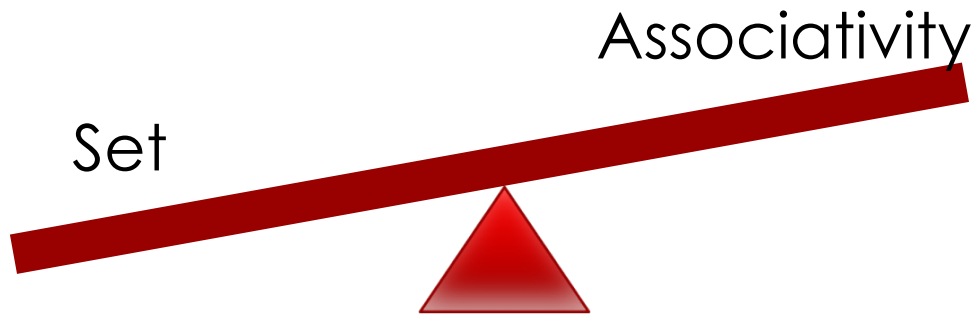


<http://synergy.ece.gatech.edu/>



SEESAW: Set Enhanced Superpage Aware caching

Mayank Parasar^Σ, Abhishek
Bhattacharjee^Ω, Tushar Krishna^Σ

^ΣSchool of Electrical and Computer Engineering
Georgia Institute of Technology



^ΩDepartment of Computer Science
Rutgers University

mparasar3@gatech.edu

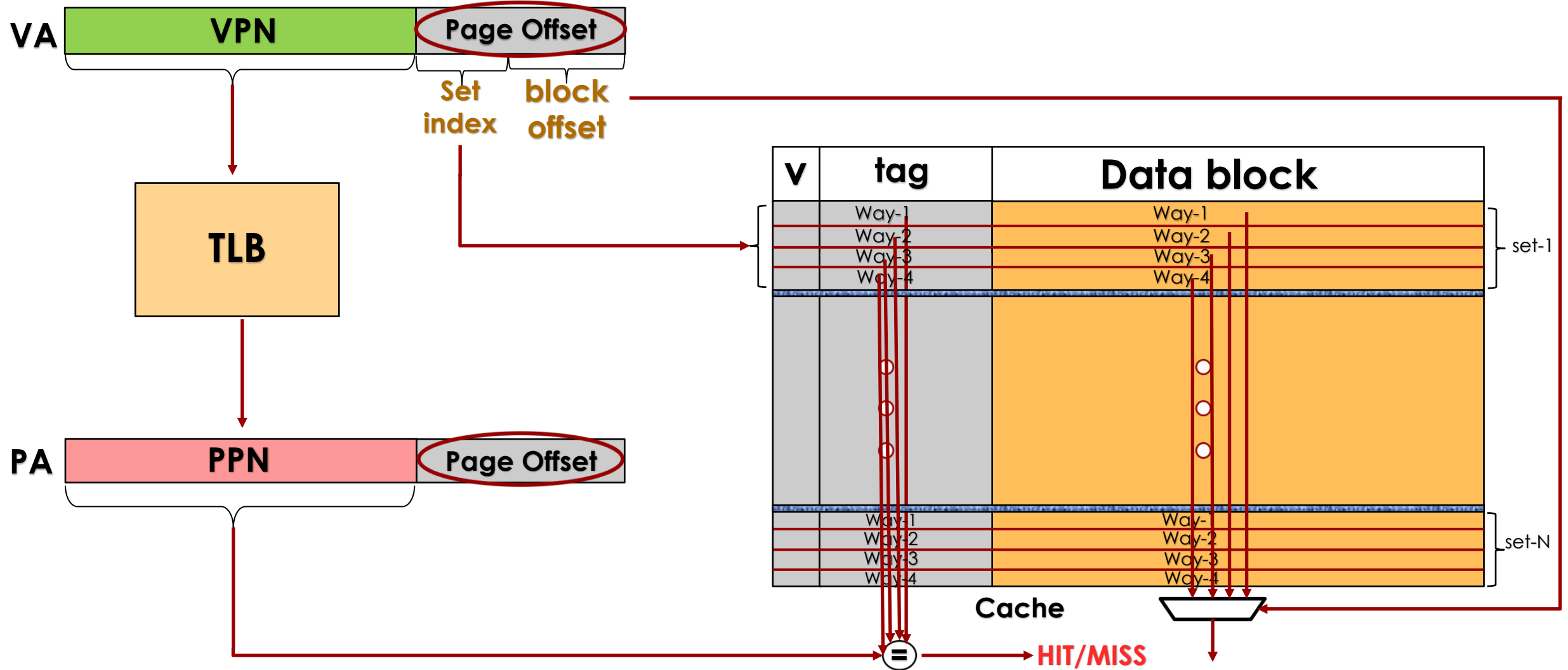
Outline

- **Motivation**
- SEESAW: Concept
- SEESAW: Micro-architecture
- Evaluation Methodology
- Results
- Conclusion

L1 Cache Characteristics

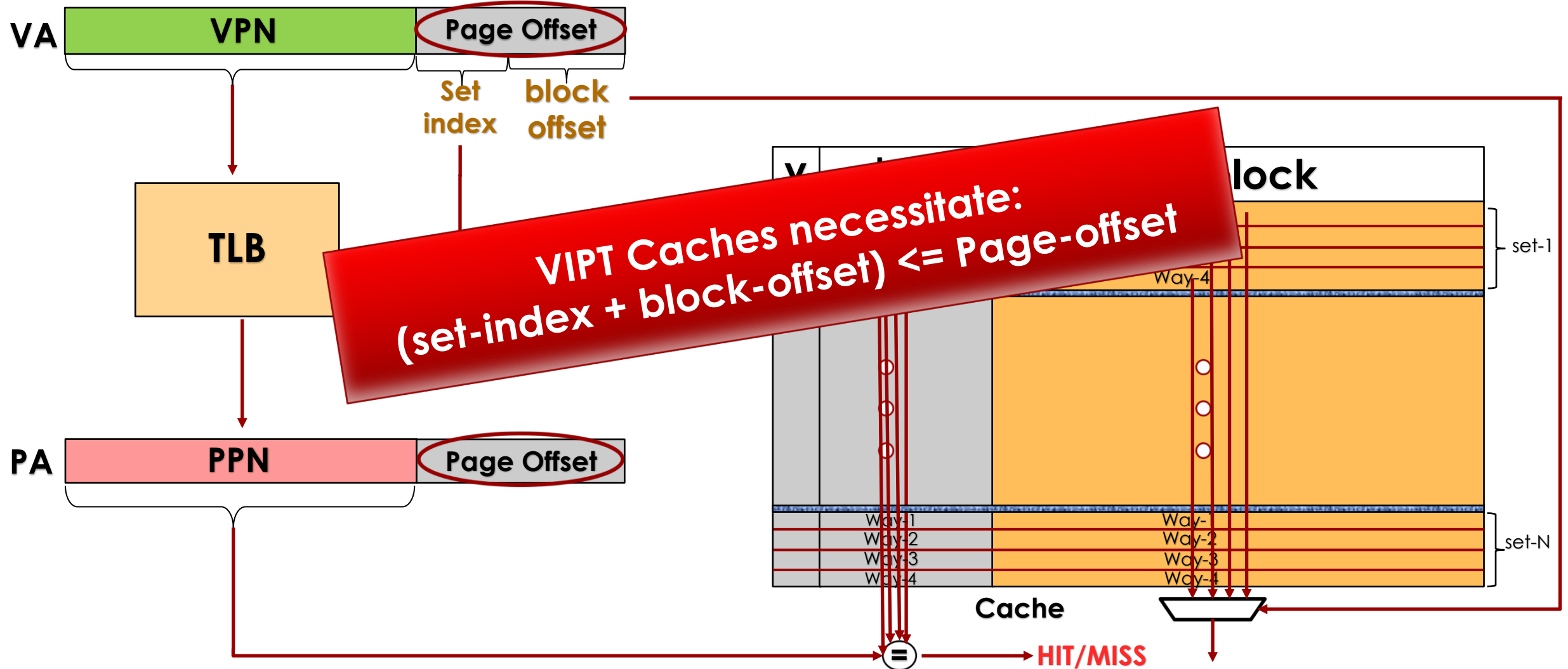
	Ideal-Cache	VIPT-Cache
Fast lookup		
High hit-rate		
Energy Efficiency		

Virtually Indexed Physically Tagged [VIPT] Cache

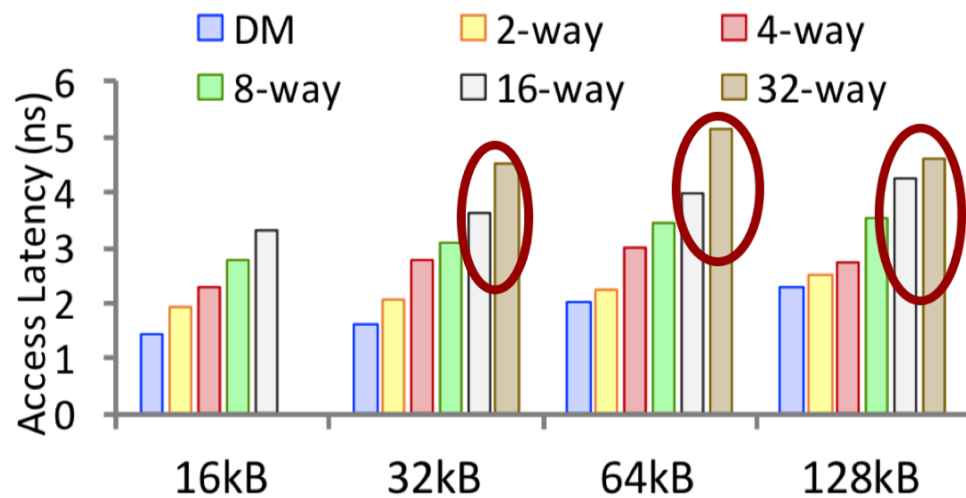


Virtually Indexed Physically Tagged [VIPT] Cache

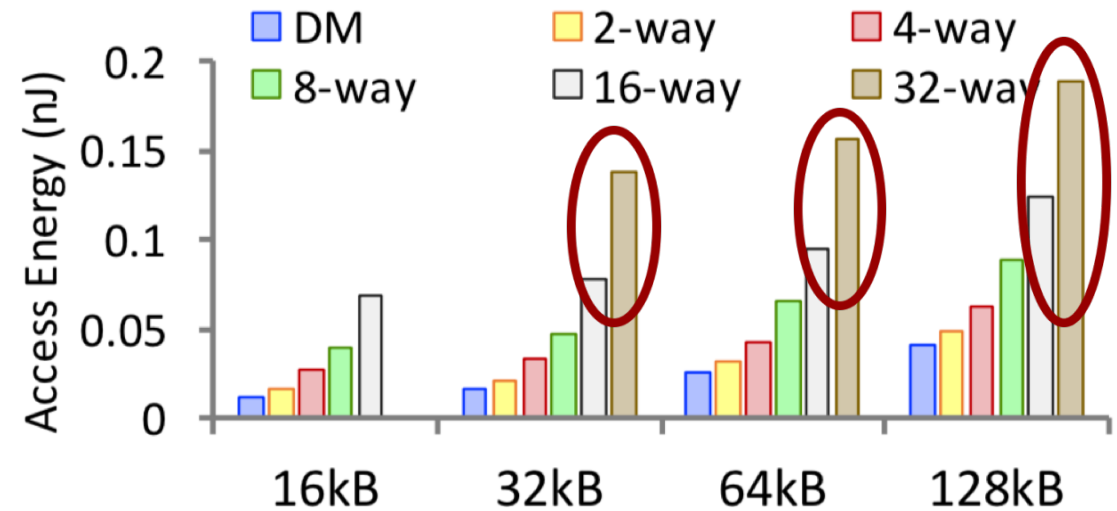
5



Impact of Associativity on Access Latency and Energy of cache

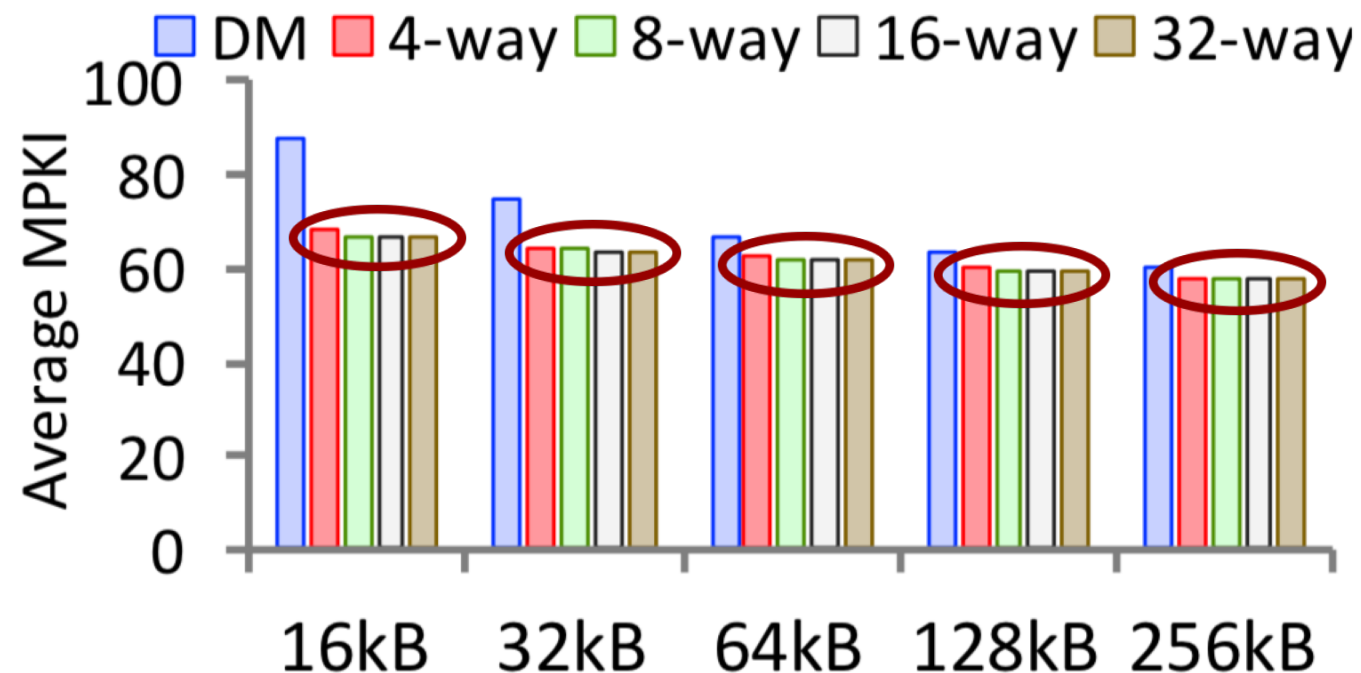


Cache Access Latency









Cache Access Energy

Effect of associativity on MPKI of cache



High Associativity hurts latency and energy without commensurately improving hit rate

Revisiting L1 Cache Characteristics for VIPT Cache

	Ideal-Cache	VIPT-Cache
Fast lookup		 ?
High hit-rate		
Energy Efficiency		 ?

**Virtual
memory!**

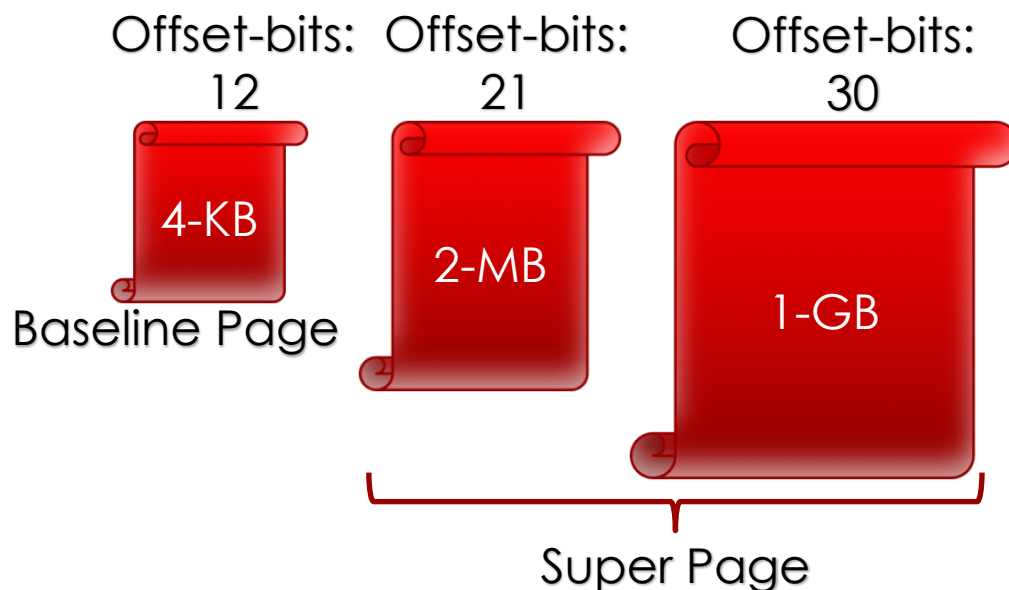
**Virtual
memory!**

Opportunity: Superpage

Is it possible to relax constraints of Traditional VIPT cache?

Yes

How ?

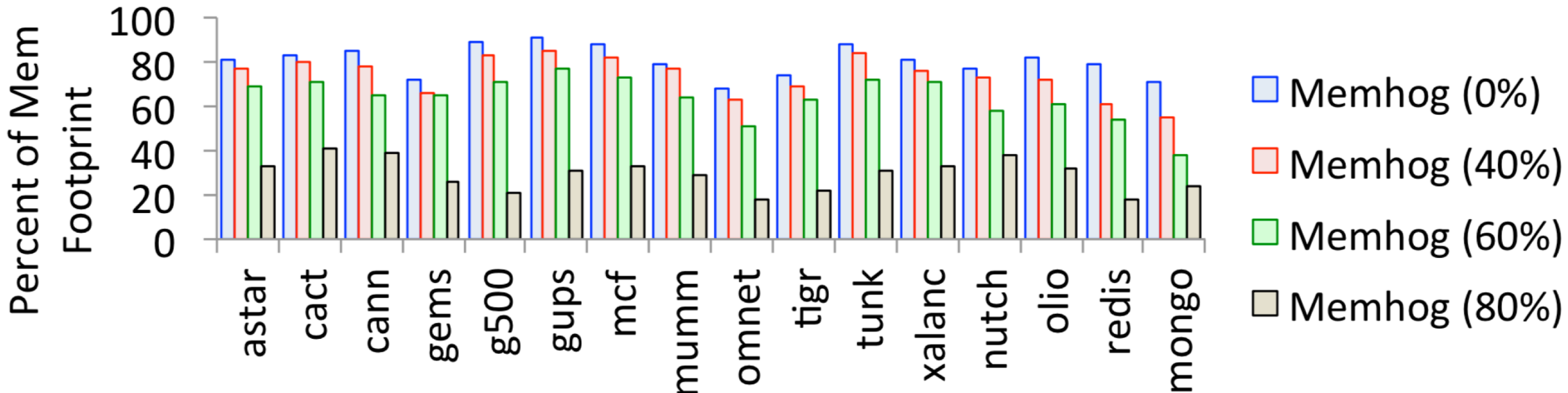


More page-offset bits for superpage!

HW and OS Support for Superpages
in modern processors

Prevalence of superpages in modern OSES under memory fragmentation

10



Ran on 32-core; Sandybridge; 32 GB RAM
Memhog causes memory fragmentation; higher %age indicates higher fragmentation

Outline

- Motivation
- **SEESAW: Concept**
- SEESAW: Micro-architecture
- Evaluation Methodology
- Results
- Conclusion

SEESAW: Concept

	v	tag	Data block
Set:1		Way-1	Way-1
		Way-2	Way-2
		Way-3	Way-3
Set:2		Way-1	Way-1
		Way-2	Way-2
		Way-3	Way-3
Set:3		Way-1	Way-1
		Way-2	Way-2
		Way-3	Way-3
		○	○
		○	○

Less-sets
More-associativity

super-page

Base-page

	v	tag	Data block
Set:1		Way-1	Way-1
Set:2		Way-1	Way-1
Set:3		Way-1	Way-1
Set:4		Way-1	Way-1
Set:5		Way-1	Way-1
Set:6		Way-1	Way-1
Set:7		Way-1	Way-1
Set:8		Way-1	Way-1
Set:9		Way-1	Way-1
		○	○
		○	○

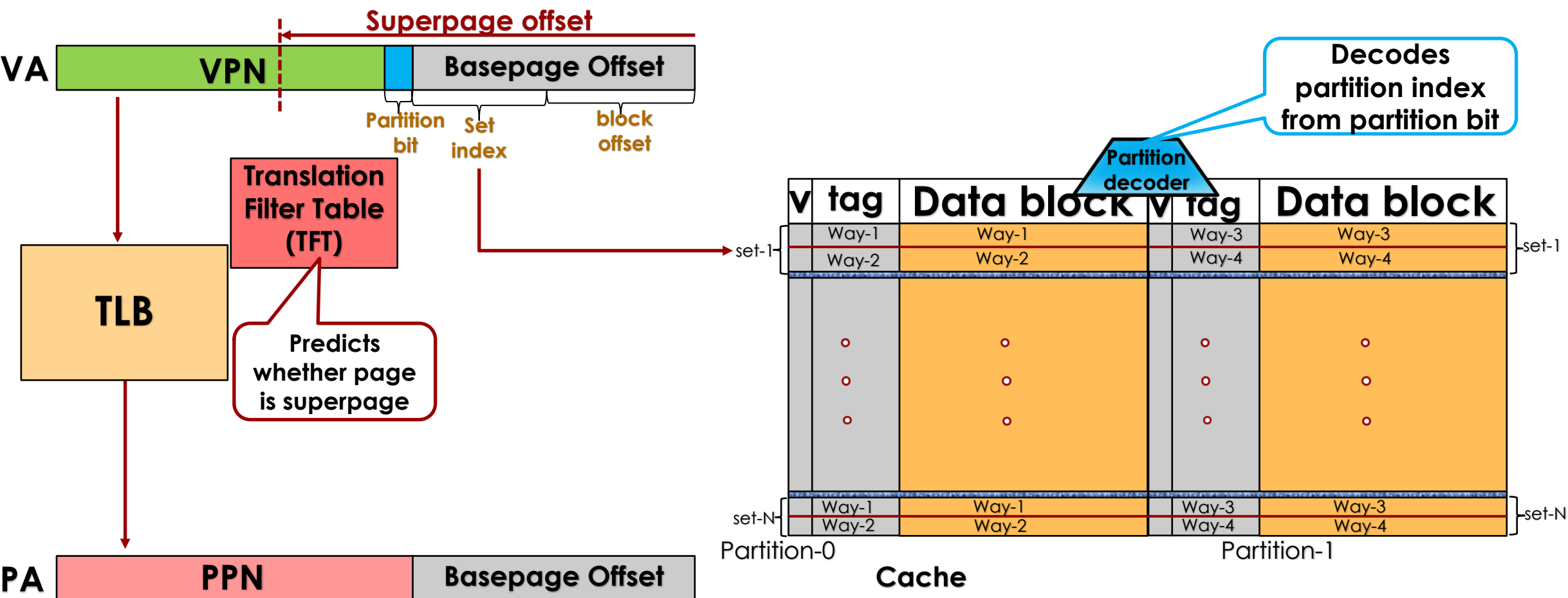
More-sets
Less-associativity

Faster
Energy-Efficient

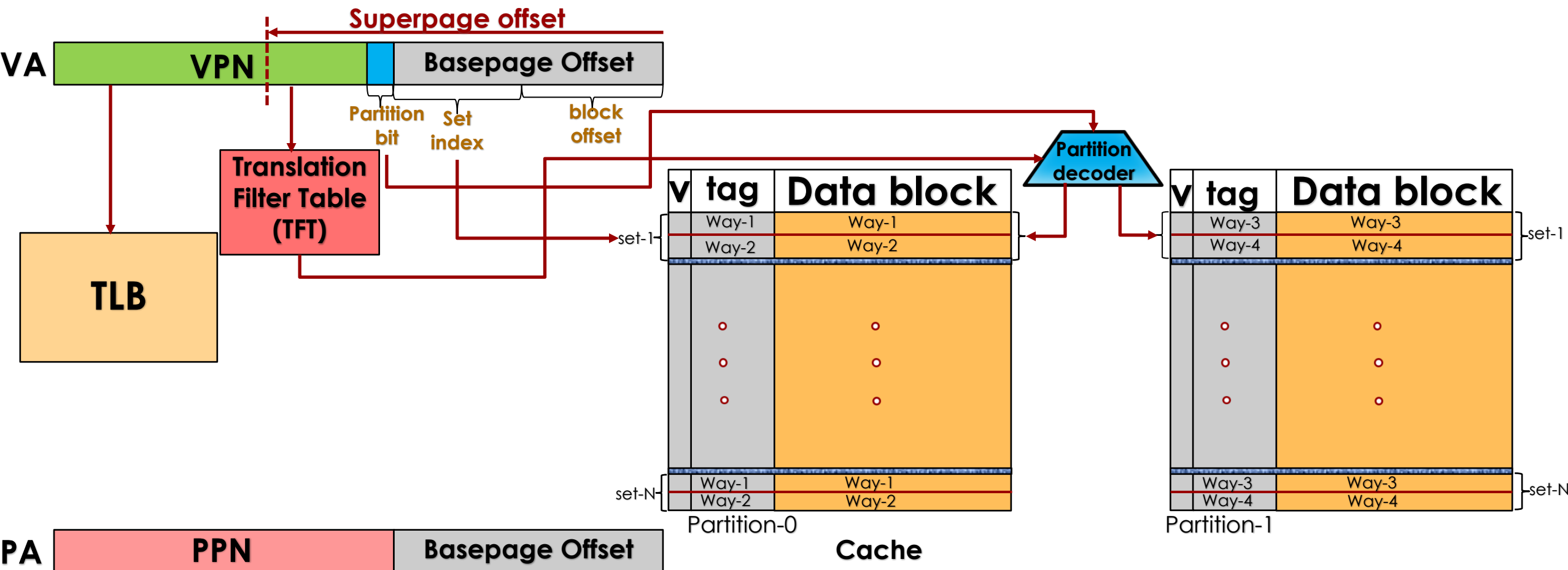
Outline

- Motivation
- SEESAW: Concept
- **SEESAW: Micro-architecture**
- Evaluation Methodology
- Results
- Conclusion

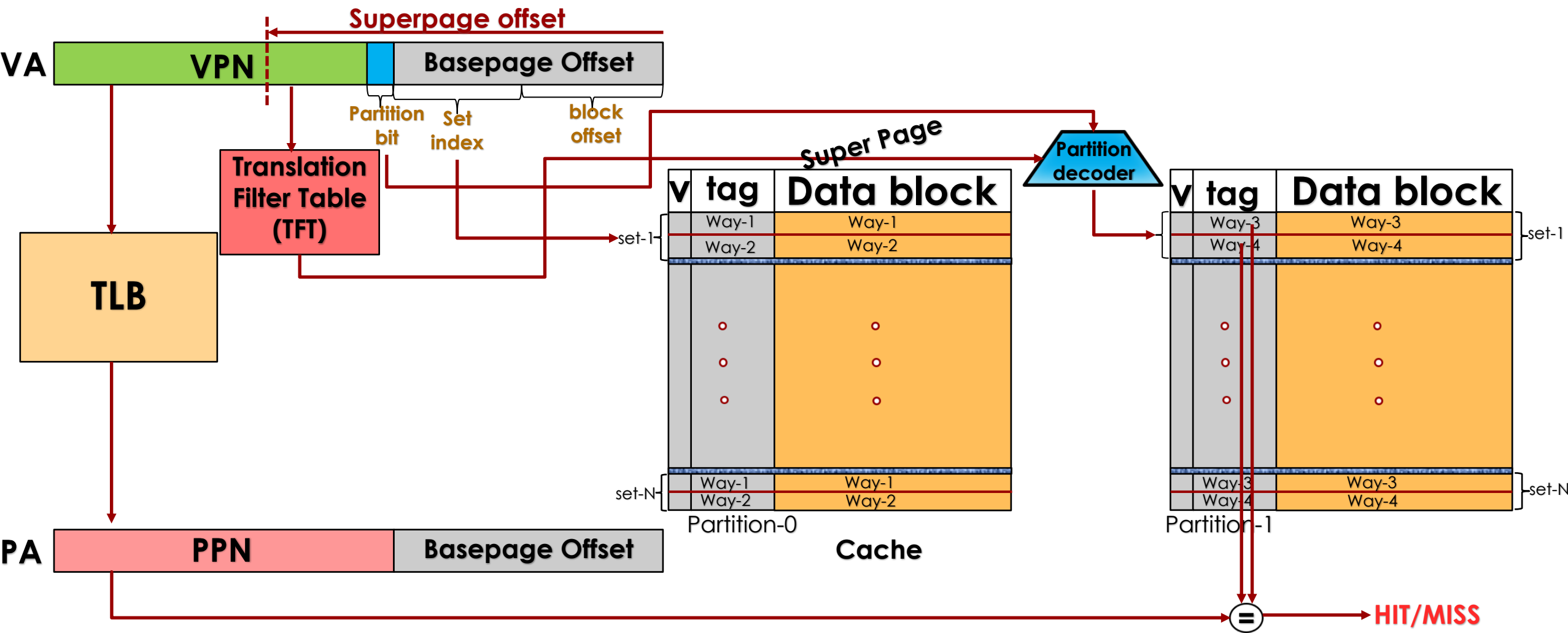
SEESAW: Micro-architecture



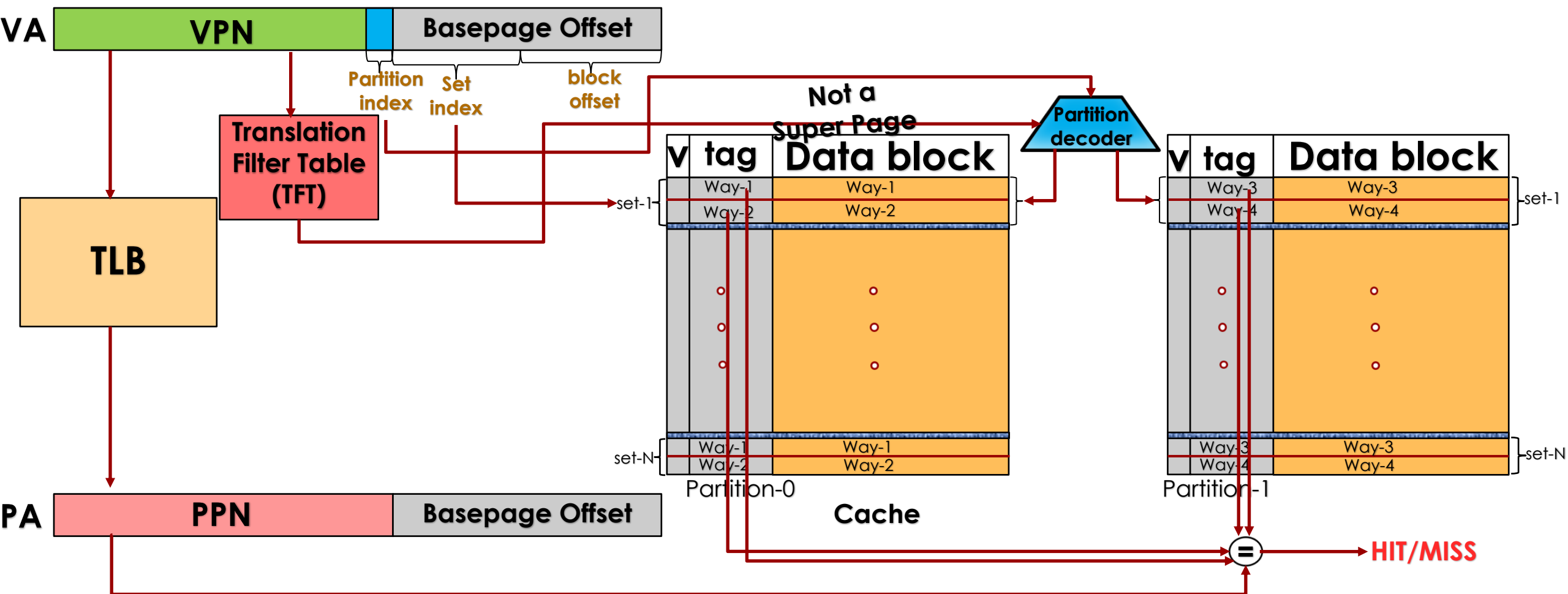
SEESAW: Micro-architecture



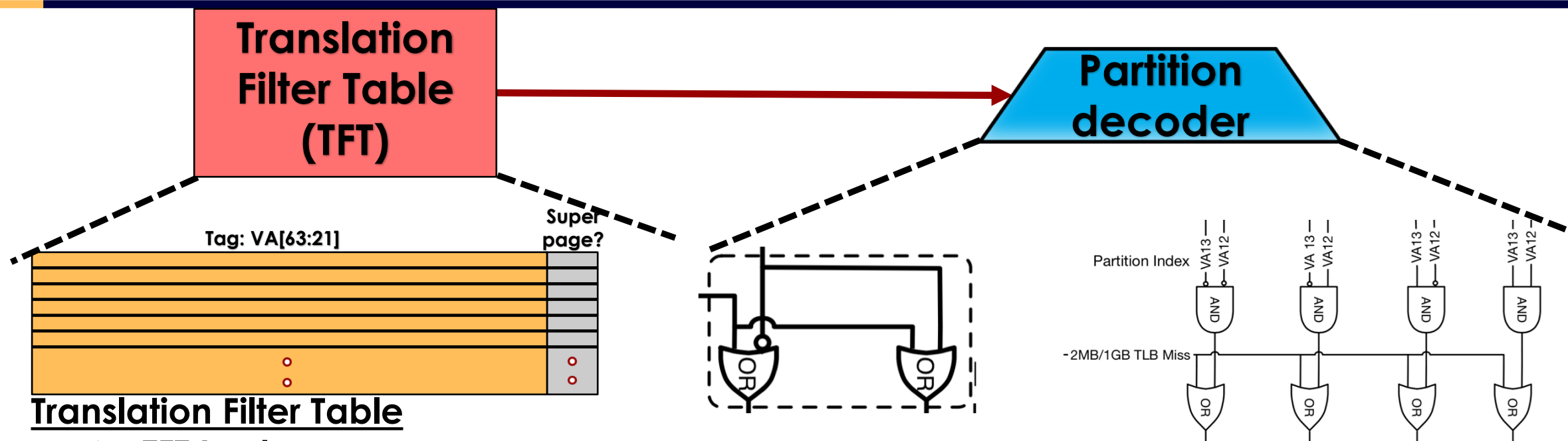
SEESAW: Superpage access



SEESAW: Basepage access



SEESAW: TFT and Partition Decoder

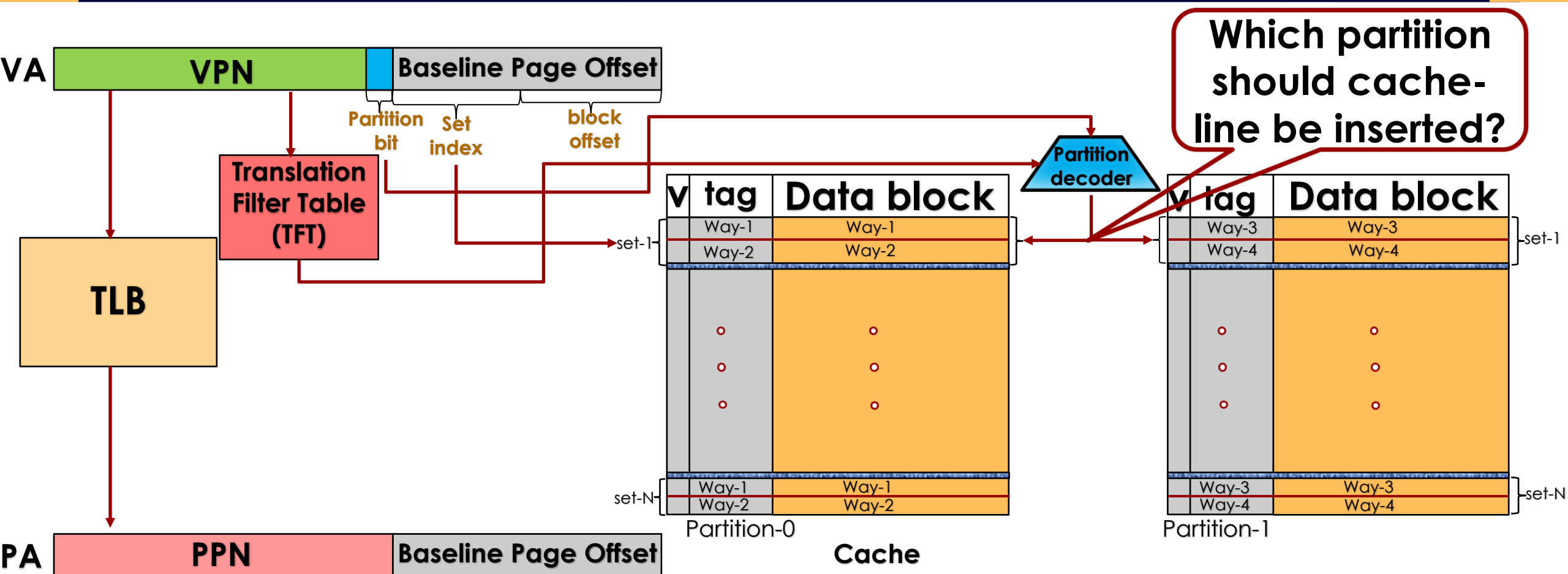


- **TFT Lookup**
 - Direct mapped
 - False negative due to size
- **TFT Update**
 - VA misprediction
 - 2MB L1-TLB fill
 - 2MB L1-TLB Invalidation

Partition Decoder

- For 32kB Cache
- For 64kB Cache

SEESAW: Cache line insertion policy



SEESAW: Cache line insertion policy

- 4way-8way
 - Superpage miss: victim within the partition
 - Basepage miss: victim within the set
- 4way
 - Uses LRU within the associated partition
 - Avoid installing the same line twice
 - Saves energy

SEESAW: System Level Optimization

- Cache coherence
 - Cache coherence lookups use physical address
 - Snoopy provide higher energy benefits over Directory based coherence
- Page table modifications
 - Superpage splintered into multiple basepages
 - Multiple basepages promoted to superpages

Outline

- Motivation
- SEESAW: Concept
- SEESAW: Micro-architecture
- **Evaluation Methodology**
- Results
- Conclusion

SEESAW: Simulated system

CPU Models	
Out-of-Order	~Intel Sandybridge: 168-entry ROB, 54-entry Instruction Scheduler, 16 byte I-fetches per cycle
In-order	~Intel Atom: Dual-Issue, 16-stage pipeline
Memory System	
L1 Cache	Private Split L1I (32kB) + L1D (Table 3)
TLB (Atom)	L1 (64-entry for 4kB, 32-entry for 2MB), 512-entry L2
TLB (Sbridge)	Split L1 (128-entry for 4kB, 16-entry for 2MB)
LLC	Unified, 24MB
DRAM	4GB, 51ns round-trip access latency
System Parameters	
Technology	22nm
Frequency	1.33 GHz, 2.80 GHz, 4.0 GHz
Cores	32, 64, 128
Coherence	MOESI directory

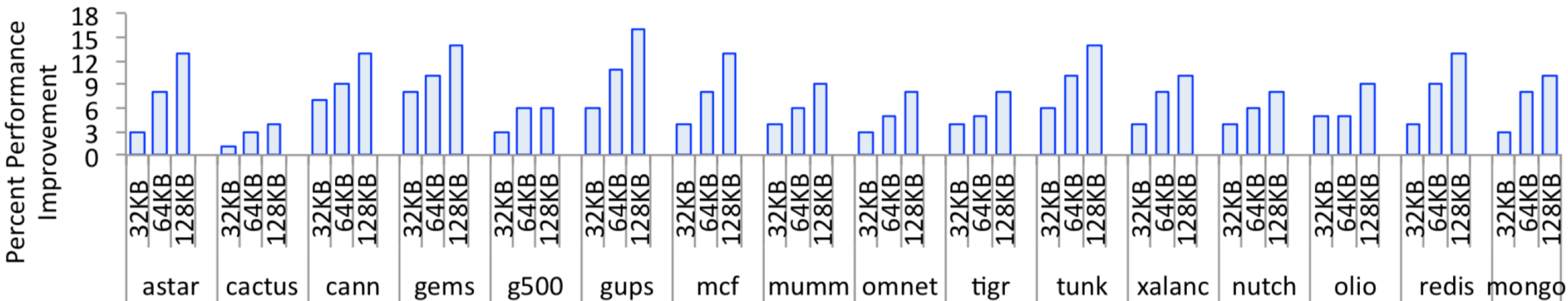
SEESAW: Workloads

- Spec
- Parsec
- Cloudsuite
 - Tunkrank
- Biobench
 - Mummer
 - Tiger
- MongoDB
- Server Workload
 - graph500
 - Nutch Hadoop
- Social-event web service
 - Olia
- Key value store
 - Redis

Outline

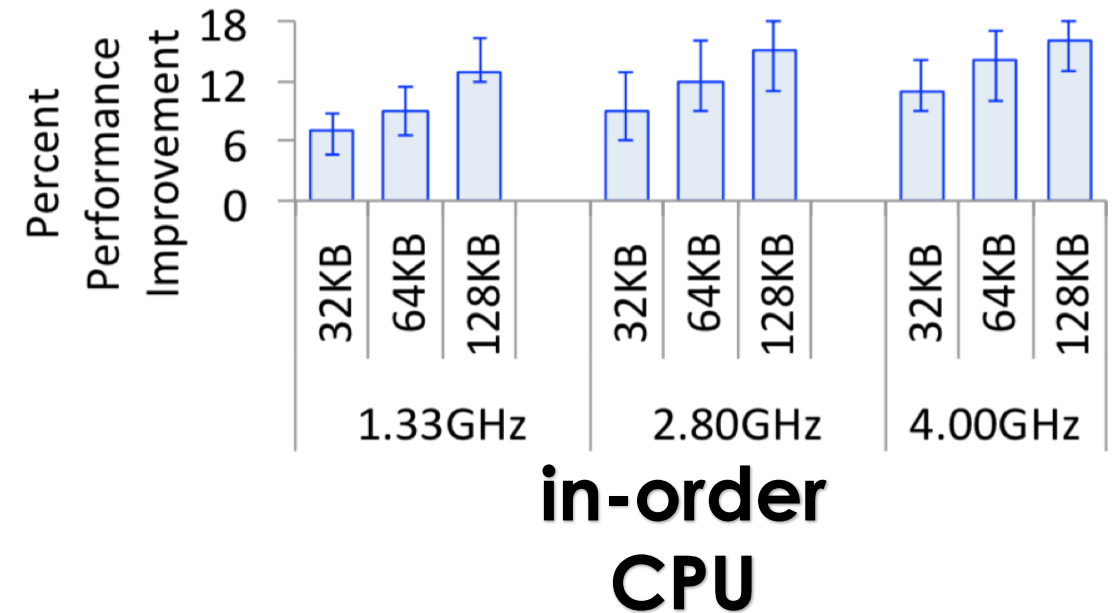
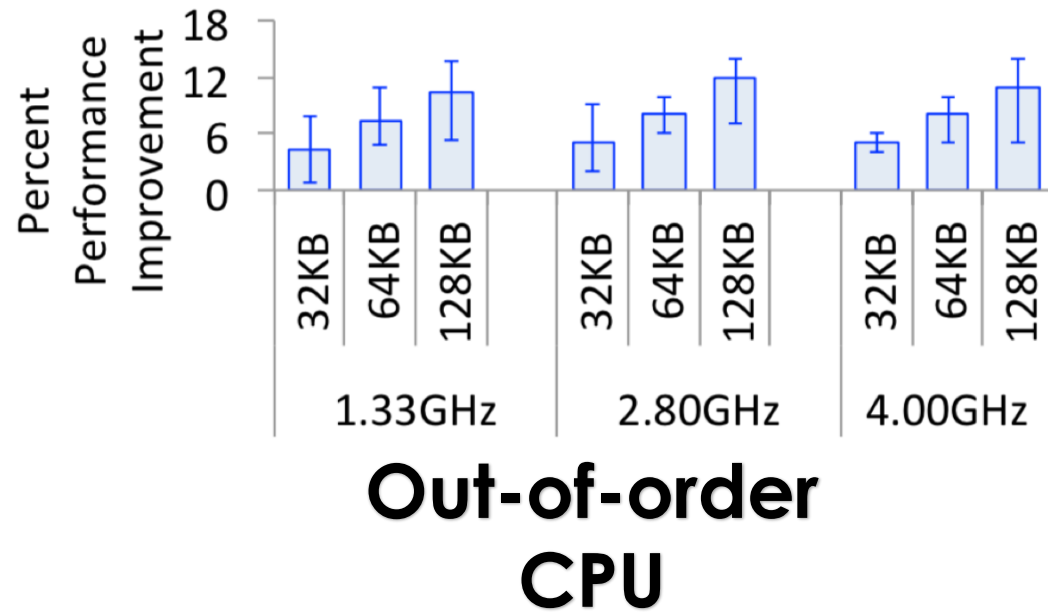
- Motivation
- SEESAW: Concept
- SEESAW: Micro-architecture
- Evaluation Methodology
- **Results**
- Conclusion

SEESAW: Performance improvement



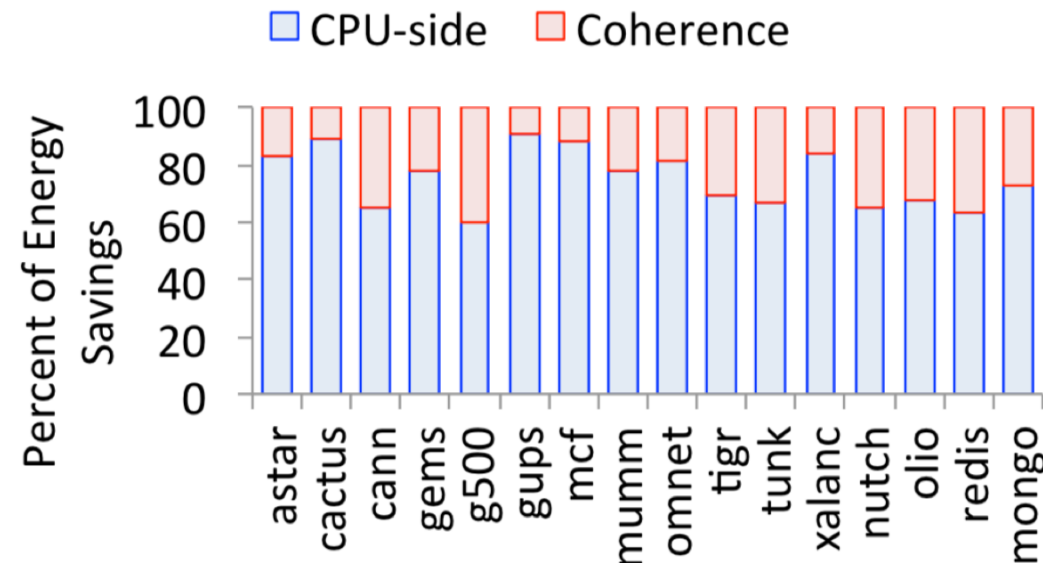
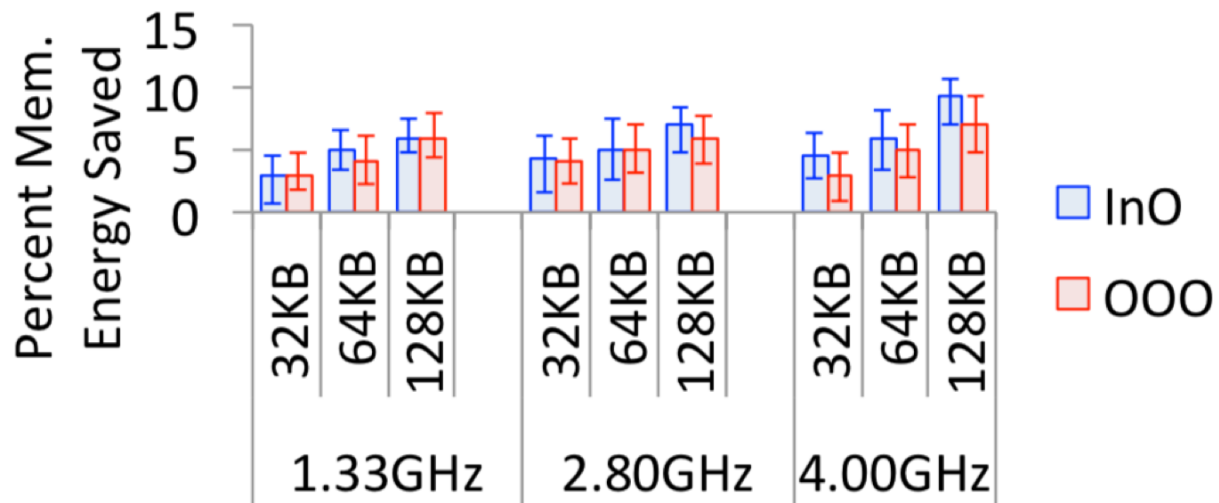
SEESAW observes 3-10% better runtime over baseline

SEESAW: Performance improvement



~10% performance improvement
for 64kB cache in OoO CPUs

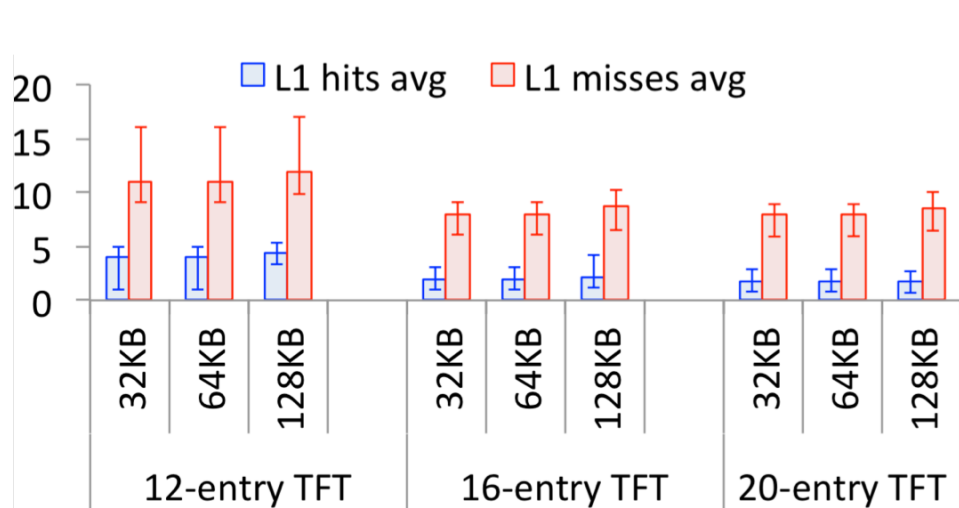
SEESAW: Energy savings



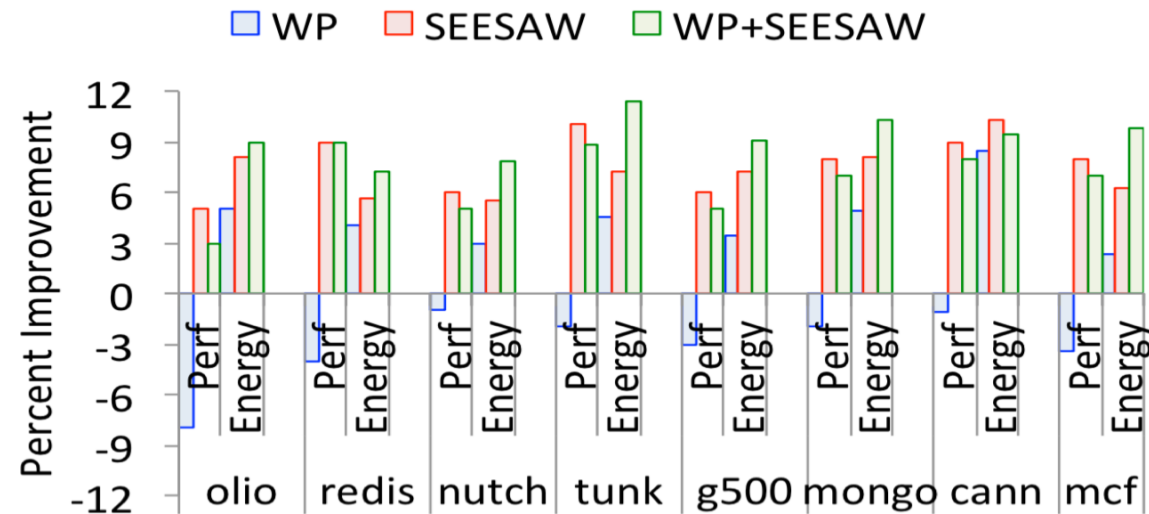
10-20% more energy savings over CPUs using baseline VIPT caches!

Approx. one-third of energy savings from coherence

SEESAW: TFT analysis and Way-Prediction



TFT Analysis



SEESAW + Way-prediction










16-entry TFT drives miss-rate under 10%

SEESAW+WP shows symbiotic behavior

Outline

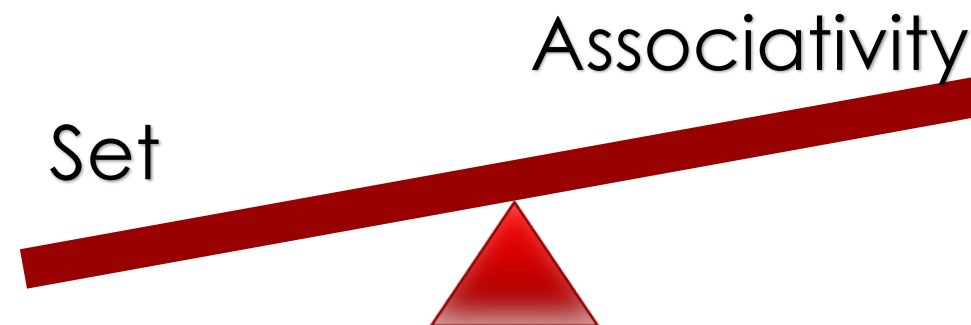
- Motivation
- SEESAW: Concept
- SEESAW: Micro-architecture
- Evaluation Methodology
- Results
- **Conclusion**

Revisiting L1 Cache Characteristic

	Ideal-Cache	VIPT-Cache	SEESAW Cache
Fast lookup			
High hit-rate			
Energy Efficiency			

SEESAW: Conclusion

- L1 caches are optimized for latency
 - VIPT imposes indirect restriction on number of sets in a L1 cache, increasing associativity
 - There is non-linear relation between associativity and access latency/energy of the L1 cache
- Superpages are often used in modern OSes
 - SEESAW provides low-associative access to superpages, providing both latency and energy benefits
 - Up to 10 % performance improvement and 20 % energy reduction in modern workloads
- SEESAW has extremely low-overhead and is readily implementable



Thank you!