ICAC 2015

July 07-10 2015
Grenoble, France

Integrity Protection for Big Data Processing with Dynamic Redundancy Computation

Zhimin Gao[1], Nicholas DeSalvo[1], Pham Dang Khoa[1], Seung Hun Kim[1,2], Lei Xu[1], Won Woo Ro[2], Rakesh M.Verma [1], and Weidong Shi[1]

{zgao5, nsdesalvo,pdkhoa, skim76, lxu13,rmverma2, wshi3}@uh.edu, wro@yonsei.ac.kr

University of Houston, TX, United State[1] Yonsei University, Seoul, Korea[2]

Introduction

- The development of cloud computing technology provides an adequate platform for big data applications.
- Most existing works focus on protection of data privacy but integrity protection of the processing procedure receives little attention, which may lead the big data application user to wrong conclusions and cause serious consequences.
- Solution: An integrity protection solution for big data processing in cloud environments using reputation based redundancy computation. The implementation and experiment results show that the solution only adds limited cost to achieve integrity protection and is practical for real world applications.

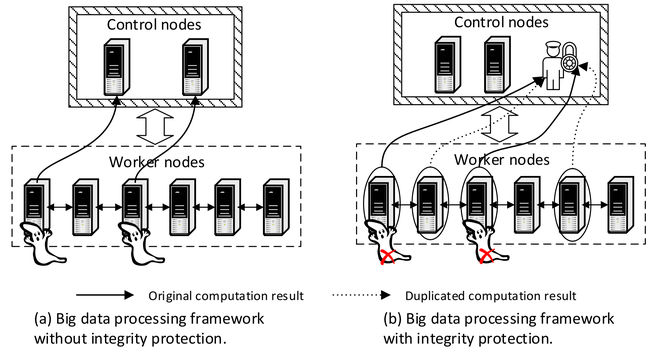
Major Contributions

Our contributions in this work are summarized as follows:

- A design of reputation based trust rating system for big data processing;
- A proposal of dynamic policy for redundant computation to enhance integrity;
- A demonstration of the feasibility of the proposed scheme using the implementation results with MapReduce framework.

Motivation and the Approach

- Cloud computing is turning from a concept to a mature technology;
- Integrity of data processing procedure is a big concern when adopting cloud computing
  - If an adversary damages the integrity of this process by modifying, deleting, or inserting into the intermediate results, it may lead to a totally wrong conclusions;
  - Most of existing protection techniques focus on privacy/confidentiality.

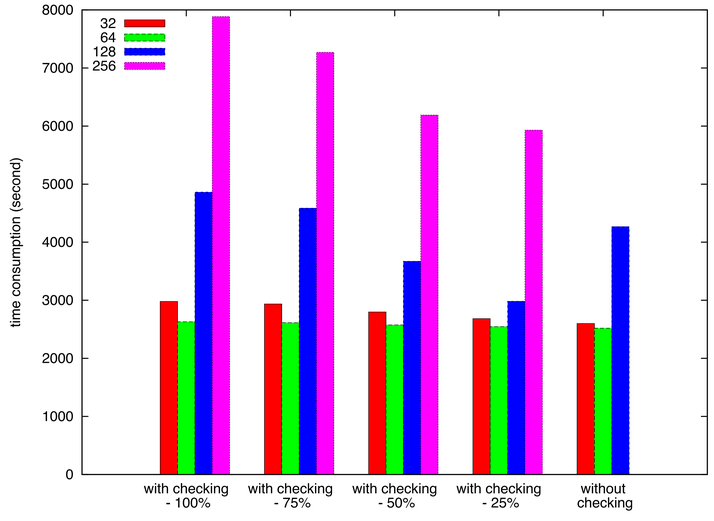


(a) Big data processing framework without integrity protection.

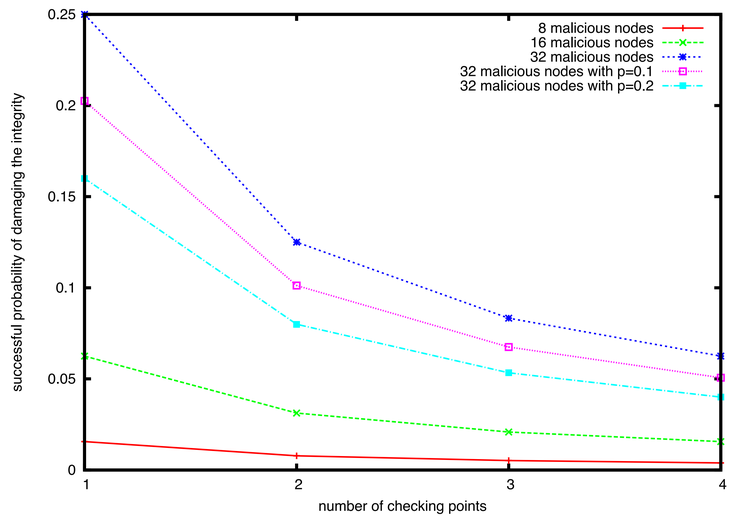
(b) Big data processing framework with integrity protection.

Big data integrity protection and threat model. The control nodes are trusted while the worker nodes may be compromised. In a nutshell, control nodes schedule some redundancy computation on different worker nodes and figure out potentially malicious worker nodes by comparing their computation results.

Performance Evaluation



Comparison of time consumptions. Each group of columns indicates a certain ratio of duplicated nodes. The columns in the group stand for the time consumption by running different number of mappers.



The simulation results of the detection probability. The probability that a malicious worker node is not detected decrease quickly when more checks are done.