Memory Access Pattern-Aware DRAM Performance Model for Multi-core Systems

ISPASS 2011

Hyojin Choi*, Jongbok Lee+, and Wonyong Sung*

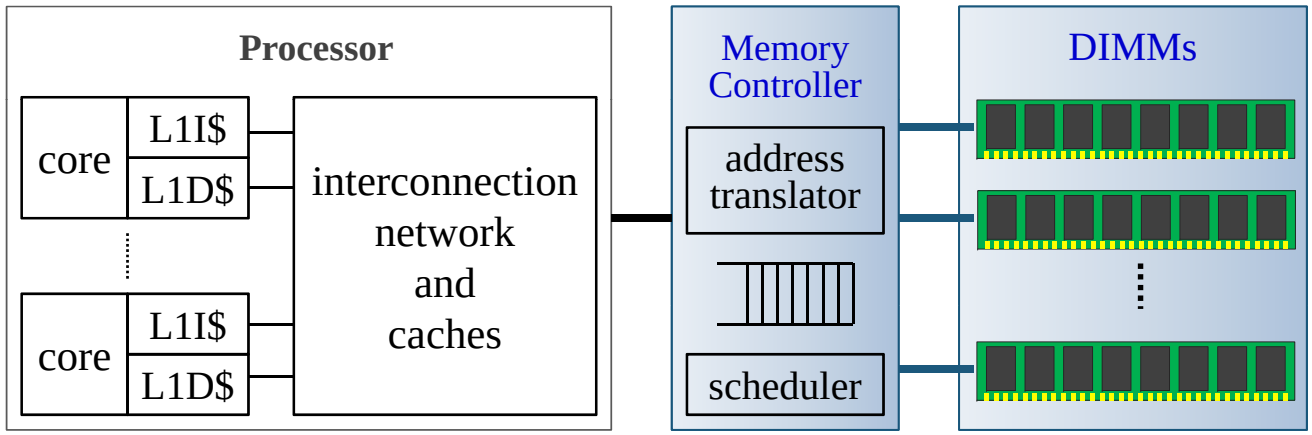
hjchoi@dsp.snu.ac.kr, jblee@hansung.ac.kr, wysung@snu.ac.kr

*Seoul National University, +Hansung University

Seoul, Korea

Introduction

- The memory-wall problem in multi-core era
  - The rate at which memory traffic is generated by an increasing number of cores is growing faster than the rate at which it can be serviced.

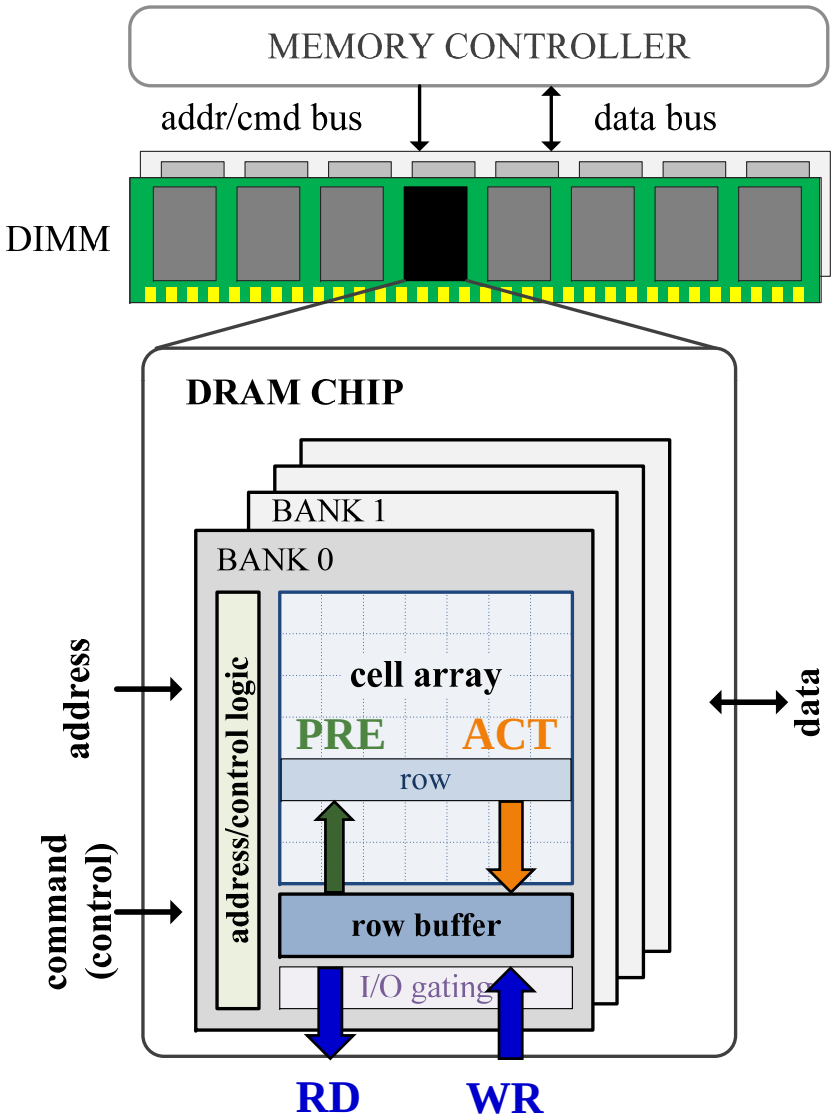


- Our research focuses on main memory subsystem design.
- This paper proposes an analytical DRAM performance model.

Outline

- Background
- Motivation
- Approach
- Objective
- Modeling Bank Busy Time
  - Minimum inter-command delays
  - Pattern parameters
  - Average bank busy time
- Evaluation Results
- Concluding Remarks

DRAM architecture

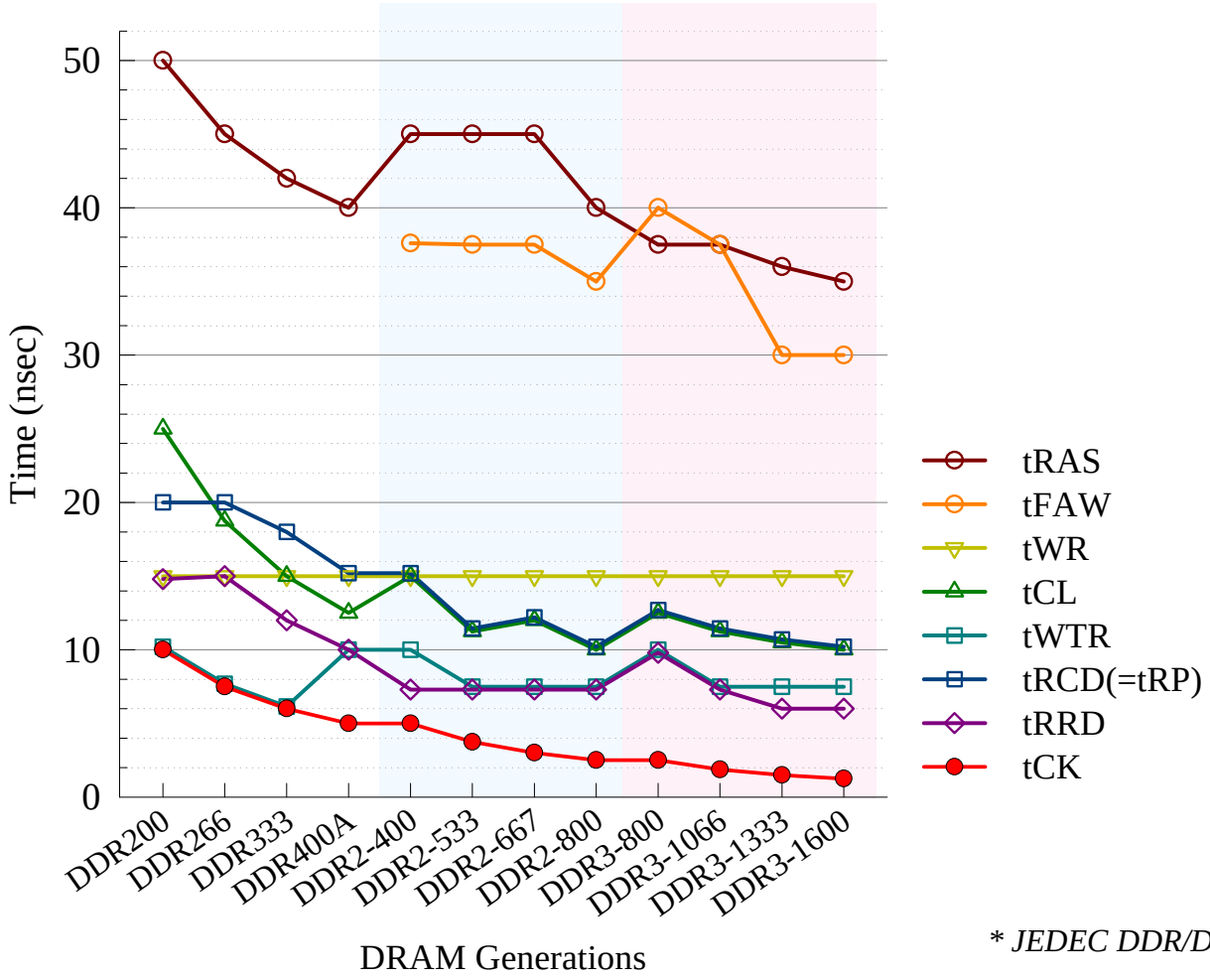


- Multiple banks (typically 4 or 8)
  - Each bank has cell array, row-buffer, and address/control logics
  - The address, command and data buses are shared by all banks

DRAM operations

- ***Activate (ACT)***
  - an entire row data is read from the cell array and stored to the row-buffer (row-buffer is open)
- ***Precharge (PRE)***
  - the contents of the row-buffer are restored to cell array (row-buffer is closed) and bitlines are precharged
- ***Read (RD)*** or ***write (WR)***
  - from/to the row-buffer

DRAM timing trends



** JEDEC DDR/DDR2/DDR3 Standards*

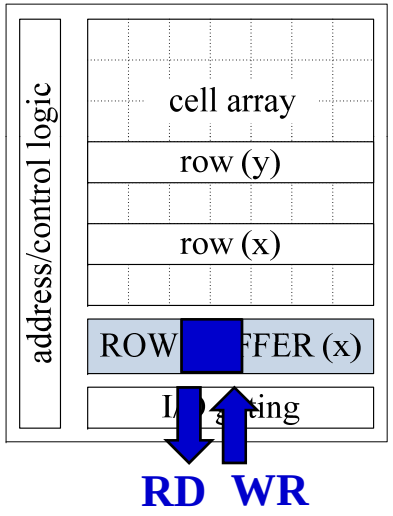
- The goal is to find out an analytical model which can show the impact of each DRAM timing on the performance.

Challenge

- DRAM access performance depends on a program's memory access behavior

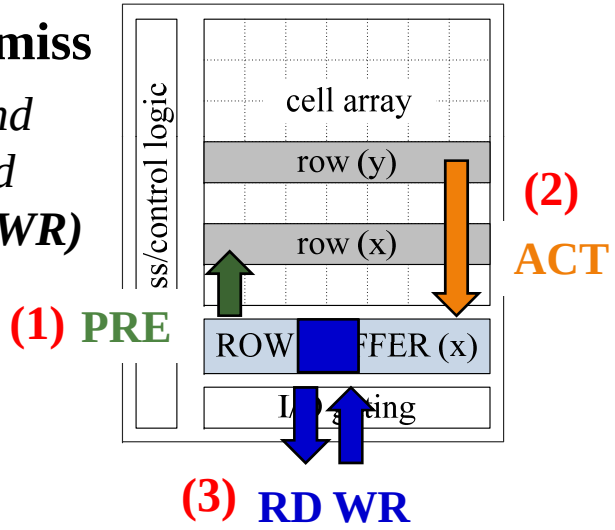
(a) row-buffer hit

row(x) is stored and
row(x) is requested
⇒ ***RD (WR)***



(b) row-buffer miss

row(x) is stored and
row(y) is requested
⇒***PRE-ACT-RD(WR)***



- The DRAM command chain generated to serve a memory request depends on the incoming request and on the row-buffer status (open or closed, row index if opened), which is determined by the previously serviced requests.

Objective

- To find out an analytical model which has a form of

$$\chi = f(\boldsymbol{w}, \boldsymbol{\tau})$$

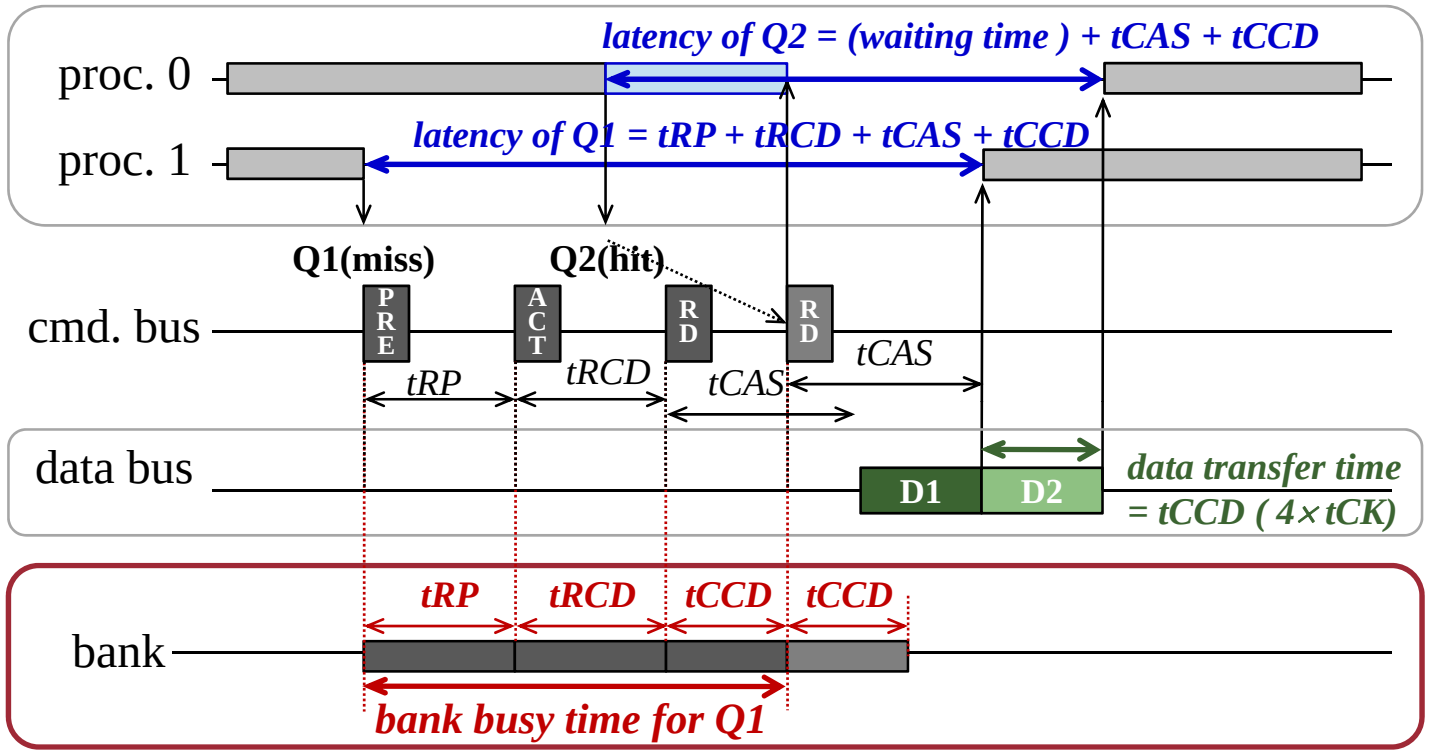
  - χ : performance metric
  - $\boldsymbol{w}$: characteristics of memory access behavior
  - $\boldsymbol{\tau}$: DRAM timings such as *tRP*, *tRCD*, *tRAS*, *tCCD*, …
  - f : a simple function of $\boldsymbol{w}$ and $\boldsymbol{\tau}$

- Key questions
  - What is the performance metric ?
  - How to characterize the memory access behavior of a program ?
  - What is the relationship between input parameters and the performance metric ?

Assumptions

- 1) One memory request is serviced by one column command
  - All memory references are cache misses.
  - cache block size = 64 Bytes, data bus width = 64 bits, burst length = 8
- 2) There are four DRAM commands: PRE, ACT, RD and WR
  - The effect of REF (refresh) to the access performance is negligible.
  - RDAP/WRAP (auto-precharge after RD/WR) are not generated when the memory controller adopts the open policy.
- 3) Open policy for row-buffer management
  - row-buffer misses $\rightarrow$ PRE-ACT-RD, PRE-ACT-WR
  - row-buffer hits $\rightarrow$ RD, WR
- 4) First-Ready First-Come First Served (FR-FCFS) scheduling
  - The row-buffer hit requests are prioritized miss ones to maximize data bus utilization.

Approach

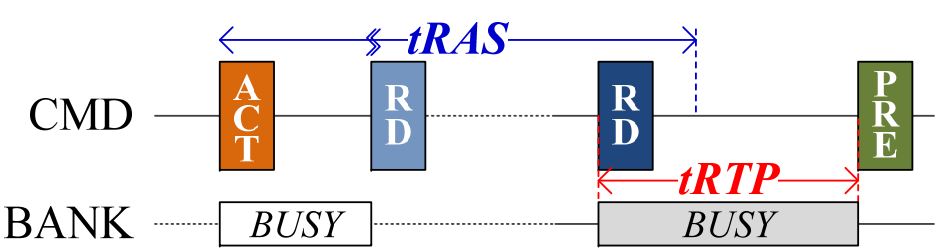


- Memory access latency includes the queuing delay.
- Data transfer time is related with only *tCK* among DRAM timings.

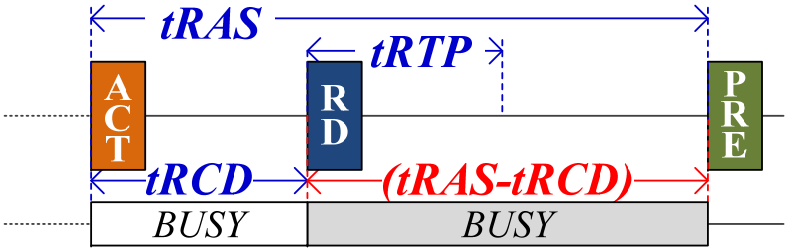
- Modeling the time needed for a bank to service DRAM commands
- → bank busy time

Bank busy time

- A bank is said to ***busy*** when it is not possible for the memory controller to issue any command to the bank due to timing constraints. Otherwise, a bank is in ***idle*** status.
- Considerations:
  - 1) simple : PRE (→*tRP*), ACT (→*tRCD*)
  - 2) dependency on the command that follows
    - ⇒ *in a pair-wise fashion (minimum inter-command delays)*
    - ex) RD-RD (→ *tCCD*) vs. RD-WR (→ *tRTW*)
  - 3) multiple timing constraints on PRE
    - ex) RD-PRE : it depends on the number RDs between ACT-PRE



(a) RD-PRE (→ *tRTP*)



(b) RD-PRE (→ *tRAS-tRCD*)

Minimum inter-command delays

- The minimum inter-command delay can be defined for all possible DRAM command pairs based on DRAM timing constraints defined in the data sheet

DRAM command pair	min. inter-command delay
PRE-ACT	tRP
ACT-WR, ACT-RD	$tRCD$
WR-PRE	$tCWL + tCCD + tWR$
RD(x)-PRE	$tRAS - tRCD - (x-1)tCCD$
RD(others)-PRE	$tRTP$
WR-WR	$tCCD$
RD-WR	$tRTW$
WR-RD	$tCWL + tCCD + tWTR$
RD-RD	$tCCD$

  - RD(x) represents the consecutive x RD commands (x=1, …, m)
    - $m = \lceil$(tRAS-tRCD-tRTP)/tCCD$\rceil$ (m=2, 3, 3, and 4 for DDR3-800/-1066/-1333/-1600)
  - RD(others) means the row-buffer miss cases which are not included in WR-PRE and RD(x)-PRE

Pattern parameters

- := the number of occurrences of each DRAM command pair
  - They can be interpreted as characteristics of memory access streams
    - cf) open-policy is assumed for the row-buffer management policy.

DRAM command pair	pattern parameters	main memory requests
PRE-ACT ACT-WR, ACT-RD	N_m	row-buffer misses
WR-PRE	N_{wp}	row-buffer miss after write
RD(x)-PRE	N_{rx}	row-buffer miss after x consecutive reads w/o write
RD(others)-PRE	N_{rt}	other cases for row-buffer miss
WR-WR	N_{ww}	write/hit request after write
RD-WR	N_{rw}	write/hit request after read
WR-RD	N_{wr}	read/hit request after write
RD-RD	N_{rr}	read/hit request after read

- the number of row-buffer misses (N_m) = $N_{wp} + N_{rx} + N_{rt}$
- the number of row-buffer hits = $N_{ww} + N_{rw} + N_{wr} + N_{rr}$

The proposed model

- The bank busy time is a linear combination of the minimum inter-command delays and pattern parameters.

$$\textit{Bank busy time} = \sum_{i=1}^{n} N_i \times D_i$$

$$\begin{aligned} S_{\text{busy}} = {} & tRP \cdot N_m \\ & + tRCD \cdot (N_m - \sum_{x=1}^{m} N_{rx}) \\ & + tCCD \cdot (N_{ww} + N_{wr} + N_{rr} + N_{wp} \\ & \quad - \sum_{x=1}^{m} (x-1) N_{rx}) \\ & + tCWL \cdot (N_{wr} + N_{wp}) \\ & + tRTW \cdot N_{rw} + tWTR \cdot N_{wr} \\ & + tRAS \cdot \sum_{x=1}^{m} N_{rx} \\ & + tWR \cdot N_{wp} + tRTP \cdot N_{rt}. \end{aligned}$$

Average bank busy time

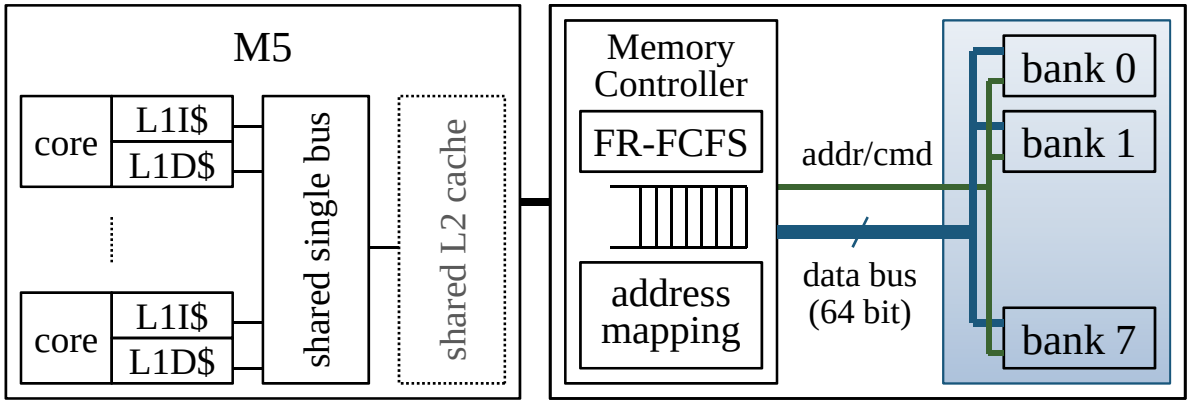
- := the bank busy time per a memory request
  - N : the number of memory requests to a bank during program execution

$$\textit{Average bank busy time} = w_0 \cdot tRP + w_1 \cdot tRCD + w_2 \cdot tCCD + w_3 \cdot tCWL + w_4 \cdot tRTW + w_5 \cdot tWTR + w_6 \cdot tRAS + w_7 \cdot tWR + w_8 \cdot tRTP$$

- , where

$$w_0 = N_m/N \quad \text{(row-buffer miss ratio)}$$
$$w_1 = (N_m - \sum_{x=1}^{m} N_{rx})/N$$
$$w_2 = (N_{ww} + N_{wr} + N_{rr} + N_{wp} - \sum_{x=1}^{m} (x-1)N_{rx})/N$$
$$w_3 = (N_{wr} + N_{wp})/N$$
$$w_4 = N_{rw}/N$$
$$w_5 = N_{wr}/N$$
$$w_6 = \sum_{x=1}^{m} N_{rx}/N$$
$$w_7 = N_{wp}/N$$
$$w_8 = N_{rt}/N.$$

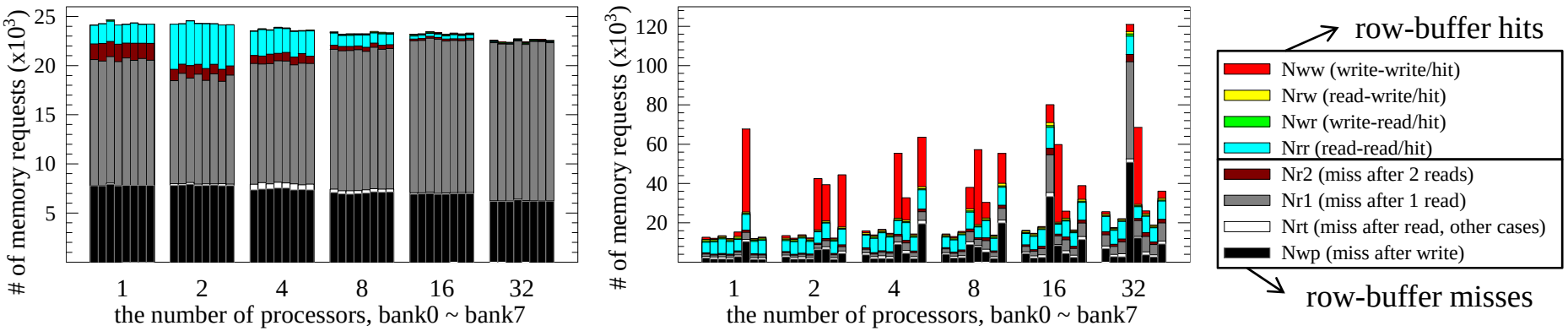
Experimental setup



kernel/application	description
FFT.MT	matrix transpose (512×512)
FFT.MM	matrix multiplication (512×512)
OceanContig	grid size : 258×258
Cholesky	input: tk23.O
LUContig	matrix size: 512×512
Raytrace	input: teapot.env
FMM	2048 particles

- Architecture simulator configuration (M5)
  - in-order processor model (P=1,2,…,64), 2 GHz
  - L1 cache : private, separate, 64 KB, 2-way, 64 Bytes, 1 cycle
  - L2 cache : shared, unified, 512 KB, 2-way, 64 Bytes, 20 cycles
  - shared bus with no overhead
- Main memory subsystem
  - a cycle-accurate DRAM timing simulator extension for M5
  - memory controller: FR-FCFS, [row:bank:col], open-policy
  - 2 Gbytes, 8 banks, DDR3-800/-1066/-1333/-1600, data bus width : 64 bit
- Seven multi-threaded workloads from SPLASH-2 benchmark

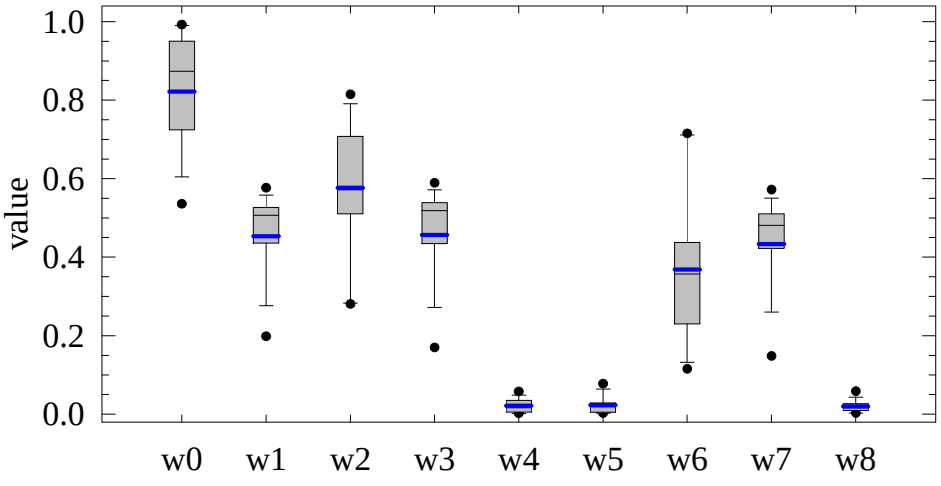
(1) Pattern parameters



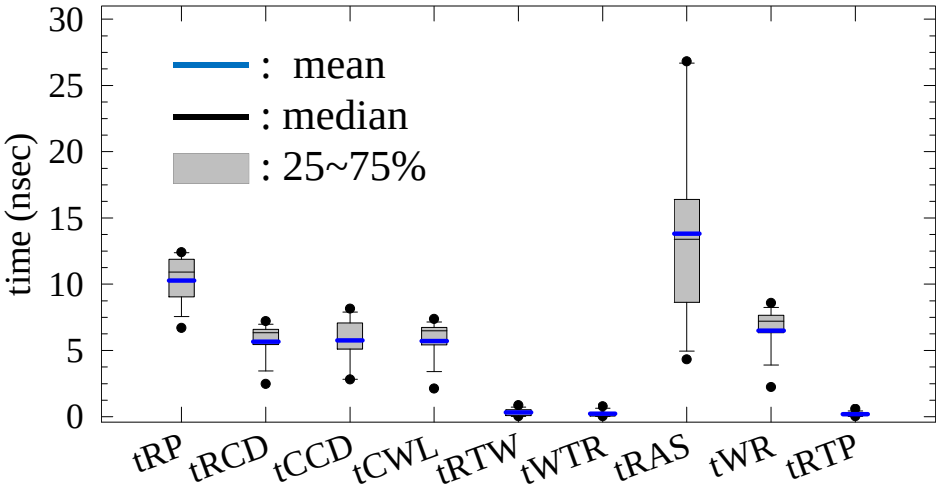
(a) FFT.MT (b) Raytrace

- The pattern parameters are obtained during the simulation as shown in the figure.
  - Other results are included in the paper.
- Selecting representative pattern parameters for a workload.
  - when the memory accesses are distributed non-uniformly across banks.
  - 1) select a bank that has the maximum number of requests
  - 2) use the pattern parameters of that bank

(2) Impact of DRAM timings on the bank busy time



(a) weights

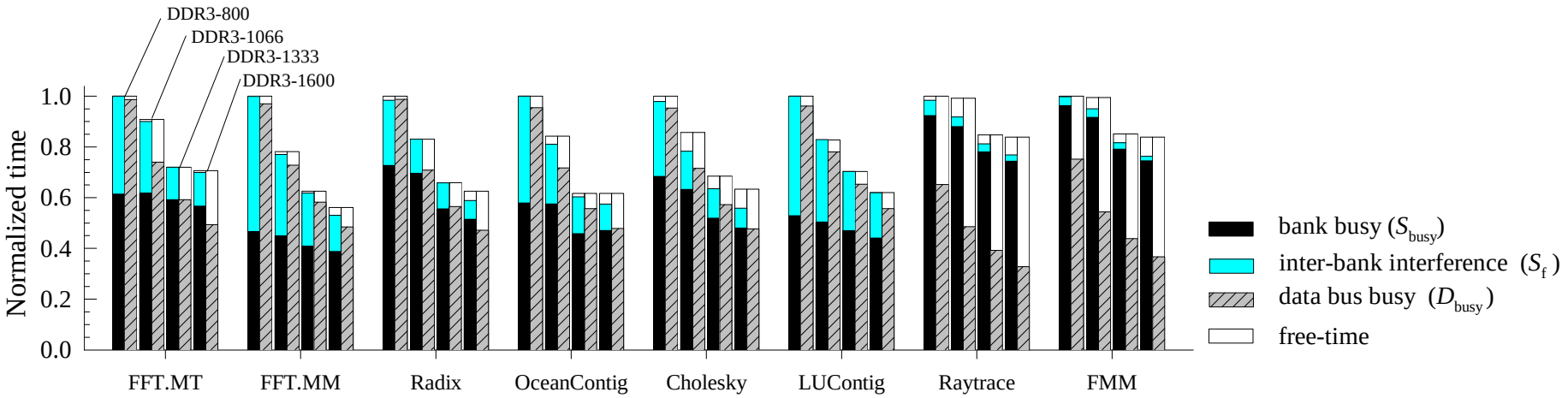


(b) weighted DRAM timings for DDR3-800

DDR3-800 timing (nsec)		weight (w_i)	weighted timing of avg. bank busy time (%)
tRP	12.5	w_0 = 0.56 ~ 0.99	17 ~ 24 %
tRAS	37.5	w_6 = 0.11 ~ 0.72	24 ~ 36 %
tCCD	10.0	w_2 = 0.27 ~ 0.82	6 ~ 17 %

cf) average bank busy time $= \boldsymbol{w_0 \cdot tRP} + w_1 \cdot tRCD + \boldsymbol{w_2 \cdot tCCD} + w_3 \cdot tCWL + w_4 \cdot tRTW + w_5 \cdot tWTR + \boldsymbol{w_6 \cdot tRAS} + w_7 \cdot tWR + w_8 \cdot tRTP$

(3) Sensitivity to DRAM clock frequency



- P=64 and without shared L2 cache (assuming intensive DRAM accesses)
- S_f := bank idle time due to inter-bank interference (measured)
- Normalized to DDR3-800 model of each workload.

(Raytrace, FMM are excluded)	**DDR3-800 (400 MHz)**	**DDR3-1600 (800 MHz)**	**diff (%)**
Execution time	1.00	0.63	- 37 %
Data transfer time ($\boldsymbol{D_{busy}}$)	0.97	0.49	- 50 %
Inter-bank interference ($\boldsymbol{S_f}$)	0.39	0.12	- 70 %
Bank busy ($\boldsymbol{S_{busy}}$)	0.60	0.48	- 20 %

Concluding remarks

- The proposed model enables quantitative analysis of the impact of DRAM timings on the access performance.
- The pattern parameters employed capture the characteristics of memory access behavior.
- It is expected to be a useful tool for providing DRAM timing guidelines in the early design stage of next DRAM standards.
- We plan to extend the model to include the amount of time delays due to inter-bank interference in our future work.