Parallel Data Mining

Jason Ma & Kanishk Tantia
Introduction to High Performance and Parallel Computing
Harvey Mudd College
jyma@g.hmc.edu & ktantia@g.hmc.edu

Abstract

1 Introduction

Data storage needs have grown exponentially as more data is increasingly digitized. IBM estimates that 2.5 Quintillion bytes of data are created daily. IDC research estimates that digital data growth will grow at a compound growth rate of 11.7% through 2020. As of 2011, the demand for more storage has outpaced the growth of digital storage means, and by 2020, it is estimated that the demand will outpace the growth of digital storage by over 15,000 Exabytes of data per year.

Data Deduplication is an efficient manner in which data usage can be reduced, and has slowly gained traction over the last decade. Assuming that a large amount of the data being stored is redundant or "junk" data, data storage needs can be met by simply never storing multiple copies of data and instead just providing the same set of bytes when the data is required.

2 Data Processing

3 Algorithms Tested

3.1 FastCDC

3.2 RabinCDC

3.3 AECDC

4 Experimental Approach

4.1 Practical Traces

4.2 Synthetic Traces

5 Results

6 Conclusions

7 References