Block Remap with Turnoff: A Variation-Tolerant Cache Design Technique

Mohammed Abid Hussain
IIIT-H, INDIA
abid@research.iiit.ac.in

Madhu Mutyam
IIT-M, INDIA
madhu@cse.iitm.ac.in

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Introduction

- What are process variations?
- Why are process variations so important in the present technology?
- Types of process variations
  - Die to die
  - With in die

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Effects of process variations on a cache

- Access time and leakage power increases.
- In 130 nm technology due to process variations
  - 30% variation in the maximum allowable frequency
  - 5 fold increase in the leakage power dissipated.

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Worst case design techniques

- Access the whole cache with worst case access latency
  - Problem
    - Inefficient, if the percentage of affected sets are less
- Turn off all the affected portions and access the remaining part with low latency
  - Problem
    - Inefficient, if the percentage of affected sets are more

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Block Remap with Turnoff technique (BRT)

Two steps

1) Block Remap

2) Block Turnoff

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Block Remap

- Main idea
  - Distribute all the affected blocks among all the sets as equally as possible
- Implementation
  - One remap register per way. Final index is obtained by XOR-ing the original index with the remap register
  - During the pre-processing the remap register contents are chosen such that each set gets almost equal number of affected blocks
  - Remap register contents are restricted to either all zeros or a one hot code to reduce the preprocessing time

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Block Turnoff

- Main Idea
  - If the affected, high latency blocks in a set are turned off, then the remaining un-affected blocks of the set can be accessed with low latency
- Implementation
  - Turnoff all the affected blocks in a set if there exists at least one defect free block in it and access that set with low latency
  - Affected blocks in a set are not turned off if there is no defect free block in their set, and the set is accessed with high latency

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Example

	Way 0	Way 1	Way 2	Way 3
Set 0	0	0	0	0
Set 1	1	1	1	1
Set 2	2	2	2	2
Set 3	3	3	3	3
Set 4	4	4	4	4
Set 5	5	5	5	5
Set 6	6	6	6	6
Set 7	7	7	7	7

(a) CAS

	Way 0	Way 1	Way 2	Way 3
	0	2	4	2
	1	3	5	3
	2	0	6	0
	3	1	7	1
	4	6	0	6
	5	7	1	7
	6	4	2	4
	7	5	3	5

(b) BRT

Optimum register remap = 000 010 100 010

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

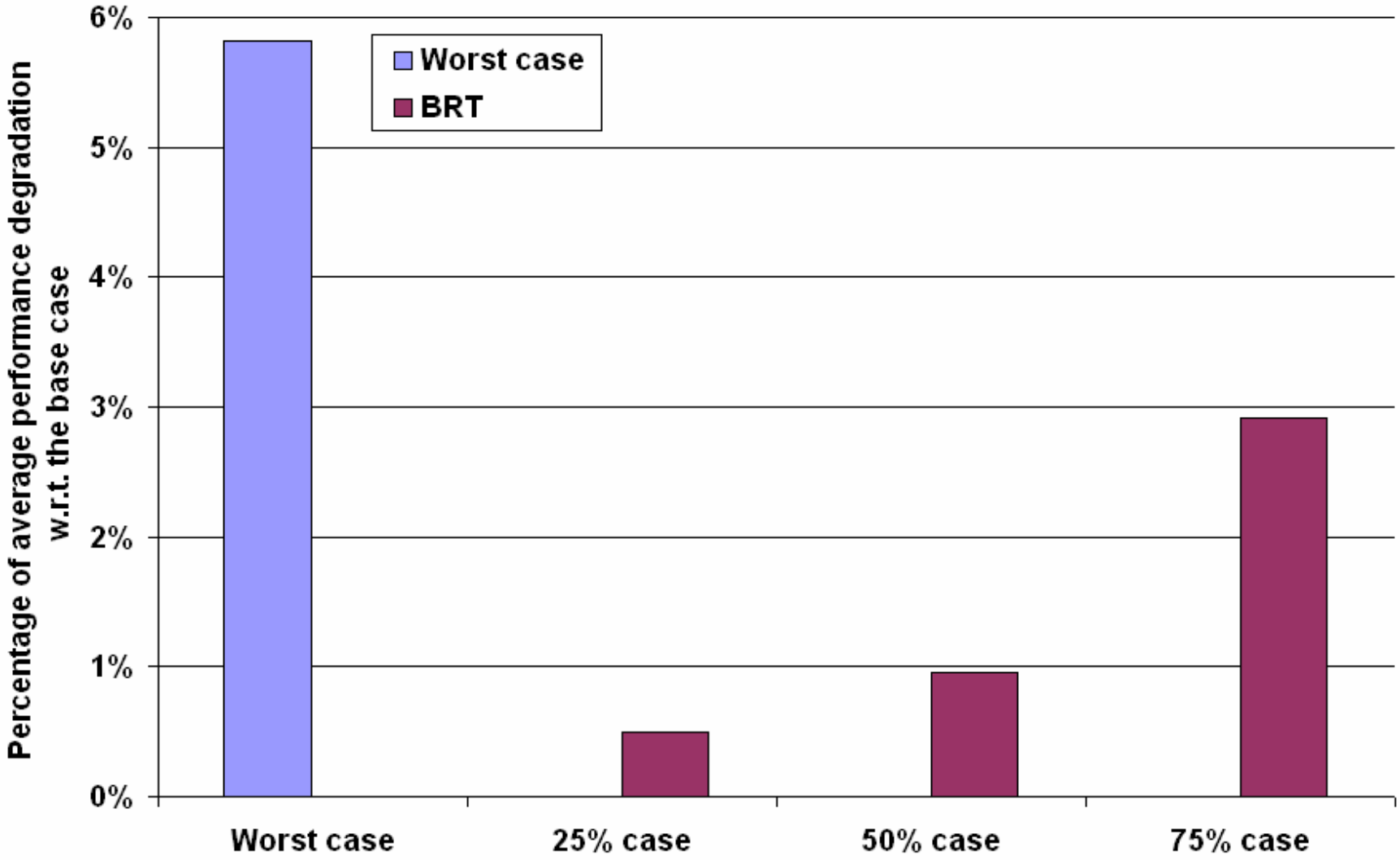
Dealing with variable cycle access latency

- To deal with non uniform access latency, we use the 2 delta stride based latency predictor.
- The dependant instructions are issued based on the predicted access latency of the load instruction.
- If the prediction is wrong, the pipeline is flushed and the instructions are executed again.

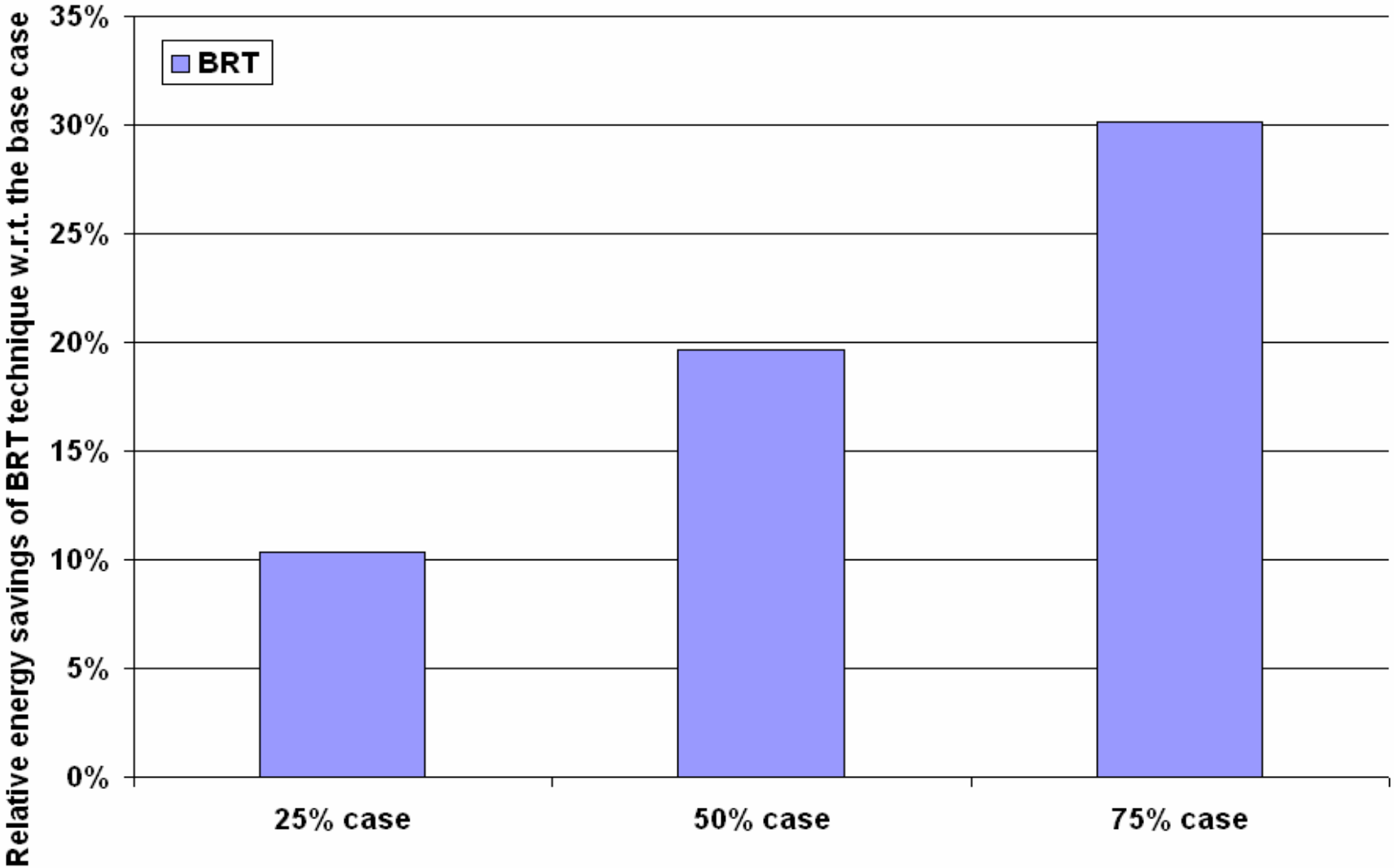
Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Results


Percentage of average performance degradation w.r.t. the base case
6%
5%
4%
3%
2%
1%
0%
Worst case
BRT
Worst case
25% case
50% case
75% case

Results (cont'd)



Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Conclusion

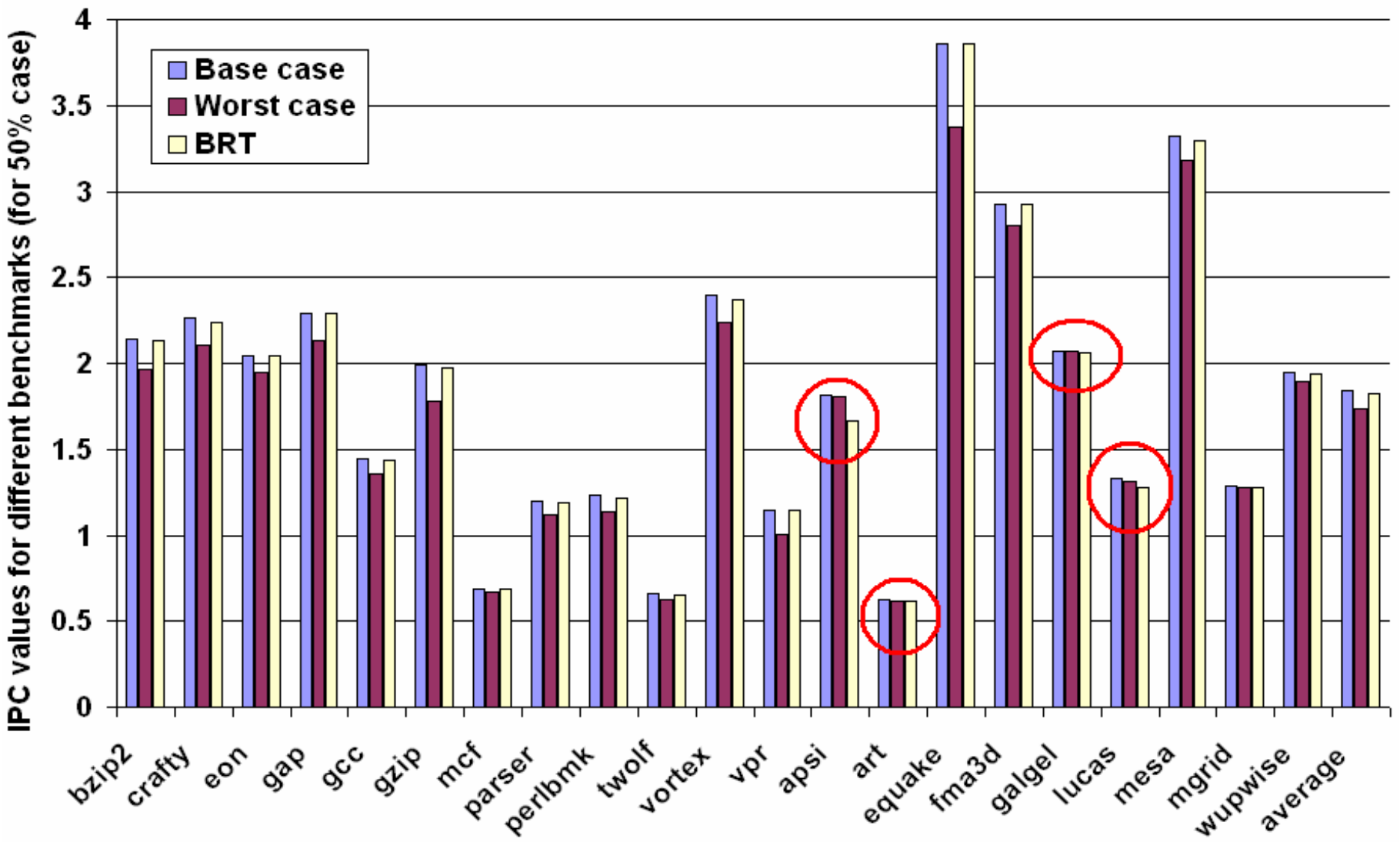
- Process variations pose a serious problem in advanced technologies, leading to reduced chip yield.

- BRT technique reduces
  - Reduces performance penalty
  - Reduces excess leakage power dissipation

Outline

- Introduction
- Effects of process variations on a cache
- Worst case design techniques
- Block Remap with Turnoff technique
  - Block Remap
  - Block Turnoff
  - Example
  - Dealing with variable cycle access latency
- Results
- Conclusion
- Future work

Future work


IPC values for different benchmarks (for 50% case)
Base case
Worst case
BRT
4
3.5
3
2.5
2
1.5
1
0.5
0
bzip2
crafty
eon
gap
gcc
gzip
mcf
parser
perlbmk
twolf
vortex
vpr
apsi
art
equake
fma3d
galgel
lucas
mesa
mgrid
wupwise
average

Thank you.

Any questions?