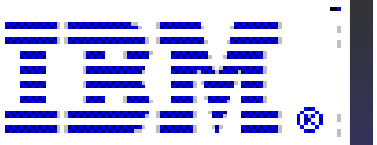


Hybrid Cache Architecture (HCA) with Disparate Memory Technologies

Xiaoxia Wu[†], Jian Li[‡], Lixin Zhang[‡],

Evan Speight[‡], Ram Rajamony[‡], Yuan Xie[†]

[†] Pennsylvania State University

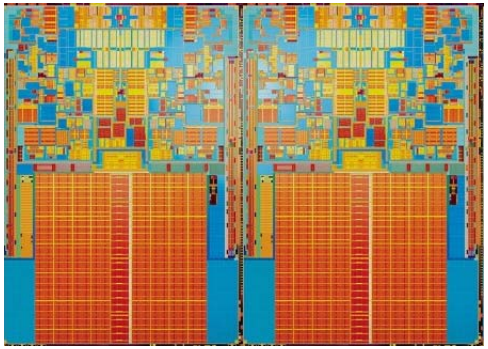
[‡] IBM Austin Research Laboratory

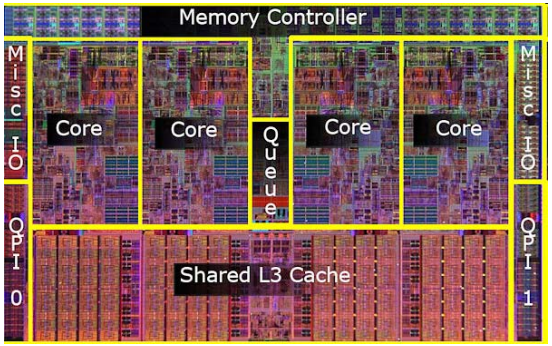
Acknowledgement: Elmootazbellah (Mootaz) Elnozahy, Hung Le, Balaram Sinharoy, William (Bill) J. Starke, and Chung-Lung Kevin Shum

Outline

- Motivation and Introduction
- Methodology
- Level based Hybrid Cache Architecture
- Region based Hybrid Cache Architecture
- 3D Hybrid Cache Stacking
- Conclusion

Introduction





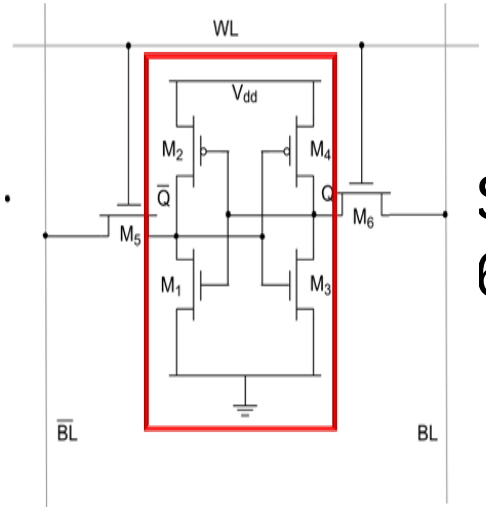
Traditional SRAM-based Cache Architecture

- Limited size with CMP: cache-core balance
- Leakage power
- More cache levels: design overhead, coherence
- Non-Uniform Cache Architecture (wire delay)

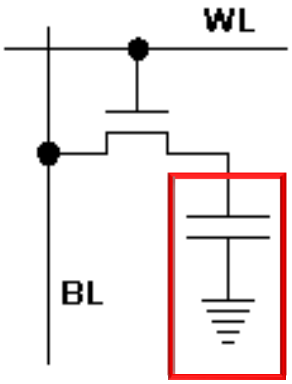
Improve cache power-performance with Emerging Memory Technologies, under the same chip area/footprint

- Embedded DRAM
- Magnetic RAM
- Phase Change RAM
- Three-dimensional space

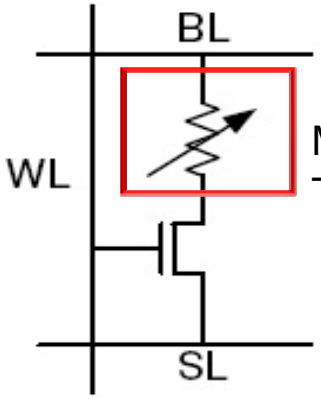
Different Memory Technologies



SRAM
6T structure

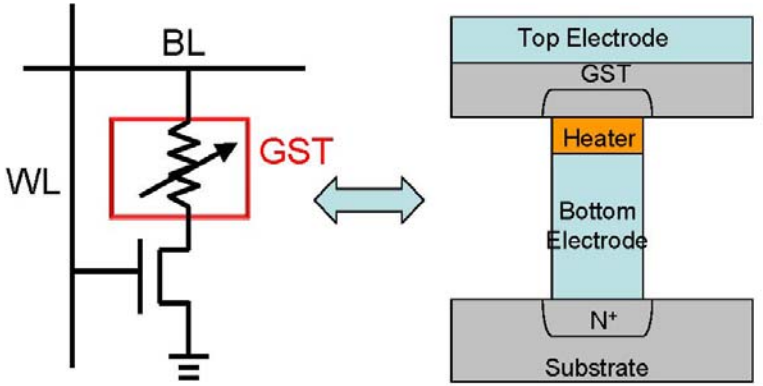


DRAM
1T1C structure



MTJ (Magnetic Tunnel Junction)

Magnetic RAM
1T1J structure



Phase Change RAM
1T1J structure

Comparisons

	SRAM	eDRAM	MRAM	PRAM
Density (ratio)	Low (1)	High (4)	High (4)	High(16)
Dynamic Power	Low	Medium	Low for read; High for write	Medium for read; High for write
Leakage Power	High	Medium	Low	Low
Speed	Very Fast	Fast	Fast for read; Slow for write	Slow for read; Very slow for write
Non-volatility	No	No	Yes	Yes
Scalability	Yes	Yes	Yes	Yes
Endurance	10^{16}	10^{16}	$>10^{15}$	10^{8}

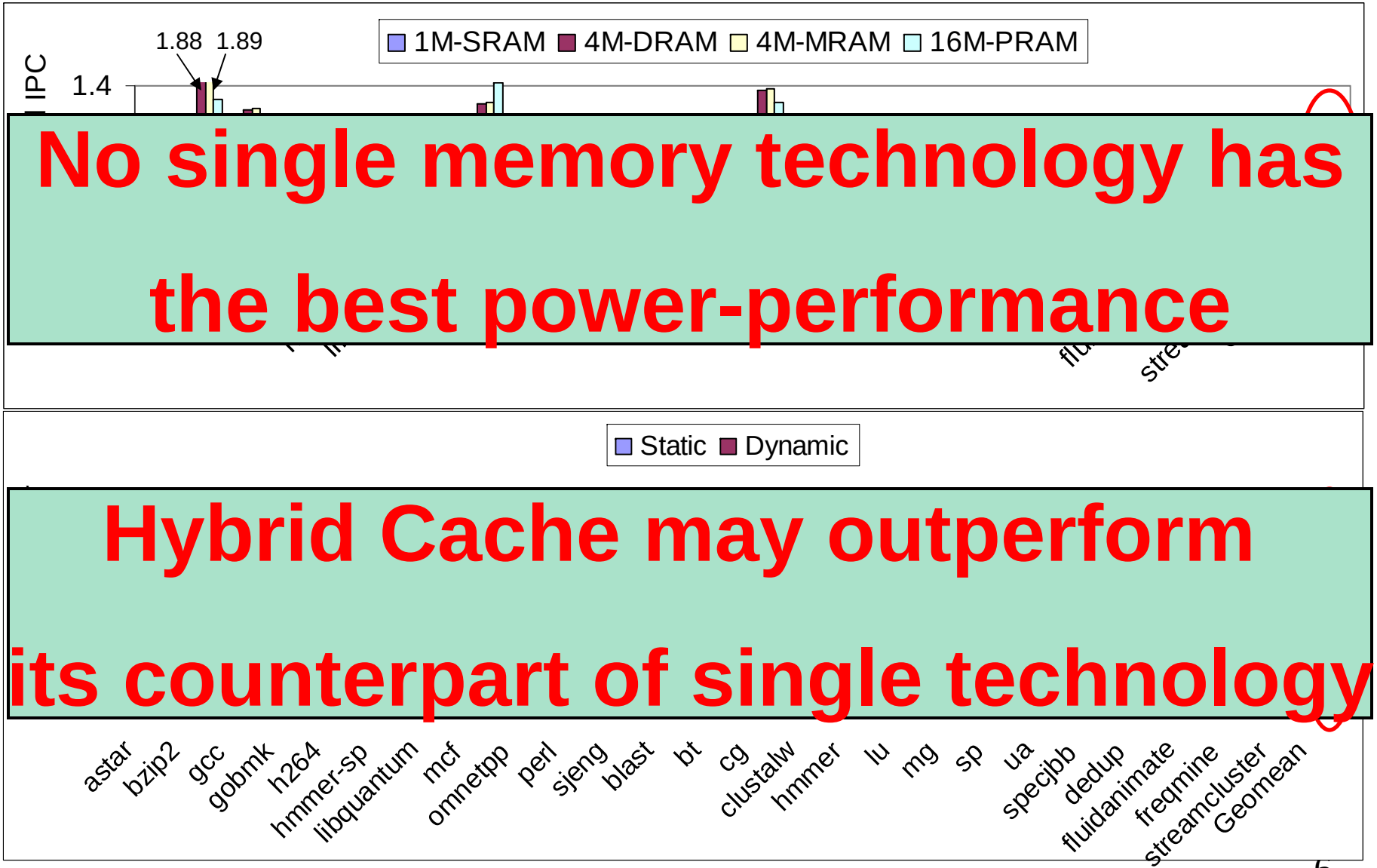
Reduce Cache miss rate

Increase hit latency

Low leakage power

High dynamic power

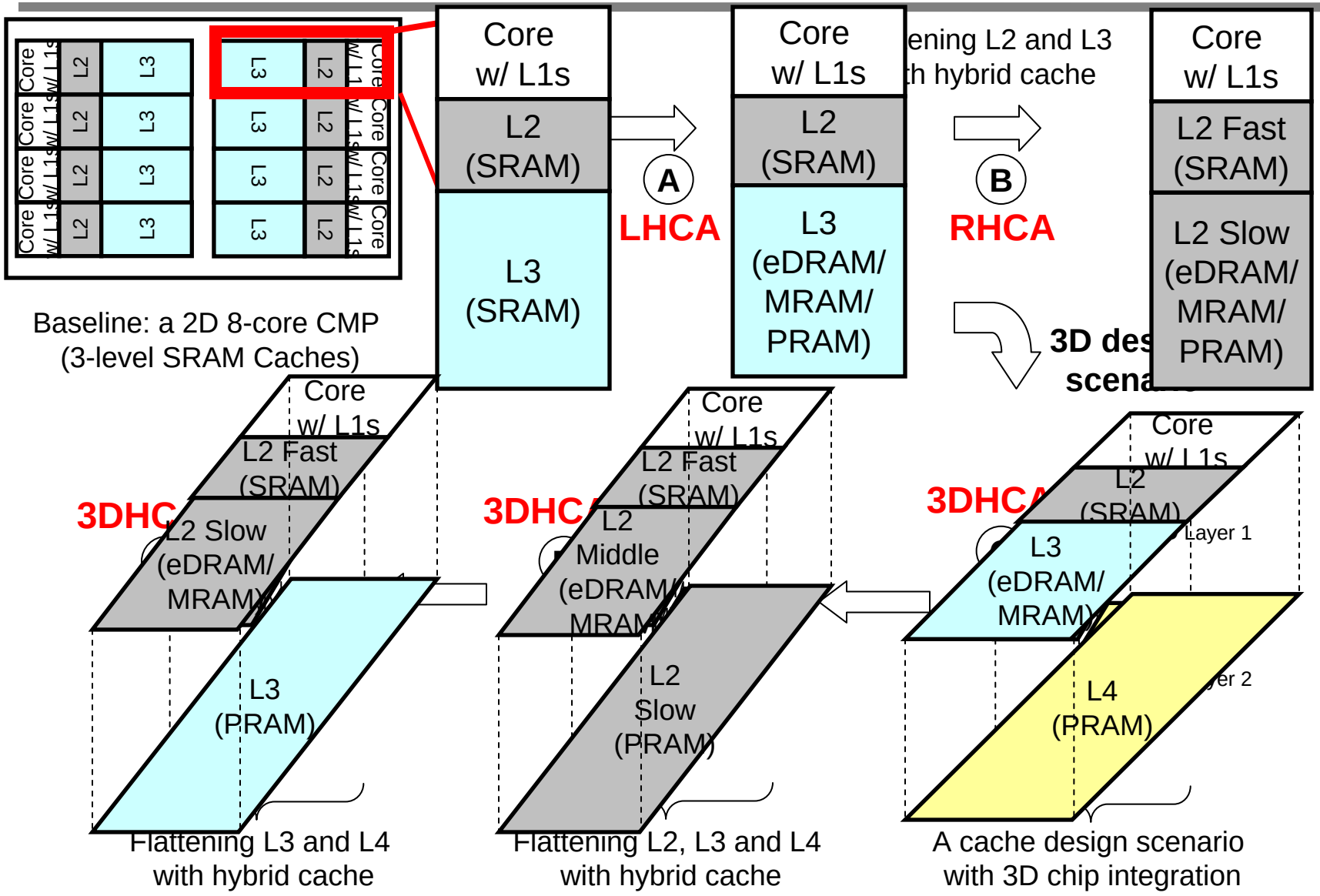
Motivation


1.88
1.89
1M-SRAM
4M-DRAM
4M-MRAM
16M-PRAM
IPC
1.4
No single memory technology has the best power-performance
Static
Dynamic
Hybrid Cache may outperform its counterpart of single technology
astar
bzip2
gcc
gobmk
h264
hmmer-sp
libquantum
mcf
omnetpp
perl
sjeng
blast
bt
cg
clustalw
hmmer
lu
mg
sp
ua
specjbb
dedup
fluidanimate
freqmine
streamcluster
Geomean

Outline

- Introduction and Motivation
- *Methodology*
- Level based Hybrid Cache Architecture
- Region based Hybrid Cache Architecture
- 3D Hybrid Cache Stacking
- Conclusions

Evaluation Methodology



Evaluation Setup

Cache	Density	Latency (cycles)	Dyn. eng (nJ)	Static power (W)
SRAM(1MB)	1	8	0.388	1.36
eDRAM(4MB)	4	24	0.72	0.4
MRAM(4MB)	4	Read:20, write:60	Read:0.4 write:2.3	0.15
PRAM(16MB)	16	Read:40 write:200	Read:0.8 write:1.5	0.3

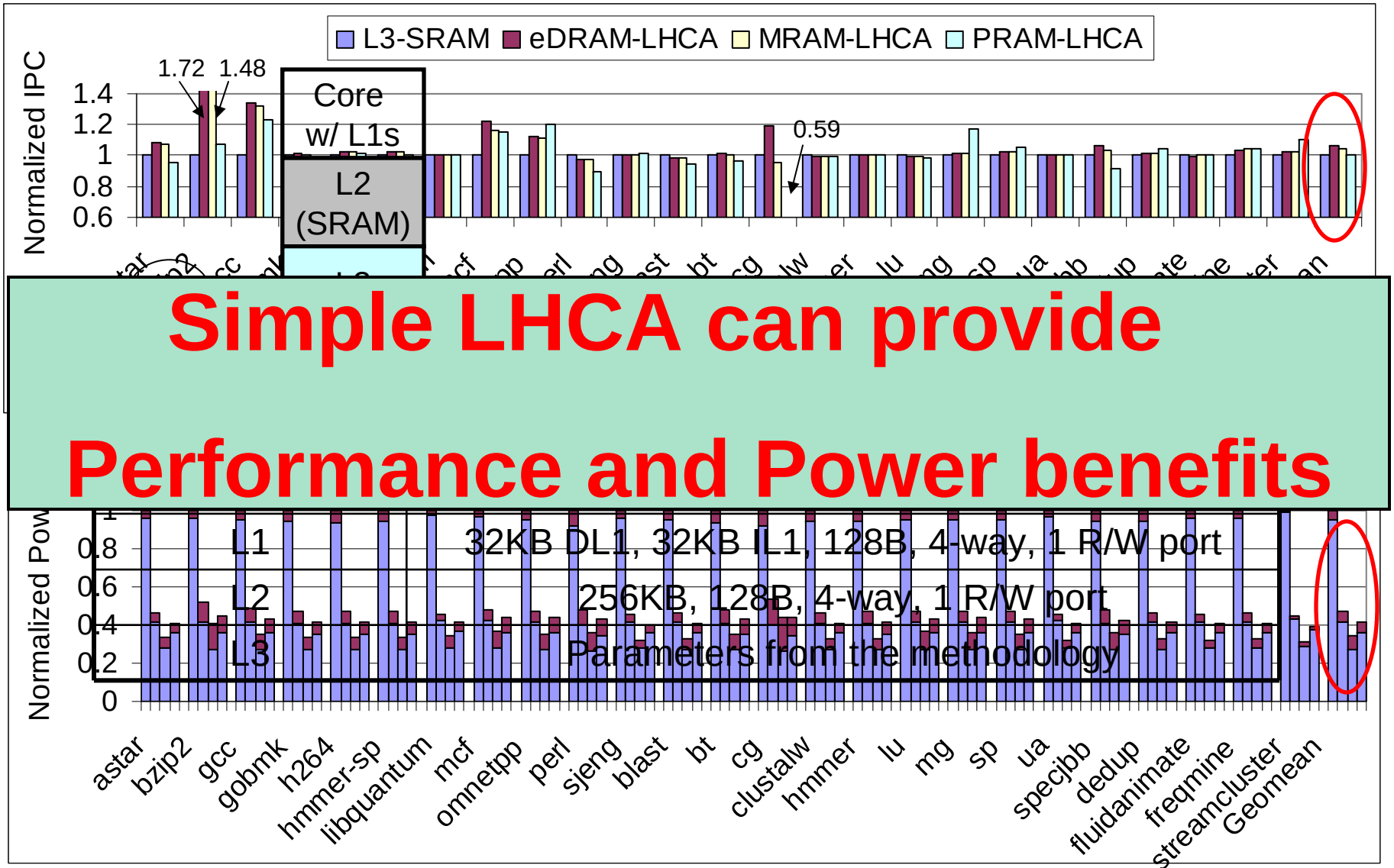
Processor	8-way issue out-of-order, 8-core, 4GHz
L1	32KB DL1, 32KB IL1, 128B, 4-way, 1 R/W port
L2/L3/L4	See corresponding design cases
Memory	400 cycles latency

- Benchmarks: SpecInt06, Specjbb, NAS, Bioperf, Parsec
- Simulator: SystemSim full system simulator

Outline

- Introduction and Motivation
- Methodology
- *Level based Hybrid Cache Architecture*
- Intra-Level Hybrid Cache Architecture
- 3D Hybrid Cache Stacking
- Conclusions

LHCA: Performance and Power



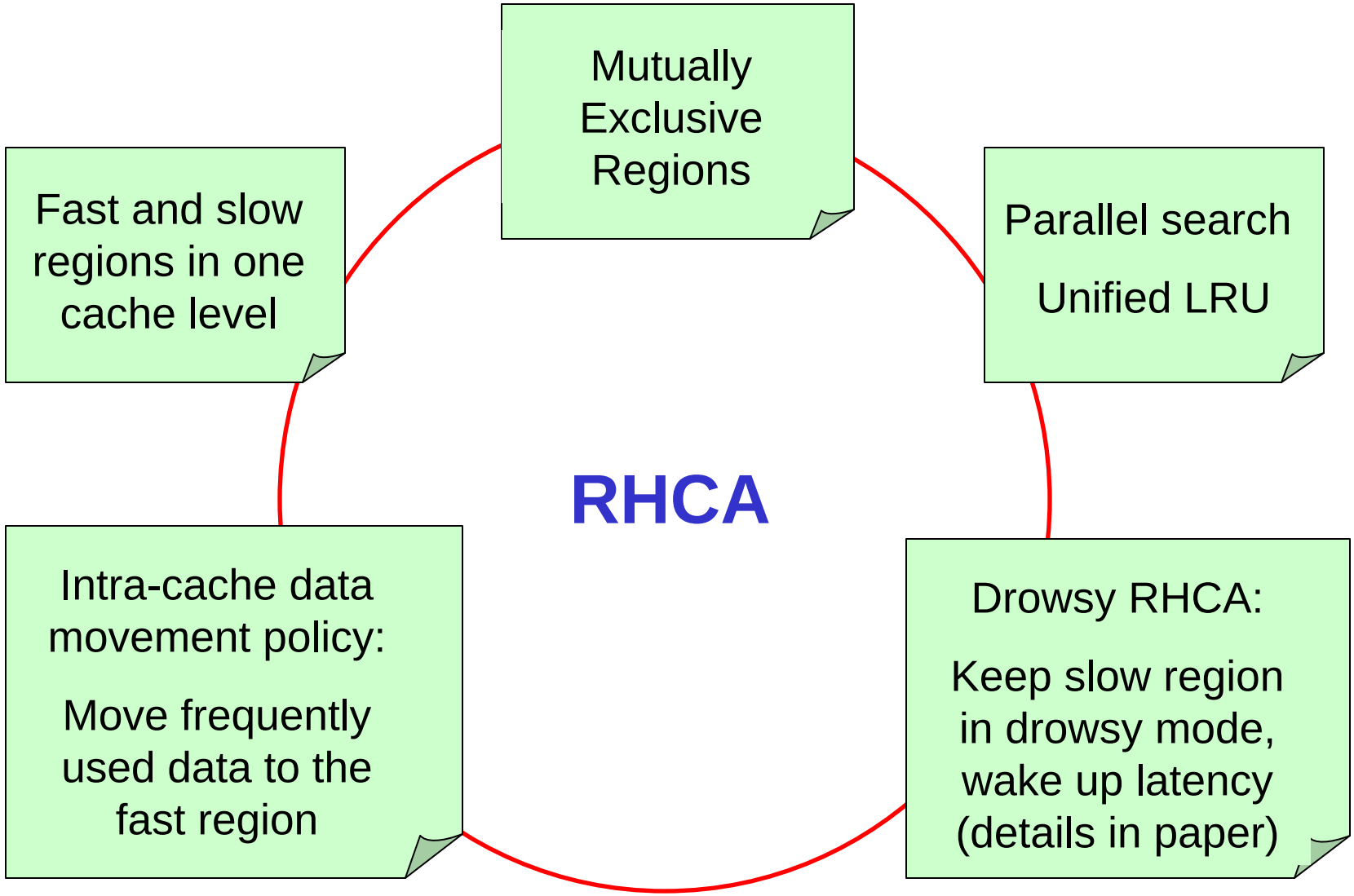
Simple LHCA can provide

Performance and Power benefits

Outline

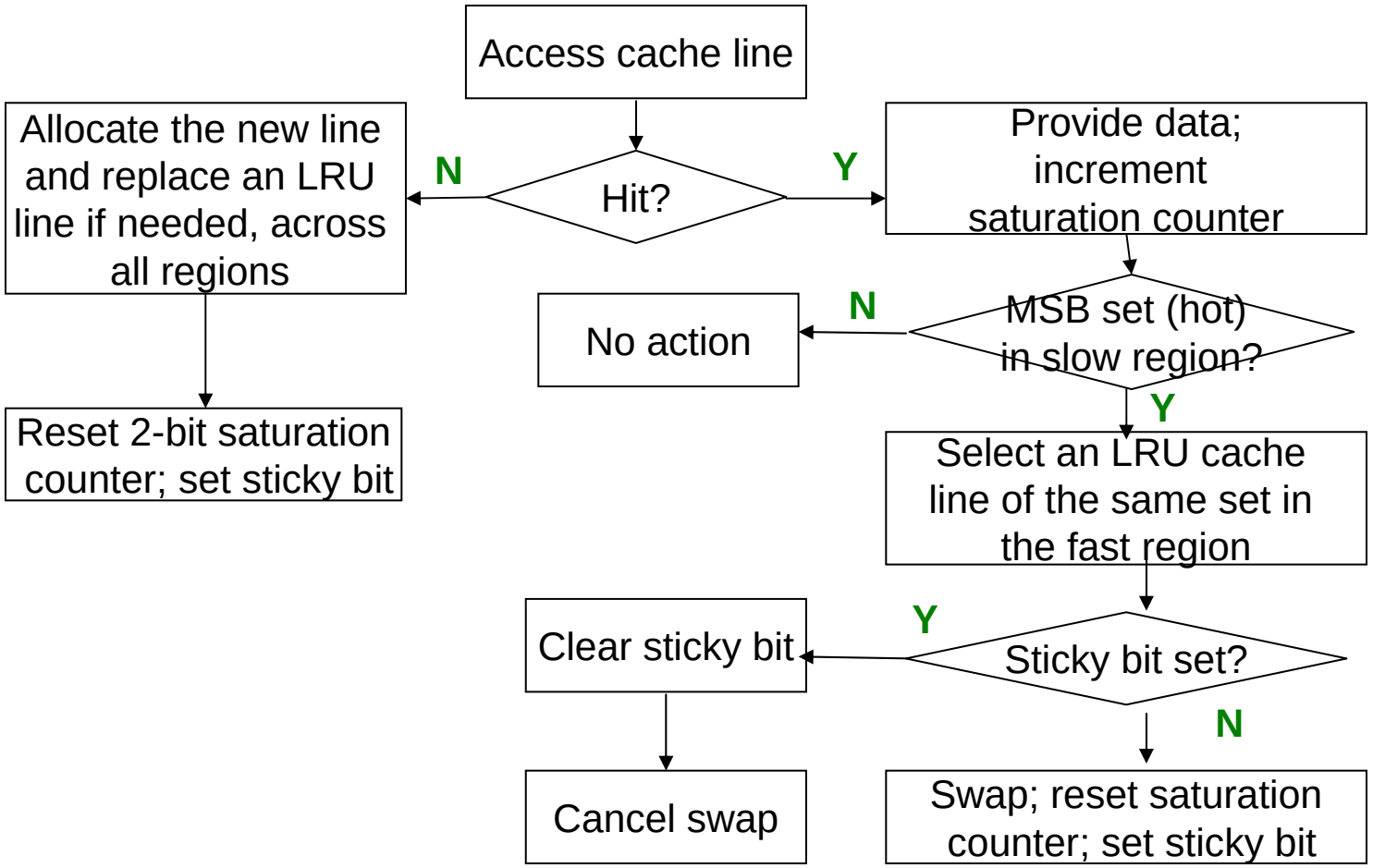
- Introduction and Motivation
- Methodology
- Level based Hybrid Cache Architecture
- *Region based Hybrid Cache Architecture*
- 3D Hybrid Cache Stacking
- Conclusions

RHCA Features

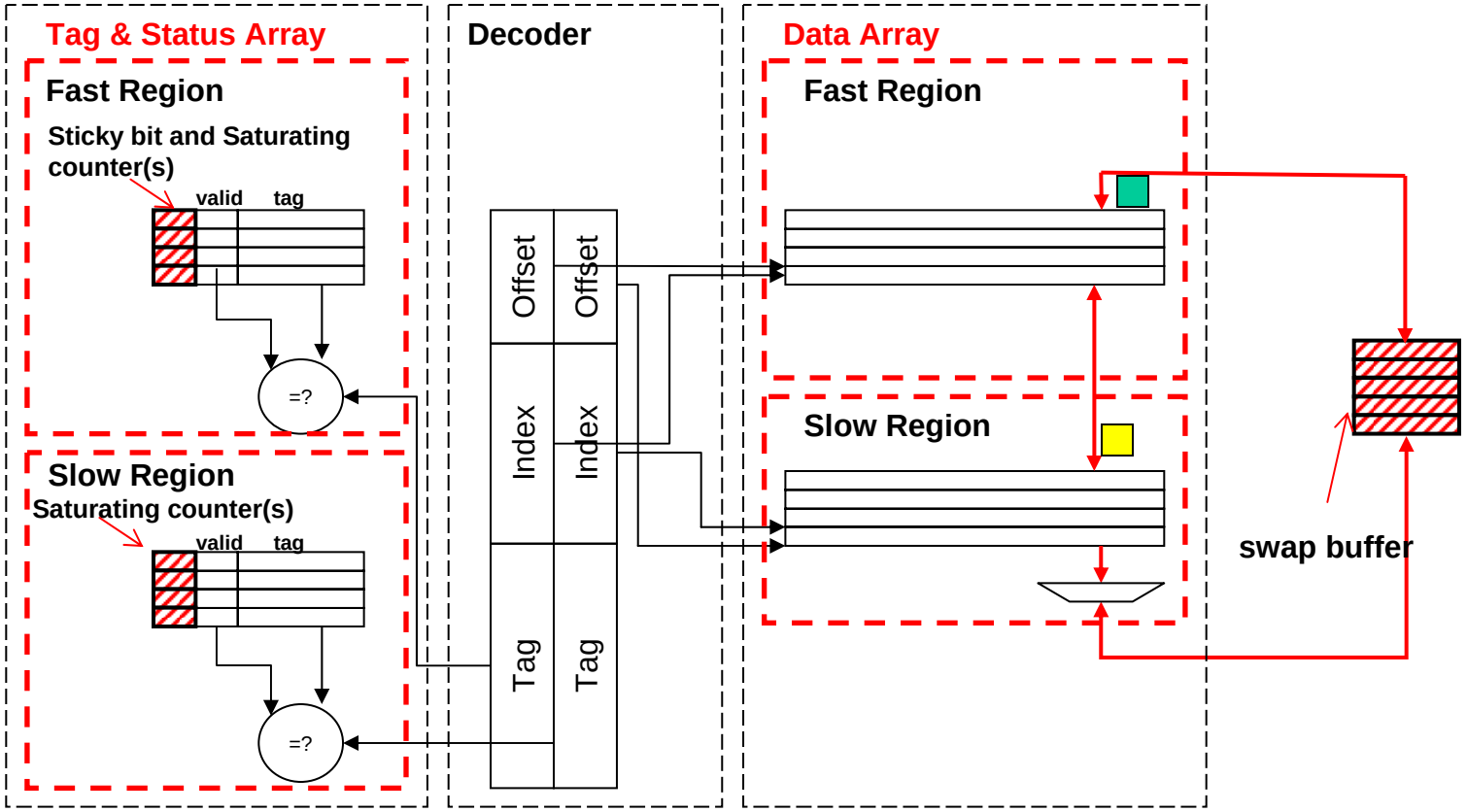


RHCA Policy

- Cache line migration policy



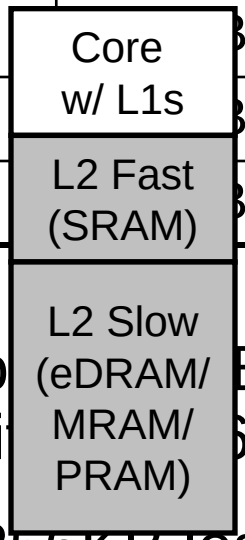
RHCA Hardware Support



- Hardware support
  - Saturating counter in slow and sticky bit in fast, swap buffer
  - Minimum hardware support: 1-bit sticky bit in fast region

RHCA Configuration

RHCA (fast+slow)	Fast region	L2 total size (latency)
SRAM+eDRAM	(6 cycles)	4MB (24 cycles)
SRAM+MRAM	(6 cycles)	4MB (r:20, w:60)
SRAM+PRAM	(6 cycles)	16MB (r:40, w:200)

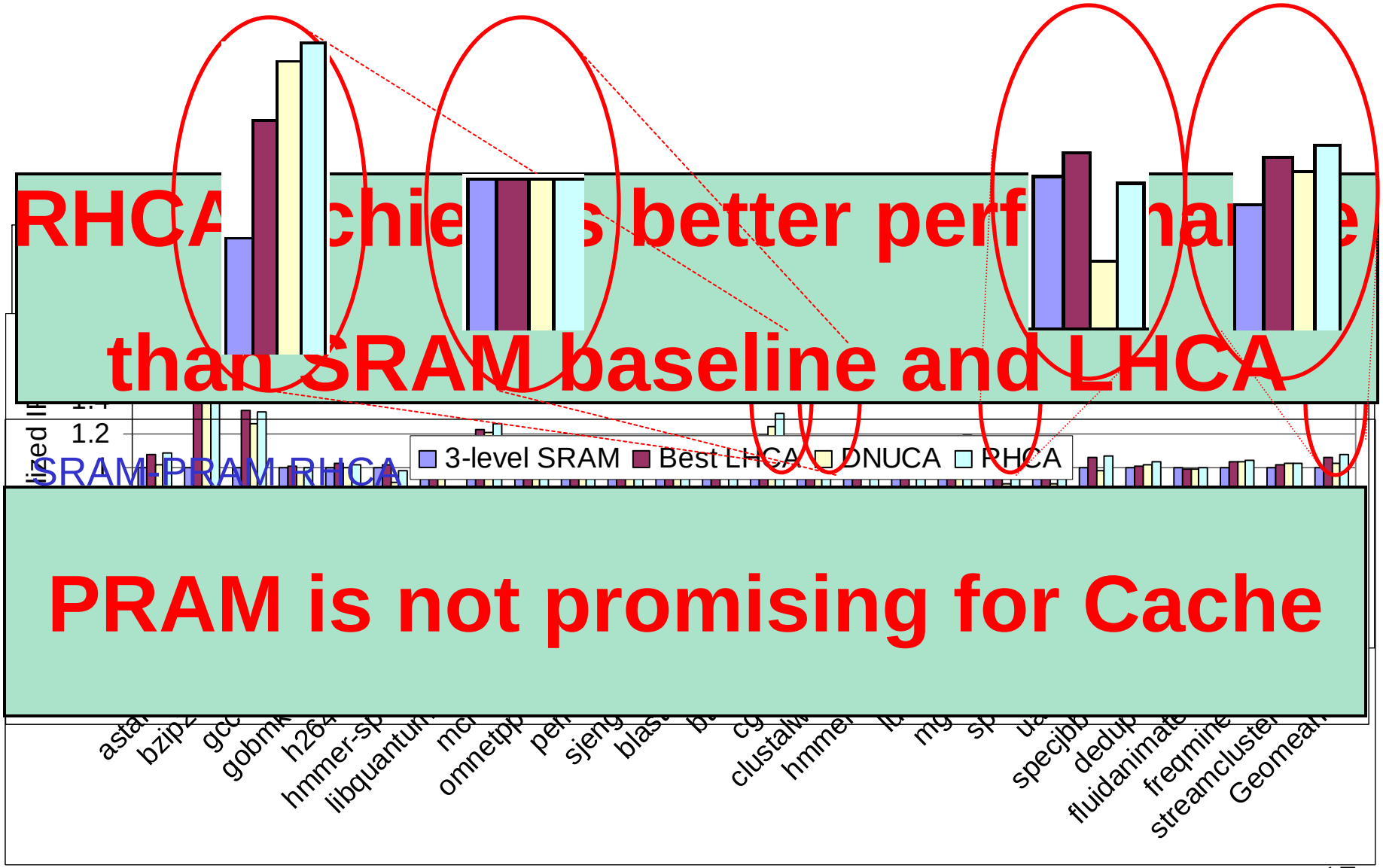


B

RHCA

- Slow regio B/bank, 1 r/w port, block size 128B, associativi , 64
- RHCA is 256KB less size than corresponding LHCA
  - Avoid odd-sized cache
- DNUCA policy: more fine grained, move a line to a closer bank on each hit, bank-based, same size

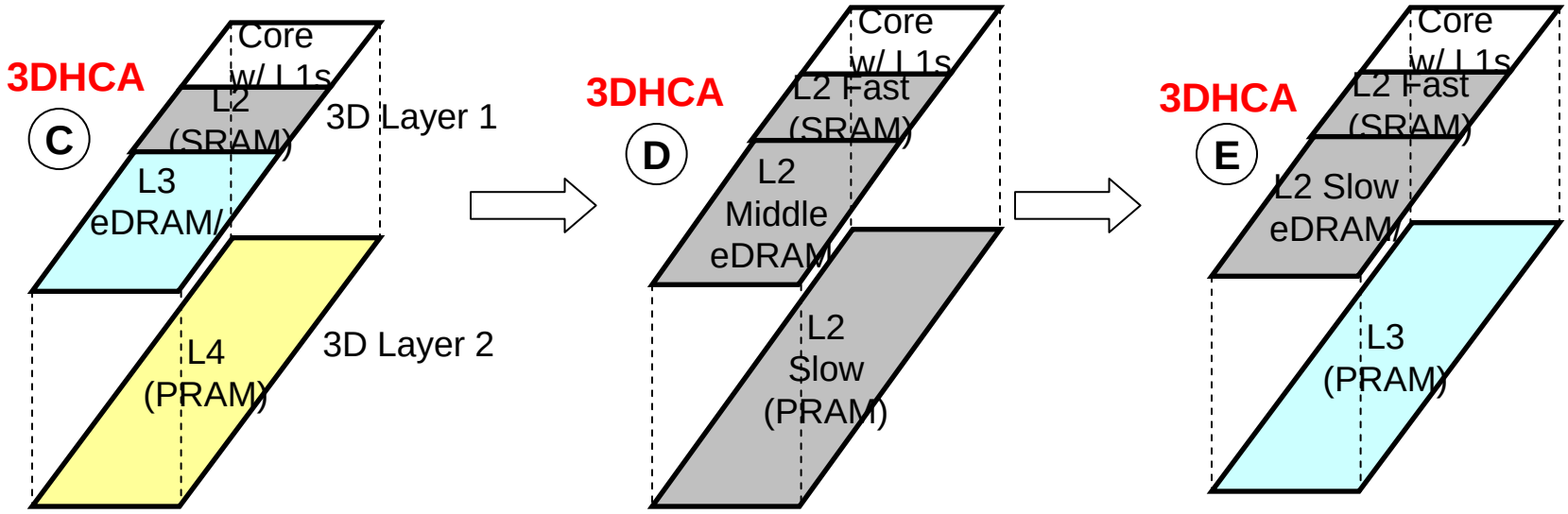
RHCA Result


RHCA achieves better performance than SRAM baseline and LHCA
SRAM-PRAM-RHCA
PRAM is not promising for Cache
1.2
3-level SRAM
Best LHCA
DNUCA
RHCA
bzip2
gcc
h264
hmmer-sp
libquantum
omnetpp
sjeng
cg
clustalw
hmmer
mg
specjbb
dedup
fluidanimate
freqmine
streamcluster
Geomean

Outline

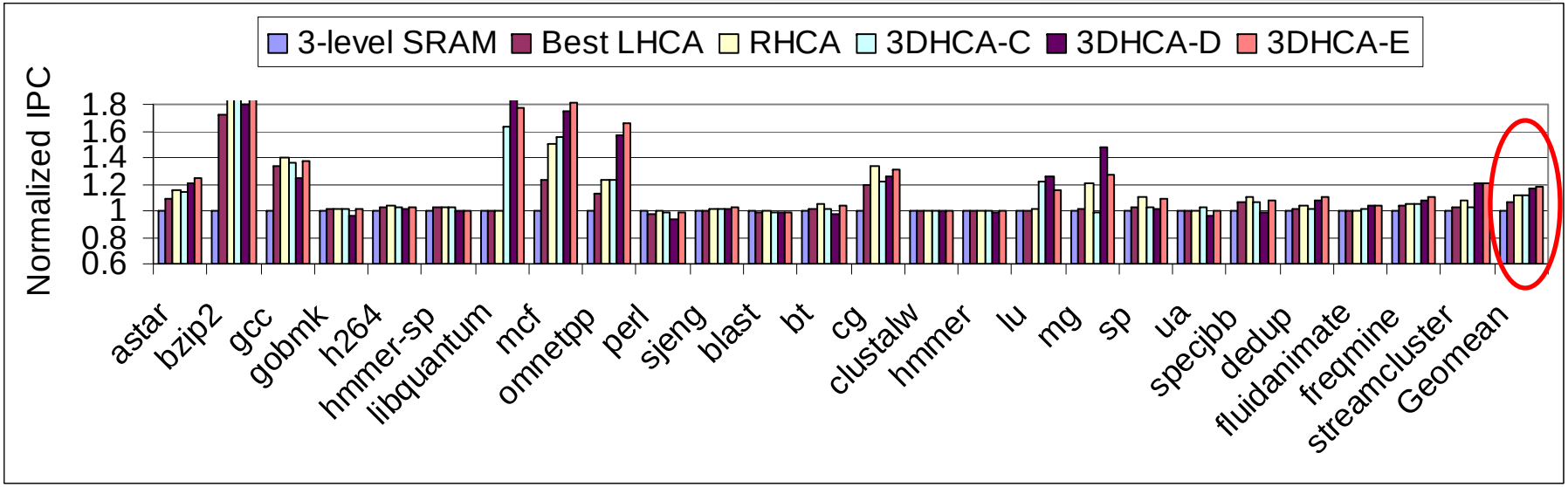
- Introduction and Motivation
- Methodology
- Level based Hybrid Cache Architecture
- Region based Hybrid Cache Architecture
- *3D Hybrid Cache Stacking*
- Conclusions

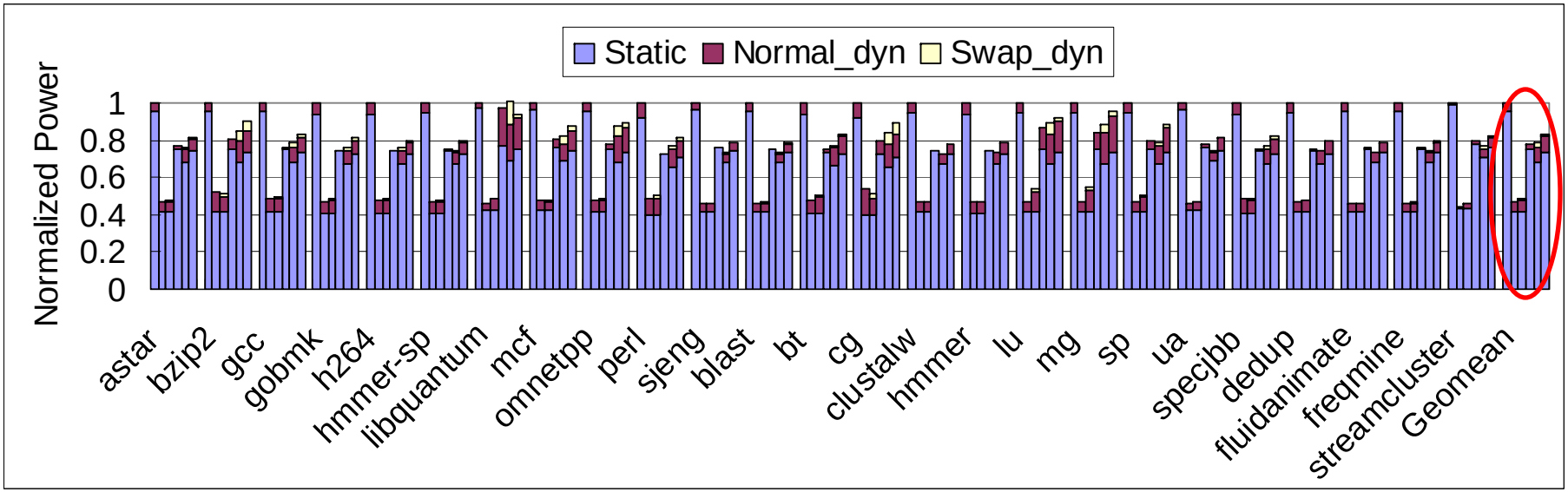
3DHCA-configuration



- 3DHCA-C (3D LHCA): 256KB L2 SRAM, 4M L3 eDRAM, 32M L4 PRAM
- 3DHCA-D: 32M L2 fast, middle, slow region (3D RHCA)
  - Data in slow region can be moved to fast and middle regions
- 3DHCA-E: 4M L2 fast+slow region, 32M L3 PRAM (LHCA+RHCA)

3DHCA-result





Conclusion

- Hybrid cache architecture is promising to improve cache power-performance under same chip area/footprint
- RHCA and LHCA achieve better power-performance than SRAM-based design
- RHCA outperforms LHCA with minimal hardware support
- 3DHCA achieves better performance than LHCA and RHCA, while still maintains lower power than 2D SRAM baseline