Novel Latency Bounds for Distributed Coded Storage

Jean-Francois Chamberland
Parimal Parag

Electrical and Computer Engineering
Texas A&M University

Electrical Communication Engineering
Indian Institute of Science

Information Theory and Applications
Feb 13, 2018

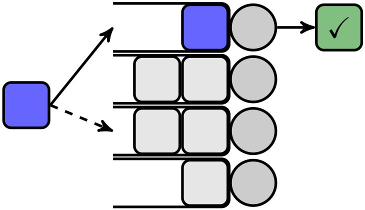
Building a Stronger Cloud

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

Inspirational Prior Work



Power of 2 Choices

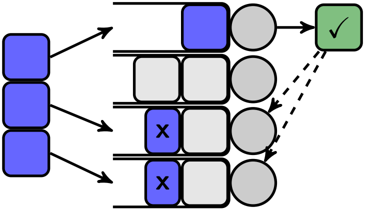
- FIFO; Info – d queues
- 1 copy w/o feedback
- Exponential gain, $d = 2$

e.g.: Karp, Luby, Meyer auf der Heide, (1992);

Adler, Chakrabarti, Mitzenmacher, Rasmussen (1995);

Vvedenskaya, Dobrushin, Karpelevich (1996);

Mitzenmacher (2001)



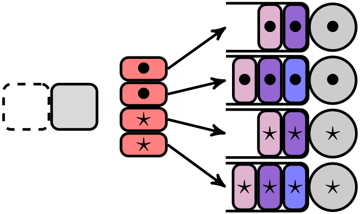
Redundancy-d Systems

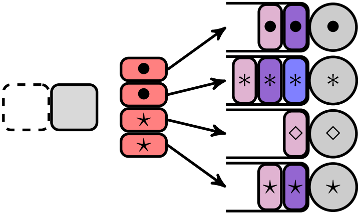
- FIFO; Info – none
- d copies w cancellation
- Exact queue distribution

e.g.: Gardner, Zbarsky, Doroudi, Harchol-Balter, Hyytiä, Scheller-Wolf (2015);

Gardner, Harchol-Balter, Scheller-Wolf, Velednitsky, Zbarsky (2016)

Duplication versus MDS Coding





Queueing Analysis

- Minimize expected delay
- MDS outperforms Repetition
- Elusive exact expression

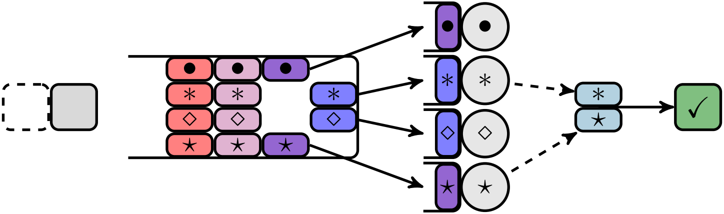
Canonical Example

- Four servers
- Two distinct pieces of information
- Find bounds

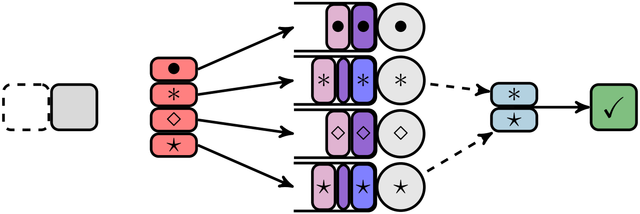
e.g.: Joshi, Liu, Soljanin (2012, 2014), Shah, Lee, Ramchandran (2013), Joshi, Soljanin, Wornell (2015), Sun, Zheng, Koksal, Kim, Shroff (2015), Kadhe, Soljanin, Sprintson (2016), Li, Ramamoorthy, Srikant (2016)

Model Variations for Distributed Storage

Centralized MDS Queue without Replication

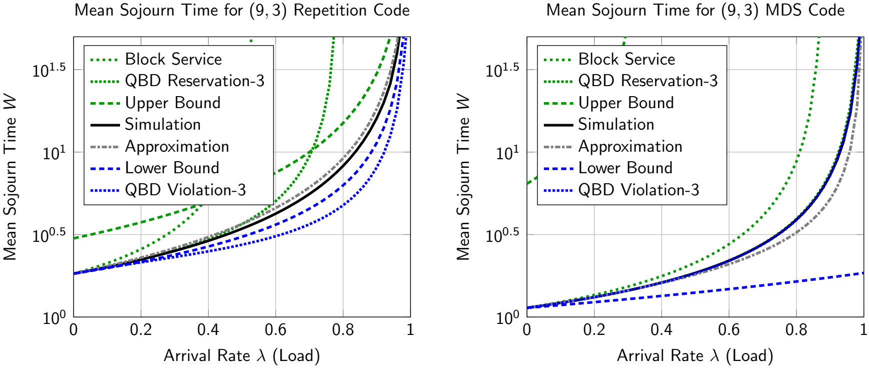


Distributed (n, k) Fork-Join Model with MDS Coding



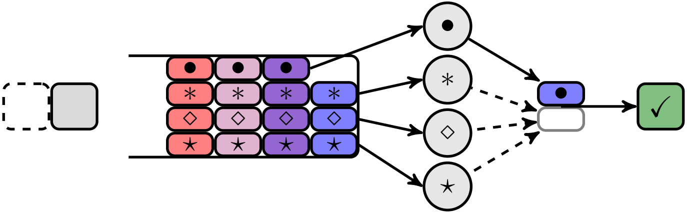
e.g.: Lee, Shah, Huang, Ramchandran (2017)

Mean Sojourn Time



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

Adopted Model: Priority Policy with MDS Coding



Assumptions

- FIFO, k out of n copies
- Information: global loads
- Feedback: cancellation
- MDS or replication

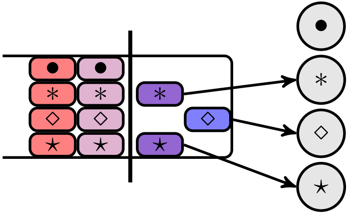
Challenges

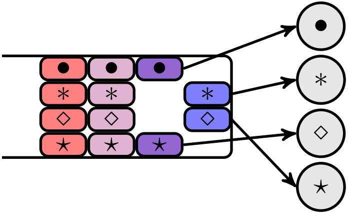
- Intricate QBD Markov process
- Infinite states in n dimensions
- Tightly coupled transitions

Parimal Parag, JFC (ITA 2013, ITA 2018), Parimal Parag, Archana Bura, JFC (ITA 2017, INFOCOM 2017)

gratias: Kannan Ramchandran, Salim El Rouayheb

Establishing Lower and Upper Bounds





MDS-Reservation(t)

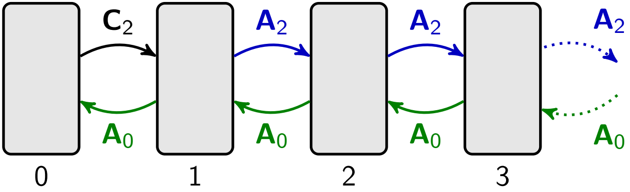
- Restriction on depth of scheduler
- Reduces dimension of chain
- Upper bound on $\mathrm{E}[T]$

MDS-Violation(t)

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

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

Aggregate System – Level Abstraction

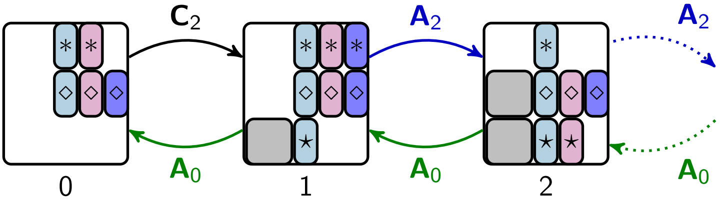


Transition Operator

$$\begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

- **Block partitioning** far more important than entries of submatrices
- $\mathbf{C}_1$ and $\mathbf{C}_2$ account for boundary conditions

Aggregate System – Stationary Distribution



Chapman-Kolmogorov Equations

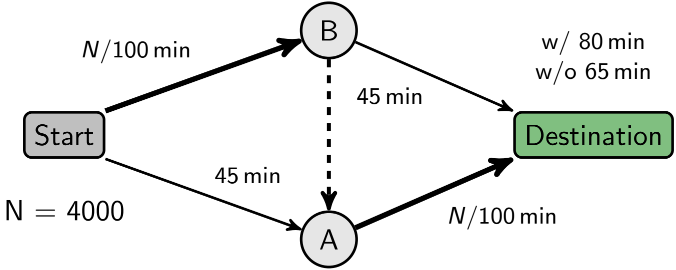
Stationary distribution, denoted $\pi = (\pi_0, \pi_1, \pi_2, \ldots,)$ with

$$\pi_q = \big(\, \Pr(s_1, q), \ldots, \Pr(s_k, q) \,\big)$$

is unique solution to balance equations

$$\pi_{\mathbf{q}} = \pi_{\mathbf{q-1}}\mathbf{A}_2 + \pi_{\mathbf{q}}\mathbf{A}_1 + \pi_{\mathbf{q+1}}\mathbf{A}_0$$

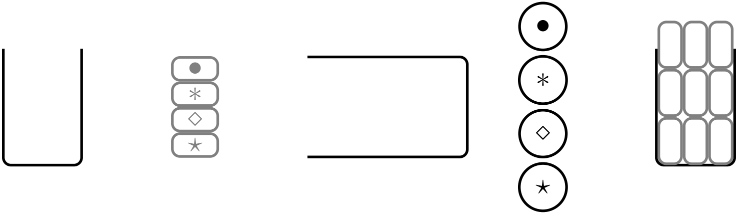
The Cautionary Tale of Braess's Paradox



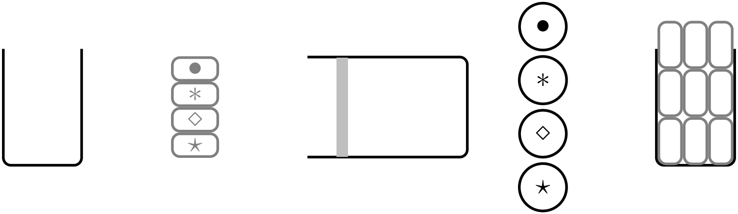
"For each point of a road network, let there be given the number of cars starting from it and the destination of the cars. Under these conditions, one wishes to estimate the distribution of traffic flow. [...] If every driver takes the path that looks most favorable to them, the resultant running times need not be minimal. Furthermore, it is indicated by an example that an extension of the road network may cause a redistribution of the traffic that results in longer individual running times."

Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

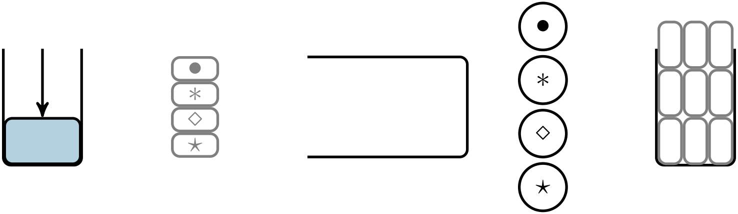


Eviction/Violation Lower Bound

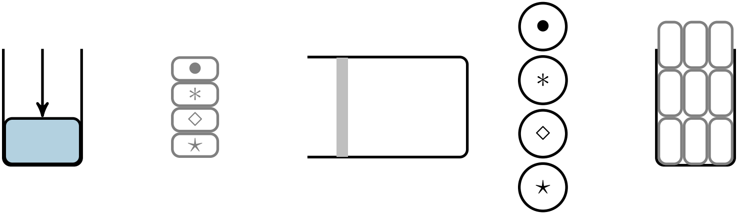


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

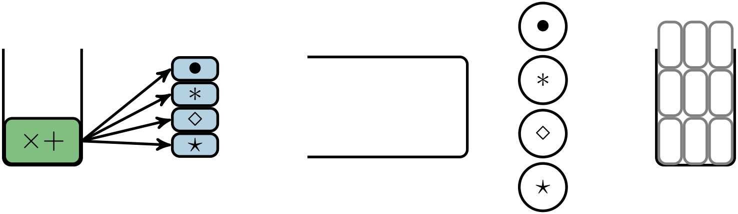


Eviction/Violation Lower Bound

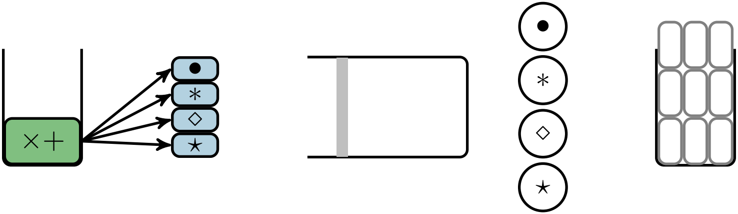


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

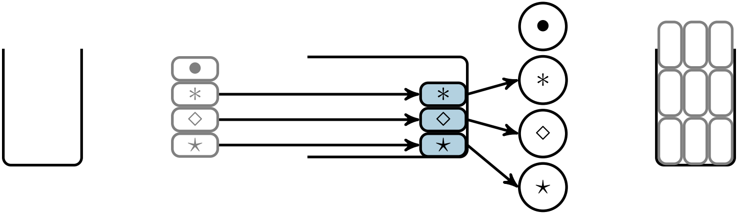


Eviction/Violation Lower Bound

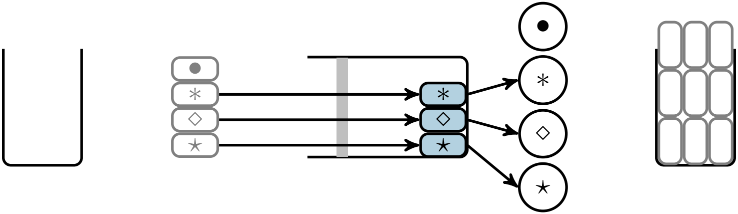


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

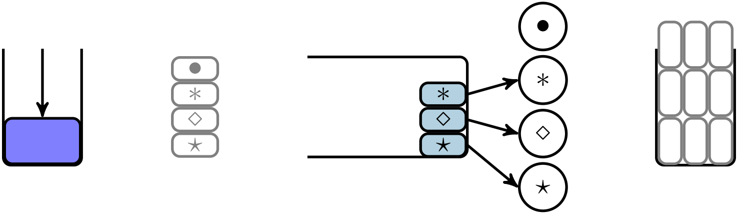


Eviction/Violation Lower Bound

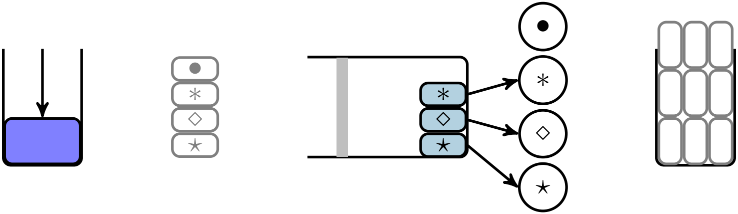


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

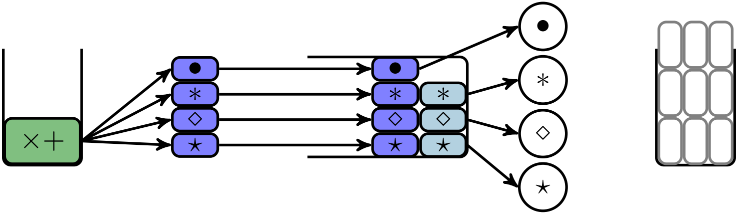


Eviction/Violation Lower Bound

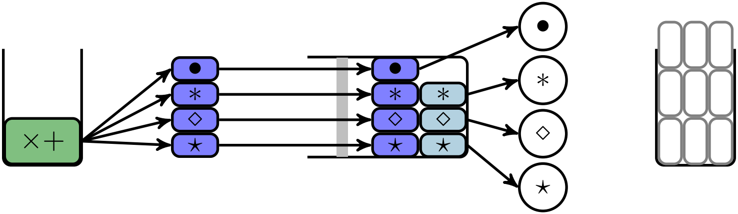


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

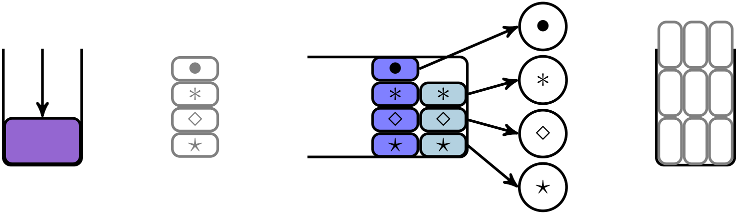


Eviction/Violation Lower Bound

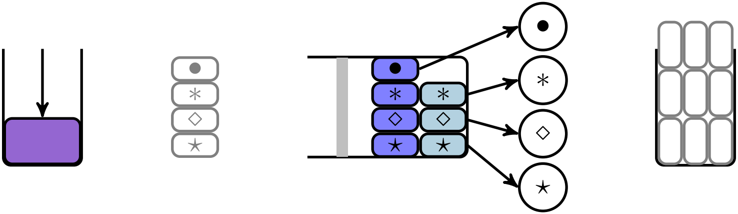


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

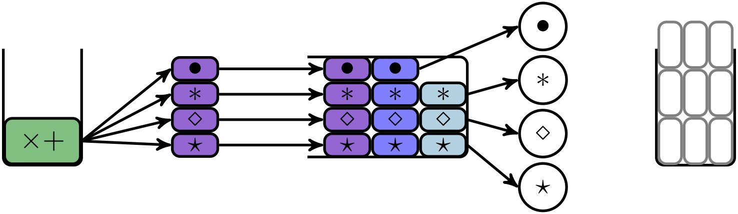


Eviction/Violation Lower Bound

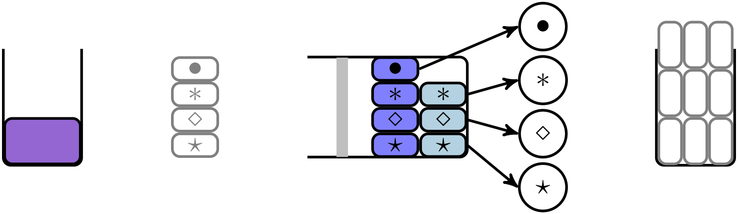


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

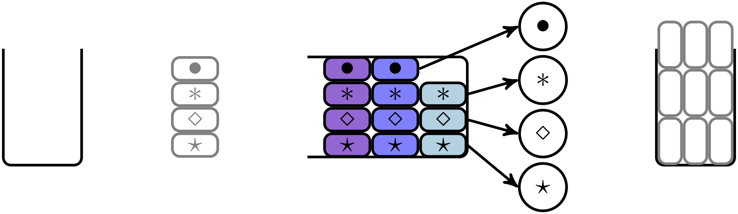


Eviction/Violation Lower Bound

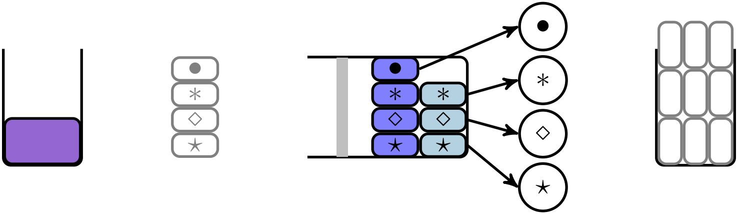


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

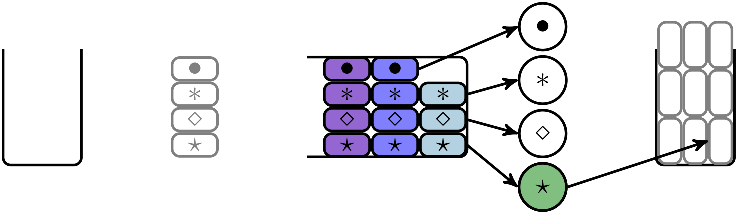


Eviction/Violation Lower Bound

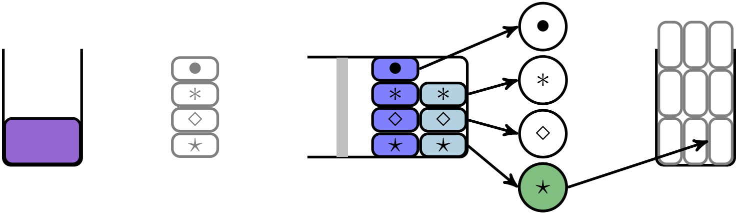


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

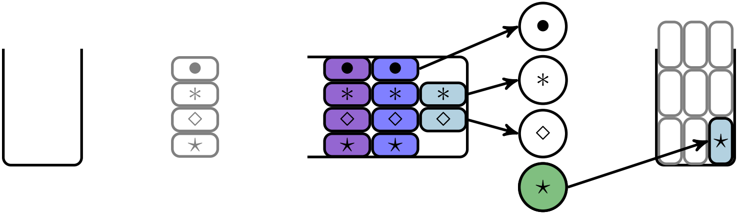


Eviction/Violation Lower Bound

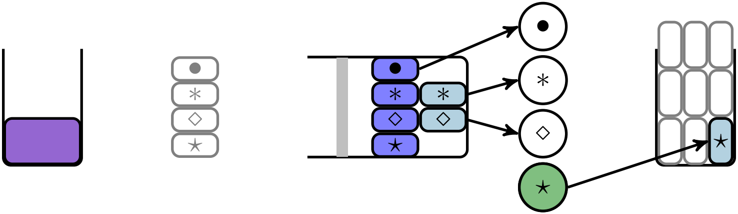


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

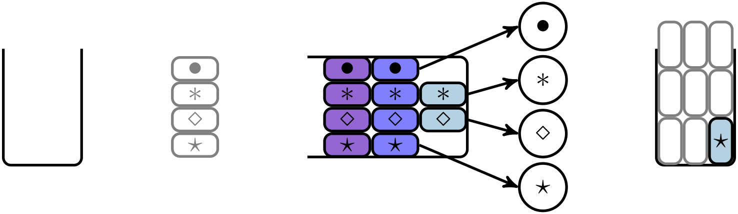


Eviction/Violation Lower Bound

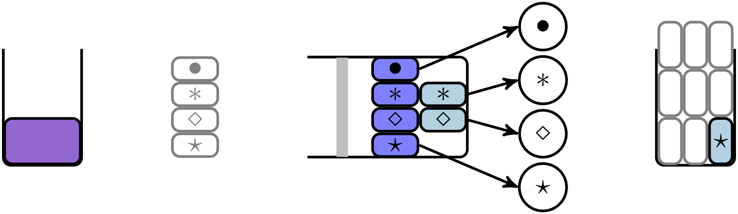


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

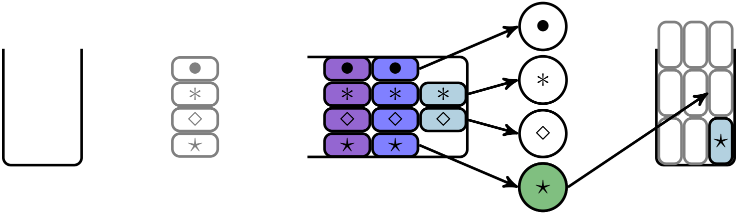


Eviction/Violation Lower Bound

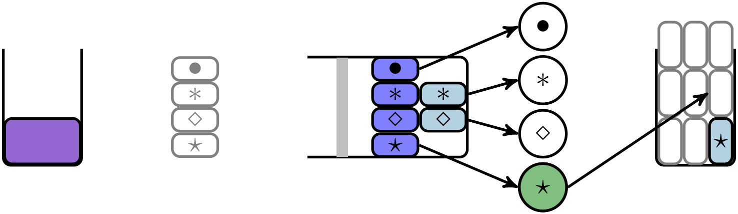


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

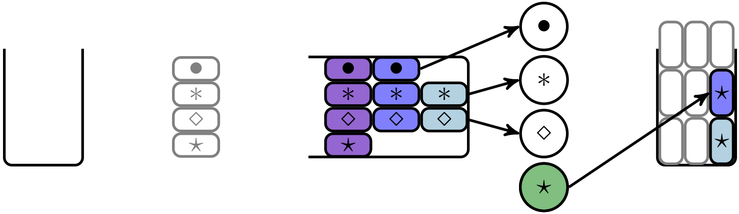


Eviction/Violation Lower Bound

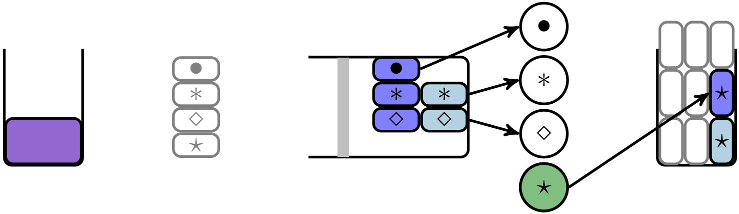


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

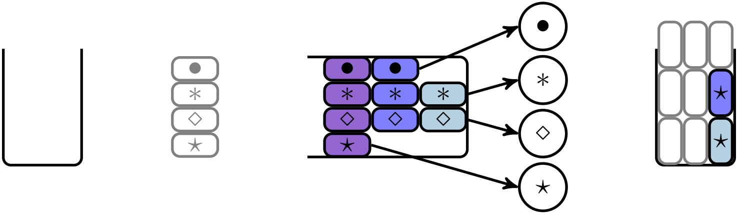


Eviction/Violation Lower Bound

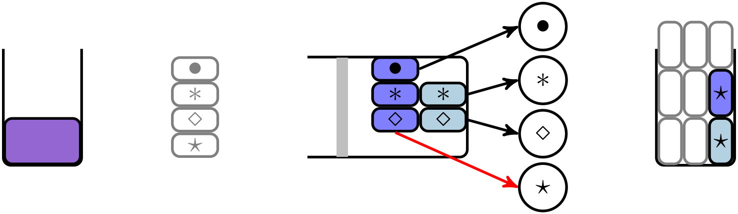


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

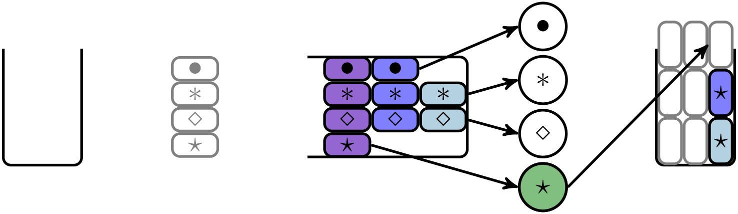


Eviction/Violation Lower Bound

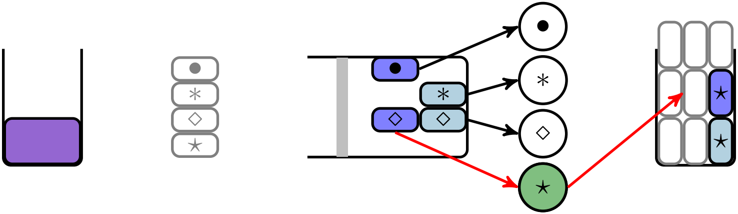


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

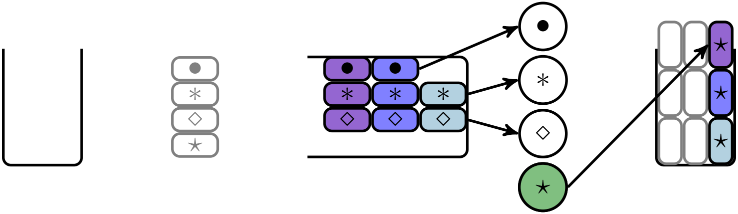


Eviction/Violation Lower Bound

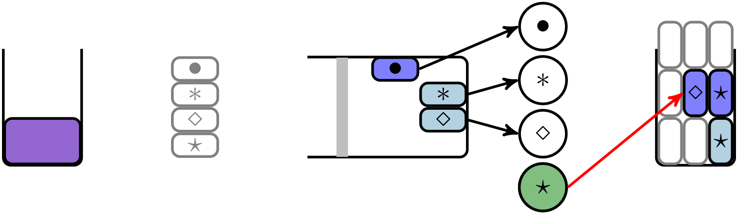


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

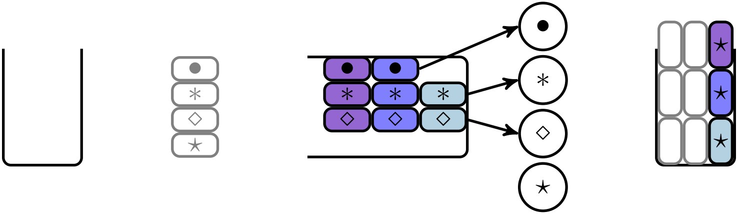


Eviction/Violation Lower Bound

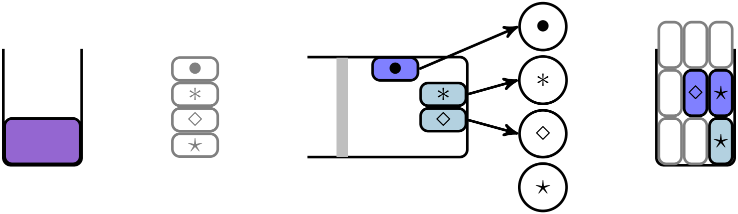


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

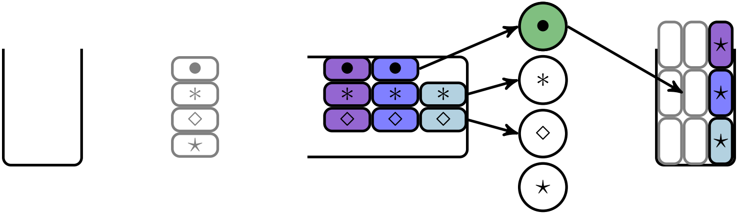


Eviction/Violation Lower Bound

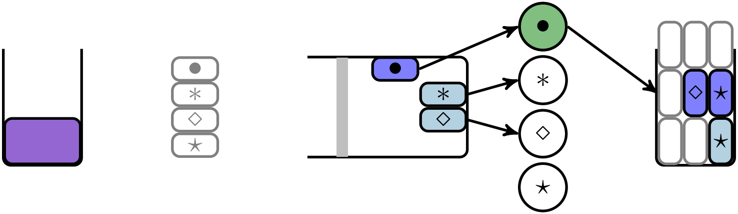


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

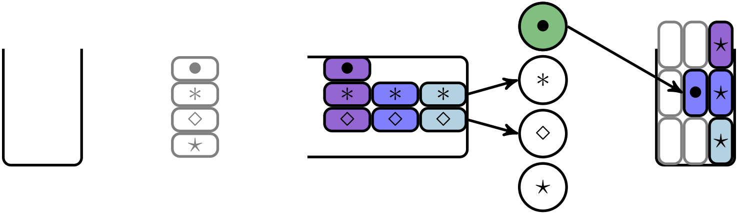


Eviction/Violation Lower Bound

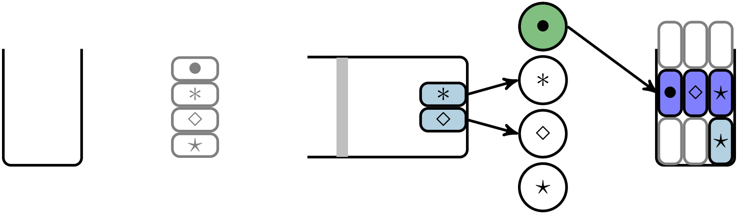


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

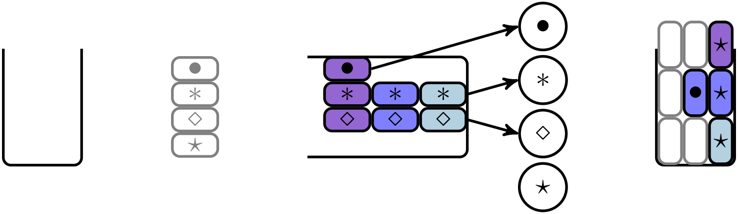


Eviction/Violation Lower Bound

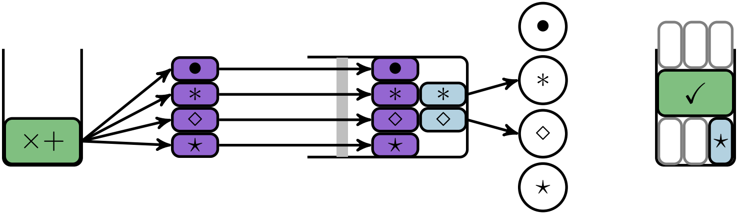


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

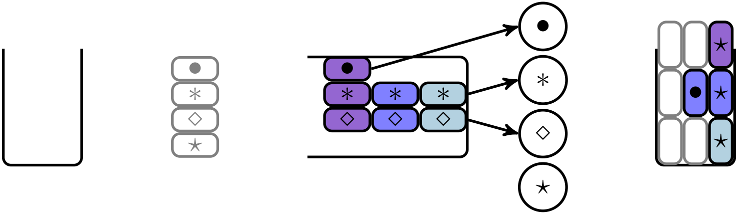


Eviction/Violation Lower Bound

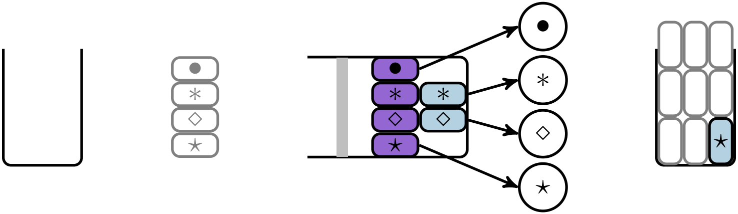


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

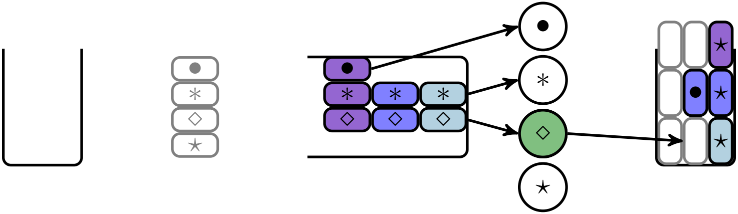


Eviction/Violation Lower Bound

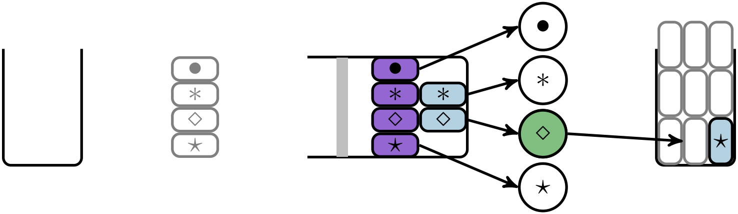


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

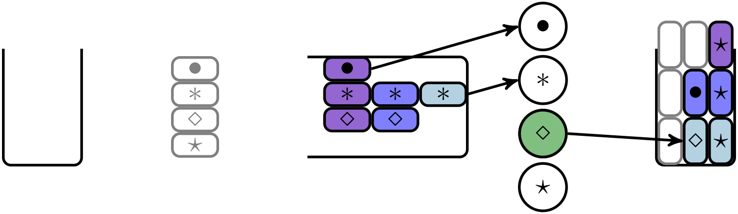


Eviction/Violation Lower Bound

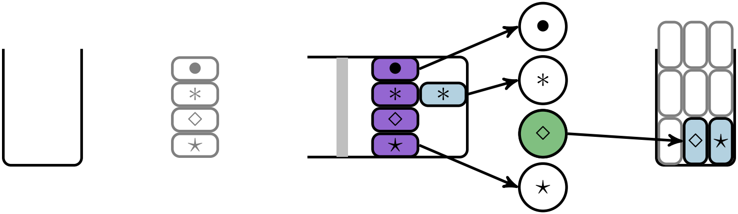


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

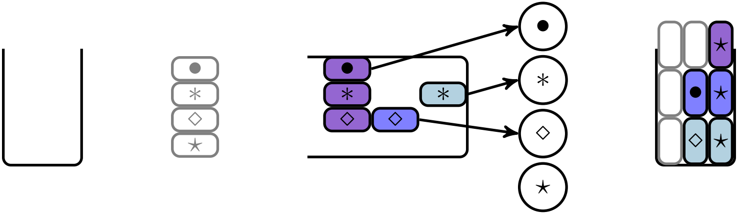


Eviction/Violation Lower Bound

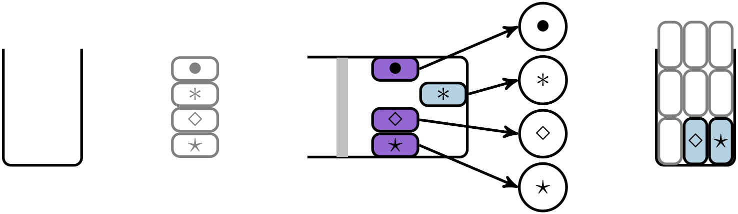


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

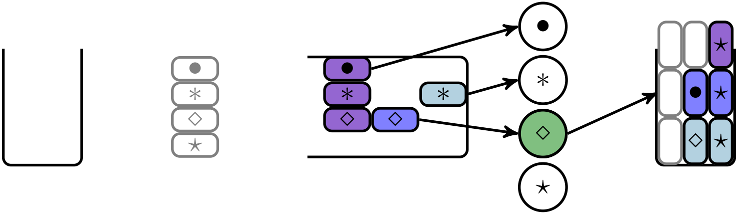


Eviction/Violation Lower Bound

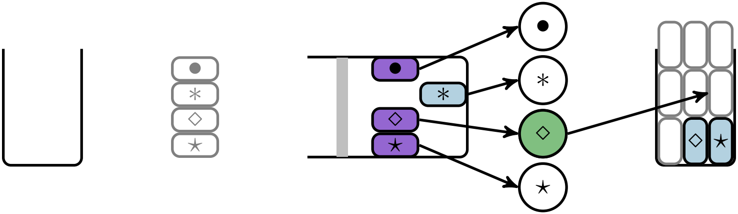


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

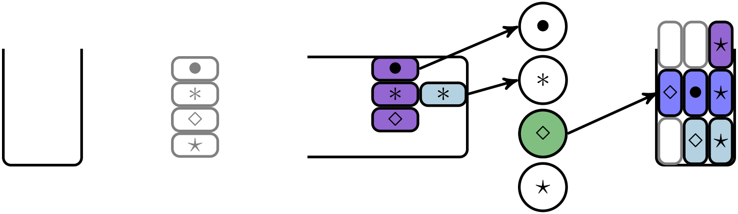


Eviction/Violation Lower Bound

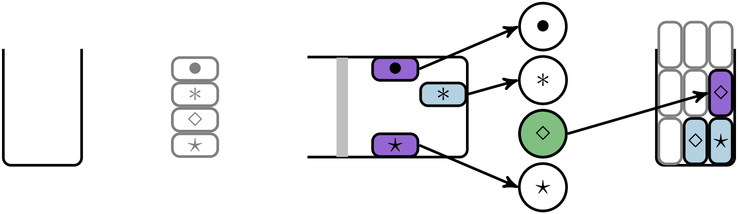


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

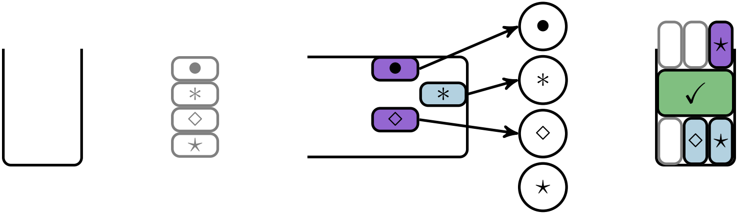


Eviction/Violation Lower Bound

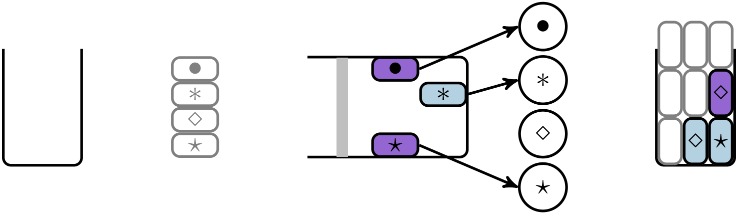


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

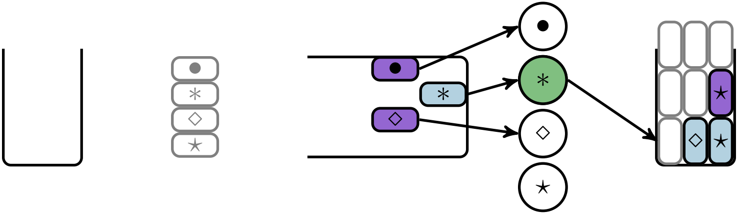


Eviction/Violation Lower Bound

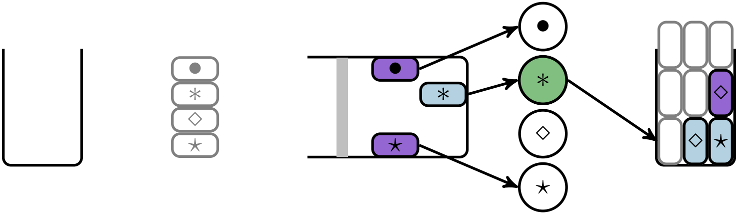


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

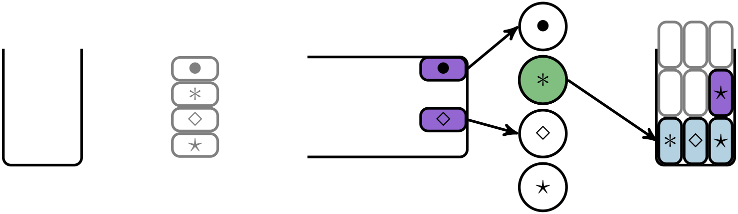


Eviction/Violation Lower Bound

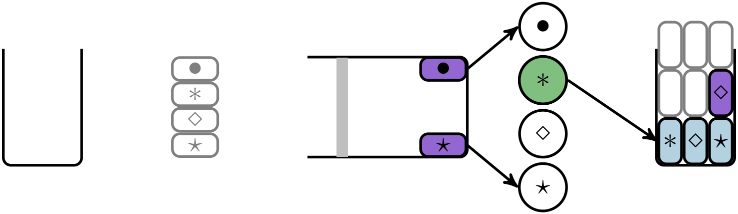


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

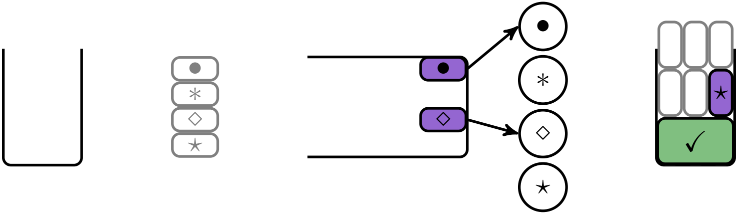


Eviction/Violation Lower Bound

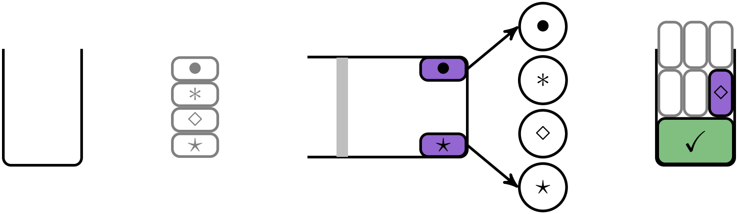


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

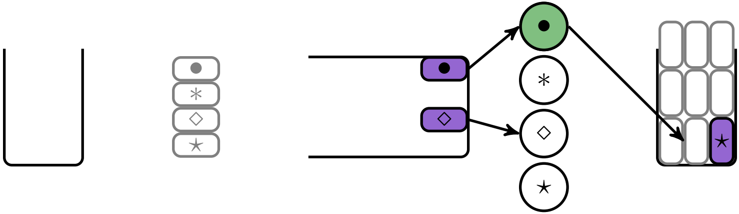


Eviction/Violation Lower Bound

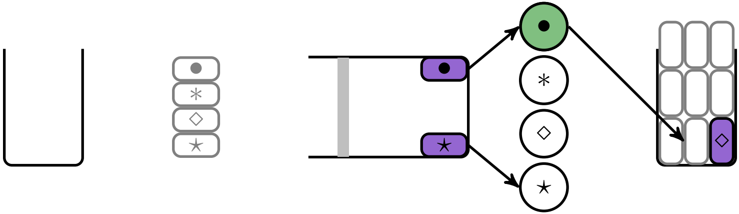


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

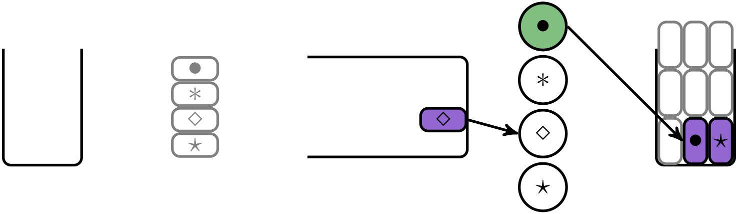


Eviction/Violation Lower Bound

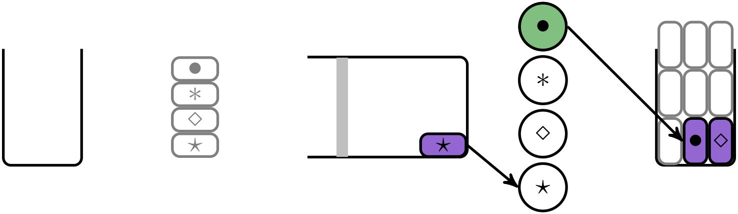


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage

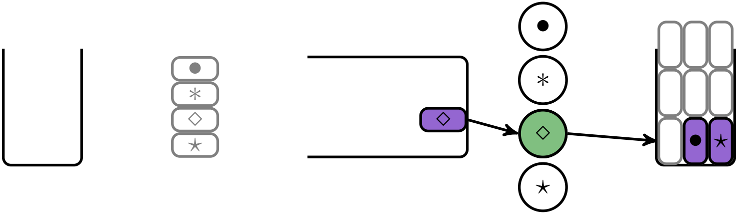


Eviction/Violation Lower Bound

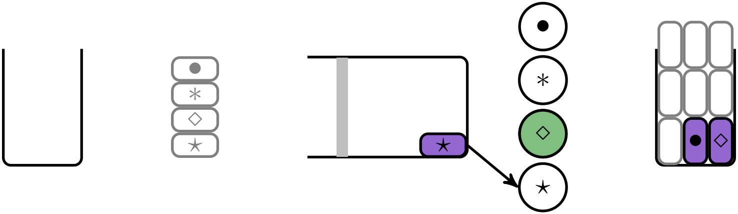


Sample Path Failure of Eviction/Violation Bound

Regular Distributed Coded Storage



Eviction/Violation Lower Bound



Beyond Sample-Path Dominance – System Model

File storage

- Media file partitioned into k pieces of equal size
- Data is encoded and stored on cloud servers

Arrivals Process

- Every request wants entire media file
- Poisson arrival process with rate λ

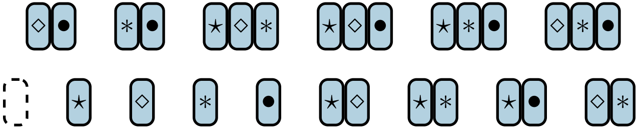
Completion Time

- Elapsed time form request to completion of service

Service Structure

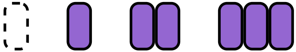
- Independence across servers
- Renewal process
- Exponentially service distribution
- Normalized rate

State Space Structure

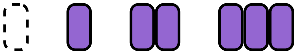


Keeping Track of Partially Fulfilled Requests

- State of partially fulfilled requests becomes large
- MDS coding and priority scheduling induce special structure: newer request have subset of older requests
- Leverage symmetry and focus on number of users with given number of pieces



State Space Collapse



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

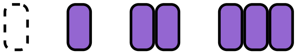
where $Y_i(t)$ is number of requests with i symbols

Results

- $\mathbf{Y}(t)$ is Markov
- Define $\phi_j(y) = \sum_{i=0}^{j} y_i$
- Define workload dominance (partial order)

$$y \leq_{\mathrm{w}} \tilde{y} \quad \text{iff} \quad \phi_j(y) \leq \phi_j(\tilde{y}) \quad \forall j$$

State Transitions of Collapsed System



Preservation of Workload Dominance

Workload dominance for two system states is **preserved** under coincident arrival of new requests, and concurrent delivery of data fragments at **a same level** in respective chains of useful servers

Expected Queue Lengths

For distributed storage with symmetric coding, fork-join queues, and FCFS service, **expected queue length** of QBD Violation-θ process $\mathrm{E}\left[\|\underline{Y}(t)\|_1\right]$ is less than or equal to expected queue length of original process $\mathrm{E}\left[\|Y(t)\|_1\right]$ at any $t \geq 0$

Summary and Discussion

Main Contributions

- Showcase that QBD-Violation, QBD-Eviction need not be sample path lower bounds
- Identify fundamental structure of coded storage systems under symmetric coding
- Introduce suitable partial order for system comparison
- Establish lower and upper bounds for expected queue lengths