Ming Hsieh Department of Electrical Engineering

# Novel Codes for Cloud Storage

Maheswaran Sathiamoorthy*, Megasthenis Asteris*,
Dimitris Papailiopoulos*, Alexandros G. Dimakis*,
Ramkumar Vadali**, Scott Chen**, Dhruba Borthakur**
* - USC, ** - Facebook

USC Viterbi School of Engineering

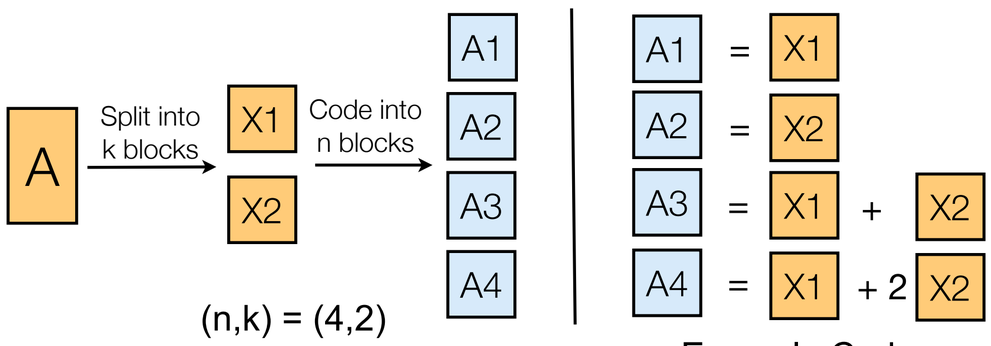
## Cloud Storage

- Thousands of machines in data centers
- Failure is the norm rather than the exception
- To provide reliability, replication is used.
  - each file is stored multiple times (generally thrice).
  - replication costly, so could coding be an alternative?
- Hadoop: Distributed storage system, widely used.

## Erasure Coding

(n,k) coding:

- Split file in k blocks and then code into n blocks
- Any k of the n blocks enough to recover the file
- *e.g.*: Reed Solomon codes




### The Good

- Lower storage
- Higher reliability

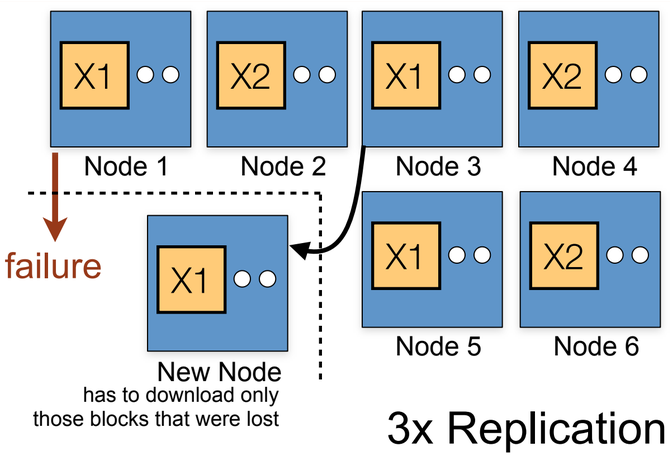
### The Bad

- Difficult to implement
- HDFS-RAID: solved!

### The Ugly

- Repair Problem!

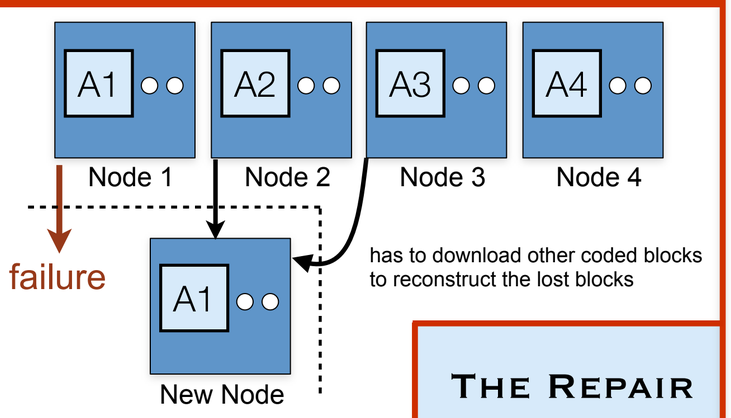
## Replication




3x Replication

Can tolerate 2 failures
Storage overhead: 200%
*Repair* overhead: 0%

## Coded Storage




(4,2) Coding

Can tolerate 2 failures
Storage overhead: 100%
*Repair* overhead: 100%

### The Repair Problem:

(n,k) code requires k blocks to be read for repairing one block

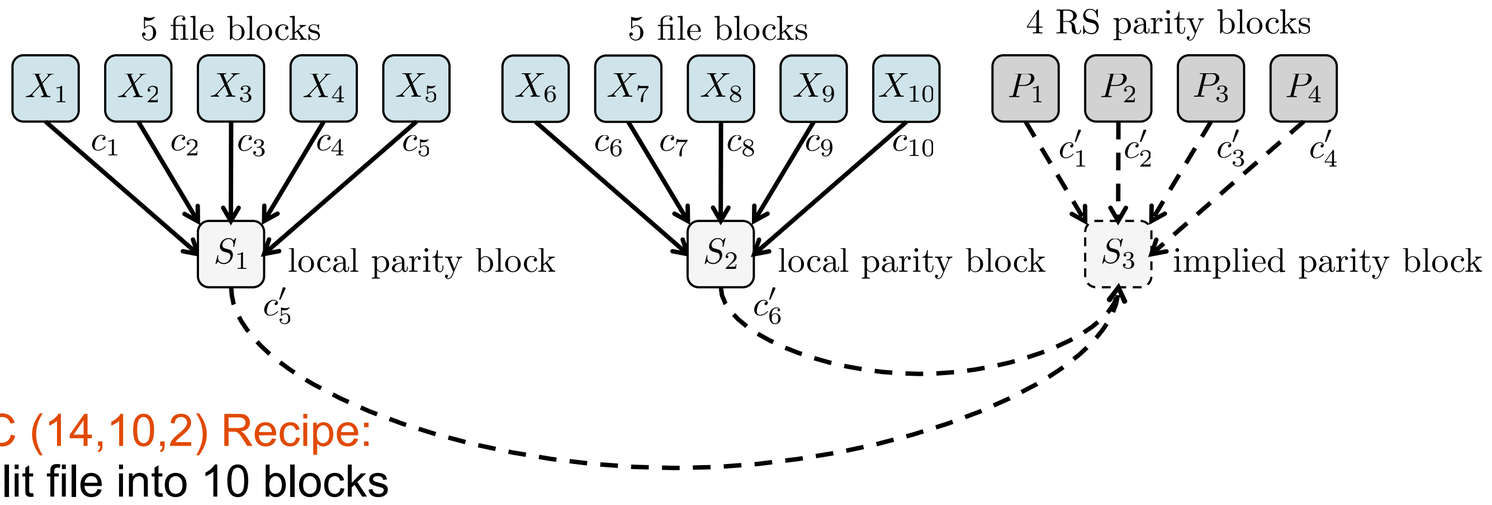
## Facebook clusters

- Facebook has one of the largest Hadoop clusters
  - >3000 machines
  - >30 PB logical data
- 3x replication costly
- Facebook uses HDFS-RAID with (14, 10) Reed Solomon Code which we will call HDFS-RS.
- Every lost block needs 10 more blocks for repair!
- Network is bottlenecked
- So only 8% of cold data is encoded

## Xorbas

- LRC Codes to the rescue!
- Facebook's HDFS-RAID is open source.
- We used it to develop our own version of Hadoop called HDFS-Xorbas*, which implements LRC codes
- We use a (14, 10, 2) LRC Code
- 2/14 = 14% extra storage
  - but mitigates the repair problem

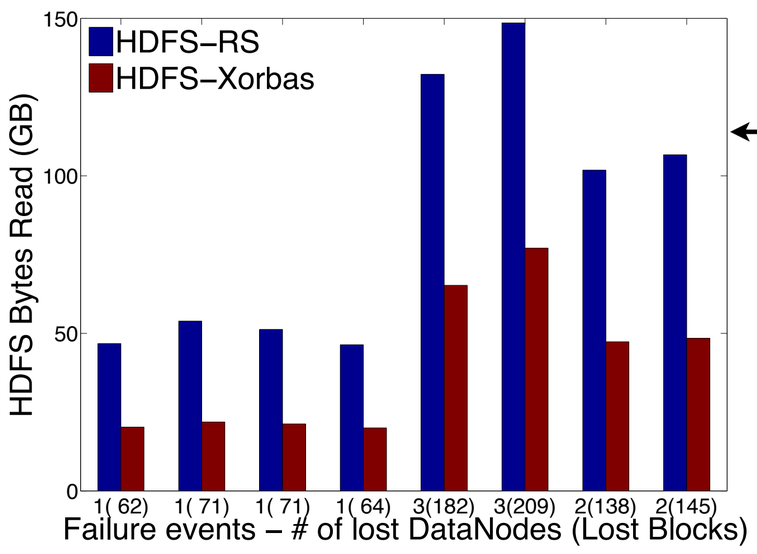
### Locally Repairable Codes (LRC)




LRC (14,10,2) Recipe:

- Split file into 10 blocks
- Then create 4 extra parity blocks using Reed Solomon encoding
- Use the first 5 and second 5 file blocks to create 2 "local parity blocks" S1 & S2
- One lost block requires only 5 others blocks to be read.

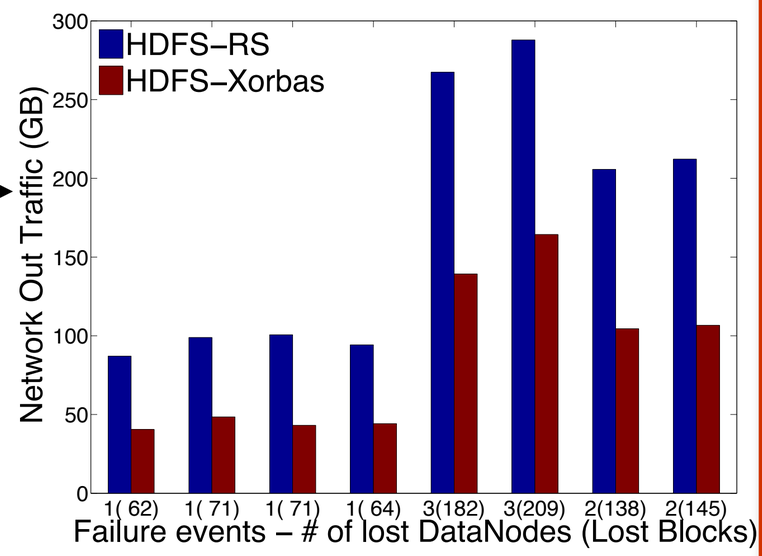
## Experiments and Results




**Experiments over 50 node clusters from Amazon S3**

Bytes read from the system for repair

Network used for repair




### The Verdict

- Use 14% more storage as compared to RS (14, 10) code
- But repair uses
  - half the network
  - and half the disk
- Also provides better reliability!

msathiam@usc.edu

*-code opensourced at http://github.com/madiator/hadoop-20

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering