Introduction to Regenerating Codes

Benjamin Renard

McGill University

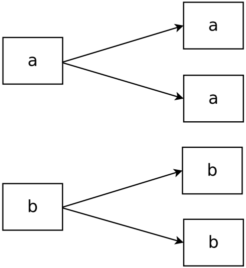
March 12, 2012

Basics of Distributed Storage

- Reliable access to data through redundancy.
- Data stored on individually unreliable nodes.
- Should be resistant to nodes failures.
- Each system comes along with a list of failures it can handle.
- Main elements:
    - Source: gives the file to the system.
    - Nodes: store fragments of the file.
    - Data Collector (DC): want to retrieve the file.

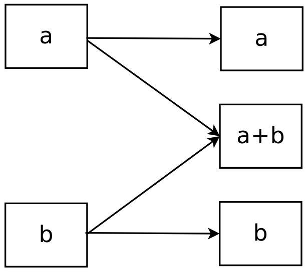
Replication versus erasure coding

- Replication to ensure redundancy:
  - file is divided into fragments
  - each fragment stored in a node
  - each node is replicated *r* times



Replication versus erasure coding

- Erasure coding:
  - we can store *linear combination* of fragments
  - use of Maximum Distance Separable (MDS) codes
  - optimal in term of redundancy-reliability trade-off



MDS codes

Definition

If Σ is a field and $C \in \Sigma^n$ is a subspace of Σ^n, then C is said to be a *linear code*. The elements of C are called codewords.
If $c_1, \ldots, c_k$ is a basis of C, then k is the *dimension* of C.

C is denoted as an $[n, k]$ code.

Definition

The *minimum distance* of a code C, $\Delta(C)$, is the minimum Hamming distance between two distinct codewords of C.

MDS codes

Theorem (Singleton Bound)

If C is an $[n, k]$ linear code then:

$$\Delta(C) \leq n - k + 1 \tag{1}$$

Definition

A linear code that meets the singleton bound is called a MDS (Maximum Distance Separable) code.

Erasure coding

- File of size $\mathcal{M}$ is divided in k fragments of size $\mathcal{M}/k$.
- Use a $[n, k]$ MDS code to encode k fragments into n fragments (of same size)

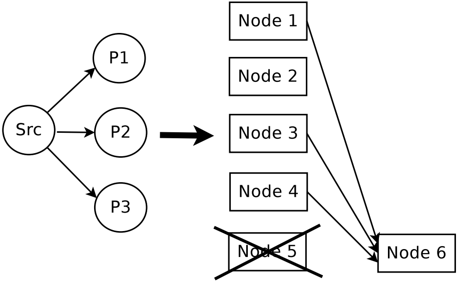
Theorem (MDS property)

The original file can be recovered from any set of k coded fragments.

Repair Bandwidth

- Repair process: when a node fails, we need to create a new one.
- Repair bandwidth: total bandwidth needed for the procedure.
- Replication: copy an existing replicate.
  - Repair bandwidth : $\mathcal{M}/k$
- Erasure coding: *naive* method
  - Recreate the whole file and then create a new fragment.
  - Repair bandwidth: $\mathcal{M}$.
  - *Exact* repair: we replace with same node.
  - *Functional* repair: only maintain the MDS property.

Regenerating Codes - Notations



- We consider a $[n, k]$ linear code over finite field $\mathbb{F}_q$.
- Each node stores α bits.
- A new node connects to d nodes, $k \leq d \leq n - 1$.
- A new node downloads β bits from each node.
- The repair bandwidth is $\gamma = d\beta$.

Storage-Bandwidth tradeoff

Theorem (Regenerating Codes)

The points $(n, k, d, \alpha, \gamma)$ *are feasible iif* $\alpha \geq \alpha^*(n, k, d, \gamma)$, [a] *where*

$$\alpha^*(n,k,d,\gamma) = \begin{cases} \frac{\mathcal{M}}{k} & \gamma \in [f(0), +\infty) \\ \frac{\mathcal{M}-g(i)\gamma}{k-i} & \gamma \in [f(i), f(i-1)) \end{cases} \quad (2)$$

$$f(i) \triangleq \frac{2\mathcal{M}d}{(2k-i-1)i + 2k(d-k+1)}$$

$$g(i) \triangleq \frac{(2d-2k+i-1)i}{2d}, \text{ with } i < k$$

Codes that achieve $\alpha = \alpha^*(n, k, d, \gamma)$ *are called* regenerating codes.

[a] A.G. Dimakis *et al*, *Network Coding for Distributed Storage System*, IEEE Trans. on Information Theory, vol.59, no.9, September 2010

Storage-Bandwidth tradeoff

Corollary

For d, n, k given, the minimum repair bandwidth γ is given by:

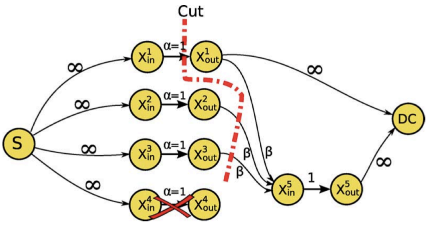
$$\gamma_{\text{min}} = f(k-1) = \frac{2\mathcal{M}d}{2kd - k^2 + k} \tag{3}$$

Important observation: $\gamma = d\beta$ decreasing function of d.

Storage-Bandwidth tradeoff

Proof (sketch).

Let us define the information flow graph:



The min-cut needs to be at least the object size $\mathcal{M}$. Use results from network coding.

Important: need a very large field size (which depends on the graph size).

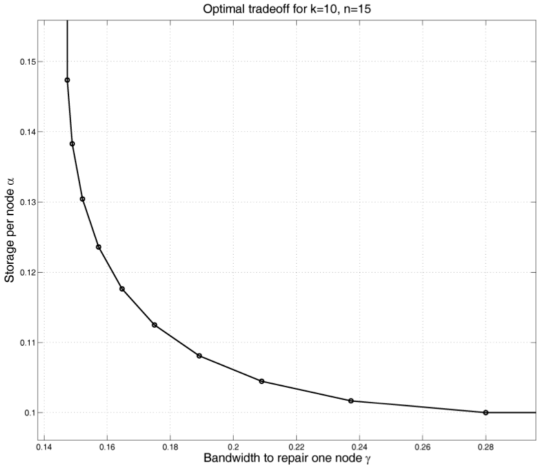


Figure: Optimal tradeoff curve between α and γ for $k = 10, n = 15$ with $\mathcal{M} = 1, d = n - 1$. Traditional erasure coding is $(\gamma = 1, \alpha = 0.1)$. (From Dimakis *et al*, 2010)

Minimum-Storage Regenerating (MSR) codes

- MSR codes are obtained by minimizing α:

$$(\alpha_{MSR}, \gamma_{MSR}) = \left(\frac{\mathcal{M}}{k}, \frac{\mathcal{M}d}{k(d-k+1)} \right) \tag{4}$$

- If $d = k$, then $\gamma_{MSR} = \mathcal{M}$: cannot avoid naive method.
- If $d = n-1$, $\gamma_{MSR}^{\min} = \frac{\mathcal{M}}{k} \cdot \frac{n-1}{n-k}$
- Equivalent with MDS codes.

Minimum-Bandwidth Regenerating (MBR) codes

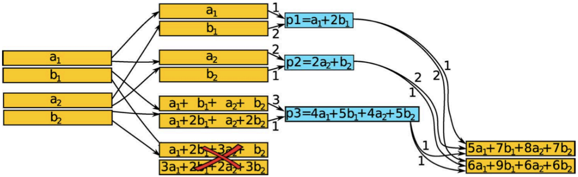
- MBR codes are obtained by minimizing γ:

$$(\alpha_{MBR}, \gamma_{MBR}) = \left(\frac{2\mathcal{M}d}{2kd - k^2 + k}, \frac{2\mathcal{M}d}{2kd - k^2 + k} \right) \tag{5}$$

- Note that $\alpha_{MBR} = \gamma_{MBR}$.
- If $d = n - 1$, $\alpha_{MBR}^{\min} = \gamma_{MBR}^{\min} = \frac{\mathcal{M}}{k} \cdot \frac{2n-2}{2n-k-1}$
- We have to allow a little more storage in order to decrease the repair bandwidth.

Functional repair - Example [1]

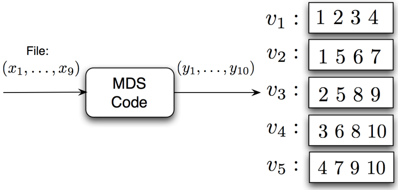
Example for a $[4,2]$ MSR code with $\mathcal{M} = 2$ Mb.



Repair bandwidth: 1.5 Mb. This code is also optimal as $\frac{\mathcal{M}}{k}\frac{n-1}{n-k} = 1.5$ Mb and $\frac{\mathcal{M}}{k} = 0.5$ Mb .

[1] A.G. Dimakis *et al*, *Network Coding for Distributed Storage System*, IEEE Trans. on Information Theory, vol.59, no.9, September 2010

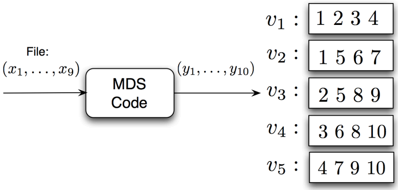
Exact Repair - Example [2]

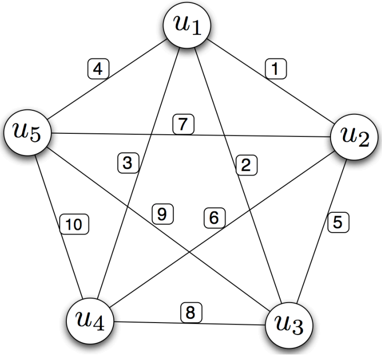


- DSS (Distributed Storage System) description: $[n, k, d] = [5, 3, 4]$
- MDS code: $(10, 9)$ parity-check code
- Nodes: $v_1, \ldots, v_5$.
- Achieves the storage-bandwidth tradeoff: MBR code.

[2] K.V. Rashmi *et al*, *Explicit Construction of Optimal Exact Regenerating Codes for Distributed Storage*, Allerton Conference, September 2009

Exact Repair - Example [2]





[2] K.V. Rashmi *et al*, *Explicit Construction of Optimal Exact Regenerating Codes for Distributed Storage*, Allerton Conference, September 2009

Exact Uncoded Repair [3]

- When read/write bandwidth of the nodes is the bottleneck of the system.
- Uncoded repair: survivor node reads only the data he sends.
- No network coding.

- Allow for a *repair table*:
  - available at all time
  - describe any recovery procedure: who should be contacted and what should they give
  - network complication

- General construction: concatenation of a MDS code and a *fractional repair* code.

[3] S.E. Rouayheb and K. Ramchandran, *Fractional Repetition Codes for Repair in Distributed Storage Systems*, Allerton Conference, September 2010

Some other results

- Rashmi *et al*, 2009: explicit construction for exact MBR codes for $d = n - 1$ and any k and MSR codes for $d = k + 1$ and any n.
- Rouayheb and Ramchandran, 2010: introduction of *uncoded* repair with some construction mechanisms.
- Rashmi *et al*, 2010: proof (with one possible exception) that exact regeneration can only be obtained for MBR.
- Rashmi *et al*, 2011: explicit constructions of MBR codes for all feasible values of $[n, k, d]$ and MSR codes for all $[n, k, d \geq 2k - 2]$.

Conclusion

- 'Regenerating codes' reduce the repair bandwidth while maintaining the optimality of redundancy-reliability.
- This reduction comes along with a augmentation of the storage.
- Construction techniques have been developed for some (but not all) possibilities.
- Many possible application: storage in wireless sensor networks, data centres...
- But yet no existing system as some issues still need to be handled: data integrity, security...

Thank you!

Questions ?