International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **2**, February **2011**

 ISSN 1349-4198
pp. 915–925

AN EFFICIENT SNAPSHOT INDEXING METHOD FOR BLOCK-LEVEL BACKUP DATA IN REPLICATION SYSTEM

Guangjun Wu[1,2], Binxing Fang[1,2], Xiangzhan Yu[1], Xiaochun Yun[2] and Shupeng Wang[2]

[1]Research Center of Computer Network and Information Security Technology
Harbin Institute of Technology
Harbin, 150001, P. R. China
wuguangjun@gmail.com

[2]Institute of Computing Technology
Chinese Academy of Sciences
Beijing, 100029, P. R. China

Received October 2009; revised March 2010

Abstract. *Snapshot query can provide a clean copy for restoration in replication system. Current backup techniques produce large volume of backup data. Up till now, there was no efficient indexing method to retrieve a snapshot from updating log in block-level replication system. In this paper, we present a novel indexing method that is capable of querying any version of historical snapshot without introducing additional backup operations. A detailed mathematical model is used to analyze the complexity of our implementation. Under the guide of mathematical analysis, we present Hierarchical Clustering Snapshot Indexing Method (HCSIM). HCSIM includes snapshot query algorithm, version deleting rules and concurrency control policy. Extensive experiments have been carried out to show that the new indexing method can support snapshot query at low cost, both in theory and practice. Performance evaluation of the novel indexing method indicates that the implementation is optimal in current snapshot indexing methods.*
Keywords: Snapshot, Indexing, Version management, Block-level, Storage

1. **Introduction.** Data corruption is an important issue for organizations in face of all kinds of failures and disasters, such as man-made errors, virus attacks, firmware malfunctions and even site failures [1]. Many data protection techniques have been proposed, such as multi-version file system [2,3], snapshot-based backup techniques [4,5], and even Continuous Data Protection (CDP) [6-8]. The common replication methods are Copy-On-Write (COW), Redirect-On-Write (ROW) and writing stream duplications. These techniques can reduce the lost data, compared with traditional backup methods [9], and decrease the value of RPO (Recovery Point Objective). Some of these techniques can even provide 24×7 High Availability (HA) service for critical business. But all of them will produce large volume size of backup data in the long lived context. If there was no efficient indexing method to support data resiliency, the backup data would be unrecoverable and meaningless for HA service. Data resiliency needs to restore backup data correctly and efficiently. A fundamental requirement is to retrieve a latest clean copy from the backup data for restoration. Block-level replication techniques are widely concerned in recent publications [6-10]. They use updating log to store the changed blocks continuously or periodically. The recovery process is first to find the latest available full snapshot and apply the changes in the updating log in redo or undo sequence. In the study of [11], snapshot was even introduced into TRAP CDP log for robustness, and [12] further combined event knowledge database with each clean copy record and restored the corrupted