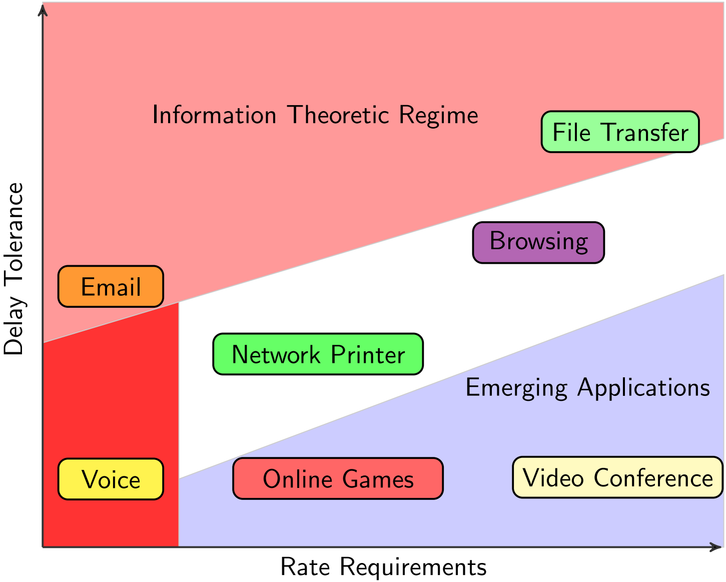
Latency Analysis for Distributed Storage

Parimal Parag
Archana Bura
Jean-Francois Chamberland

Electrical Communication Engineering
Indian Institute of Science

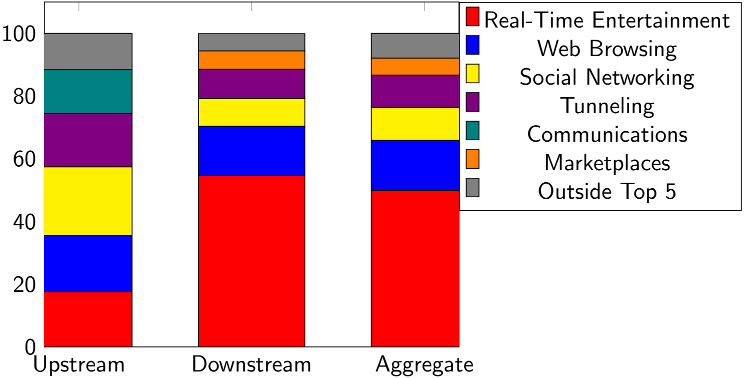
IBM Research, Bangalore
Feb 10, 2017

Evolving Digital Landscape



Dominant traffic on Internet

Peak Period Traffic Composition (North America)



- Real-Time Entertainment: 62% for fixed access and 43% for mobile access[1]

[1] https://www.sandvine.com/trends/global-internet-phenomena

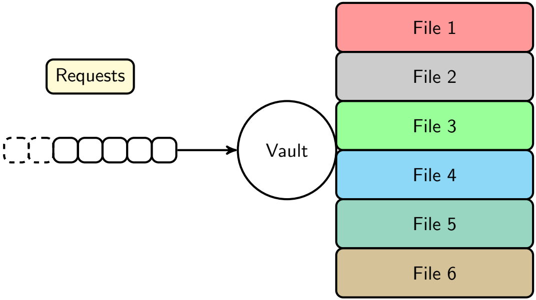
Building a Stronger Cloud

Cloud Readiness Characteristics

- Network access and broadband ubiquity
- Download and upload speeds
- Delays experienced by users are due to high network and server latencies

Reducing delay in delivering packets to and from the cloud is crucial to delivering advanced services

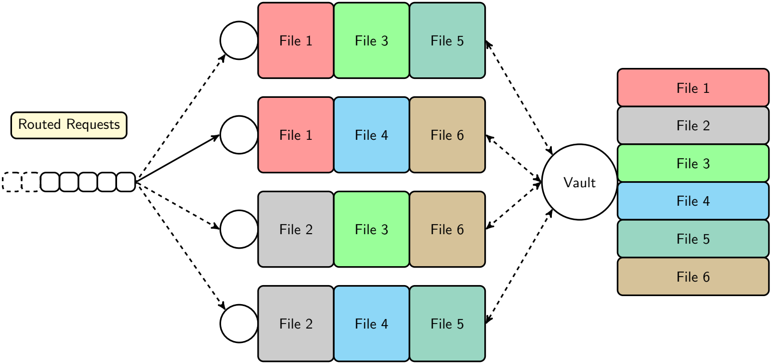
Centralized Paradigm – Media Vault



Potential Issues with Centralized Scheme

- Traffic load: Vault must handle all requests
- Service rate: Large storage entails longer access time
- Not robust to hardware failures or malicious attacks

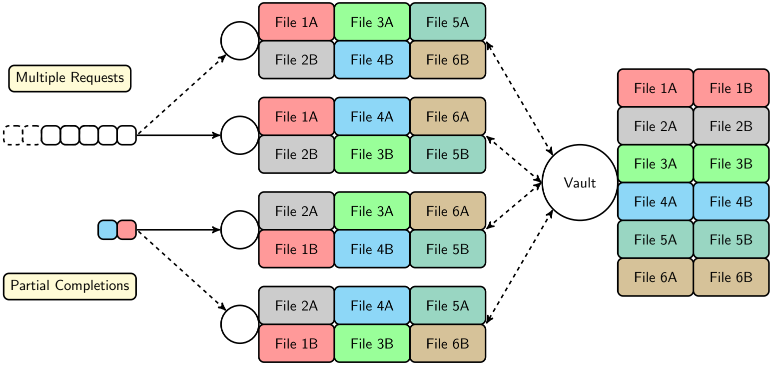
Established Solutions – Content Delivery Network



Congestion Prevention and Outage Protection

- Mirroring content with local servers
- Media file on multiple servers

Load Balancing through File Fragmentation



Shared Coherent Access

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

Coding for Distributed Storage Systems

Pertinent References (very incomplete)

N. B. Shah, K. Lee, and K. Ramchandran, "When do redundant requests reduce latency?" IEEE Trans. Commun., 2016.

G. Joshi, Y. Liu, and E. Soljanin, "On the delay-storage trade-off in content download from coded distributed storage systems" IEEE Journ. Spec. Areas. Commun., 2014.

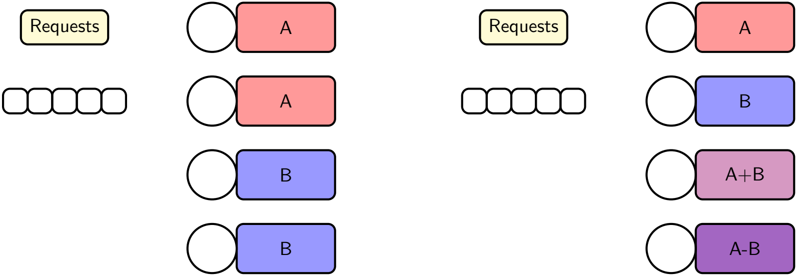
Network Coding for Distributed Storage Systems at IEEE TIT 2010 by Dimakis, Godfrey, Wu, Wainwright, and Ramchandran

A. Eryilmaz, A. Ozdaglar, M. Médard, and E. Ahmed, "On the delay and throughput gains of coding in unreliable networks," IEEE Trans. Info. Theory, 2008.

D. Wang, D. Silva, F. R. Kschischang, "Robust Network Coding in the Presence of Untrusted Nodes", IEEE Trans. Info. Theory, 2010.

A. Dimakis, K. Ramchandran, Y. Wu, C. Suh, "A Survey on Network Codes for Distributed Storage", Proceedings of IEEE, 2011.

Problem Statement



Question

For a single message with k fragments, how should one encode fragments and store them at the distributed storage nodes to reduce mean access time? Does coding offer any latency gains?

Answer

Coded storage offers scaling gains over replication.

System Model

File storage

- Each media file divided into k pieces
- Pieces encoded and stored on n servers

Arrival of requests

- Each request wants entire media file
- Poisson arrival of requests with rate λ

Time in the system

- Till the reception of whole file

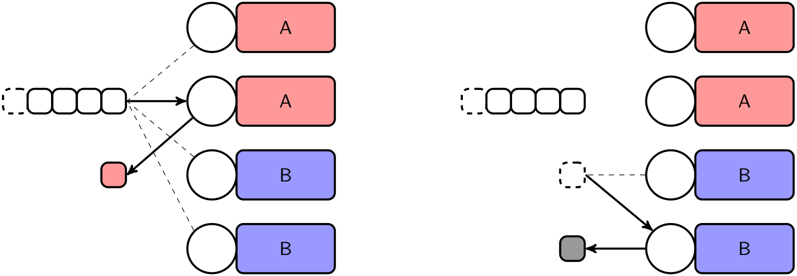
Service at each server

- IID exponential service time with rate k/n

Replication: Distribute Pieces across Servers

Typical Sequence for Replication Scheme

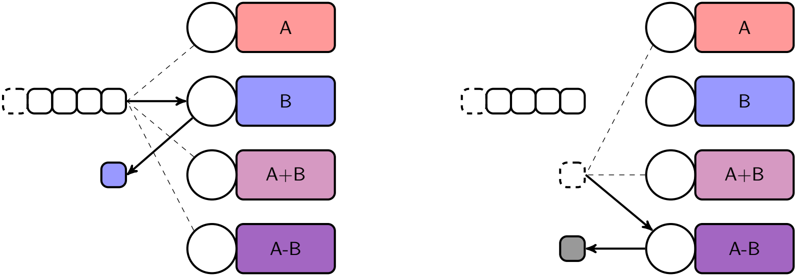
- Obtain first piece from any server
- Get second piece from constrained set



Network Coding: Create Independent Blocks

Typical Sequence for Coded Scheme

- Obtain first piece from any server
- Get second piece from complement set



State Space

Replication

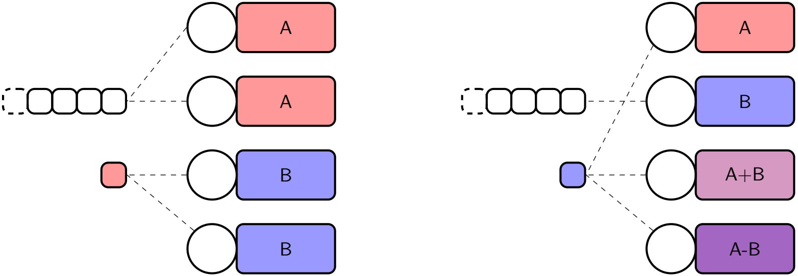
- Number of requests $Y_S(t)$ with subset of information symbols $S \subset [k]$ at time t
- $\bar{Y}(t) = \{Y_S(t) : S \subset [k]\}$ is a Markov process

Coding

- Number of requests $Y_S(t)$ with subset of information symbols $S \subset [n]$ at time t
- $\bar{Y}(t) = \{Y_S(t) : S \subset [n], |S| < k\}$ is a Markov process

Scheduling Model

- Parallel processing at all “useful” servers
- Non-useful servers stop serving



State Space Reduction

Theorem

For the repetition and coding schemes under priority scheduling and parallel processing model, the collection

$$\mathcal{S}(t) = \{S : Y_S(t) > 0, |S| < k\}$$

of information subsets at any time t is totally ordered in terms of set inclusion.

Corollary

Let $Y_i(t)$ be the number of requests with i information symbols at time t, then

$$Y(t) = (Y_0(t), Y_1(t), \ldots, Y_{k-1}(t)),$$

is a Markov process.

State Transitions

Arrival

- Unit increase in $Y_0(t) = Y_0(t-) + 1$ with rate λ

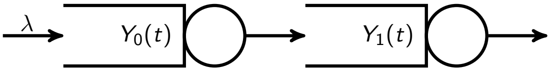
Getting additional symbol

- Unit increase in $Y_i(t) = Y_i(t-) + 1$
- Unit decrease in $Y_{i-1}(t) = Y_{i-1}(t-) - 1$

Getting last remaining symbol

- Unit decrease in $Y_{k-1}(t) = Y_{k-1}(t-) - 1$

Tandem Queue Interpretation



Replication

- If all states non-empty
- Number of useful servers available to level i are n/k
- Service time of each server is iid exponential with rate k/n
- Service rate at ith level is

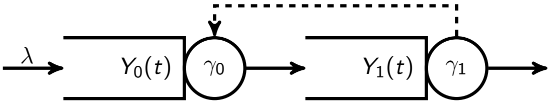
$$\gamma_i = 1, \qquad i = 0, \dots, k-1.$$

Coding

- If all states non-empty
- One useful server available to level $i \neq k-1$
- Service time of each server is iid exponential with rate k/n
- Service rate at i th level is

$$\gamma_i = \begin{cases} \frac{k}{n} & i < k-1, \\ \frac{k}{n}(n-k+1) & i = k-1, \end{cases}$$

State Transition Rates



Pooled Tandem Queue

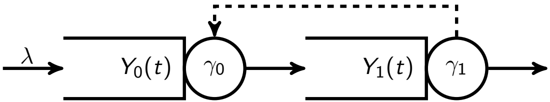
- Next occupied information level

$$l_i(t) = k \wedge \{l > i : Y_l(t) > 0\}$$

- All useful servers for level i are helping levels above it
- All useful severs for level i that are available are below $l_i(y)$
- Aggregate service available at level i is

$$\sum_{j=i}^{l_i(t)-1} \gamma_j$$

Multi-dimensional Markov Process



Generator Matrix

- Generator matrix for the Markov process $Y(t)$

$$Q(y, y+e_0) = \lambda,$$

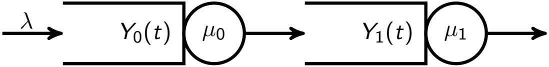
$$Q(y, y-e_i+e_{i+1}) = \sum_{j=i}^{l_i(y)-1} \gamma_j 1\{y_i > 0\},\ i = 0, \dots, k-2$$

$$Q(y, y-e_{k-1}) = \gamma_{k-1} 1\{y_{k-1} > 0\}.$$

- No known technique to compute stationary distribution of multi-dimensional Markov processes

Bounding and Separating

Theorem



If $\lambda < \min \mu_i$, then the tandem queue has a product form distribution,

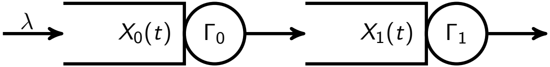
$$\pi(y) = \prod_{i=0}^{k-1} \frac{\lambda}{\mu_i} \left(1 - \frac{\lambda}{\mu_i}\right)^{y_i}$$

Lemma

The transition rates $Q(y, e_{i+1}(y))$ are bounded by

$$\gamma_i < \sum_{j=i}^{l_i(y)-1} \gamma_j < \sum_{j=i}^{k-1} \gamma_j \triangleq \Gamma_i .$$

Lower Bounding Tandem Queue

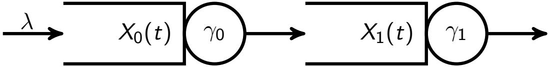


Theorem

Each queue in the lower bounding system has Poisson arrival rate λ and independent exponential service time Γ_i, and hence the stationary distribution is

$$\pi(y) = \prod_{i=0}^{k-1} \frac{\lambda}{\Gamma_i} \left(1 - \frac{\lambda}{\Gamma_i}\right)^{y_i}$$

Upper Bounding Tandem Queue

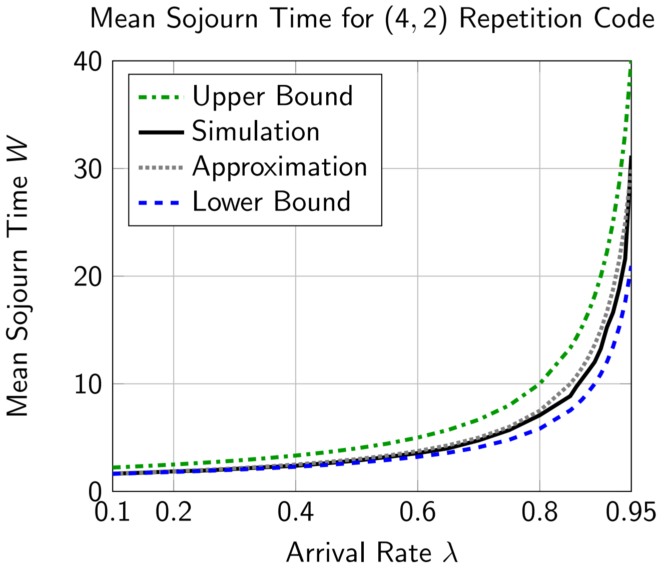


Theorem

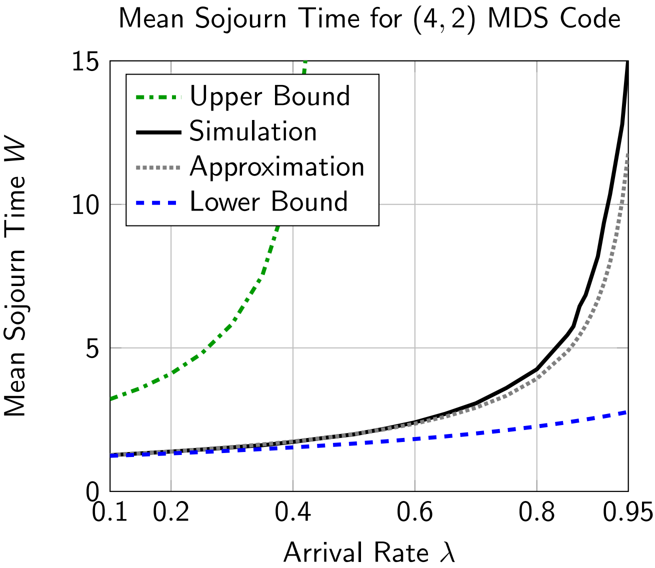
Each queue in the upper bounding system has Poisson arrival rate λ and independent exponential service time γ_i, and hence the stationary distribution is

$$\pi(y) = \prod_{i=0}^{k-1} \frac{\lambda}{\Gamma_i} \left(1 - \frac{\lambda}{\Gamma_i}\right)^{y_i}$$

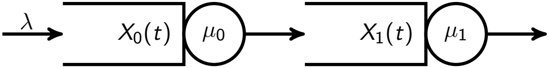
Bounds for Replication



Bounds for Coding



Approximating Pooled Tandem Queue



Independent Approximation

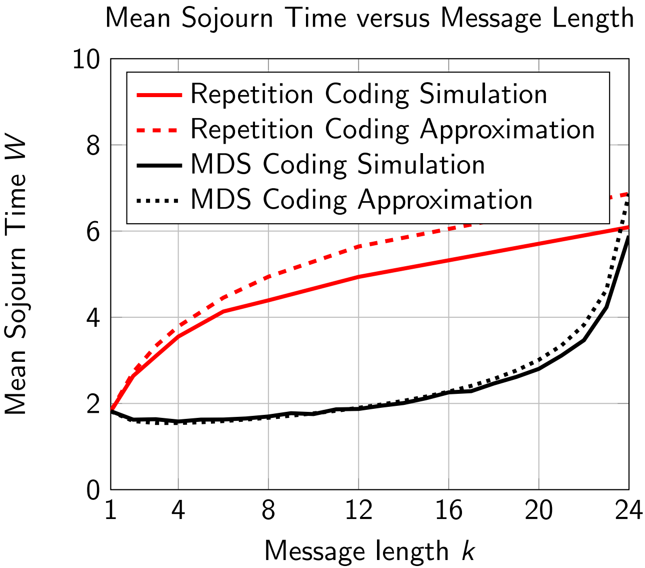
Each queue has Poisson arrival rate λ and independent exponential service time μ_i such that

$$\mu_i = \begin{cases} \gamma_{k-1} & i = k-1, \\ \gamma_i + \mu_{i+1}\pi_{i+1}(0) & \end{cases}$$

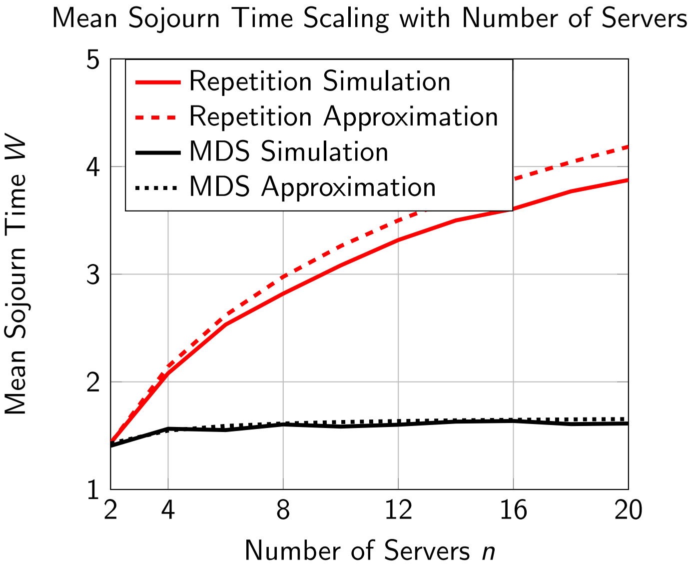
Then the service rate can be written as

$$\mu_i = \Gamma_i - (k - i + 1)\lambda.$$

Comparing Replication vs MDS Coding



Comparing Replication vs MDS Coding



Discussion and Concluding Remarks

Main Contributions

- Analytical framework for study of distributed computation and storage systems
- Upper and lower bounds to analyse replication and MDS codes
- A tight closed-form approximation to study distributed storage codes
- MDS codes are better suited for large distributed systems
- Mean access time is better for MDS codes for all code-rates

Thank You