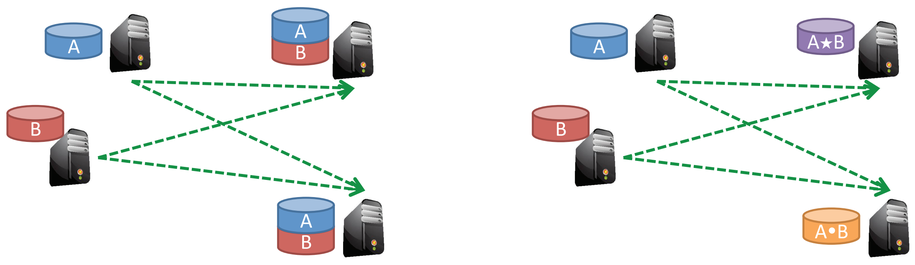
DISKREDUCE: RAIDING THE CLOUD - GROUPING CHOICES

Bin Fan, Wittawat Tantisiriroj, Lin Xiao, Garth Gibson (CMU)

OVERVIEW

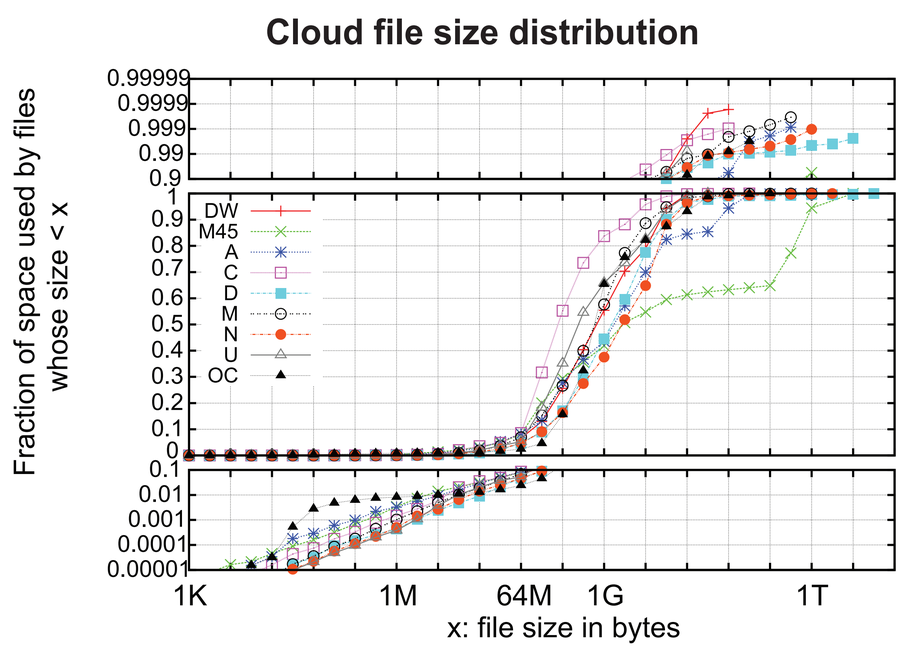
Google FS/ HDFS on Data Intensive Scalable Computers

- **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**

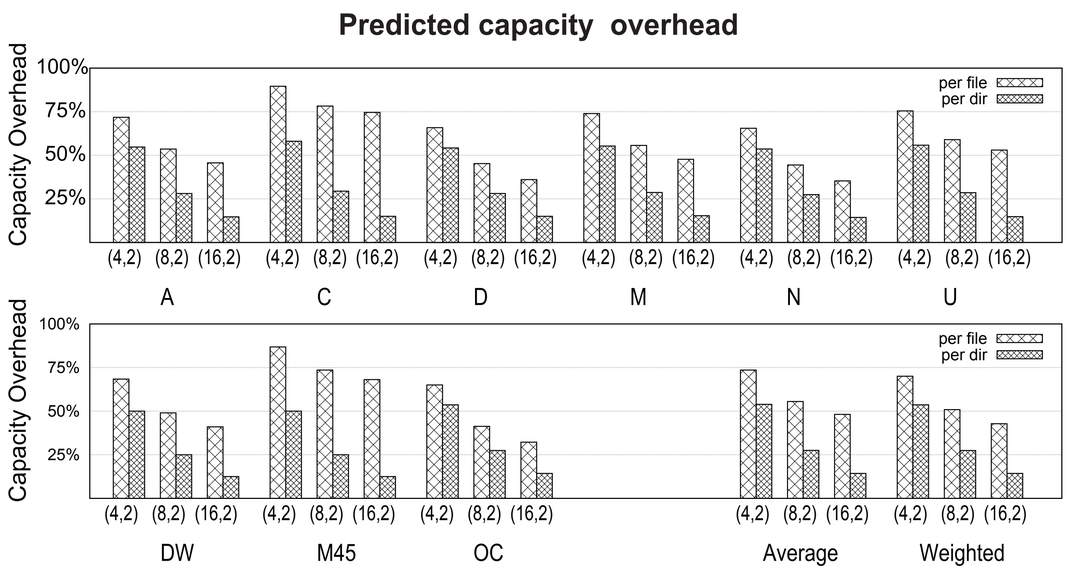


GFS / HDFS DiskReduce

CAPACITY OVERHEAD



- **Across 9 file systems (1.5 - 21 PB), 30 - 80% of the storage used by files smaller than 1 GB (size of 16 blocks, 64 MB each)**
- **Since each block is large (64 MB by default), there will be few blocks in per-file RAID sets**



RAID set (N,M): N data blocks+ M check blocks

Weighted scales average by size of cluster

- **With 8 data blocks in a RAID set, per-file RAID 6 requires about 56% capacity overhead while per-dir RAID 6 requires only 28% overhead (25% is minimal)**

RAID SET DEFINITIONS

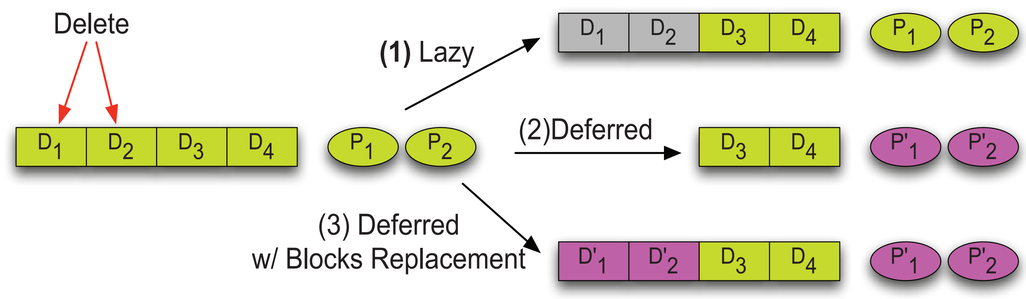
RAID Per-file: blocks in a RAID set are from the same file

+ **Simple**
- **Too much overhead**

RAID Across-files: blocks in a RAID set can be from different files

+ **Per-directory RAID 6 can achieve lower overhead**
- **Small write problem - potential read-modify-write to update parity blocks on file deletion**

HANDLING DELETED BLOCKS

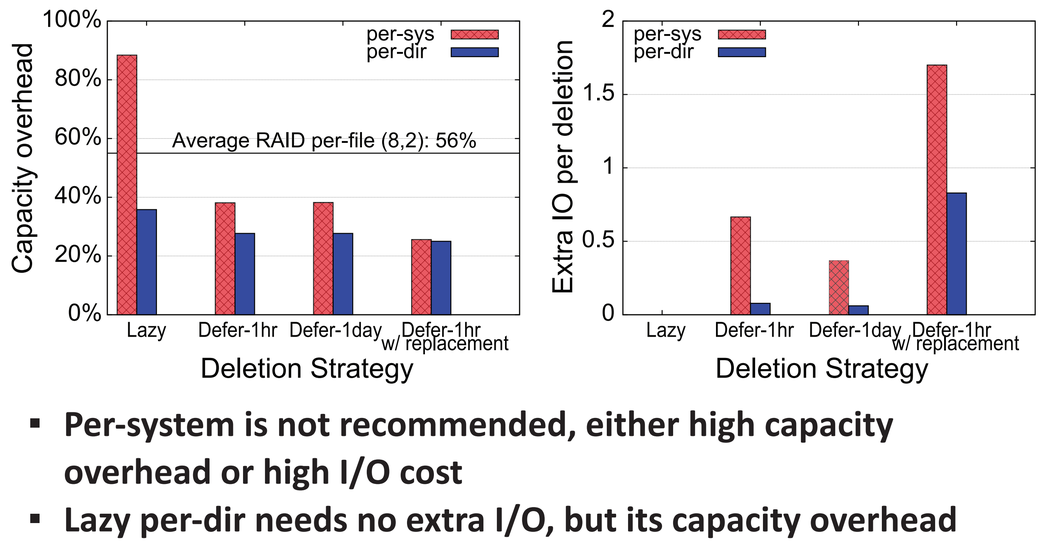


3 ways to handle deletion:

1. **Lazy Deletion**
   - **Never recalculate RAID equation so only recover space when all blocks in a RAID set are deleted**
2. **Deferred Deletion**
   - **After a timeout to allow all blocks to be deleted naturally, recalculate RAID over smaller data set**
3. **Deferred Deletion with Block Replacement**
   - **When recalculating check blocks, add new data blocks into existing RAID set to prevent RAID sets from becoming shorter**

TRACE ANALYSIS

Yahoo! M45 trace over 2000 hours (80 days)



- **Per-system is not recommended, either high capacity overhead or high I/O cost**
- **Lazy per-dir needs no extra I/O, but its capacity overhead (36%) is significantly larger than minimal (25%)**
- **Defer 1 day per-dir reduces capacity overhead to near minimal (28%), paying only 0.06 (64MB) I/O per block created and deleted (which at least 3 I/Os)**

Intel Science & Technology Center for Cloud Computing

YAHOO! facebook Carnegie Mellon University Georgia Tech intel PRINCETON UNIVERSITY UC Berkeley