INTERMITTENT HARDWARE ERRORS RECOVERY: MODELING AND EVALUATION

LAYALI RASHID, KARTHIK PATTABIRAMAN AND SATHISH GOPALAKRISHNAN
INTERMITTENT FAULTS-DEFINITION

• Hardware errors that appear non-deterministically at the same microarchitectural location.
• 40% of the real-world failures in processors are caused by intermittent faults [1].
CONTRIBUTIONS

• Build a model of chip multiprocessor running a parallel application using Stochastic Activity Networks.

• Propose intermittent fault models that abstract real intermittent faults at the system level.

• Evaluate the performance of a processor after applying different recovery options.
RECOVERY-MOTIVATION
Program Execution

CHKPT

Hardware Error

Problem!
RECOVERY-MOTIVATION

- Transient Hardware Error
- Problem!
- Restore to Checkpoint
- Recovery
Program Execution

Permanen\textbf{t} Hardware Error

Problem!

Checkpoint

Restore to Checkpoint

Core Reconf.

Core

Recovery

RECOVERY-MOTIVATION
Program Execution

Intermittent Hardware Error

Problem!

CHKPT

Restore to Checkpoint

Core Reconf.

Core

Recovery

RECOVERY-MOTIVATION
MODEL OVERVIEW

- Rollback-Only
- Permanent Reconfiguration
- Temporary Reconfiguration

- Base
- Exponential
- Weibull
KEY FINDINGS

• Error rate and the relative importance of the error location are the main factors in finding the best recovery for high intermittent failure rates.

• Permanent shutdown of the defective unit results in a slight improvement of the performance compared to the temporary shutdown.
PROCESSOR MODEL

Program Execution → Error Detection

Error Detection → Rollback to Checkpoint

Rollback to Checkpoint → Reconfigure?

Reconfigure? → Yes

Yes → Fine-Grained Diagnosis

Fine-Grained Diagnosis → Rollback-Delay Scenario

Rollback-Delay Scenario → Unit Shutdown-Permanent

Unit Shutdown-Permanent → Perf. Degradation

Perf. Degradation → Program Execution
PROCESSOR MODEL

1. Program Execution
2. Error Detection
3. Error
4. Reconfigure?
5. Rollback to Checkpoint
6. No
7. Reconfigure?
8. Yes
9. Fine-Grained Diagnosis
10. Unit Shutdown-Temporary
11. Full Throughput
12. Enable Unit?
13. No
14. Perf. Degradation
15. No
16. Reconfig-Temporary
FAULT MODEL-BASE FAULT MODEL

- Abstract physical fault models.
- Prune down the space of system configurations.
FAULT MODEL-EXPONENTIAL FAULT MODEL

- Abstract physical fault models.
- Prune down the space of system configurations.
FAULT MODEL - WEIBULL FAULT MODEL

- Abstract physical fault models.
- Prune down the space of system configurations.

\[ \lambda_1, \sigma \]

\[ \lambda_2 \]

\[ 1 - p \]

\[ p \]
EXPERIMENT SETUP

• Used Mobius\textsuperscript{[2]} to simulate the system for 48 hours with a confidence interval of 95%.

• Used useful work\textsuperscript{[3]} measure to model processor throughput in a certain amount of time.

• Analyzed a model of multiprocessor running coordinated checkpoint.
SYSTEM PARAMETERS

Checkpoint

Program Execution

Error

Error Detection

Rollback to Checkpoint

Reconfigure?

No

70% Accuracy

Yes

17 sec/5-60min

Perf. Degradation 0-35%

Unit Shutdown-Permanent

Fine-Grained Diagnosis

30sec/5-60min

2sec

0-60sec
RESEARCH QUESTIONS

• When should we recover from an intermittent fault by shutting down the defective component?

• For errors that are tolerated by shutting down the defective component, should the shutdown be permanent or temporary?
• Permanent/temporary reconfiguration leads to 27% more useful work than rollback-only for exponential and Weibull fault models.
What is the granularity of the disabled component that maximizes the processor’s performance?
COMPONENT RANK

• The maximum percentage of useful work that is lost when the component is disabled.

• 4-core processor, each core has two LSUs and is running a program that is using all the 8 LSUs for 60% of the time.

• Using Amdahl’s law, LSU rank is 19% or \( \frac{1}{0.4 + \frac{0.6}{0.125}} \)
• For this experiment, components with a rank of 35% or more should be disabled if diagnosed with intermittent errors.
Sensitivity to Fault Rate
RESULTS- SENSITIVITY TO FAULT RATE

• If lost useful work outweighs the rank of the defective component, then the defective component should be disabled.
KEY FINDINGS

• Error rate and the relative importance of the error location are the main factors in finding the best recovery for high intermittent failure rates.

• Permanent shutdown of the defective unit results in a slight improvement of the performance compared to the temporary shutdown.