ESTUDO DO IMPACTO DA DEDUPLICAÇÃO DE DADOS SOBRE O DESEMPENHO DE SISTEMAS DE ARQUIVOS IMPLEMENTADOS SOBRE RECURSOS EXPLORADOS DE FORMA OPORTUNISTA

Resumo

A duplicação de dados é um problema comum em sistemas de armazenamento. Este problema causa um aumento nos custos de armazenamento, impactando inclusive aplicações de e-Ciência, visto que muitas dessas aplicações manipulam grandes massas de dados. A deduplicação de dados é uma solução muito utilizada para este problema. No entanto, pouca atenção foi dada à implementação de deduplicação em ambientes oportunistas, que também trazem vantagens econômicas em relação ao melhor aproveitamento dos recursos disponíveis. Visto que as características de uma estratégia podem reduzir os ganhos conseguidos com a outra, é necessário entender como uma estratégia impacta a outra. O objetivo deste trabalho é estudar o impacto da deduplicação de dados sobre o desempenho de um sistema de arquivos distribuído implementado sobre recursos explorados de forma oportunista, em especial sobre o tempo de acesso aos arquivos. Para tal, um modelo que descreve a deduplicação no sistema de arquivos foi desenvolvido e duas análises envolvendo instâncias deste modelo foram realizadas. Tais análises revelaram que popularidade dos arquivos e o padrão de similaridades no sistema de arquivos são características que devem ser levadas em conta ao deduplicar.

Palavras-chave:Deduplicação de dados, Sistemas de arquivos, Ambientes oportunistas.

STUDY OF THE IMPACT OF DATA DEDUPLICATION ON THE PERFORMANCE OF FILE SYSTEMS IMPLEMENTED OVER RESOURCES EXPLOITED IN OPPORTUNISTIC WAYS

Abstract

Data duplication is a common problem in storage systems. This problem causes the increase of storage costs, impacting even e-Science applications, since many of these applications deal with large data sets. Data deduplication is a common solution to resolve this problem. However, little attention has been given to the implementation of deduplication in oportunistic environments, which bring economic advantages concerning the best utilization of the available resources. As the characteristics of one strategy may reduce the gains of the other, it is necessary to understand how each of the strategies impacts the other. The goal of this paper is to study the impact of the data deduplication on the performance of a distributed file system implemented over resources exploited in opportunistic ways, focusing on the file access time. To this end, a model which describes the deduplication in file systems was developed and two analyzes using instances of this model were performed. Such analyses revealed that file popularity and files similary pattern in the file system are characteristics which must be considered when deduplicating.

Keywords:Data Deduplication, File Systems, Oportunistic Environments.