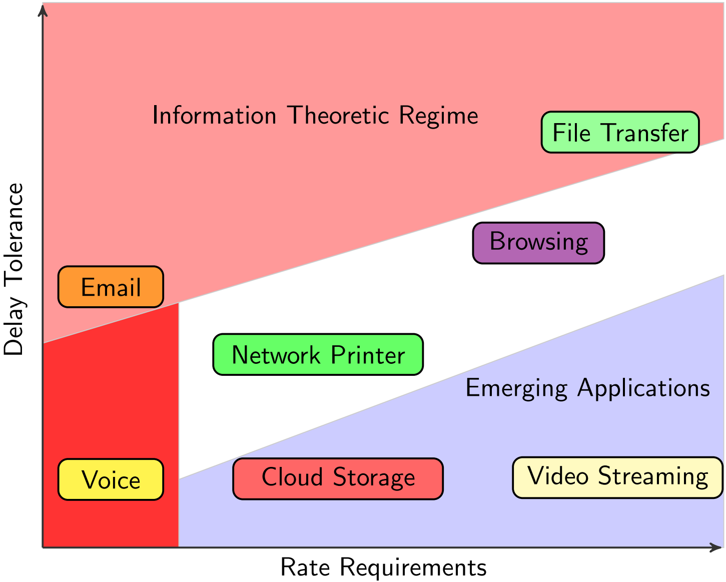
Low latency replication over memory constrained servers

Parimal Parag
Rooji Jinan
Ajay Badita
Pradeep Sarvepalli

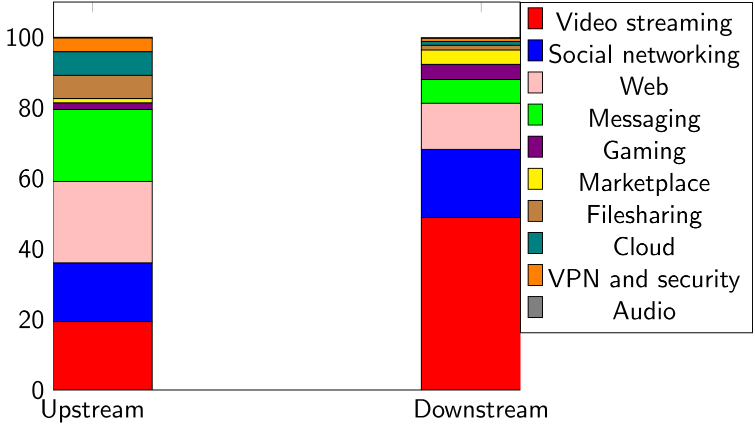
Sep 01, 2021



Evolving Digital Landscape



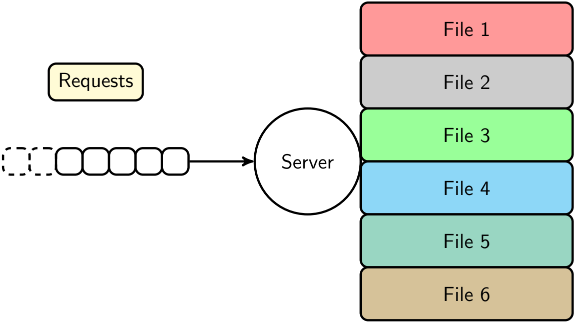
Global application traffic share 2021 [1]



---

[1] https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/2021/Phenomena/MIPR%20Q1%202021%2020

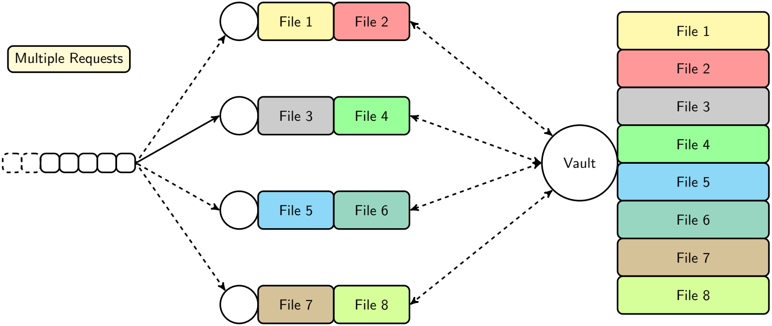
Centralized Paradigm



Potential Issues

- Not scalable with traffic load
- Susceptible to hardware failures and attacks

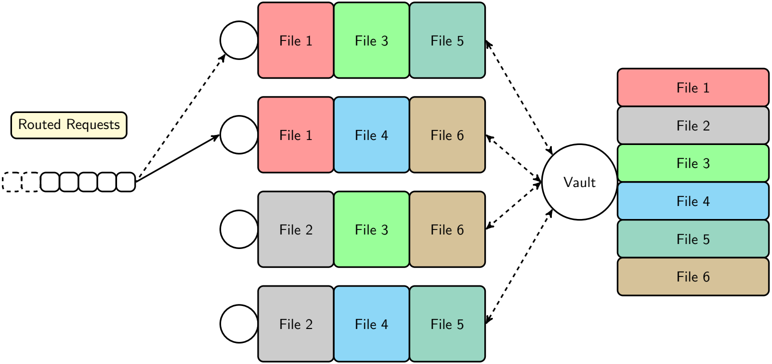
Distributed Paradigm



Potential Issues

- Susceptible to hardware failures and attacks

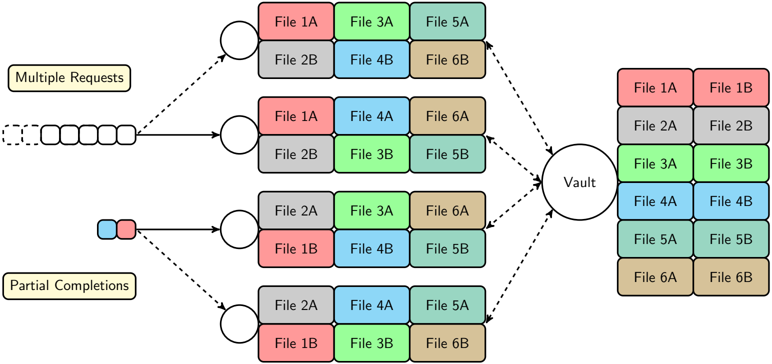
Resilience though redundancy



Latency redundancy tradeoff

- Download speedup due to parallel access
- Increased load due to redundant access

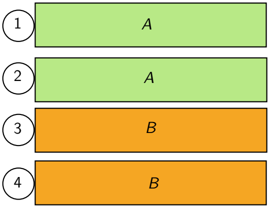
Load balancing through file fragmentation

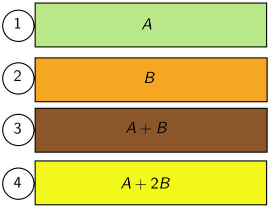


Shared coherent access

- Availability and better content distribution
- File segments on multiple servers

Coded Storage for single file

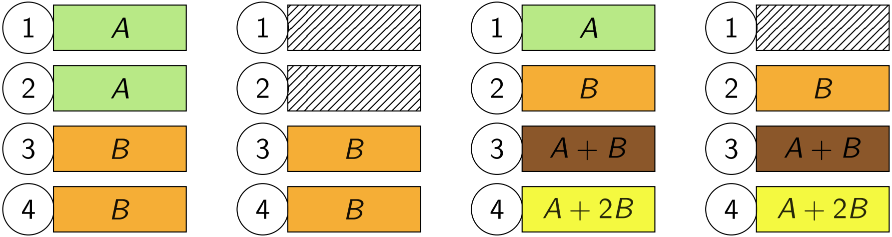




Single file divided into V fragments

- encoded into VR fragments
- each coded fragment stored over $B = VR$ servers
- reconstruction by set of V coded symbols

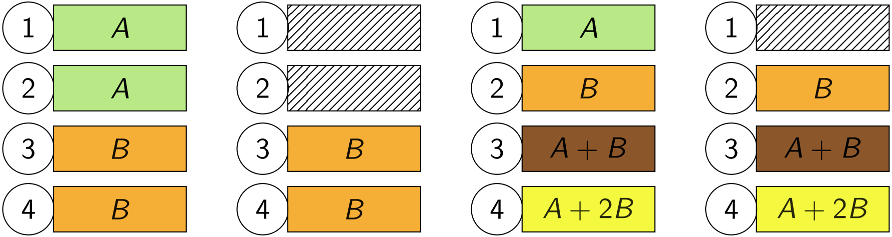
File download time



Mean file download time

- fragment downloads are *i.i.d.* and memoryless with unit rate
- parallel access from N_ℓ useful servers after ℓ downloads
- Harmonic sum of number of useful servers $\sum_{\ell=0}^{V-1} \frac{1}{N_\ell}$

File download time



Number of useful servers after ℓ downloads

- **replication:** $B - R\ell$
- **MDS coding:** $B - \ell$

Prior Work

MDS codes

Outperform replication codes in file access delay

- Huang et al(2012), Li et al(2016), Badita et al(2019)

Rateless codes

Offers near optimal performance

- Mallick et al(2019)

Staircase codes

Subfragmentation improves latency performance

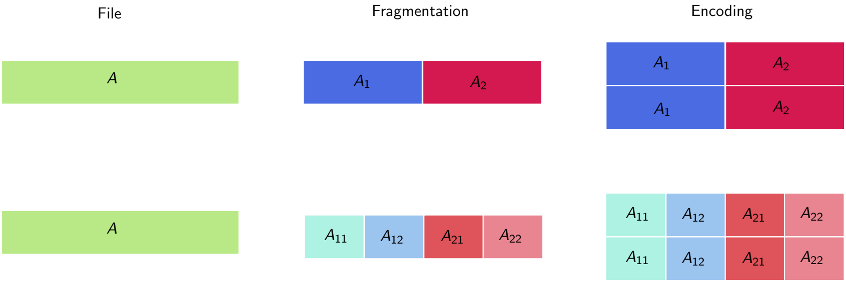
- Bitar et al(2020)

Our model

Replication codes for a file with equal sized fragmentation over multiple servers where each can store multiple file fragments

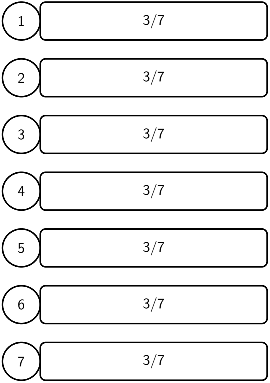
Storage model

fragmentation & encoding



- File divided into V fragments & encoded into VR fragments

Memory constrained system

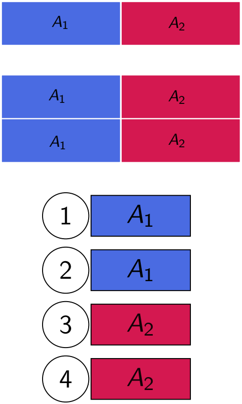


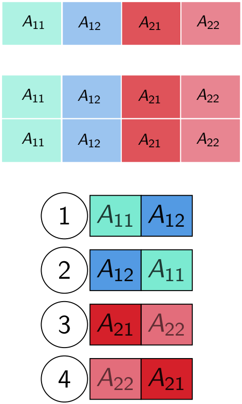
Storing αB size coded messages for a unit size message

- parallel access from all B servers
- α-fragment of message stored at each server

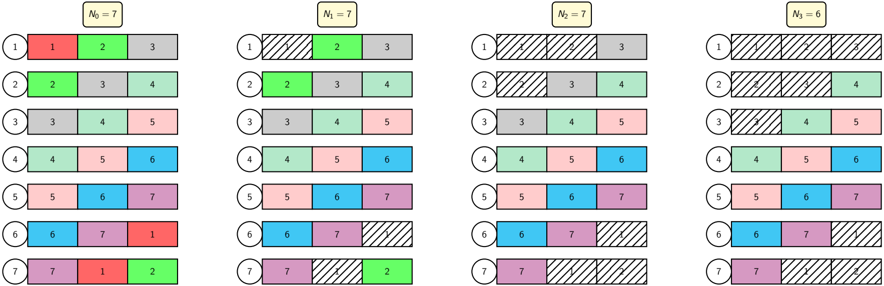
Storage model

Placement



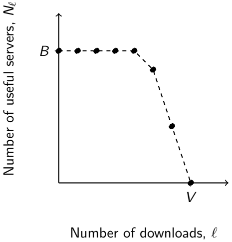


File download time



- Number of useful servers after ℓth download, N_ℓ
- Fragment download times are *i.i.d.* exponential with unit rate
- Rate of download at ℓth stage is N_ℓ
- The mean download time is $\mathbb{E}\sum_{\ell=0}^{V-1}\frac{1}{N_\ell}$

Optimality criterion



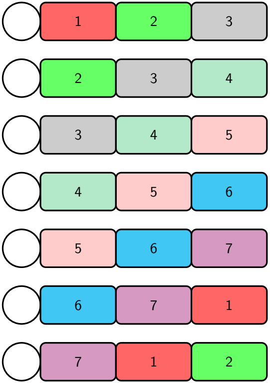
Normalized mean download time

$$\frac{1}{V}\mathbb{E}\sum_{\ell=0}^{V-1}\frac{1}{N_\ell} \geqslant \frac{1}{\frac{1}{V}\sum_{\ell=0}^{V-1}\mathbb{E}N_\ell}$$

Optimality condition for storage scheme

Maximize the normalized mean number of useful servers averaged

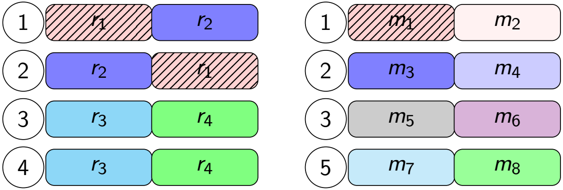
Latency optimal storage and access



A unit size divisible message $m = (m_1, \dots, m_V)$

- replicated $R = \alpha B / V$ times
- **storage:** for each fragment, where to store each replica?
- **access:** for each server, sequence of access for replicas?

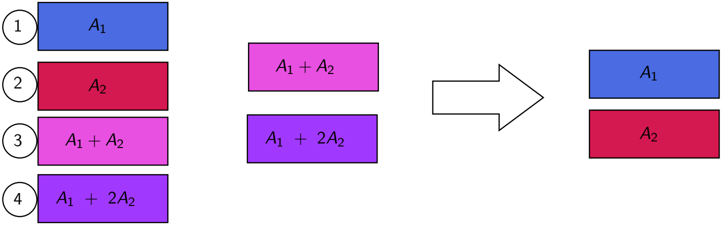
MDS coded storage



Optimality of MDS coded storage

- Sequence of number of useful servers is the largest
- Latency optimal storage code

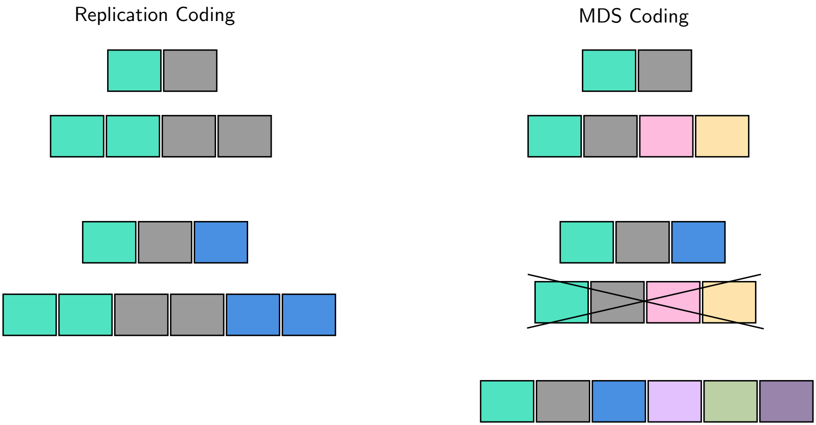
Decoding complexity



Implementation challenges

- Requires sufficiently large alphabet or large fragment sizes
- Polynomial decoding complexity that can't be parallelized

Scaling issues of MDS coding



Encoding growing data or redundancy

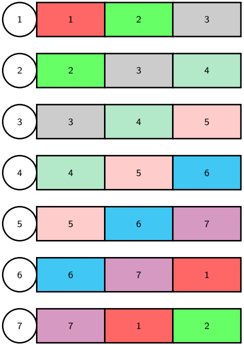
- Complete re-encoding of data blocks
- Potential data loss waiting for sufficient data blocks

Replication coded storage

α-(V, R) replication coded storage over B servers

$$\mathcal{S} \triangleq \{(S_1, S_2, \ldots, S_B) : |S_b| = \alpha V \text{ for all } b, \alpha = R/B\}.$$

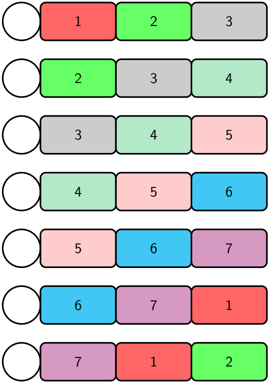
$\frac{3}{7} - (7, 3)$ replicated storage



1	1	2	3
2	2	3	4
3	3	4	5
4	4	5	6
5	5	6	7
6	6	7	1
7	7	1	2

- Fragment sets $S_1 = \{1, 2, 3\}, S_2 = \{2, 3, 4\}, \ldots$

Problem statement

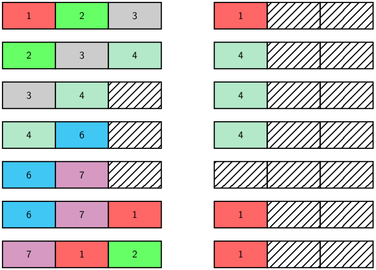


1	2	3
2	3	4
3	4	5
4	5	6
5	6	7
6	7	1
7	1	2

Find optimal storage scheme

$$S^* = \arg\max_{S \in \mathcal{S}} \frac{1}{V} \sum_{\ell=0}^{V-1} \mathbb{E} N_\ell .$$

Upper bound on number of useful servers N_ℓ

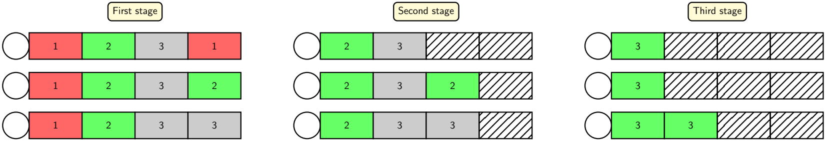


Upper bound

- For $m \triangleq \lceil B/R \rceil$, we have $N_\ell \leqslant B\mathbb{1}_{\{\ell \leqslant V-m\}} + (V-\ell)R\mathbb{1}_{\{\ell > V-m\}}$
- Normalized average of number of useful servers is upper bounded as

$$\frac{1}{BV}\sum_{\ell=0}^{V-1} N_\ell \leqslant 1 - \frac{(m+1)}{2V}.$$

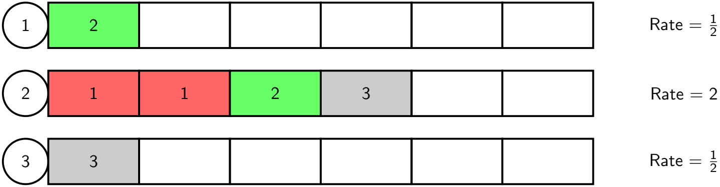
Trivial case: $\alpha \geqslant 1$



Replication as good as MDS without memory constraint

- Each server can store all the fragments
- All servers remain useful throughout
- What if $\alpha < 1$?

Randomized (B, V, R) replication coded storage

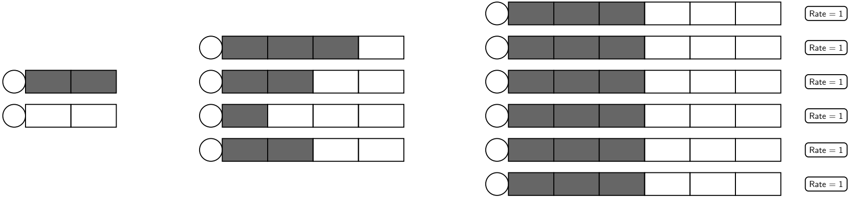


Place the fragments on randomly chosen servers

- Each server can store all coded VR fragments
- Exponential download rate $\propto$ the number of stored fragments

Asymptotically an α-(V, R) storage

- As V is increased with R/B fixed
- normalized storage at any server converges to $\alpha = R/B$
- service rate of servers converge to unity for almost all downloads



Asymptotic optimality

The randomized (B, V, R) storage scheme is an α-(V, R) storage scheme asymptotically in V.

Performance of Random Replication Storage

i.i.d. random storage vector Θ where $P\{\Theta_{vr} \neq b\} = (1 - 1/B)$

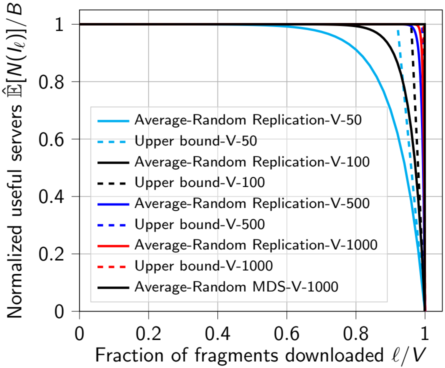
- $N_\ell = B - \sum_{b \in [B]} \prod_{v \notin I_\ell} \prod_{r \in [R]} \mathbb{1}_{\{\Theta_{vr} \neq b\}}$.
- $\frac{1}{BV}\mathbb{E}N_\ell = \frac{1}{V}\left(1 - \left(1 - \frac{1}{B}\right)^{R(V-\ell)}\right)$

Mean number of useful servers

For the random (B, V, R) replication storage ensemble,

$$\frac{1}{BV}\sum_{\ell=0}^{V-1}\mathbb{E}N_\ell = 1 - \frac{\left(1 - \frac{1}{B}\right)\left(1 - (1 - \frac{1}{B})^{RV}\right)}{V\left(1 - (1 - \frac{1}{B})^{R}\right)}$$

Numerical Results



Conclusion

- We studied codes for distributed storage system with storage constraints and file subfragmentation for achieving low latency
- For exponential download times, we proposed to maximize mean number of useful servers instead of minimizing latency
- We show that MDS codes are optimal
- When there are no memory constraints at the server, replication coded file can be optimally placed
- When servers have memory constraints, we show that replication coding combined with probabilistic placement are optimal asymptotically

Practical storage and access

- Placement of coded fragments depends on certain properties of storage codes
- Optimal access sequence is a Markov decision process
- We have heuristic solution to both questions
- Optimal placement remains open

Acknowledgements

References

- R. Jinan, A. Badita, P. Sarvepalli, P. Parag. Low latency replication coded storage over memory-constrained servers. ISIT 2021.
- R. Jinan, A. Badita, P. Sarvepalli, P. Parag. Latency optimal storage and scheduling of replicated fragments for memory-constrained servers. arXiv, Sep. 2020. Under review at TIT.
- A. Badita, P. Parag, and J.-F. Chamberland. Latency analysis for distributed coded storage systems. IEEE Transactions on Information Theory. 65(8):4683–4698, Aug 2019.
- Vaneet Aggarwal and Tian Lan. Modeling and optimization of latency in erasure-coded storage systems. Foundations and Trends in Communications and Information Theory. Vol. 18, Issue 3, pp 380–525, 2021.