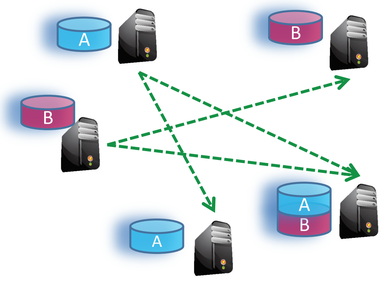
DiskReduce: Making Room for More Data on DISCs

Bin Fan, Wittawat Tantisiriroj, Lin Xiao, Garth Gibson

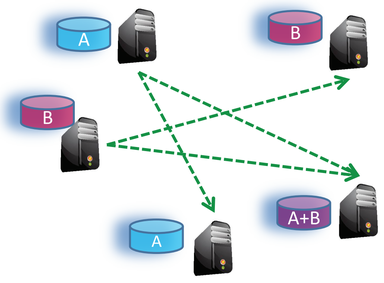
Overview

Google FS/ HDFS on Data Intensive Scalable Computers

- **Triplication can recover from 2 failures but it trades 200% extra storage for this redundancy**
- **Parity saves storage and tolerates the loss of any two nodes**



GFS



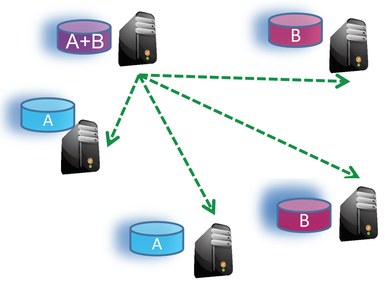
DiskReduce

Complexity

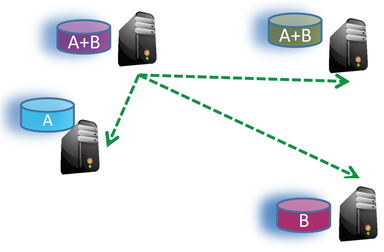
Yahoo! & Cloudera nervous of cluster RAID

- **In traditional RAID, encoding and reconstruction are inline with critical data path**
- **Based on HDFS approach, DiskReduce is "simple" in design:**
  - **Triplicate data blocks initially**
  - **Use asynchronous background process to encode and to reconstruct**

Encoding Choice



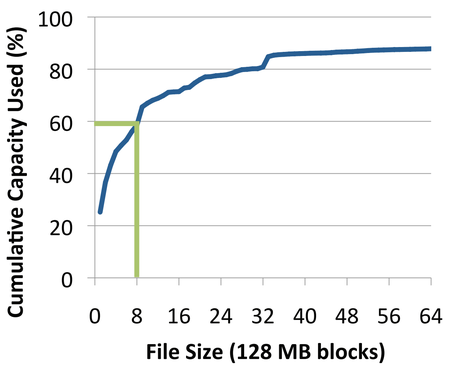
RAID5 + Mirror: two copies and one parity



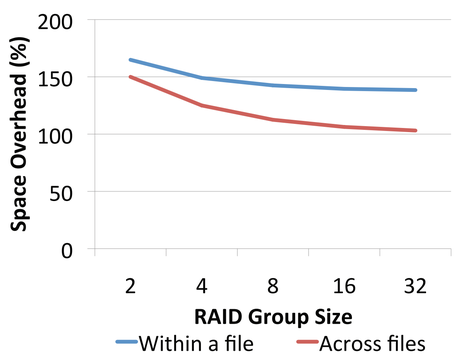
RAID6: double protection code

Encoding Efficient & Deletion

Cloud file size distribution (Yahoo! M45)



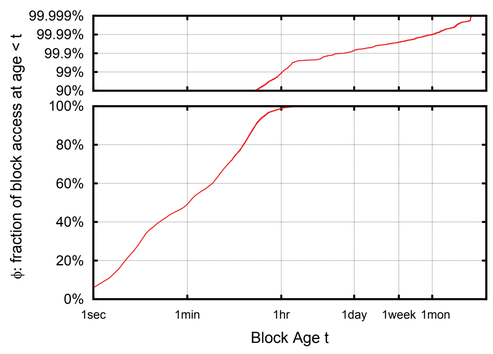
Space overhead after encoded



- **Large percentage of small files (e.g. ~60% of files in M45 HDFS clusters have 8 blocks or less)**
- **Parity only internal to file is not space effective**
- **Partial group deletion, e.g. one block, either frees no space or needs parity to be recomputed**

Asynchronous Encoding

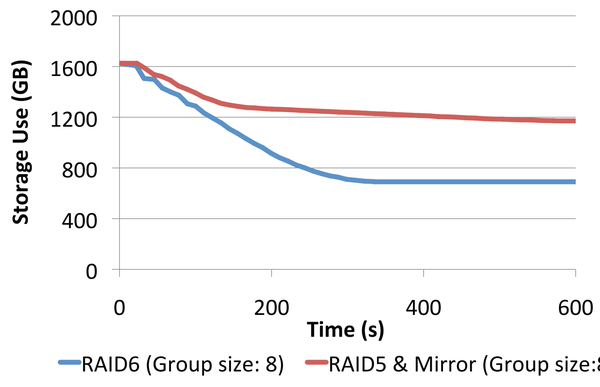
Cloud data access pattern (Yahoo! M45)



- **99% of data accesses happen within the first hour of a data block's lifetime**
- **Delaying encoding for one hour benefits most data access with multiple copies**
- **With 2TB disks, one hour at 25MB/s per disk, a workload of continuous writing, would consume only 6% of each disks' capacity**

Prototype

- **32 nodes (two quad-core 2.83GHz Xeon, 16GB memory, four 7200 rpm SATA 1TB disk, 10 Gigabit Ethernet)**



- **Overhead is reduced from 200% to 113% by "RAID 5 and mirror" and to 25% by "RAID 6" schemes**

Status and Plan

Tech Transfer

- **Based on a talk about DiskReduce v1, Dhruba Borthakur of Facebook has implemented a variant of RAID5 + Mirror in HDFS**

Prototype Status

- **We have a prototype of RAID5 + Mirror and RAID6 in HDFS/Hadoop v0.20.0**

Plan

- **Online reconstruction to provide degraded mode read**
- **Support and optimize deletion cleanup**
- **Analyze additional traces**

Carnegie Mellon

pdsi

Parallel Data Laboratory

R09