Thermal-aware Dynamic Checkpoint Interval Tuning for High Performance Computing

Pei Li, Mulya Agung (Graduate School of Information Sciences, Tohoku University)
Muhammad Alfian Amrizal, Ryusuke Egawa, Hiroyuki Takizawa (Cyberscience Center, Tohoku University)

Cyberscience Center

TOHOKU UNIVERSITY

Introduction

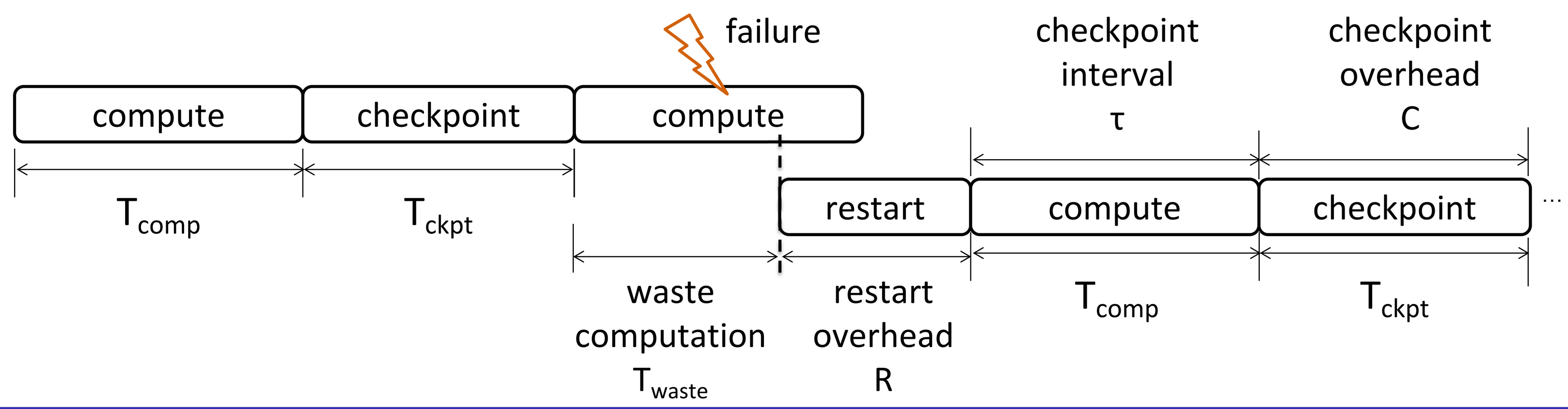
High Performance Computing (HPC) & Checkpoint/Restart

HPC systems are getting faster and larger

Users want to submit more complex and long-running jobs. The application execution will face a higher probability of encountering failures because of longer execution time.

Fault-tolerance mechanisms are required

Checkpoint/Restart (CPR) is one of the most commonly used fault-tolerance mechanisms.



Optimal Checkpoint Interval

CPR needs additional time overhead for storing **a large amount of data** to a stable storage.

More frequent checkpoints: will delay application's progress.

Less frequent checkpoints: will not be able to protect the execution from failures effectively.

Optimal checkpoint interval must be carefully selected to minimize the total execution time.

Problem Statement

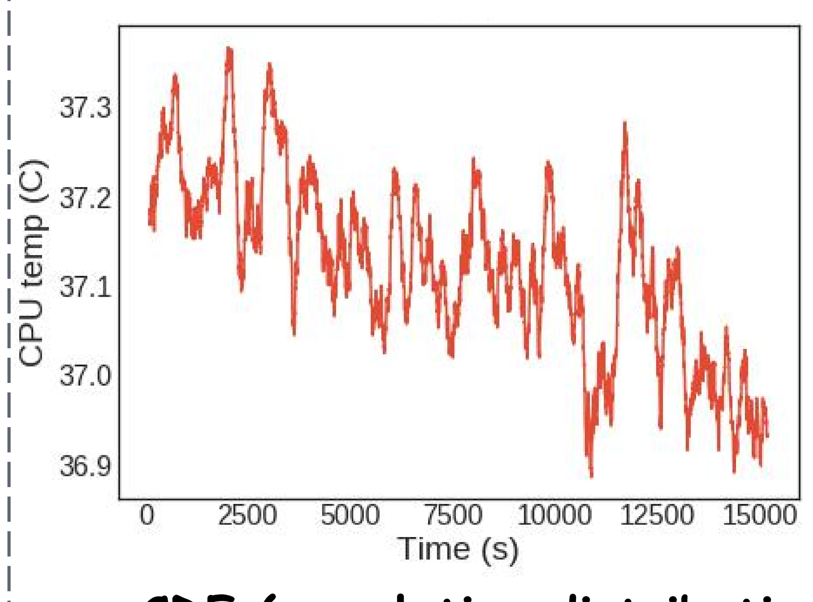
- Many related researches assume that failures are exponentially distributed with a constant failure rate, **λ**.
- Previous work has demonstrated that the reliability of systems can be affected by many factors [1].
  - One of the determinant factors is the ***operating temperatures*** of the system. **λ** dynamically changes with the change of operating temperature.

Therefore, deriving an *optimal checkpoint interval* by assuming an exponential distribution of failure with a constant λ, might not lead to a minimal execution time.

[1] Tang, Kun, et al. "Power-capping aware checkpointing: On the interplay among power-capping, temperature, reliability, performance, and energy." Dependable Systems and Networks (DSN), 2016 46th Annual IEEE/IFIP International Conference on. IEEE, 2016.

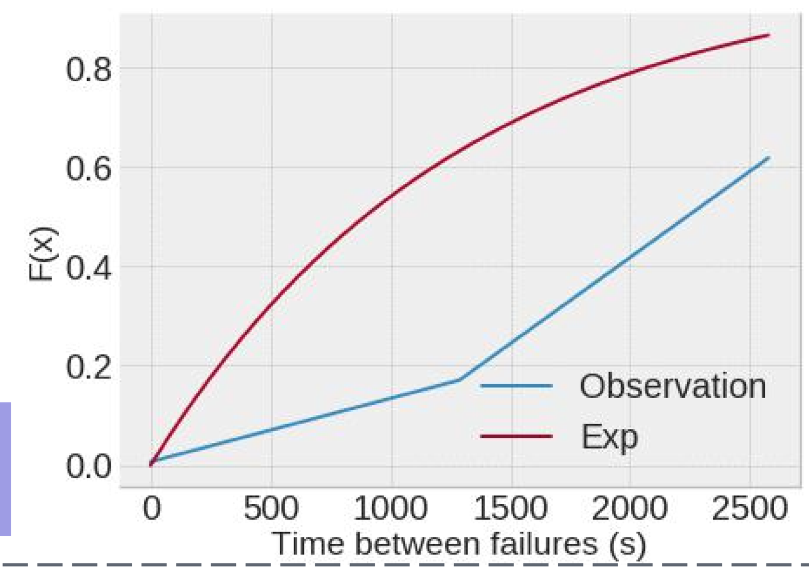
Monitoring Results

Conjugate Gradient (CG)



Temperature monitoring results of the CG application (left)

Temperature changes during the execution. Higher temperatures are shown at the beginning of the execution.

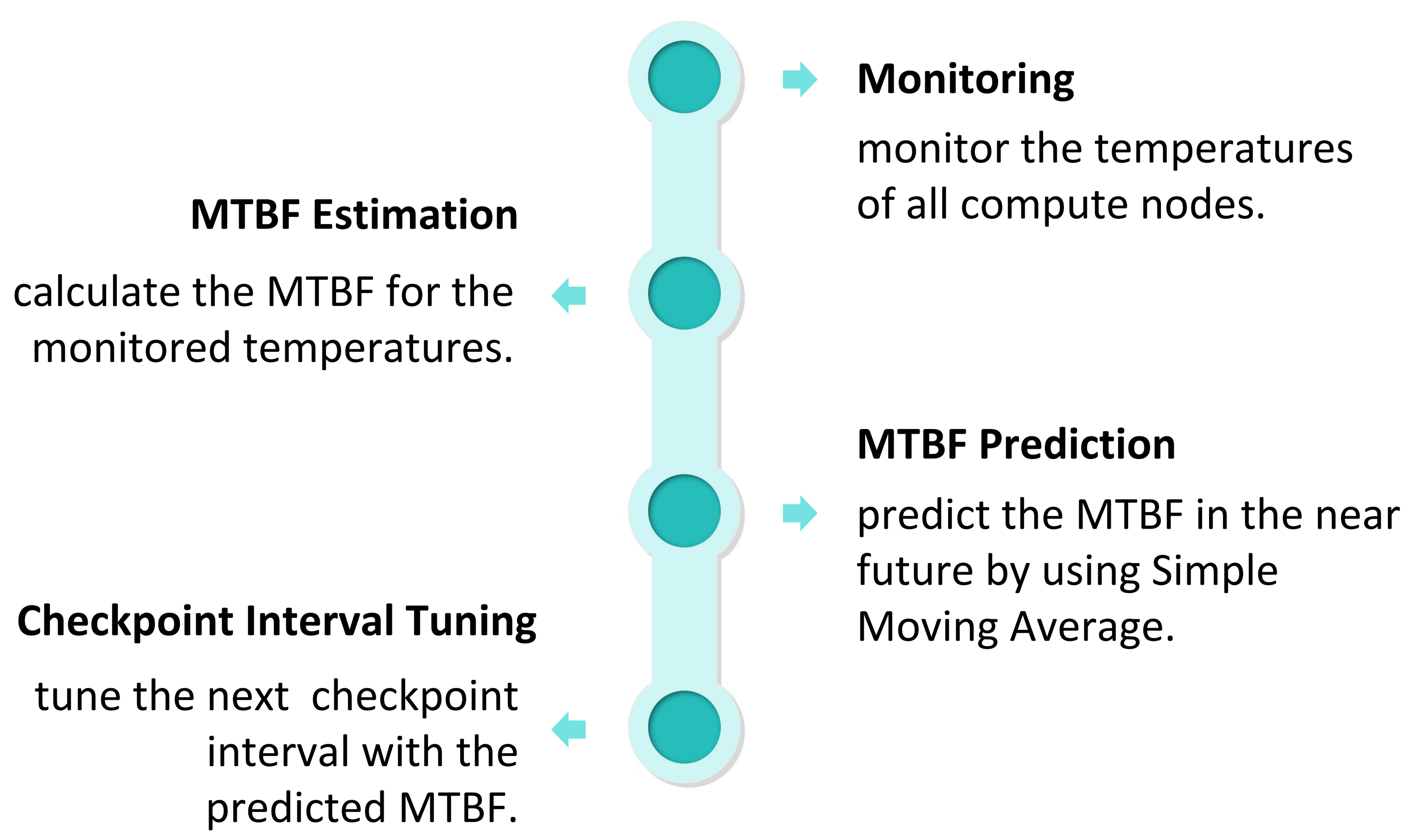


CDF (cumulative distribution function) of the failures for the CG application (right)

The exponential distribution does not fit well observations.

Proposed Method

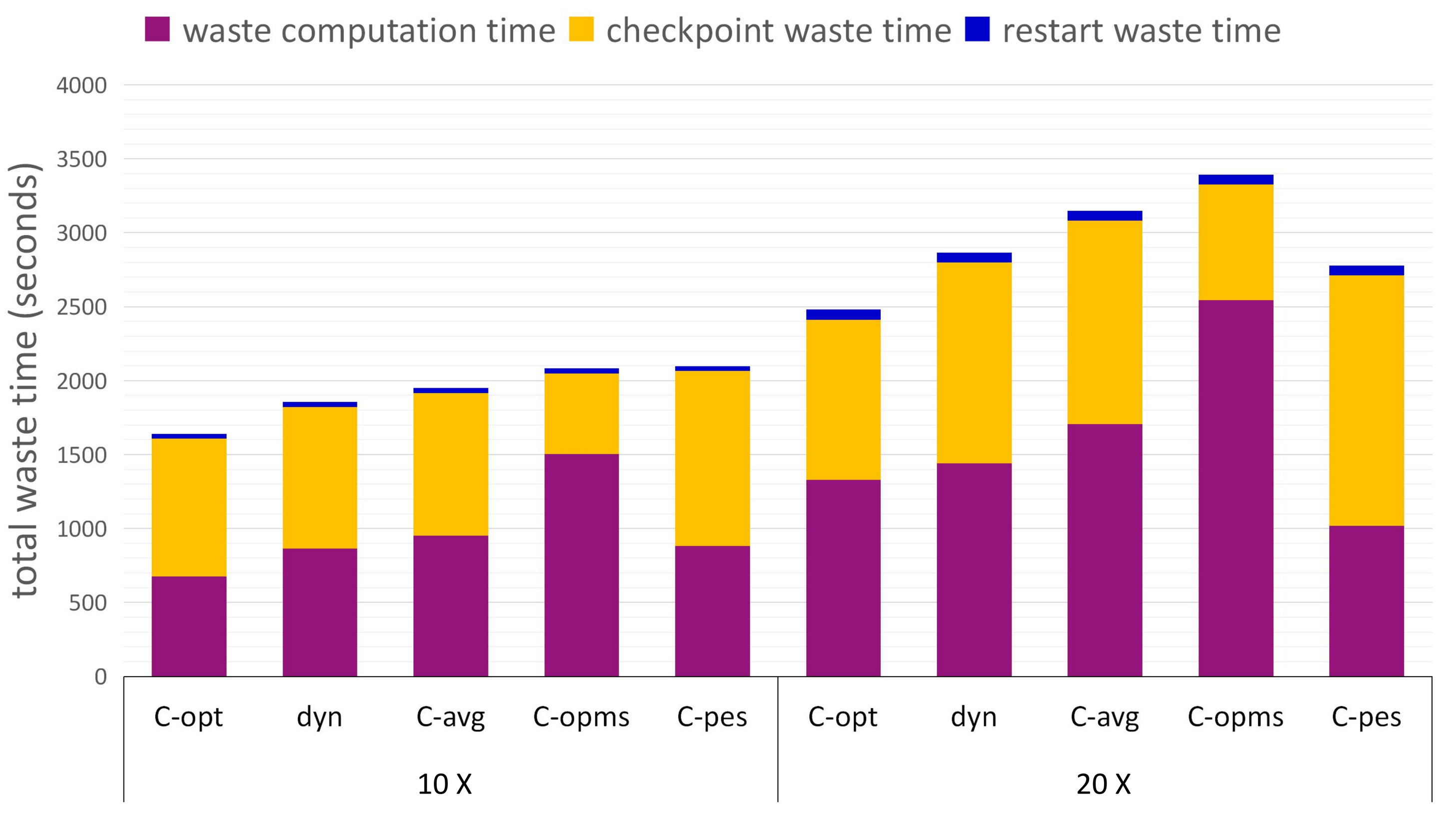
Thermal-aware Dynamic Checkpoint Interval Tuning



Evaluation

Experimental Setup	
OS	Linux 4.4 (Ubuntu)
Processor	Intel Core i7-6700 @ 3.40GHz
Number of cores	4

Checkpoint/Restart Setup	
checkpoint overhead	1 s
restart overhead	2 s



* To simulate a large scale system, the failure rates are multiplied by factors of 10 and 20.

Dynamic (dyn): The proposed method.
Constant Optimal (C-opt): The optimal checkpoint interval is obtained by using a brute-force search.
Constant Average (C-avg): The checkpoint interval is determined by the average failure rate during the execution.
Constant Optimistic (C-opms): The checkpoint interval is determined by the minimum failure rate.
Constant Pessimistic (C-pes): The checkpoint interval is determined by the maximum failure rate.

Conclusions & Future Work

- **The evaluation results show that the proposed method can achieve a comparable performance to the Constant Optimal method (brute-force search) which is near-ideal.**
- **In the future, this work will focus on improving the method for larger systems and distributed applications.**

HPC Asia 2018

(E-mail) {lipei, agung, alfian}@sc.cc.tohoku.ac.jp, {egawa, takizawa}@tohoku.ac.jp