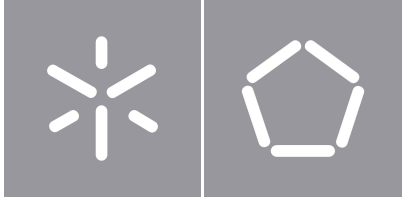
University of Minho
School of Engineering

Diogo Alexandre Correia Marques

Realistic Benchmarking of Data Deduplication and Compression Systems

september 2025



University of Minho
School of Engineering

Diogo Alexandre Correia Marques

Realistic Benchmarking of Data Deduplication and Compression Systems

Master's Dissertation in Informatics Engineering

Dissertation supervised by
João Tiago Medeiros Paulo

september 2025

Abstract

Write abstract here (in English)

Keywords keywords, here, comma, separated

Resumo

A deduplicação de dados corresponde a uma técnica para identificar e remover conteúdos duplicados em sistemas de armazenamento, contribuindo assim para uma melhor utilização do espaço disponível e consequente redução de custos. De facto, os sistemas modernos combniam ainda técnicas de compressão para extrair maior densidade dos dados e armazenar somente o extritamente necessário.

Ao suportarem várias técnicas de manipulação de dados, a avaliação destes sistemas torna-se cada vez mais complexa, dado que as workloads necessitam de responder a uma série de critérios que validem a deduplicação e compressão em simultâneo, sem esquecer que as caracteristicas comuns entres sistemas devem continuar no alvo da avaliação, em particular a localidade espacial e temporal dos acessos.

No entanto, os benchmarks disponíveis à comunidade (fio, vdbench) apenas permitem uma manipulação parcial dos níveis de entropia e deduplicação. Sem mencionar que a simulação de traces é demasiado simplista e torna-se impraticável em sistemas modernos por estes serem demasiodo rápidos a concluir o trace e não existir uma forma trivial de extender o mesmo e preservar as suas caracteristicas.

Além disso, no sentido de extrarir o máximo de performance, alguns sistemas disponibilizam unicamente interfaces de I/O baixo nível, tal como SPDK, o torna ainda mais complicada a execução de workloads, pois os benchmarks anteriormente referidos não suportam diretamente tais protocolos para comunicação com o disco.

Posto isto, esta dissertação tem por objetivo desenvolver um benchmark para sistemas de armazenamento, sendo este capaz de suportar diversas interfaces de I/O, bem como a geração de workloads realistas que permitam a recolha de metricas relevantes para a avaliação do sistema, permitindo assim a identifcação gargalos de desempenho e impactos associados às caracteristica do sistema de armazenamento.

Palavras-chave sistema de armazenamento, interface de I/O, deduplicação, compressão, workload realista

Contents

List of Figures

List of Tables

Acronyms

FIO Flexible I/O Tester.

SPDK Storage Performance Development Kit.

vbdev virtual block device.

Chapter 1

Introduction

1.1 Problem Statment and Objectives

1.2 Contributions

1.3 Document Structure