Cache Size Selection for Performance, Energy and Reliability of Time-Constrained Systems

Y. Cai*, M. T. Schmitz†, A. Ejlali†,

B. M. Al-Hashimi†, S. M. Reddy*

*Electrical and Comp. Engineering
University of Iowa
United States

†Electronics and Comp. Science
University of Southampton
United Kingdom

Overview

- Introduction
  - The affection of the cache size on performance, energy and reliability
- The models
  - transient fault model
  - performability model
  - cache energy model
- Simulation setup
- Experimental results
- Conclusions

Introduction

- Performance vs cache size:

  Generally, with the same cache configuration (same block size, same ways, …), larger the cache size, higher the processor performance. [1]

- Energy vs cache size:

  - Energy per cache access increases with the cache size [23].

  - The number of cache access is both application and cache size dependent.

- Reliability vs cache size:

  Complex, introduced with the fault model later

Introduction

- Previous work
  - only reducing the cache energy consumption. [3], [4], [5] …
  - only enhancing the cache reliability. [9], [12] …
  - considering both but not from the cache size perspective. [10]
- Our work

  examining the jointly effect of the cache size on performance, energy and reliability.
- Study method

  Simulation based on cycle-accurate simulator.

Models

- Fault model

- Cause: alpha particles [9]

- Result: bit-flip in the cache [19]

- Feature: transient, tolerated by re-execution [13]

- Uniformly distributed in space, Poisson distributed in time [9, 14]

- Reliability vs cache size

large cache size: more faults,

but more slack to re-execute

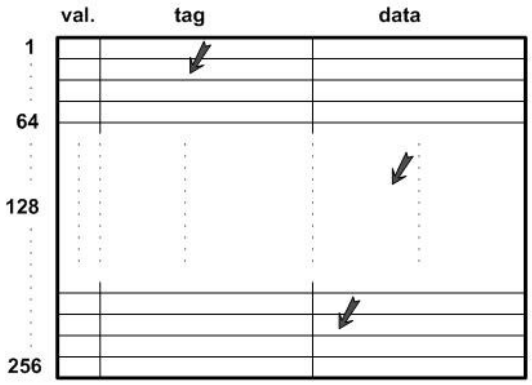
small cache size: less faults,

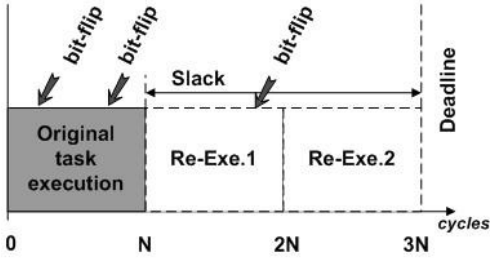
but less slack to re-execute

cache size: 256 lines

3 faults

slack for 2 re-executions

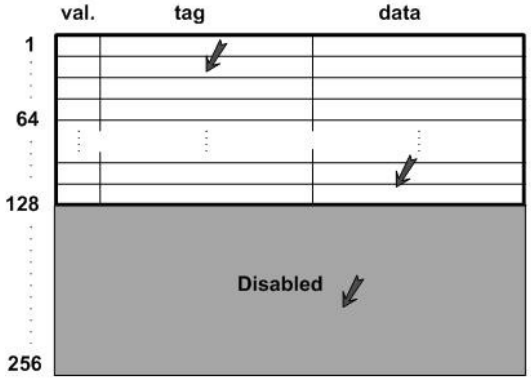


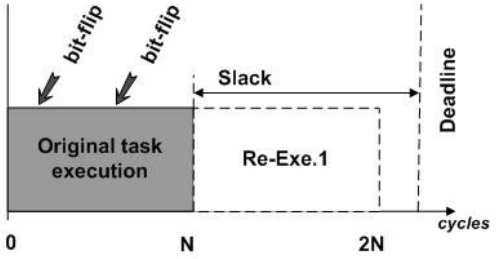


cache size: 128 lines

2 faults

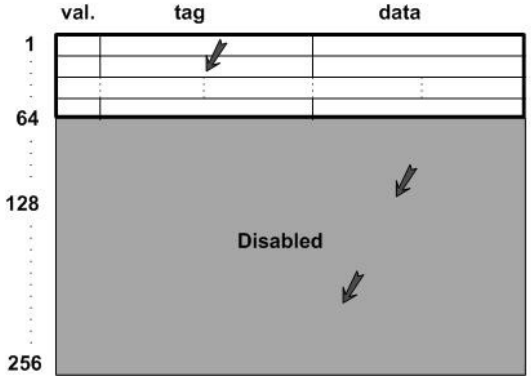
slack for 1 re-execution

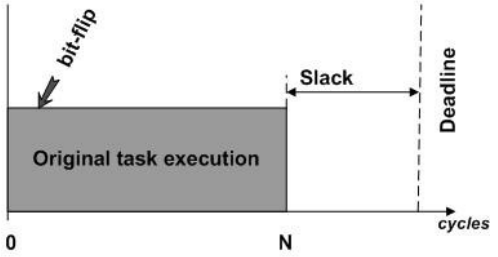




cache size: 64 lines

1 faults
Insufficient slack for re-execution





- Performability model [14]

  - Definition:

  The probability of executing a task correctly within the time-constraint

  - Feature:

  measure the performance and reliability together

  - Derivation:

  number of possible re-executions:

$$k = \left\lfloor \frac{D}{N/f} \right\rfloor - 1 = \left\lfloor \frac{D \times f}{N} \right\rfloor - 1$$

  Where *D* is the time constraint, *f* is the frequency and *N* is the clock cycles a task needs to be executed

The probability of at least one error during the execution: [14]

$$\rho_e = 1 - e^{\frac{-\lambda_{error} \times N}{f}} = 1 - e^{\frac{-VF \times \lambda_{fault} \times N}{f}}$$

Where

VF: vulnerability factor, the ratio between the number of errors and faults (faults do not necessarily cause errors)

λ_{error} : error rate, product of *VF* and λ_{fault}

λ_{fault} : fault rate, constant, measured at sea level [20]

Performability: [14]

$$P = 1 - \rho_e^{k+1} = 1 - \left(1 - e^{\frac{\lambda_{error} \times N}{f}}\right)^{\left\lfloor \frac{D \times f}{N} \right\rfloor}$$

- Energy model

$$E = E_{read} \times N_{read} + E_{write} \times N_{write}$$

where E_{read} / E_{write} is the energy consumption per read/write access, N_{read} / N_{write} is the number of cache read/write accesses.

Simulation setup

Simulator: MPARM [24], cycle-accurate, ARM7 microprocessor

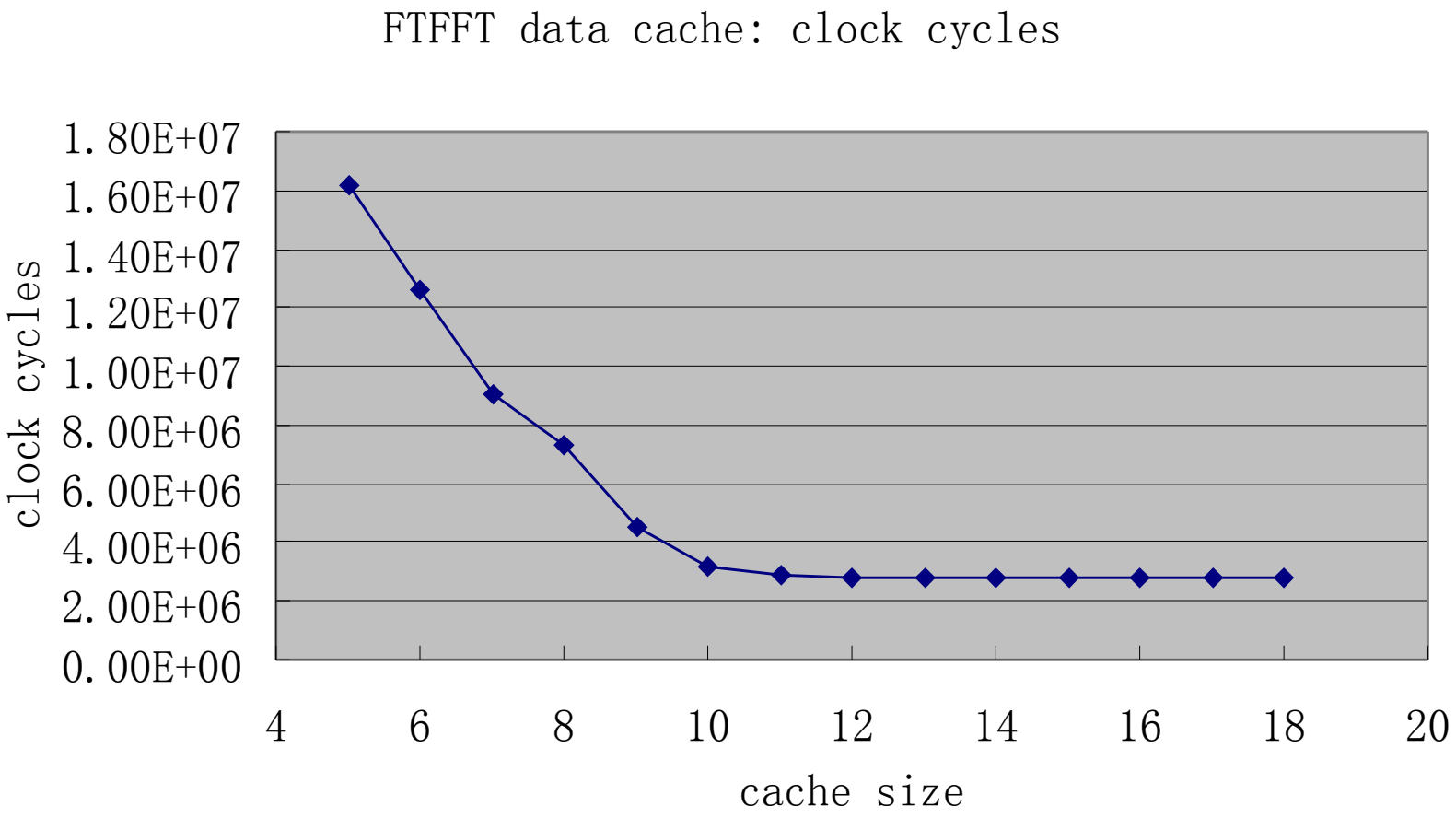
Cache configuration: separated data and instruction cache, maximum size of 256K bytes, minimum size of 32 bytes

Fault injection: inject faults into the cache during the execution and count the number of error results to obtain the vulnerability factor, which is used to compute the performability.

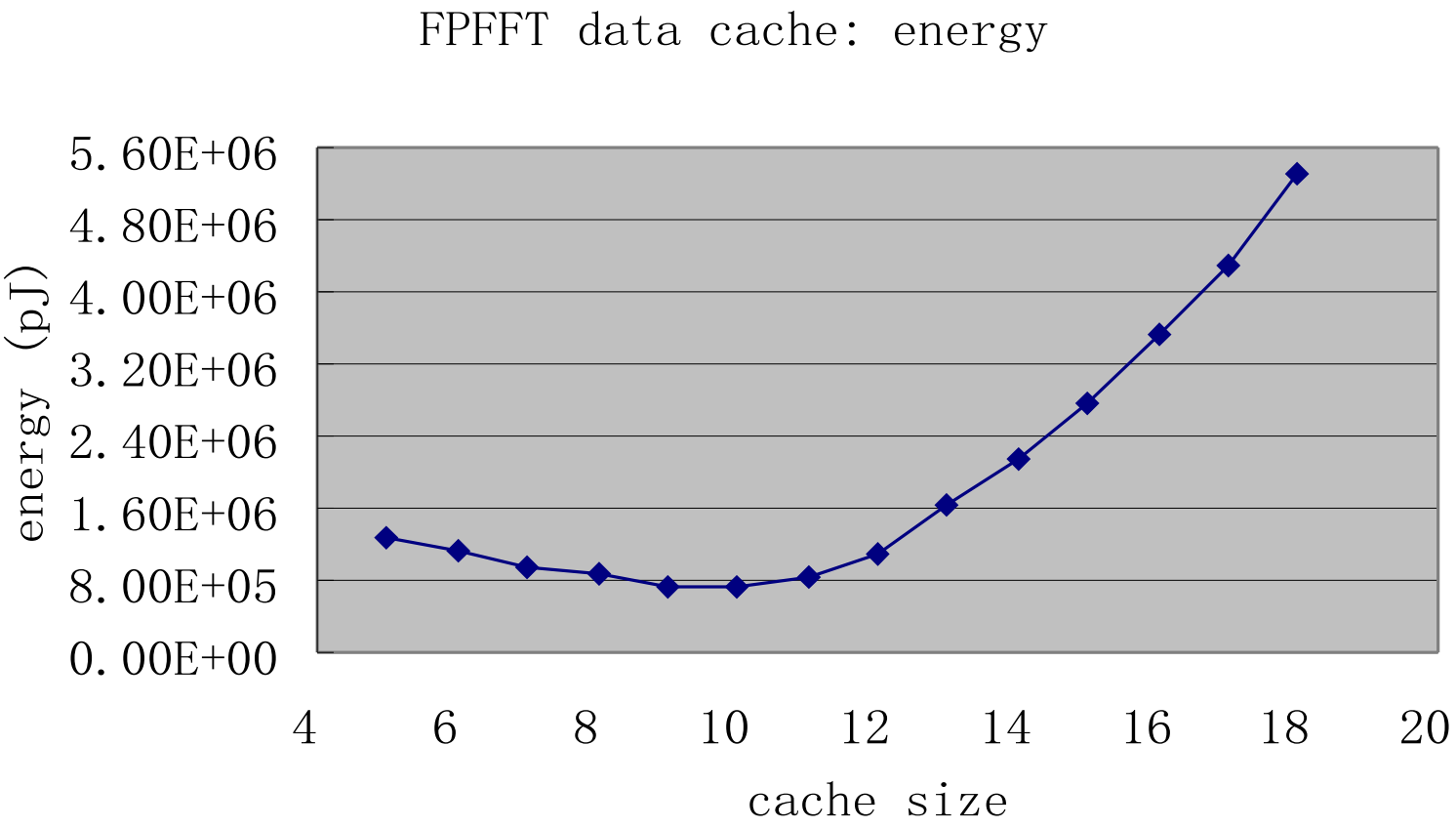
Simulation setup

- Benchmarks:
  - fixed point FFT (FPFFT)
  - cyclic redundancy check (CRC)
  - matrix multiplication (MM)
  - matrix addition (MA)
  - quick sort (QSORT)

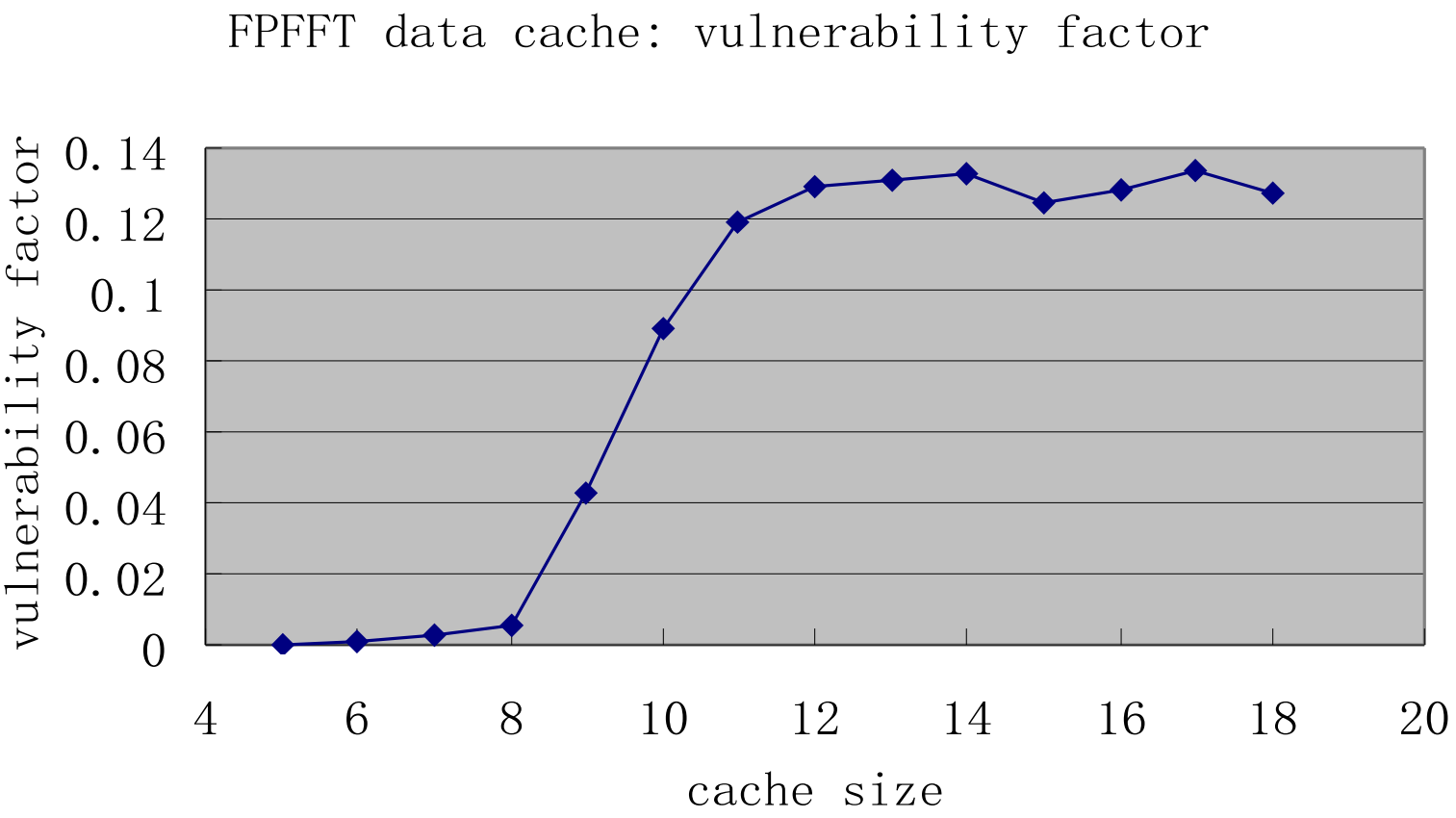
Experimental Results



Experimental Results



Experimental Results



Experimental Results

FPFFT data cache: performability

Cache size	Number of 9's	Digits after 9
5	6	89742
6	6	74106
7	6	59974
8	12	52998
9	16	69592
10	26	81057
11	26	52011
12	26	39275
⋮		
18	26	42730

Experimental Results

For FPFFT benchmark, 2^{10} bytes is the optimal data cache size in terms of both energy and performability.

- Other benchmarks

  CRC: Pareto-optimal set { 2^9, 2^{10} }

  MM: optimal size 2^{10}

  MA: optimal size 2^9

  QSORT: optimal size 2^9

- Instruction cache

  FPFFT: Pareto-optimal set { 2^9 , 2^{10} }

  CRC: optimal size 2^9

  MM: optimal size 2^8

  MA: Pareto-optimal set { 2^7, 2^8 }

  QSORT: Pareto-optimal set { 2^8, 2^9 }

Conclusions

- Jointly impaction of cache size selection on performance, energy and reliability is studied through simulation
- Performability is used to combine the analysis of the performance and reliability
- Cache size should be carefully selected to find optimal energy/performability trade-off points

Thank you!

For further questions:

Yuan Cai (yuan-cai@uiowa.edu)