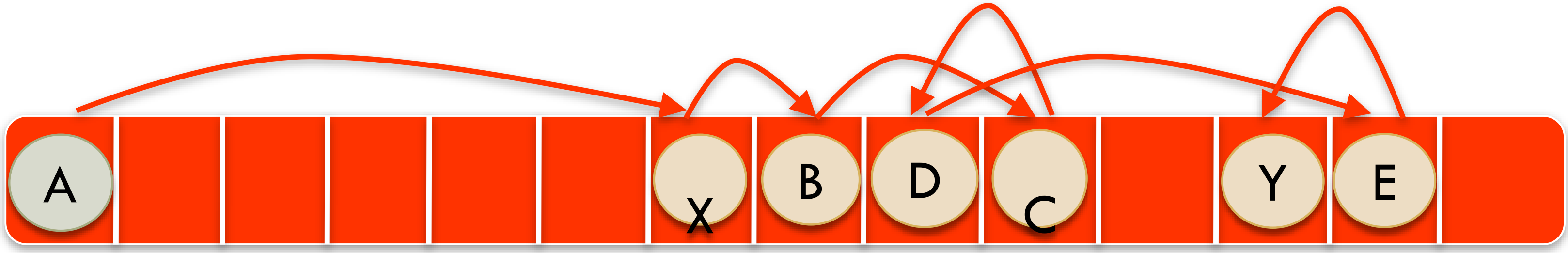
Wenisch

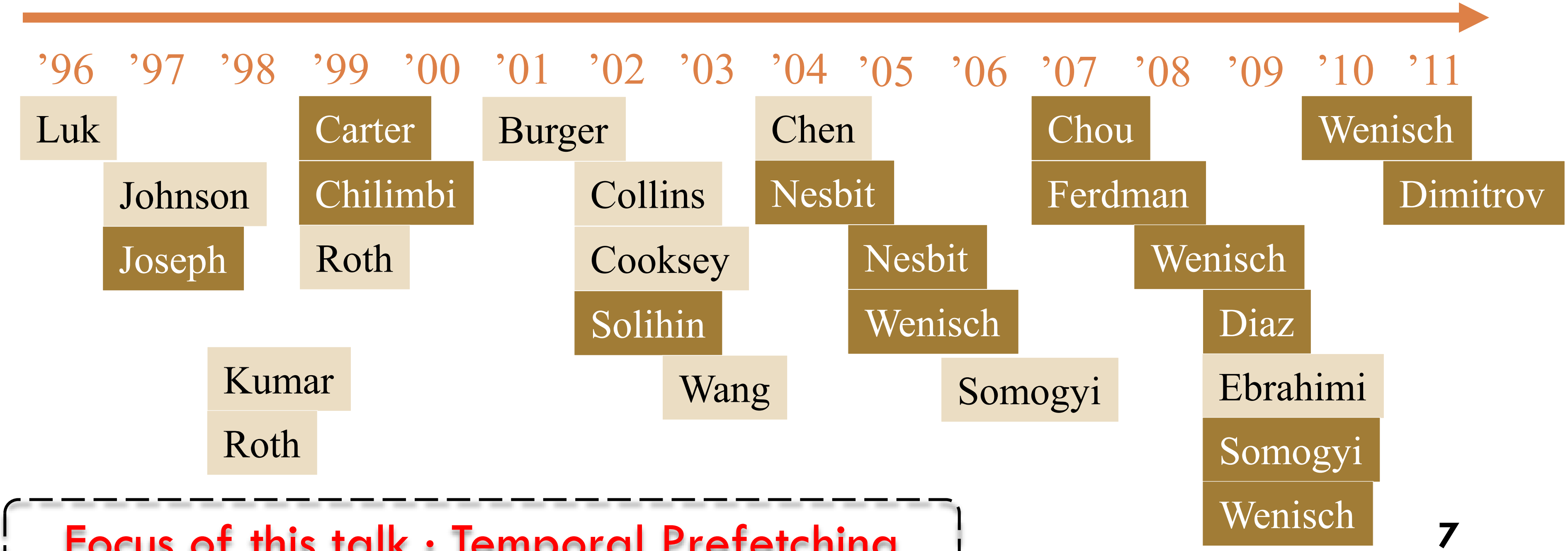
Diaz
Ebrahimi
Somogyi
Wenisch

Wenisch

Dimitrov

The Problem : Irregular Prefetching





'96 '97 '98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11

Luk

Johnson

Joseph

Kumar

Roth

Carter

Chilimbi

Roth

Burger

Collins

Cooksey

Solihin

Wang

Chen

Nesbit

Nesbit

Wenisch

Somogyi

Chou

Ferdman

Wenisch

Diaz

Ebrahimi

Somogyi

Wenisch

Wenisch

Dimitrov

Focus of this talk : Temporal Prefetching

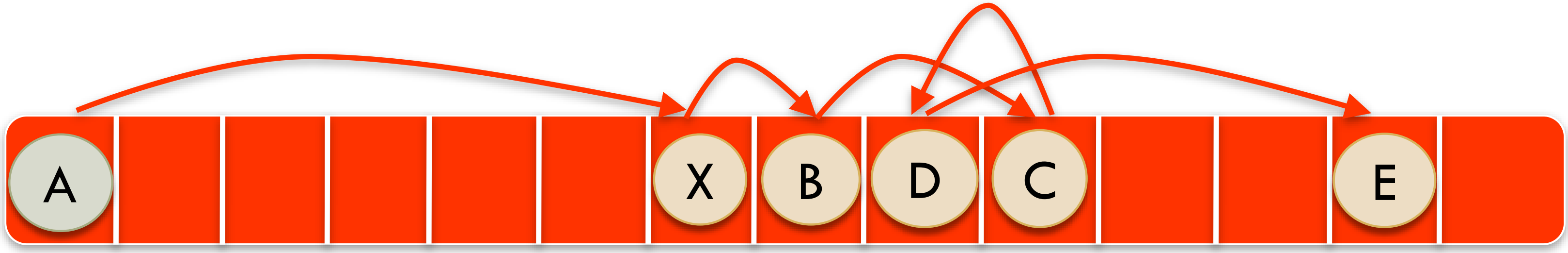
Our Contribution

- Significantly advance the state-of-the-art in temporal prefetching

	Previous Best	Our Solution
Speedup	8.3%	23.1%
Accuracy	58.6%	93.7%

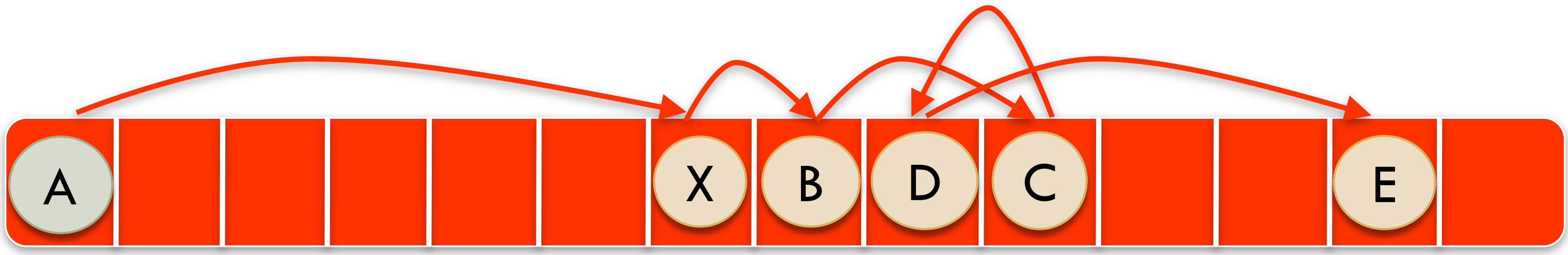
- How do we achieve this improvement?
  - Address Correlation [Grunwald 97]
  - PC-localization [Nesbit 04, Somogyi 06]

Address Correlation



- A and X are *temporally correlated if the* occurrence of A is very likely to be followed by X
- *Temporal streams:* Sequences of correlated accesses
  - Highly variable lengths from 2 to several hundreds [Chilimbi 02, Wenisch 08]

Address Correlation is Expensive

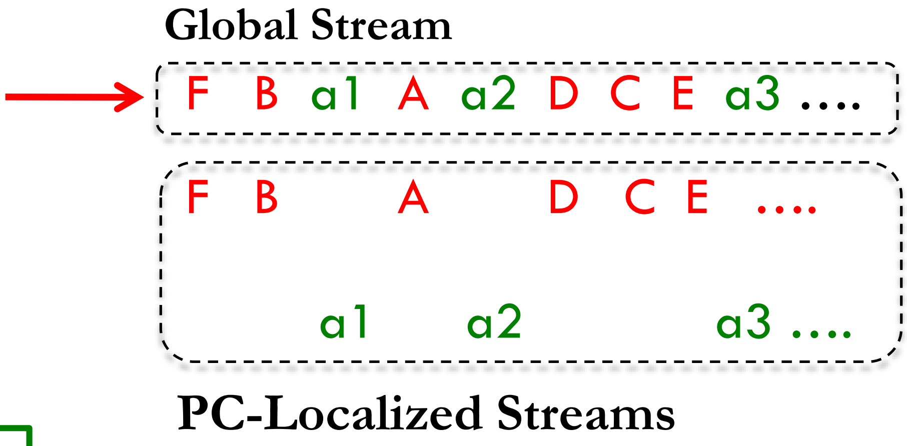


- Large storage requirements

- *Weaker forms of Correlation:* Learn correlated deltas between consecutive memory addresses
  - Correlate distance(A,X) with distance(X,B)
  - Sacrifice performance for lower storage requirements

PC-Localization

```
while ( ! end )
{
    read tree->next;
    if( condition)
        read linked_list->next;
 }
```



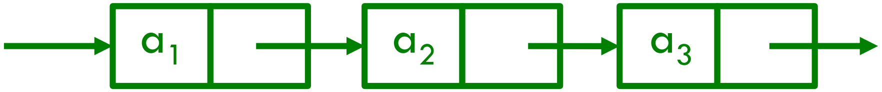
Global Stream

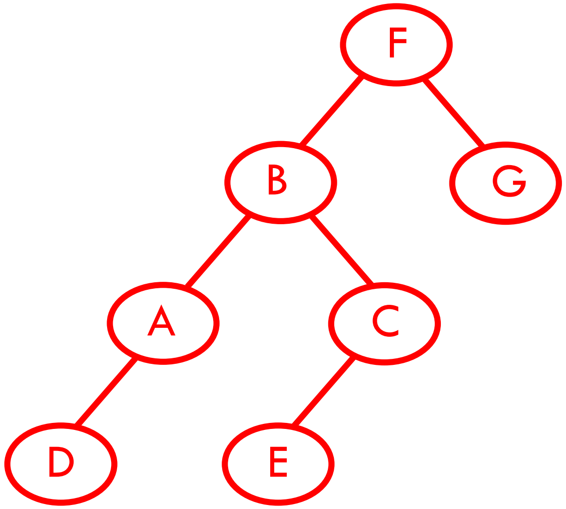
F B a1 A a2 D C E a3 ….

F B A D C E ….

a1 a2 a3 ….

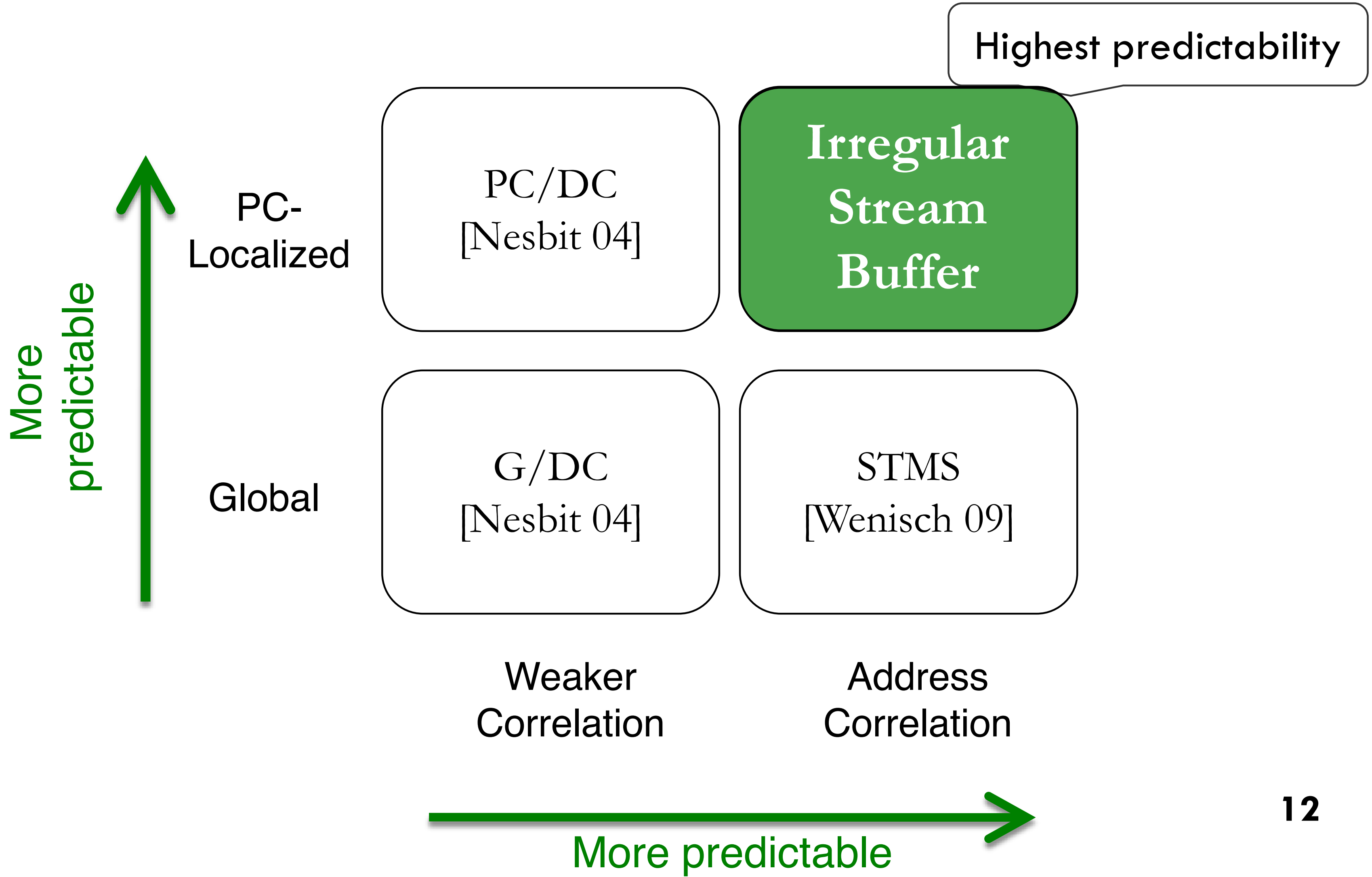
PC-Localized Streams





- PC-localization: Segregate the global stream by the load instruction’s PC
- PC-localized streams are more predictable!

Design Space for Temporal Prefetching

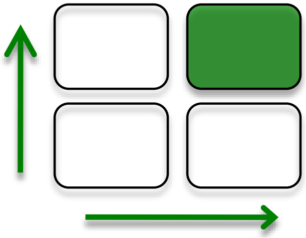


Highest predictability

	Weaker Correlation	Address Correlation
PC-Localized	PC/DC [Nesbit 04]	Irregular Stream Buffer
Global	G/DC [Nesbit 04]	STMS [Wenisch 09]

Our Contribution

- Significantly advance the state-of-the-art
- How do we achieve this improvement?
  - First to combine address correlation and PC-localization
  - Two more benefits!



- Why is it hard to combine address correlation and PC-localization?
  - Prefetcher Implementation
    - Global History Buffer (GHB)

Address Correlation with Global History Buffer

Global Access Stream:

A X B C D Y E … A X B C …

History Buffer
A
X
B
C
D
Y
E
…
A
X
B
C

History Buffer

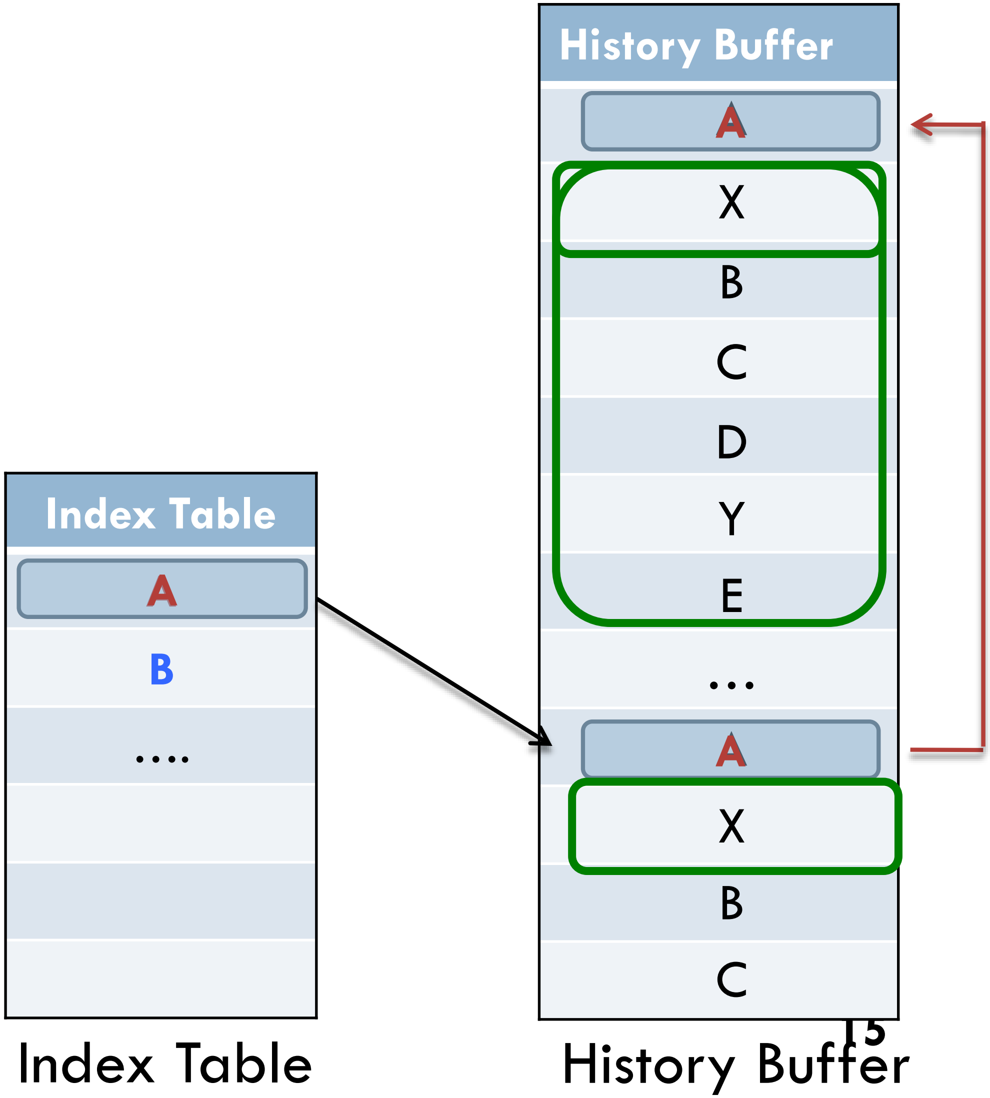
Address Correlation with Global History Buffer

Global Access Stream:

A X B C D Y E … A X B C …

Input : A

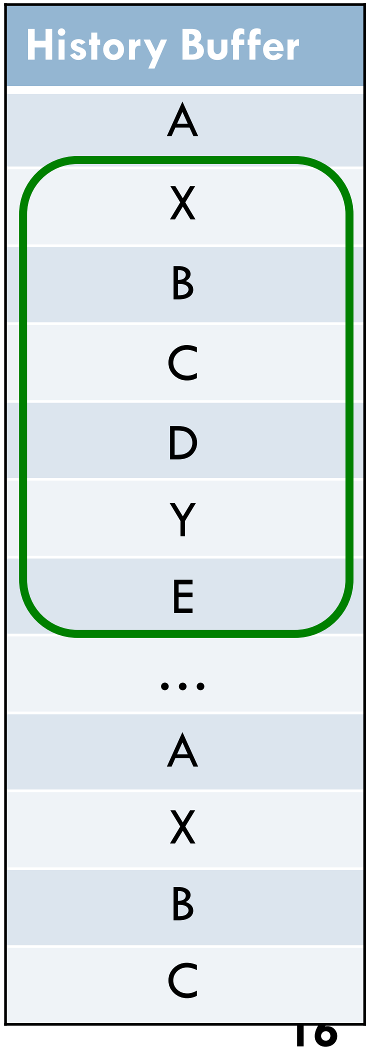
Predict X



Address Correlation with Global History Buffer

Large history buffer

- 10-100 MB [Wenisch 09]
- Stored off-chip



History Buffer
A
X
B
C
D
Y
E
…
A
X
B
C

History Buffer

PC-localization with Global History Buffer

Global Access Stream :

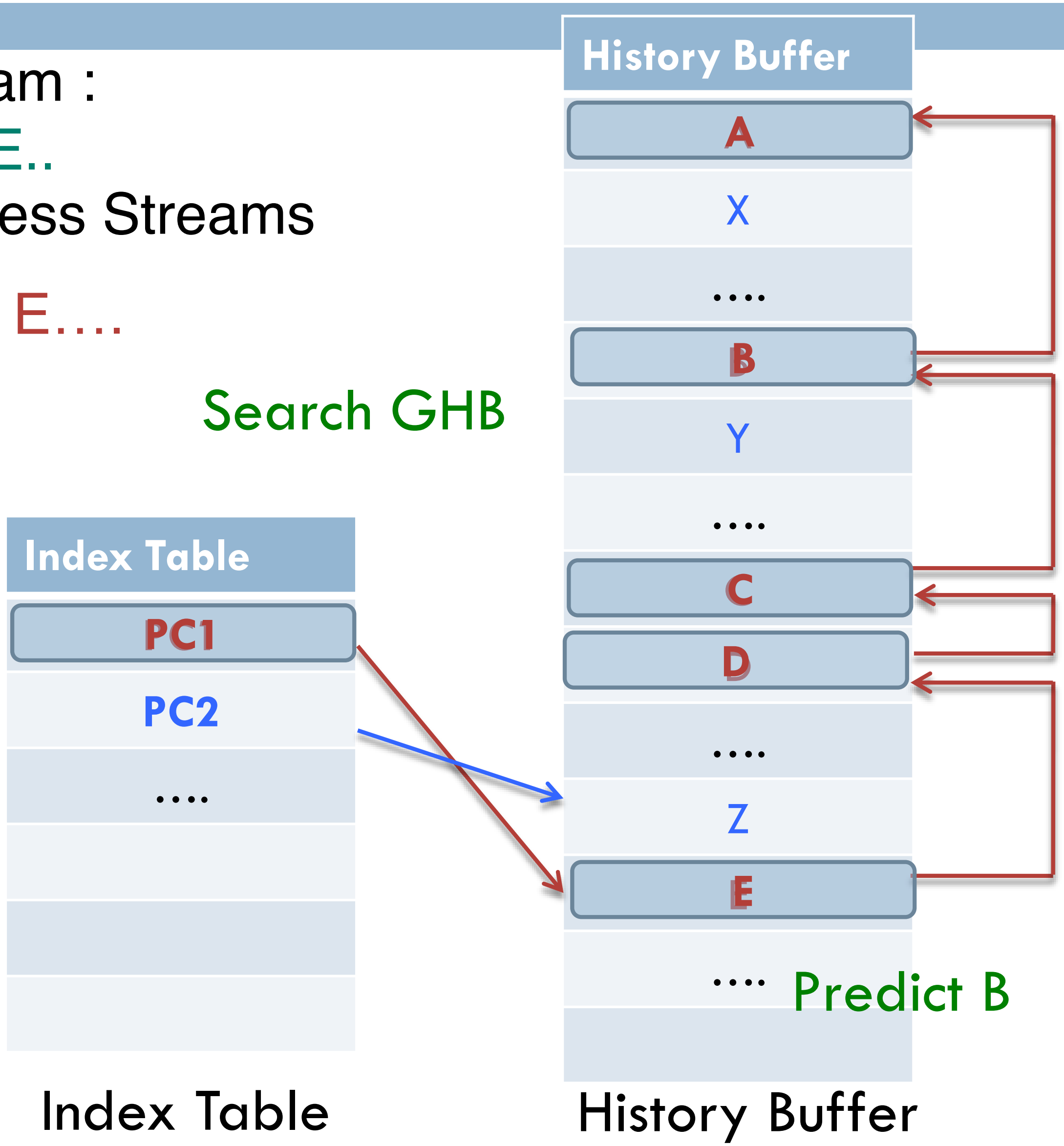
A, X, B, Y, C, D, Z, E..

PC-localized access Streams

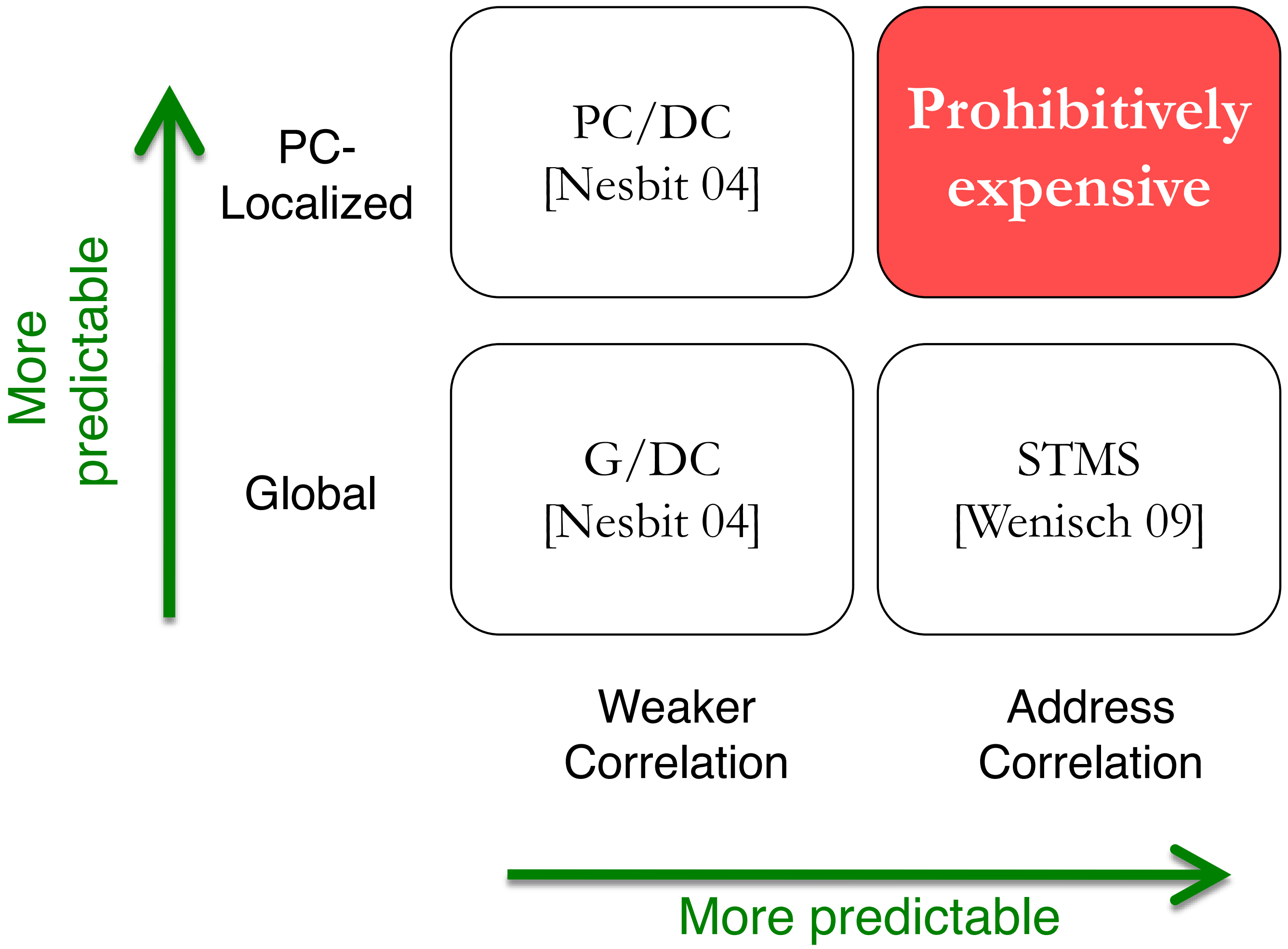
PC1 : A, B, C, D, E….

PC2 : X, Y, Z

Input : PC1, A

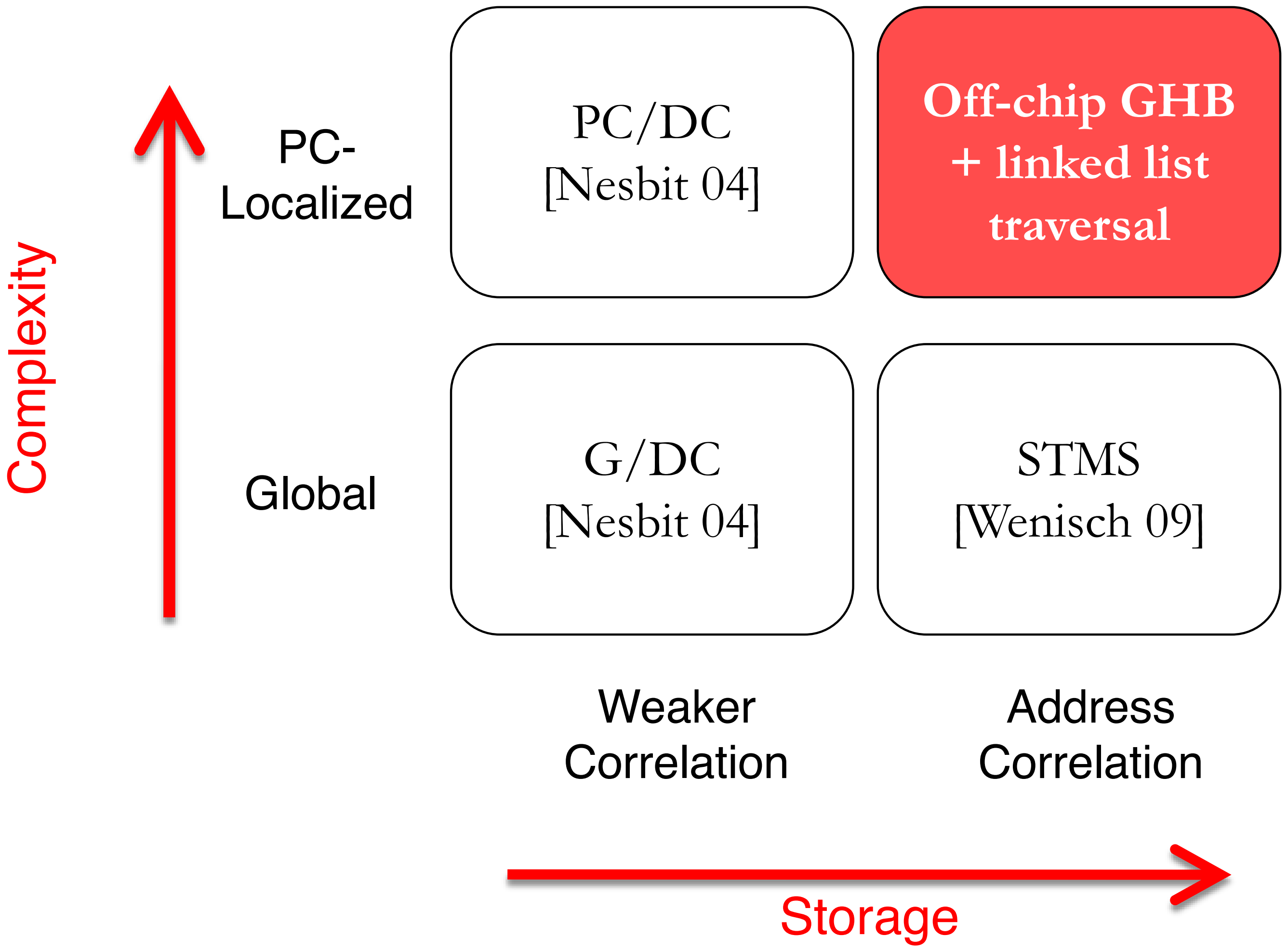


Limitation of GHB-based Solutions



	Weaker Correlation	Address Correlation
PC-Localized	PC/DC [Nesbit 04]	Prohibitively expensive
Global	G/DC [Nesbit 04]	STMS [Wenisch 09]

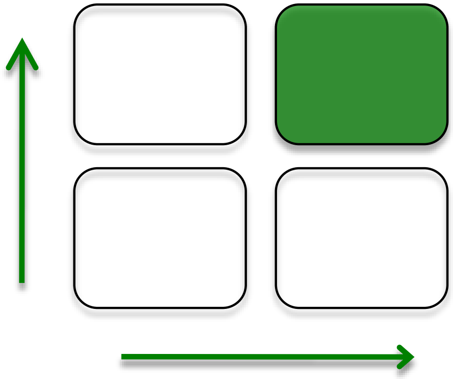
Limitation of GHB-based Solutions



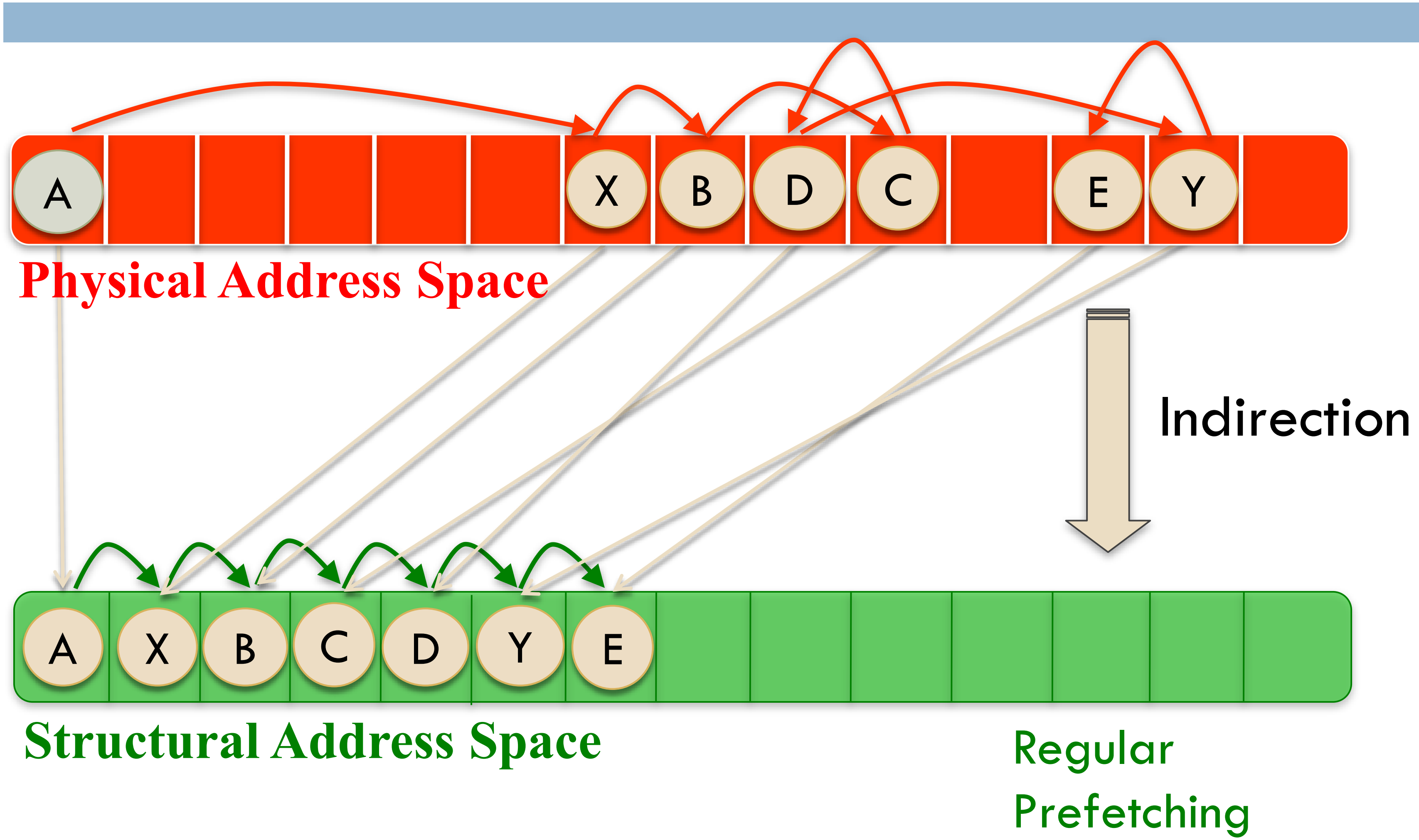
	Weaker Correlation	Address Correlation
PC-Localized	PC/DC [Nesbit 04]	Off-chip GHB + linked list traversal
Global	G/DC [Nesbit 04]	STMS [Wenisch 09]

Outline

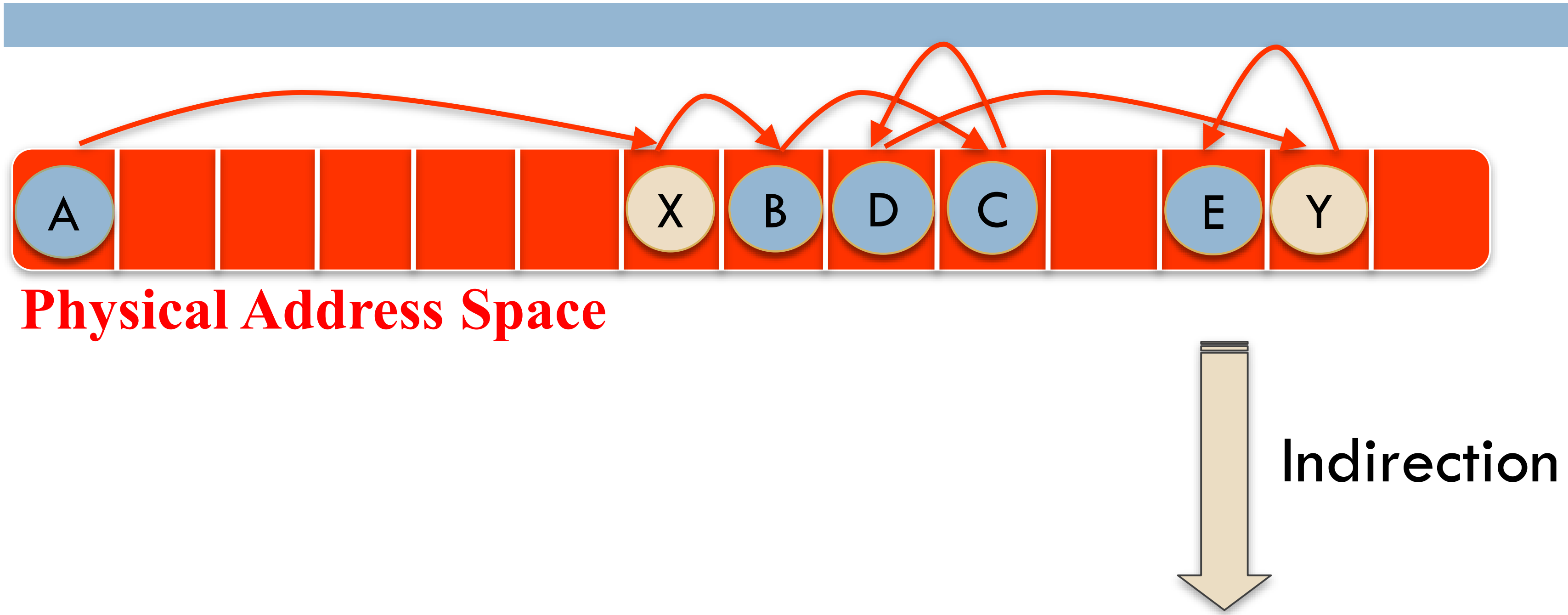
- Motivation
  - GHB-based solutions are limited
- Our Solution
  - Replace the GHB with a better organization
- Evaluation
- Future Work
- Conclusions



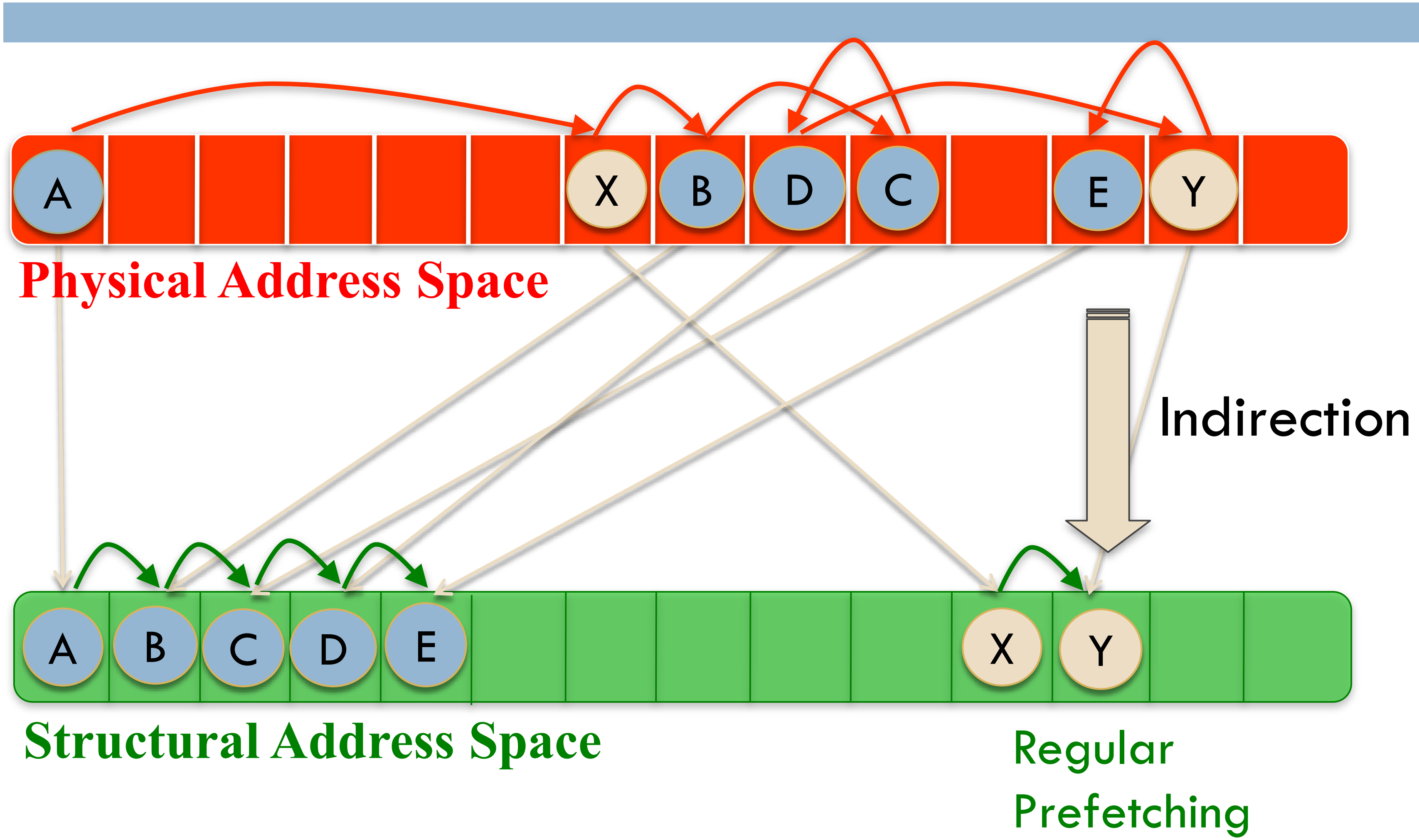
Our Solution



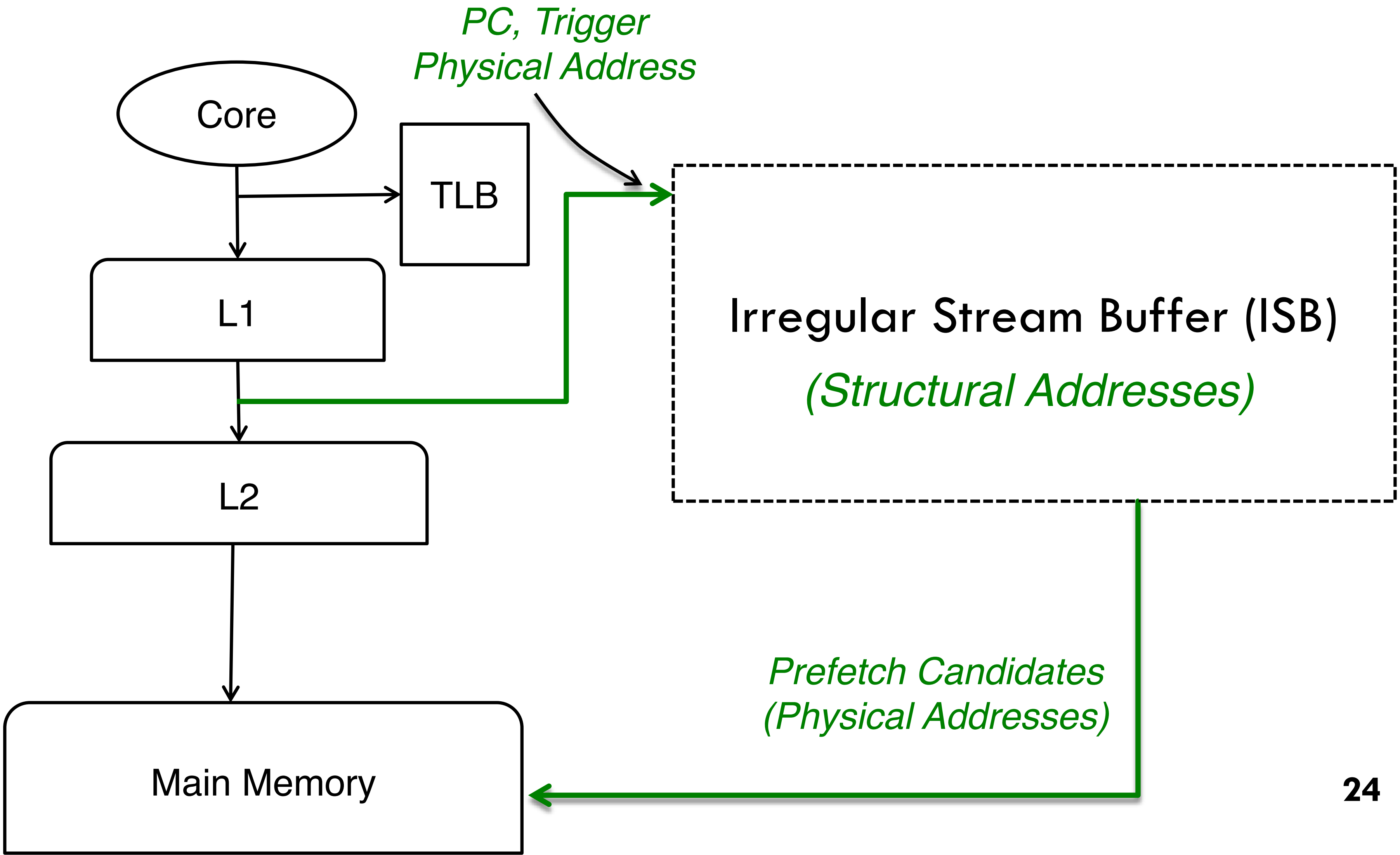
PC-Localization



PC-Localization



Irregular Stream Buffer (ISB)



Creating Structural Addresses

PC-Localized Stream : A

Physical Address Space

Physical to Structural Mapping

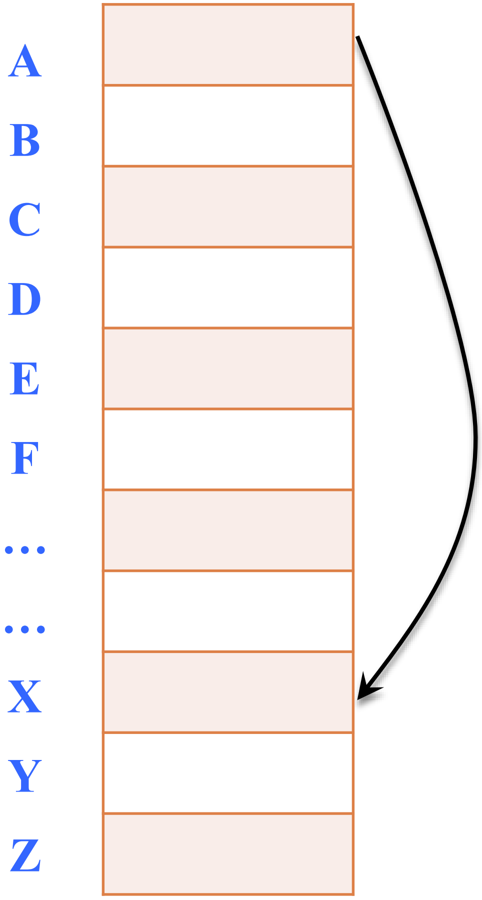
Physical Address	Structural Address
A	1
B	
C	
D	
E	
F	
…	
X	
Y	
Z	

Creating Structural Addresses

PC-Localized Stream : A X

Correlated pair :
(A, X)

Physical
Address Space



Physical to Structural Mapping

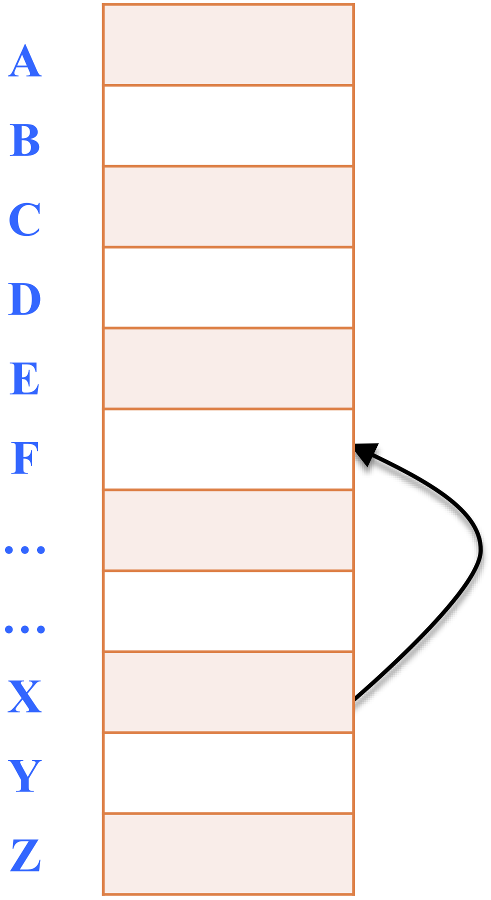
Physical Address	Structural Address
A	1
B	
C	
D	
E	
F	
…	
X	2
Y	
Z	

Creating Structural Addresses

PC-Localized Stream : A X F

Correlated pair : (X, F)

Physical Address Space



Physical to Structural Mapping

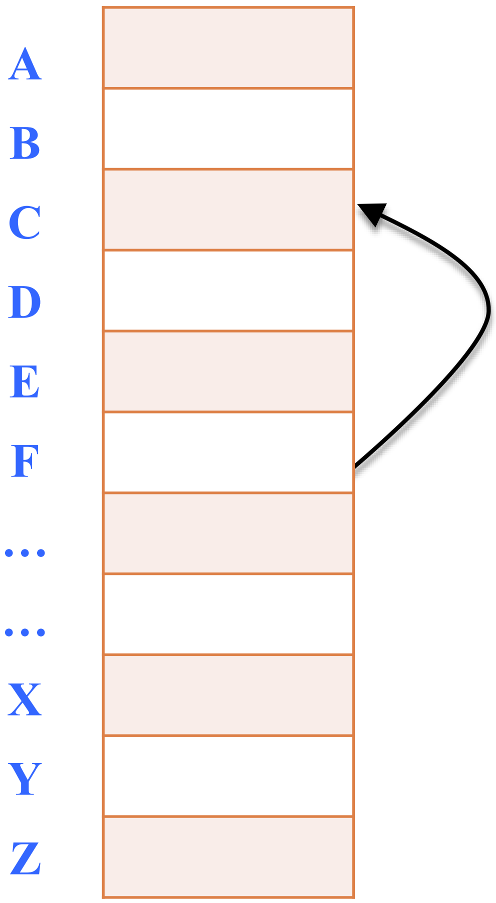
Physical Address	Structural Address
A	1
B	
C	
D	
E	
F	3
…	
X	2
Y	
Z	

Creating Structural Addresses

PC-Localized Stream : A X F C

Correlated pair : (F, C)

Physical Address Space



Physical to Structural Mapping

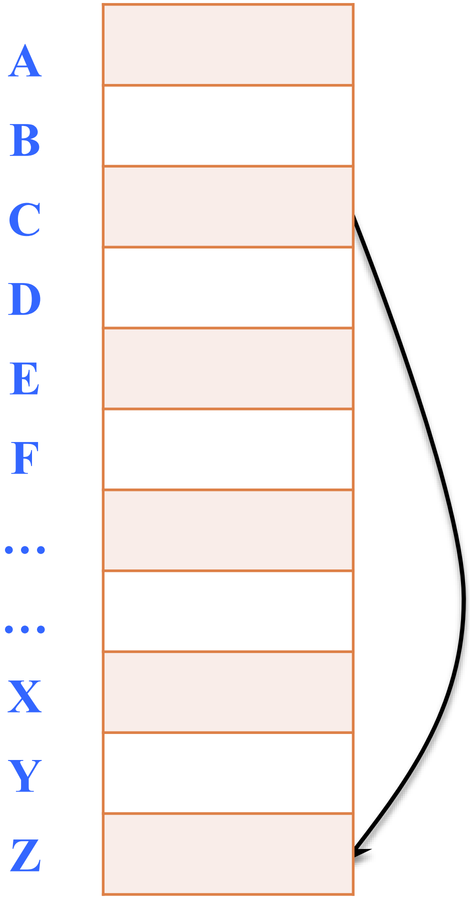
Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
…	
X	2
Y	
Z	

Creating Structural Addresses

PC-Localized Stream : A X F C Z

Correlated pair : (C, Z)

Physical Address Space



Physical to Structural Mapping

Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
…	
X	2
Y	
Z	5

Prediction in the Structural Address Space

Trigger Address: X

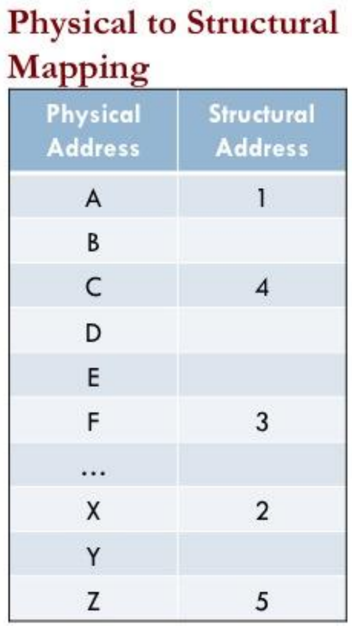
Physical to Structural Mapping

Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
…	
X	2
Y	
Z	5

Structural to Physical Mapping

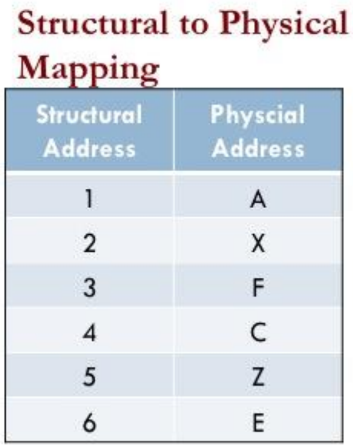
Structural Address	Physcial Address
1	A
2	X
3	F
4	C
5	Z
6	E

Irregular Stream Buffer (ISB)



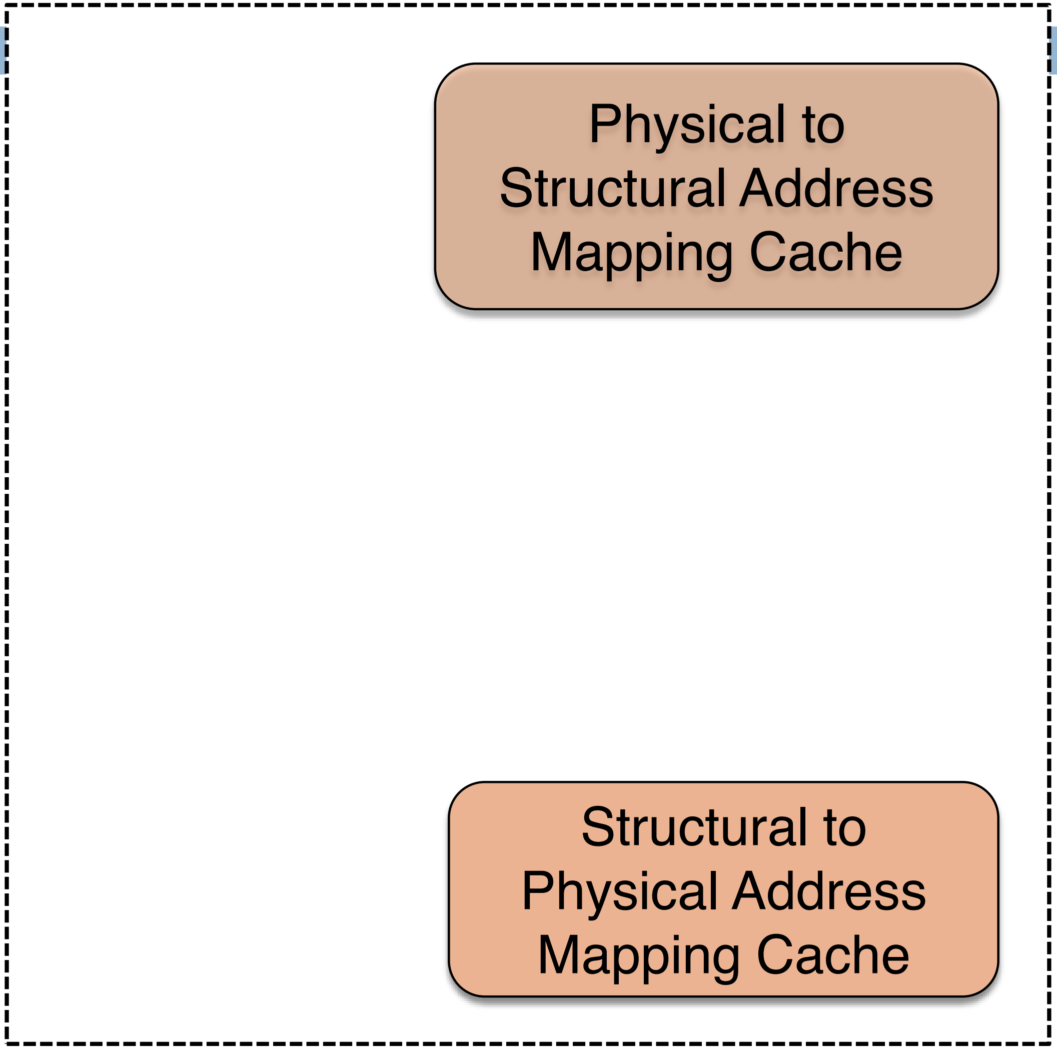
Physical to Structural Mapping

Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
...	
X	2
Y	
Z	5



Structural to Physical Mapping

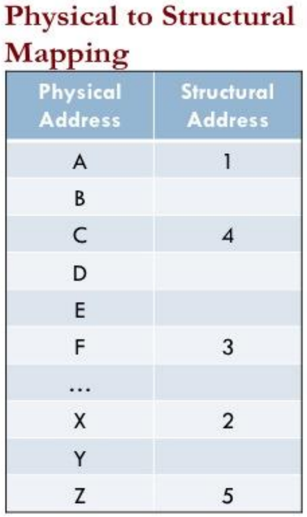
Structural Address	Physcial Address
1	A
2	X
3	F
4	C
5	Z
6	E



Physical to Structural Address Mapping Cache

Structural to Physical Address Mapping Cache

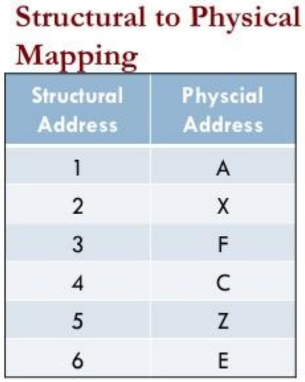
Irregular Stream Buffer (ISB)



Physical to Structural Mapping

Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
…	
X	2
Y	
Z	5

Physical to Structural Address Mapping Cache

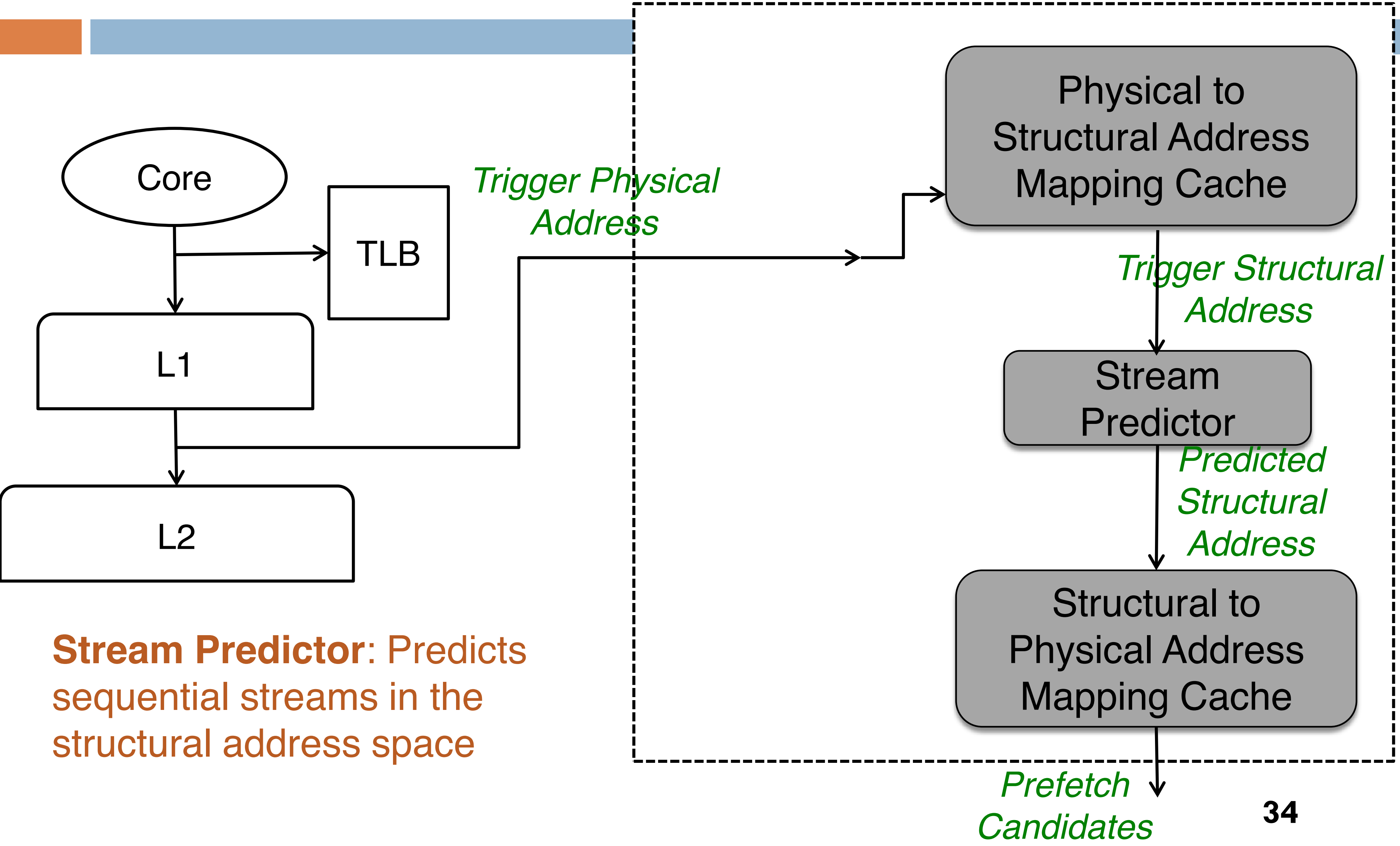


Structural to Physical Mapping

Structural Address	Physcial Address
1	A
2	X
3	F
4	C
5	Z
6	E

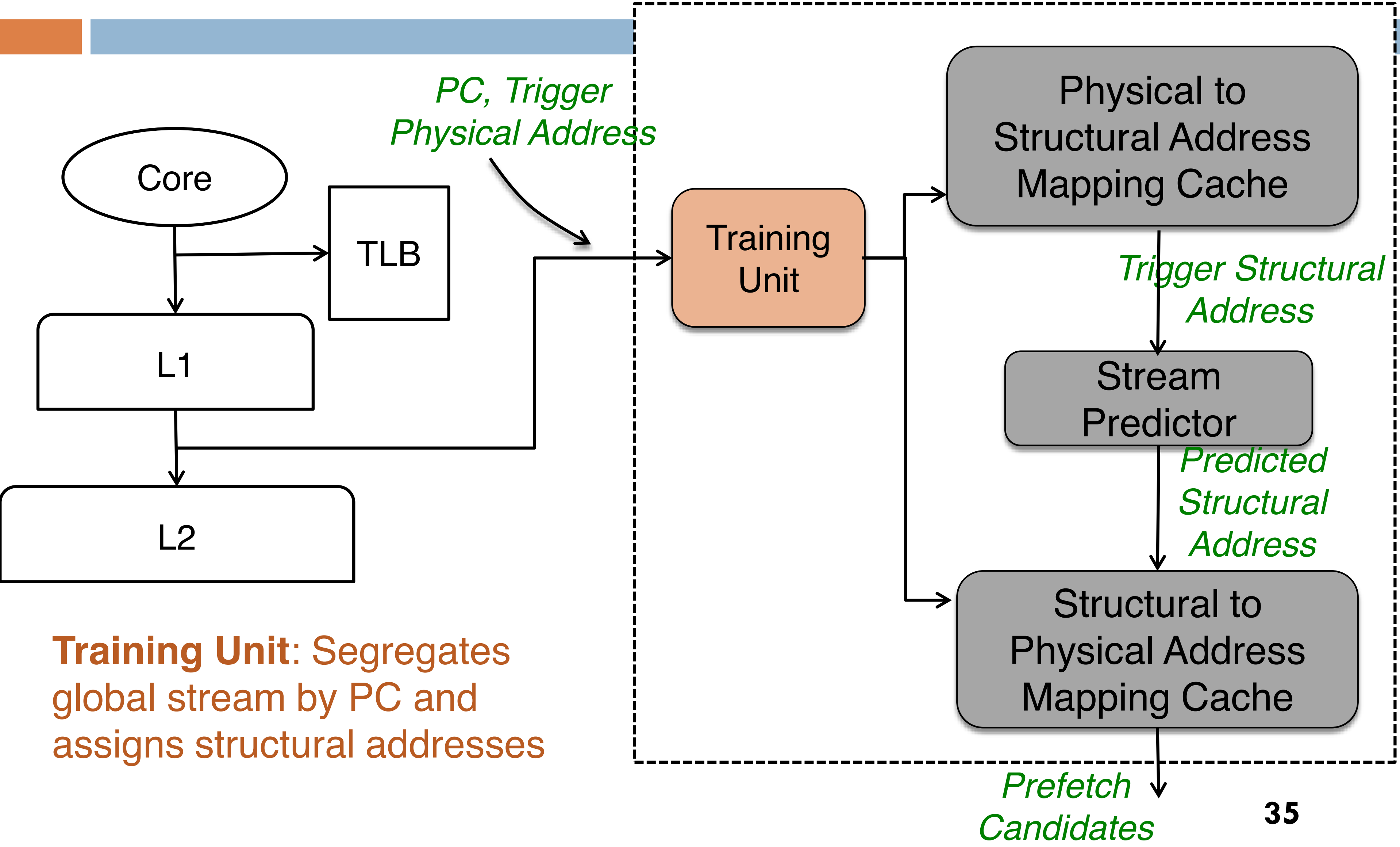
Structural to Physical Address Mapping Cache

Irregular Stream Buffer (ISB)


Core
TLB
L1
L2
Trigger Physical Address
Physical to Structural Address Mapping Cache
Trigger Structural Address
Stream Predictor
Predicted Structural Address
Structural to Physical Address Mapping Cache
Prefetch Candidates

Stream Predictor: Predicts sequential streams in the structural address space

Irregular Stream Buffer (ISB)



Training Unit: Segregates global stream by PC and assigns structural addresses

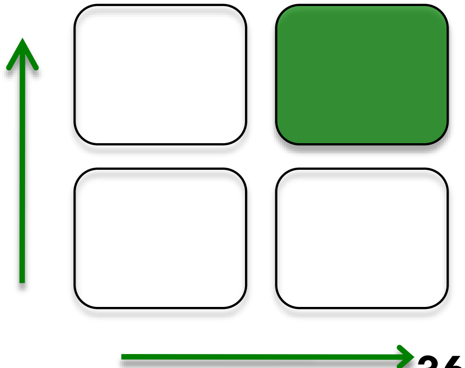
Revisiting Our Contributions

- Significantly advance the state-of-the-art in temporal prefetching

- How do we achieve this improvement?
  - First to combine address correlation and PC-localization
  - Two more benefits!

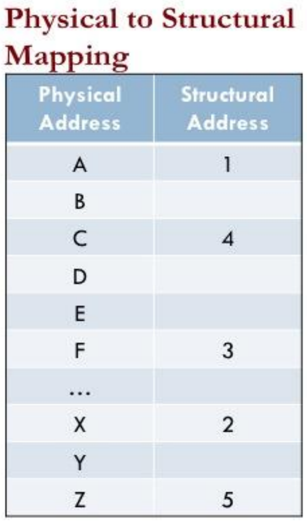
High coverage

High accuracy



Meta-data Management

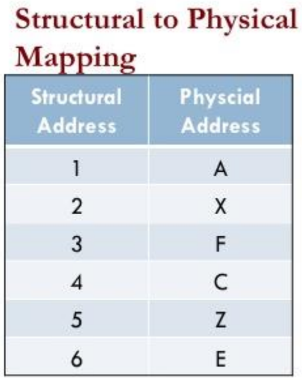
But the table is too big to be stored on-chip!



Physical to Structural Mapping

Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
…	
X	2
Y	
Z	5

Physical to Structural Address Mapping Cache



Structural to Physical Mapping

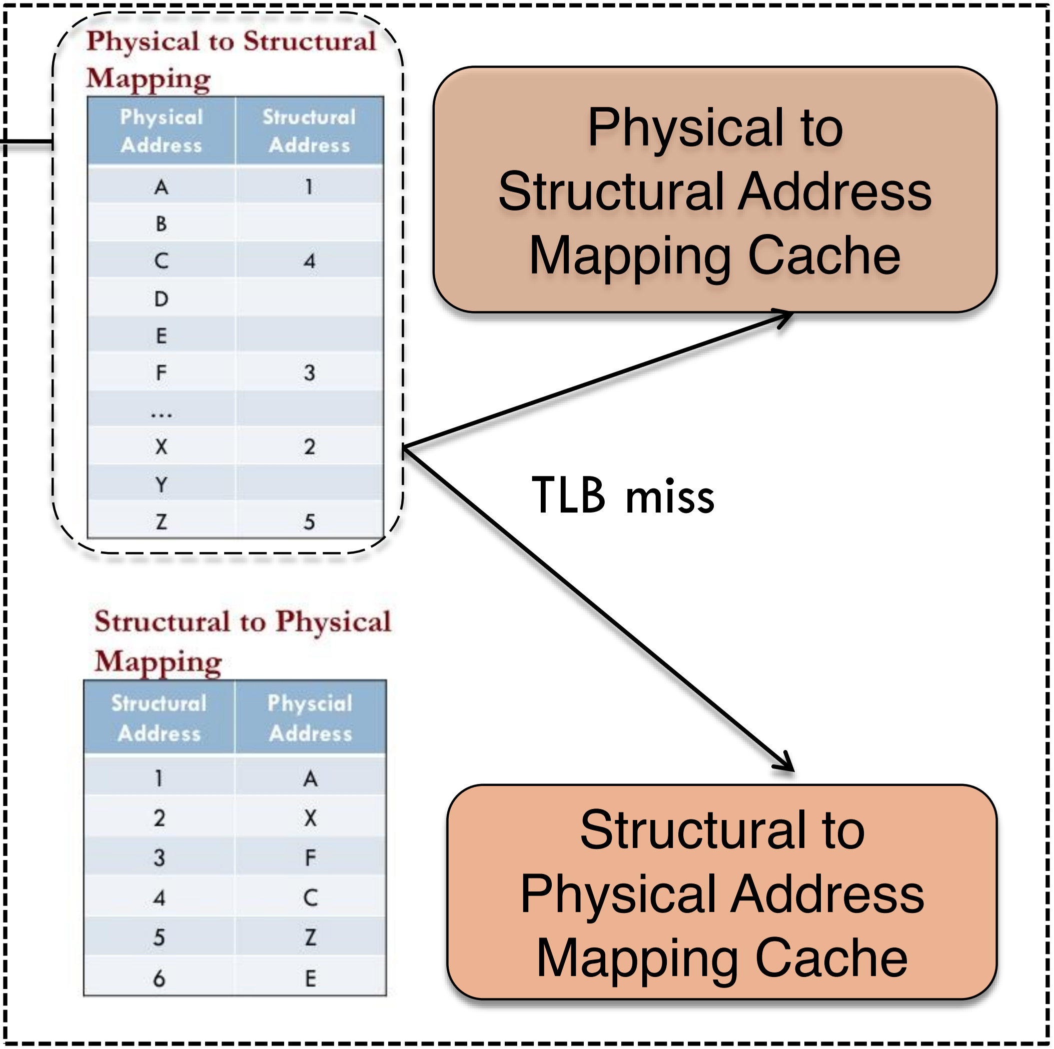
Structural Address	Physcial Address
1	A
2	X
3	F
4	C
5	Z
6	E

Structural to Physical Address Mapping Cache

New Caching Scheme

We already know how to cache the virtual to physical mappings!

Off-chip



Physical to Structural Mapping

Physical Address	Structural Address
A	1
B	
C	4
D	
E	
F	3
…	
X	2
Y	
Z	5

Structural to Physical Mapping

Structural Address	Physcial Address
1	A
2	X
3	F
4	C
5	Z
6	E

Physical to Structural Address Mapping Cache
TLB miss
Structural to Physical Address Mapping Cache

New Caching Scheme

- Mapping cached on-chip only for TLB resident data
  - Small on-chip budget
- Movement of off-chip meta-data synchronized with TLB misses
  - Hides latency of accessing off-chip meta-data hidden
  - Reduces memory traffic
- This caching scheme is not possible with the GHB

Revisiting Our Contributions

- Significantly advance the state-of-the-art in temporal prefetching

- How do we achieve this improvement?
  - First to combine address correlation and PC-localization
  - Enables a novel TLB-synchronized caching scheme
  -

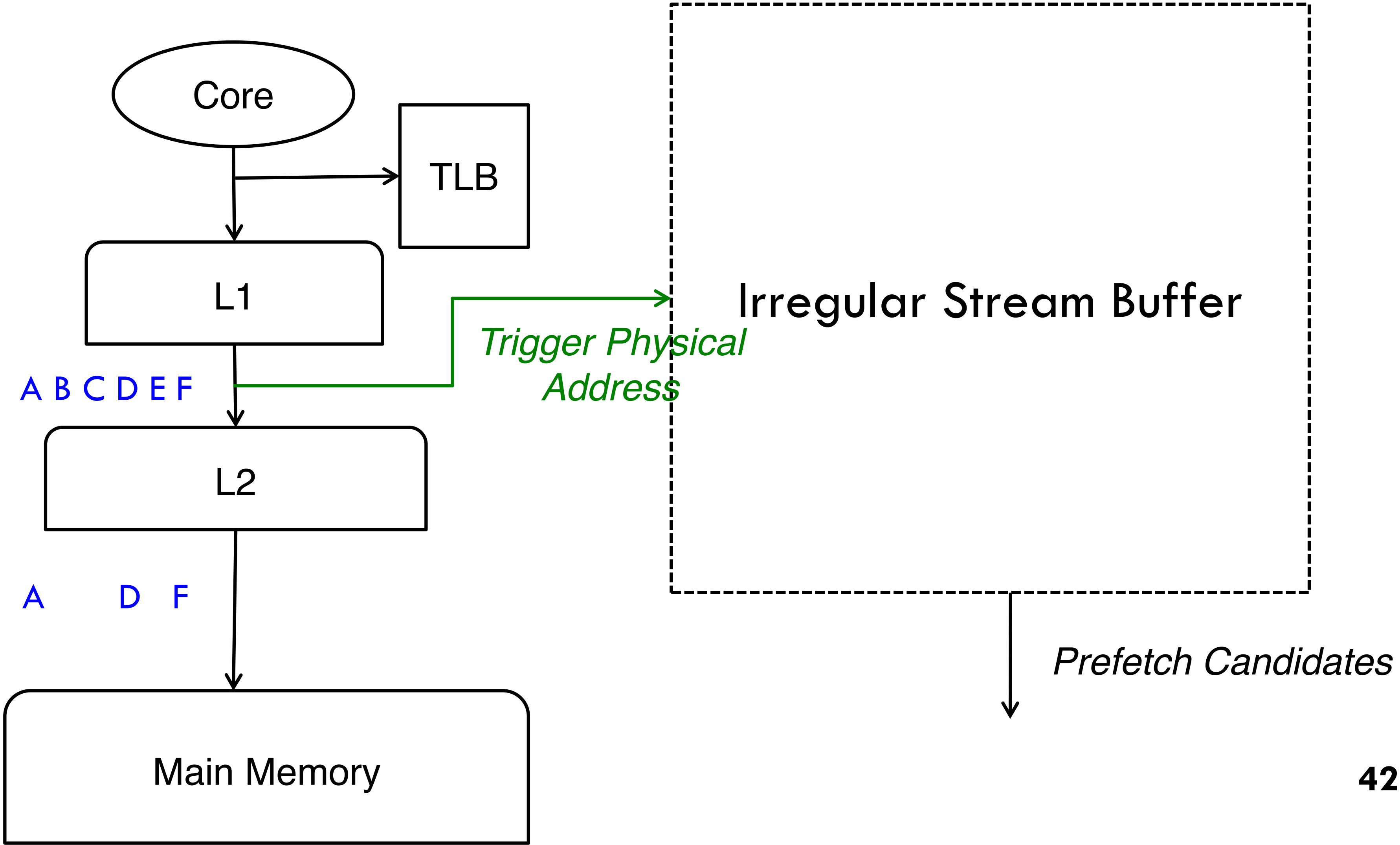
High coverage

High accuracy

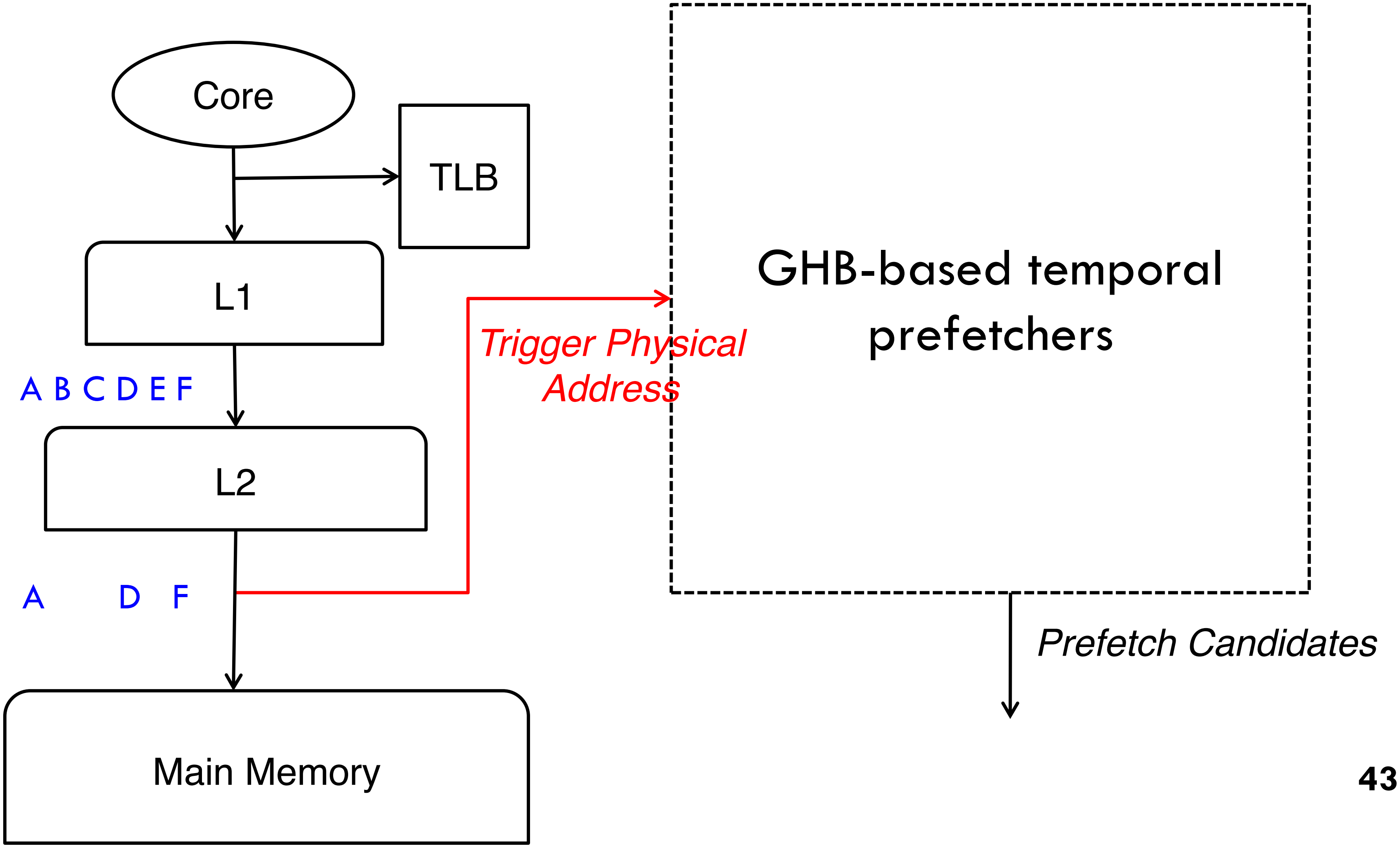
Low traffic overhead

Low on-chip storage

Training on the L2 Access Stream



Train on the L2 Access Stream



Revisiting Our Contributions

- Significantly advance the state-of-the-art in temporal prefetching

- How do we achieve this improvement?
  - First to combine address correlation and PC-localization
  - Enables a novel TLB-synchronized caching scheme
  - Allows the ISB to train on the L2 access stream

High coverage

High accuracy

Low traffic overhead

Low on-chip storage

Outline

- Motivation
  - GHB-based solutions are limited
- Our Solution
  - Replace the GHB with a better organization
- Evaluation
- Future Work
- Conclusions

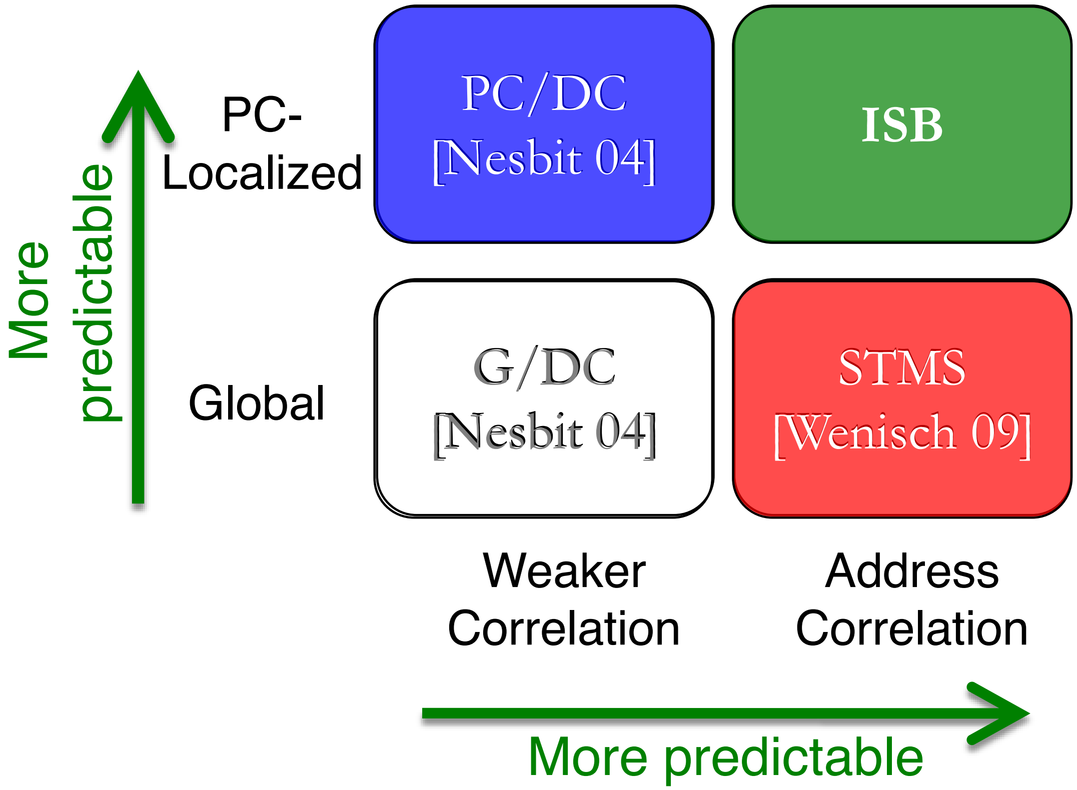
Methodology

- MARSSx86 simulator
  - Cycle-accurate out-of-order code
  - L1 Data Cache – 64 KB
  - L2 Data Cache – 2 MB
  - DTLB – 128 entries (TLB miss latency modeled realistically)
  - Main memory 140 cycles (DRAM queue contention accurately modeled)
- Benchmarks – Irregular subset of Spec2006
  - 20-25 SimPoints per benchmarks
  - Each SimPoint simulates 250M instruction

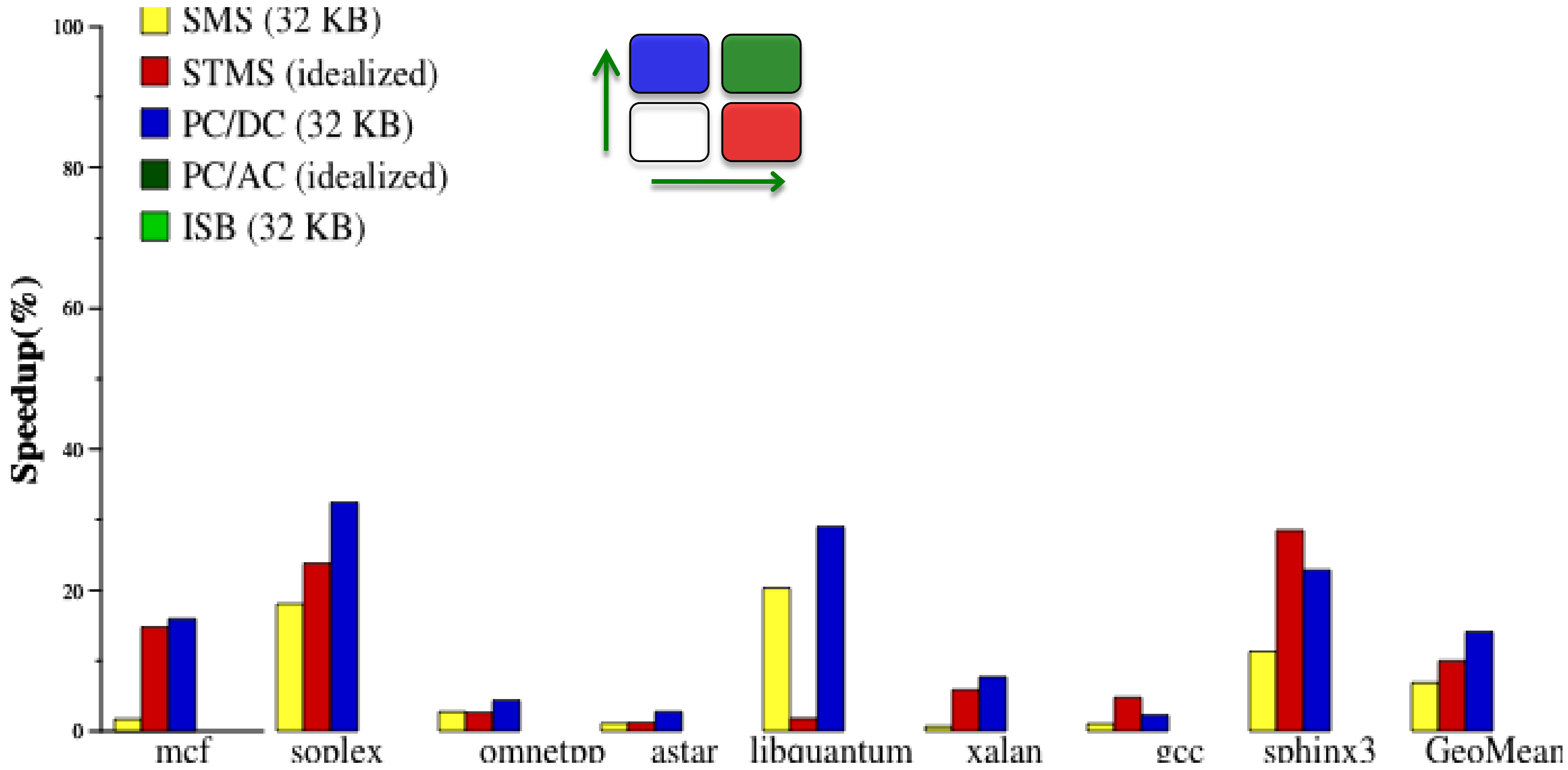
Comparison with Irregular Prefetchers

- 🟨 SMS (32 KB)
- 🟥 STMS
- 🟦 PC/DC
- ■ PC/AC
- 🟩 ISB

Limit Study : A completely impractical GHB-based combination of PC-localization and address correlation

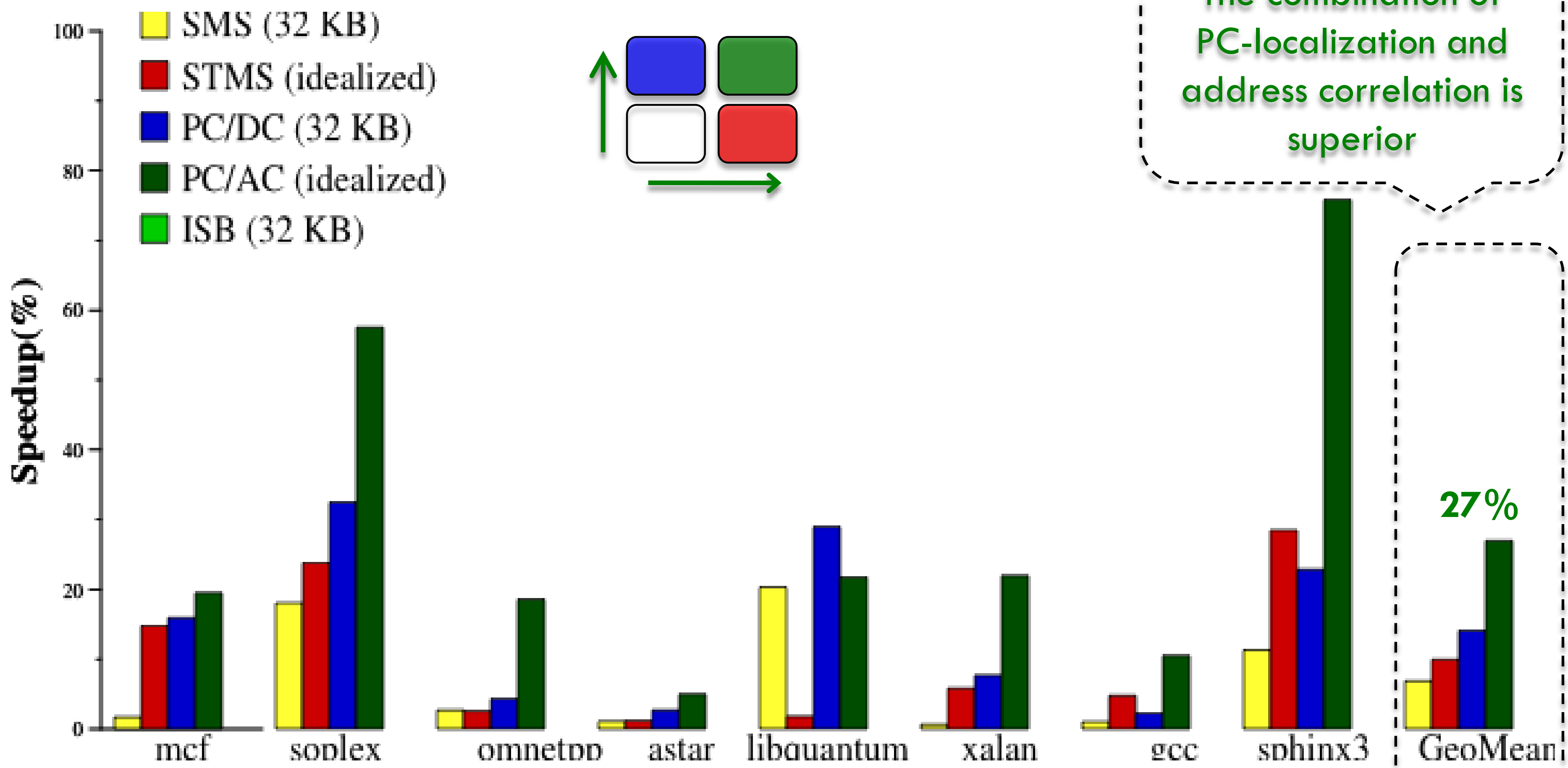


Comparison with Irregular Prefetchers



Speedup due to irregular prefetchers over a baseline with no prefetching

Comparison with Irregular Prefetchers

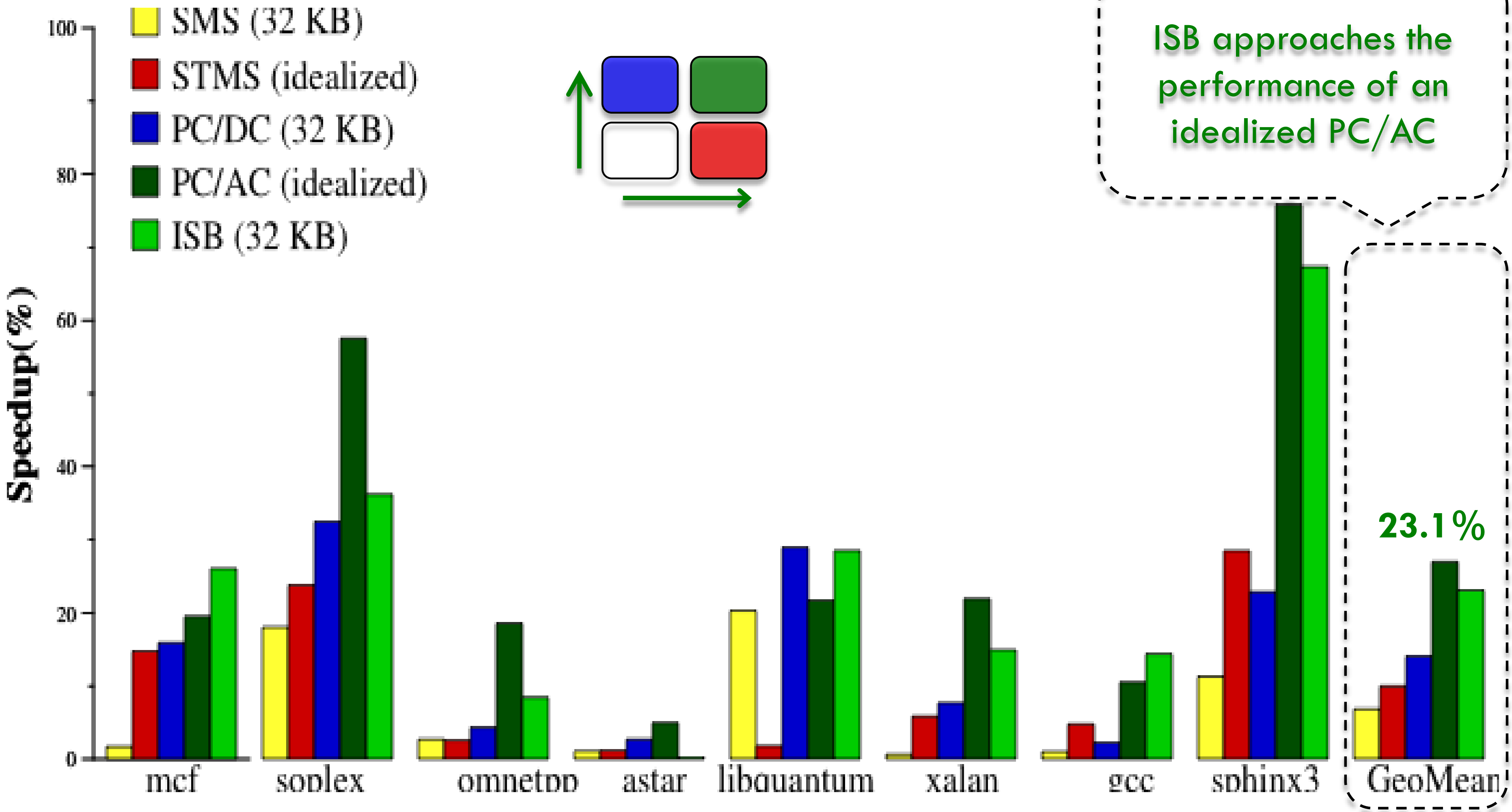


The combination of PC-localization and address correlation is superior

27%

Speedup due to irregular prefetchers over a baseline with no prefetching

Comparison with Irregular Prefetchers



ISB approaches the performance of an idealized PC/AC

Speedup due to irregular prefetchers over a baseline with no prefetching

Recall Our Contributions

- Significantly advance the state-of-the-art in temporal prefetching

- How do we achieve this improvement?
  - First to combine address correlation and PC-localization
  - Enables a novel TLB-synchronized caching scheme
  - Train on the L2 access stream

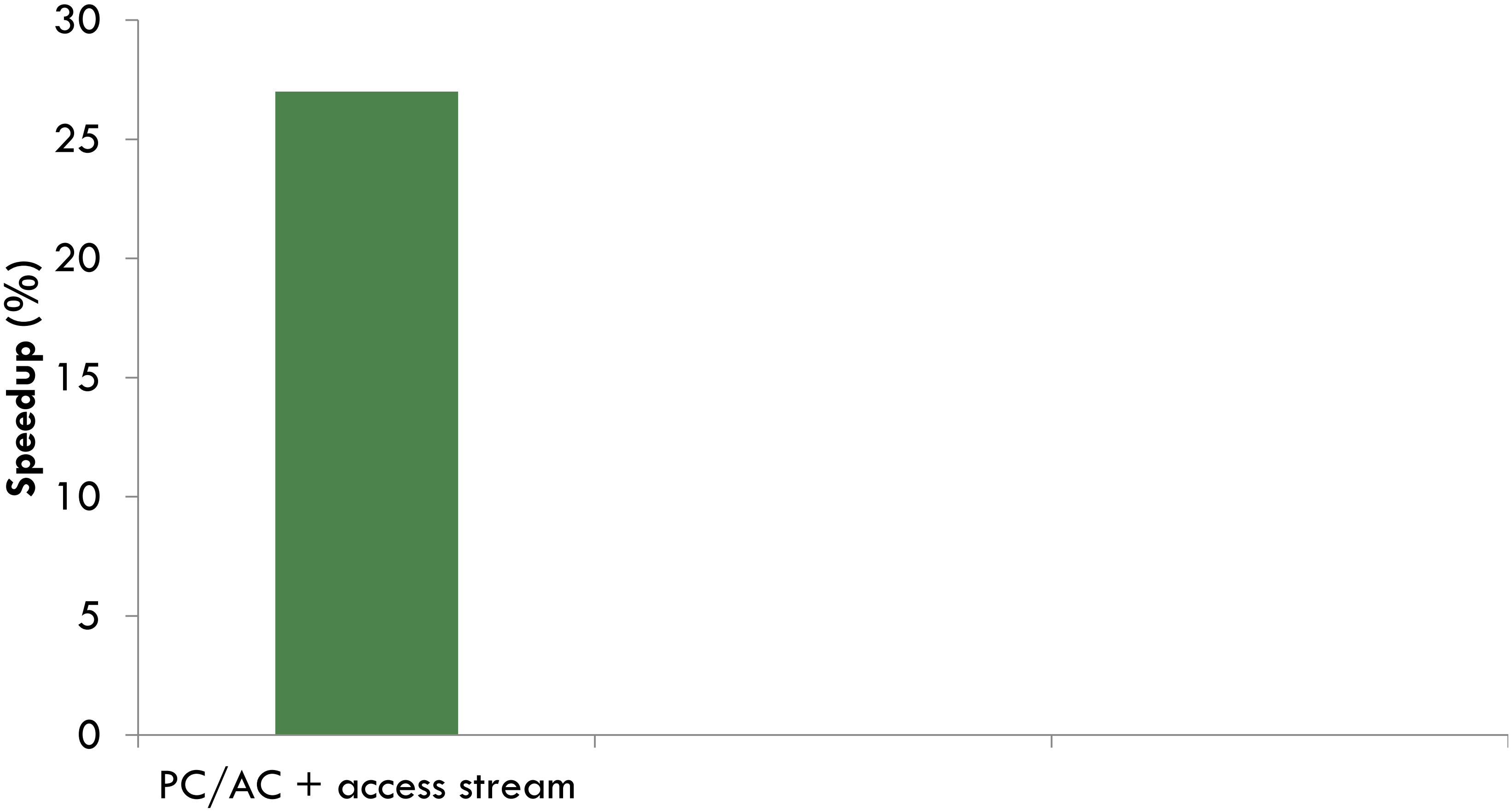
High coverage

High accuracy

Low traffic overhead

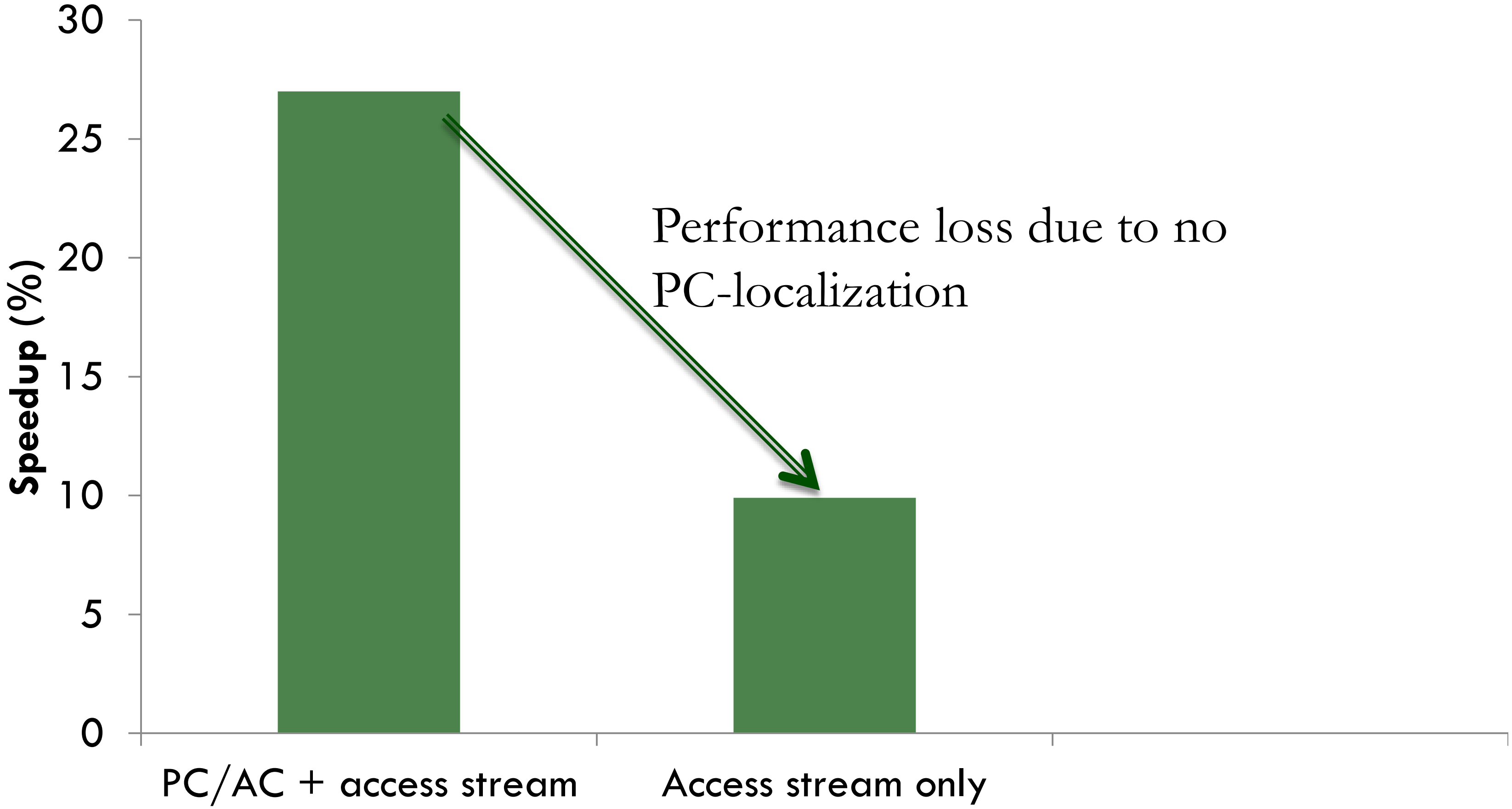
Low on-chip storage

Why does the ISB win?



PC-Localization	✓
Access Stream	✓

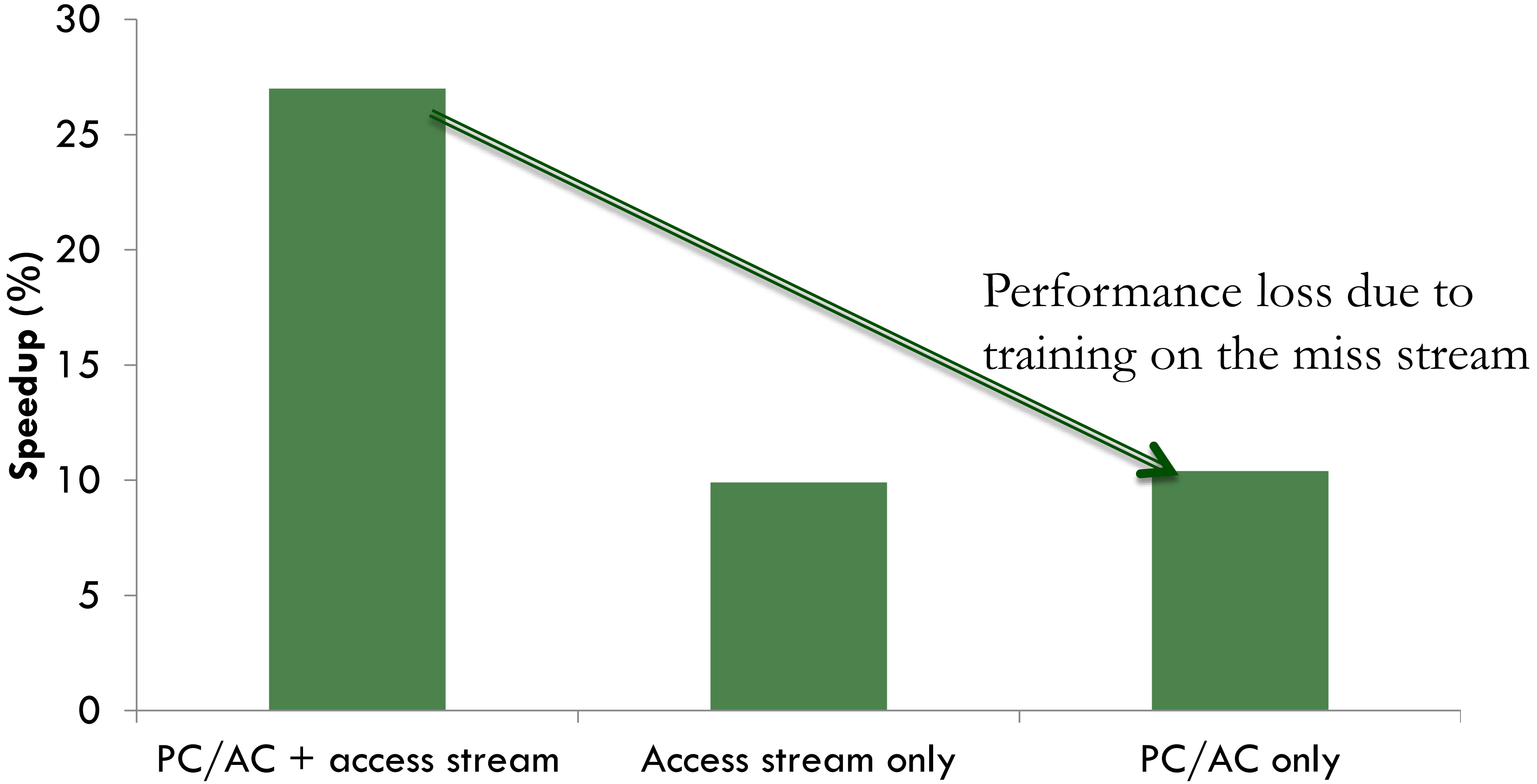
Why does the ISB win?



Performance loss due to no PC-localization

PC-Localization	✓	✗
Access Stream	✓	✓

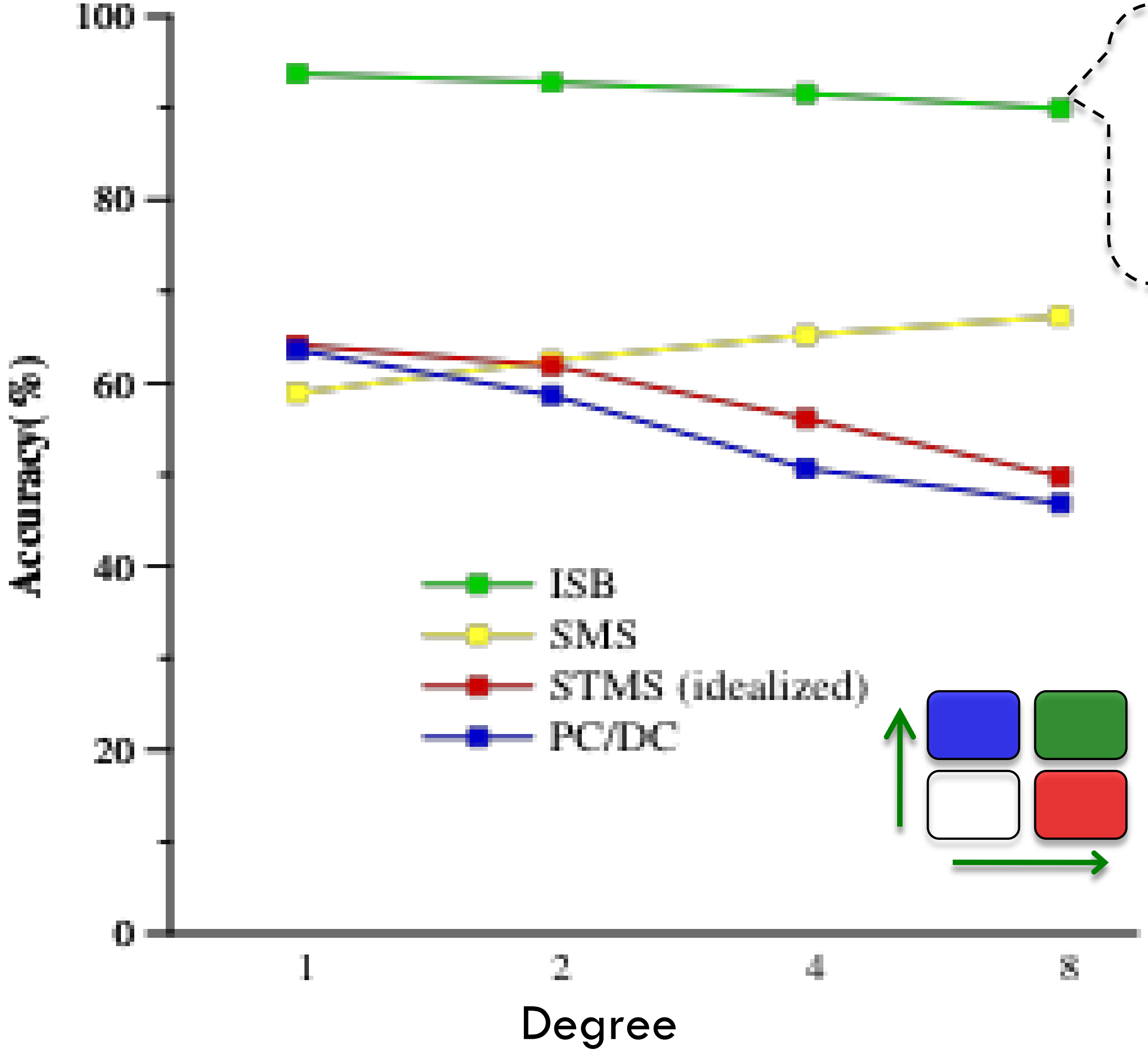
Why does the ISB win?



Performance loss due to training on the miss stream

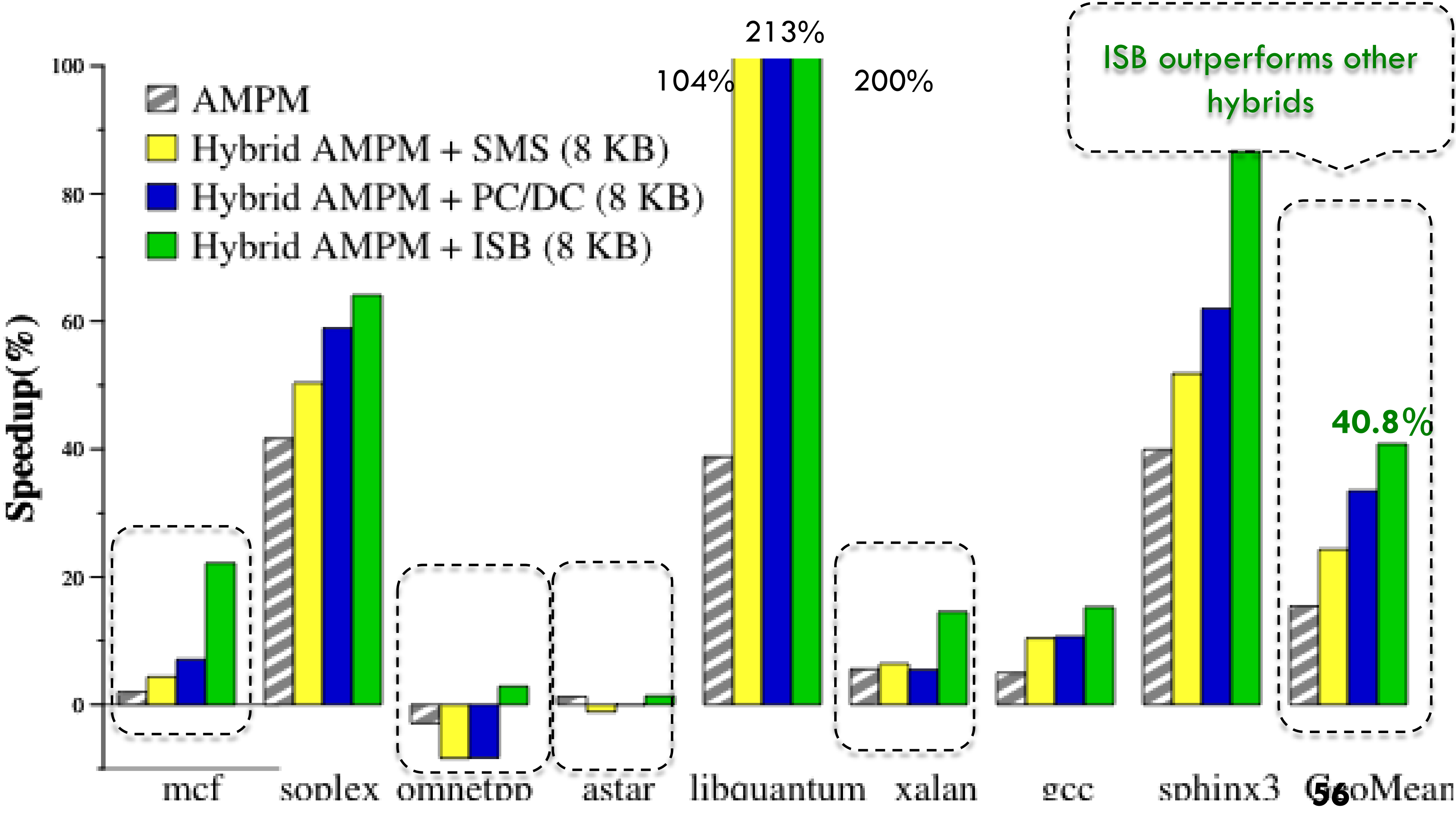
PC-Localization	✓	✗	✓
Access Stream	✓	✓	✗

Accuracy comparison



ISB is over 90% accurate even at higher degrees

Hybrid Prefetchers



ISB outperforms other hybrids

Hardware Cost

- Address Mapping Caches for TLB-resident data
  - 32 KB on-chip storage for TLB with 128 entries
  - 8 KB sufficient in a hybrid setting

- Memory traffic Overhead
  - Average meta-data traffic: 8.4%
  - Average traffic due to useless prefetch requests: 6.3%

	mcf	soplex	omnetpp	astar	libq	xalan	gcc	sphinx
Meta-data Traffic	5.7%	3.8%	5.7%	12%	1.6%	12.6%	11.3%	3.9%

Future Work

- Combining spatial and temporal prefetching
  - STeMS - Spatial-temporal Prefetching [Somogyi 2009]
  - Temporal component can be enhanced using the ISB
- Evaluation on commercial workloads
  - Bigger memory footprints don't affect the ISB's on-chip storage
  - Optimizations to improve TLB reach, such as superpages, affect ISB's on-chip storage

Conclusions

- Replace the GHB with a new organization
- Significantly improves the state-of-the-art
  - First to combine address correlation and PC-localization
  - Enables a novel TLB-synchronized caching scheme
  - Train on the L2 access stream

Thank You !

High coverage

High accuracy

Low traffic overhead

Low on-chip storage