EFFICIENT SLEPIAN-WOLF BASED PROOF OF RETRIEVABILITY USING SPLITTING AND PARTITIONING SCHEME FOR CLOUD STORAGE

TAN CHOON BENG

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2018

EFFICIENT SLEPIAN-WOLF BASED PROOF OF RETRIEVABILITY USING SPLITTING AND PARTITIONING SCHEME FOR CLOUD STORAGE

TAN CHOON BENG

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2018

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL: **EFFICIENT SLEPIAN-WOLF BASED PROOF OF RETRIEVABILITY USING SPLITTING AND PARTITIONING SCHEME FOR CLOUD STORAGE**

IJAZAH: **IJAZAH SARJANA SAINS (SAINS KOMPUTER)**

Saya **TAN CHOON BENG**, Sesi **2017-2018**, mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:

1. Tesis ini adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (/):

□	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972)
□	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
/	TIDAK TERHAD	

Disahkan Oleh,

TAN CHOON BENG
MI1611010T

(Tandatangan Pustakawan)

Tarikh: 7 Jun 2018

(Dr. Mohd Hanafi Ahmad Hijazi)
Penyelia

DECLARATION

I hereby declare that the material in this thesis is my own work except for certain quotations, equations, summaries, definitions, and references, which have been duly acknowledged.

22 DECEMBER 2017

....

TAN CHOON BENG

MI1611010T

CERTIFICATION

NAME	:	**TAN CHOON BENG**
MATRIC NO.	:	**MI1611010T**
TITLE	:	**EFFICIENT SLEPIAN-WOLF BASED PROOF OF RETRIEVABILITY USING SPLITTING AND PARTITIONING SCHEME FOR CLOUD STORAGE**
DEGREE	:	**MASTER OF SCIENCE (COMPUTER SCIENCE)**
VIVA DATE	:	**21 MAY 2018**

CERTIFIED BY:

SUPERVISOR **Signature**
Dr. Mohd. Hanafi Ahmad Hijazi

____

ACKNOWLEDGEMENT

The completion of this master thesis is impossible without the kindness, help and guidance of several grateful parties. There are not much people who willing to stand by our side, encourage and support us to achieve our goals, regardless of favorable or unfavorable situations. Here, I would like take this opportunity to express my deepest gratitude and appreciation to them whom have been constantly assisting and supporting me along the way.

First of all, I would like express my sincere thankfulness to my main advisor, Dr. Mohd Hanafi Ahmad Hijazi who has been serving as a senior lecturer in Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS). He is my advisor since my Bachelor degree. Without his patience, motivation and continuous supports and advices, I am afraid that I would not able to make this far gratefully.

I would like to thank my co-advisor Associate Professor Yuto Lim from Japan Advanced Institute of Science and Technology (JAIST). I am grateful for his precious comments and advices which have always broaden my horizons on scientific research. His dedication in supervising always inspire me towards a new and feasible direction throughout my research.

Last but not least, I would like to express my deepest gratitude to my beloved family for mental support throughout my Master degree, although we were always a sea apart. Sincerely, I devote a thousand thanks to my caring family.

Tan Choon Beng
22 December 2017

ABSTRACT

Cloud storage is an online storage service offered by Cloud Service Provider (CSP), where client's data is hosted on cloud servers' side. However, as client does not have physical access to outsourced data, cloud storage sometime labelled as untrustworthy or semi-trustworthy. To ensure cloud data integrity and availability, which are the precondition for the existence of a cloud storage system, a protocol known as Proof of Retrievability (PoR) is introduced. PoR allows cloud storage to proof to the client that the stored data is intact and fully retrievable. Recently, Slepian-Wolf Based Proof of Retrievability (SW-PoR) was introduced to provide cost-efficient and time consistent exact repair mechanism for erroneous outsourced data. However, to achieve maximum resiliency for data correctness, encoding process of SW-PoR requires considerably long computational time compared to the conventional storage method involving replication. Hence, this research proposed two viable solutions as extension to SW-PoR to address this limitation. The solutions are named as Partial Binary Encoding for SW-PoR (PBE-SW-PoR) and Optimized SW-PoR (Opti-SW-PoR). PBE-SW-PoR allows part of the data, *A*, to be encoded by SW-PoR while the other part of the data, *B*, is secured by adapting Cyclic Redundancy Check (CRC) and replication. Opti-SW-PoR adapted the concept of partitioning to reduce computation time of SW-PoR. Simulation was conducted to evaluate the performance of the proposed solutions by means of comparison to the original SW-PoR scheme in term of computation time. In the simulation, PBE-SW-PoR and Opti-SW-PoR showed significant reduction with respect to total computation time compared to the original SW-PoR. At data size of 1,000 file blocks, original SW-PoR recorded 835,205.4 seconds of total computation time. In comparison to original SW-PoR, PBE-SW-PoR shorten the total computation time by 89.72% while Opti-SW-PoR shorten the total computation time by 99.99%.

ABSTRAK

BUKTI KEBOLEHULANGAN SEMULA BERASASKAN SLEPIAN-WOLF YANG CEKAP DENGAN TEKNIK PEMISAHAN DAN PARTISI UNTUK DATA AWAN

Penyimpanan awan merupakan satu perkhidmatan penyimpanan data atas talian yang disediakan oleh Pembekal Perkhidmatan Awan (CSP), di mana data pelanggan dihoskan di server awan, di mana pelanggan perlu membayar sejumlah bayaran kepada CSP berdasarkan kadar penggunaan. Namun, disebabkan pelanggan tiada akses fizikal ke data penyumberan luar, penyimpanan awan kerap dikatakan kurang dipercayai atau semi-dipercayai. Demi memastikan integriti dan ketersediaan data awan yang menjadi prasyarat kewujudan sesuatu sistem penyimpanan awan, protokol yang dikenali sebagai Bukti Kebolehulangan Semula (PoR) telah diperkenalkan. PoR membenarkan penyimpanan awan untuk membuktikan kepada pelanggan bahawa data yang disimpan adalah utuh dan dapat dikembalikan dengan sepenuhnya. Sejak kebelakangan ini, Bukti Kebolehulangan Semula Berasaskan Slepian-Wolf (SW-PoR) telah diperkenalkan demi menyediakan kos efektif dan masa yang konsisten dalam pembaikan tepat mekanisme untuk data penyumberan luar yang rosak. Namun, untuk mencapai daya tahan lasak yang maksima untuk ketepatan data, proses pengekodan dalam SW-PoR memerlukan masa komputasi yang agak panjang berbanding dengan kaedah penyimpanan data secara tradisional seperti replikasi. Oleh itu, penyelidikan ini bercadangkan dua kaedah penyelesaian yang berpotensi sebagai kerja lanjutan SW-PoR untuk menangani isu ini. Kaedah-kaedah penyelesaian tersebut dinamakan Pengekodan Binari Separa untuk SW-PoR (PBE-SW-PoR) dan Optimum SW-PoR (Opti-SW-PoR). PBE_SW-PoR membolehkan sebahagian data sahaja, A, dikodkan, manakala bahagian yang selebihya, B, dijamin dengan mengadaptasi cek redundansi kitaran (CRC) dan replikasi. Opti-SW-PoR mengadaptasi konsep pembahagian untuk mengurangkan masa komputasi SW-PoR. Simulasi telah dijalankan untuk menilai prestasi kaedah-kaedah peneyelesaian yang dicadangkan tersebut dengan cara perbandingan dengan SW-PoR asal dari segi masa komputasi. Dalam simulasi itu, PBE-SW-PoR dan Opti-SW-PoR menunjukkan penggurangan jumlah masa komputasi yang menonjol. Pada saiz data 1,000 blok fail, SW-PoR asal memakan jumlah masa sebanyak 835,205.4 saat. Berbanding dengan itu, PBE-SW-PoR dan Opti-SW-PoR masing-masing memendekkan jumlah masa komputasi sebanyak 89.72% dan 99.99%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AES	-	Advanced Encryption Standard
bit	-	binary digit
BLS signature	-	Boneh-Lynn-Shacham signature
CRC	-	Cyclic Redundancy Check
CSP	-	Cloud Service Provider
DOJ	-	Department of Justice
EC	-	Erasure Coding
ECC	-	Error Correcting Codes
GB	-	Giga Bytes
GHz	-	Giga Hertz
kb	-	kilobits
MAC	-	Message Authentication Code
Mb	-	Megabits
NC	-	Network Coding
NSA	-	National Security Agency
Opti-SW-PoR	-	Optimized SW-PoR
OS	-	Operating System
PBE-SW-PoR	-	Partial Binary Encoding for SW-PoR
PDP	-	Provable Data Possession
PoR	-	Proof of Retrievability
PoW	-	Proof of Ownership
PRF	-	Pseudorandom Functions
RAM	-	Random Access Memory
SSL	-	Secure Sockets Layer
SW-PoR	-	Slepian-Wolf based Proof of Retrievability
TLS	-	Transport Layer Security
TPA	-	Third Party Auditor
XOR	-	exclusive OR

LIST OF SYMBOLS

$\lvert F \rvert$	-	file size
$\lvert P \rvert$	-	number of elements in the list of permutation
$\oplus$	-	XOR
A	-	Part A of data in PBE-SW-PoR
B	-	Part B of data in PBE-SW-PoR
c	-	coded block
$\hat{c}$	-	metadata
C (m, 3) or $\binom{m}{3}$	-	Combination of m choose 3 (Binomial Coefficient)
crc	-	size of CRC bits
F	-	file
m	-	number of file blocks
m'	-	number of file blocks per partition
n	-	number of XORs
n'	-	number of XORs per partition
p	-	number of partitions
q	-	number of file blocks in A
r	-	redundancy per file block
rep	-	total size of replicates of CRC added B
s	-	size of a pair of coded block and metadata
S	-	storage cost
$serv$	-	number of servers
$\overline{w}$	-	file block
$\lvert\overline{w}\rvert$	-	block size

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Cloud computing is a distributed, online based service provided by Cloud Service Provider (CSP), where online resources such as storage spaces and computing power are shared among cloud users (Mell and Grance, 2011). Due to the ubiquitousness and accessibility, the demand on cloud computing has increased over the years (Statista, 2017) while the highest global spending is on the Infrastructure as a Service (IaaS) (Srinivas, Reddy, and Qyser, 2014). Although the storage trend has shifted to cloud storage due to its advantages such as efficient telecommute, ubiquitous data storage and backup, as well as disaster recovery services (Baciu, 2015), but some organisations still using traditional data storage method via local servers due to some factors like vendor lock-in, reliability, privacy, pricing, interoperability, and security concern (Quest, 2015) (Nedelcu, Stefanet, Tamasescu, Tintoiu, and Vezeanu, 2015). Nonetheless, the ubiquitousness of cloud computing leads to high probability of cloud incidents like unauthorised access, confidential disclosure and financial loss. For instances, several well-known gigantic CSPs such as Amazon, Google, Microsoft and Sony have suffered from cloud incidents (Ko, Lee, and Rajan, 2013). Hence, cloud information security is one of the key factors that should be addressed to reduce and solve most of the cloud incidents worldwide, while promoting global cloud adoption securely.

Information security is identified as the key factor as well as the main concern towards the adoption of cloud in many corporations. It is composed of three main components, namely confidentially, integrity and availability (ISACA, 2015). Confidentiality is defined as preserving authorized restrictions on access and disclosure, including means for protecting privacy and proprietary information; integrity is defined as the guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity; availability is defined as ensuring timely and reliable access to and use of information (ISACA, 2015).

With respect to the cloud data security concern, three cloud storage protocols were introduced:

i. Proof of Ownership (PoW),
ii. Provable Data Possession (PDP), and
iii. Proof of Retrievability (PoR).

Figure 1.1 illustrates the difference between PoW, PDP and PoR in a general view. PoW is a protocol used by cloud storage to ensure the stored data is only made available and accessible by legitimate users or true data owner only. Examples of PoW schemes are (Halevi, Harnik, Pinkas, and Shulman-Peleg, 2011), (Yu, Chen, and Chao, 2015), (Hur, Koo, Shin, and Kang, 2016), and (Lorena and Agustin, 2015). On the other hand, PDP schemes such as (Ateniese, Burns, and Herring, 2007), (Mukundan, Madria, and Linderman, 2014), (Lin, Shen, Chen, and Sheldon, 2017), and (Wang, Wu, Qin, Tang, and Susilo, 2017) were introduced to allow the storage server to proof to its data client that the stored data is correctly stored by the servers. As PDP does not provide recovery to erroneous data, PoR was introduced by (Juels and Kaliski Jr., 2007) to address this issue. Hence, in term of data retrievability, PoR is better than PDP. Nonetheless, PoR might have a higher computation time in data encoding and decoding for error recovery mechanism. The early work of PoR (Juels and Kaliski Jr., 2007) have a limitation in term of number of times challenge-proof (data audit) are allowed to be conducted. To overcome this limitation, an improved PoR scheme was introduced by (Shacham and Waters, 2008) that apply erasure coding (EC) to allow recovery in case of data corruption and no constraint on the number of times challenge-proof (data audit) allowed to be conducted. However, it provides unbounded number of PoR challenges to ensure data integrity.

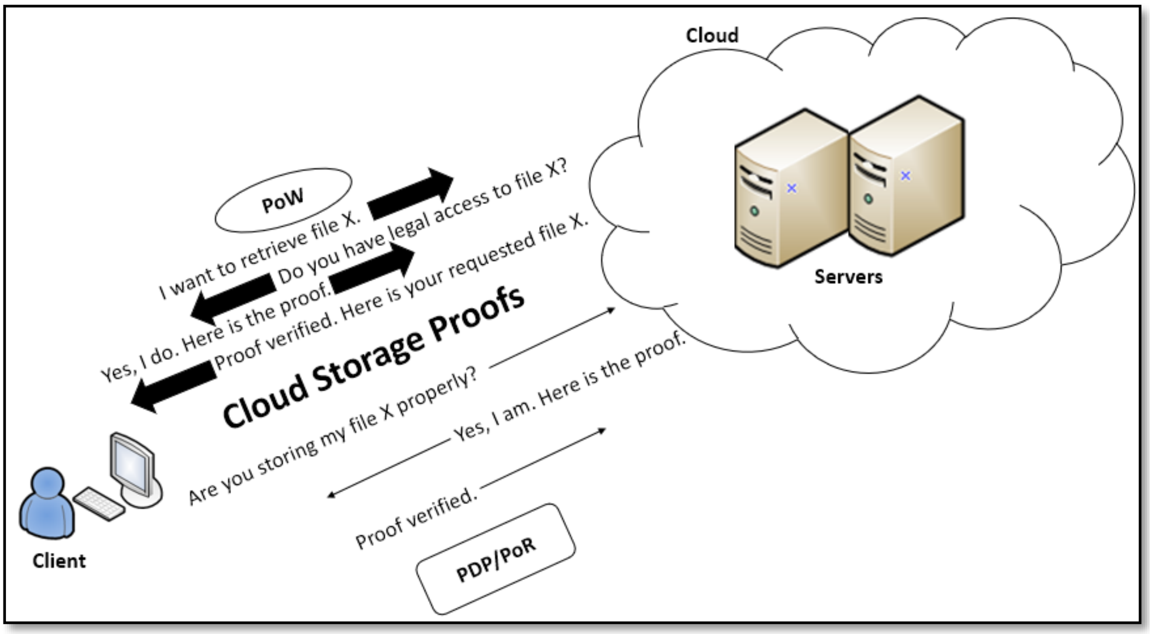
Cloud
Servers
PoW
I want to retrieve file X.
Do you have legal access to file X?
Yes, I do. Here is the proof.
Proof verified. Here is your requested file X.
Cloud Storage Proofs
Are you storing my file X properly?
Yes, I am. Here is the proof.
Proof verified.
PDP/PoR
Client

Figure 1.1: Difference between PoW versus PDP and PoR

Conventional data storage that involves replications requires abundant of storage spaces for data storage whereas cloud storage protocols like PDP and PoR schemes consumed less. However, PDP and PoR schemes requires data preprocessing, usually known as data encoding, before data is stored in distributed servers on cloud. Correspondingly, the conventional data storage method such as replication need only a minimal or even does not need any data preprocessing before the storing process take place. Eventually, PDP and PoR will perform slower than conventional data storage method like replication. Hence, performance in term of computation time of cloud storage protocols like PDP and PoR schemes is always the main concern in the construction of the scheme itself. It leaves a research gap on cloud storage schemes for achieving more efficient and secure PDP and PoR schemes.

1.2 Problem Statements

Recently, a variation of POR known as Slepian-Wolf based Proof of Retrievability (SW-PoR), designed for cloud storage to ensure the distributed cloud data is intact (Thao, Kho, and Lim, 2014). Intact of data means the stored data is in a condition where it is not damaged or not impaired in any way to ensure data completeness. There are three functions in SW-PoR scheme namely Encode, Retrieve and Repair. Details of these three functions will be described in Chapter 2.

SW-PoR scheme has been shown to be able to fully recover data even if data erroneous or lost occurs with its Repair function (Thao *et al.*, 2014) . Furthermore, SW-PoR also proven to be secure and efficient compared to linear network coding scheme for encode, retrieve and repair functions when simulated on different sizes of data up to 150 kilobits (kb) (Thao *et al.*, 2014). With respect to the real-world cloud system, the size of the data used in the reported simulation (Thao *et al.*, 2014) which is only 150 kb, is very difficult to represent the real cloud environment as the size data might be as small as few hundred kb of word document, up to few hundred megabits (Mb) of video file. The SW-PoR computation time is still an issue that need to be addressed. The time consumed by SW-PoR increases linearly when the file size increases for retrieve function, but the encode function rather experienced an exponential increment. When larger data is involved, the encoding time of SW-PoR scheme would be long, causing upload-to-storage time increases greatly. As a result, data integrity might be lost due to the un-trust of client to CSP, as trustworthiness of data is an important component in data integrity (Bertino and Lim, 2010).

1.3 Research Motivation

Storing a file in cloud does not end at uploading the file successfully, but it must also consider security measures such as encryption, error correction and recovery measures. Hence, the process to store data in cloud requires synchronization time for data preprocessing before actual storage take place (Google, 2017a). Although it is common for cloud servers to consume some time for data preprocessing, but the length of the time used is always an issue. In most cases, a more secure and complex process of storing a file in cloud servers requires a longer synchronizing time (Google, 2017b).

Ironically, the longer the time for a file to stay in transit for preprocessing, the longer the time the file to be stored exposed to security threats like eavesdropping and other malicious activities. An excellent cloud storage in term of security and trustworthy requires efficient algorithms for the computation, particularly when the growth rate of data size is exponential. Hence, the motivation of the work presented in this thesis is to have a viable approach that allow shorter computation time of PoR scheme, specifically SW-PoR scheme, for cloud storage.

1.4 Research Questions

To achieve the motivation stated in the foregoing section, two research questions were identified:

1. What improvement can be conducted in order to reduce the computation time of SW-PoR without compromising its performance?
2. Will the proposed improved SW-PoR compute efficiently on larger size data that uses an accepted amount of time to complete?

1.5 Research Objectives

The objectives of this research, derived from the identified research questions, are listed as follows:

1. To reduce the computation time of SW-PoR encoding using two proposed schemes, namely Partial Binary Encoding for SW-PoR (PBE-SW-PoR) and Optimized SW-PoR (Opti-SW-PoR).

2. To apply the two proposed schemes in (1) on larger data in simulation similar to actual cloud environment.
3. To compare and analyse the performance of the proposed scheme(s) in term of computation time, storage cost, overhead, resiliency and redundancy against the original SW-POR.

1.6 Research Scope

This research is an extension work to SW-PoR scheme (Thao *et al.*, 2014) to improve its computational performance. Performance evaluation for the proposed schemes is based on simulation. Data sizes simulated are ranged from 200 file blocks to 1000 file blocks (1Mb), where each file block composed of 1024 bits. Larger data sizes are not simulated as original SW-PoR encoding would require extremely long time to complete, which is explained in Chapter 3. Simulation is conducted using personal computer for the mentioned data sizes to imitate real cloud environment. Machine specification used for simulation are as follow: Intel i5-3210M processor, 2.50 GHz, 8GB of RAM, Windows 10 (64-bit) OS.

1.7 Research Methodology

In order to achieve the research objectives, a methodology consists of four main phases is designed. Figure 1.2 illustrates the methodology employed in this research.