A Hybrid Buffer Design with STT-MRAM for On-Chip Interconnects

Department of Computer Science & Engineering

Hyunjun Jang, Baik Song An, Nikhil Kulkarni, Ki Hwan Yum, and Eun Jung Kim

Dept. of Computer Science & Engineering

Texas A&M University

Outline

- Background of NoC
- Motivation of selecting STT-MRAM
- Challenges in using STT-MRAM
- Approaches
  - Hybrid Buffer Design
  - Simple & Lazy Migration Scheme
- Performance and Power Evaluation
- Conclusions

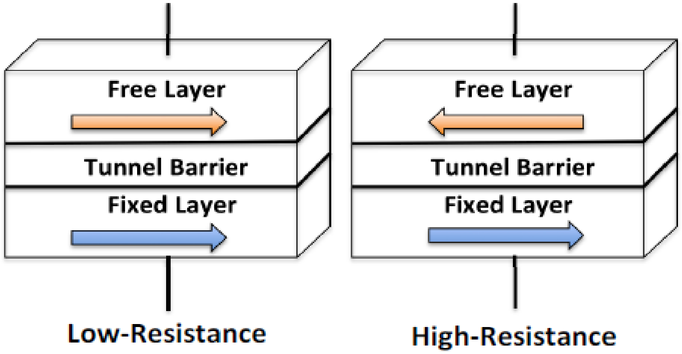
Networks-on-Chip (NoCs)

- NoCs for Large-Scale Chip Multi-Processors (CMPs)
- Packet-Switching Networks
  - Switch-based interconnects
    - Scalable
    - More suitable for large-scale Multi-Processor Systems

But, Power & Area Budgets in On-Chip Networks are very Limited

Why STT-MRAM in NoCs

- Near-zero leakage power compared to SRAM or DRAM
- Much higher density than SRAM (more than 4xs)
- Much higher endurance compared to other Non-volatile memories e.g., PCM, or Flash
  - Tolerate much more frequent write accesses



STT-MRAM bit storage (MTJ)

Weaknesses of STT-MRAM

- Long write latency compared to SRAM
  - More than 10 cycles
- High write power compared to SRAM
  - More than 8xs

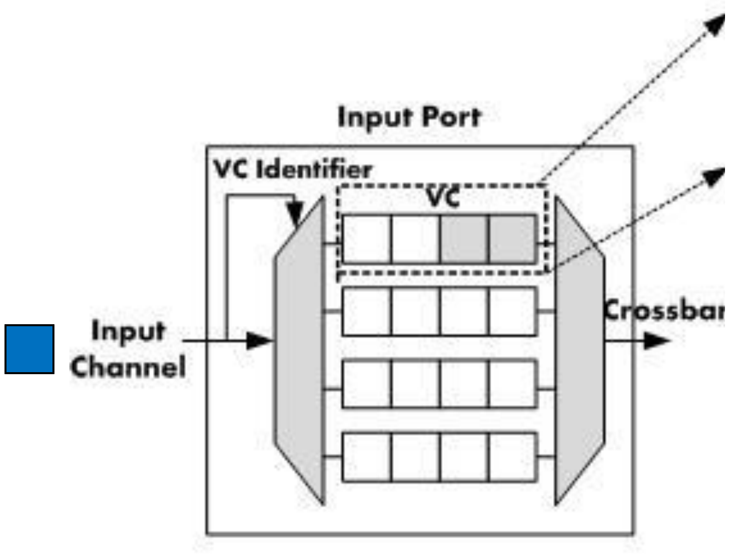
To exploit the benefits of STT-MRAM, these challenges should be addressed first

Approaches

- Hiding the Long Write Latency, while Maximizing Area Efficiency
  - SRAM + STT-MRAM Hybrid Buffer Design
- Sacrificing the Retention Time
  - From 10yrs to 10ms
  - Accordingly, latency also changes: 3.2 ns →1.8ns, which is corresponding to 6 cycles in 3GHz clock frequency
- Reducing the Dynamic Write Power
  - Adaptive flit migration scheme in hybrid buffer considering current SRAM buffer occupancy

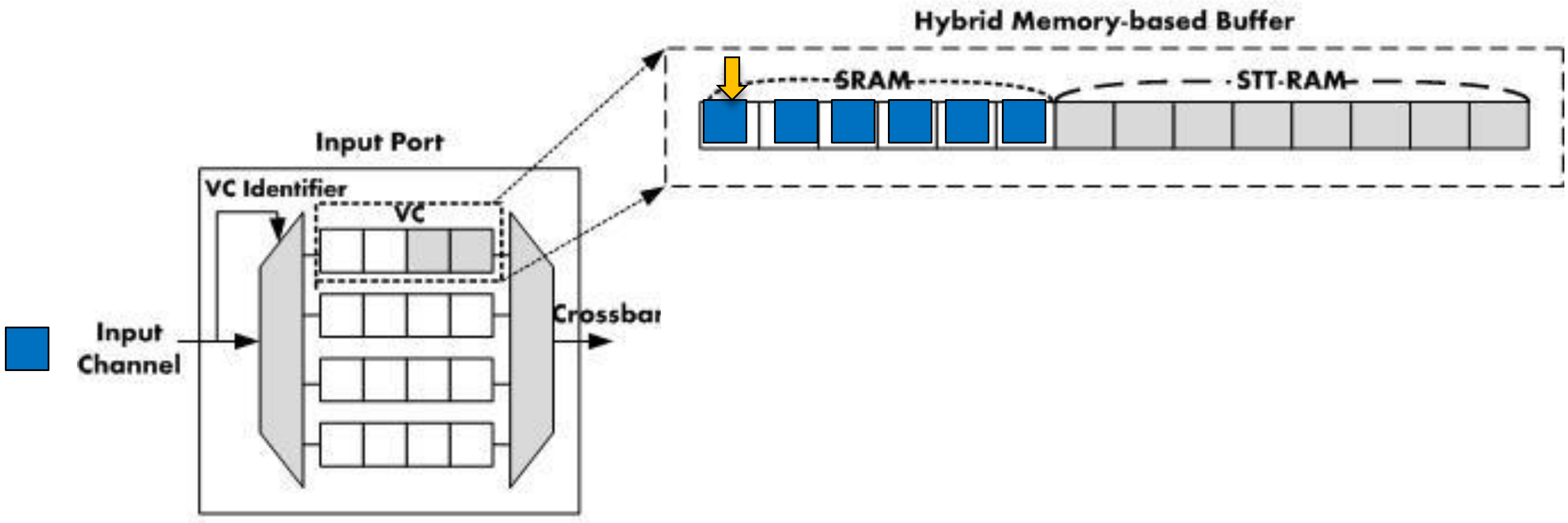
Hybrid Buffer Design

- Hiding the Long Write Latency (write lat = 6cycles)



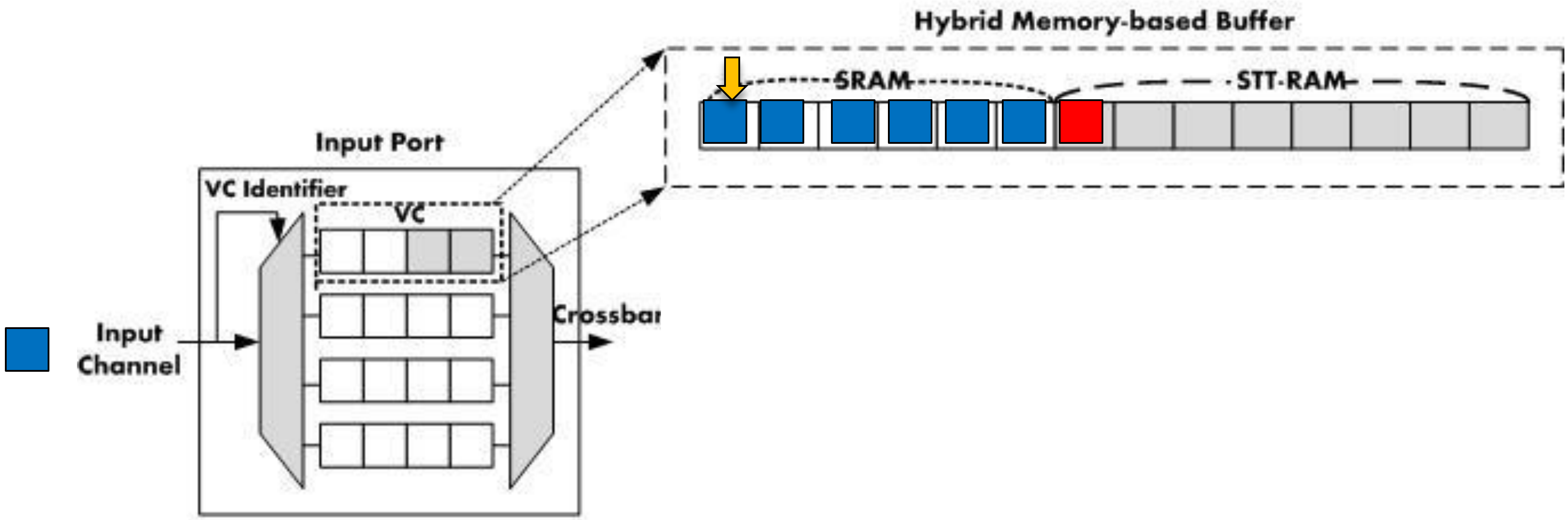
Hybrid Buffer Design

- ❑ Hiding the Long Write Latency (write lat = 6cycles)



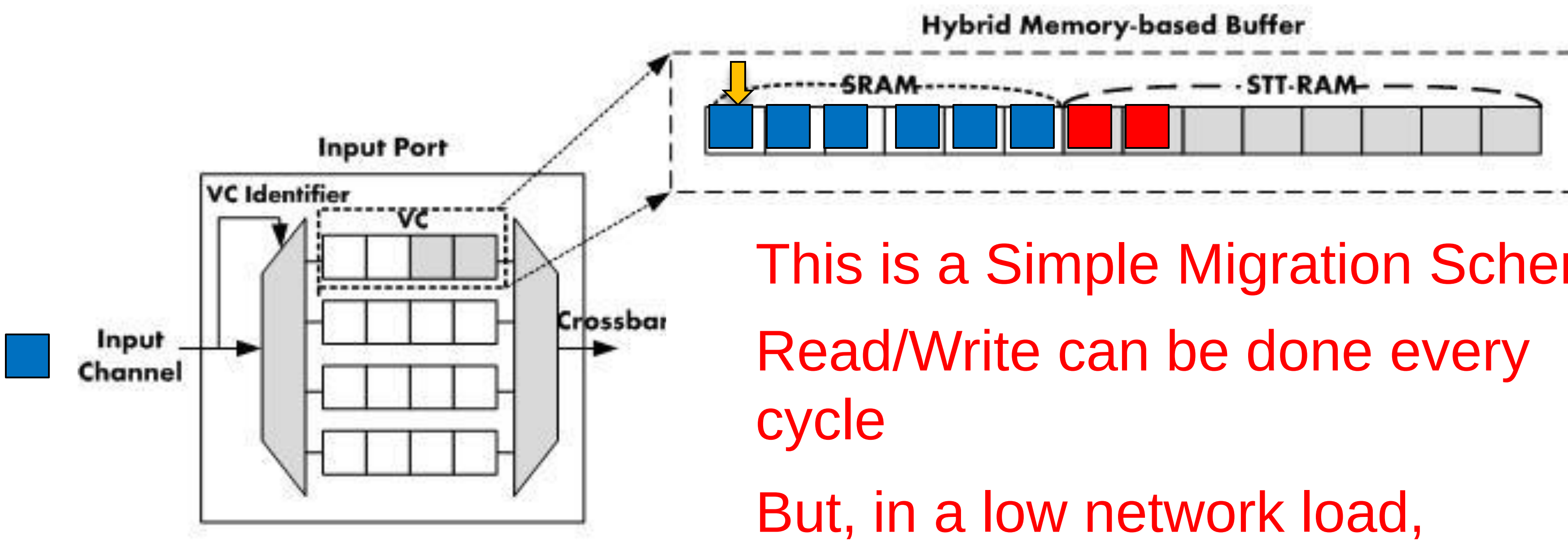
Hybrid Buffer Design

- Hiding the Long Write Latency (write lat = 6cycles)



Hybrid Buffer Design

❑ Hiding the Long Write Latency (write lat = 6cycles)



This is a Simple Migration Scheme

Read/Write can be done every cycle

But, in a low network load, migration energy is unnecessarily wasted

Reducing Dynamic Power Consumption

Department of Computer Science & Engineering

❑ Lazy Migration Scheme

- IF (SRAM Buffer Occupancy >= Threshold)
  - Start migrating flits to STT-MRAM
- ELSE
  - Maintain flits in SRAM

# of flits/ buffer size

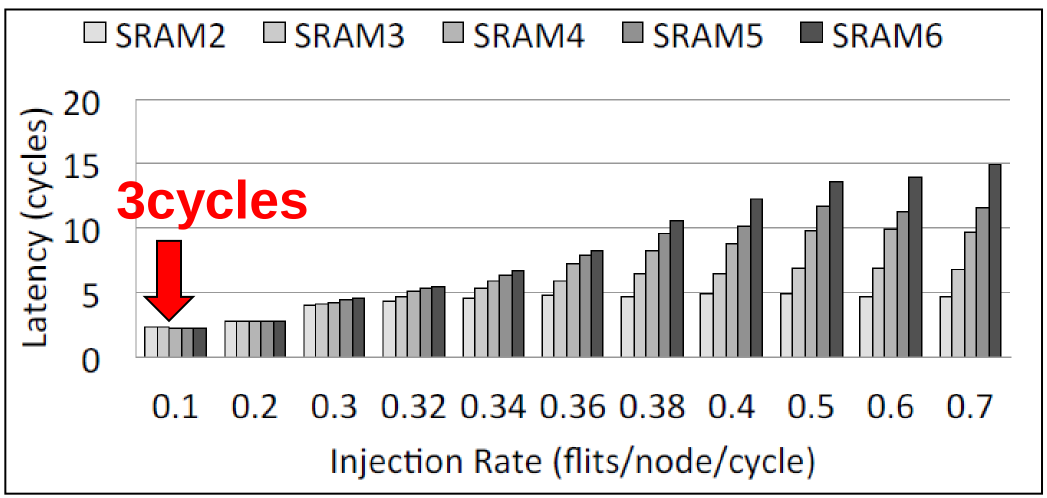
- e.g. threshold in SRAM4 case : 0%, 25%, 50%, 75%

ref. Credit-based Flow Control

- Only considers SRAM buffer in credit management

Front-end SRAM Buffer Size

- In our experiment, Flits written into buffer stay at least 3 cycles in each on-chip router (Intra-router latency)
- It is possible to reduce front-end SRAM from 6 to 3
  - Thus, we can replace more SRAM with STT-MRAM



Various Hybrid Buffer Configurations

Department of Computer Science & Engineering

- STT-MRAM is 4xs denser than SRAM
- Therefore, under the same area budget, 1 SRAM space can be replaced with 4 STT-MRAM space
- So, under the baseline SRAM6 space,
  - SRAM5-STT4
  - SRAM4-STT8
  - SRAM3-STT12
  - SRAM2-STT16

All these 4 different hybrid configurations have same area budget (SRAM6)

Performed experiments to find best hybrid buffer configuration

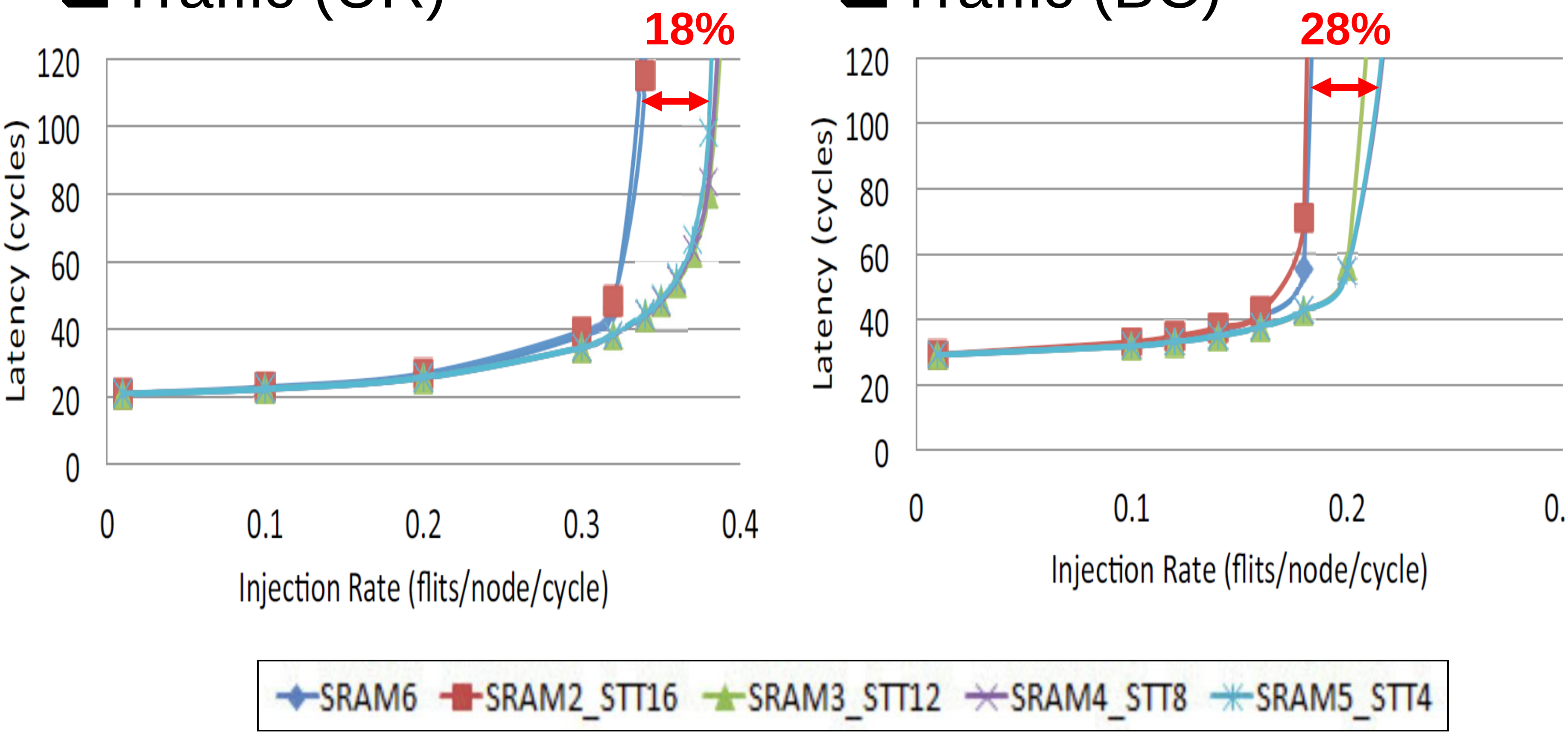
Performance/Power Evaluation

- **Performance Model**: Cycle-accurate on-chip network simulator
  - Models all router pipeline stages in detail
- **Power Model**: Orion for both dynamic and leakage power estimation

Topology	**8×8 Mesh**, 2D-Torus, Flattened BFly
Routing	**XY**, O1TURN
# of VC/Port	**4**
Buffer Depth/VC (Same area budget)	**SRAM6(baseline)**, SRAM5-STT4, SRAM4-STT8, SRAM3-STT12, SRAM2-STT16
Packet Length	**4 flits** (128bits/flit)
Synthetic Traffic, Benchmark	**UR**, BC, NN, Splash-2
SRAM Read, Write Energy	5.25 (pJ/flit), 5.25 (pJ/flit)
SRAM Read, Write Latency	**1cycle for Read and Write**
STT Read, Write Energy	3.826 (pJ/flit), 40.0 (pJ/flit)
STT Read, Write Latency	**1 cycle for Read**, **6 cycles for Write**

Performance Analysis - Different Traffic

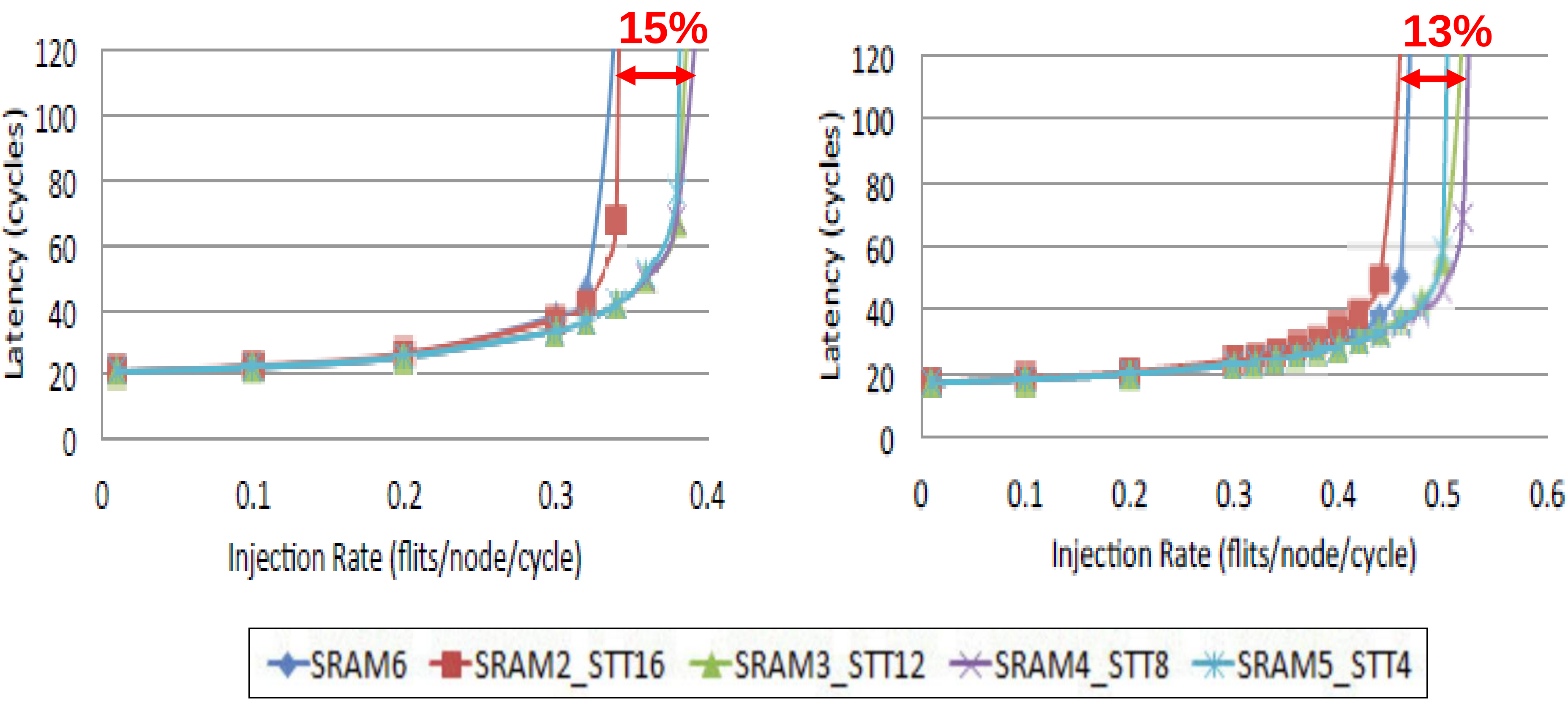
- Traffic (UR)
- Traffic (BC)



Performance Analysis - Different Routing, Topology

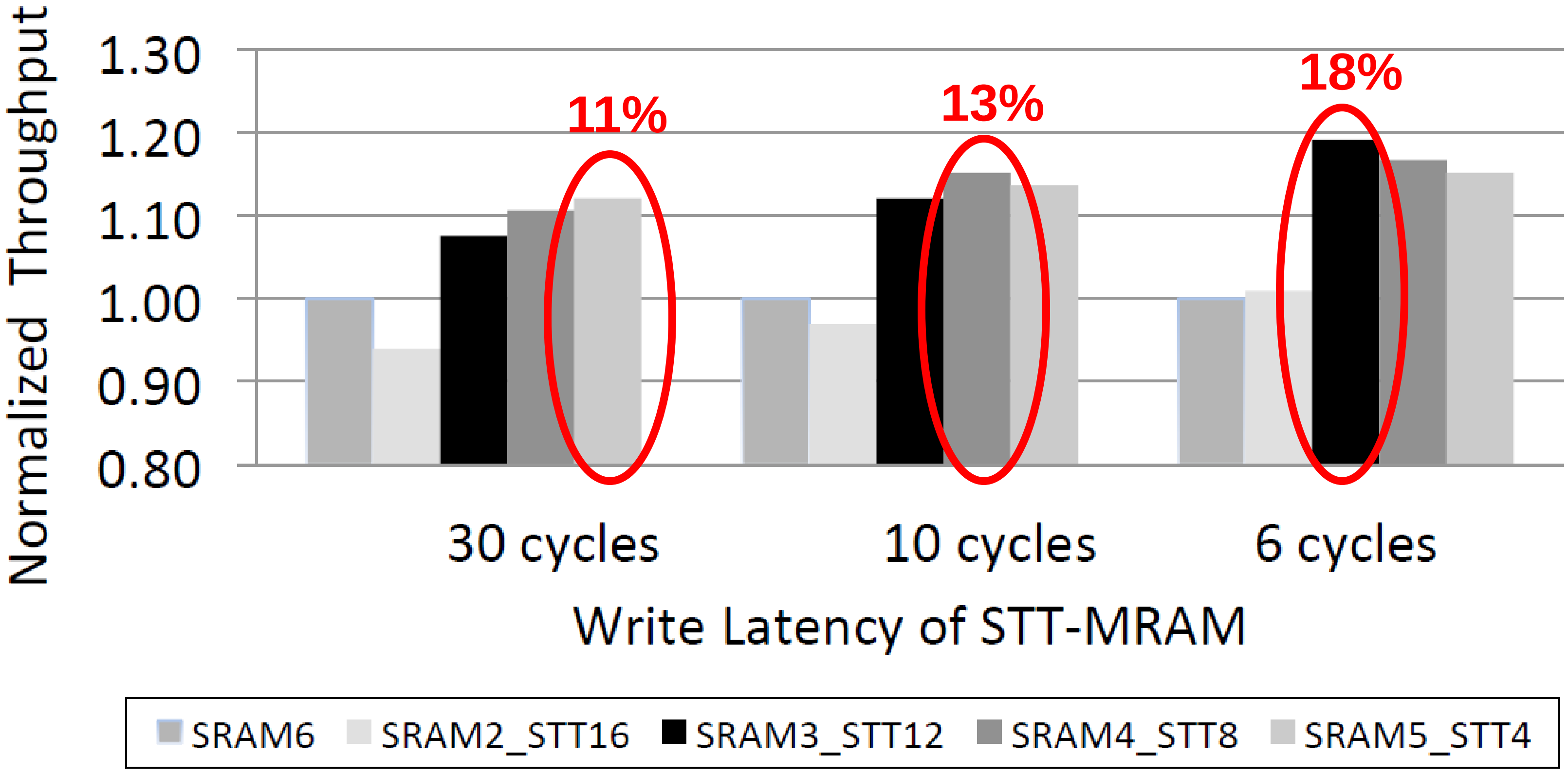
Department of Computer Science & Engineering

- Routing (O1TURN)
- Topology (2D-Torus)



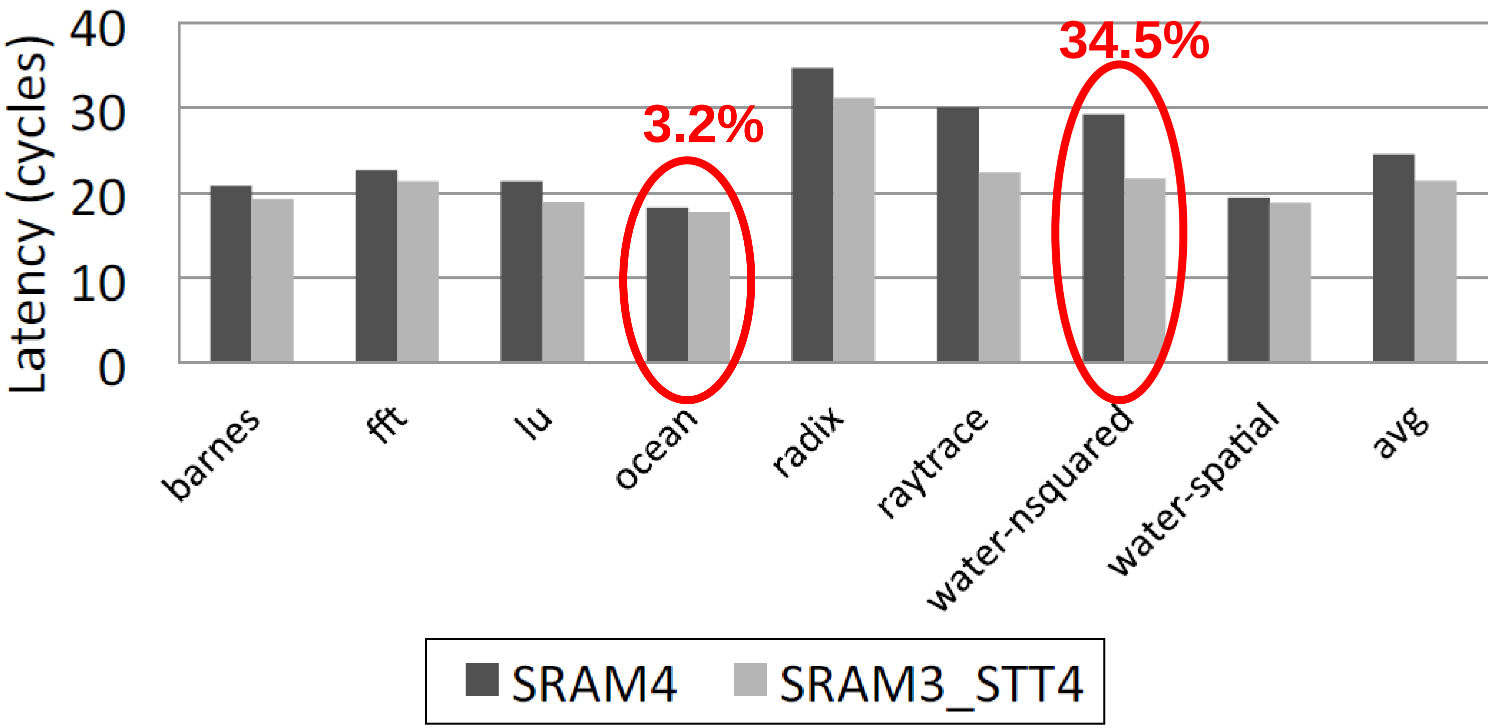
Performance Analysis
- Various STT Write latencies

- Write latencies (30, 10, 6 cycles)



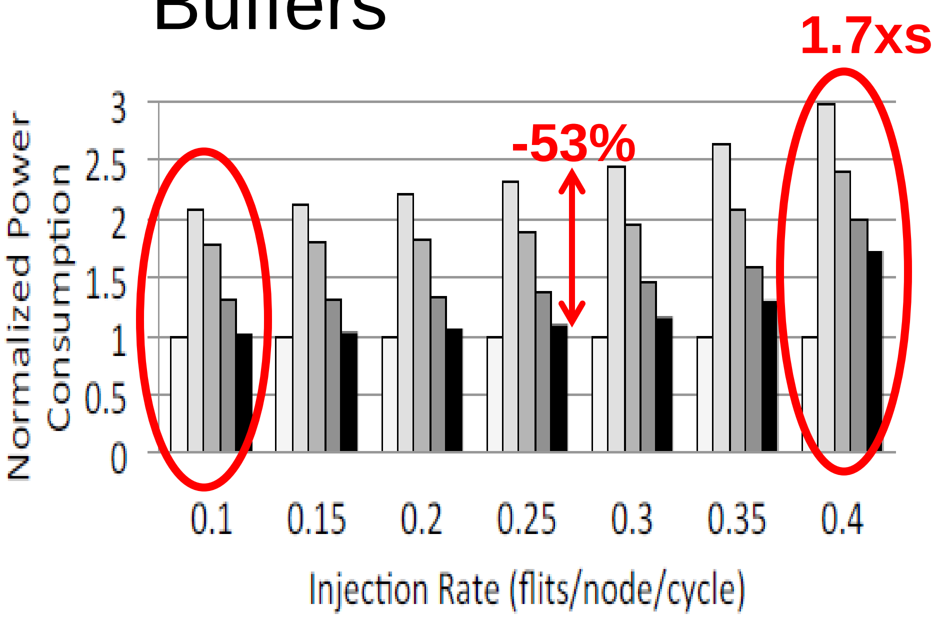
Performance Analysis - Benchmark Test

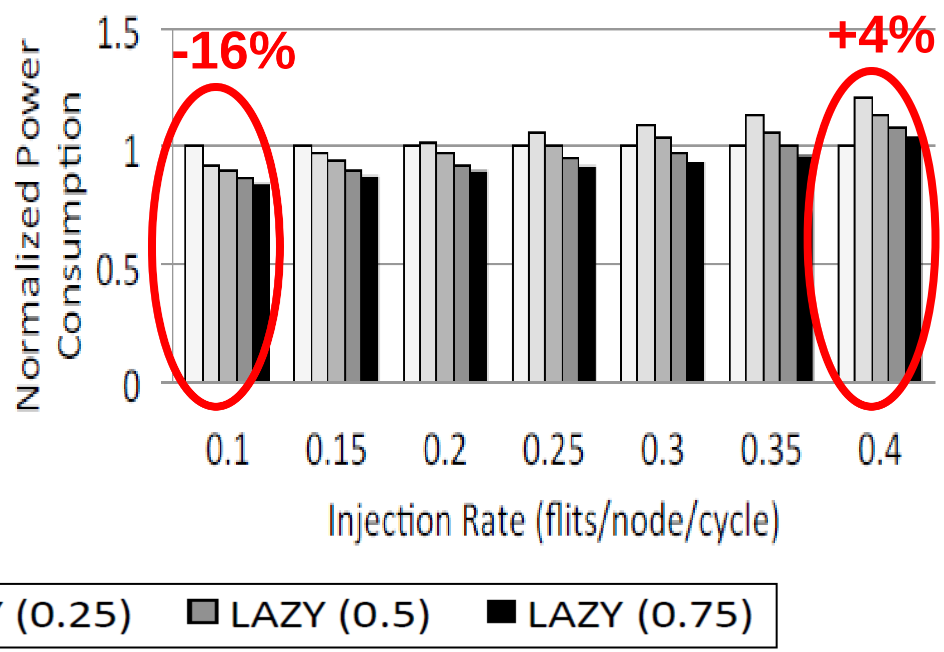
- SPLASH-2 parallel benchmarks



Power Analysis

- Dynamic Power consumption of Input Buffers
- Dynamic + Leakage Power consumption of on-chip routers





Conclusions

- Hybrid Buffer Design with STT-MRAM
  - Provide more buffer space under the same area budget
  - Throughput-efficient
- Performance Improvement
  - 21% on average in synthetic workloads
  - 14% on average in SPLASH-2 parallel benchmarks
- Power Savings
  - Lazy migration scheme reduces power by 61% on average compared to simple migration scheme