University of Pittsburgh

STD-TLB: A STT-RAM-based Dynamically-configurable Translation Lookaside Buffer for GPU Architectures

Xiaoxiao Liu, Yong Li, Yaojun Zhang,
A.K. Jones, Yiran Chen
University of Pittsburgh

Introduction

- GPUs begin to have TLB to support virtual memory addressing.

- The performance of GPUs is more sensitive to the capacity of TLBs:
  - heavier memory accesses;
  - Limited by large SRAM cell area.

- In this work:
  - Proposed a STT-RAM-based dynamically-configurable TLB (STDTLB) by leveraging differential sensing technique;
  - Reconfigurable based on real-time application need.

Outline

- TLBs in GPU
- STT-RAM and differential Sensing
- STT-RAM-based dynamically-configurable TLB (STDTLB)
- Evaluation and Comparison
- Conclusion

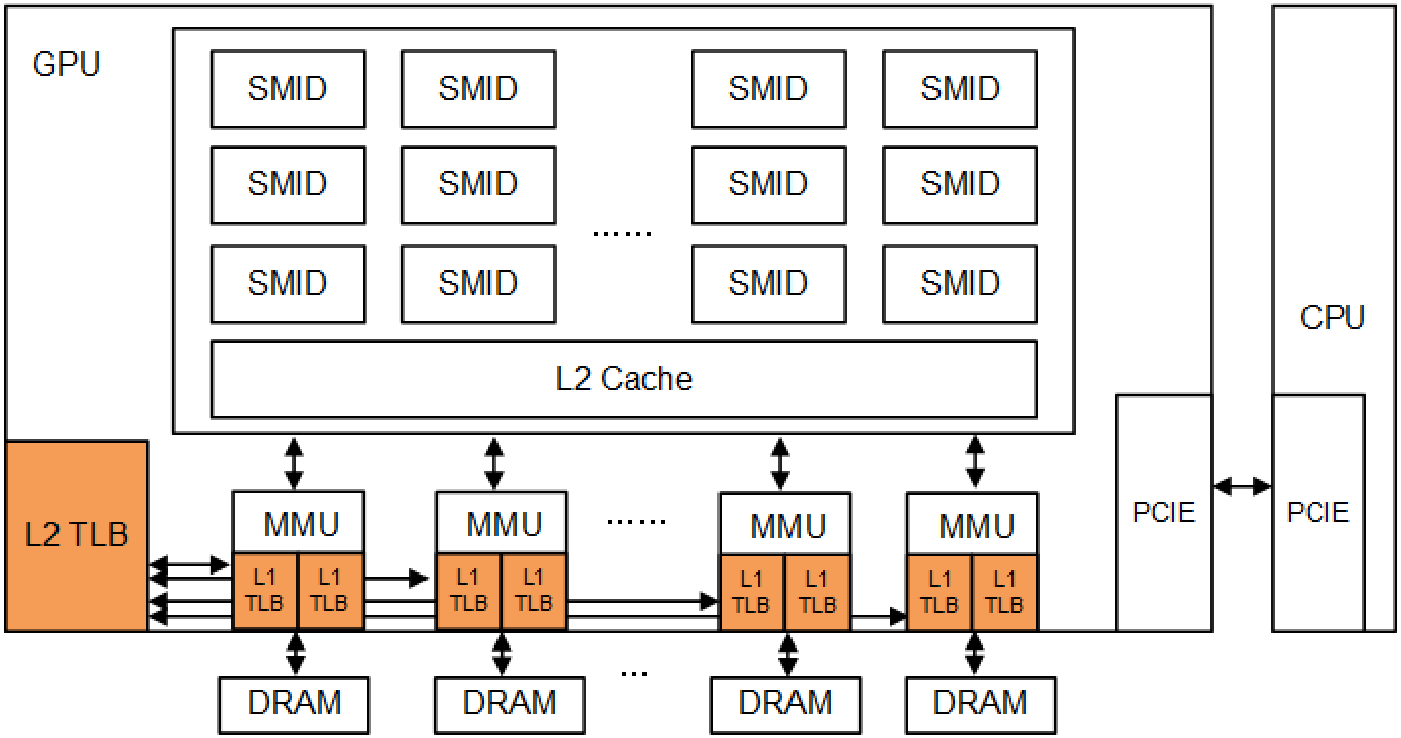
TLBs in GPU

- **Translation Lookaside Buffer(TLB)**
  - **Store virtual-to-physical page addresses;**
  - **Speedup address translation.**

- **GPU begins to support unified memory space:**
  - **Nvidia’s Fermi architecture**
  - **AMD’s Graphic Core Next(GCN) architecture**

TLBs in GPU

- **Overview of memory hierarchy in GPU**



TLBs in GPUs

- **Compared to CPUs, significantly larger TLB is required to retain enough physical page addresses and fast to achieve a high address translation performance.**

Processor		Memory	L2 TLB
GPU	GT200 (Fermi)	Texture	4096 entries
		Global	8192 entries
CPU	I7	Global	512 entries

GPU TLB capacity is limited by the large cell area of SRAM!

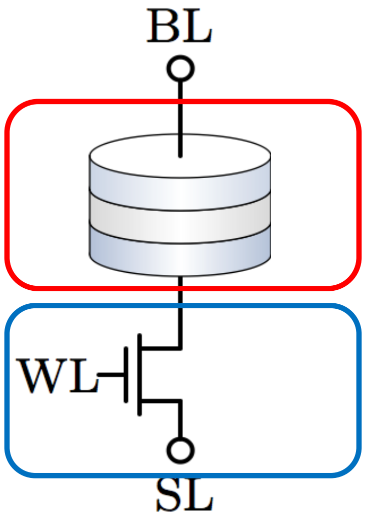
Need new technologies!

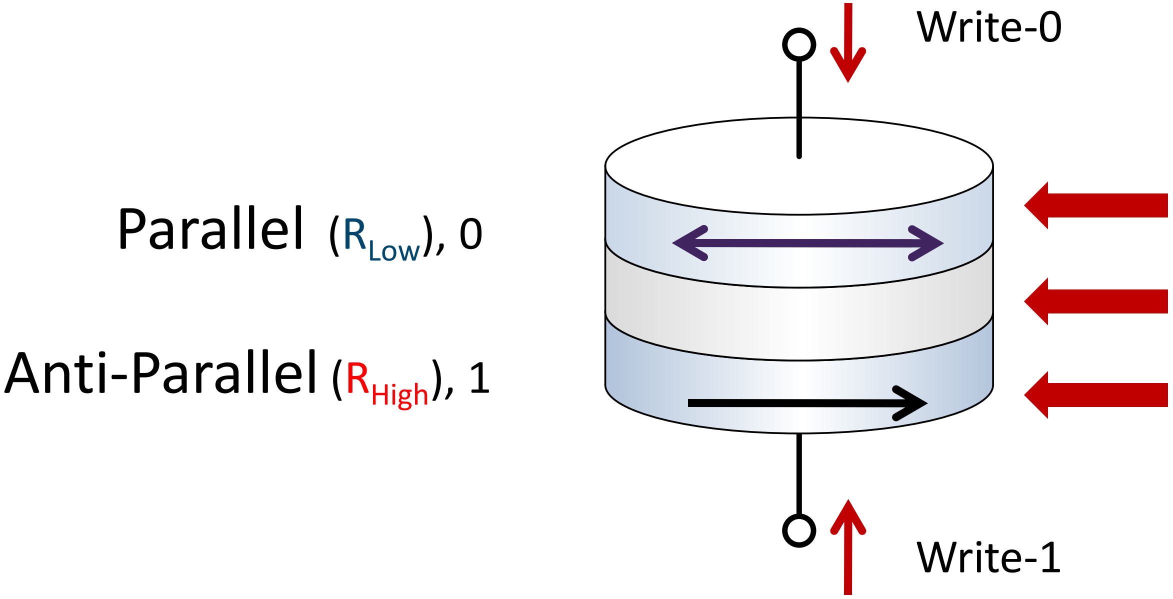
Outline

- TLBs in GPU
- STT-RAM and differential Sensing
- STT-RAM-based dynamically-configurable TLB (STDTLB)
- Evaluation and Comparison
- Conclusion

STT-RAM Basics

- **STT-RAM Cell:**
  - **Transistor and MTJ (Magnetic Tunnel Junction);**
  - **Denoted as standard STT-RAM (1T1J).**
- **MTJ:**
  - **Free Layer and Ref. Layer;**
  - **Read: Direction → Resistance;**
  - **Write: Current → Direction.**





Free Layer

Barrier

Reference Layer

STT-RAM Basics

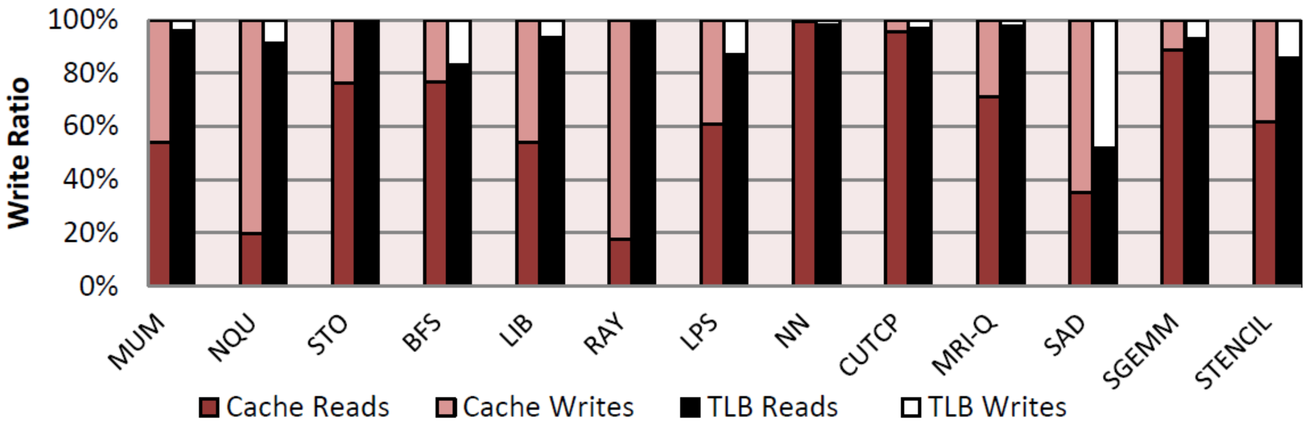
- Comparison with SRAM

	SRAM	STT-RAM
Area	✗	✓✓
Read Speed	✓	✗
Write Speed	✓	✗✗

- Unbalanced R/W access makes TLB well suited to be built with STT-RAM.

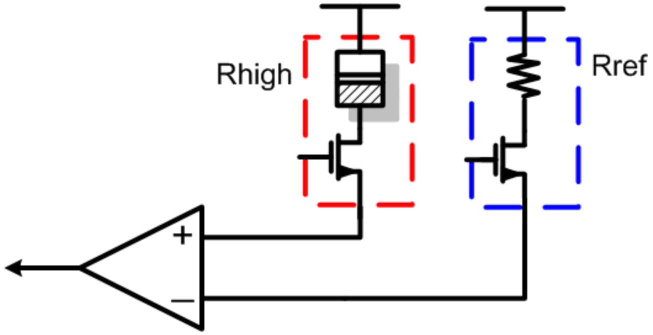
Need faster read!

Don’t care write!

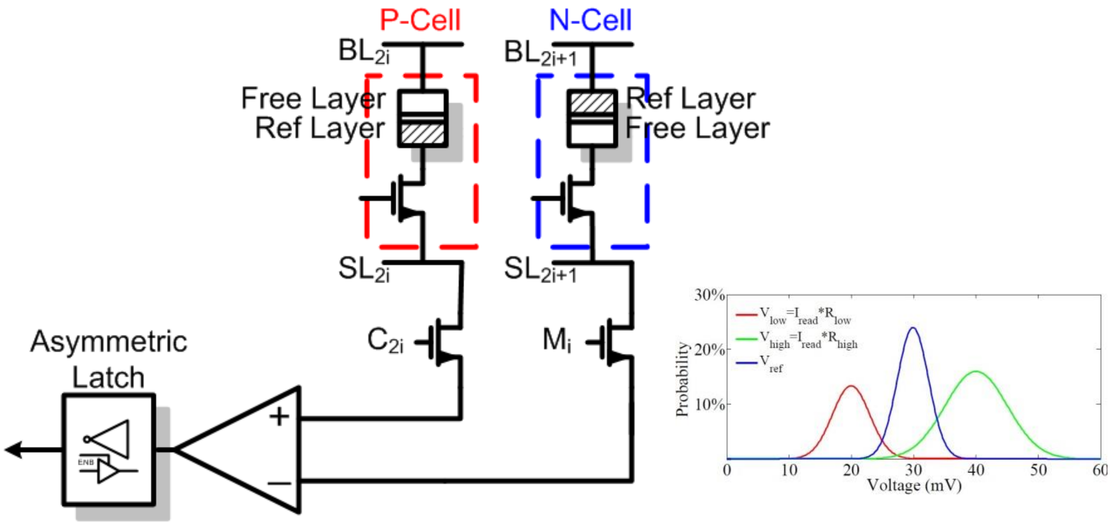


STT-RAM with Differential Sensing(2T2J)

- **Write:**
  - **Inversion of the data is written into an adjacent cell.**
- **Read:**
  - **Compare the resistance of these two complimentary cells instead of reference cell**
  - **Increasing sensing margin to get faster read by scarifying capacity**



Standard STT-RAM



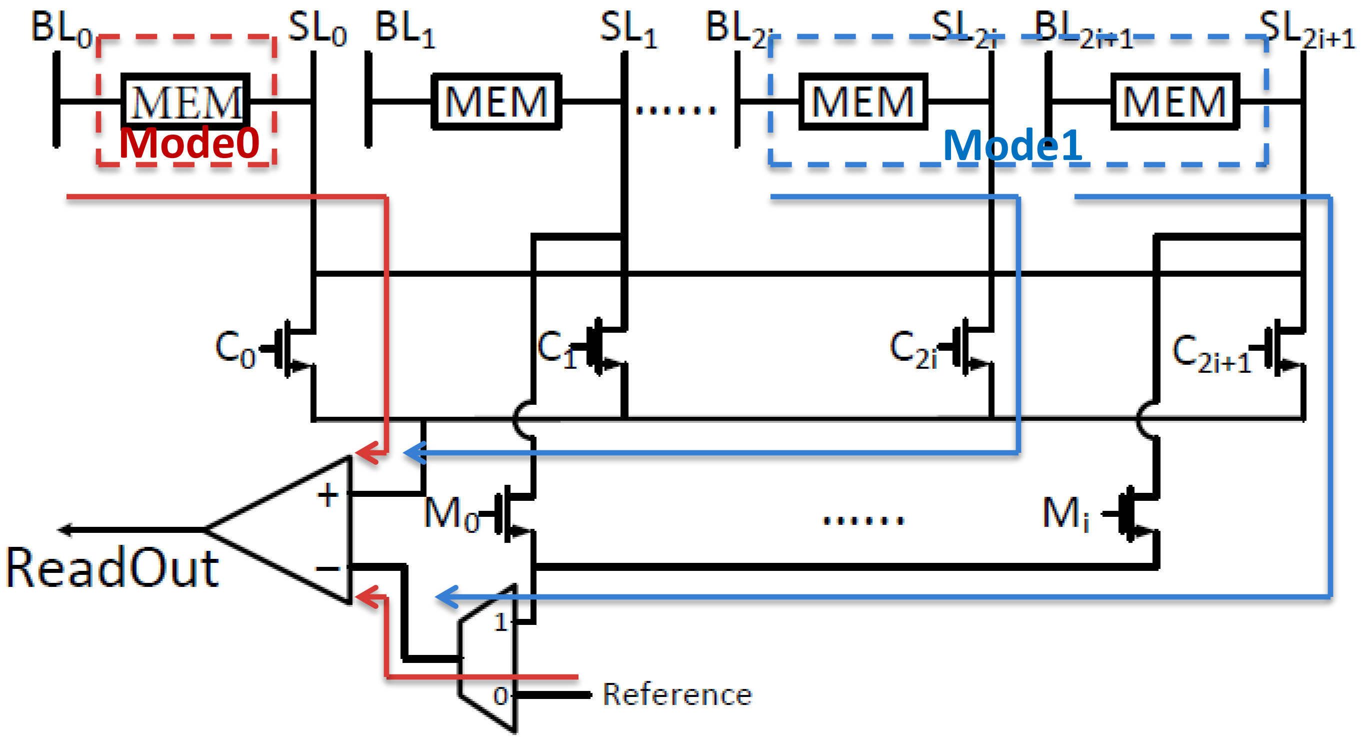
Differential Sensing

Outline

- TLBs in GPU
- STT-RAM and differential Sensing
- STT-RAM-based dynamically-configurable TLB (STDTLB)
- Evaluation and Comparison
- Conclusion

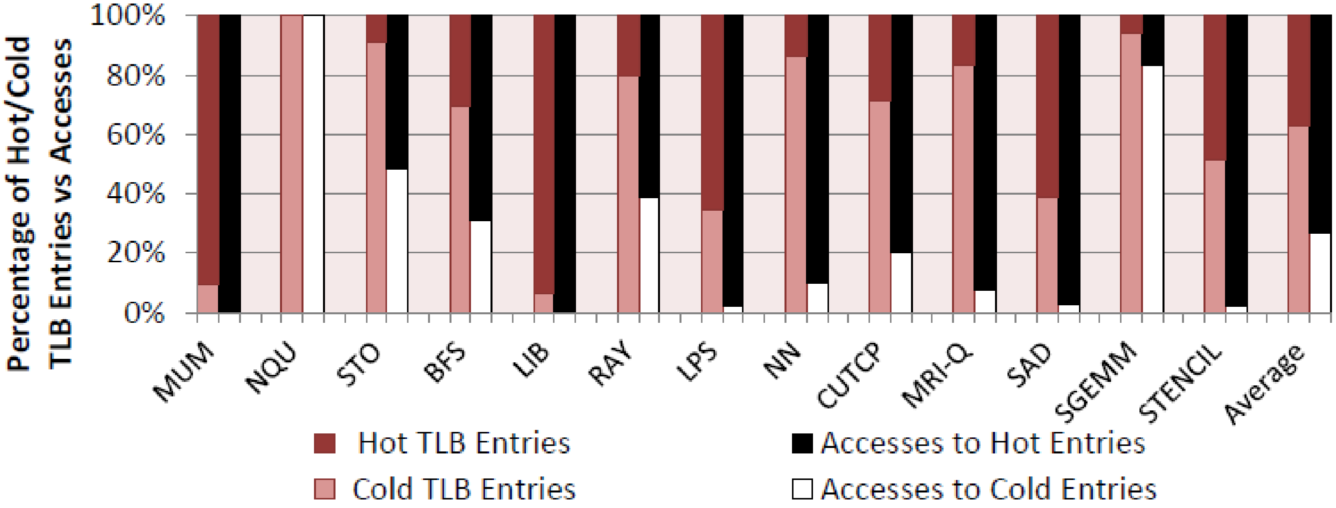
STT-RAM-based dynamically-configurable TLB

- Reconfigurable differential sensing circuit:
  - Mode0: 1T1J mode -> High capacity -> infrequent access
  - Mode1: 2T2J mode -> High performance -> frequent access



STT-RAM-based dynamically-configurable TLB

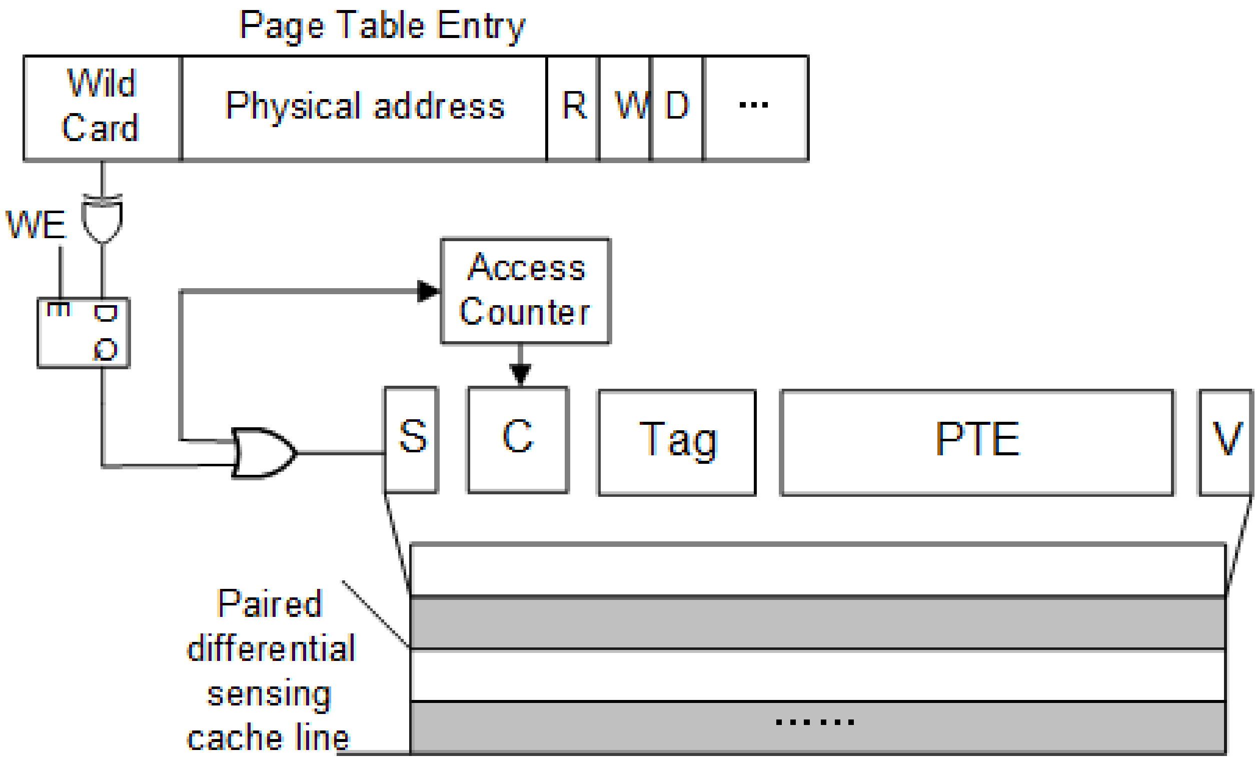
- **Read access to TLB is very unbalanced:**
  - **Group TLB entries as "Hot/Cold" based on their access frequency;**
  - **More than 75% read happens in "Hot" entries.**



- **"Hot/Cold" entries should store in 2T2J/1T1J separately to achieve better performance.**

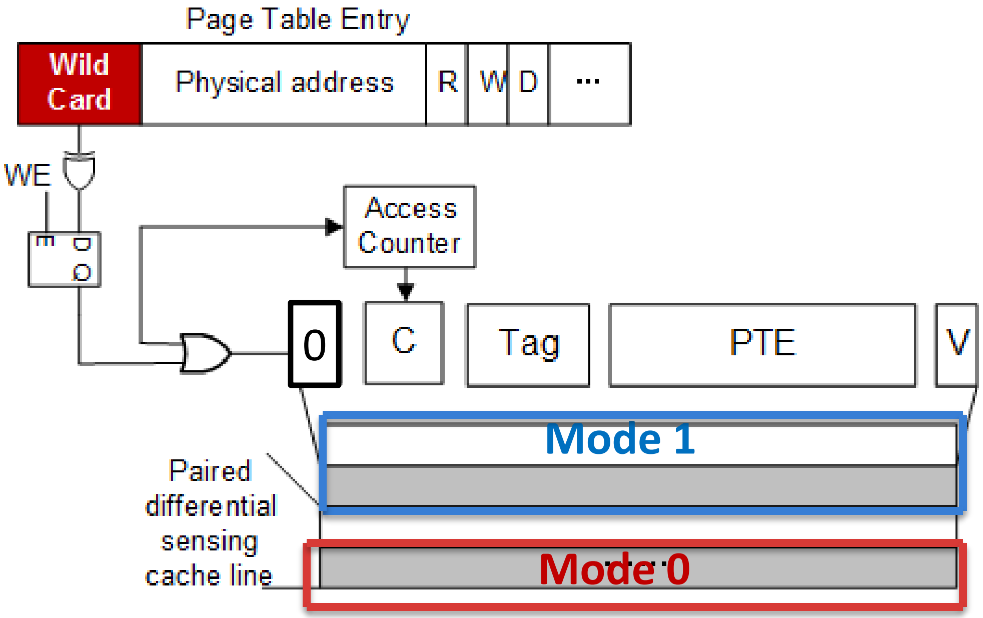
STT-RAM-based dynamically-configurable TLB

- **Organization of STD-TLB design**



STT-RAM-based dynamically-configurable TLB

- **STD-TLB Working Mode Management:**
  - **Initialization: Big pages are always “hot”.**
    - **If WildCard>0, write the entry in high performance mode(2T2J)**
    - **If WildCard=0, program in high capacity mode(1T1J)**

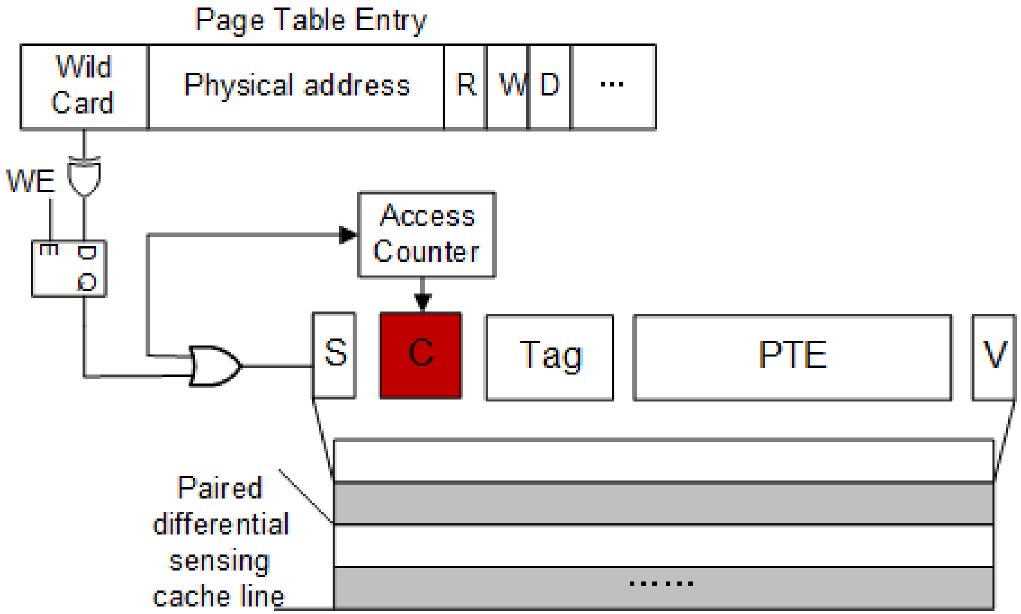


$$\text{Pagesize} = 4KByte \times 2^{WildCard}$$

WildCard > 0	WildCard = 0
⬇	⬇
Big page	Small page
⬇	⬇
Mode1(2T2J)	Mode0(1T1J)

STT-RAM-based dynamically-configurable TLB

- **STD-TLB Working Mode Management:**
  - **Dynamic reconfiguration:**
    - **Frequent-access small page: Upgrade to high performance mode (2T2J) when Counter reaches Threshold**
    - **“Demote” instead of “eviction”**

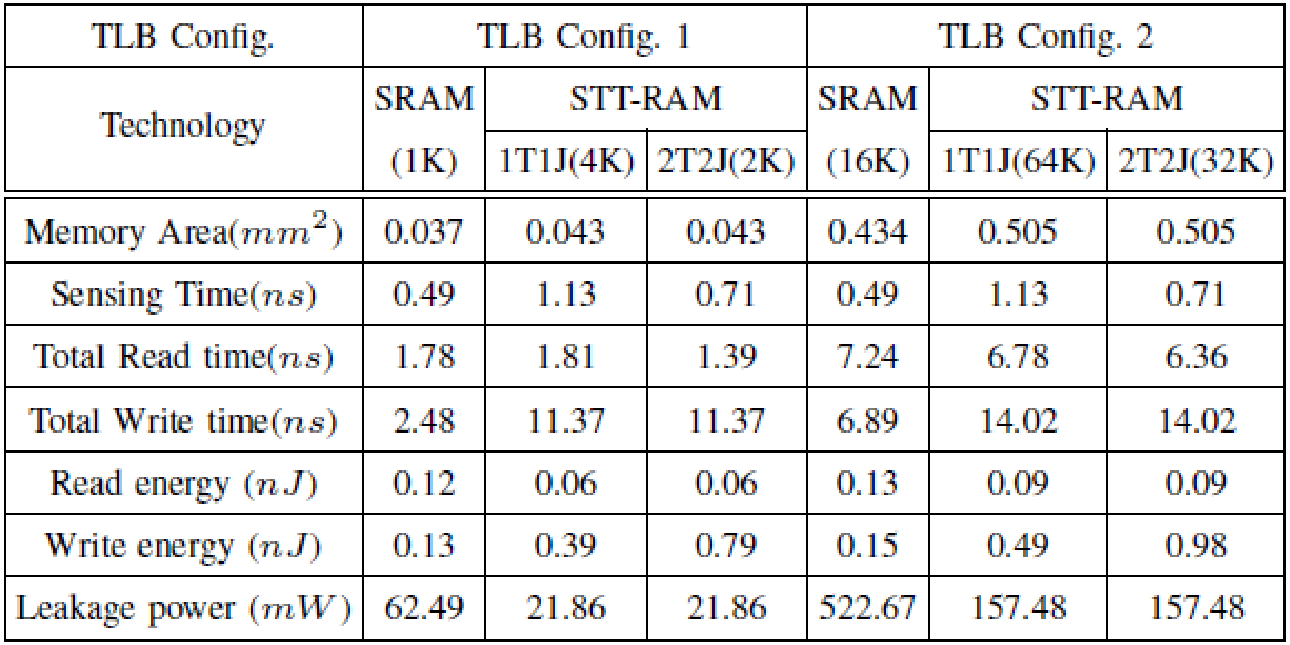


Outline

- TLBs in GPU
- STT-RAM and differential Sensing
- STT-RAM-based dynamically-configurable TLB (STDTLB)
- Evaluation and Comparison
- Conclusion

Evaluation and Comparison

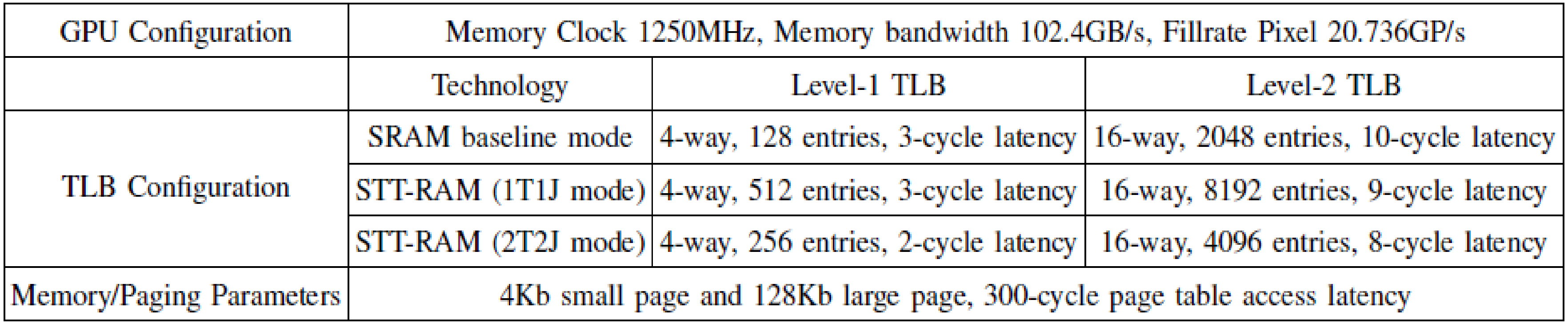
- **TLB parameters:**
  - **45nm technology**
  - **Based on CACTI and SPICE simulation**



TLB Config.	TLB Config. 1			TLB Config. 2		
Technology	SRAM	STT-RAM		SRAM	STT-RAM	
	(1K)	1T1J(4K)	2T2J(2K)	(16K)	1T1J(64K)	2T2J(32K)
Memory Area(mm^2)	0.037	0.043	0.043	0.434	0.505	0.505
Sensing Time(ns)	0.49	1.13	0.71	0.49	1.13	0.71
Total Read time(ns)	1.78	1.81	1.39	7.24	6.78	6.36
Total Write time(ns)	2.48	11.37	11.37	6.89	14.02	14.02
Read energy (nJ)	0.12	0.06	0.06	0.13	0.09	0.09
Write energy (nJ)	0.13	0.39	0.79	0.15	0.49	0.98
Leakage power (mW)	62.49	21.86	21.86	522.67	157.48	157.48

Evaluation and Comparison

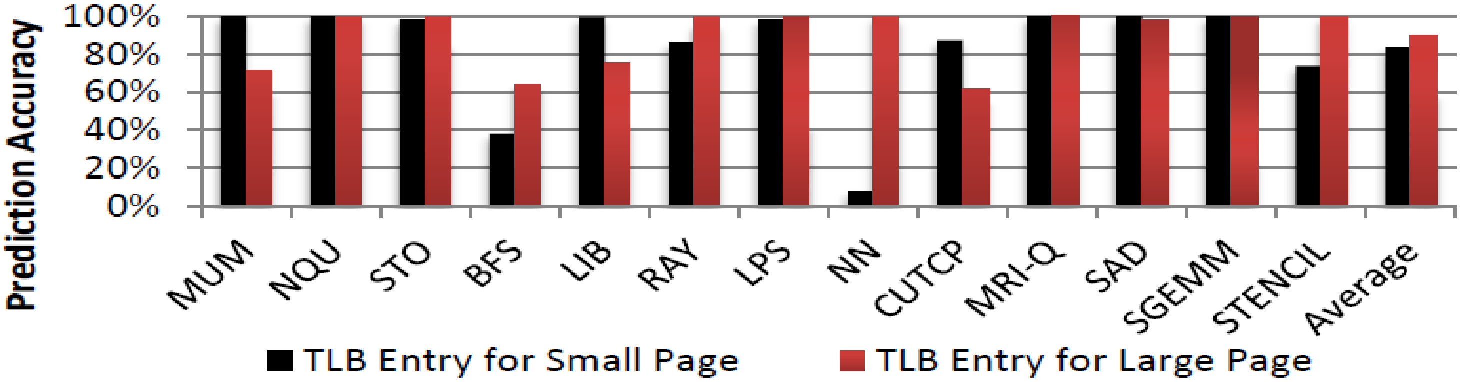
- System configuration
  - GPGPUsim simulator
  - Nvidia Quadro FX5800 as baseline architecture



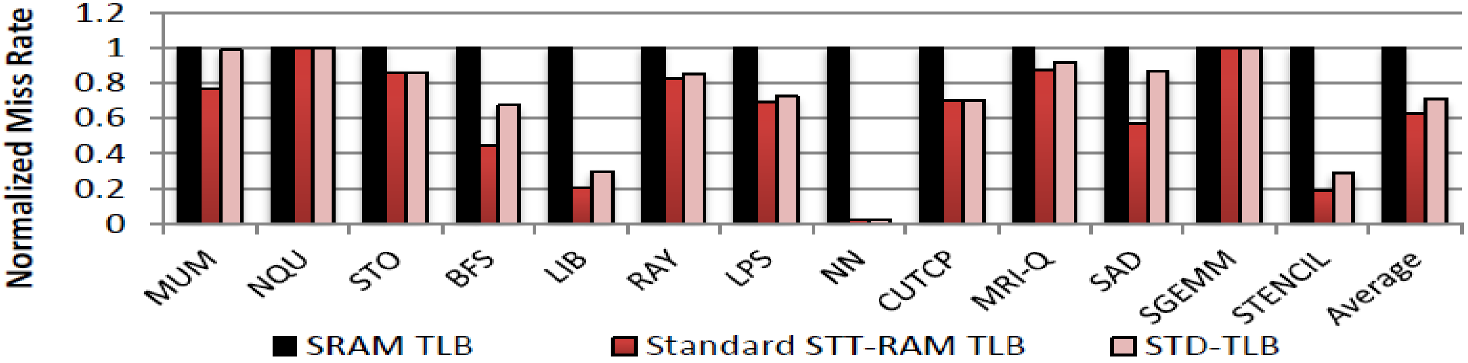
GPU Configuration	Memory Clock 1250MHz, Memory bandwidth 102.4GB/s, Fillrate Pixel 20.736GP/s		
	Technology	Level-1 TLB	Level-2 TLB
TLB Configuration	SRAM baseline mode	4-way, 128 entries, 3-cycle latency	16-way, 2048 entries, 10-cycle latency
	STT-RAM (1T1J mode)	4-way, 512 entries, 3-cycle latency	16-way, 8192 entries, 9-cycle latency
	STT-RAM (2T2J mode)	4-way, 256 entries, 2-cycle latency	16-way, 4096 entries, 8-cycle latency
Memory/Paging Parameters	4Kb small page and 128Kb large page, 300-cycle page table access latency		

Evaluation

- **Mode Configuration Accuracy(> 86%)**
  - **Our mode selection mechanism captures averagely 83% of ”cold” small pages and 90% of ”hot” large pages**

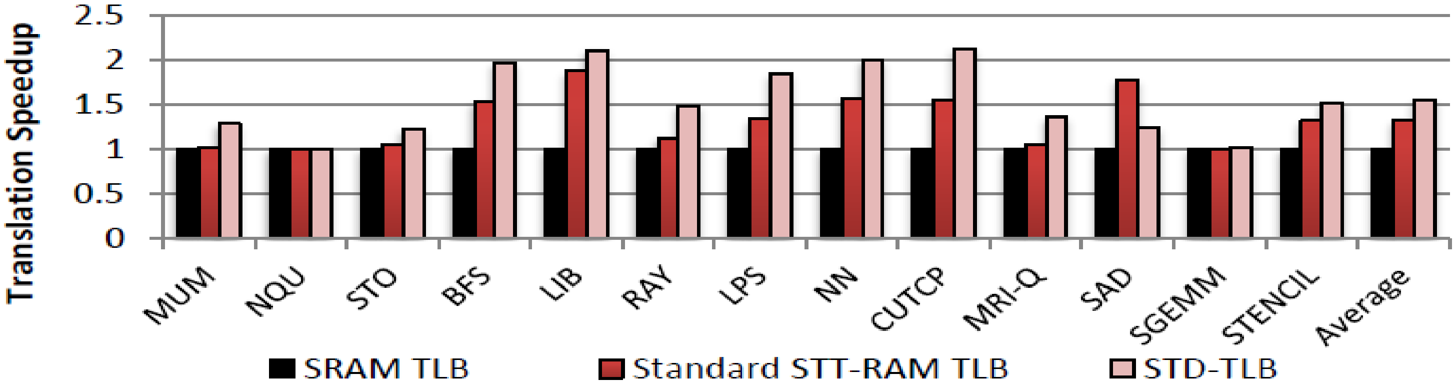


- **TLB miss rate**
  - **Standard STT-RAM TLB reduces 38% of miss rate, while STD-TLB reduces 30%**

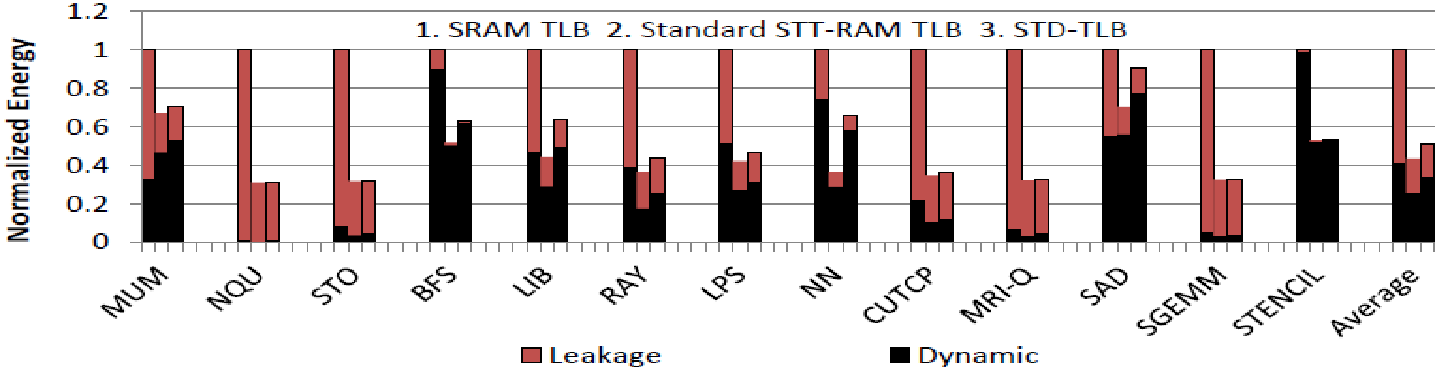


Evaluation

- **Performance**
  - **TLB translation performance improved 32% by standard STT-RAM TLB and 55% by STD-TLB**

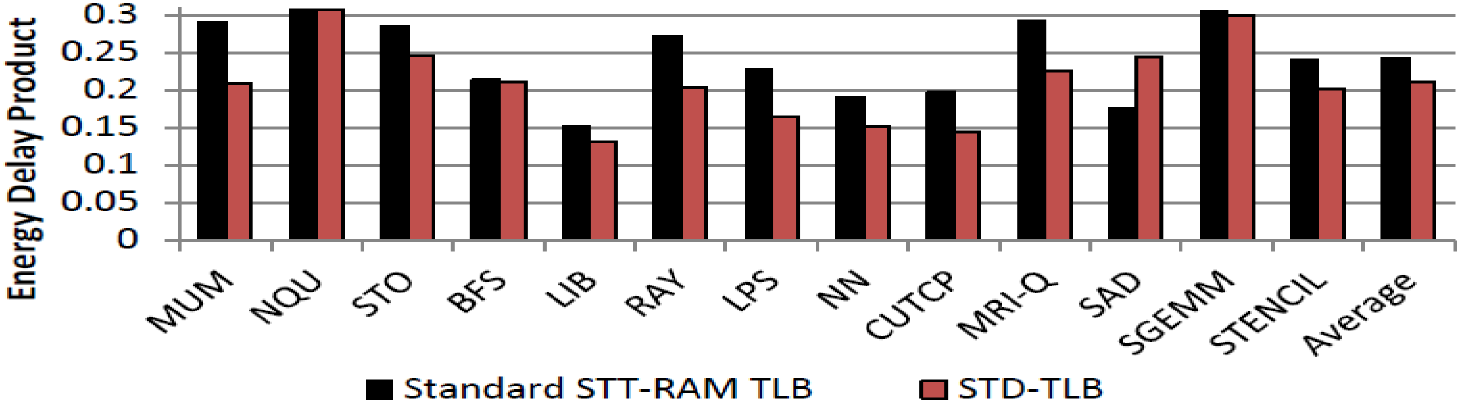


- **Energy**
  - **Standard STT-RAM TLB/STD-TLB achieve 57%/49% energy savings**

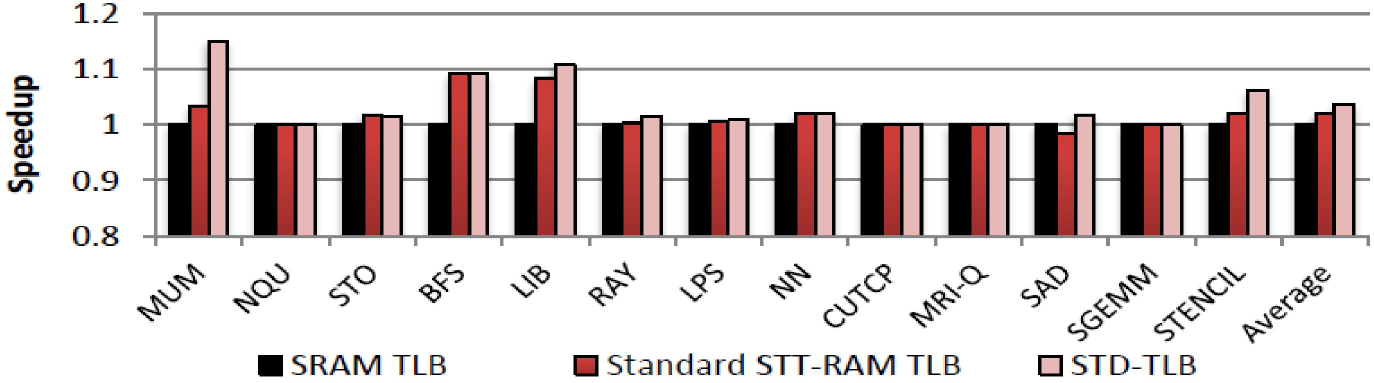


Evaluation

- **Energy Delay Product**
  - **Improve 75%/80% by standard STT-RAM TLB/STD-TLB**



- **Overall System Performance Speedup**
  - **STD-TLB achieves 4% and 2% system performance speedup compared to SRAM TLB and standard STT-RAM TLB**



Outline

- TLBs in GPU
- STT-RAM and differential Sensing
- STT-RAM-based dynamically-configurable TLB (STDTLB)
- Evaluation and Comparison
- Conclusion

Conclusion

- **Proposed the use of STT-RAM in TLB for new virtually addressed GPUs.**

- **Besides standard STT-RAM-based TLB, we also presented a novel STT-RAM-based dynamically-configurable TLB (STD-TLB) that leverages high read speed of STT-RAM differential sensing and dynamic configuration to further improve performance of the heavily accessed pages.**

- **Expect STD-TLB to provide a dramatic system benefit in modern GPGPU and Heterogeneous system.**

THANKS.

QUESTION?