STLAC: A Spatial and Temporal Locality-Aware Cache and Network-on-Chip Codesign for Tiled Many-core Systems

Mingyu Wang and Zhaolin Li

Institute of Microelectronics, Tsinghua University, Beijing 100084, China

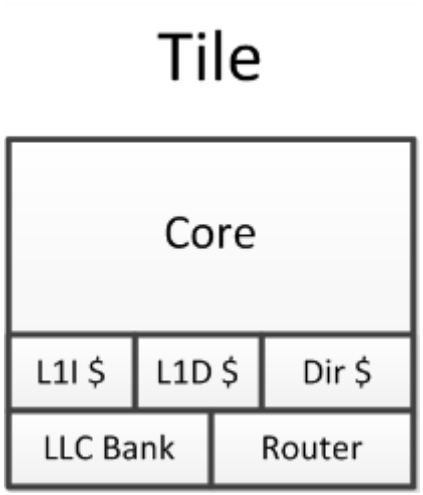
Email: wang-my12@mails.tsinghua.edu.cn

Outline

- Introduction and Motivation
- Design of STLAC
- Evaluation and Results
- Conclusions

Introduction

- Target architectures
- Tiled many-core architectures
  - large last level cache resources
  - meshed network-on-chip (NoC)
  - good power distribution
  - good scalability.



Tiled many-core architectures are widely used in multimedia applications and scientific computing.

Introduction

- Challenges for Design:
  - Cache capacity interference problem
  - Difference workloads → Difference memory access behaviors

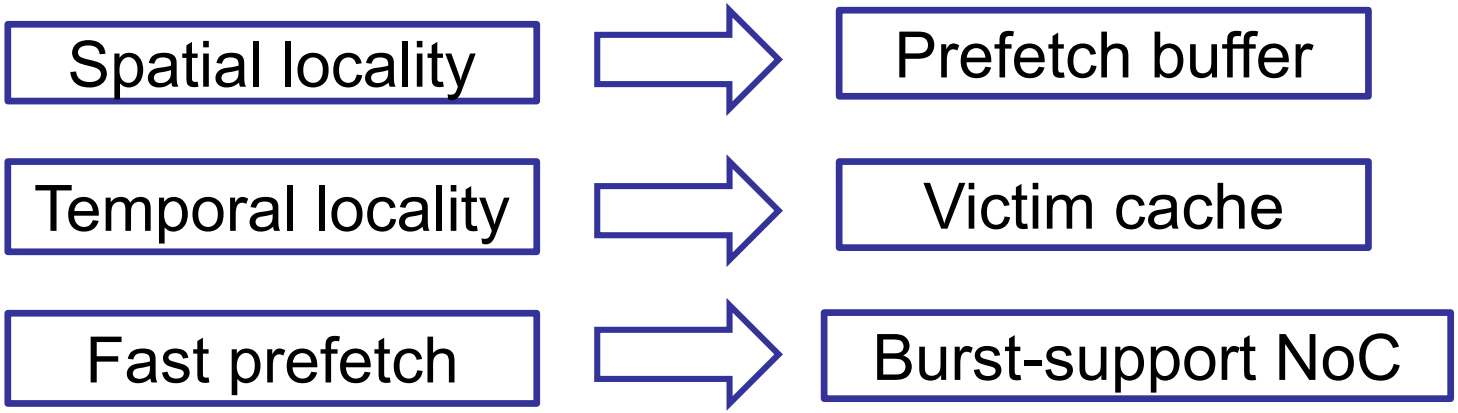
    (e.g. stream-like workloads with low temporal locality)
  - Not-uniform cache access (NUCA) effect.
  - Especially notable in NoC-linked cache banks

Introduction

- Motivation
  - The spatial and temporal locality of workloads are the root causes for cache designs to overcome the memory wall problem.
  - Difference workloads have different locality features.
  - To solve the cache capacity interference problem and NUCA effect from a view of cache and NoC codesign.

Introduction

- Motivation
  - Prefetch buffer speculates the data blocks in subsequent addresses to exploit the spatial locality.
  - Victim cache collects the evicted data blocks from the upper memory hierarchy to exploit the temporal locality.

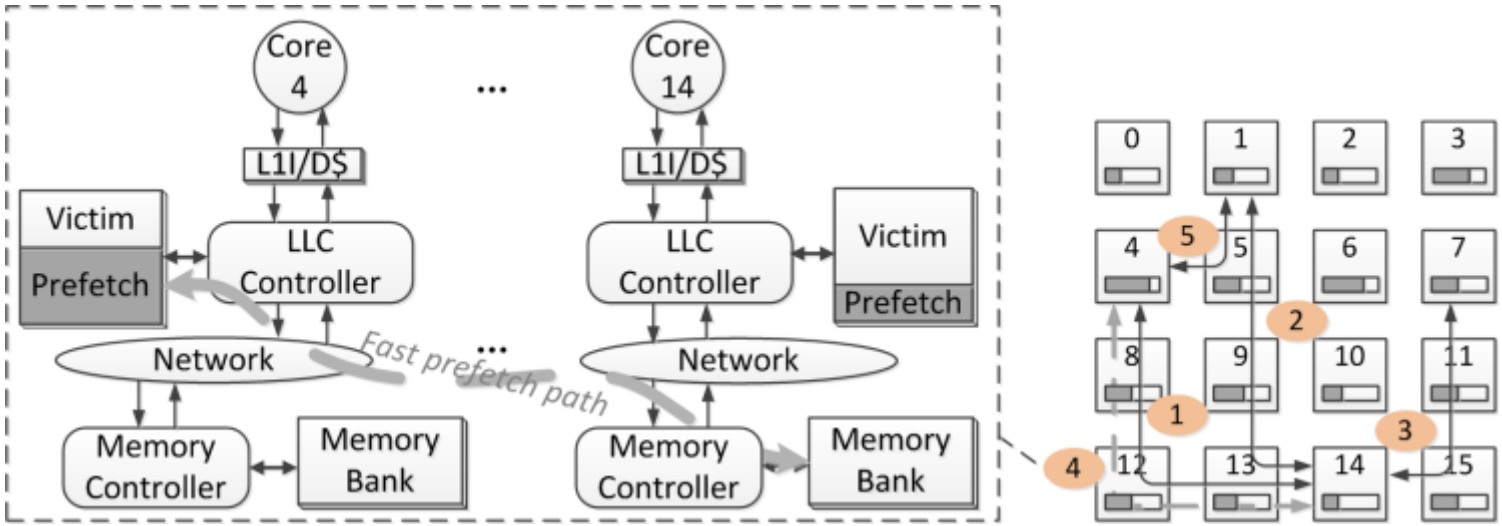


Proposed STLAC

- Key Idea:
  - To dynamically partition the last level cache (LLC) as data prefetch buffer or victim cache for locality prediction.
  - To exploit a hybrid burst-support NoC for fast data prefetch and to save the network usage.
  - To explore more optimization opportunities from the cache and NoC codesign.

Proposed STLAC

- Overview of the Codesign

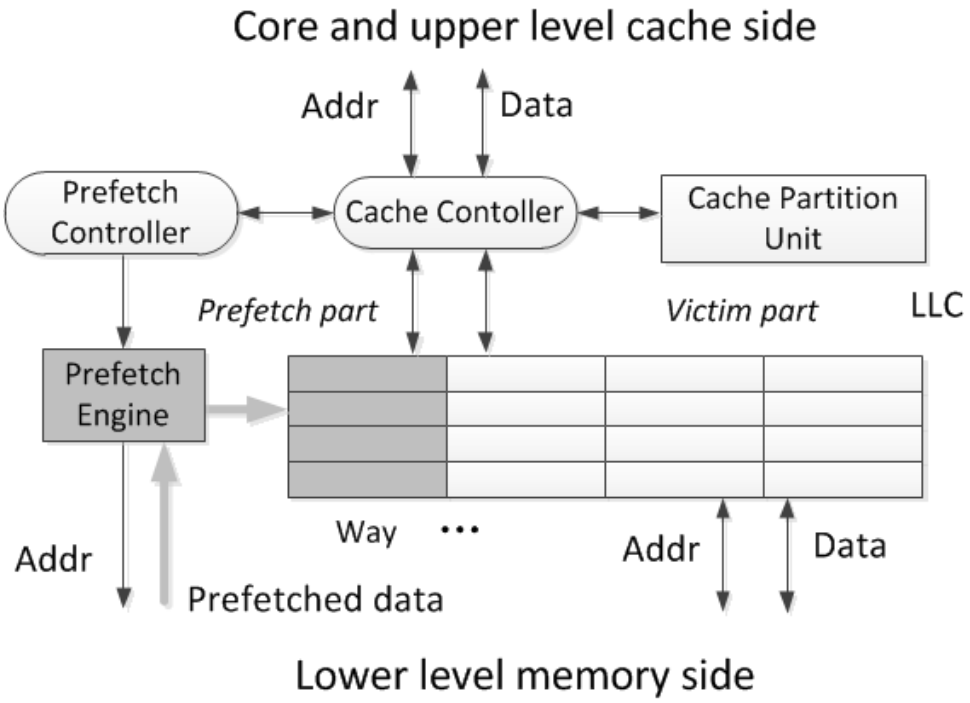


①②③⑤: normal data access behaviors.

④: fast data prefetch is issued between node 4 and node 14 via the burst-support NoC .

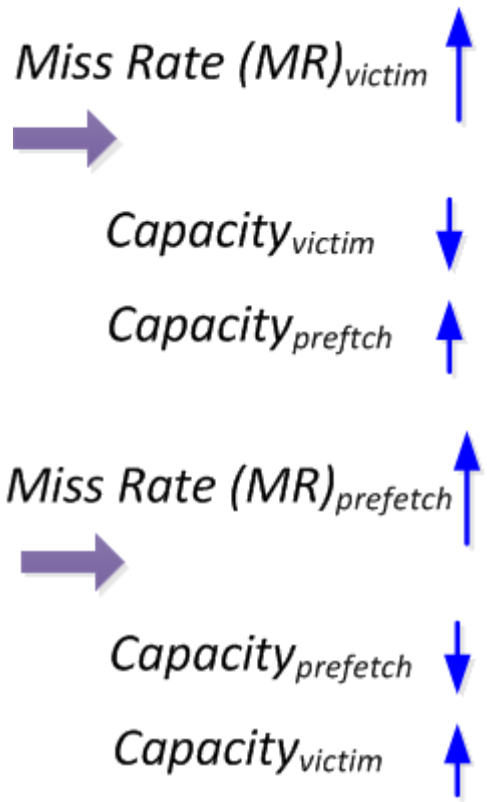
Proposed STLAC

- Adaptive Cache Resource Partition
  - to dynamically adjust the ratio between the victim part and prefetch part
  - cache partition is operated at way-granularity and executed periodically



Proposed STLAC

- Cache Partition Algorithm (CPA)

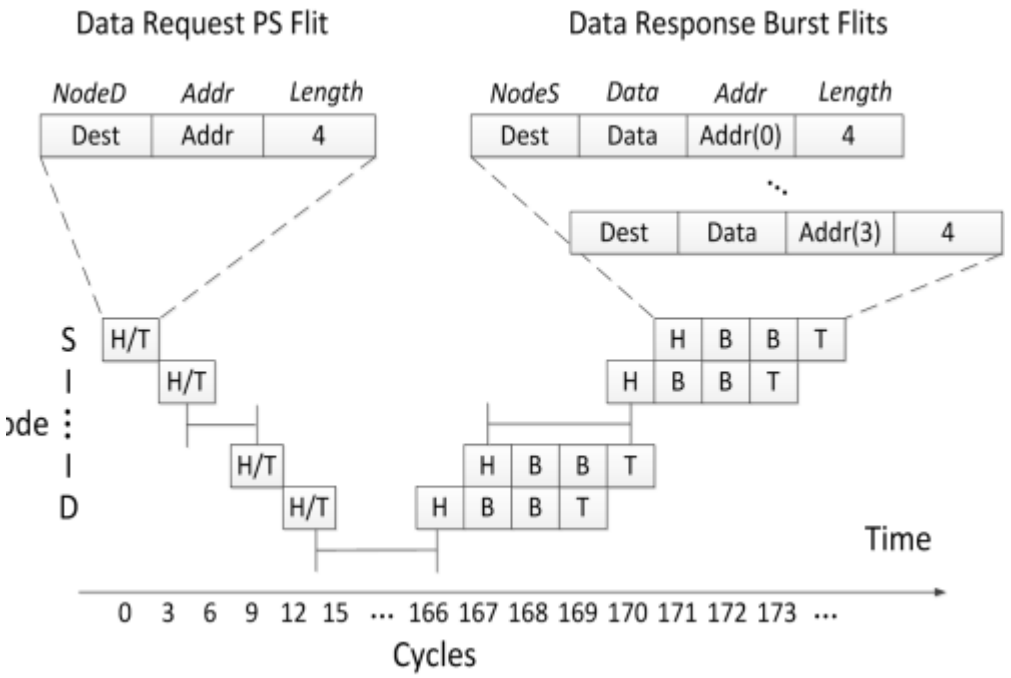


Algorithm 1 Cache Partition Algorithm (CPA)

1: **set** $Capacity_v = Capacity_p$;
reset all Counters and MR Profilers
{cache configuration and counter initiate}
2: **if** $MR_v - MR_p \geq threshold$ **then**
3: $Capacity_v - -$;
{**shrink** the capacity of victim cache}
4: $Capacity_p + +$;
{**expand** the capacity of prefetch buffer}
5: **configure** the LRU way of the victim cache as prefetch data buffer;
6: **else if** $MR_v - MR_p \leq threshold$ **then**
7: $Capacity_v + +$;
{**expand** the capacity of victim cache}
8: **configure** the LRU way of the prefetch data buffer as victim cache;
9: $Capacity_p - -$;
{**shrink** the capacity of prefetch buffer}
10: **else**
11: **Return**;
{optimal cache configuration achieved}
12: **end if**

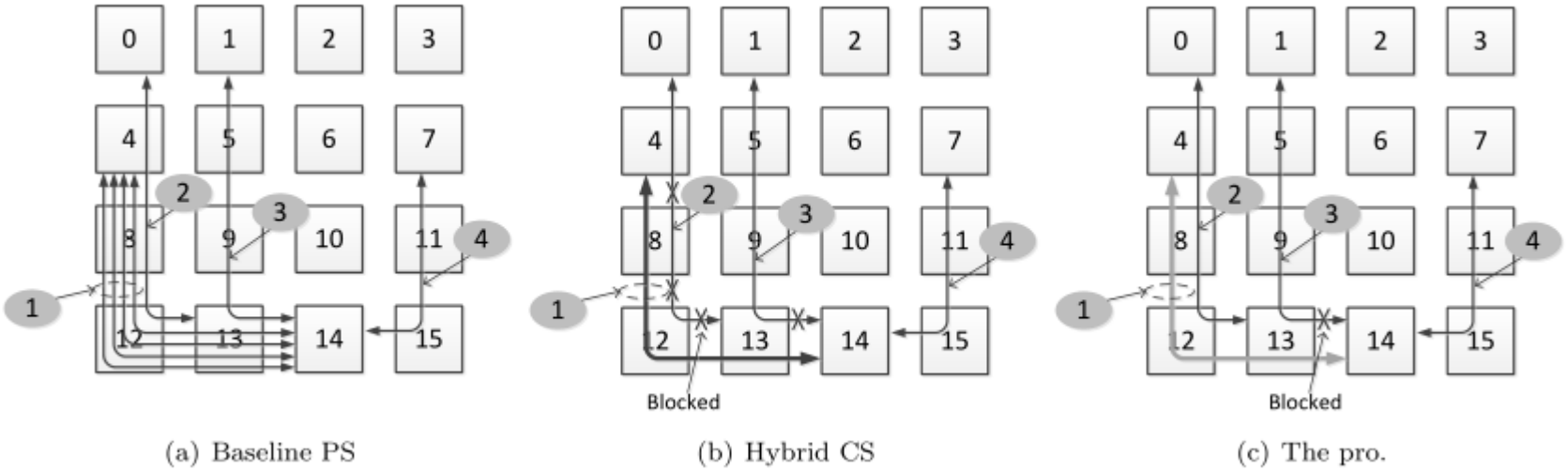
Proposed STLAC

- Fast Data Prefetch using Burst-support NoC
  - To prefetch remote data as fast as possible without breaking the data continuity.
  - To speculate the data blocks of incremental addresses in destination nodes.
  - To avoid frequently sending separate request or response flits from source or destination nodes.



Proposed STLAC

- Example of the hybrid data access

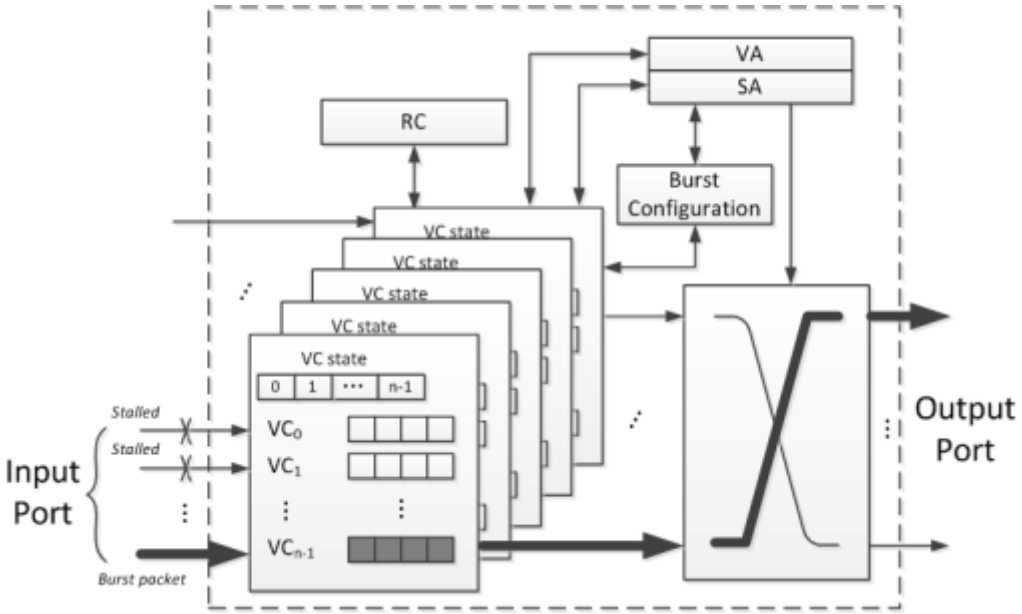


(a) Baseline PS (b) Hybrid CS (c) The pro.

① represents a 4-length data prefetch with packet-switched (PS), circuit switch (CS) and the burst-support NoC connections. ② ③ ④ represent other normal PS connections.

Proposed STLAC

- Architecture of the burst-support router
  - N-field virtual channel (VC) state register is added.
  - Higher priority is assigned to burst packets.
  - Switch will be reserved for the entire burst packets to keep the data continuity.



Proposed STLAC

- Architecture of the burst-support router
- Starvation avoidance
  - The initial age of the normal PS flits is set to 0 and burst flits have the higher priority than the PS flits with age 1.
  - After PS transmission is blocked for a predetermined number of cycles, the age of PS flits will be increased. PS flits with age 1 have the same priority with burst flits.

Evaluation and Results

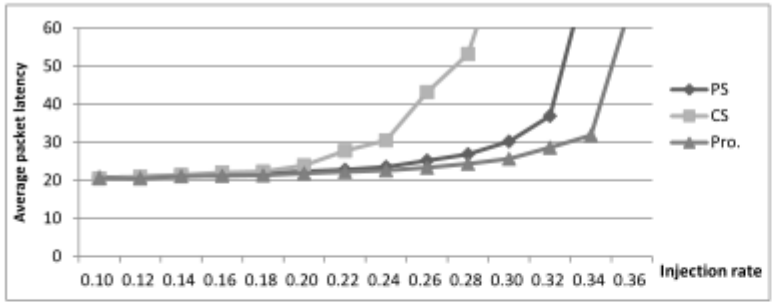
- Methodology
  - *NoC simulator Booksim2.0*
  - *4x4 2D mesh, 128-bit channels, DOR*
  - *FeS2 multiprocessor simulator*

TABLE I: Configuration Parameters

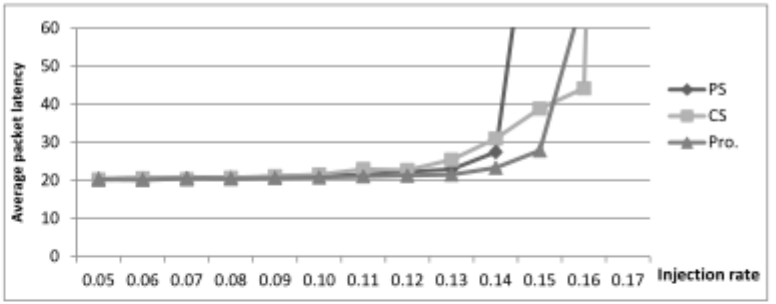
CPU type	16-core, x86
OS	Fedora 64-bit
L1I/D cache	16kBytes, 4-way, private
L2 cache	128kBytes/slice, 8-way, STLAC
CPA threshold	10%
Block size	64Bytes
Prefetch length	4
Coherence protocol	MOESI Distributed Directory
NoC topology	4×4, 2D Mesh
Routing algorithm	dimension-order
Baseline router	3-stage pipeline
Main memory	512MBytes, 150 cycles latency

Evaluation and Results

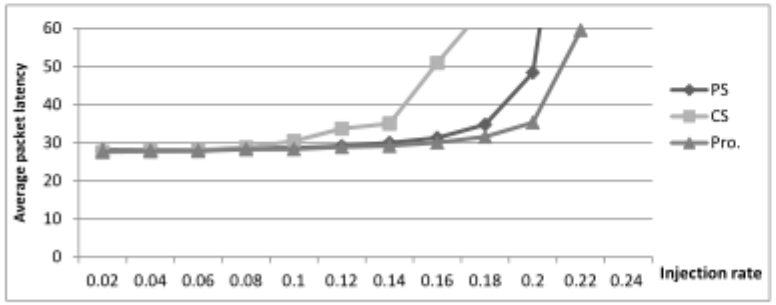
- Burst-support NoC Simulation Results



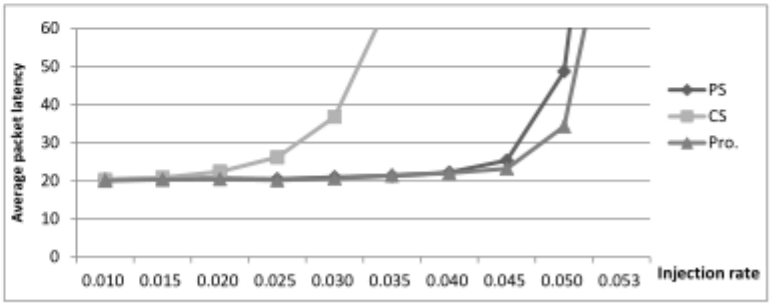
(a) uniform



(b) transpose



(c) bitcomp



(d) hotspot

5% burst inject rate

Evaluation and Results

- Full System Simulation Results
  - Miss rate reduction results

- The cache partition is executed every 1M cycles.
- About 40% off-chip misses are reduced on average.
- This reduction rises to 50% if the running workloads have good spatial and temporal locality.

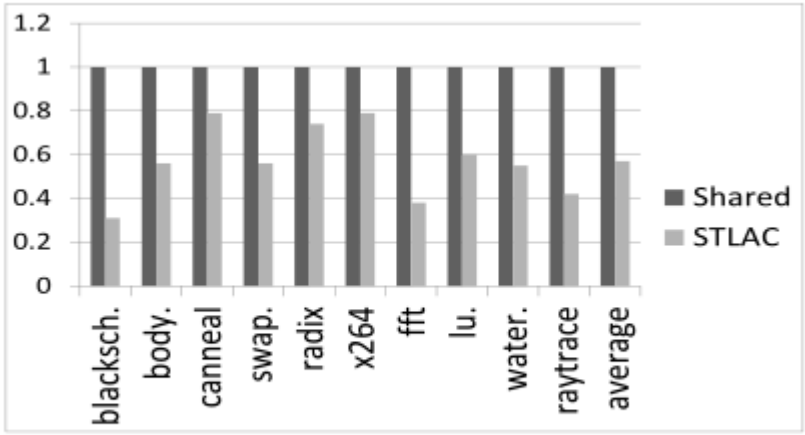


Fig. 7.: Normalized off-chip misses results.

Evaluation and Results

- Full System Simulation Results
  - Hit distributions to show the locality difference

- High hit rate in prefetch part shows good spatial locality.
- High hit rate in victim part shows good temporal locality.
- Cache pollution is relieved because the prefetch part and victim part is managed separately.

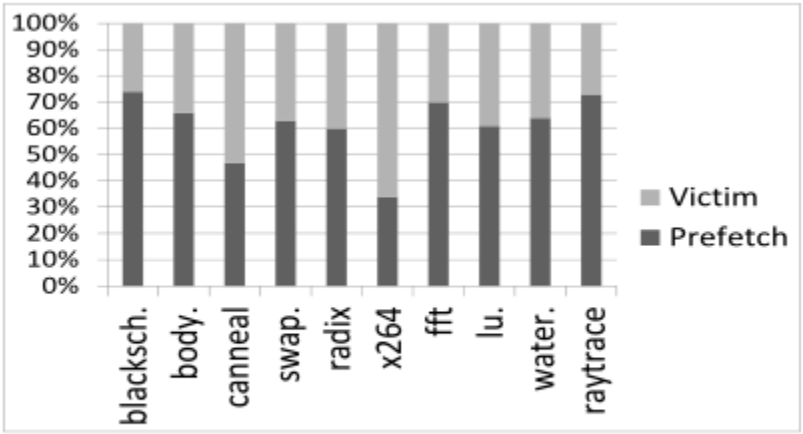


Fig. 8.: Hit distributions for different benchmarks.

Evaluation and Results

- Full System Simulation Results
  - Network usage reduction

➢ The total number of injected flits into the network is taken as the metric for network usage evaluation.

➢ The network usage is saved by 7.6% on average.

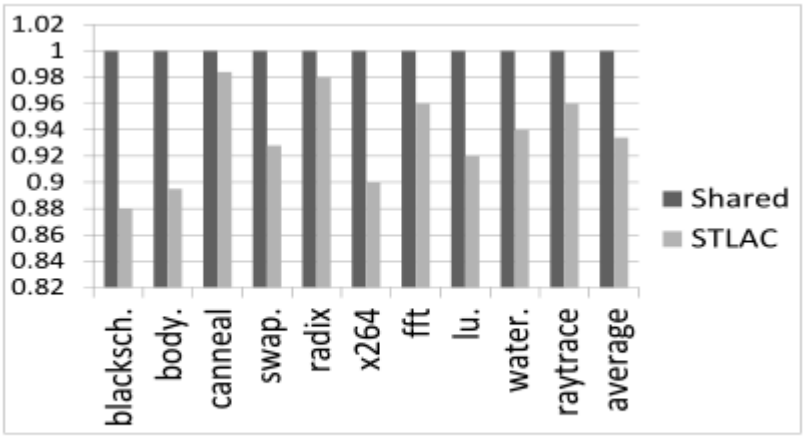


Fig. 9.: Normalized network usage results.

Evaluation and Results

- Full System Simulation Results
  - Performance results

➢ The total runtime (cycles) is taken as the metric for performance evaluation.

➢ About 15% performance is improved on average.

➢ This improvement even rises to nearly 30% for these workloads with good spatial and temporal locality.

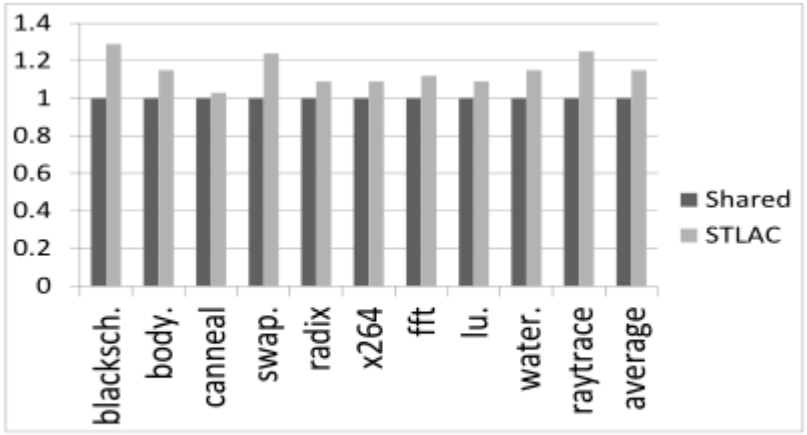


Fig. 10.: Normalized performance results.

Conclusions

- In this work
  - Adaptive cache resources partition by taking advantage of the differences of the spatial and temporal locality.
  - Fast prefetch is realized in the proposed hybrid burst-support network to save the on-chip network usage.
  - Explore the opportunities for performance improvement from a view of cache and NoC codesign.
- Future work
  - More discussion about energy consumption.
  - Further optimization on NoC router for network latency reduction (e.g. routing algorithm).

Thank you for your attention!