MCSoC 15th IEEE International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2022)
Universiti Sains Malaysia, Penang, Malaysia, December 19-22,2022

Traffic-Aware Energy-Efficient Hybrid Input Buffer Design for On-Chip Routers*

Yujie Gao[1], Yuan He[2,1], Xiaohan Yue[1], Haiyan Jiang[1], Xibo Wang[1]

[1]Shenyang University of Technology, Liaoning, China
[2]Keio University, Yokohama, Japan

*This work was supported, in part, by JSPS KAKENHI with Grant JP20K23315, and by the Okawa Foundation for Information and Telecommunications under Grant 21-04.

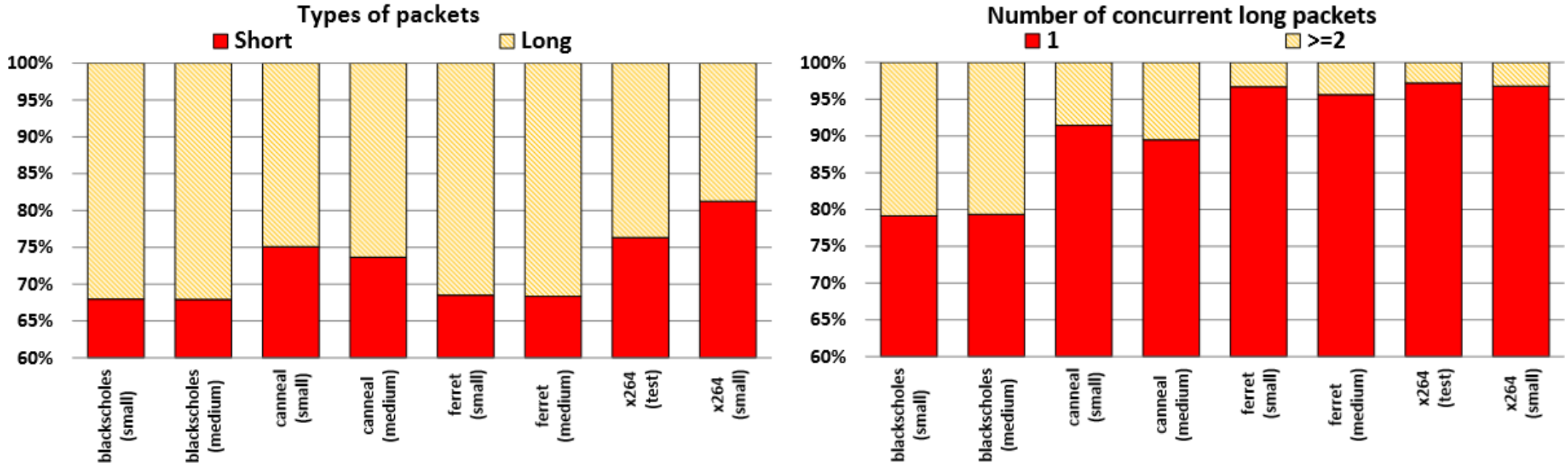
Agenda

- Background
- Motivations
- Our proposal: **The Traffic-Aware Hybrid Input Buffer Design**
- Evaluations
  - Methodology
  - Results
- Conclusions

Background

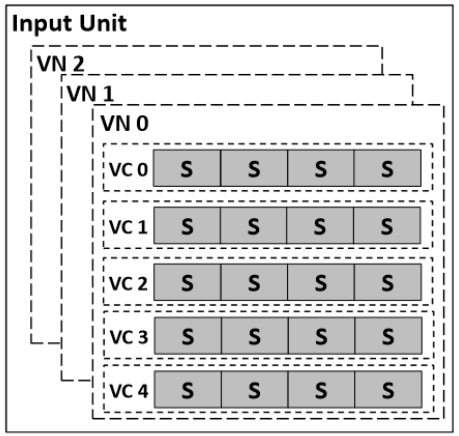
- Multicore/Manycore dominate today's computer systems
  - End of Dennard scaling and the continuous demand for higher perf.
  - More and more processor cores/devices will be piled on a single chip
- Needs for scalable/efficient on-chip interconnections (NoCs)
  - Size and complexity of NoCs grows as the number of cores/devices increases
  - Performance and power consumption of NoCs are critical to the system
- Up to date, most NoC designs employ Virtual Channels (VCs) for better utilization of the link bandwidth
  - Each physical port in a router has several VCs while each VC is a set of buffers used to store network traffic
  - NoC power is dominated by such VCs (buffers)
    - Accounted for more than half of the static and over 80% of the dynamic power consumed by the NoC

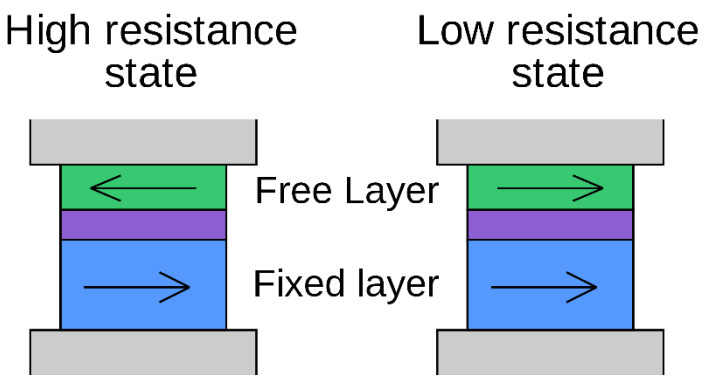
Motivations



- Characterizing the traffic in NoCs with certain applications from PARSEC
  - From the LHS figure: short packets (1 flit) dominate.
  - From the RHS figure: when there is a long packet in an input unit, it is likely the only long packet in the input unit
- These two observations told us:
  - The first buffer elements in all VCs are more frequently used than other buffer elements
  - When an input unit holds long packets, it is most likely there is only 1 VC that is fully active

Motivations (cont.)

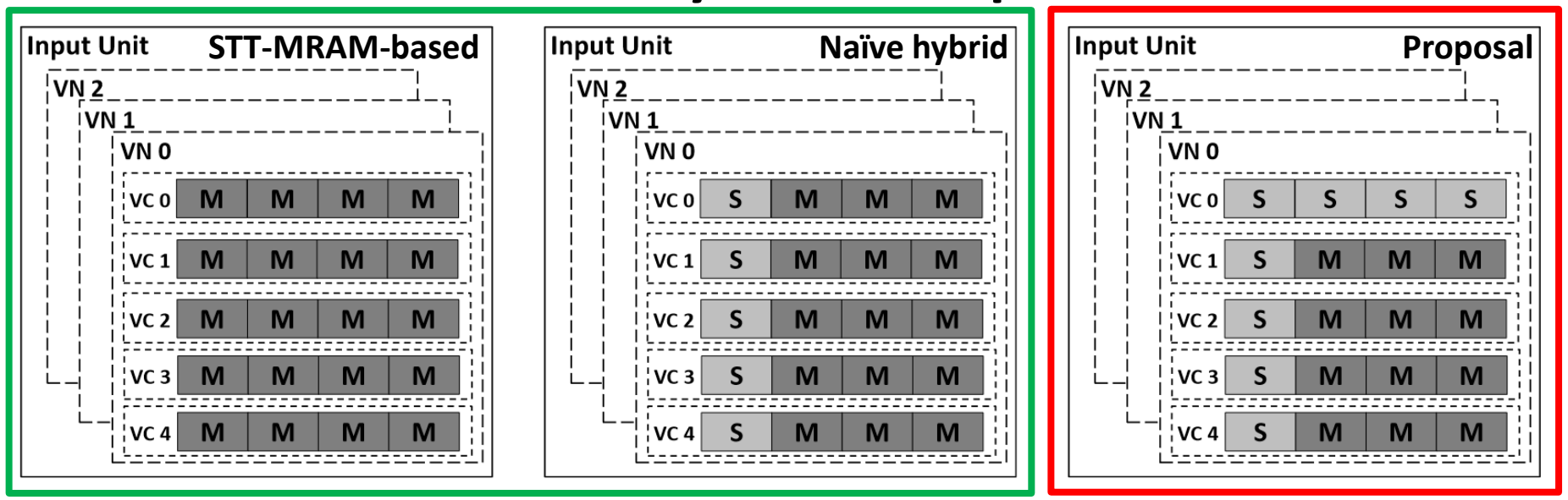




- If the first buffer elements in all VCs are more frequently used than other buffer elements
- And when an input unit holds long packets, it is most likely the case that only one VC is fully active

- Do we still need to implement the input buffer with fast but leaky SRAM devices?
- Do we have another choice? STT-MRAM
  - Low leakage
  - High density
  - Fast read

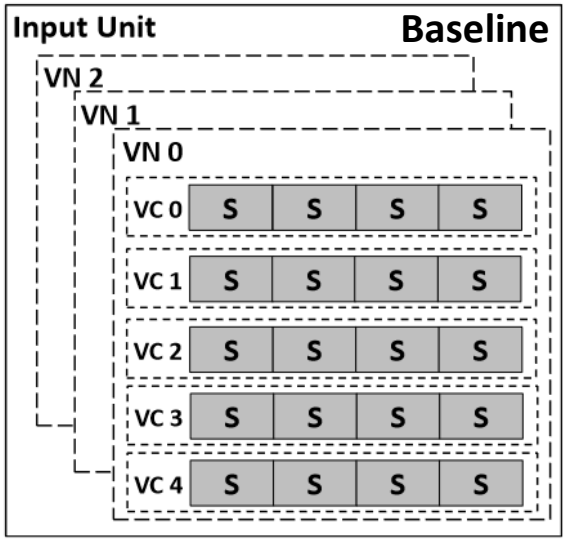
The Traffic-Aware Hybrid Input Buffer Design

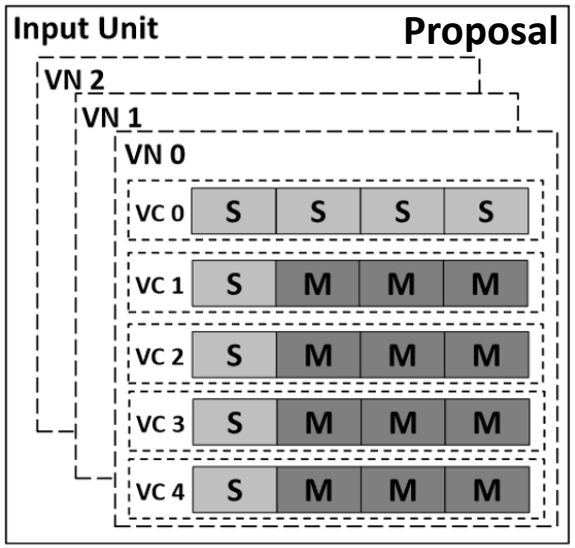


S: SRAM
M: STT-MRAM

- To retain the performance for short packets
  - The first buffer elements of all VCs are kept to SRAM
- To retain the performance for majority of the long packets
  - VC0 is also kept to SRAM
- Other buffer elements are instead implemented with STT-MRAM devices
  - We assume the peripheral circuits can be shared by both devices
- Two other designs are also added for comparisons
  - STT-MRAM-based
  - Naïve hybrid

Qualitative Comparisons between Baseline and Proposal

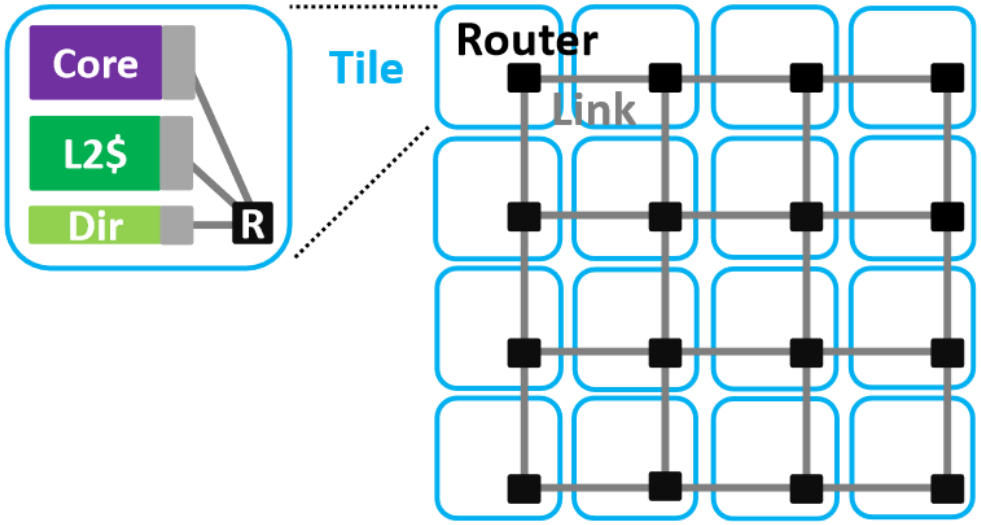




S: SRAM
M: STT-MRAM

- Performance: Read and write to these hybrid VCs for long packets will be slower when compared to the baseline design, but this is not frequent
- Static power: Consumes smaller amount of static power than the baseline design
- Dynamic power: Reduces the amount of SRAM accesses, thus reducing its dynamic power
- Energy: Consumes less energy when compared to the baseline design
- Area: The area consumption is smaller

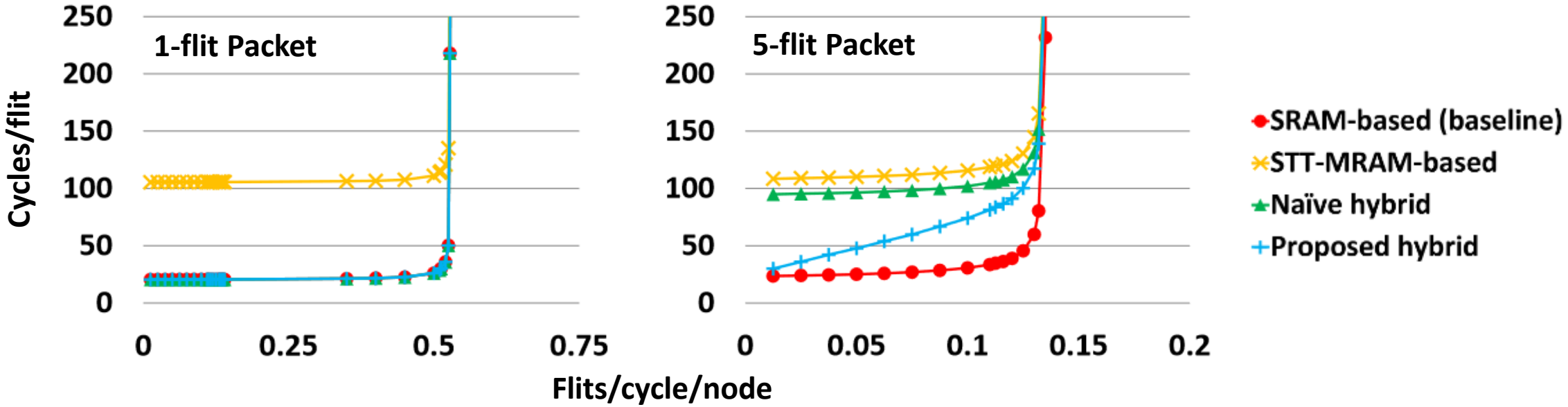
Evaluation Methodology



System parameters	
Number of cores:	16
Topology:	4 × 4 mesh
Processor:	3 GHz, In-order
L1 I/D cache:	32 KB per Processor, 2-way set associative, 2 cycles per access
L2 cache:	256 KB per Bank, 8-way set associative, 20 cycles per access
Cache line:	64 Bytes
Main memory:	8 GB, 180 cycles per access
Coherence protocol:	MOESI, Directory
Link:	128-bit, 1 cycle traversal
Packet:	128-bit control (short), 640-bit data (long)
Router:	3 GHz, 4 cycles pipeline, 103.15 mW static power, 1534.26 mW peak dynamic power
Virtual channel:	5 per Virtual network
Virtual network:	3 per Physical link
Routing algorithm:	X-Y routing
Process technology:	22 nm
Vdd:	1 V
STT-MRAM parameters	
Capacity:	64 bytes (same as a VC)
Access latency:	5 cycles per read, 31 cycles per write
Access energy:	31.64 pJ per read, 128.8 pJ per write
Leakage power:	534.52 uW

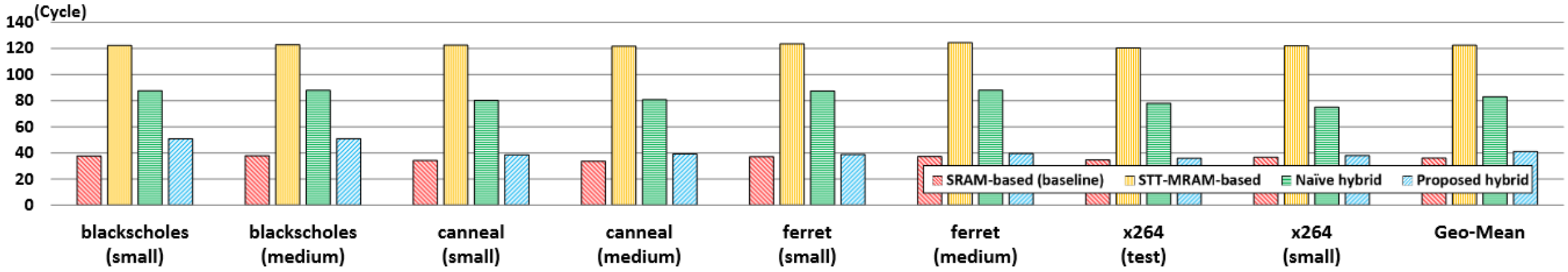
- Simulated on a 16-core system
  - gem5 with GARNET (network), McPAT (power) and NVSim (STT-MRAM)
- Network under 2D mesh topology
  - 128-bit link width
  - X-Y routing
- Detailed parameters in the RHS table
- Workloads from PARSEC

Evaluation Results with Synthetic Traffic



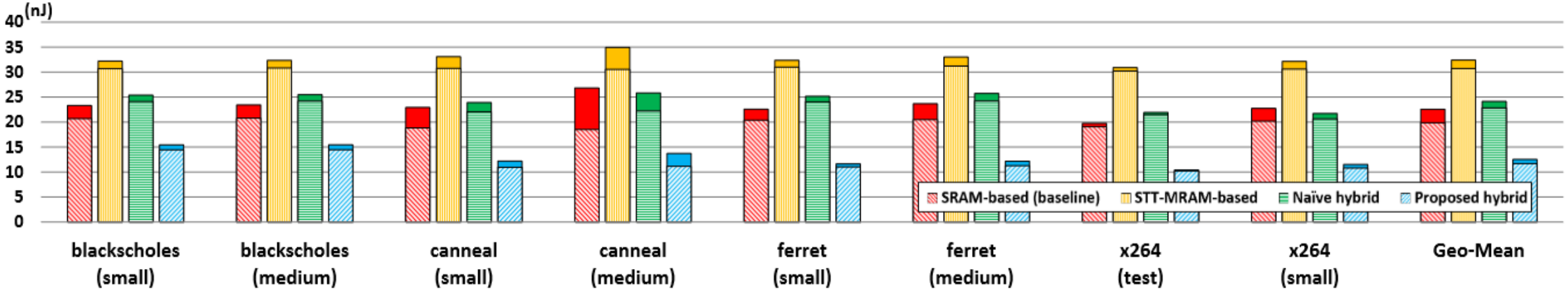
- With 1-flit packets, naive hybrid and our proposal perform as well as the baseline
  - Only SRAM-based buffer elements are accessed
- With 5-flit packets, only our proposal performs similar to the baseline when the injection rate is low
  - Our proposal has a complete SRAM-based
- With our proposal, latency rapidly exacerbates and its performance approaches the STT-MRAM-based and the naive hybrid designs when injection rate increases
  - A consequence of having more and more long packets in the input units

Evaluation Results with Application Traffic (Latency/flit)



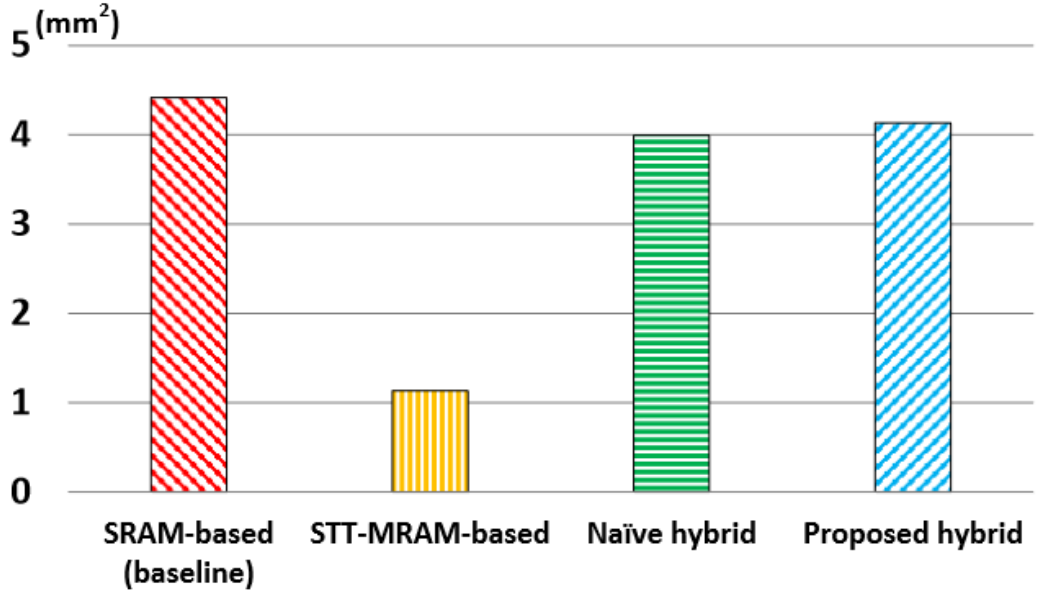
- Our proposal uses slower memory technology, but it achieves similar latency to the baseline design
  - It is roughly 1 cycle slower than the baseline design in “x264” workloads with an average slowdown of 13.9%
  - This has proved that our traffic-aware hybrid design works well
- Both the STT-MRAM-based and the naive hybrid designs have seriously suffered from the large access latency to STT-MRAM-based buffers
  - The naive hybrid design is better than the STT-MRAM-based one
  - As only long packets are slowed down in this case

Evaluation Results with Application Traffic (Energy/flit)



- Despite of utilizing more STT-MRAM devices, the STT-MRAM-based and the naive hybrid designs consume more energy than the baseline design and our proposal
  - Since they incur much longer network latency
- When being compared to the baseline design, our proposal is nearly able to cut the energy consumption in half
  - With a geometric mean of 44.5%
  - Considering the characteristics of network traffic is the key

Area of Routers with Different Input Buffer Designs



- On-chip routers with our proposal consumes 93.6% of the area of a baseline router
- The STT-MRAM-based design is the best
  - Its buffers are completely replaced with STT-MRAM whose density is much better than SRAM
- The naive hybrid design consumes slightly less area than our proposal
  - Since they have more buffers implemented with STT-MRAM than our proposal

Conclusions

- Input buffers in routers are very critical for NoCs
  - They determine both power consumption and throughput of the network
- In this work, we proposed a novel input buffer design
  - It mixes two memory devices, SRAM and STT-MRAM
  - The frequency of long packets is utilized to optimize the accesses to input buffers
    - Results in a significant cut to the energy consumption in the network
    - Incurs very little negative effect on the latency despite of using slow but less leaky memory technology
  - Such effectiveness proves our proposal is very future-proof
    - Competitive latency/throughput
    - Reduced energy consumption