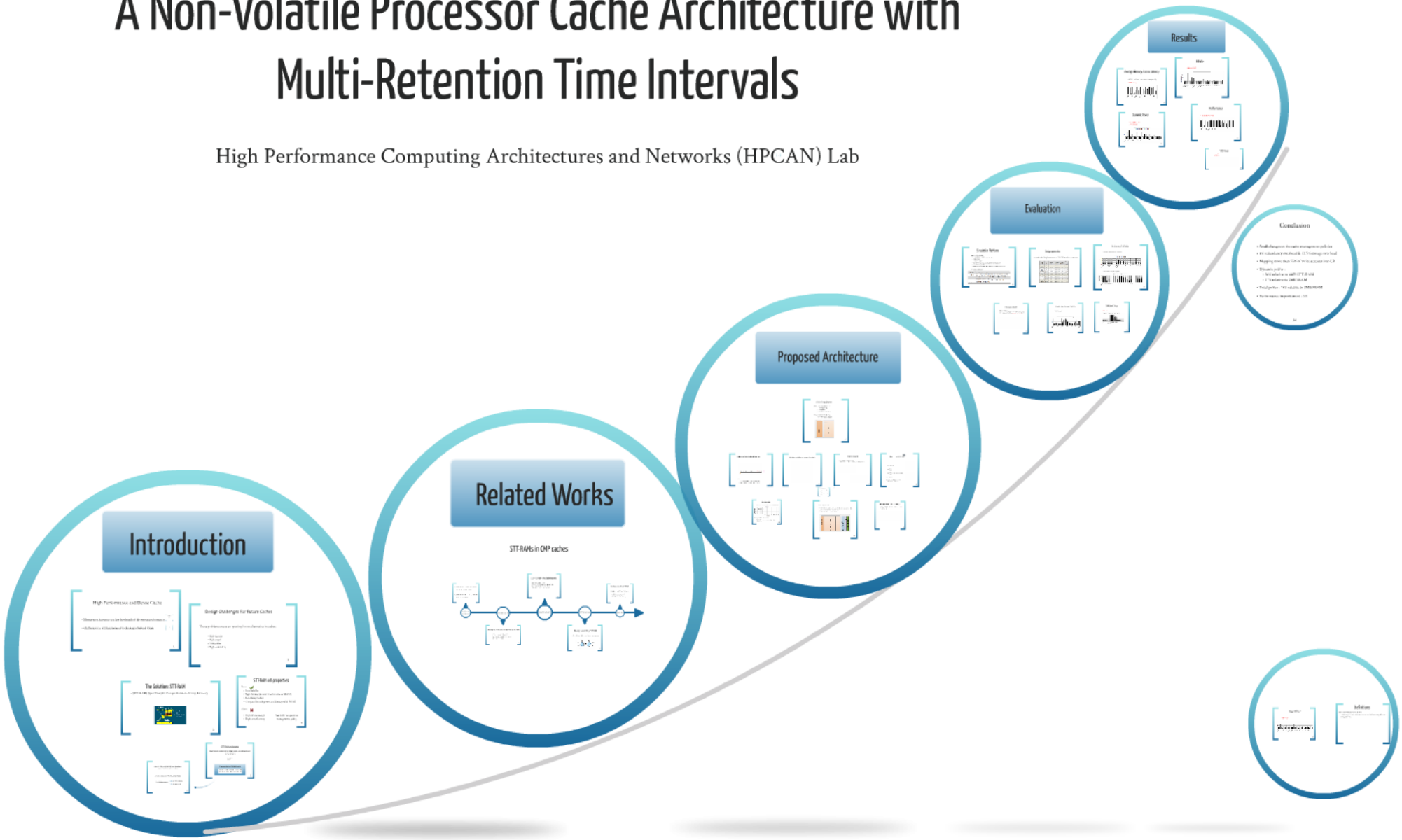
A Non-Volatile Processor Cache Architecture with Multi-Retention Time Intervals

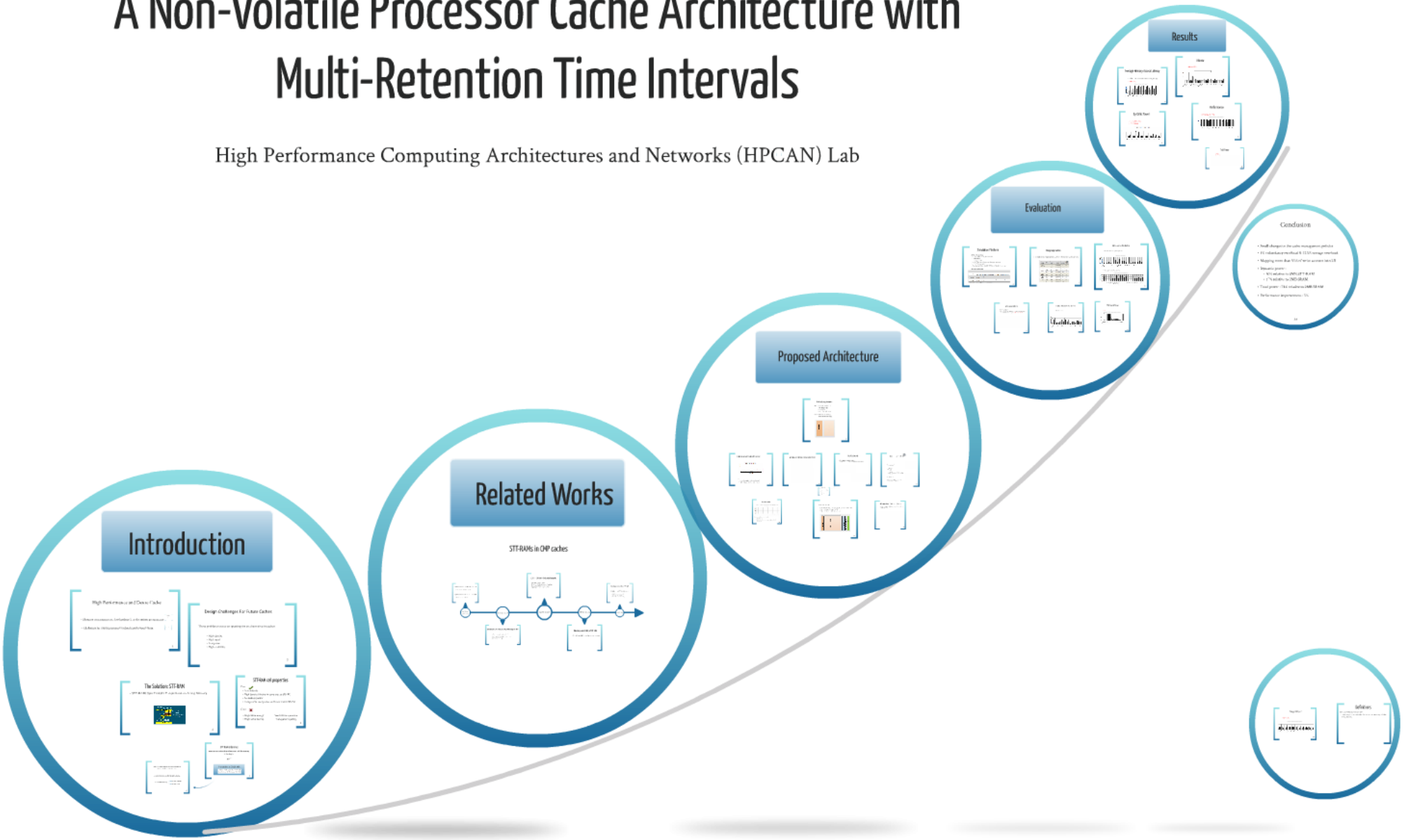
High Performance Computing Architectures and Networks (HPCAN) Lab



Prezi

A Non-Volatile Processor Cache Architecture with Multi-Retention Time Intervals

High Performance Computing Architectures and Networks (HPCAN) Lab



Prezi

Introduction

High Performance and Dense Cache

- Memory performance as a key bottleneck of the system performance
- Challenges for existing memory technologies beyond 45nm

1

Design Challenges For Future Caches

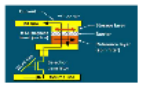
These problems create an opening for an alternative in caches

- High-density
- High speed
- Low power
- High- scalability

2

The Solution: STT-RAM

- STT-RAM: Spin Transfer Torque Random Access Memory



3

STT-RAM cell properties

Pros ✔

- Non-volatile
- High density (4x size in same area as SRAM)
- No leakage power
- Comparable read power and latency with SRAM

Cons ✖

- High write energy
- High write latency

Needs write operation management policy

4

STT-RAM endurance

much more better than other non-volatile memory technologies

BUT

Not perfect as SRAM cells

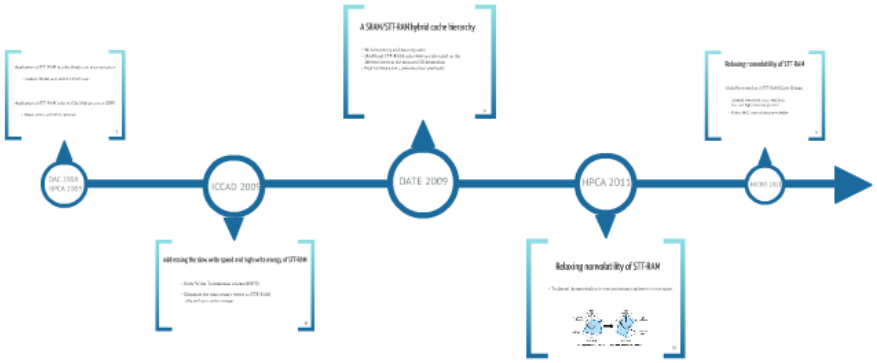
5

6

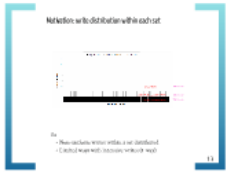
Prezi

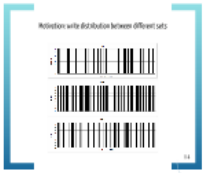
Related Works

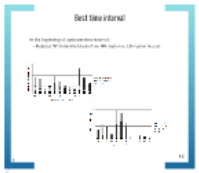
STT-RAMs in CMP caches

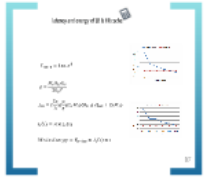


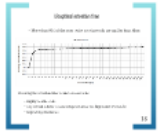
Proposed Architecture

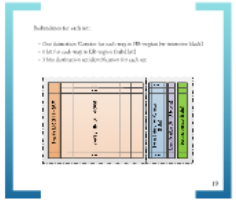


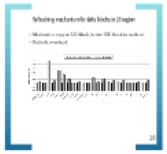




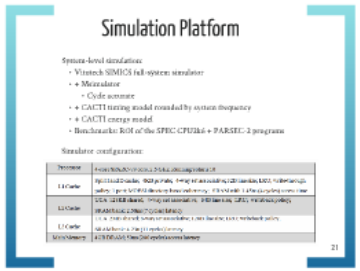


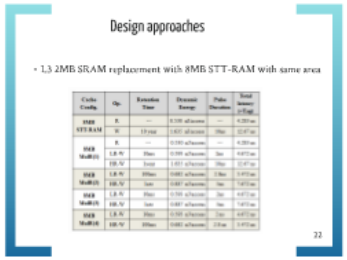


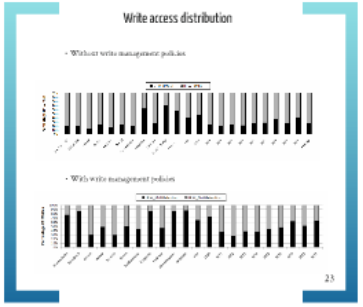


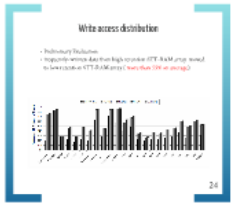


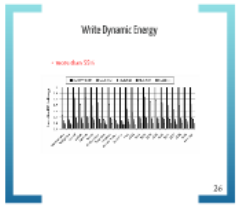
Evaluation









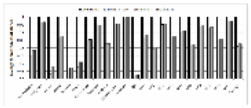


Prezi

Results

Average Memory Access Latency

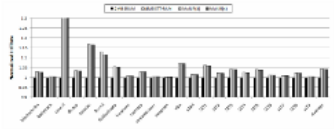
- AML = hit time + miss rate x miss penalty
- AML (20%)



28

Hitrate

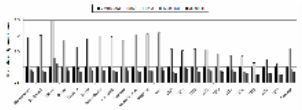
- hitrate (5%)



27

Dynamic Power

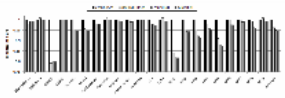
- 50% (STT-RAM)
- 15% (SRAM)



30

Performance

- Performance (5%)



29

Total Power

32

Prezi

Introduction

High Performance and Dense Cache

- Memory performance as a key bottleneck of the system performance
- Challenges for existing memory technologies beyond 45nm

1

Design Challenges For Future Caches

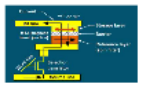
These problems create an opening for an alternative in caches

- High-density
- High speed
- Low power
- High- scalability

2

The Solution: STT-RAM

- STT-RAM: Spin Transfer Torque Random Access Memory



3

STT-RAM cell properties

Pros ✔

- Non-volatile
- High density (4x size in same area as SRAM)
- No leakage power
- Comparable read power and latency with SRAM

Cons ✖

- High write energy
- High write latency

Needs write operation management policy

4

STT-RAM endurance

much more better than other non-volatile memory technologies

BUT

Not perfect as SRAM cells

5

6

Prezi

Introduction

Prezi

High Performance and Dense Cache

- Memory performance as a key bottleneck of the system performance

- Challenges for existing memory technologies beyond 45nm

Prezi

Performance Bottleneck

- Ever increasing number of on-chip cores
- Deep pipelining in microprocessor architectures
- Ever-increasing trend of memory footprint of the programs (980MB for SPEC CPU2k6 vs. up to 200MB for SPEC CPU2k)
- Critical applications are becoming more data-centric, less compute-centric

Prezi

High Performance and Dense Cache

- Memory performance as a key bottleneck of the system performance

- Challenges for existing memory technologies beyond 45nm

Prezi

Challenges beyond 45 nm

SRAM:

- High power consumption
- Leakage increasing 10X with each technology node
- Low scalability

Prezi

High Performance and Dense Cache

- Memory performance as a key bottleneck of the system performance

- Challenges for existing memory technologies beyond 45nm

Prezi

Design Challenges For Future Caches

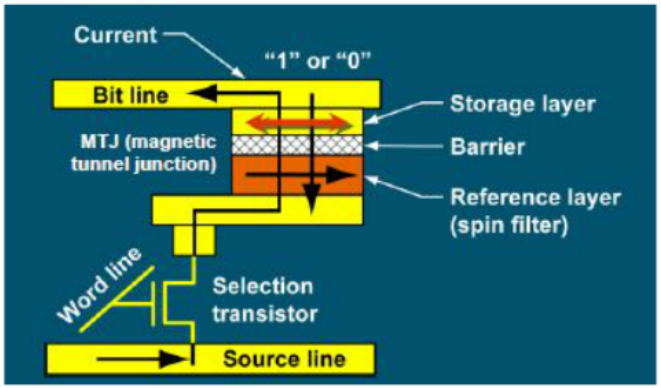
These problems create an opening for an alternative in caches

- High-density
- High speed
- Low power
- High- scalability

Prezi

The Solution: STT-RAM

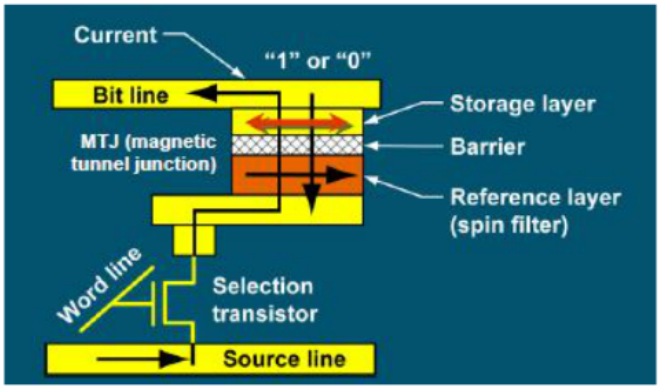
- STT-RAM: Spin Transfer Torque Random Access Memory



Prezi

The Solution: STT-RAM

- STT-RAM: Spin Transfer Torque Random Access Memory



Prezi

STT-RAM cell properties

Pros 

- Non-volatile
- High density (4x size in same area as SRAM)
- No leakage power
- Comparable read power and latency with SRAM

Cons 

- High write energy
- High write latency

Needs write operation management policy

Prezi

STT-RAM endurance

much more better than other non-volatile memory technologies

BUT

Not perfect as SRAM cells

$10^{16}\ for\ SRAM\ vs.\ 10^{15} for\ STT-RAM$

$become\ worse\ for\ MLC\ cells: under\ 10^{12}$

Prezi

Using STT-RAM as LLC in CMP could benefit from write traffic filtering by L1/L2 caches

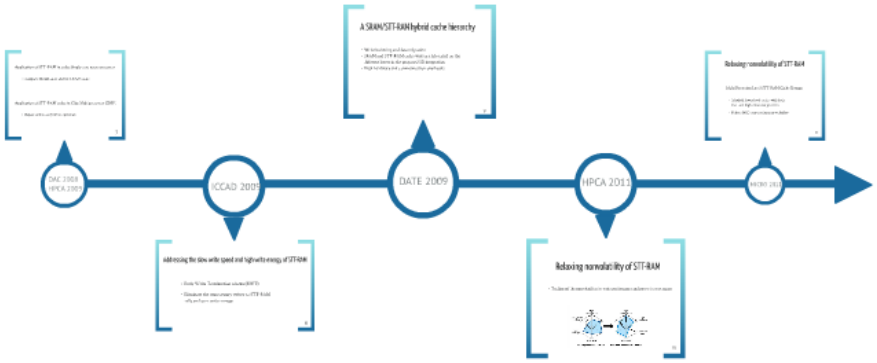
L1 caches in CMPs use write-through mechanism

- 3 level cache hierarchy
  - L1: write-through
  - L2: write-back

Prezi

Related Works

STT-RAMs in CMP caches



Related Works

STT-RAMs in CMP caches

Prezi

Application of STT-RAM in cache Single-core microprocessor

- Compare SRAM cache and STT-RAM cache

Application of STT-RAM cache in Chip Multiprocessor (CMP)

- Impact of the costly write operation

DAC 2008
HPCA 2009

Prezi

Application of STT-RAM in cache Single-core microprocessor

- Compare SRAM cache and STT-RAM cache

Application of STT-RAM cache in Chip Multiprocessor (CMP)

- Impact of the costly write operation

Prezi

ICCAD 2009

Addressing the slow write speed and high write energy of STT-RAM

- Early Write Termination scheme(EWT)
- Eliminate the unnecessary writes to STT-RAM cells and save write energy

Prezi

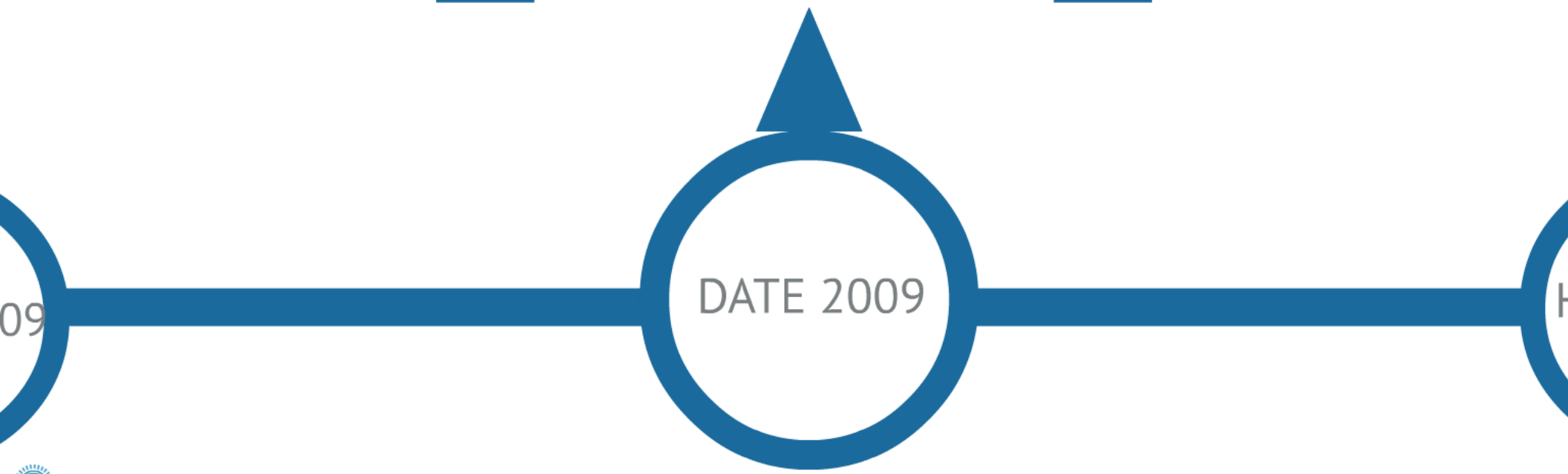
Addressing the slow write speed and high write energy of STT-RAM

- Early Write Termination scheme(EWT)
- Eliminate the unnecessary writes to STT-RAM cells and save write energy

Prezi

A SRAM/STT-RAM hybrid cache hierarchy

- Write buffering and data migration
- SRAM and STT-RAM cache ways are fabricated on the different layers in the proposed 3D integration
- High hardware and communication overheads



Prezi

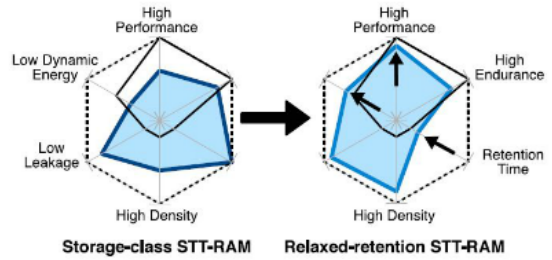
A SRAM/STT-RAM hybrid cache hierarchy

- Write buffering and data migration
- SRAM and STT-RAM cache ways are fabricated on the different layers in the proposed 3D integration
- High hardware and communication overheads

Prezi

Relaxing nonvolatility of STT-RAM

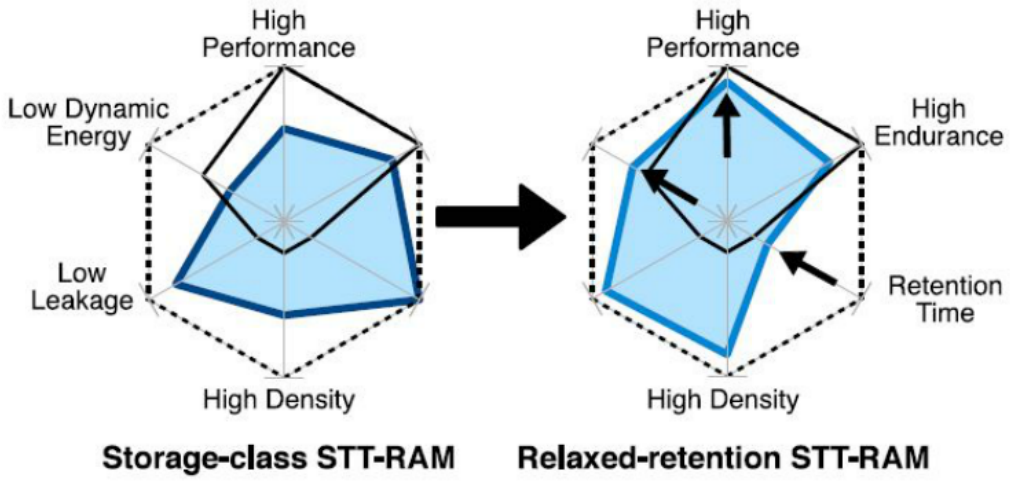
- Trading off the nonvolatility for write performance and power improvement



Prezi

Relaxing nonvolatility of STT-RAM

- Trading off the nonvolatility for write performance and power improvement



Prezi

Relaxing nonvolatility of STT-RAM

Multi Retention Level STT-RAM Cache Designs

- A hybrid lower level cache with both low- and high-retention portions
- Reduce MTJ area to relax non-volatility

11

Prezi

Relaxing nonvolatility of STT-RAM

Multi Retention Level STT-RAM Cache Designs

- A hybrid lower level cache with both low- and high-retention portions
- Reduce MTJ area to relax non-volatility

Prezi

STT-RAMs in CMP caches

DAC 2008
HPCA 2009

Application of STT-RAM in cache Single-core microprocessor

- Compare SRAM cache and STT-RAM cache

Application of STT-RAM cache in Chip Multiprocessor (CMP)

- Impact of the costly write operation

7

ICCAD 2009

Addressing the slow write speed and high write energy of STT-RAM

- Early Write Termination scheme(EWT)
- Eliminate the unnecessary writes to STT-RAM cells and save write energy

8

DATE 2009

A SRAM/STT-RAM hybrid cache hierarchy

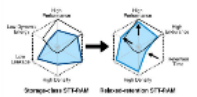
- Write buffering and data migration
- SRAM and STT-RAM cache ways are fabricated on the different layers in the proposed 3D integration
- High hardware and communication overheads

9

HPCA 2011

Relaxing nonvolatility of STT-RAM

- Trading off the nonvolatility for write performance and power improvement



10

MICRO 2011

Relaxing nonvolatility of STT-RAM

Multi Retention Level STT-RAM Cache Designs

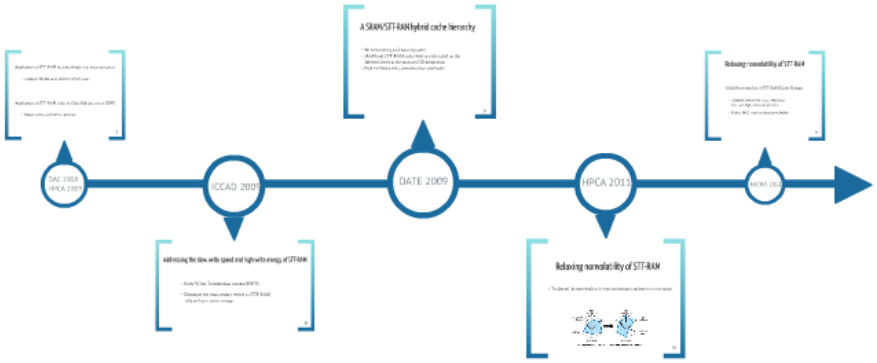
- A hybrid lower level cache with both low- and high-retention portions
- Reduce MTJ area to relax non-volatility

11

Prezi

Related Works

STT-RAMs in CMP caches



Proposed Architecture

Ideal cache requirements:

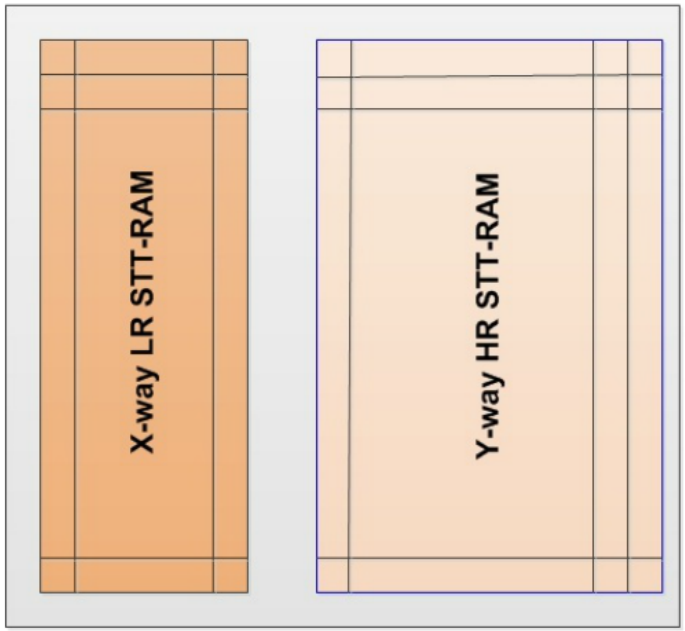
- High retention STT-RAM array
  - low leakage energy
  - non-volatility
  - low read time and energy

Prezi

Ideal cache requirements:

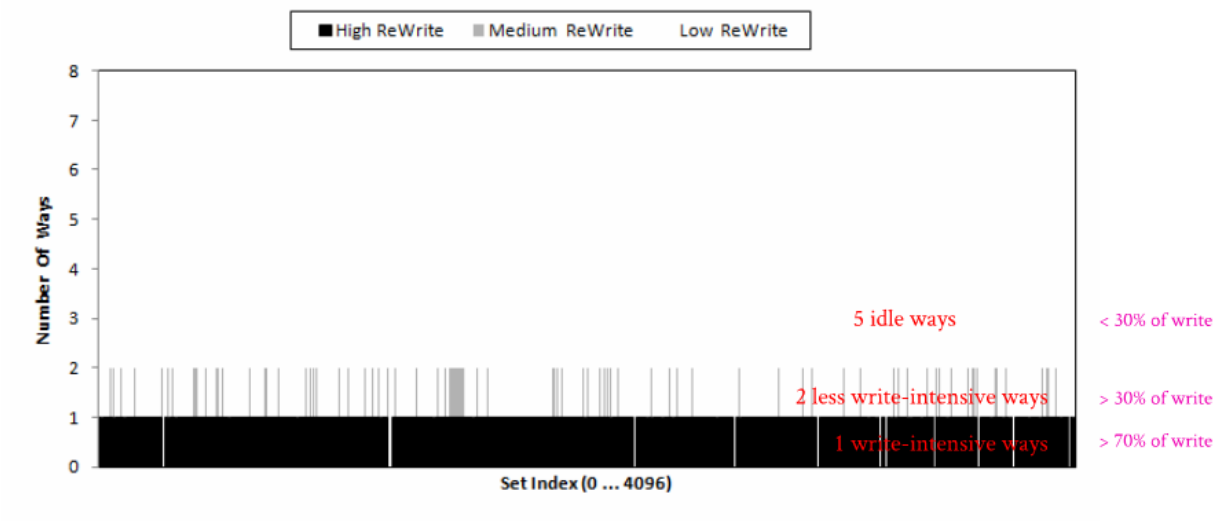
- High retention STT-RAM array
  - low leakage energy
  - non-volatility
  - low read time and energy

- Low retention STT-RAM array
  - Low write time and energy



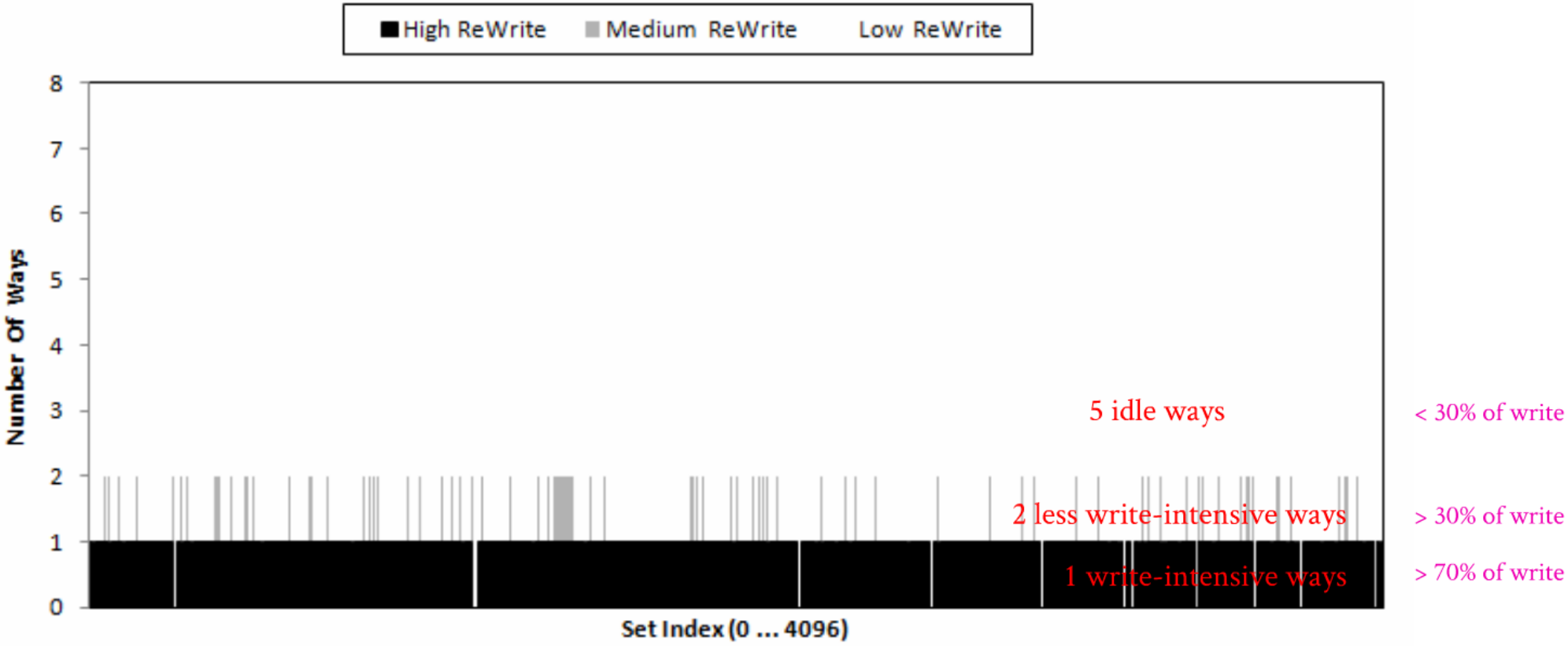
Prezi

Motivation: write distribution within each set



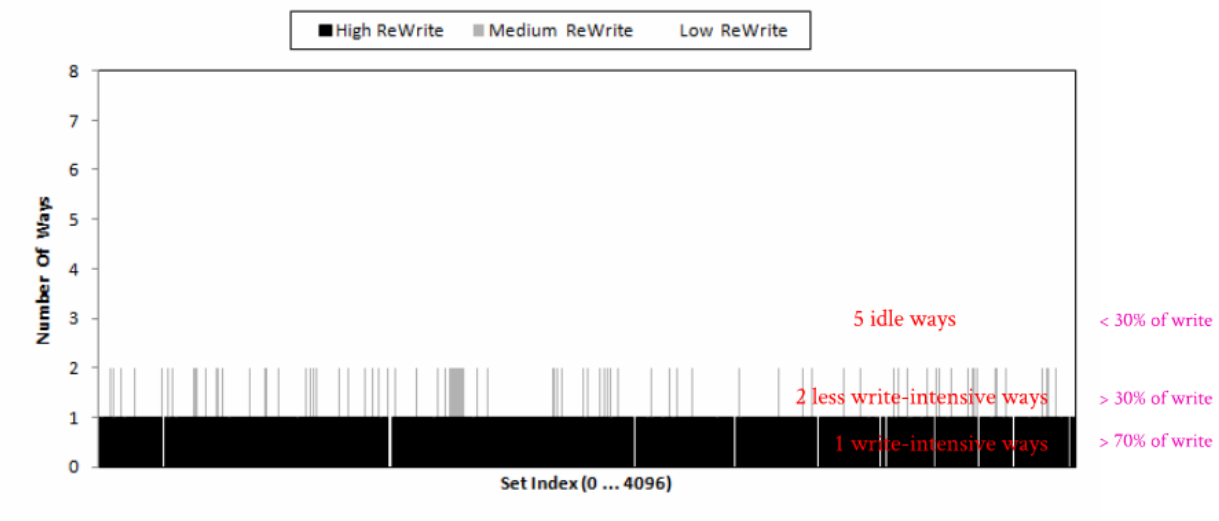
So

- Non-uniform writes within a set distributed
- Limited ways with intensive writes (1 way)


High ReWrite
Medium ReWrite
Low ReWrite
Number Of Ways
8
7
6
5
4
3
2
1
0
Set Index (0 ... 4096)
5 idle ways
< 30% of write
2 less write-intensive ways
> 30% of write
1 write-intensive ways
> 70% of write

Prezi

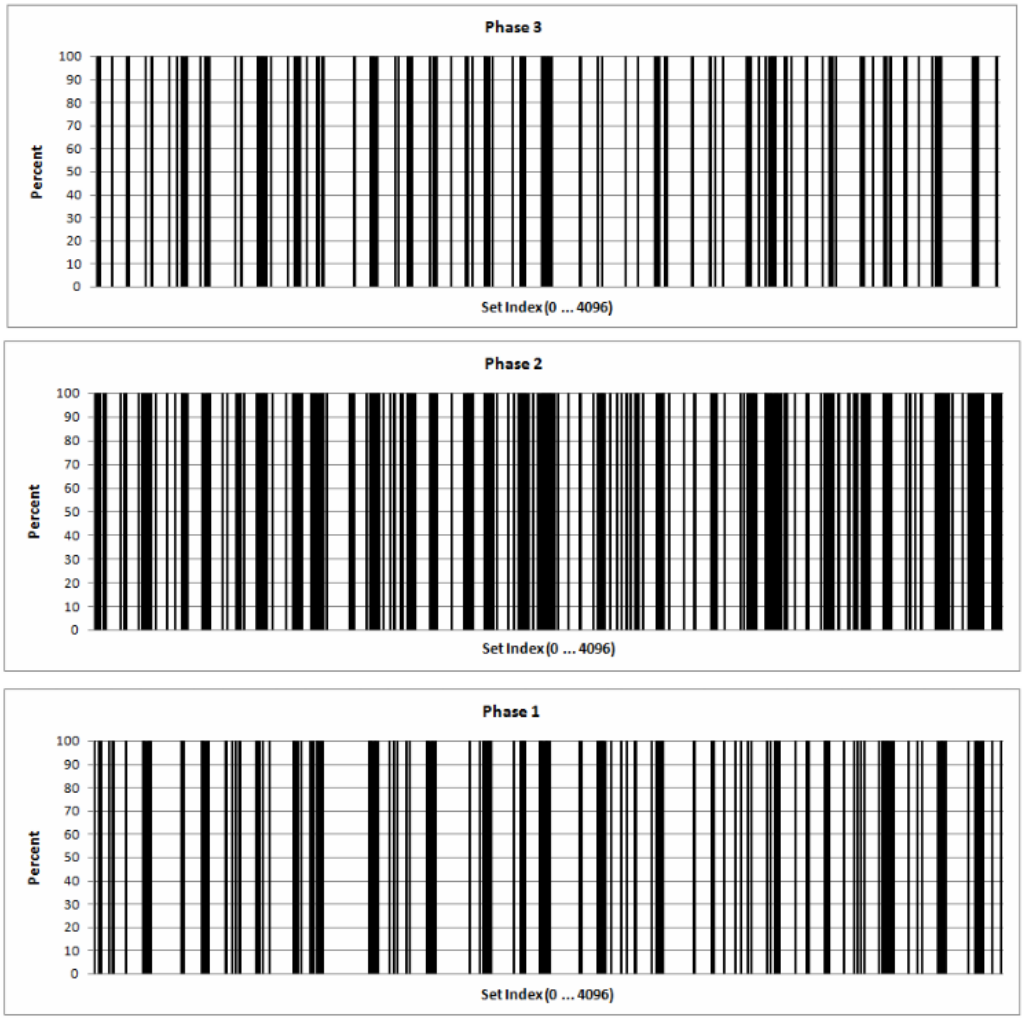
Motivation: write distribution within each set



So

- Non-uniform writes within a set distributed
- Limited ways with intensive writes (1 way)

Motivation: write distribution between different sets



Prezi

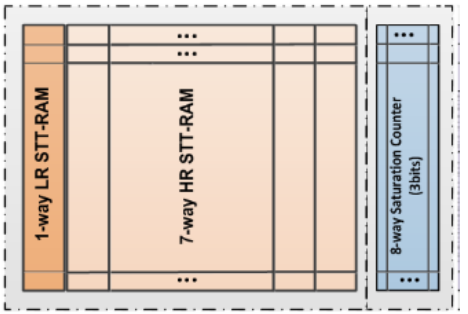
Managing writes onto cache arrays

- maintain write-stressed data into the LR STT-RAM array within each set

Improve utilization of LR-region

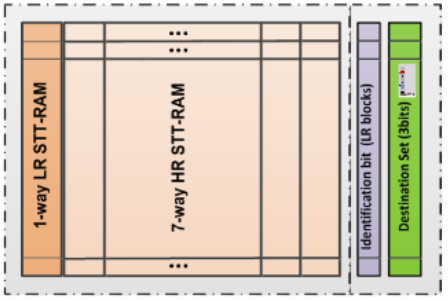
- Intra-Set Write Management
  - Per way small SCs

Intra-Set Write Management



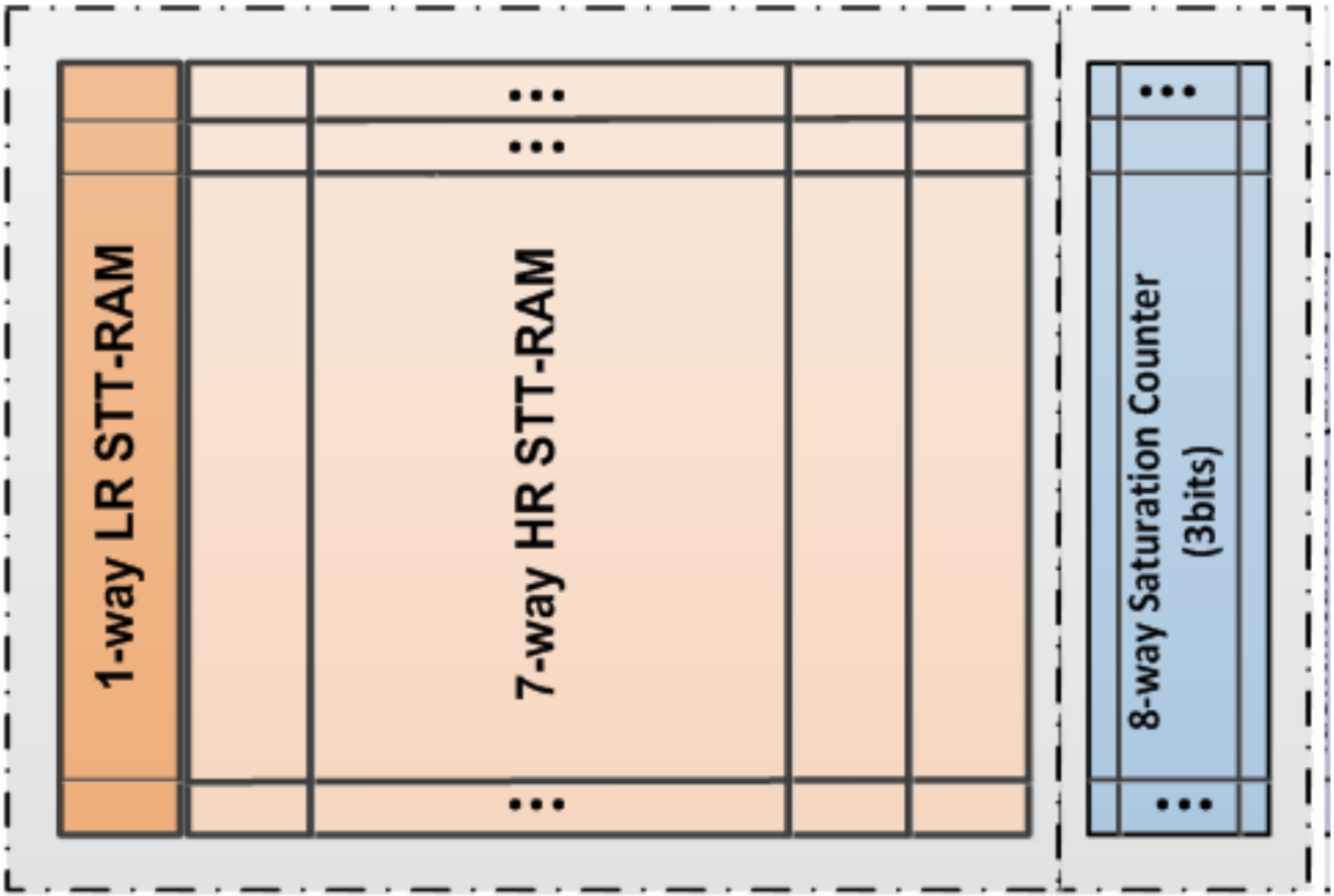
- Inter-Set Write Management
  - Semi-static destination set selection
  - Keep trace of a established redirection(3bit per set)
  - Searching cache
  - Breaking a merge

Inter-Set Write Management



Prezi

Intra-Set Write Management



Prezi

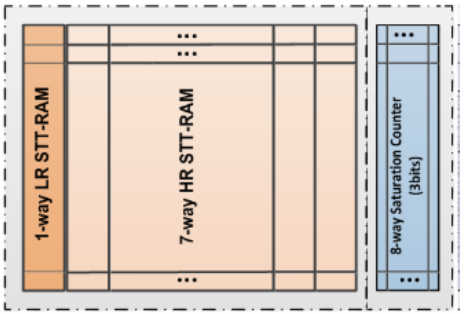
Managing writes onto cache arrays

- maintain write-stressed data into the LR STT-RAM array within each set

Improve utilization of LR-region

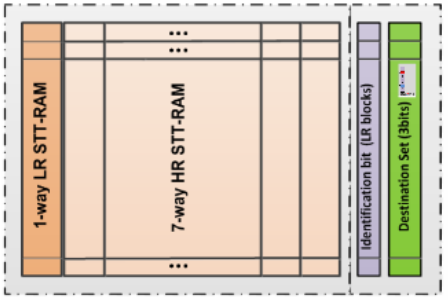
- Intra-Set Write Management
  - Per way small SCs

Intra-Set Write Management



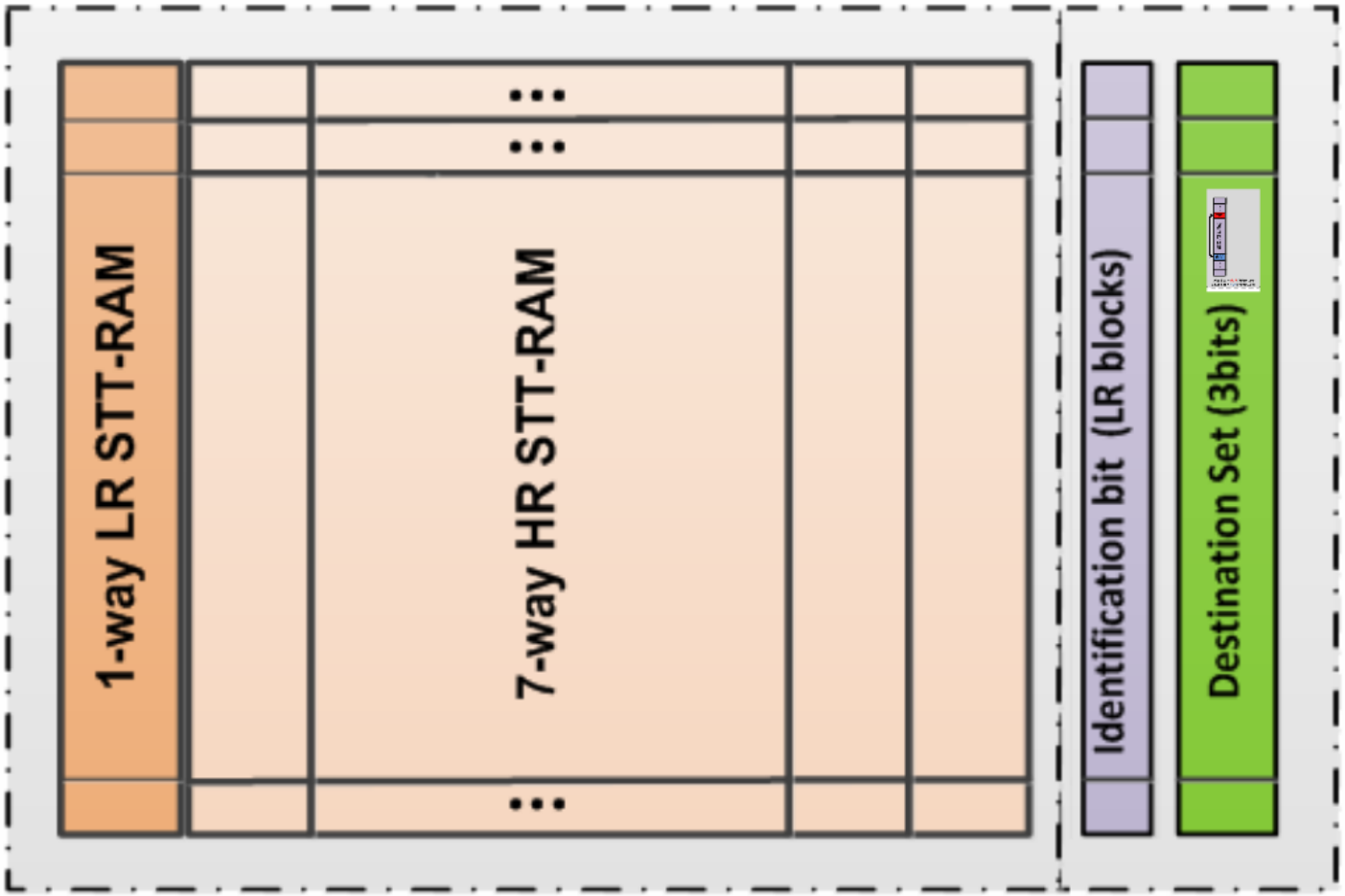
- Inter-Set Write Management
  - Semi-static destination set selection
  - Keep trace of a established redirection(3bit per set)
  - Searching cache
  - Breaking a merge

Inter-Set Write Management



Prezi

Inter-Set Write Management

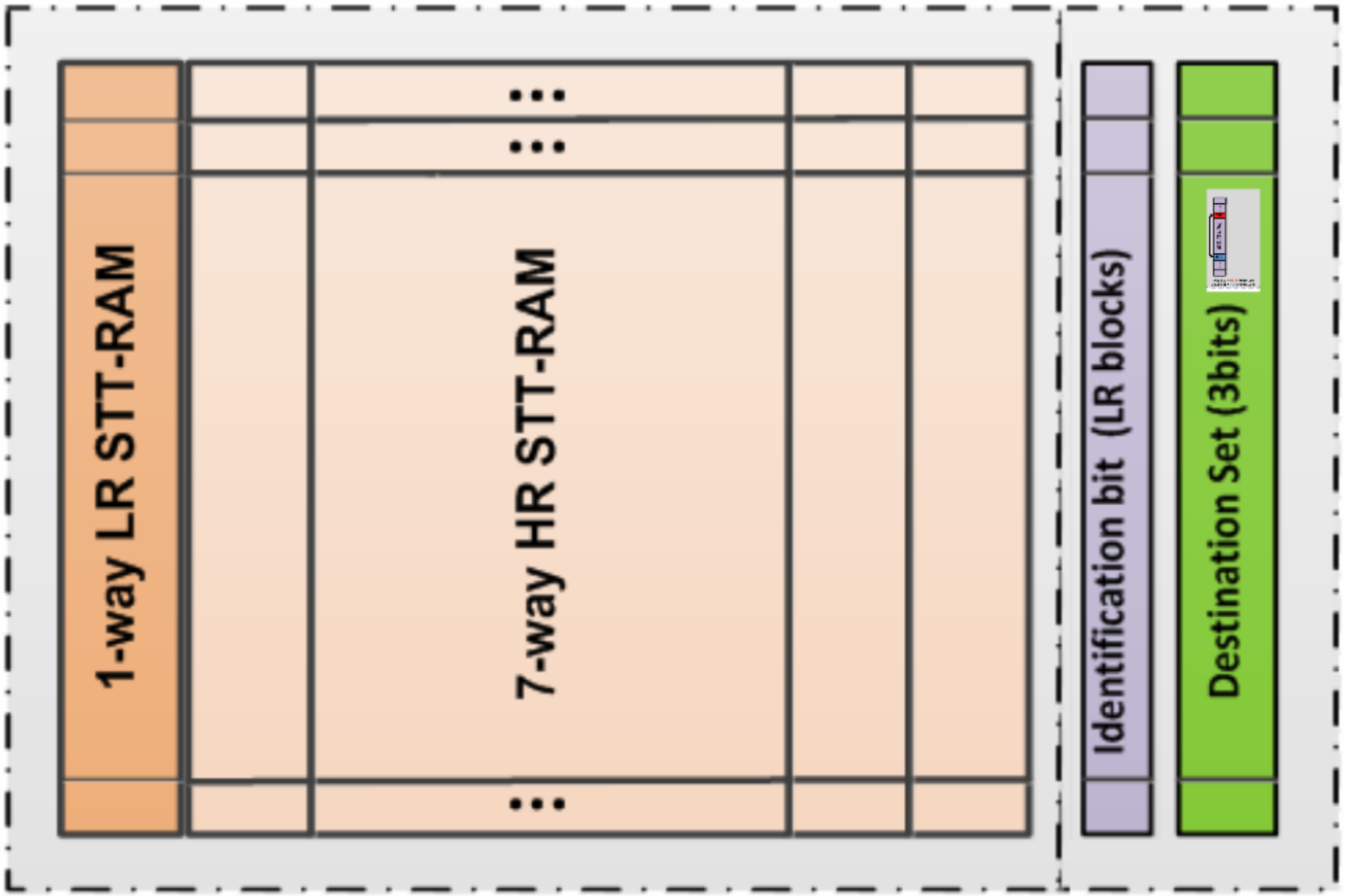


Hot Set = 01111010...011

Cold Set = 10111010...011

Prezi

Inter-Set Write Management



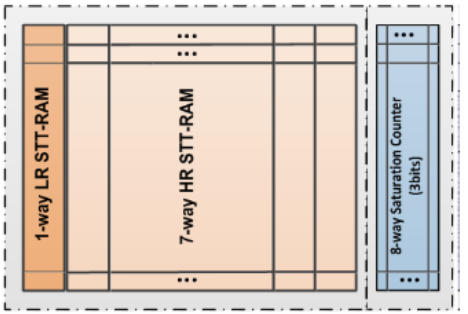
Managing writes onto cache arrays

- maintain write-stressed data into the LR STT-RAM array within each set

Improve utilization of LR-region

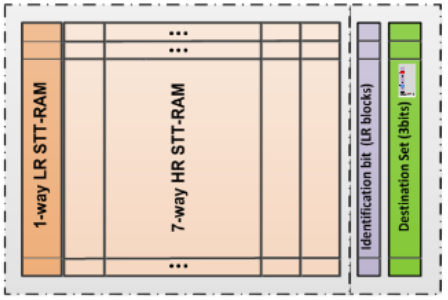
- Intra-Set Write Management
  - Per way small SCs

Intra-Set Write Management



- Inter-Set Write Management
  - Semi-static destination set selection
  - Keep trace of a established redirection(3bit per set)
  - Searching cache
  - Breaking a merge

Inter-Set Write Management

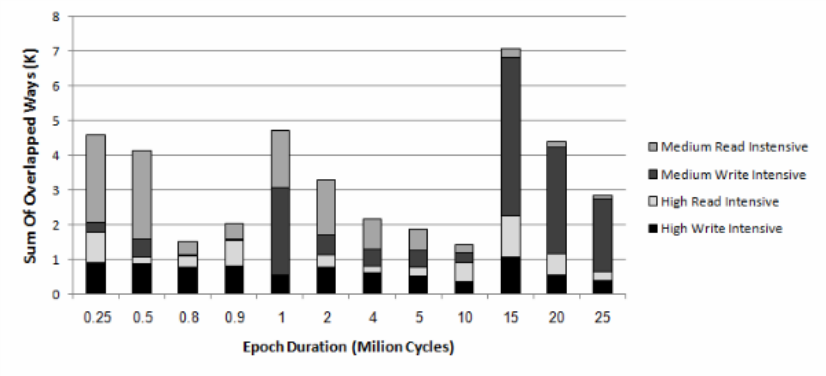


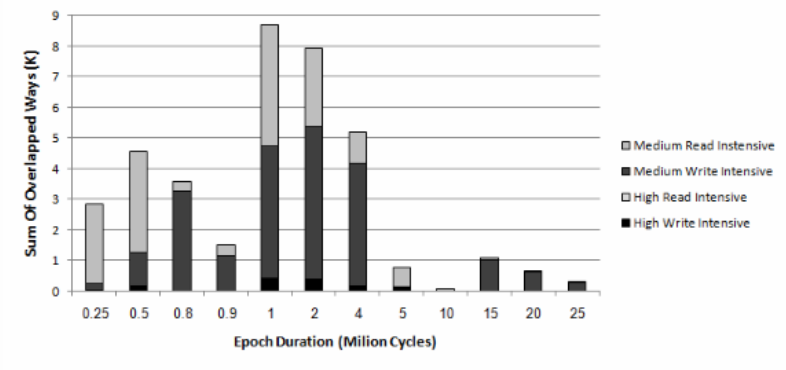
Prezi

Best time interval

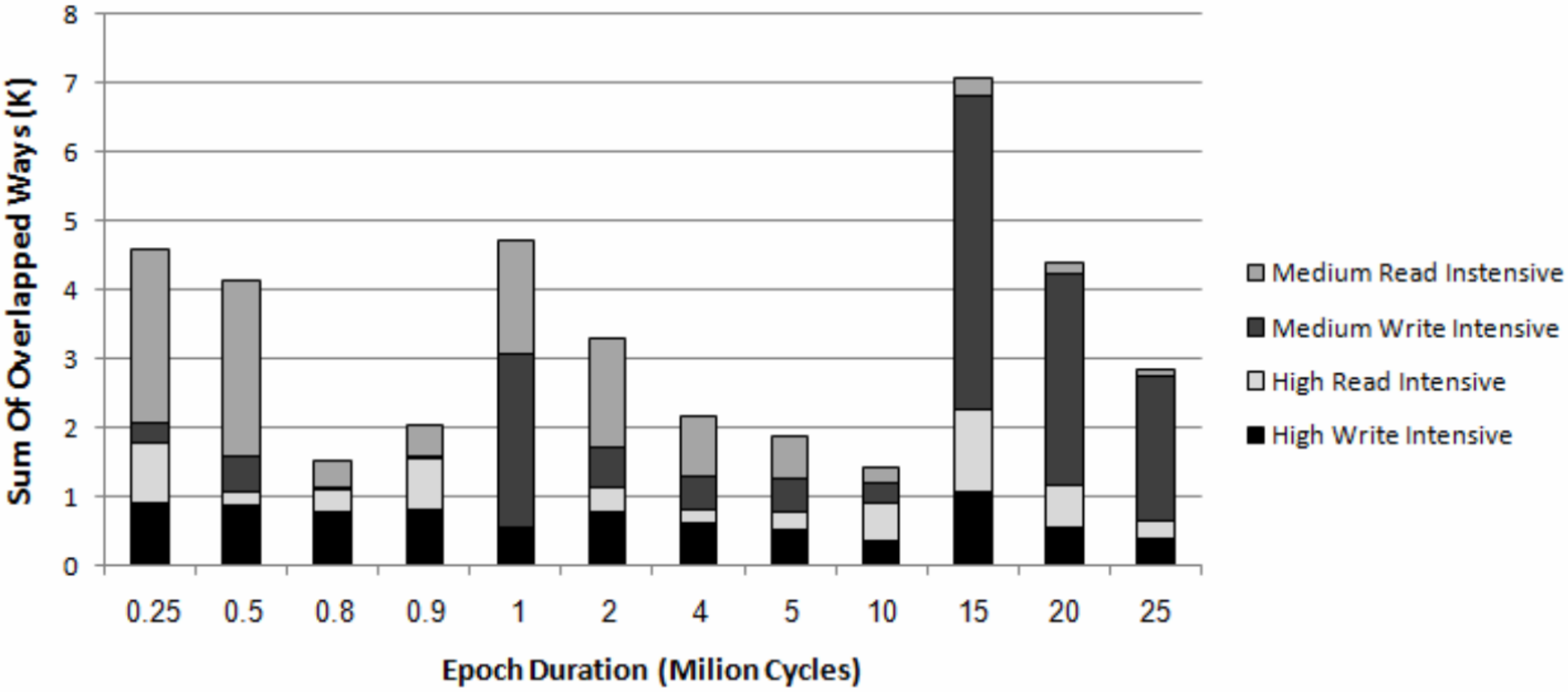
At the beginning of optimum time interval :

- Redirect W-intensive blocks from HR-region to LR-region in a set



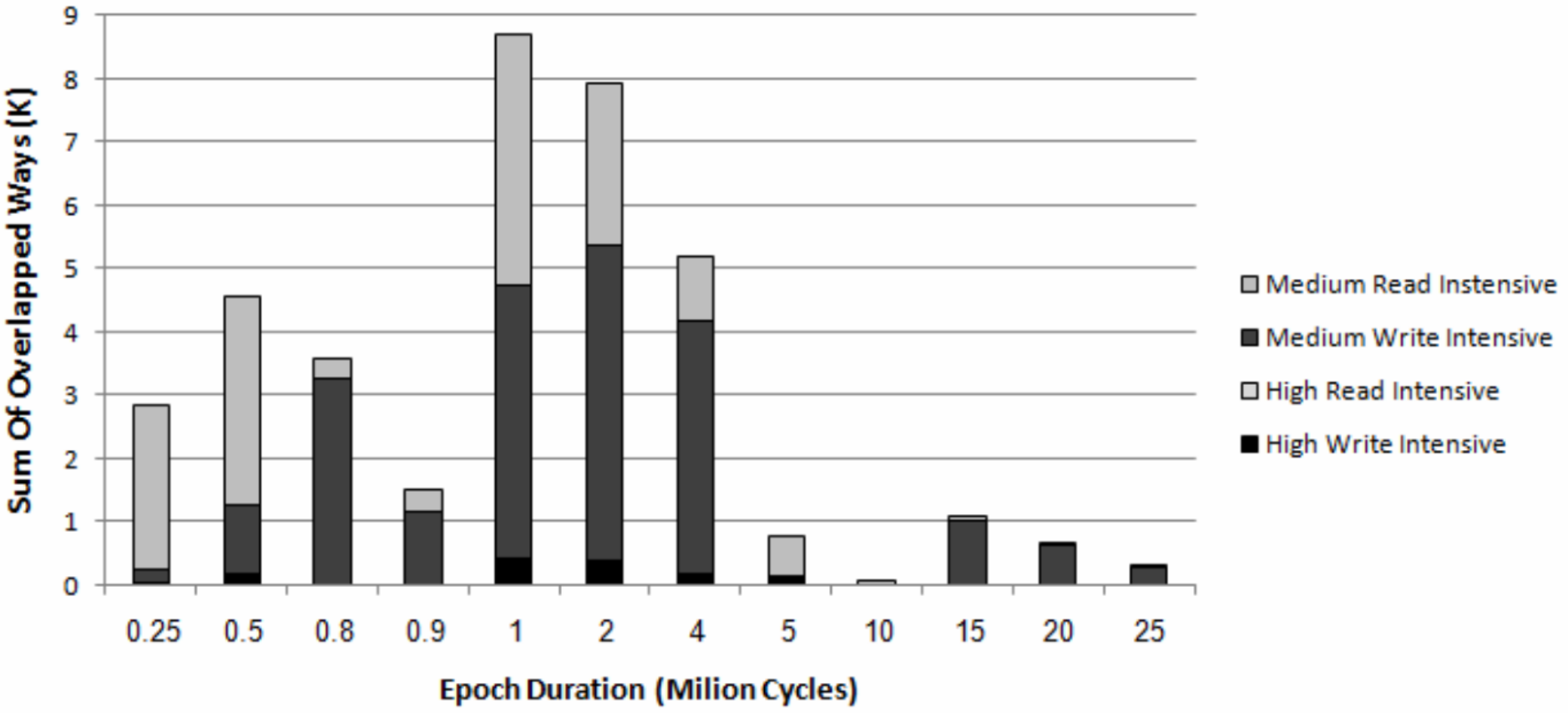


Prezi


Sum Of Overlapped Ways (K)
0
1
2
3
4
5
6
7
8
0.25
0.5
0.8
0.9
1
2
4
5
10
15
20
25
Epoch Duration (Milion Cycles)
Medium Read Instensive
Medium Write Intensive
High Read Intensive
High Write Intensive


9
8
7

Prezi

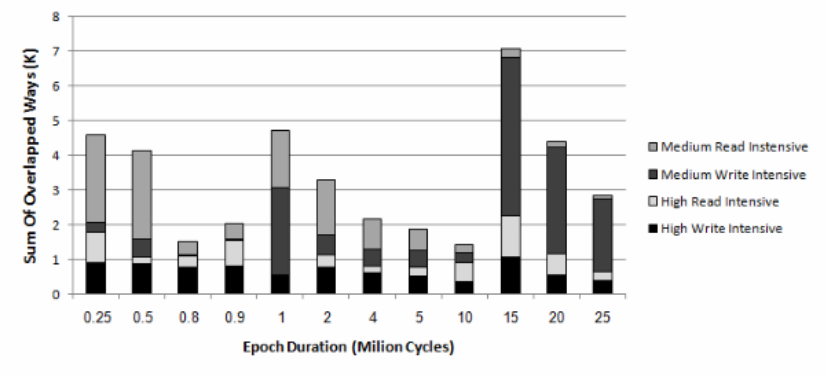

Sum Of Overlapped Ways (K)
9
8
7
6
5
4
3
2
1
0
0.25
0.5
0.8
0.9
1
2
4
5
10
15
20
25
Epoch Duration (Milion Cycles)
Medium Read Instensive
Medium Write Intensive
High Read Intensive
High Write Intensive

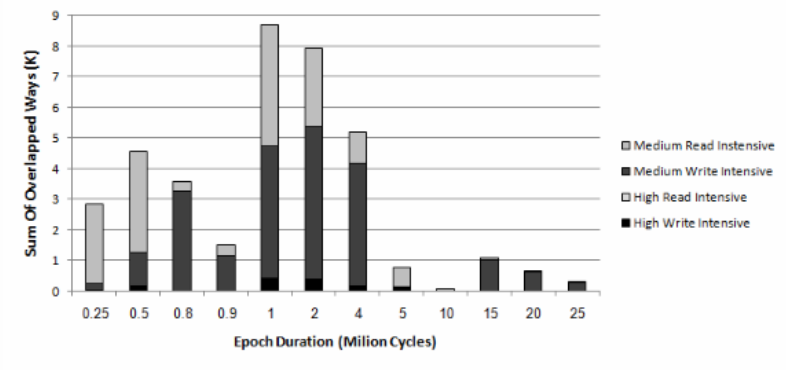
Prezi

Best time interval

At the beginning of optimum time interval :

- Redirect W-intensive blocks from HR-region to LR-region in a set





Prezi

latency and energy of LR & HR cache

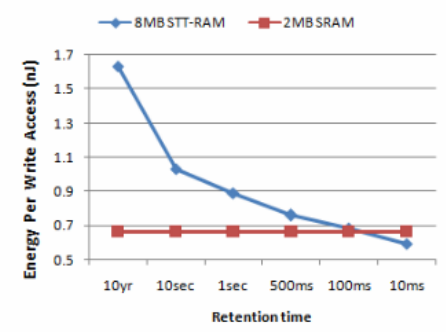
$$T_{store} = 1\text{ns}.e^{\Delta}$$

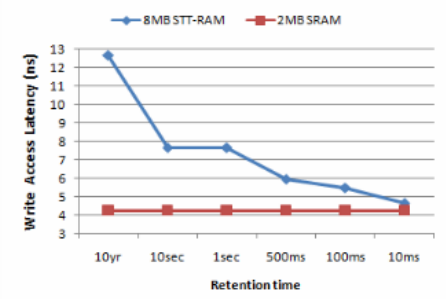
$$\Delta = \frac{M_s H_k A t_F}{2k_B T}$$

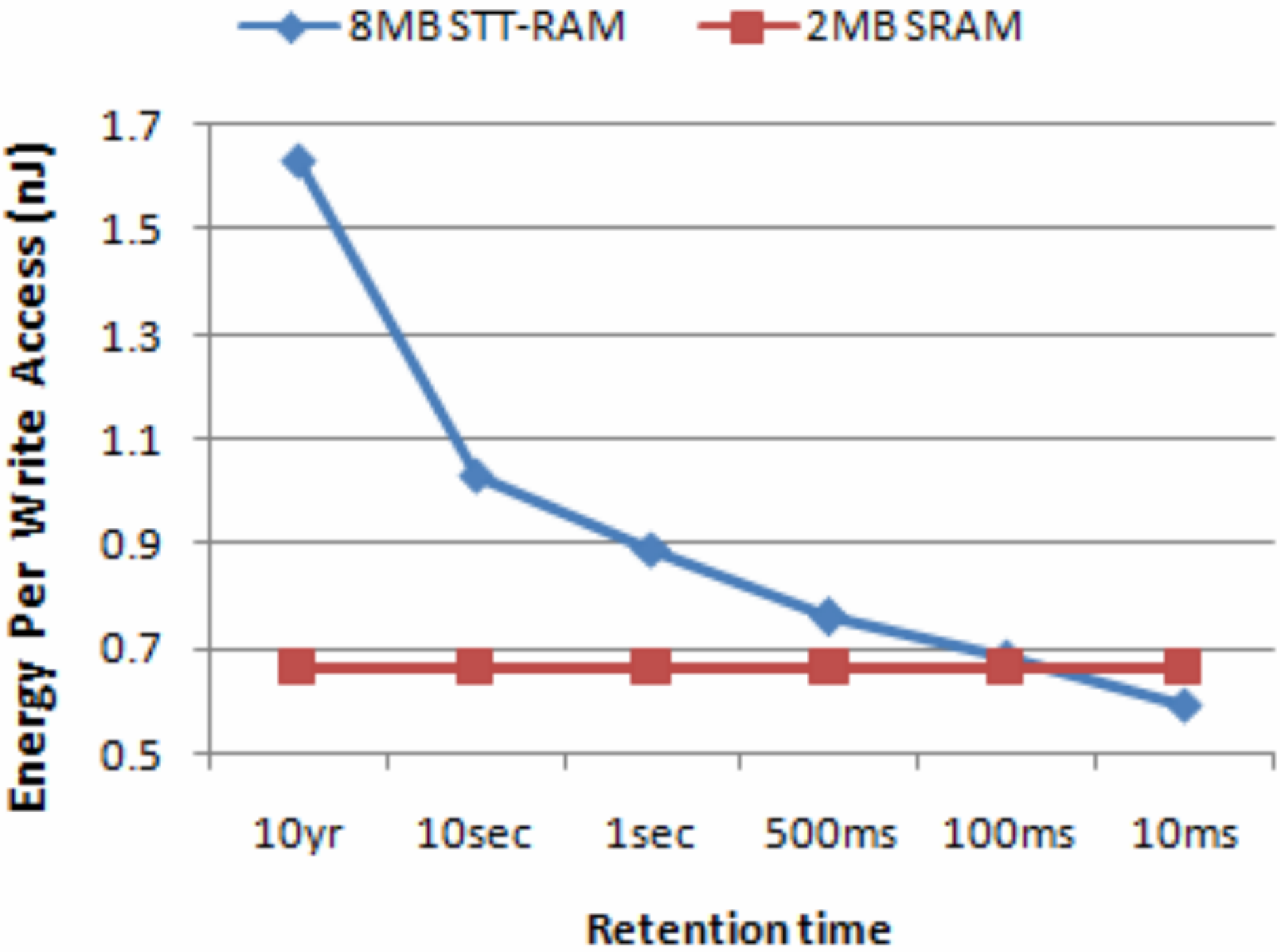
$$J_{c0} = (\frac{2e}{\text{h}})(\frac{\alpha}{\eta})(t_F M_s)(H_k \pm H_{ext} + 2\pi M_s)$$

$$I_c(\tau) = A \times J_c(\tau)$$

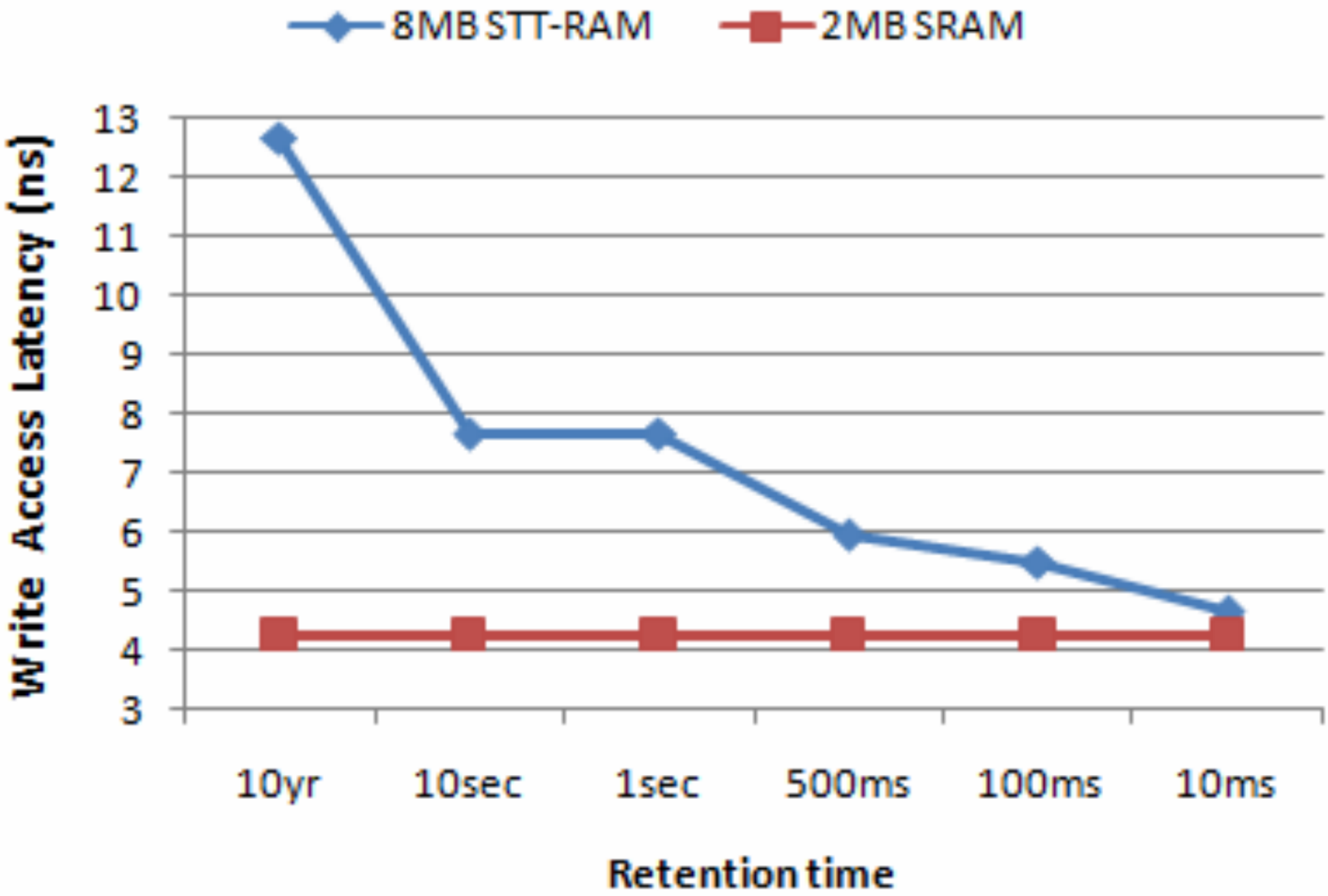
$$Write\ Energy = V_{Write} \times I_c(\tau) \times \tau$$






8MB STT-RAM
2MB SRAM
Energy Per Write Access (nJ)
1.7
1.5
1.3
1.1
0.9
0.7
0.5
10yr
10sec
1sec
500ms
100ms
10ms
Retention time

Prezi


8MB STT-RAM
2MB SRAM
Write Access Latency (ns)
13
12
11
10
9
8
7
6
5
4
3
10yr
10sec
1sec
500ms
100ms
10ms
Retention time

Prezi

latency and energy of LR & HR cache

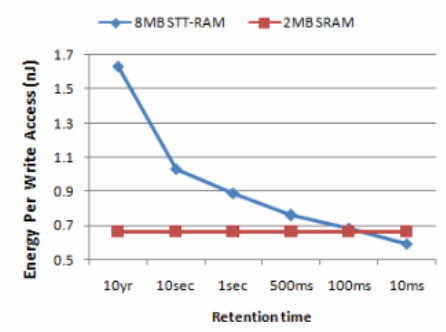
$$T_{store} = 1\text{ns}.e^{\Delta}$$

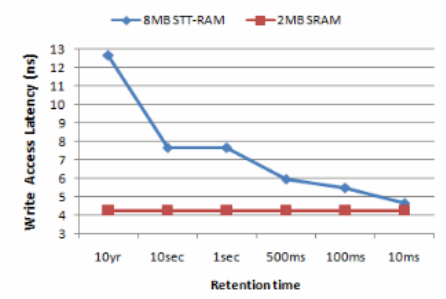
$$\Delta = \frac{M_s H_k A t_F}{2k_B T}$$

$$J_{c0} = (\frac{2e}{\text{h}})(\frac{\alpha}{\eta})(t_F M_s)(H_k \pm H_{ext} + 2\pi M_s)$$

$$I_c(\tau) = A \times J_c(\tau)$$

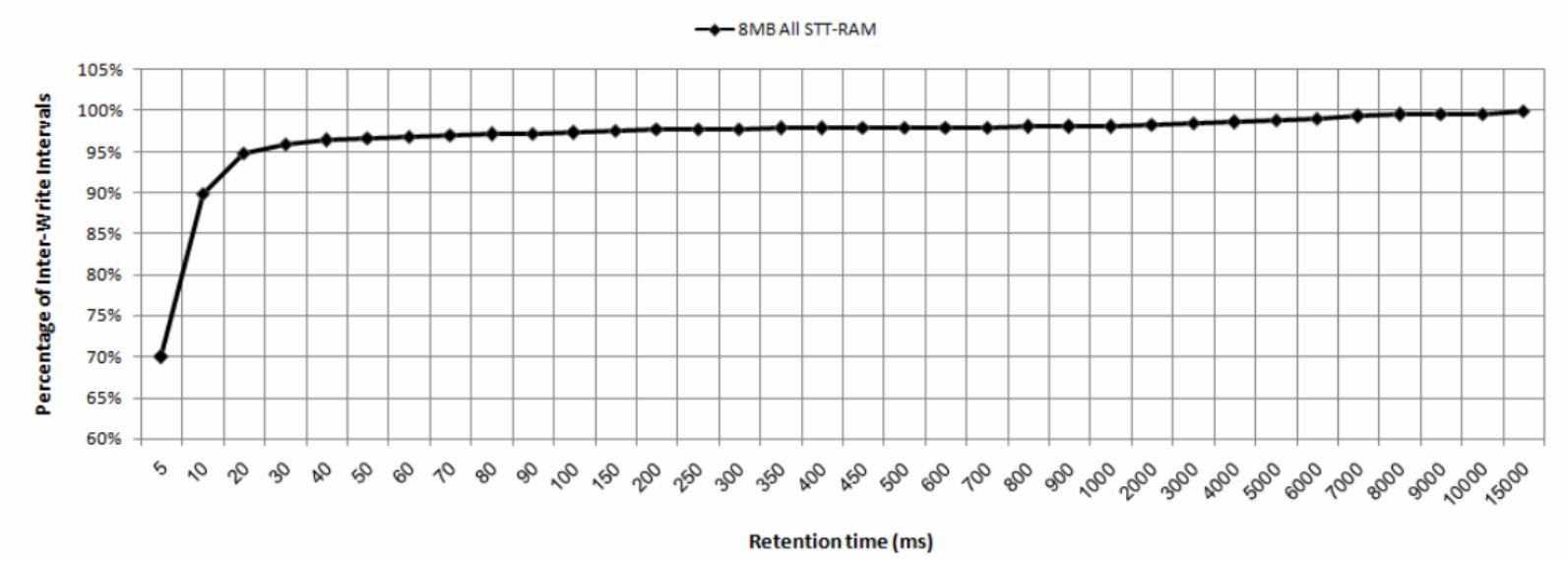
$$Write\ Energy = V_{Write} \times I_c(\tau) \times \tau$$





LR optimal retention time

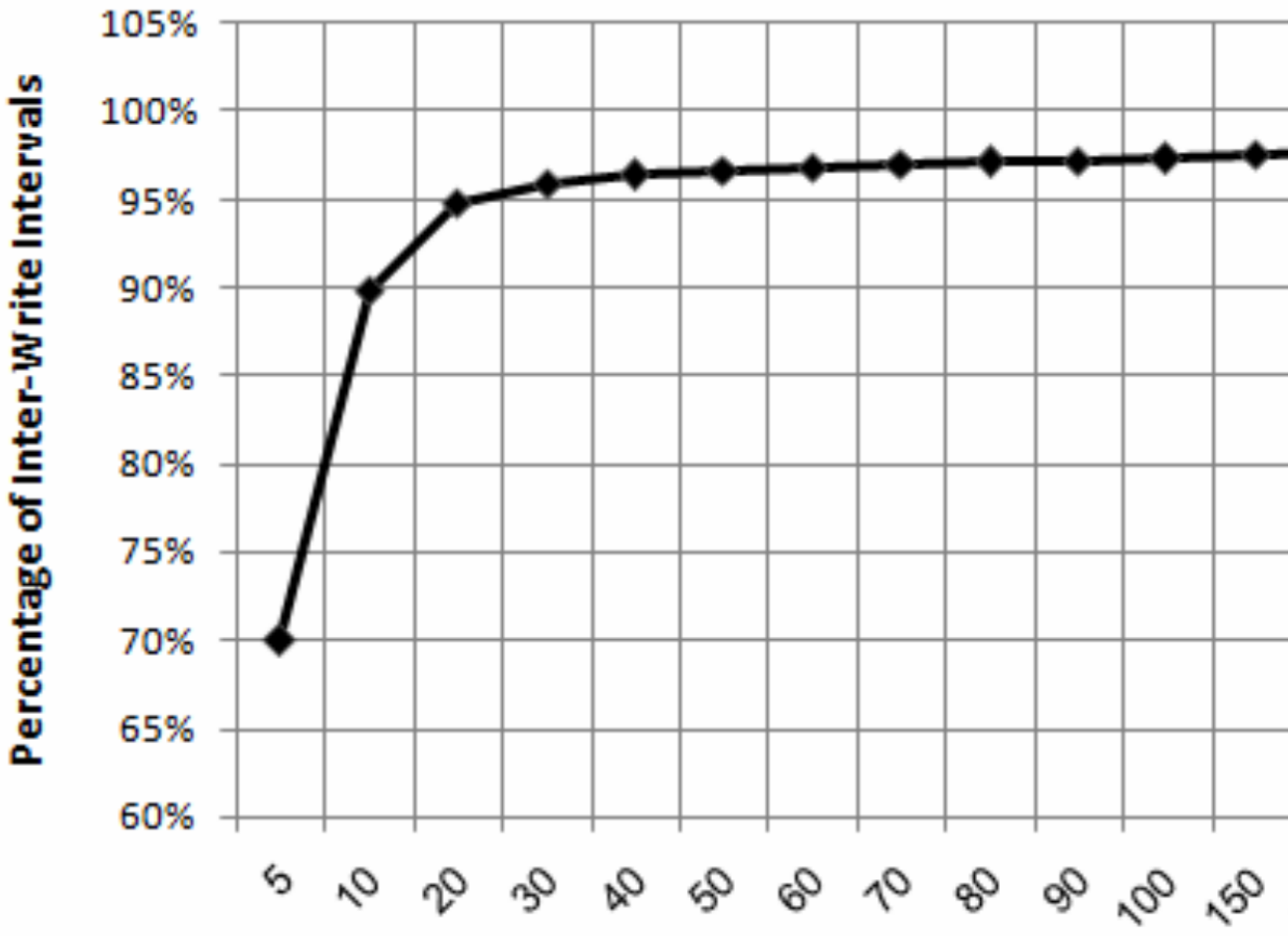
- More than 90% of the inter-write time intervals are smaller than 10ms



Reducing the retention time to micro second scale:

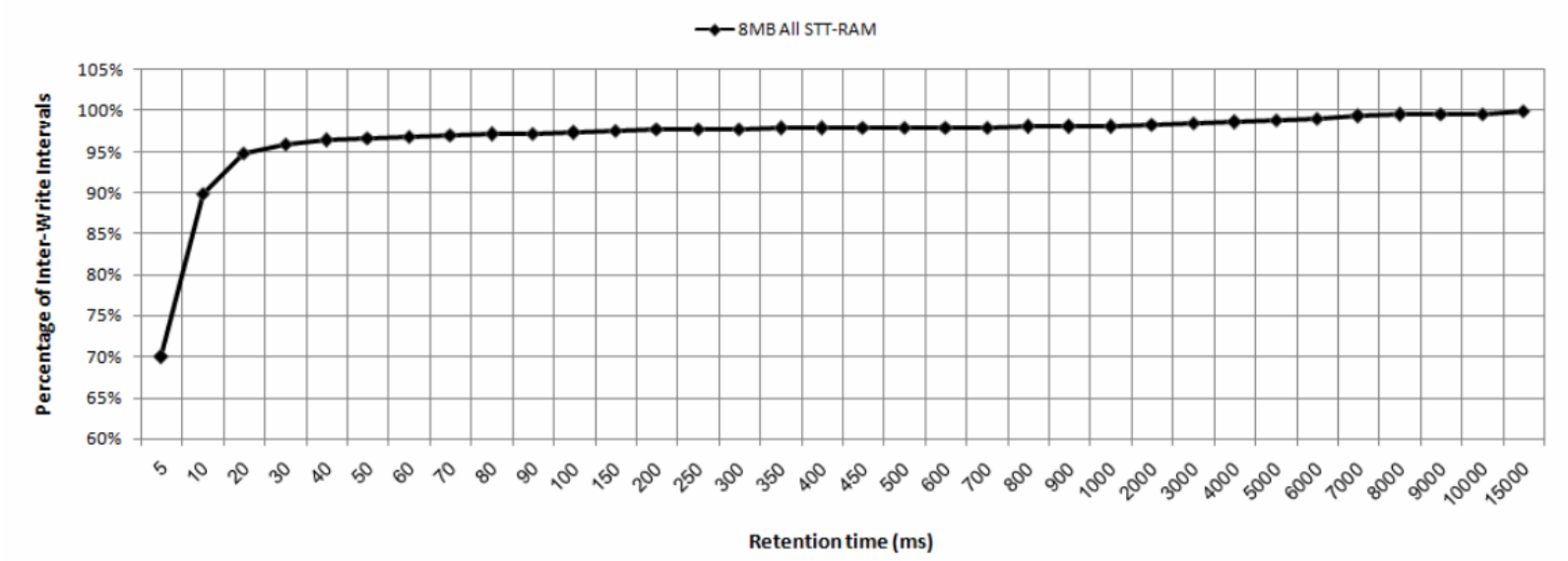
- Highly volatile cache
- Any refresh scheme becomes impractical for the large lower level cache.
- Degraded performance

Prezi


Percentage of Inter-Write Intervals
105%
100%
95%
90%
85%
80%
75%
70%
65%
60%
5
10
20
30
40
50
60
70
80
90
100
150

LR optimal retention time

- More than 90% of the inter-write time intervals are smaller than 10ms



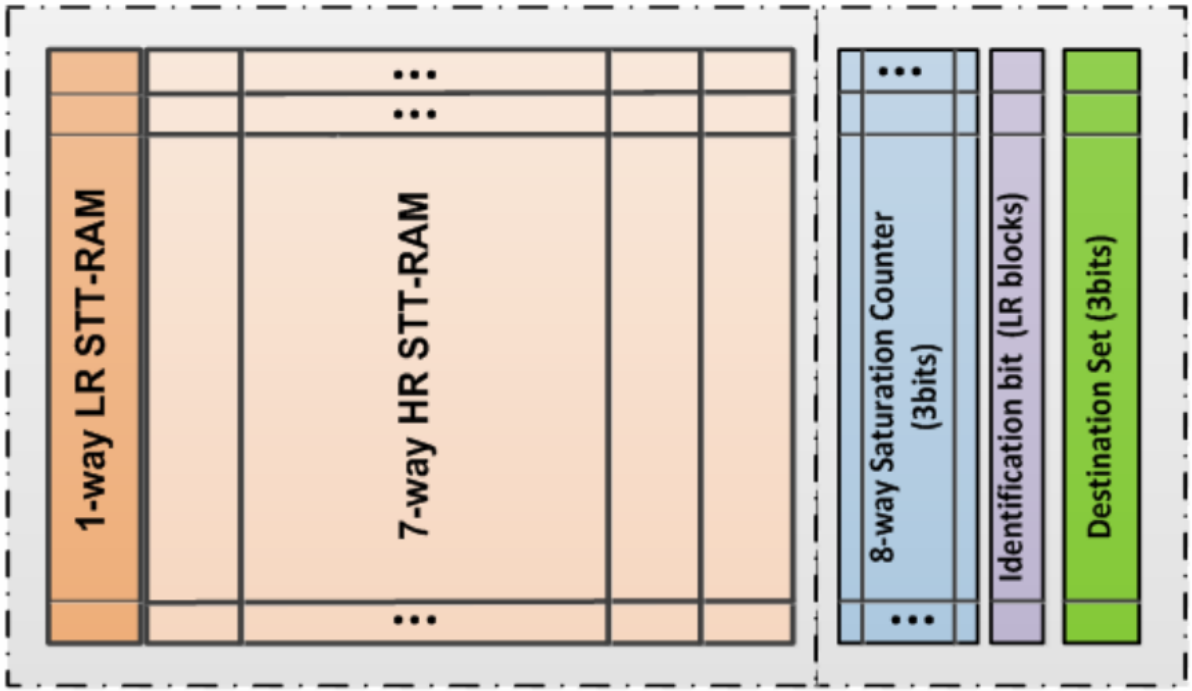
Reducing the retention time to micro second scale:

- Highly volatile cache
- Any refresh scheme becomes impractical for the large lower level cache.
- Degraded performance

Prezi

Redundancy for each set :

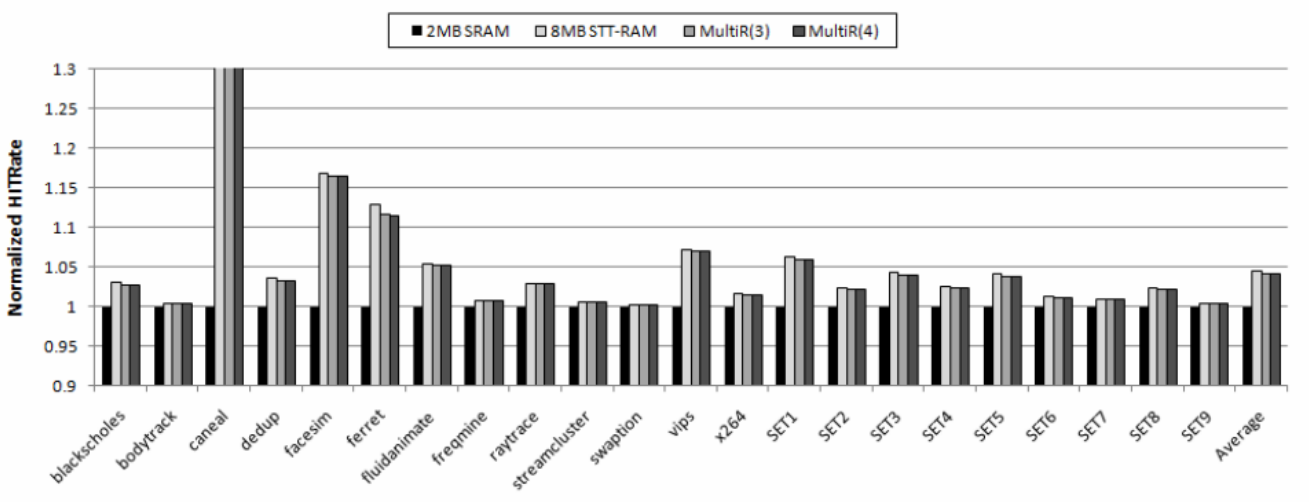
- One Saturation Counter for each way in HR-region (w-intensive block)
- 1 bit for each way in LR-region (valid bit)
- 3 bits destination set identification for each set



Prezi

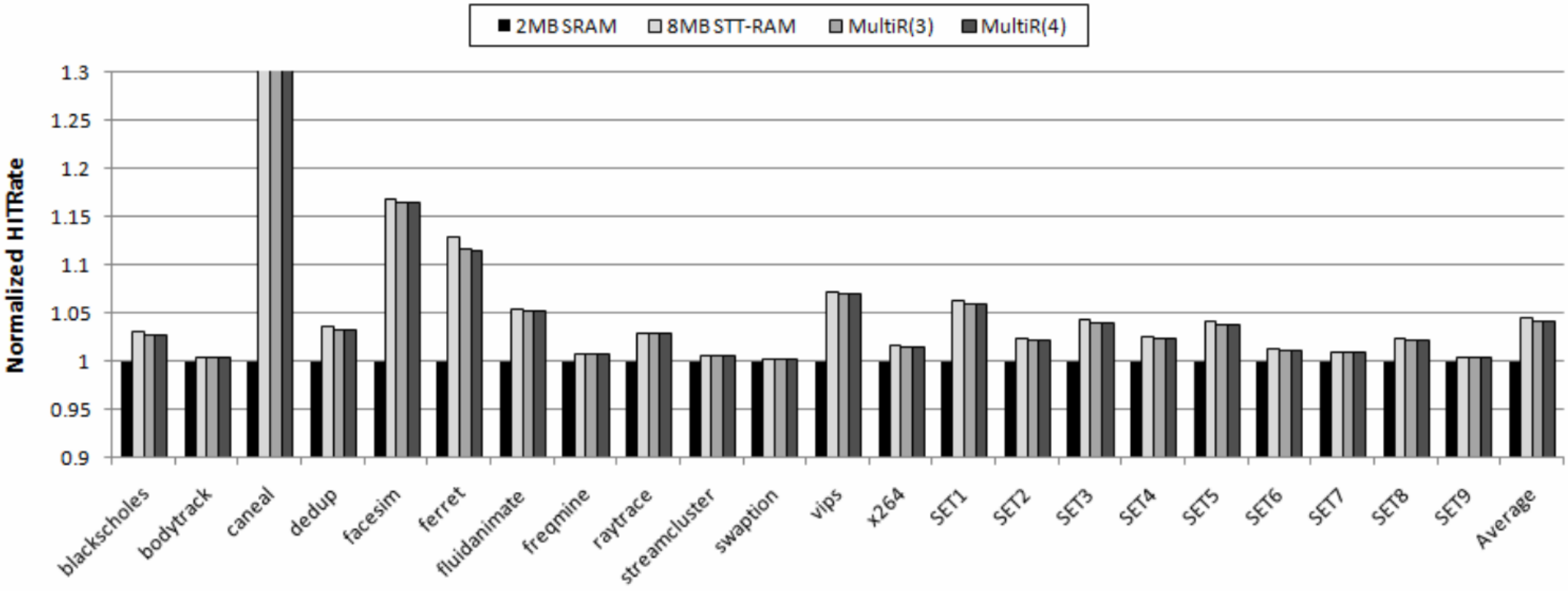
Refreshing mechanism for data blocks in LR region

- Maintain a copy of LR block in one HR block in each set
- Refresh overhead :



Prezi

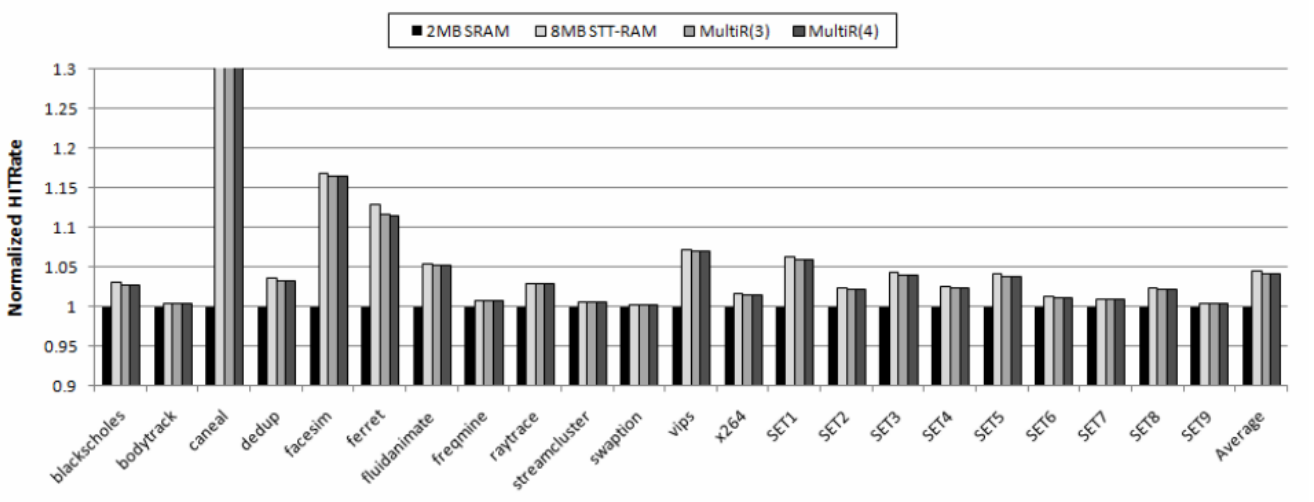
- Maintain a copy of LR block in one HR block in each set
- Refresh overhead :



Prezi

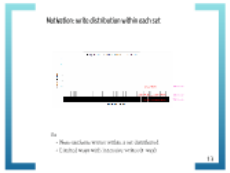
Refreshing mechanism for data blocks in LR region

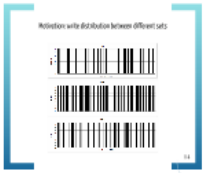
- Maintain a copy of LR block in one HR block in each set
- Refresh overhead :

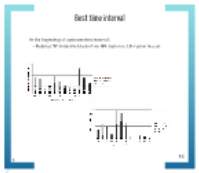


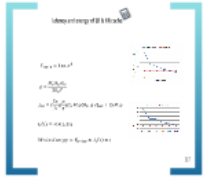
Prezi

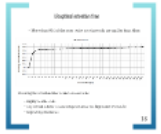
Proposed Architecture

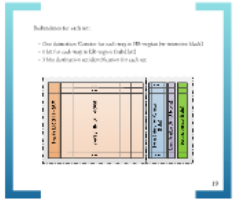


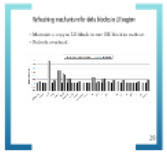




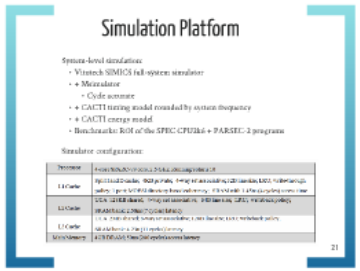


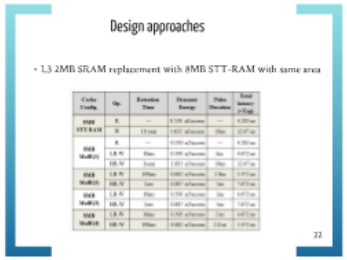


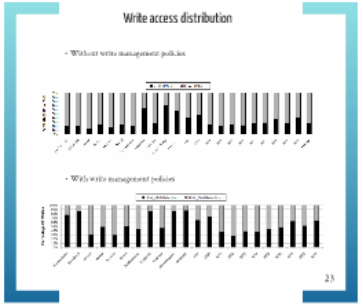


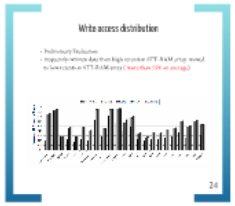


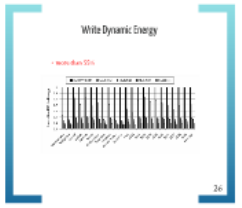
Evaluation











Prezi

Evaluation

Simulation Platform

System-level simulation:

- Vitutech SIMICS full-system simulator
- + Msimulator
  - Cycle accurate
- + CACTI timing model rounded by system frequency
- + CACTI energy model
- Benchmarks: ROI of the SPEC CPU2k6 + PARSEC-2 programs

Simulator configuration:

Processor	4-core SPARC-v9 core, 2.5 GHz, Running Solaris 10
L1 Cache	Split I and D cache; 4KB private; 4-way set associative; 32B line size; LRU; write-through policy; 1 port; MOESI directory-based coherency; SRAM with 1.45ns (4 cycles) access time
L2 Cache	UCA 128KB shared; 4-way set associative; 64B line size; LRU; writeback policy; SRAM bank: 2.58ns (7 cycles) latency
L3 Cache	UCA 2MB shared; 8-way set associative; 128B line size; LRU; writeback policy; SRAM bank: 4.23n (11 cycles) latency
Main Memory	4GB DRAM; 50ns (200 cycles) access latency

Prezi

Design approaches

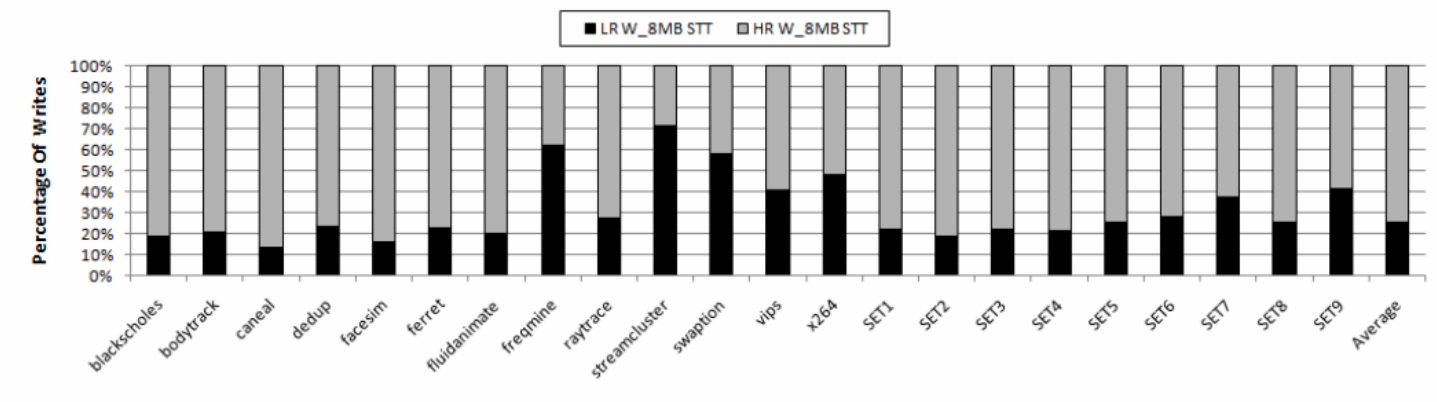
- L3 2MB SRAM replacement with 8MB STT-RAM with same area

Cache Config.	Op.	Retention Time	Dynamic Energy	Pulse Duration	Total latency (+Tag)
8MB STT-RAM	R	---	0.530 nJ/access	---	4.283 ns
	W	10 year	1.635 nJ/access	10ns	12.67 ns
8MB MulR(1)	R	---	0.530 nJ/access	---	4.283 ns
	LR-W	10ms	0.593 nJ/access	2ns	4.672 ns
	HR-W	1year	1.635 nJ/access	10ns	12.67 ns
8MB MulR(2)	LR-W	100ms	0.682 nJ/access	2.8ns	5.472 ns
	HR-W	1sec	0.887 nJ/access	5ns	7.672 ns
8MB MulR(3)	LR-W	10ms	0.593 nJ/access	2ns	4.672 ns
	HR-W	1sec	0.887 nJ/access	5ns	7.672 ns
8MB MulR(4)	LR-W	10ms	0.593 nJ/access	2 ns	4.672 ns
	HR-W	100ms	0.682 nJ/access	2.8 ns	5.472 ns

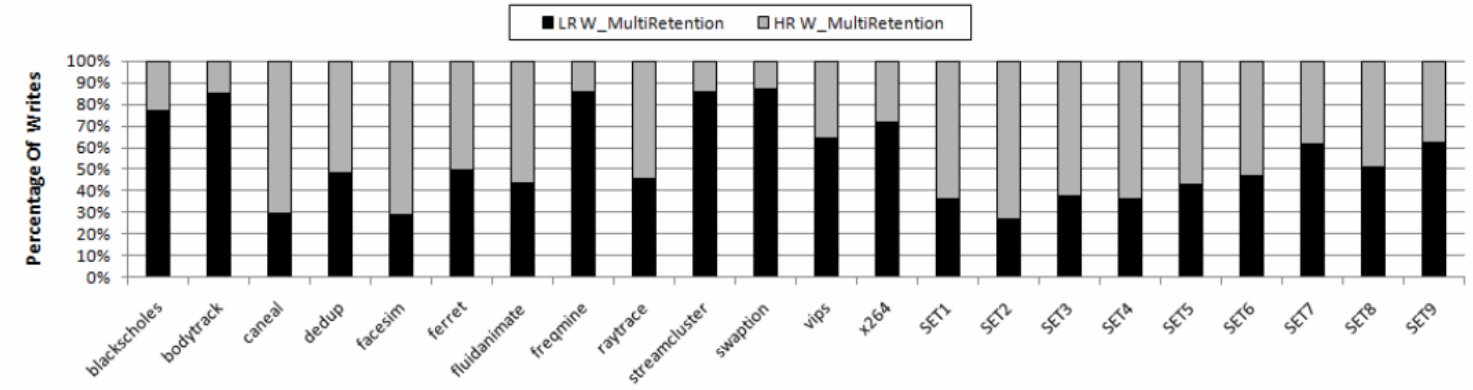
Prezi

Write access distribution

- Without write management policies



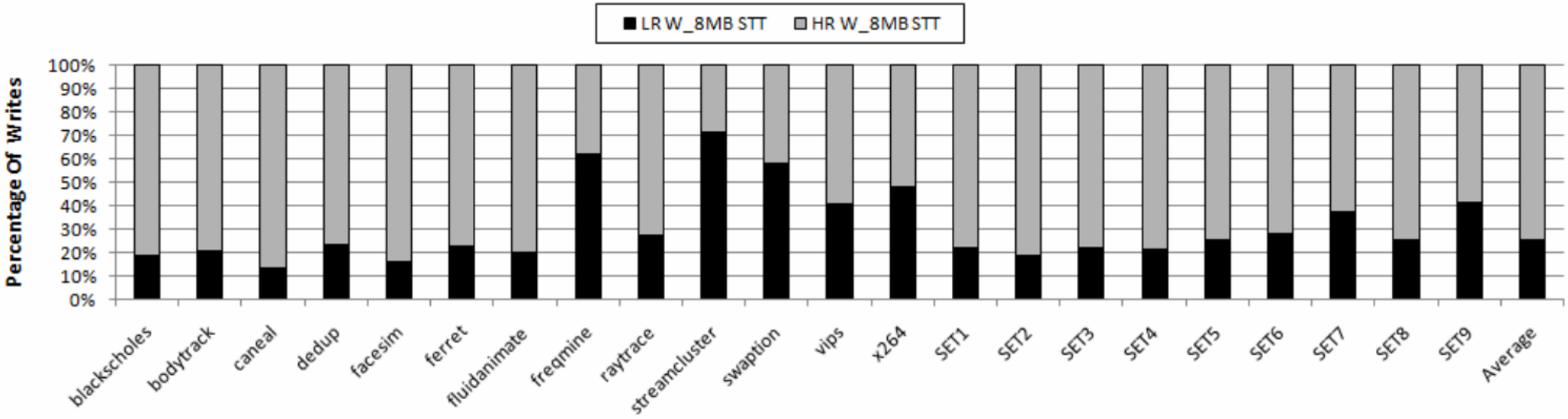
- With write management policies



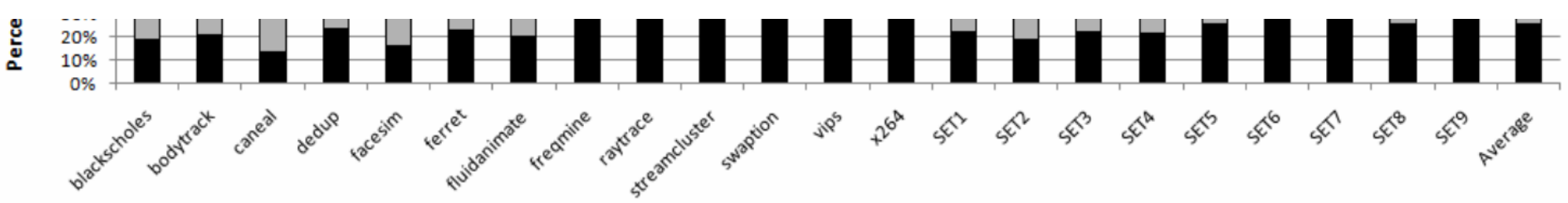
Prezi

Write access distribution

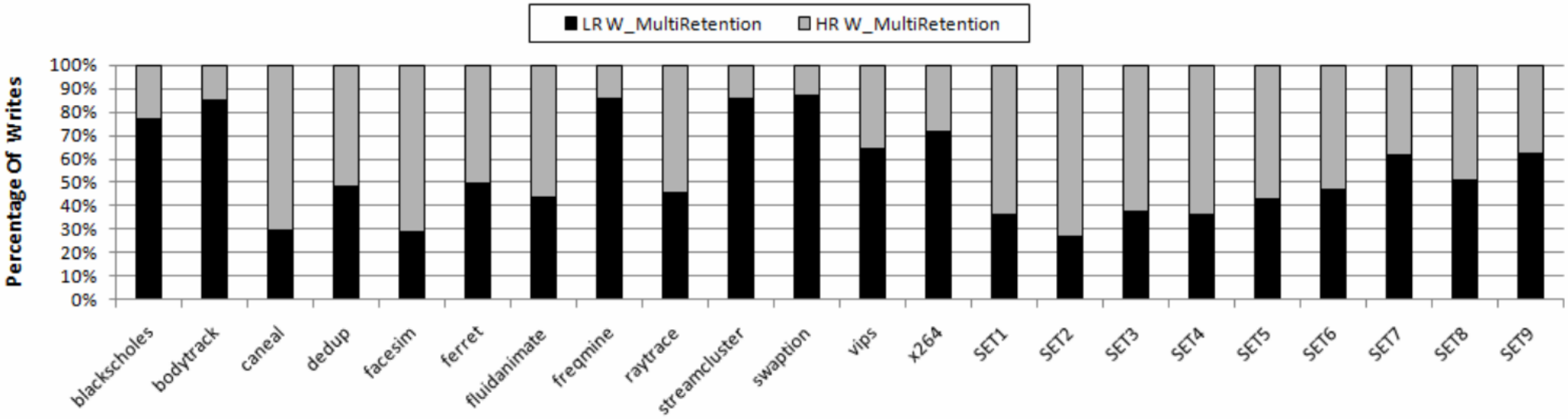
- Without write management policies



- With write management policies



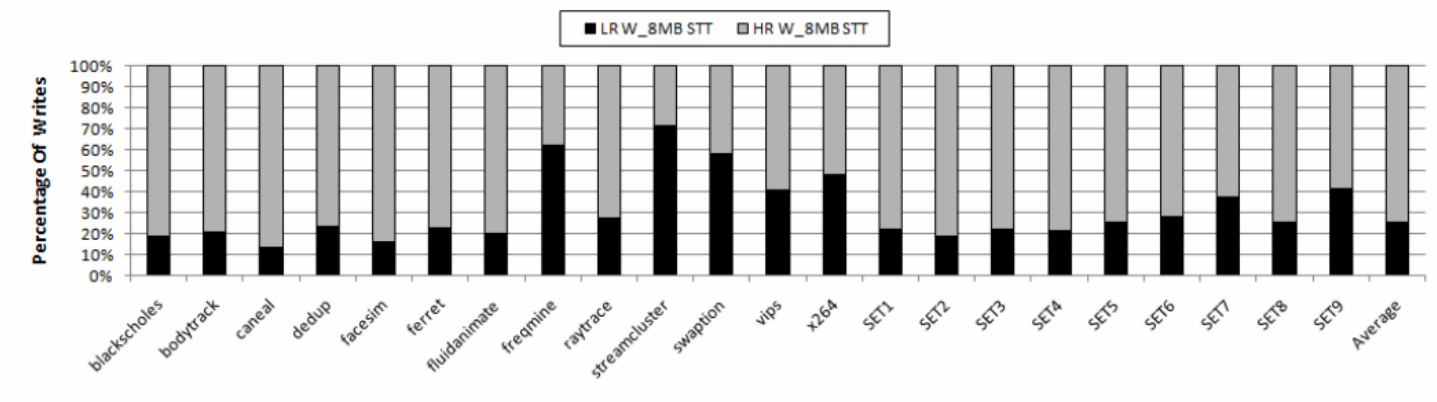
- With write management policies



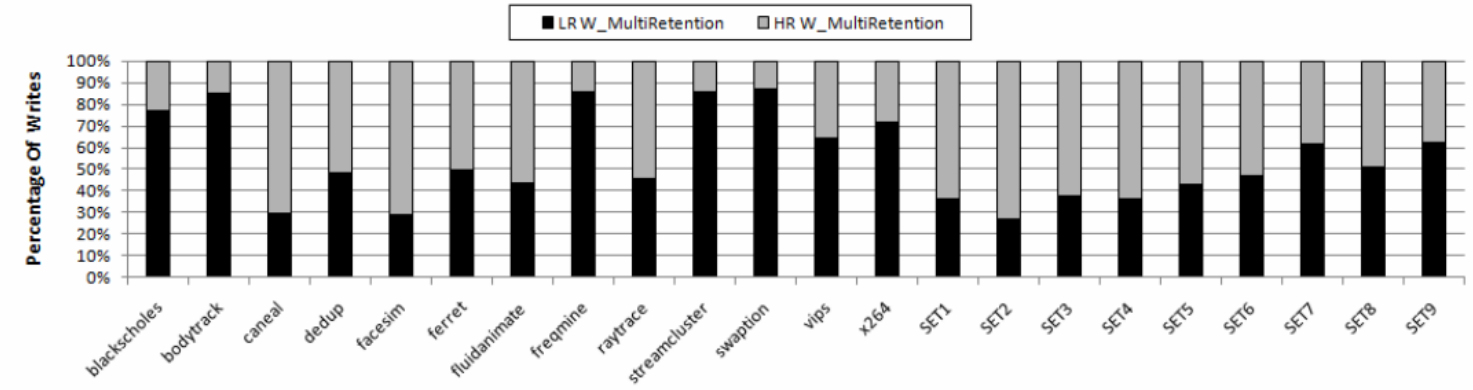
Prezi

Write access distribution

- Without write management policies



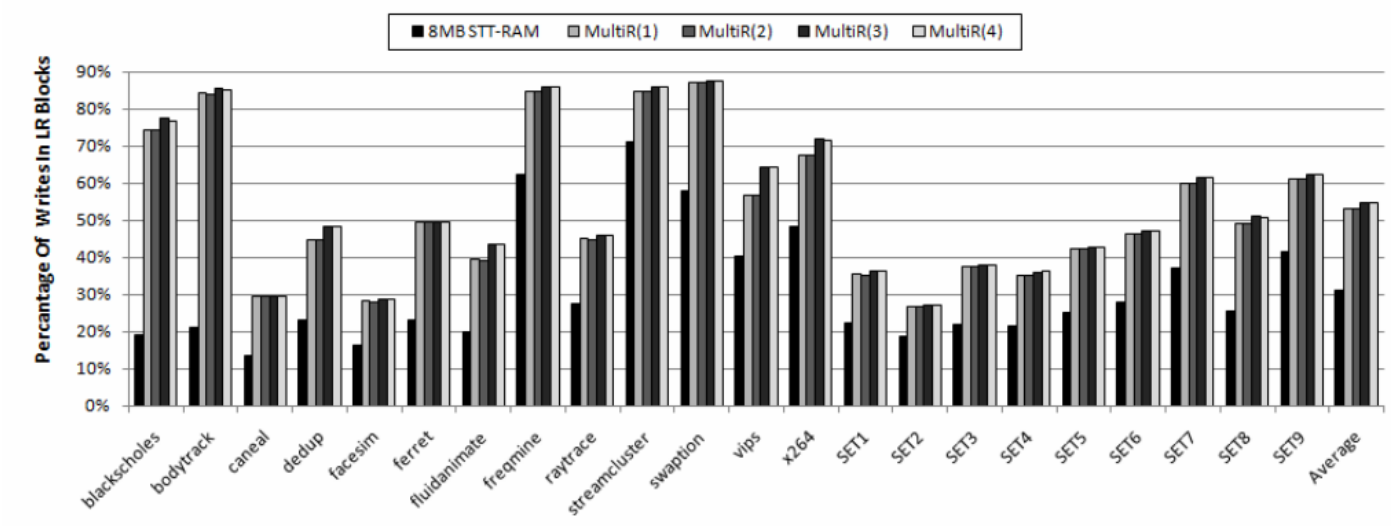
- With write management policies



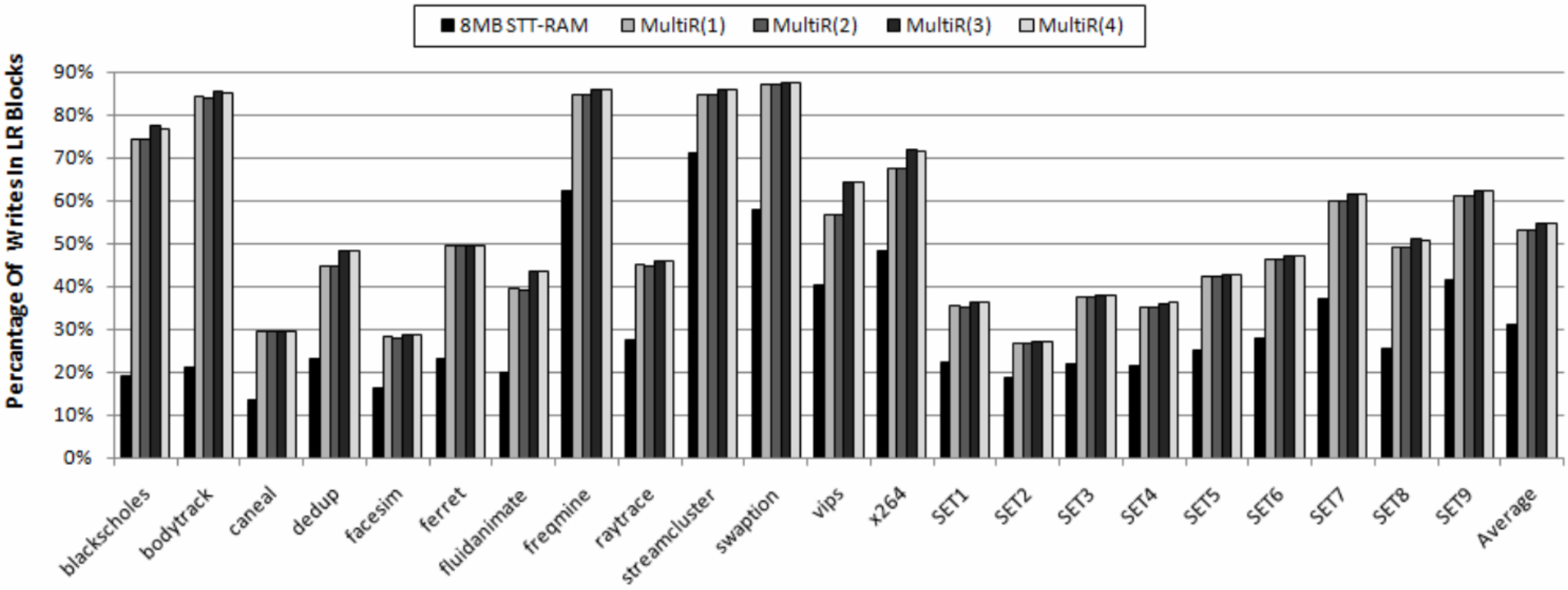
Prezi

Write access distribution

- Preliminary Evaluation
- frequently-written data from high retention STT-RAM array moved to low retention STT-RAM array (more than 55% on average)

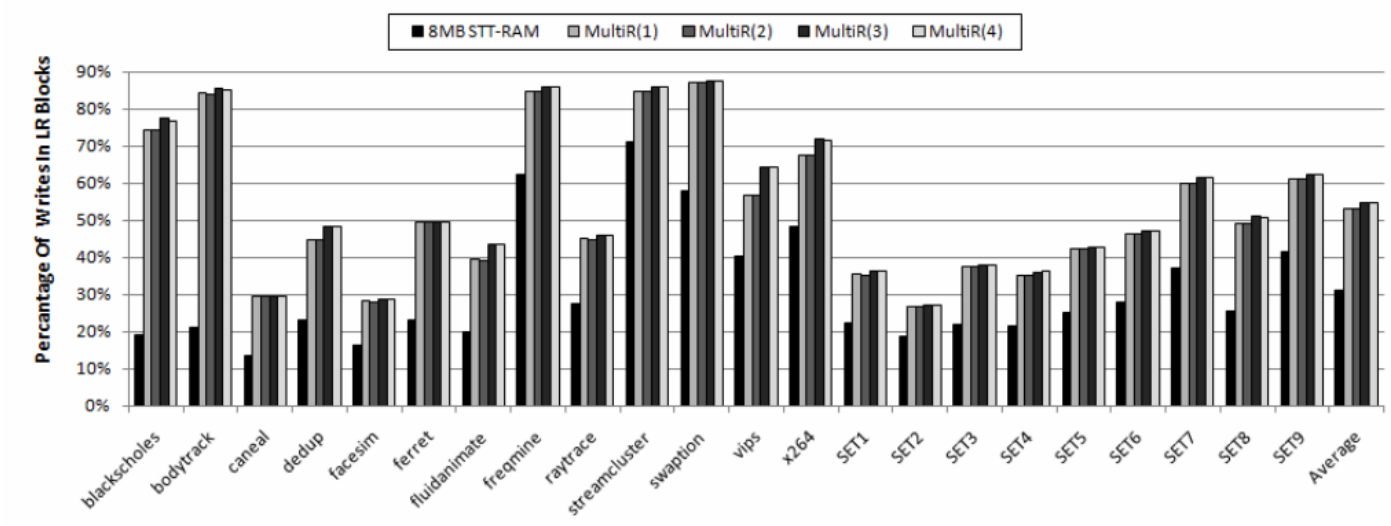


to low retention STT-RAM array (more than 55% on average)



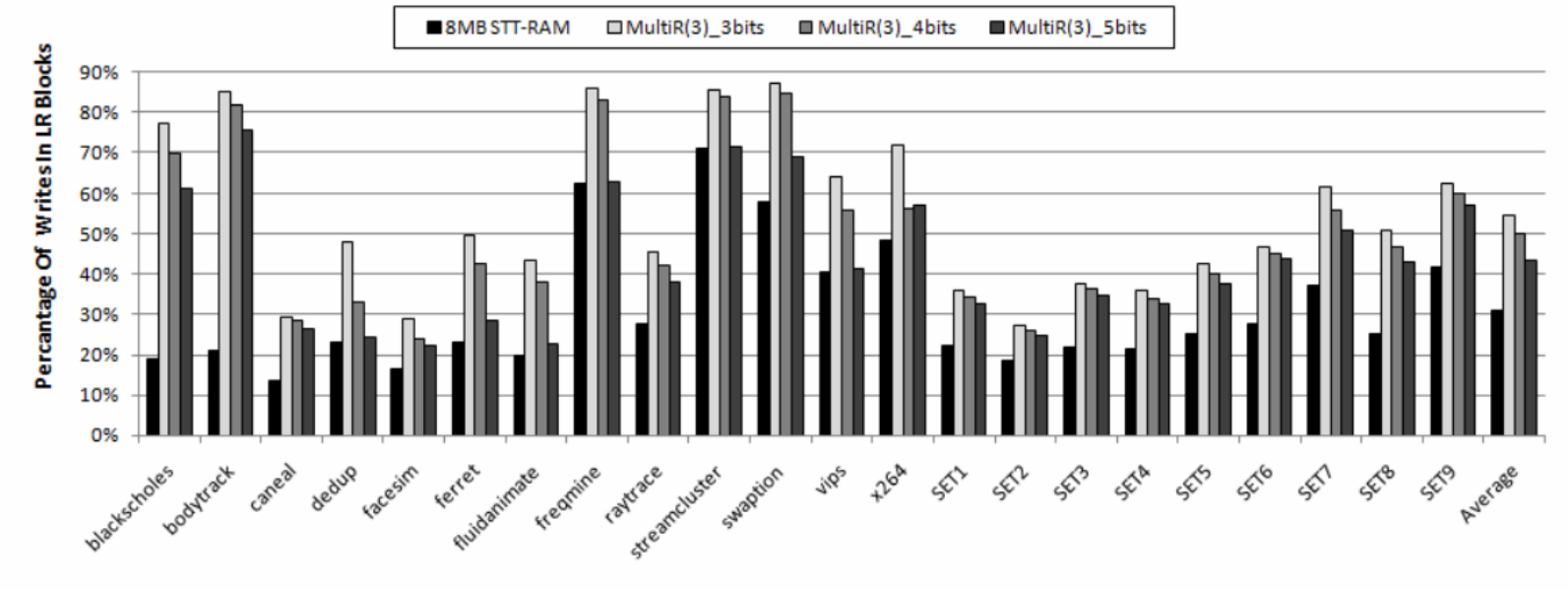
Write access distribution

- Preliminary Evaluation
- frequently-written data from high retention STT-RAM array moved to low retention STT-RAM array (more than 55% on average)



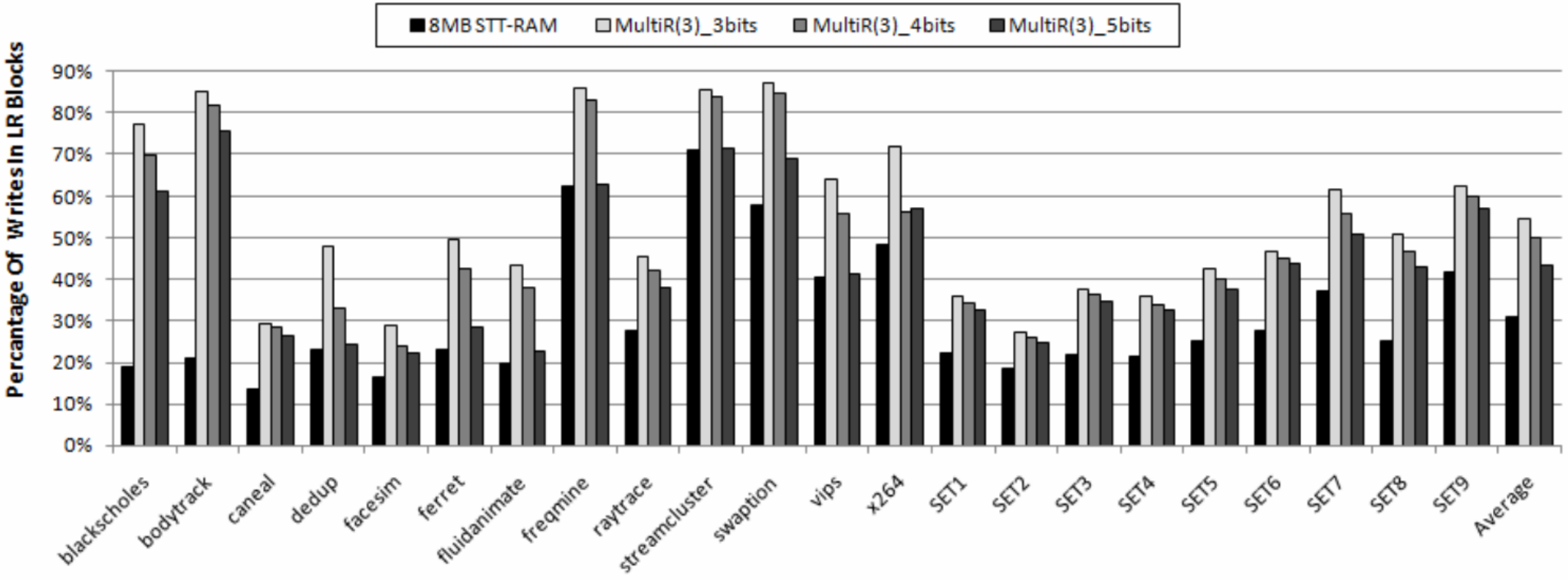
Sensivity Analysis to Saturation Counter Bits

- 3 bit saturation counter



Prezi

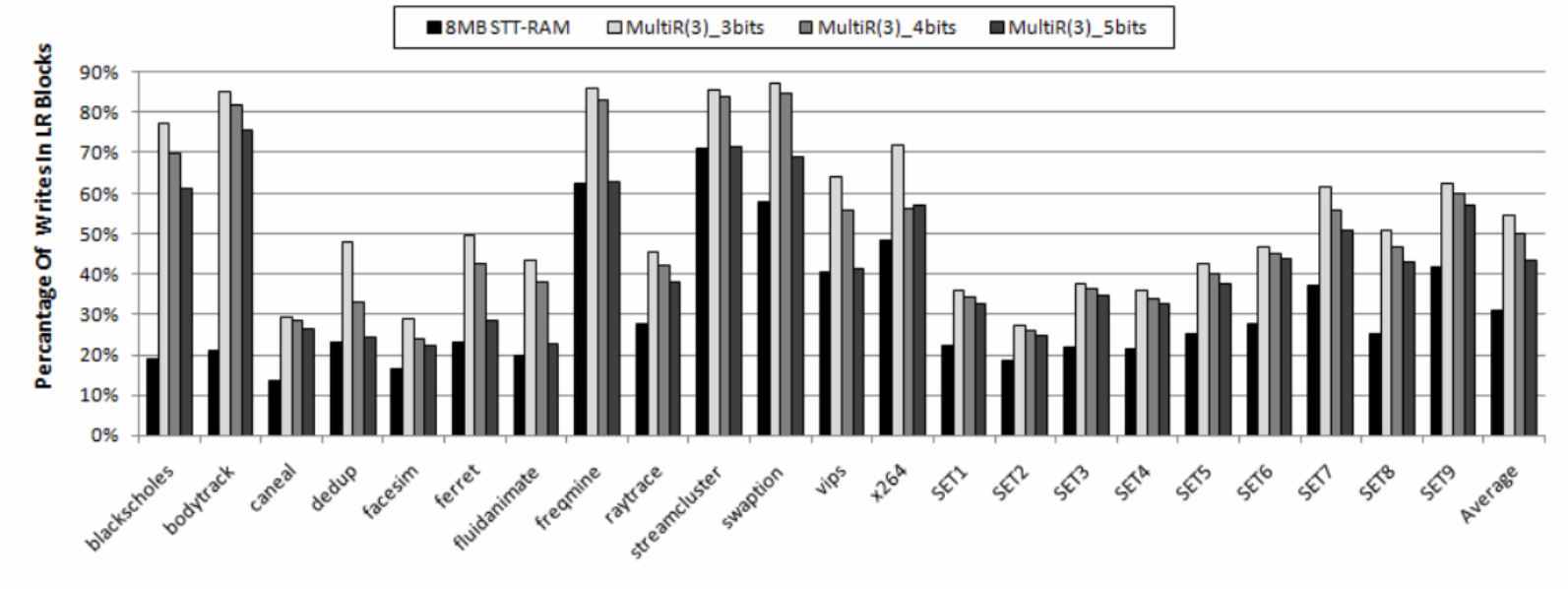
- 3 bit saturation counter



Prezi

Sensivity Analysis to Saturation Counter Bits

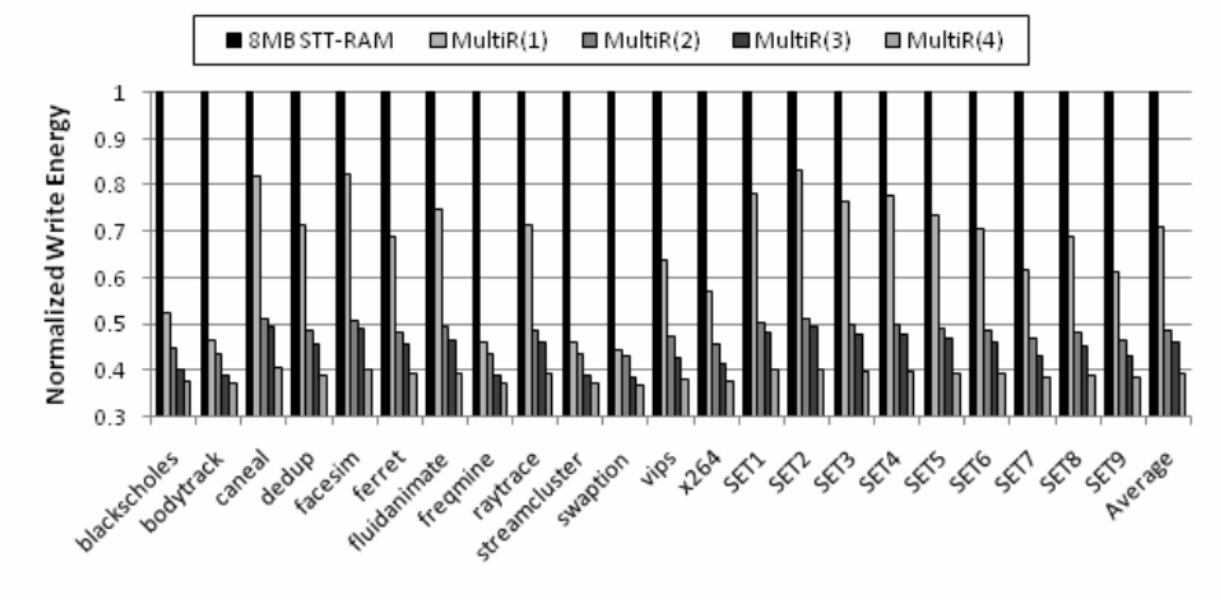
- 3 bit saturation counter



Prezi

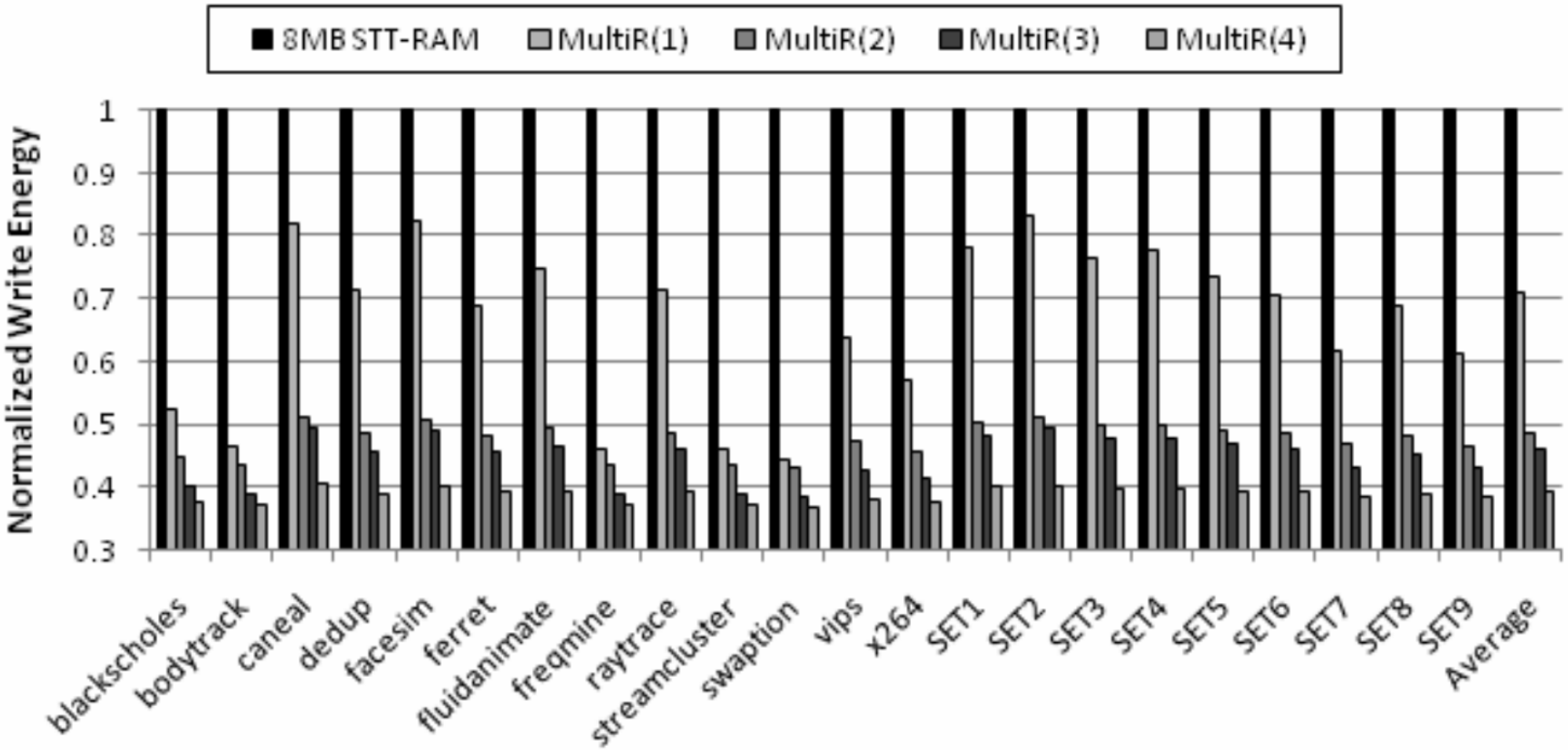
Write Dynamic Energy

- more than 55%



Prezi

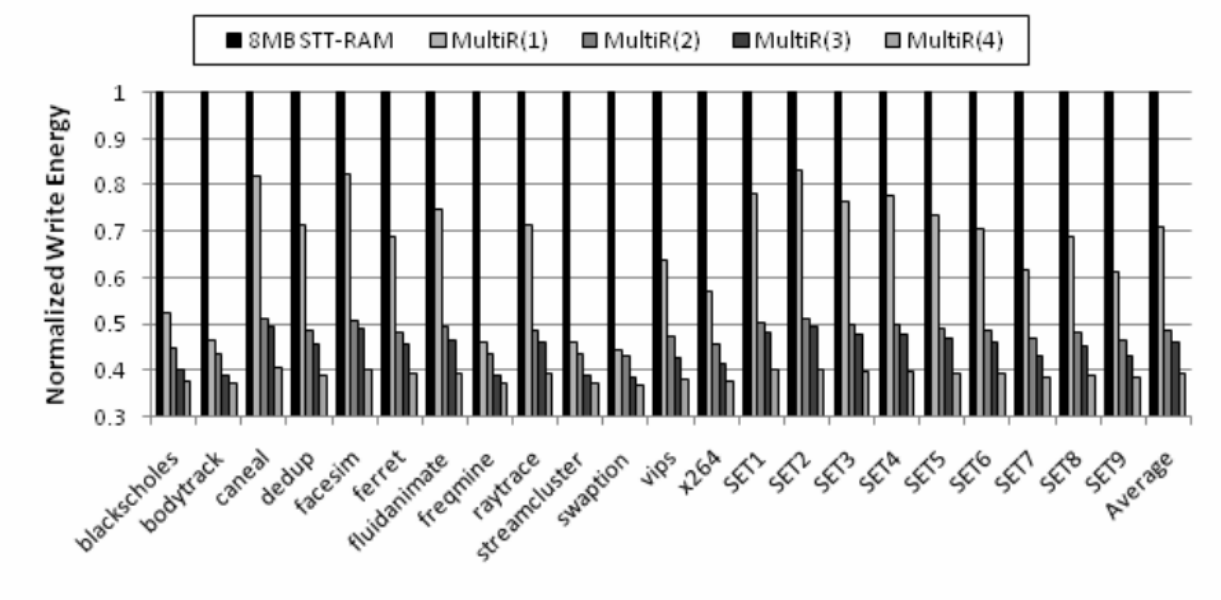
- more than 55%



Prezi

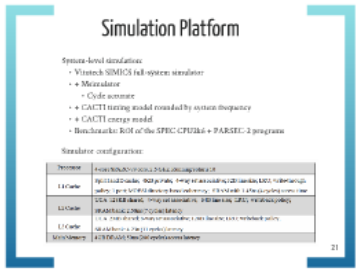
Write Dynamic Energy

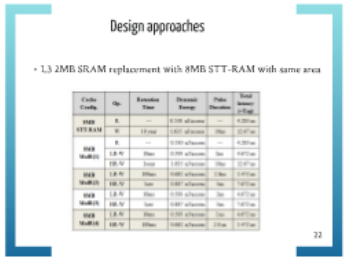
- more than 55%

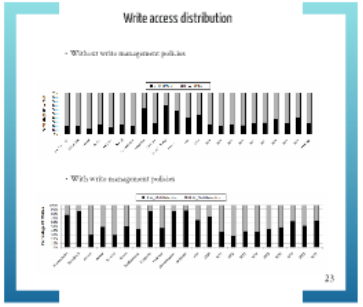


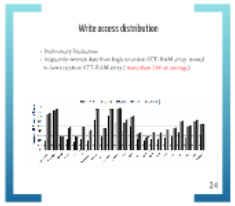
Prezi

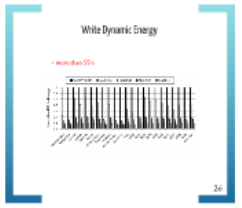
Evaluation









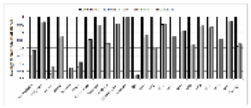


Prezi

Results

Average Memory Access Latency

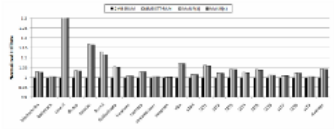
- AML = hit time + miss rate x miss penalty
- AML (20%)



28

Hitrate

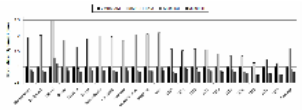
- hitrate (5%)



27

Dynamic Power

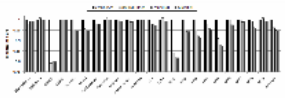
- 50% (STT-RAM)
- 15% (SRAM)



30

Performance

- Performance (5%)



29

Total Power

32

Prezi

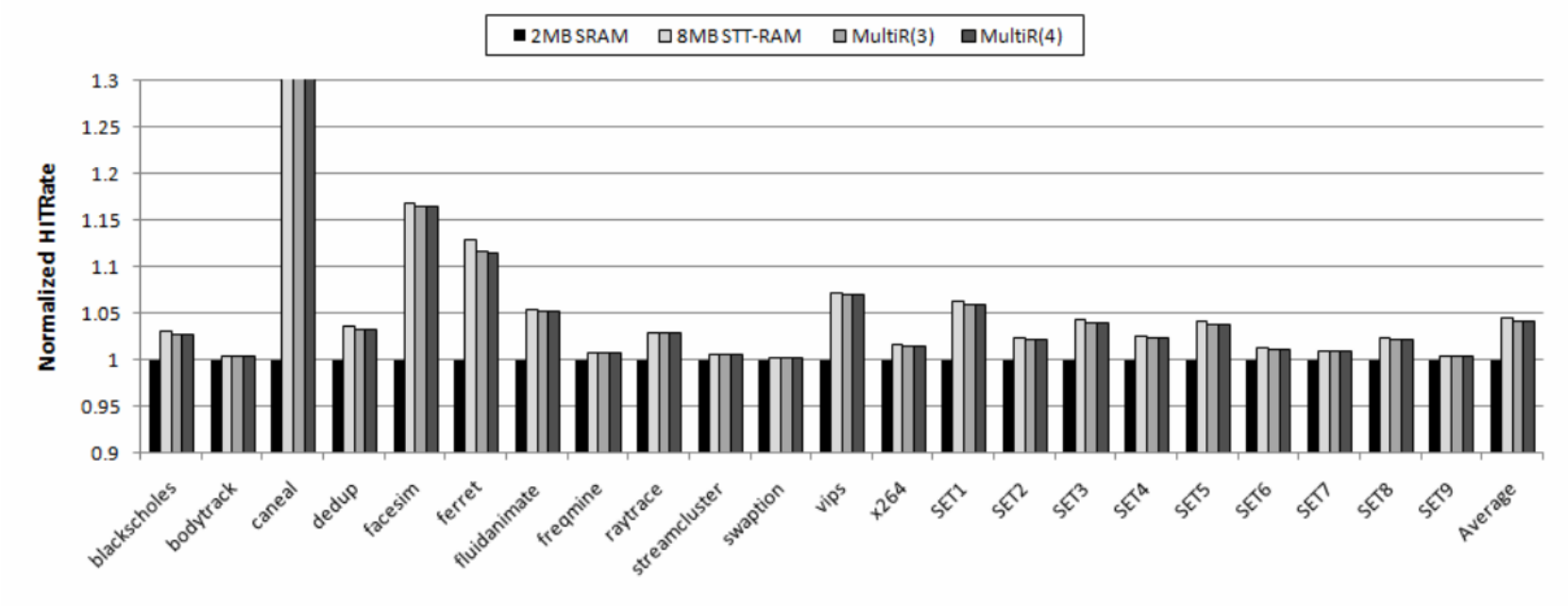
Results

Hitrate

Prezi

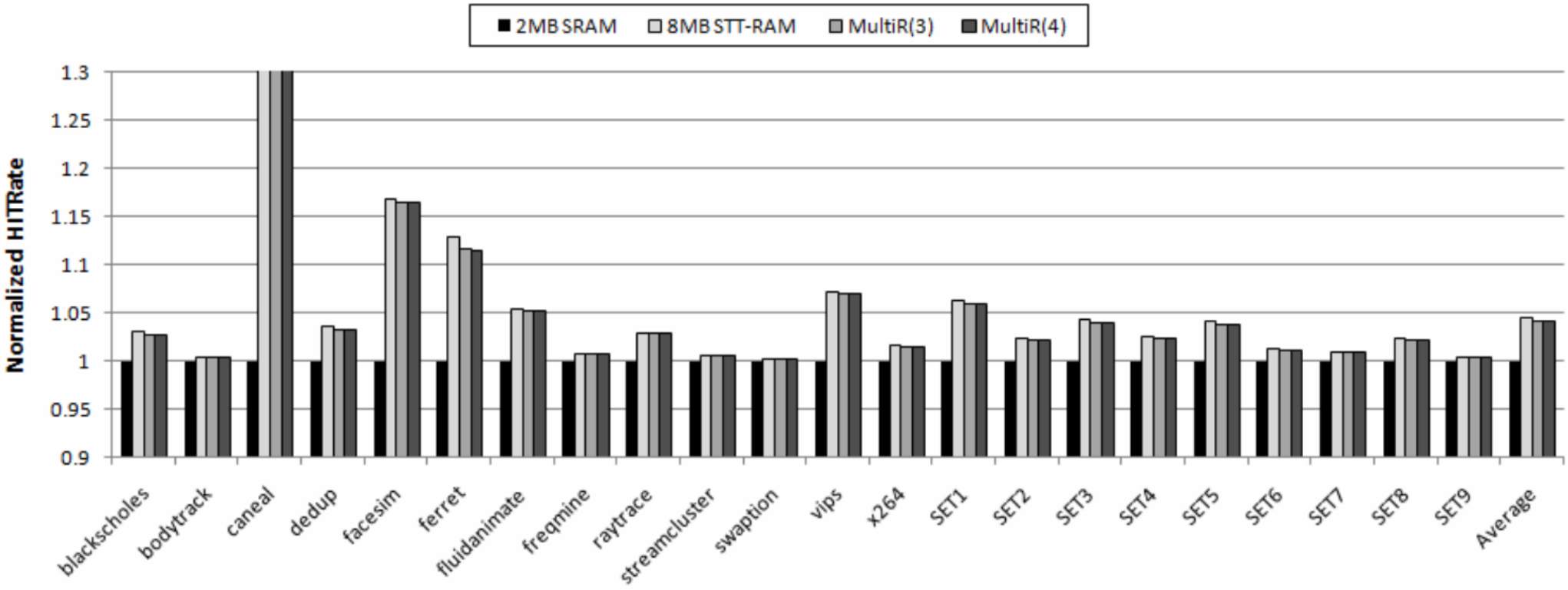
Hitrate

- hitrate (5%)



Prezi

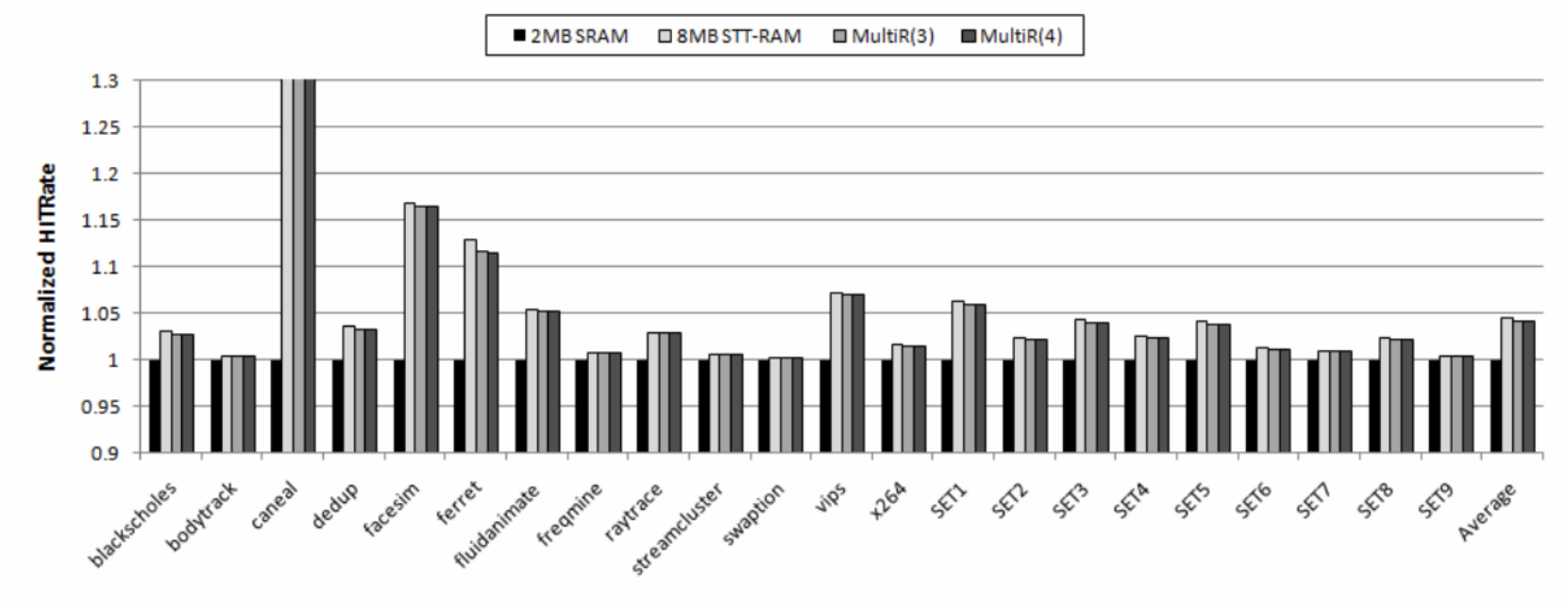
- hitrate (5%)



Prezi

Hitrate

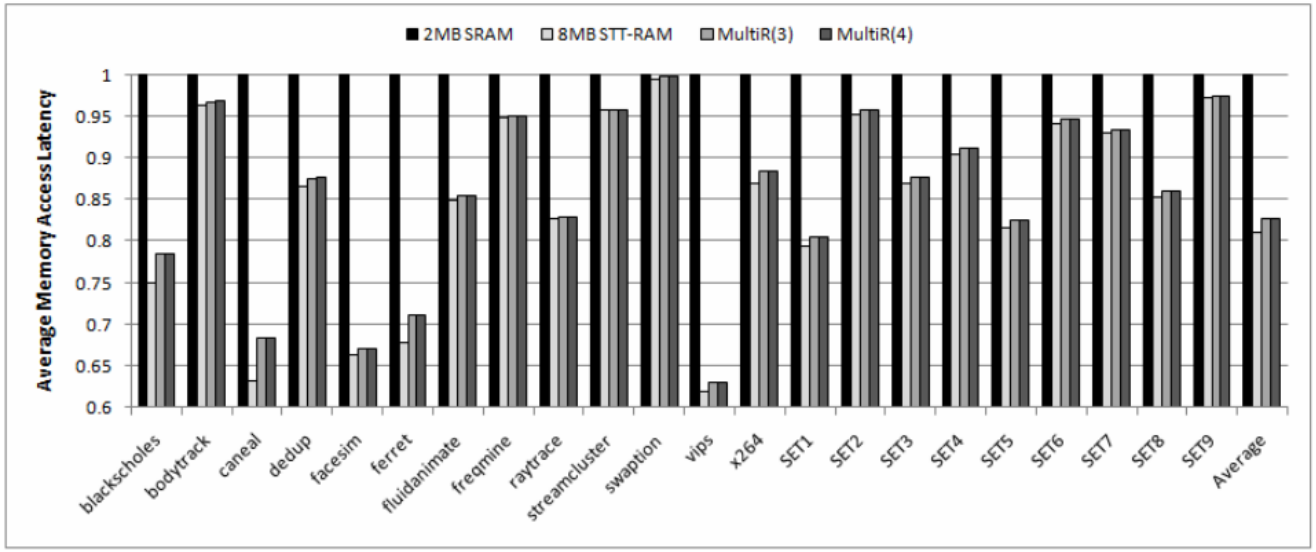
- hitrate (5%)



Prezi

Average Memory Access Latency

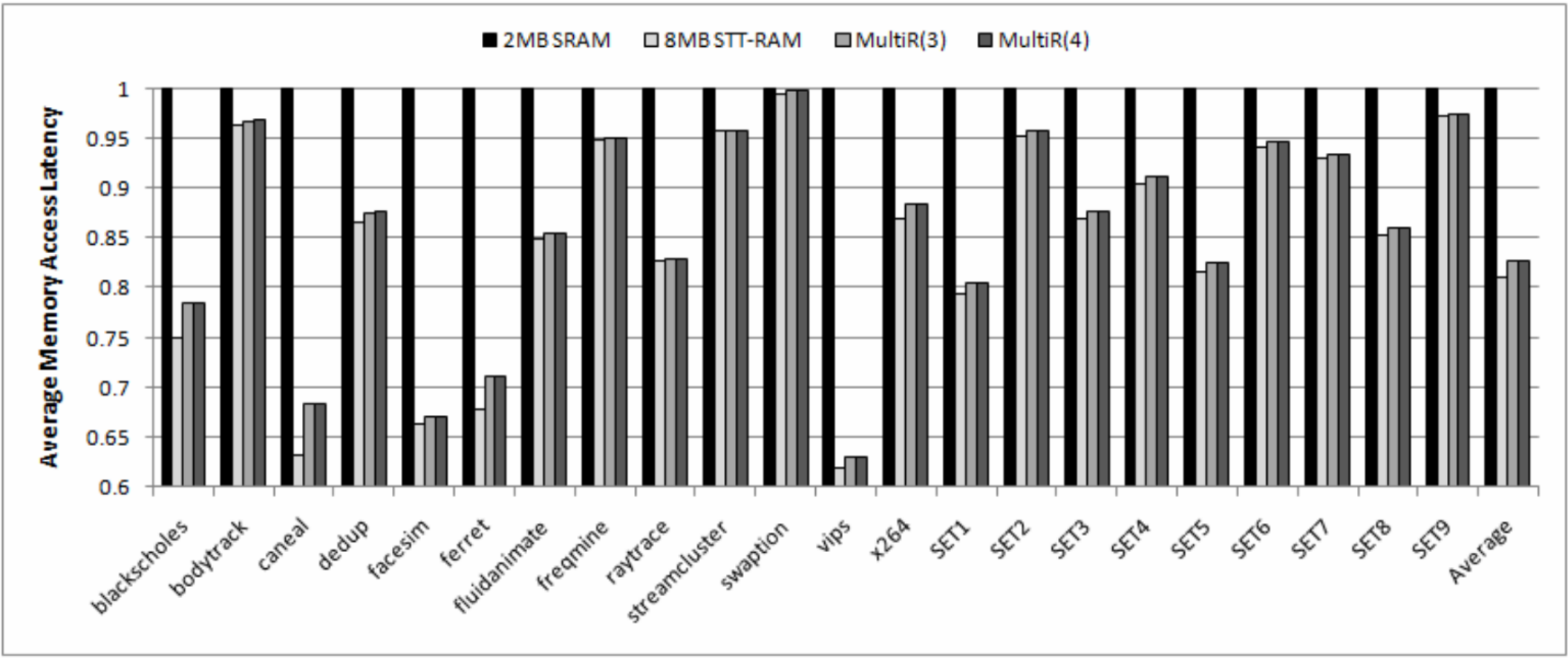
- AML = hit time + miss rate x miss penalty
- AML (20%)



Prezi

- AML = hit time + miss rate x miss penalty

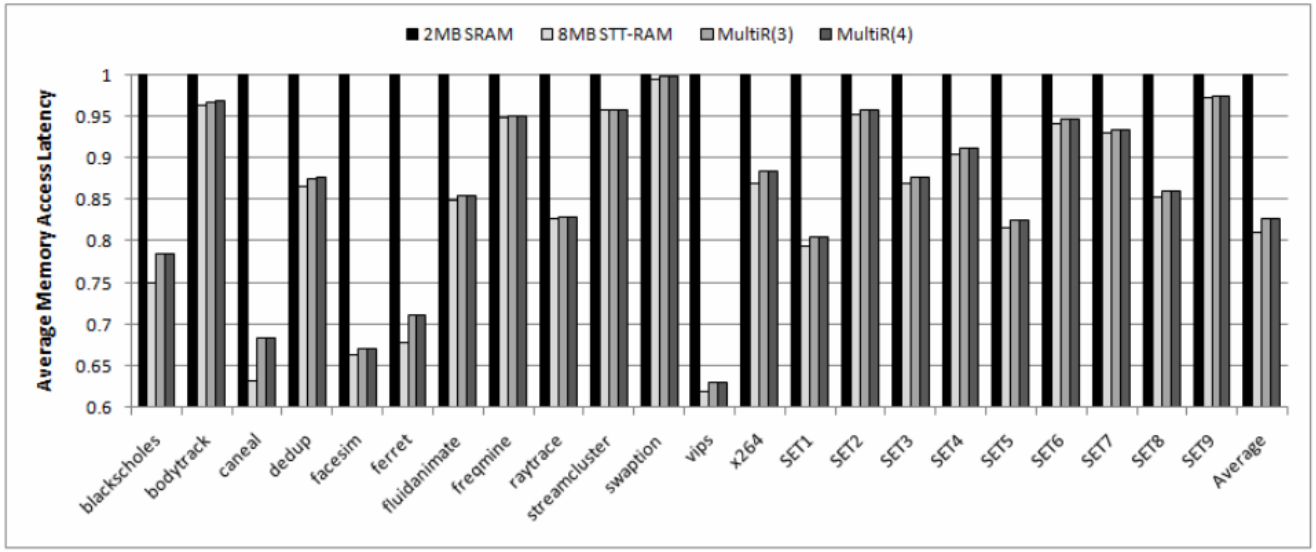
- AML (20%)



Prezi

Average Memory Access Latency

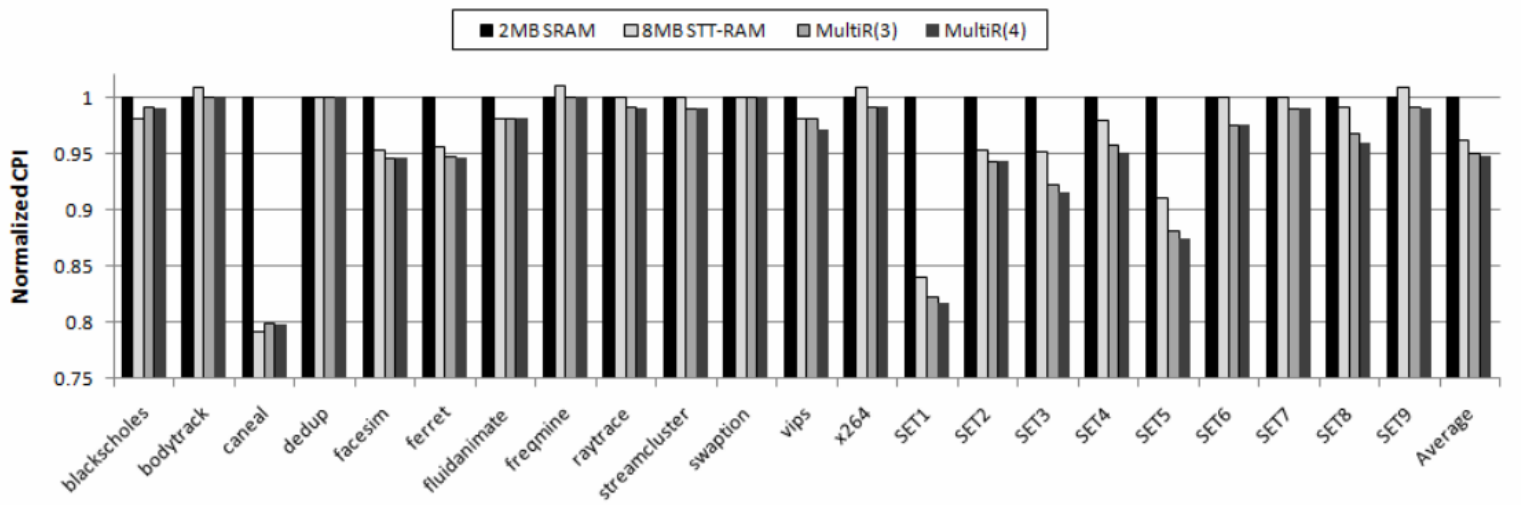
- AML = hit time + miss rate x miss penalty
- AML (20%)



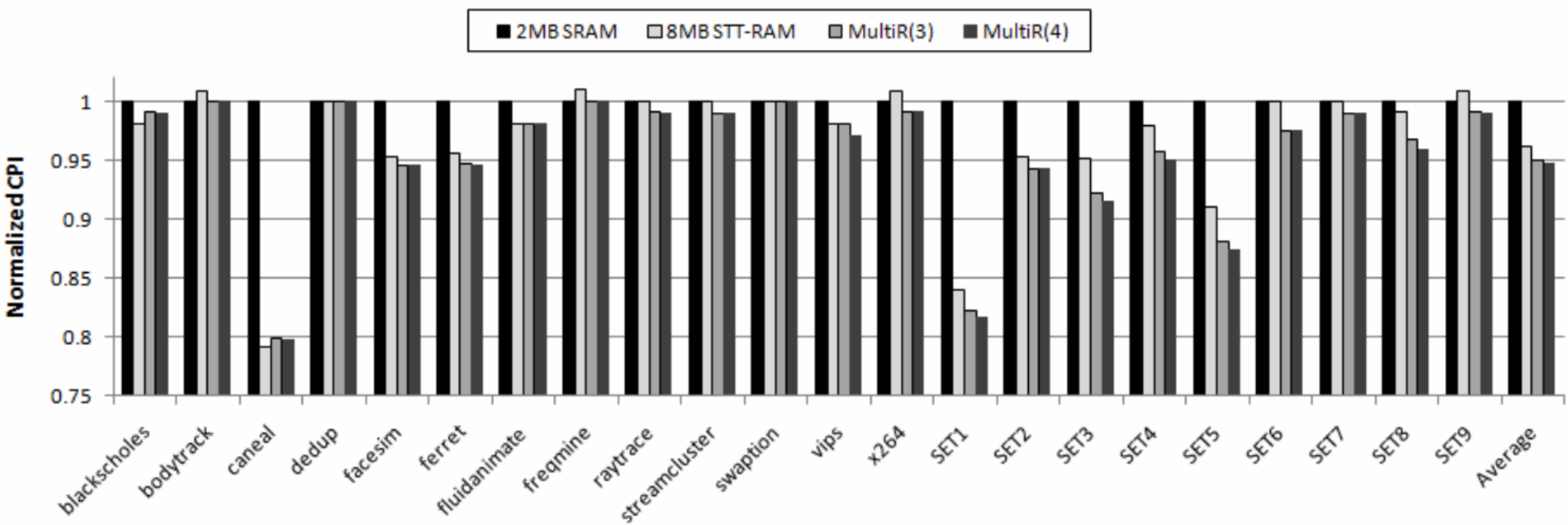
Prezi

Performance

- Performance (5%)



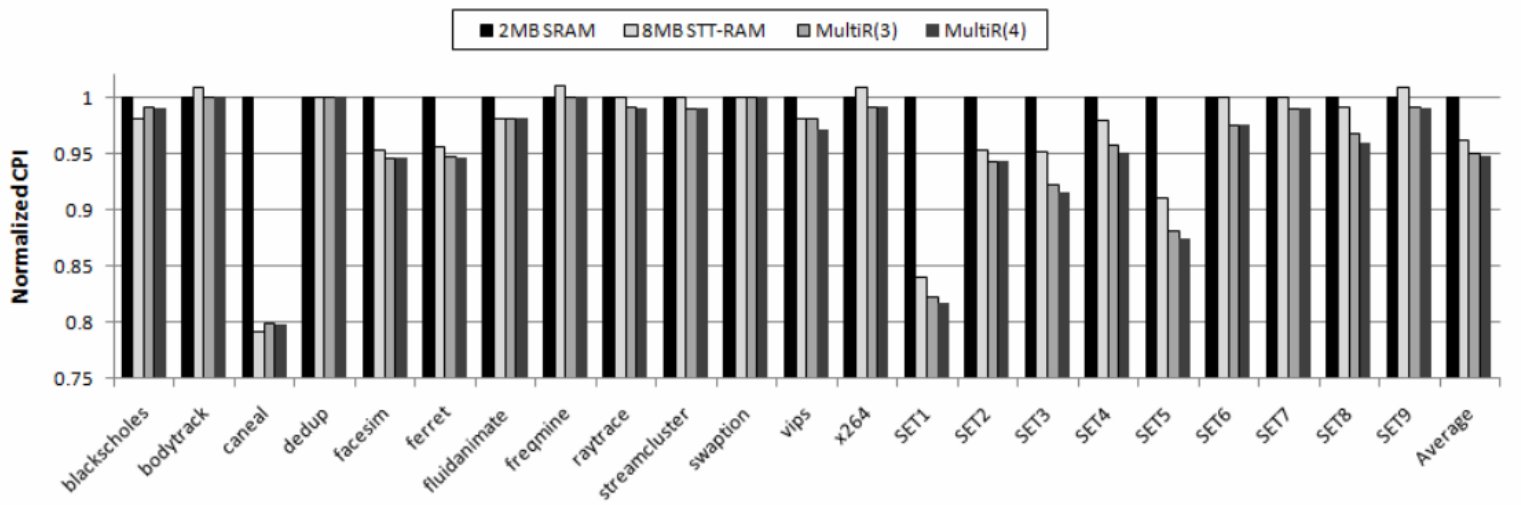
- Performance (5%)



Prezi

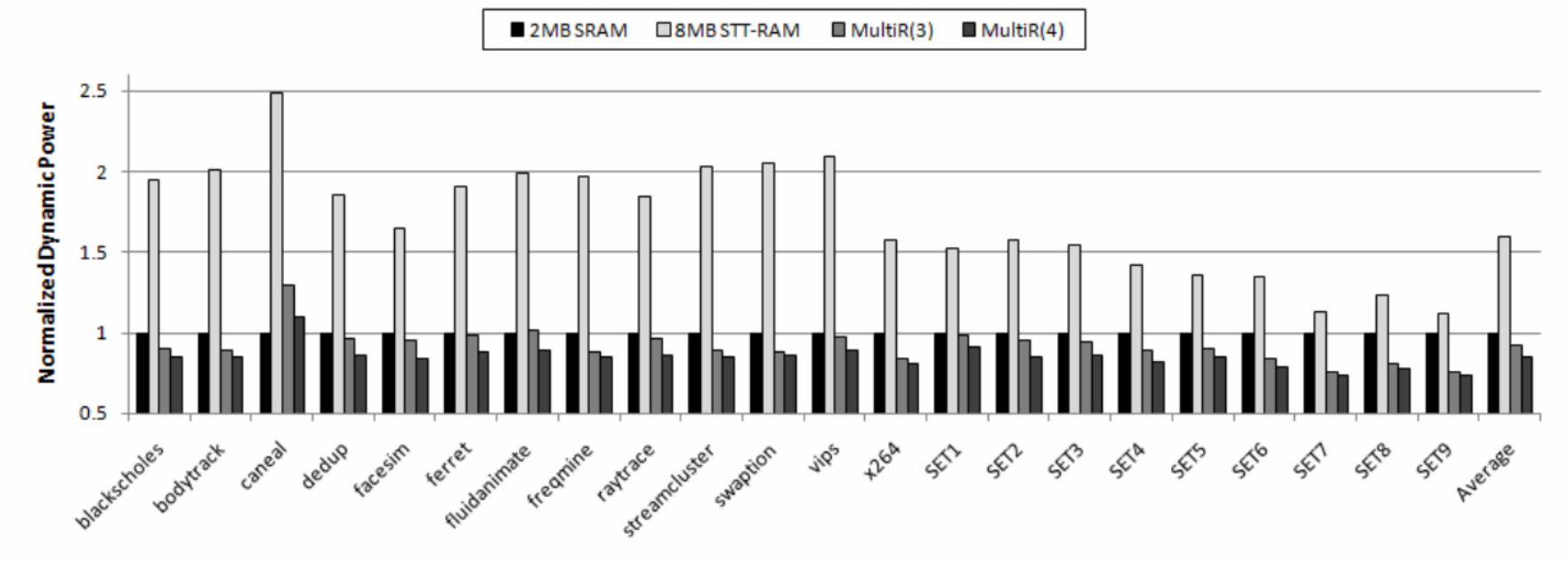
Performance

- Performance (5%)



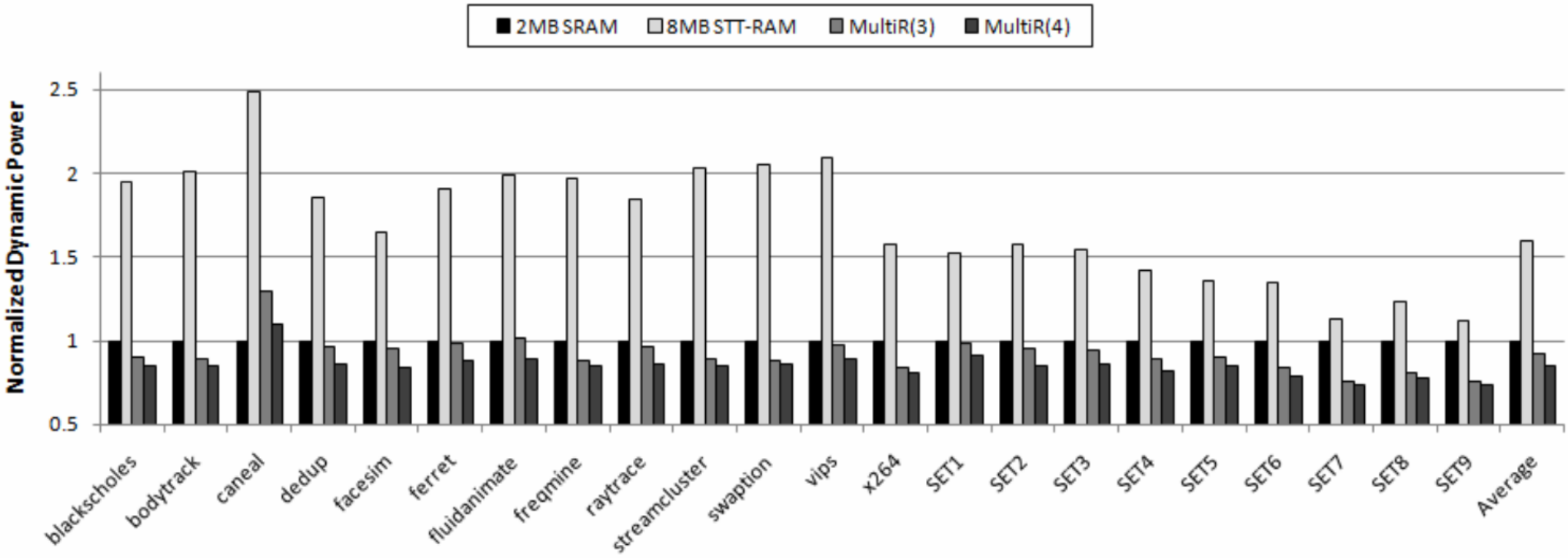
Dynamic Power

- 50% (STT-RAM)
- 15% (SRAM)



Prezi

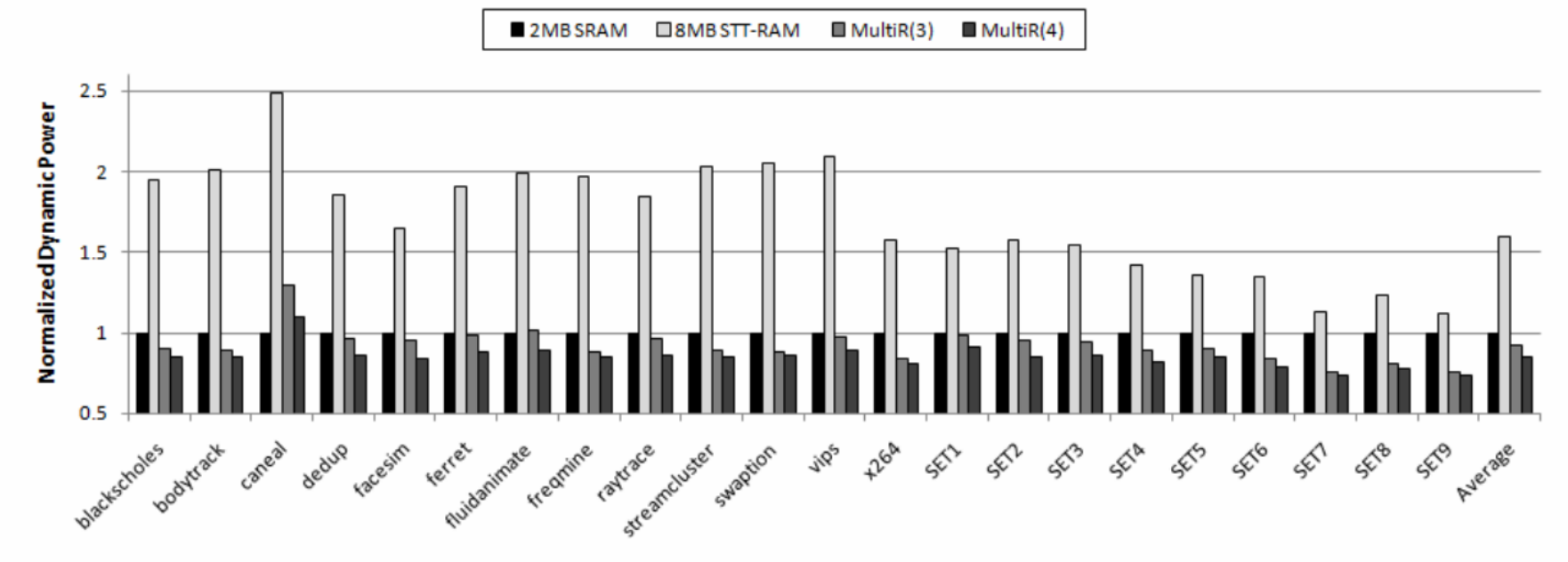
- 50% (STT-RAM)
- 15% (SRAM)



Prezi

Dynamic Power

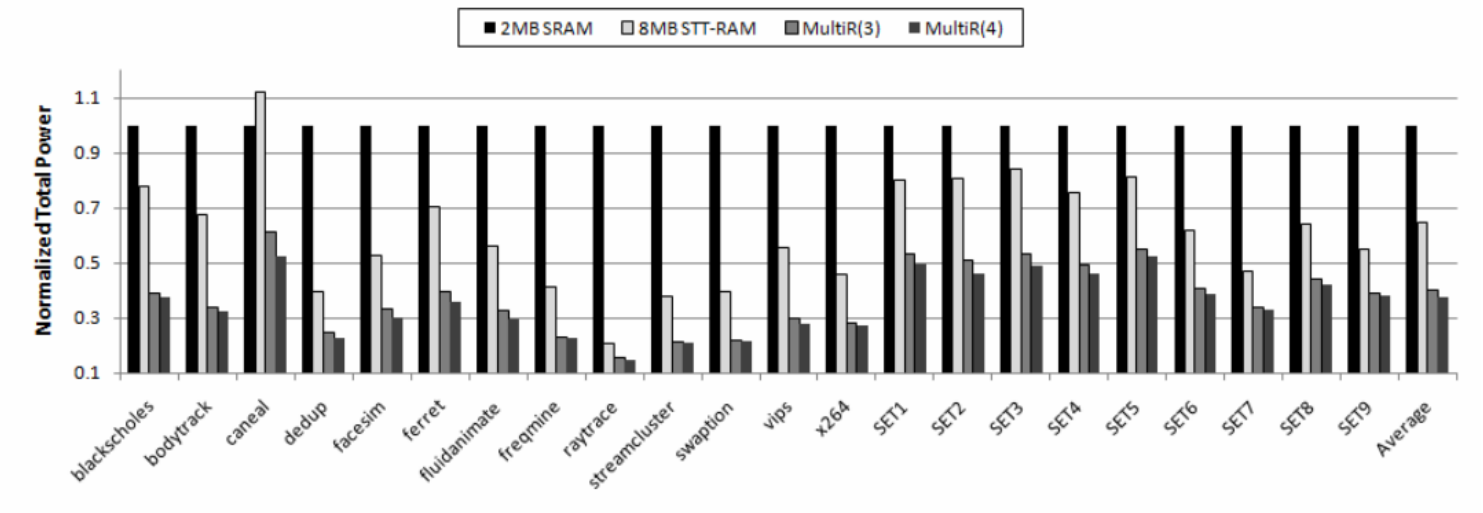
- 50% (STT-RAM)
- 15% (SRAM)



Prezi

Total Power

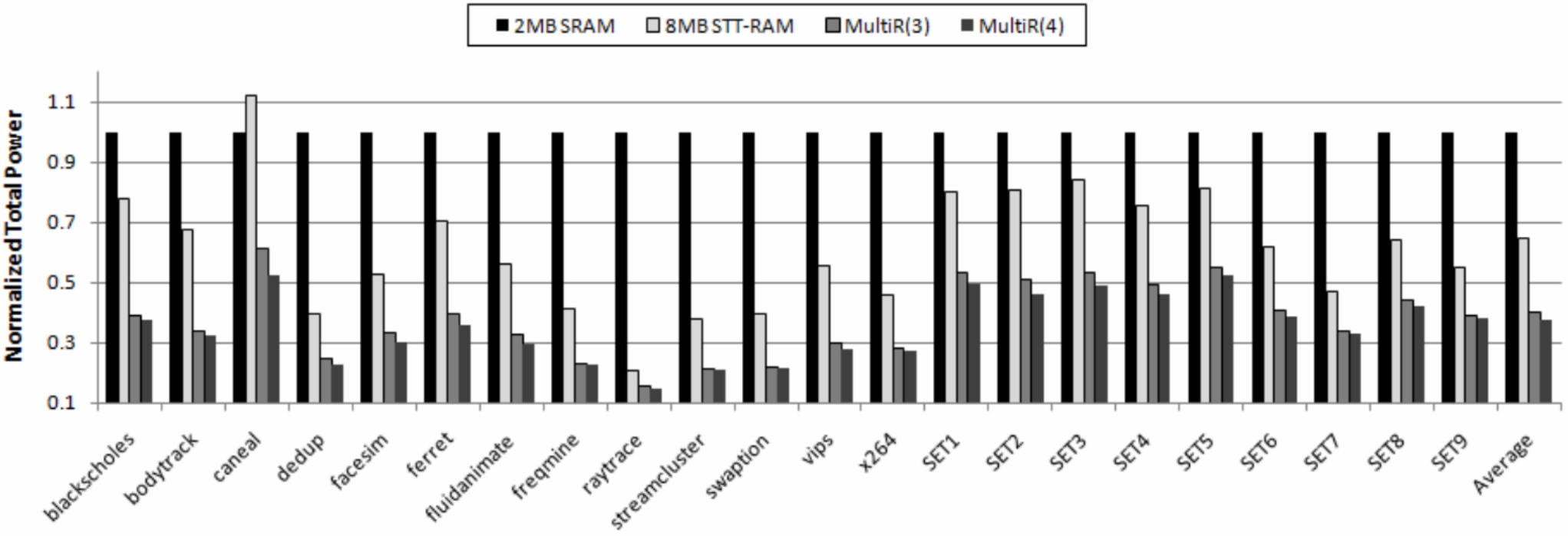
- parsec (74%)
- spec (60%)



Prezi

Total Power

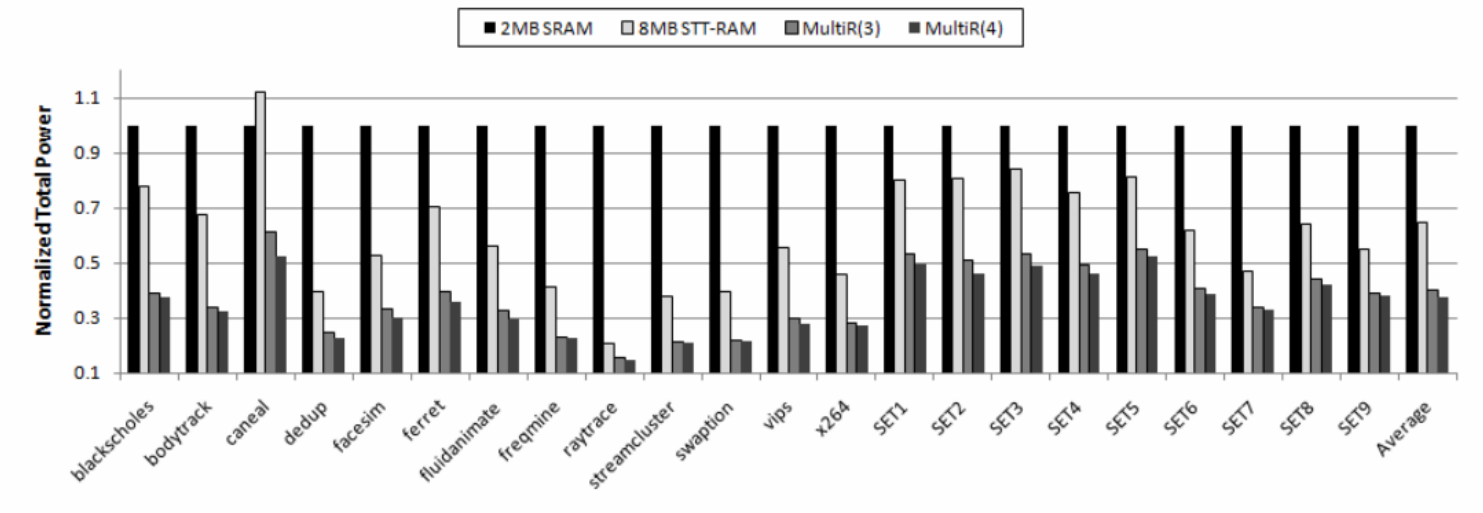
- parsec (74%)
- spec (60%)



Prezi

Total Power

- parsec (74%)
- spec (60%)

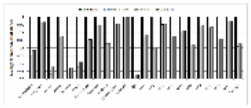


Prezi

Results

Average Memory Access Latency

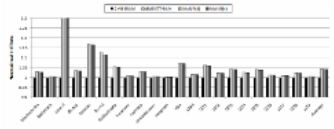
- AML = hit time + miss rate x miss penalty
- AML (20%)



28

Hitrate

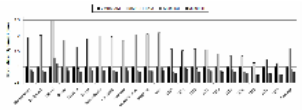
- hitrate (5%)



27

Dynamic Power

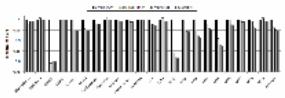
- 50% (STT-RAM)
- 15% (SRAM)



30

Performance

- Performance (5%)



29

Total Power

32

Prezi

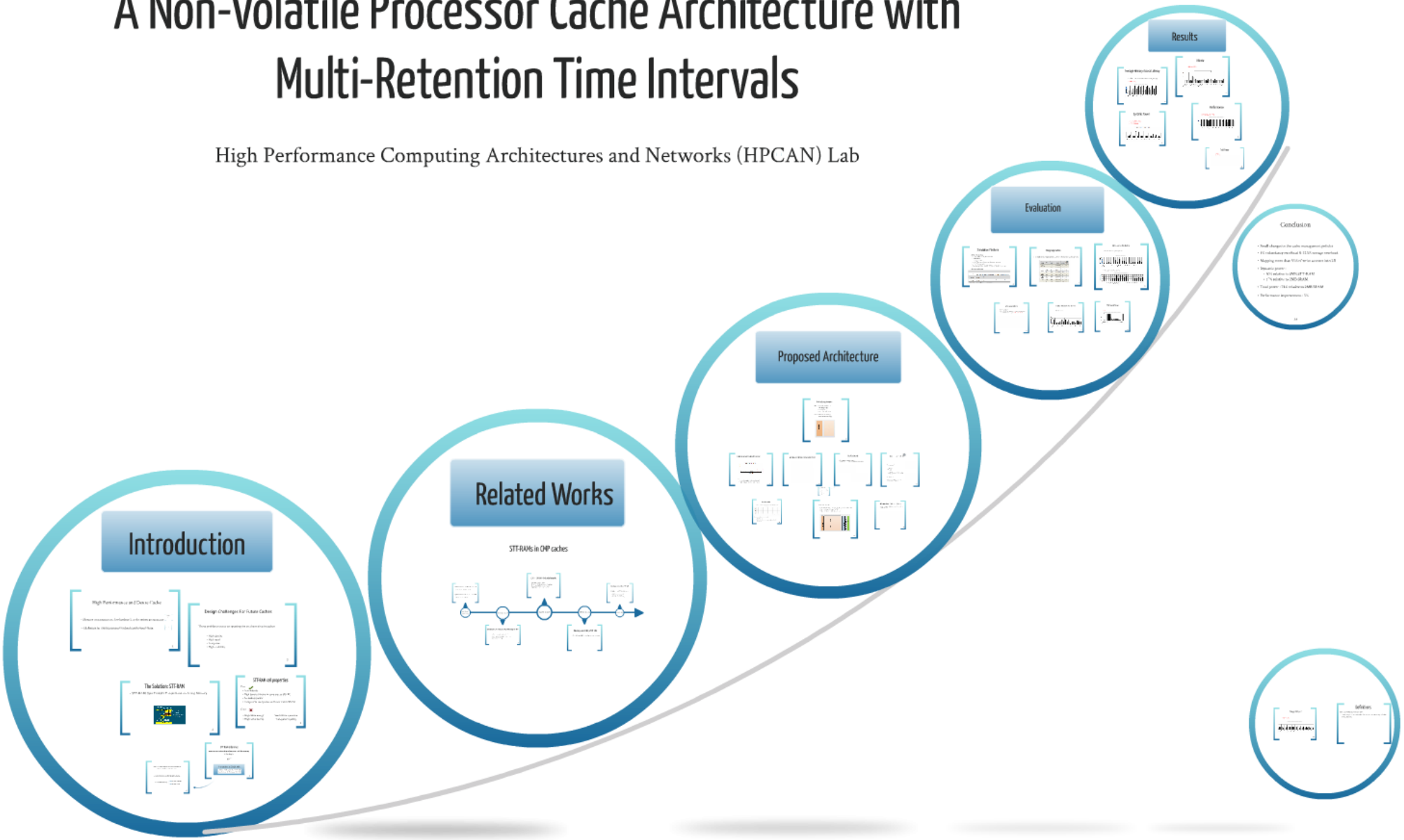
Conclusion

- Small changes to the cache management policies
- 1% redundancy overhead & 12.5% storage overhead
- Mapping more than 55% of write accesses into LR
- Dynamic power :
  - 50% relative to 8MB STT-RAM
  - 17% relative to 2MB SRAM
- Total power : 74% relative to 2MB SRAM
- Performance improvement : 5%

Prezi

A Non-Volatile Processor Cache Architecture with Multi-Retention Time Intervals

High Performance Computing Architectures and Networks (HPCAN) Lab



Prezi

Future works

- Improve utilization of LR-region
  - Increase associativity of LR-region
- Different LR & HR array capacity
  - Increase LR ways in each set
  - Saving more write energy
- Different Refreshing mechanism for LR blocks
- Inter-set redirection:
  - increase number of hot blocks redirected to LR of destination set
- Considering endurance of STT-RAM LLC :
  - uniform distribution of write accesses

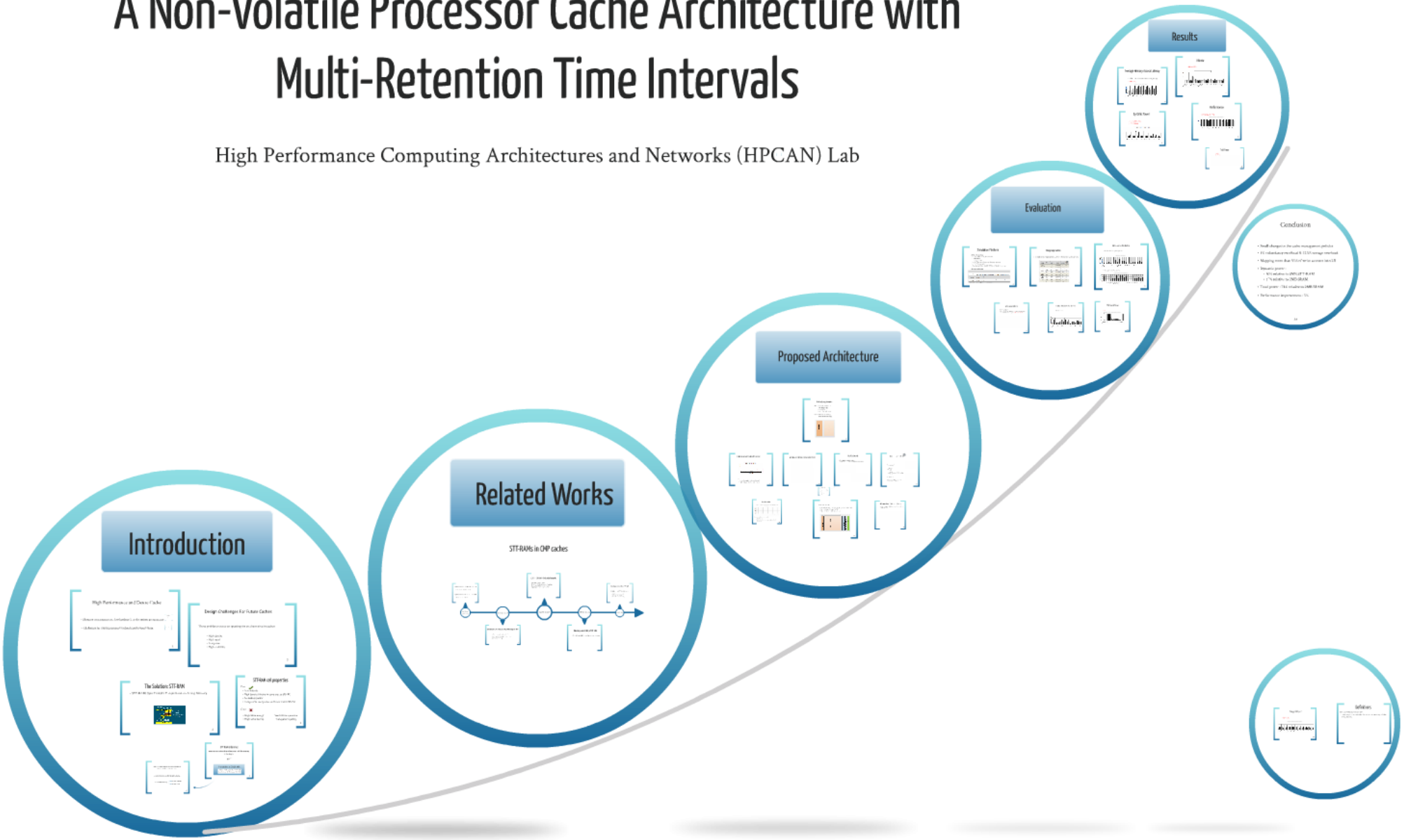
Prezi

Thanks for your attention

Prezi

A Non-Volatile Processor Cache Architecture with Multi-Retention Time Intervals

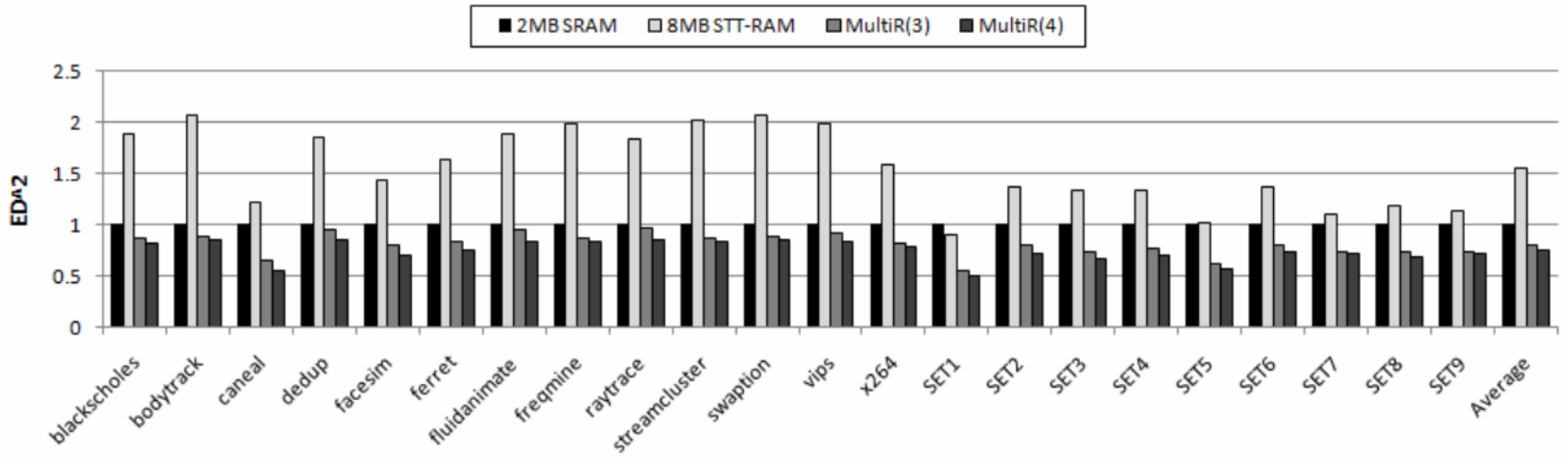
High Performance Computing Architectures and Networks (HPCAN) Lab



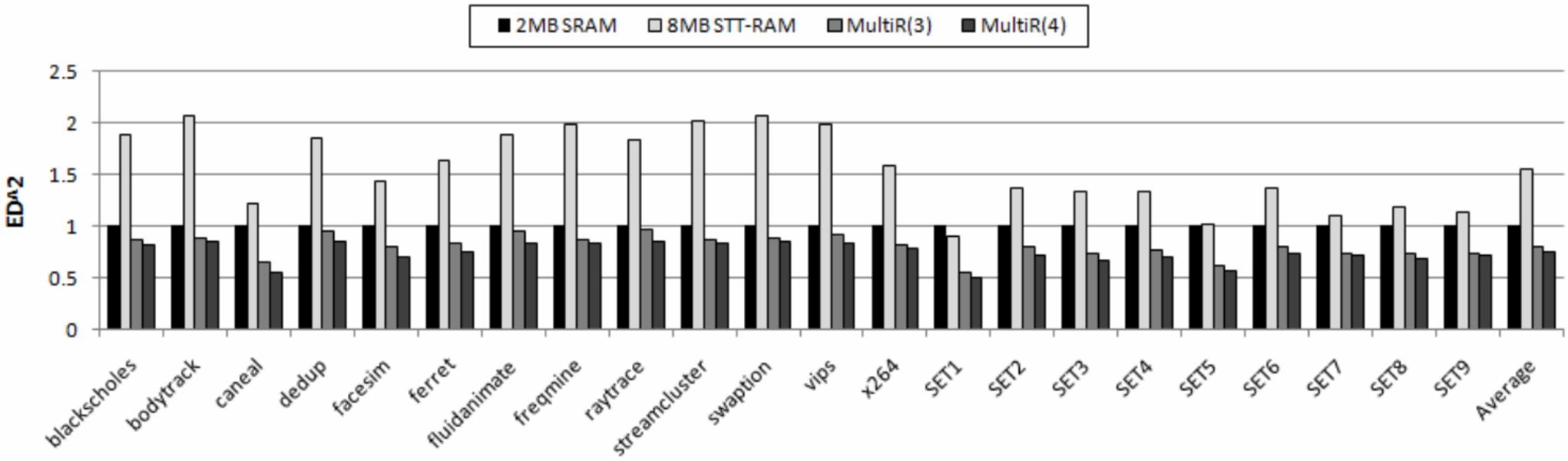
Prezi

Energy x Delay~2

- ED^2 (25%)



- ED^2 (25%)


2MB SRAM
8MB STT-RAM
MultiR(3)
MultiR(4)
ED^2
2.5
2
1.5
1
0.5
0
blackscholes
bodytrack
caneal
dedup
facesim
ferret
fluidanimate
freqmine
raytrace
streamcluster
swaption
vips
x264
SET1
SET2
SET3
SET4
SET5
SET6
SET7
SET8
SET9
Average

Prezi