Latency analysis for Distributed Storage

Parimal Parag
Archana Bura
Jean-François Chamberland

Electrical Communication Engineering
Indian Institute of Science

Electrical and Computer Engineering
Texas A&M University

National Conference on Communication
Mar 3, 2017

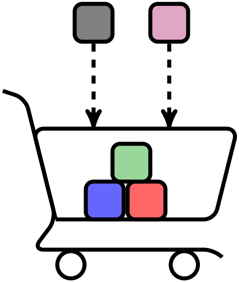
Building a Stronger Cloud

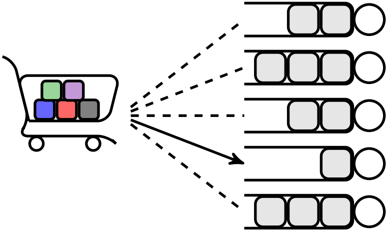
Cloud Readiness Characteristics

- 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

Supermarket Models Revisited





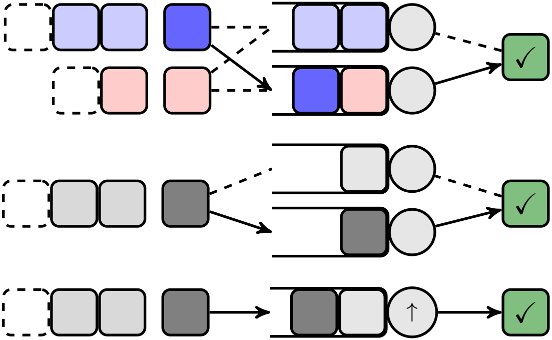
Shopping Tasks

- Acquiring listed items
- Sequence of queues
- Sum of waiting times

Checkout Process

- Select one queue
- FIFO policy
- Waiting time in 1 queue

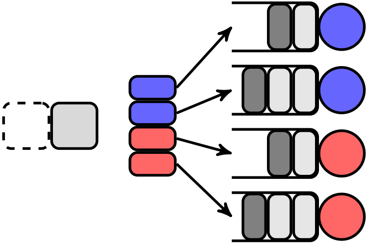
Traditional Queueing Analysis

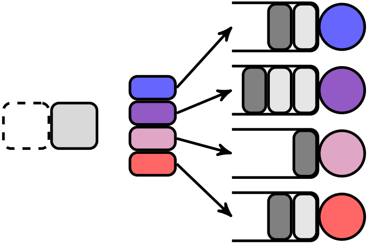


Measures for Enhancing Performance

- Improve server speed
- Increase number of server per flow
- Pool resources and load balance

Question: Duplication versus MDS Coding





Reduction of access time

- How many fragments should a single message be divided into?
- How should one encode and store at the distributed storage nodes?

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

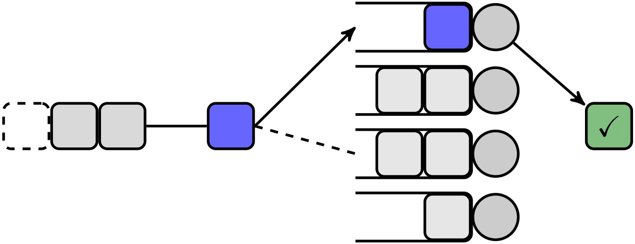
Time in the system

- Till the reception of whole file

Service at each server

- IID exponential service time with rate k/n

Supermarket Model: Power of 2 Choices



Assumptions

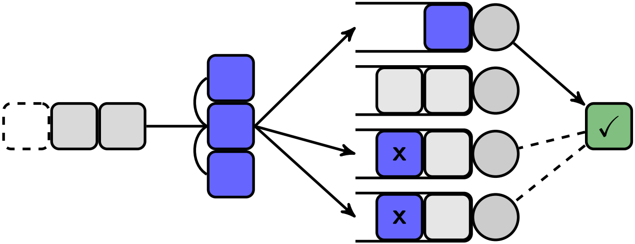
- Prior info: d queues
- FIFO, one copy
- Feedback: none

Findings

- Exponential improvements in expected time for $d = 2$ over $d = 1$
- Constant factor thereafter

exempli gratia: Karp, Luby, Meyer auf der Heide, (1992); Adler, Chakrabarti, Mitzenmacher, Rasmussen (1995); Vvedenskaya, Dobrushin, Karpelevich (1996); Mitzenmacher (2001); Ying, Srikant, Kang (2015)

Supermarket Model: Redundancy-d Systems



Assumptions

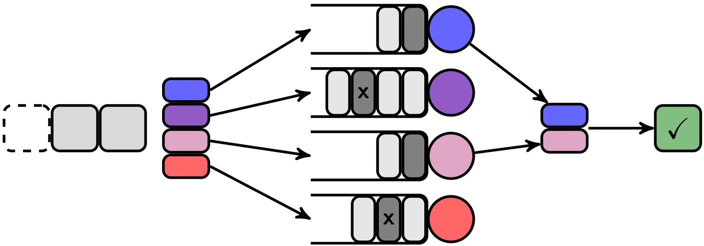
- Prior info: none†
- FIFO, d copies
- Feedback: cancellation
- Clairvoyance gain

Findings

- A little redundancy goes a long way
- Local balance equations
- Exact queue distribution

exempli gratia: Gardner, Zbarsky, Doroudi, Harchol-Balter, Hyytiä, Scheller-Wolf (2015); Gardner, Harchol-Balter, Scheller-Wolf, Velednitsky, Zbarsky (2016)

Storage Coding – (n, k) Fork-Join Model



Assumptions

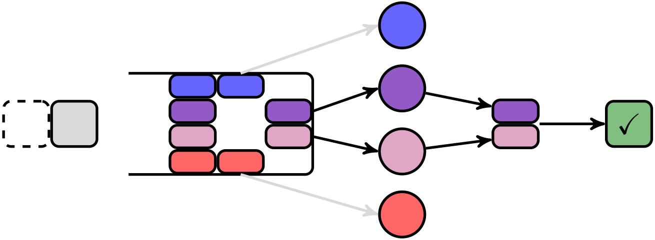
- Prior info: none†
- FIFO, k out of n copies
- Feedback: cancellation
- Clairvoyance gain

Findings

- Coding exploits diversity better than redundancy
- $\mathrm{E}[T] \leq$ split-merge
- Cascade $\leq \mathrm{E}[T]$

exempli gratia: Joshi, Liu, Soljanin (2012, 2014), Joshi, Soljanin, Wornell (2015), Sun, Zheng, Koksal, Kim, Shroff (2015), Kadhe, Soljanin, Sprintson (2016), Li, Ramamoorthy, Srikant (2016)

Storage Coding – The Centralized MDS Queue



Assumptions

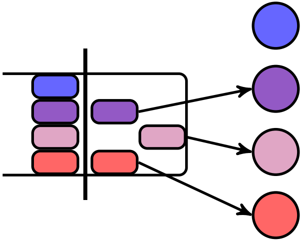
- Info: global loads
- FIFO, k out of n copies
- Feedback: cancellation

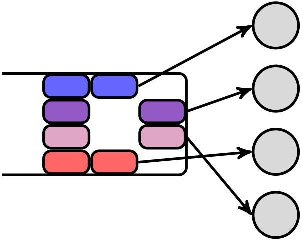
Challenges

- Intricate QBD Markov process
- Infinite states in n dimensions
- Tightly coupled transitions

exempli gratia: Shah, Lee, Ramchandran (2013), Lee, Shah, Huang, Ramchandran (2017), Vulimiri, Michel, Godfrey, Shenker (2012), Ananthanarayanan, Ghodsi, Shenker, Stoica (2012) Baccelli, Makowski, Shwartz (1989)

Storage Coding – The Centralized MDS Queue





MDS-Reservation(t)

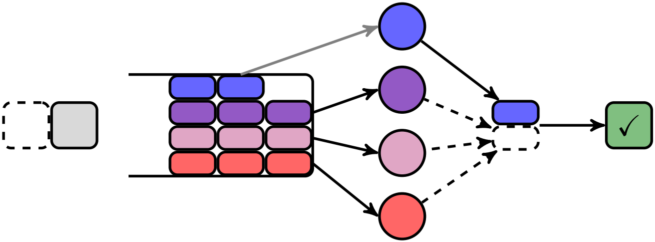
- Restriction on depth of scheduler
- Reduces dimension of chain
- Upper bound on $\mathrm{E}[T]$

MDS-Violation(t)

- Unconstained servers
- Equivalent to resource pooling without coding
- Lower bound on $\mathrm{E}[T]$

Shah, Lee, Ramchandran (2013), Lee, Shah, Huang, Ramchandran (2017)

Proposed Model: Priority Policy



Assumptions

- Info: global loads
- Policy: shortest (expected) remaining time
- k out of n copies
- Feedback: cancellation

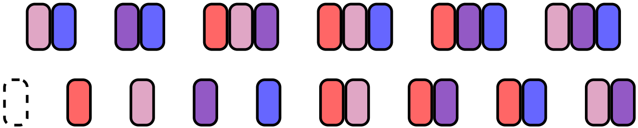
Approach

- Intricate Markov process
- State Collapse
- Framework amenable to MDS coding and duplication

PP, Jean-François Chamberland (ITA 2013), PP, Archana Bura, Jean-François Chamberland (ITA 2017, INFOCOM 2017)

gratias: Kannan Ramchandran, Salim El Rouayheb

State Space Structure



Keeping Track of Partially Fulfilled Requests

- Label distinct pieces with integers
- Element of state vector $Y_S(t)$ is number of users with given subets S of pieces

Continuous-Time Markov Chain

- $\mathbf{Y}(t) = \{Y_S(t) : S \subset [n]\}$ is a Markov process
- Markov process with local transitions

State Space Collapse

Theorem

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

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

of information subsets is totally ordered in terms of set inclusion

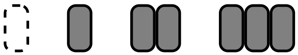
Corollary

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

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

is Markov process

State Transitions of Collapsed System



Arrival of Requests

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

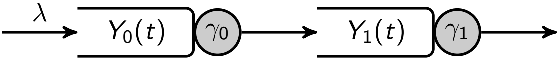
Getting Additional Symbol

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

Getting Last Missing Symbol

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

Tandem Queue Interpretation (No Empty States)



Duplication

- When all states non-empty
- No. servers available at level i is n/k
- Normalized service rate at level i

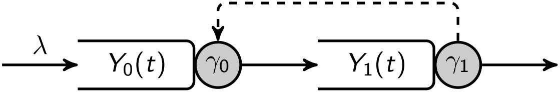
$$\gamma_i = 1 \quad i = 0, \ldots, k-1$$

MDS Coding

- When all states non-empty
- One server available at level $i \neq k-1$
- Normalized service rate at level i

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

Tandem Queue Interpretation (General Case)



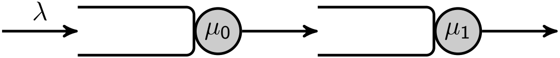
Tandem Queue with Pooled Resources

- Servers with empty buffers help upstream
- Aggregate service at level i becomes

$$\sum_{j=i}^{l_i(t)-1} \gamma_j \qquad \text{where} \qquad l_i(t) = k \wedge \{l > i : Y_l(t) > 0\}$$

- No explicit description of stationary distribution for multi-dimensional Markov process

Bounding and Separating



Theorem[†]

When $\lambda < \min \mu_i$, tandem queue has product form distribution

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

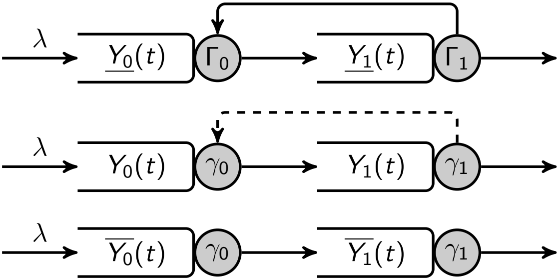
Uniform Bounds on Service Rate

Transition rates are uniformly bounded by

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

[†] F. P. Kelly, Reversibility and Stochastic Networks. New York, NY, USA: Cambridge University Press, 2011.

Bounds on Tandem Queue



Lower Bound

Higher values for service rates yield lower bound on queue distribution

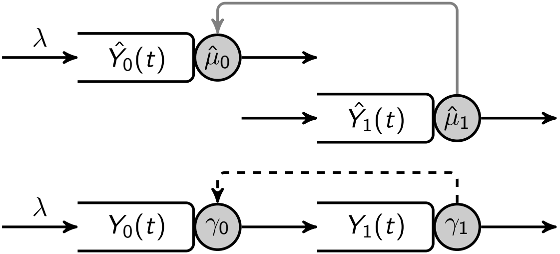
$$\underline{\pi}(y) = \prod_{i=0}^{k-1} \frac{\lambda}{\Gamma_i}\left(1 - \frac{\lambda}{\Gamma_i}\right)^{y_i}$$

Upper Bound

Lower values for service rate yield upper bound on queue distribution

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

Approximating Pooled Tandem Queue



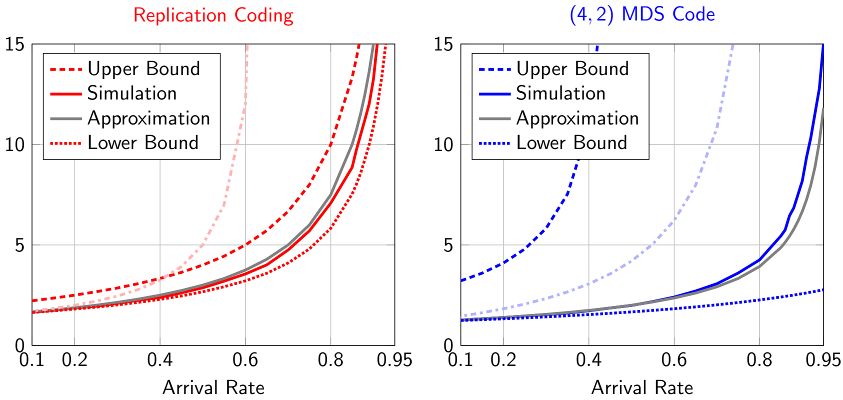
Independence Approximation with Statistical Averaging

Service rate is equal to base service rate γ_i plus cascade effect, averaged over time

$$\hat{\mu}_{k-1} = \gamma_{k-1}$$
$$\hat{\mu}_i = \gamma_i + \hat{\mu}_{i+1}\hat{\pi}_{i+1}(0)$$

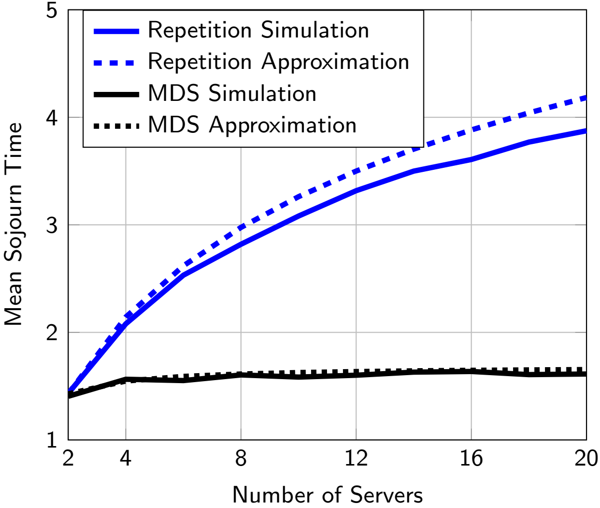
$$\hat{\pi}(y) = \prod_{i=0}^{k-1} \frac{\lambda}{\hat{\mu}_i}\left(1 - \frac{\lambda}{\hat{\mu}_i}\right)^{y_i}$$

Mean Sojourn Time



- MDS coding significantly outperforms replication
- Bounding techniques are only meaningful under light loads
- Approximation is accurate over range of loads

Comparing Replication versus MDS Coding



Arrival rate 0.3 units and coding rate $n/k = 2$

Summary and Discussion

Main Contributions

- Analytical framework for study of distributed computation and storage systems
- Upper and lower bounds to analyze 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