私有云下基于相似度聚类的重删算法研究

马柄腾，刘　丹

（电子科技大学 电子科学技术研究院，四川 成都 610054）

摘　要：随着虚拟化技术的日趋成熟与发展，越来越多的企业采用私有云平台来替代传统PC办公.然而，虚拟磁盘数据的高度重复与超大容量导致了存储空间与系统能耗的巨大浪费.为解决这一问题，首先提出了基于特征标签粗归类算法，将重删操作分散到云平台计算结点，有效地避免了传统重删算法的性能瓶颈；然后，提出了基于指纹ID相似度聚类算法，提高了各计算结点上的重删率；最后，通过实验对两种子算法进行分析，并验证了其有效性.

关键词：私有云；主存储；oVirt；重复数据删除；相似度；聚类分析

A Deduplication Algorithm Based on Similarity Clustering in Private Cloud

MA Bing-teng，LIU Dan

（Research Institute of Electronic Science and Technology， University of Electronic Science and Technology of China， Chengdu 610054, China）

Abstract：With the maturity and development of virtualization technology， more and more enterprises adopt private cloud platform to replace the traditional PC of office. The highly duplicated and overlarge virtual disk data led to a huge waste of storage space and system power consumption. To solve this problem， first， we propose a classification algorithm based on feature tag. It re-distributes to deduplication the computing nodes of cloud platform， and effectively avoid the performance bottleneck of traditional deduplication algorithms. Then， it is proposed a similarity clustering algorithm based on fingerprint ID to improve the deduplication rate. Finally， through experiments， the two sub-algorithms is analyzed， and its effectiveness is verified.

Key words：private cloud；domain storage；oVirt； deduplication；similarity；cluster analysis

作者简介：

马柄腾　男，（1990-），硕士.研究方向为分布式存储、云计算、网络安全.E-mail：mabingteng@sina.com.

刘　丹　男，(1969-)，博士，副教授.研究方向为网络安全、分布式并行计算、云计算与大数据处理.