Zurich Research Laboratory

Enhancing the Reliability of Large-Scale Data Storage Systems

Ilias Iliadis
April 26, 2018

Long-term Storage of Increasing Amount of Information

An increasing amount of information is required to be stored

- Web services
  - Email, photo sharing, web site archives
- Fixed-content repositories
  - Scientific data
  - Libraries
  - Movies
  - Music
- Regulatory compliance and legal issues
  - Sarbanes–Oxley Act of 2002 for financial services
  - Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the healthcare industry

Information needs to be stored for long periods and be retrieved reliably

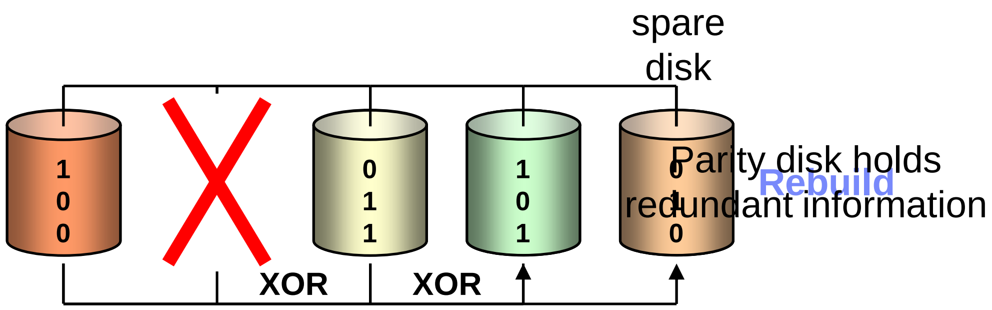
Storage

- Disk drives widely used as a storage medium in many systems
  - personal computers (desktops, laptops)
  - distributed file systems
  - database systems
  - high end storage arrays
  - archival systems
  - mobile devices
- Disks fail and need to be replaced
  - Mechanical errors
    - Wear and tear: it eventually leads to failure of moving parts
    - Drive motor can spin irregularly or fail completely
  - Electrical errors
    - A power spike or surge can damage in-drive circuits and hence lead to drive failure
  - Transport errors
    - The transport connecting the drive and host can also be problematic causing interconnection problems

Data Losses in Storage Systems

- Storage systems suffer from data losses due to
  - component failures
    - disk failures
    - node failures
  - media failures
    - unrecoverable and latent media errors

- Reliability enhanced by a large variety of redundancy and recovery schemes
  - RAID systems (**R**edundant **A**rray of **I**ndependent **D**isks)

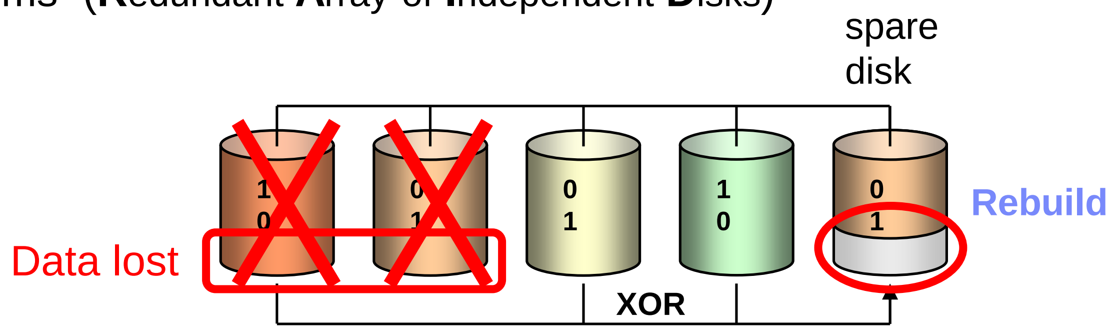


  - RAID-5: Tolerates one disk failure

Data Losses in Storage Systems

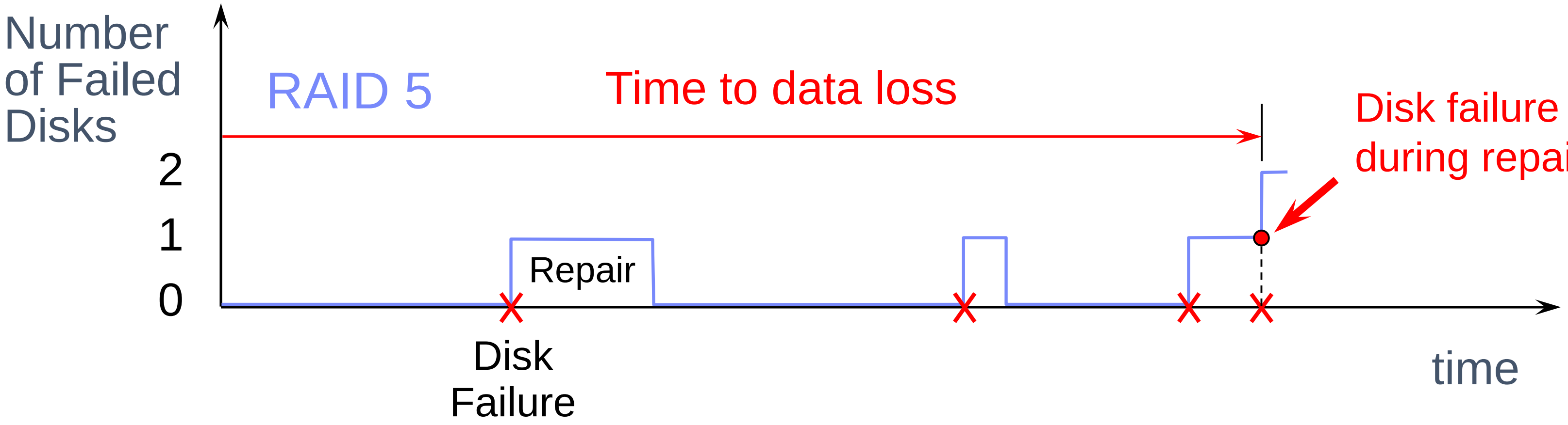
- Storage systems suffer from data losses due to
  - component failures
    - disk failures
    - node failures
  - media failures
    - unrecoverable and latent media errors

- Reliability enhanced by a large variety of redundancy and recovery schemes
  - RAID systems (**R**edundant **A**rray of **I**ndependent **D**isks)



  - RAID-5: Tolerates one disk failure
  - RAID-6: Tolerates two disk failures

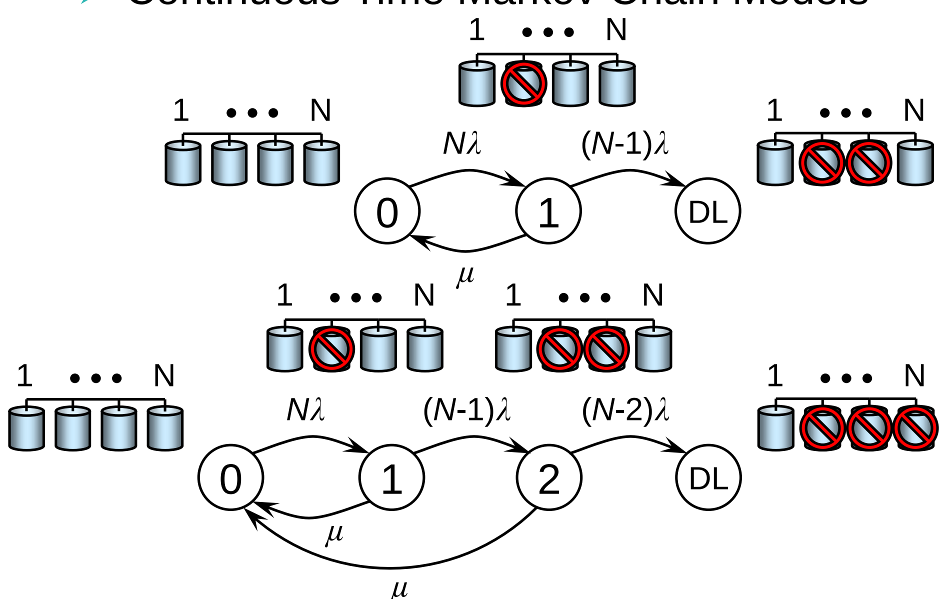
Time to Failure and MTTDL



– Reliability Metric: **MTTDL** (Mean Time to Data Loss)
  - Continuous Time Markov Chain Models

RAID 5:

RAID 6:



– λ : 1/ MTTF for disks
– μ : 1/ MTTR

$$\text{MTTDL} \simeq \frac{\mu}{N(N\text{-}1)\lambda^2}$$

[Patterson *et al*. 1988]

$$\text{MTTDL} \simeq \frac{\mu^2}{N(N\text{-}1)(N\text{-}2)\lambda^3}$$

[Chen *et al*. 1994]

original MTTDL equations

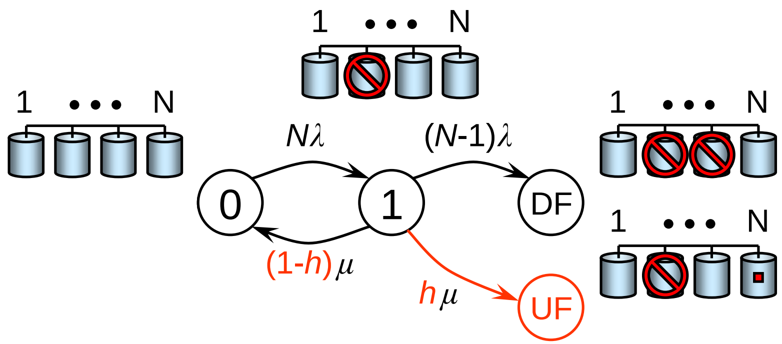
Markov Models for Unrecoverable Errors

- Parameters:
  - C_d : Disk capacity (in sectors)
  - P_s : P (unrecoverable sector error)
  - h : P (unrecoverable failure during rebuild in critical mode)
  - q : P (unrecoverable failure during RAID 6 rebuild in degraded mode)

$$h = 1 - [(1 - P_s)^{C_d}]^{(N-1)}$$

- Reliability Metric: MTTDL (Mean Time To Data Loss for the array)

RAID 5:

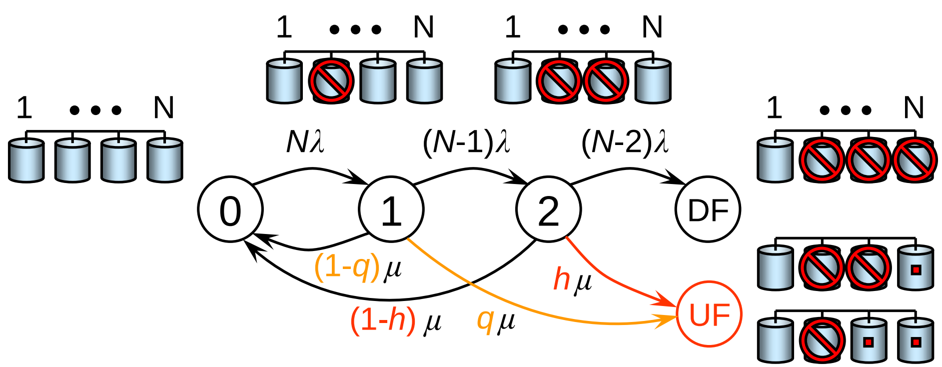


Data loss owing to:

- DF: Disk Failure
- UF: Unrecoverable Failure

$$\text{MTTDL} = \frac{(2N-1)\lambda + \mu}{N\lambda[(N-1)\lambda + \mu h]}$$

RAID 6:



$$h = (N-2) C_d P_s + O(P_s^2)$$

$$q = \binom{N-1}{2} C_d P_s^2 + O(P_s^3)$$

$q \ll h$ for $P_s \ll$

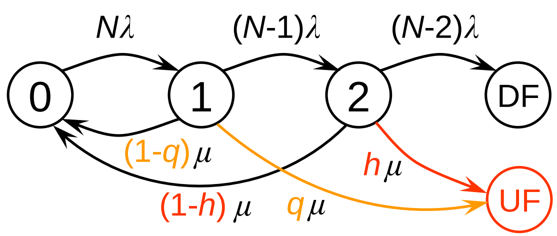
MTTDL for RAID 5 and RAID 6

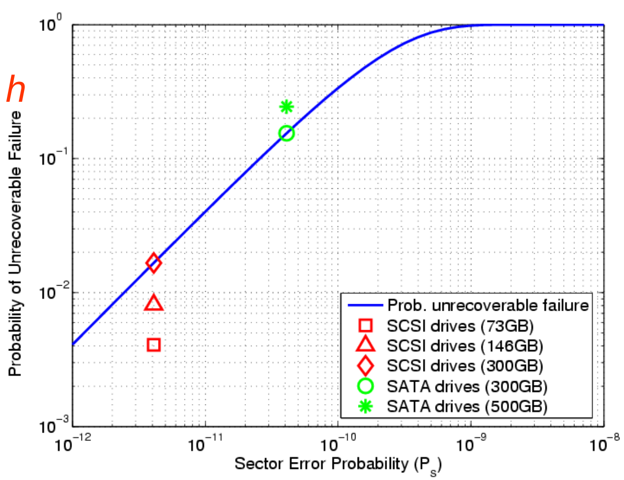
Assumptions:

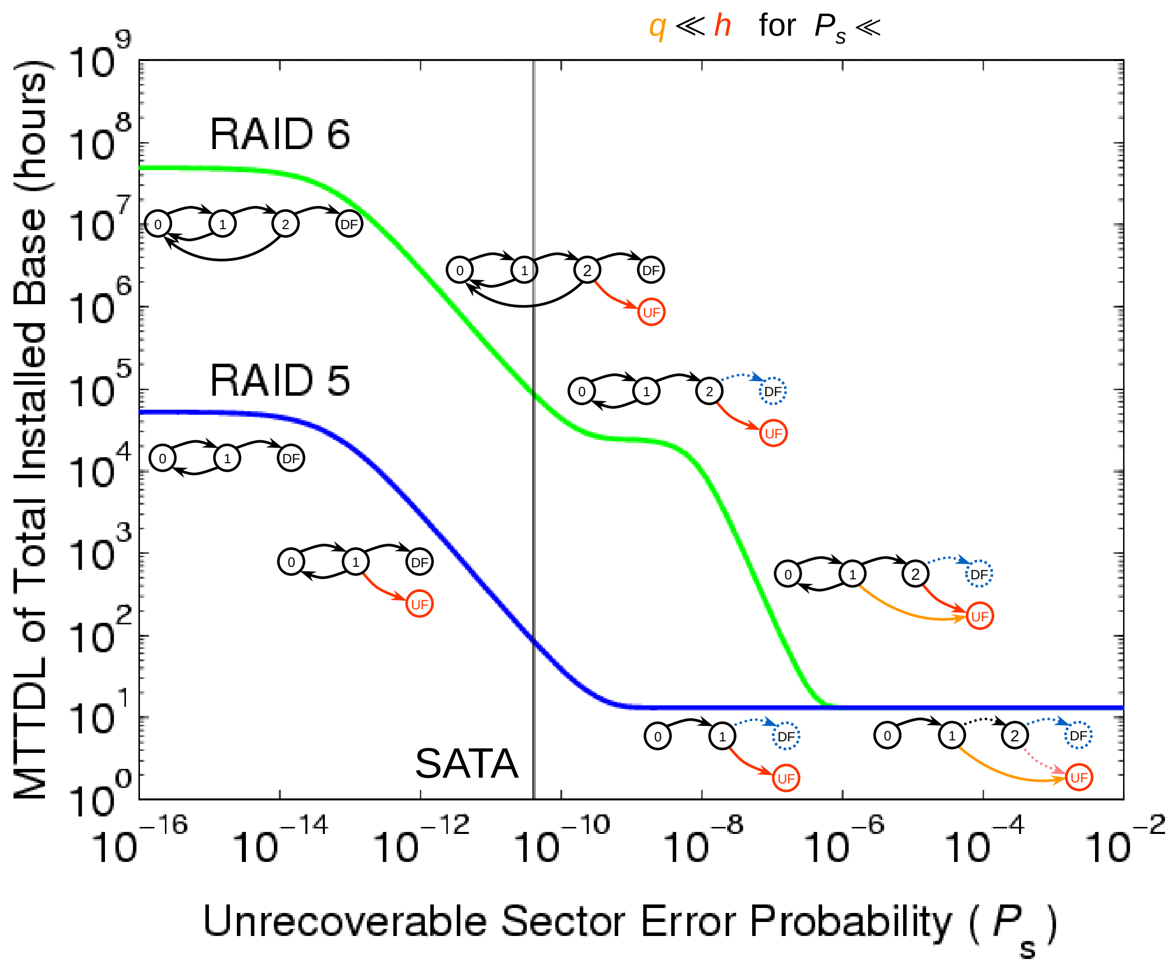
- UD : 10 PB = 10^{15} bytes user data base
- C_d : 300 GB SATA disk drive capacity
- N : 8 disks per array group for RAID 5
  16 disks per array group for RAID 6
- N_{total} : 38096 disks: 4762 arrays for RAID 5
  2381 arrays for RAID 6
- $MTTF_d$: 500 000 hours for a SATA disk
- $MTTR_d$: 17.8 hours expected repair time
- P_b : P(unrecoverable bit error) = 10^{-14} for SATA
  $\Rightarrow P_s = 4096 \times 10^{-14} = 4.096 \times 10^{-11}$

h : P (unrecoverable failure during rebuild in the critical mode)

q : P (unrecoverable failure during RAID 6 rebuild in the degraded mode)





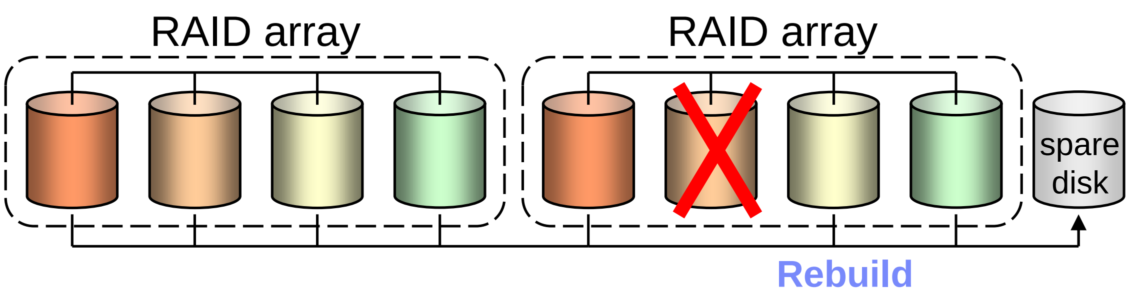


Reliability of Large-Scale Storage Systems

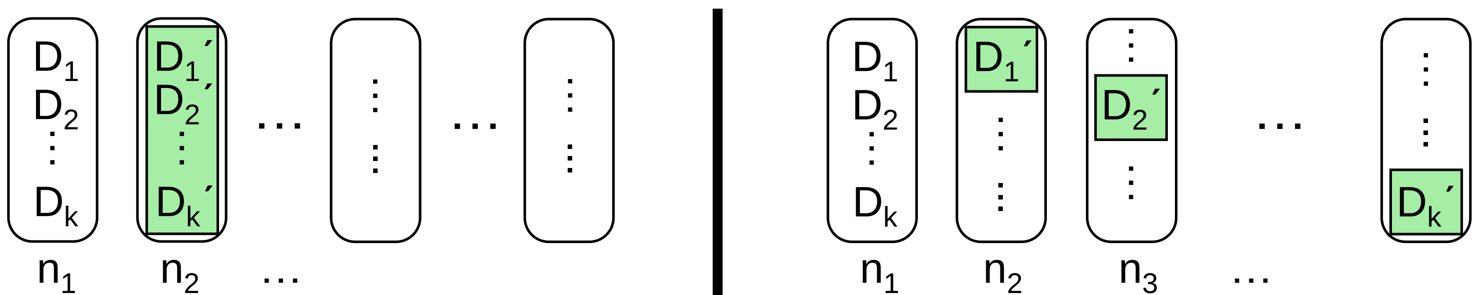
- Storage systems have become large
  - Petabytes of data in 1000s of disks in 100s of nodes
  - Device failures are daily events
- Replication is widely used to store redundant data to protect system from data loss
  - IBM XIV
  - Google File System
- Various factors affect reliability
  - Placement of replicas
    - Clustered replication vs. Distributed replication
  - Rebuild strategy / rebuild times
- Assessing system reliability is
  - Essential
  - Not trivial; RAID reliability results not applicable
- Developed enhanced models and obtained reliability expressions
  - r-way replication

Distributed Storage Systems

- **Markov models**
  - Times to disk failures and rebuild durations exponentially distributed (-)
  - MTTDL has been proven to be a useful metric for (+)
    - estimating the effect of the various parameters on system reliability
    - comparing schemes and assessing tradeoffs



Reduce vulnerability window
- Distributing data
- Distributed rebuild method



- replicated data on the same node

Clustered Placement

- replicated data on different nodes

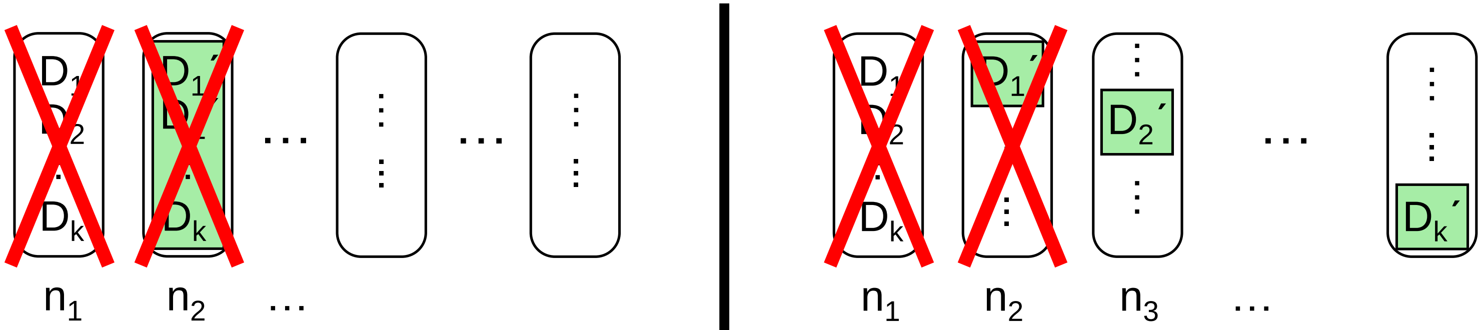
Declustered Placement

- **Non-Markov-based analysis**
  - V. Venkatesan et al. “Reliability of Clustered vs. Declustered Replica Placement in Data Storage Systems”, MASCOTS 2011
  - V. Venkatesan et al. “A General Reliability Model for Data Storage Systems”, QEST 2012

General non-exponential failure and rebuild time distributions
- MTTDL is insensitive to the failure time distributions; it depends only on the mean value

Time To Data Loss vs. Amount of Data Lost

- MTTDL measures time to data loss
  - no indication about amount of data loss
    - ➢ Consider the following example
      - Replicated data for D_1, D_2, …, D_k is placed:



- on the same node

  Clustered Placement

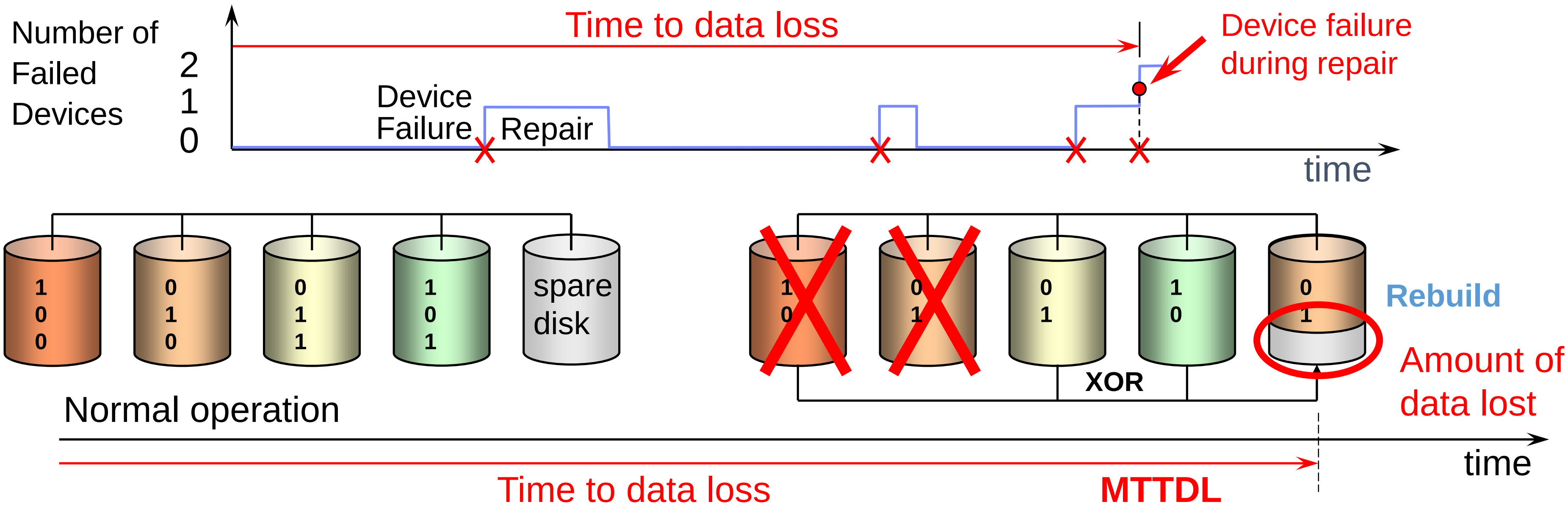
- on different nodes

  Declustered Placement

- Distinguish between data loss events involving
  - high amounts of data lost
  - low amounts of data lost

➢ Need for a measure that quantifies the amount of data lost

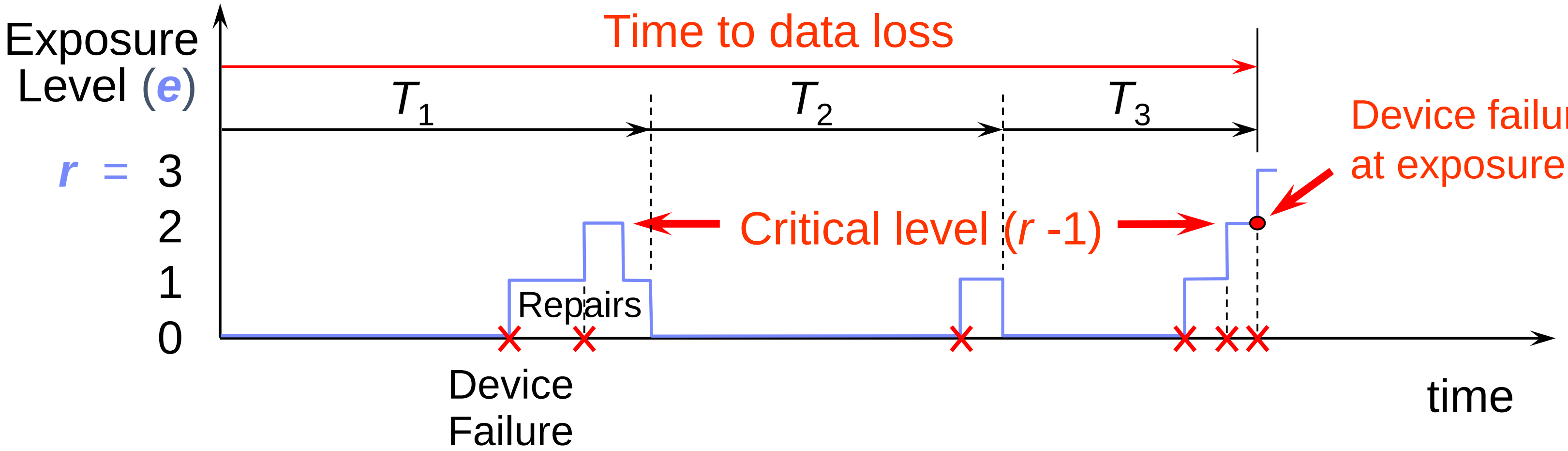
Reliability Metrics – MTTDL and EAFDL



- Data loss events documented in practice by Yahoo!, LinkedIn, Facebook and Amazon
  - Amazon S3 (Simple Storage Service) is designed to provide 99.999999999% durability of objects over a given year
    - average annual expected loss of a fraction of 10^{-11} of the data stored in the system
- Assess the implications of system design choices on the
  - frequency of data loss events
    - **Mean Time to Data Loss (MTTDL)**
  - amount of data lost
    - **Expected Annual Fraction of Data Loss (EAFDL)**

      I. Iliadis and V. Venkatesan,
      "Expected Annual Fraction of Data Loss as a Metric for Data Storage Reliability", MASCOTS 2014
  - These two metrics provide a useful profile of the magnitude and frequency of data losses

Non-Markov Analysis for EAFDL and MTTDL



- EAFDL evaluated in parallel with MTTDL
  - *r* : Replication Factor
  - *e* : Exposure Level: maximum number of copies that any data has lost
  - T_i : Cycles (Fully Operational Periods / Repair Periods)
  - P_{DL} : Probability of data loss during repair period
  - *U* : Amount of user data in system
  - *Q* : Amount of data lost upon a first-device failure

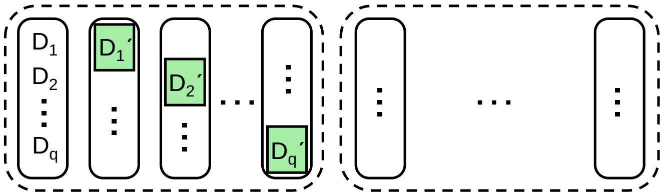
➢ MTTDL $\approx \sum_{i=1}^{m} E(T_i) \approx \dfrac{E(T)}{P_{\mathrm{DL}}}$ EAFDL = $\dfrac{E(Q)}{E(T) \cdot U}$

MTTDL and EAFDL expressions obtained using non-Markov Analysis

Theoretical Results

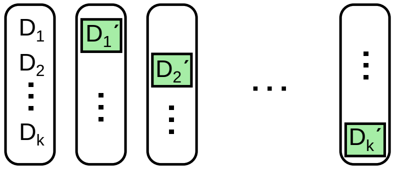
- *n* : number of storage devices — 4 to 64
- *c* : amount of data stored on each device — 12 TB
- *r* : replication factor — 2, 3, 4
- *b* : reserved rebuild bandwidth per device — 96 MB/s
- *1/λ* : mean time to failure of a storage device — 10,000 h - Weibull distributions with shape parameters greater than one
  - ➤ increasing failure rates over time
    - shape parameter = 1.5

$$\text{MTTDL} \approx \begin{cases} \left(\dfrac{b}{\lambda c}\right)^{r-1} \dfrac{1}{n\lambda}, \\ \left(\dfrac{b}{2\lambda c}\right)^{r-1} \dfrac{(r-1)!}{n\lambda} \displaystyle\prod_{e=1}^{r-2} \left(\dfrac{n-e}{r-e}\right)^{r-e-1} \end{cases}$$

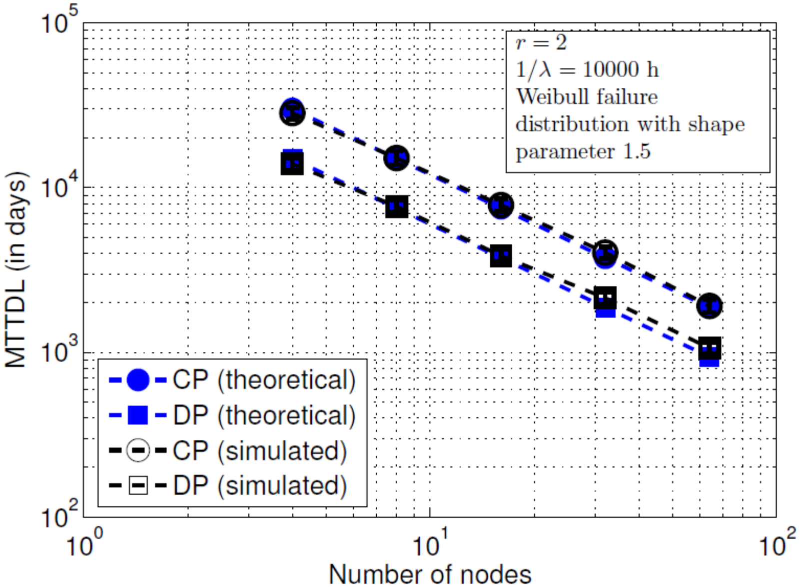


Symmetric placement

$$\text{EAFDL} \approx \begin{cases} \left(\dfrac{\lambda c}{b}\right)^{r-1} \lambda, \\ \left(\dfrac{2\lambda c}{b}\right)^{r-1} \dfrac{\lambda}{(r-1)!} \displaystyle\prod_{e=1}^{r-1} \left(\dfrac{r-e}{n-e}\right)^{r-e}, \quad \text{for DP} \end{cases}$$



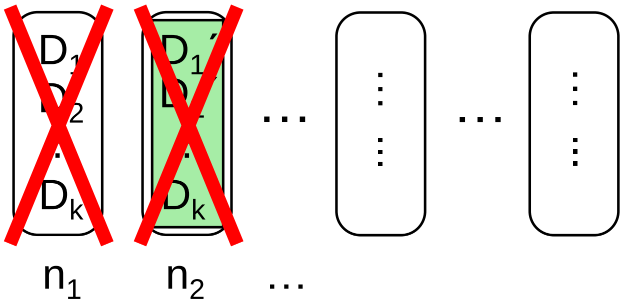
Reliability Results for Replication Factor of 2



- MTTDL
  - Declustered placement is not better than clustered one

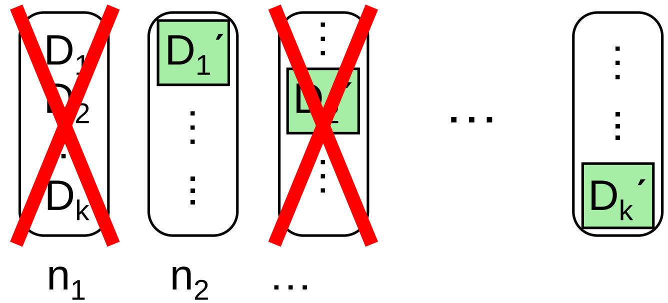
Distributed Storage Systems

Replicated data for D_1, D_2, …, D_k is placed:



- on the same node

Clustered Placement

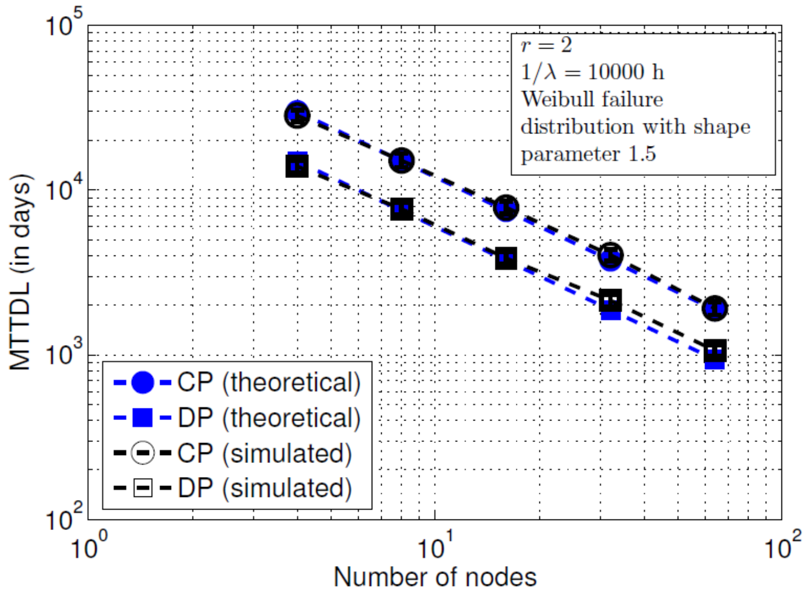


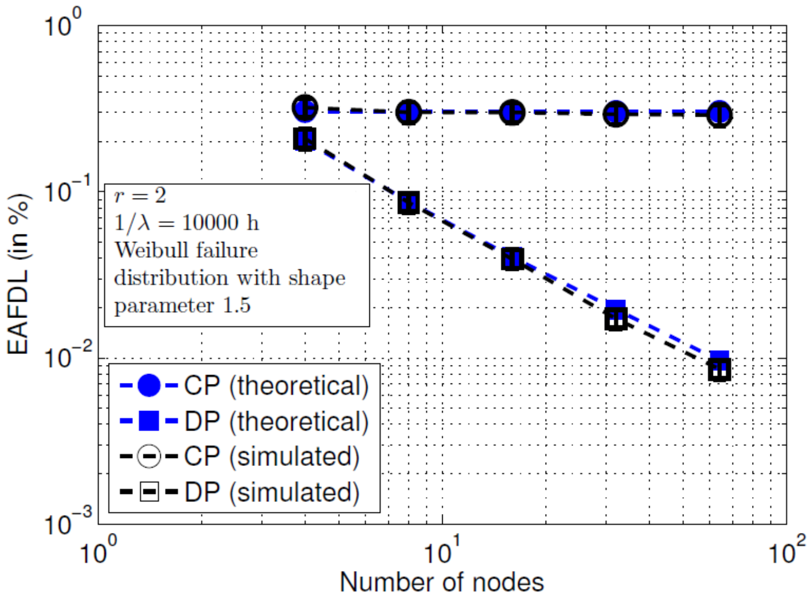
- on different nodes

Declustered Placement

- MTTDL
  - Reduced repair time (+)
    - Reduced vulnerability window
  - Increased exposure to subsequent device failures (-)
- EAFDL
  - Reduced amount of data lost (+)

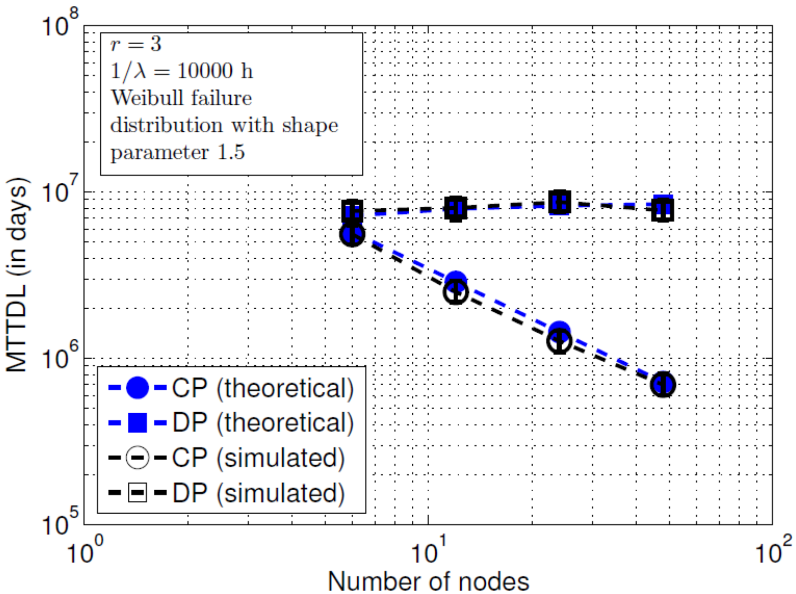
Reliability Results for Replication Factor of 2

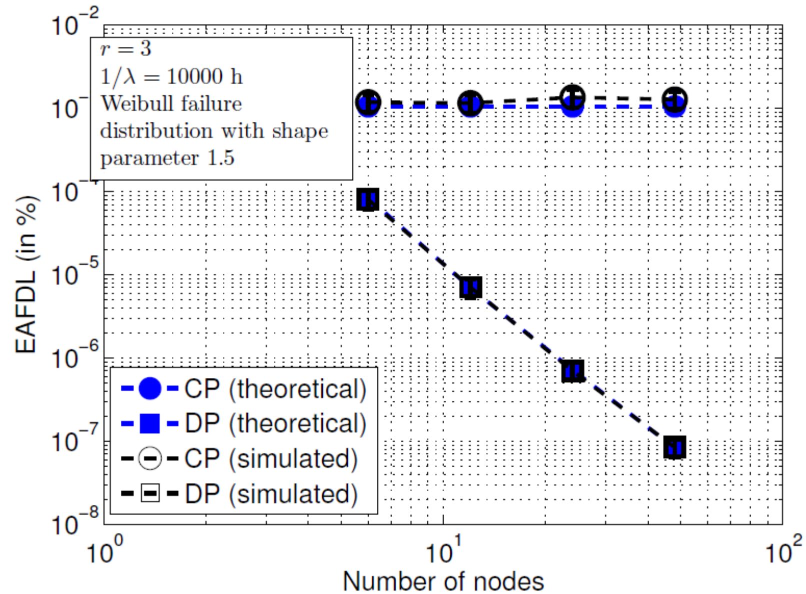




- MTTDL
  - Declustered placement not better than clustered one
- EAFDL
  - Independent of the number of nodes for clustered placement
  - Inversely proportional to the number of nodes for declustered placement
    - Declustered placement better than clustered one

Reliability Results for Replication Factor of 3





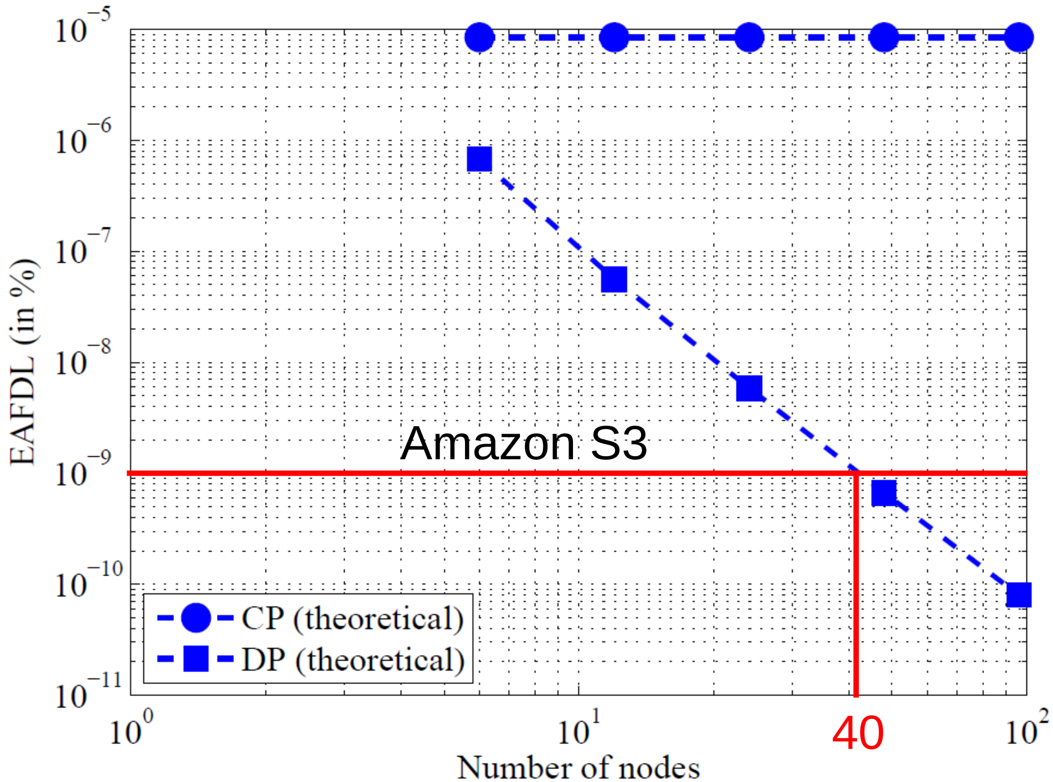
- MTTDL
  - Inversely proportional to the number of nodes for clustered placement
  - Independent of the number of nodes for declustered placement
    - Declustered placement better than clustered one
- EAFDL
  - Independent of the number of nodes for clustered placement
  - Inversely proportional to the cube of the number of nodes for declustered placement
    - Declustered placement better than clustered one

Theoretical EAFDL Results for Replication Factor of 3

c = 12 TB
b = 96 MB/s
MTTR = 35 h

MTTF = $1/\lambda$ = 50,000 h

MTTR / MTTF = 0.0007



- Theoretical results are accurate when devices are very reliable
  - MTTR/MTTF ratio is small
    - Quick assessment of EAFDL
    - No need to run lengthy simulations

Discussion

- EAFDL should be used cautiously
  - suppose EAFDL = 0.1%
  - this does not necessarily imply that 0.1% of the user data is lost each year
    - System 1: MTTDL=10 years 1% of the data lost upon loss
    - System 2: MTTDL=100 years 10% of the data lost upon loss
  - The desired reliability profile of a system depends on the
    - application
    - underlying service
  - If the requirement is that data losses should not exceed 1% in a loss event
    - only <System 1> could satisfy this requirement

Reliability of Cloud Storage Systems

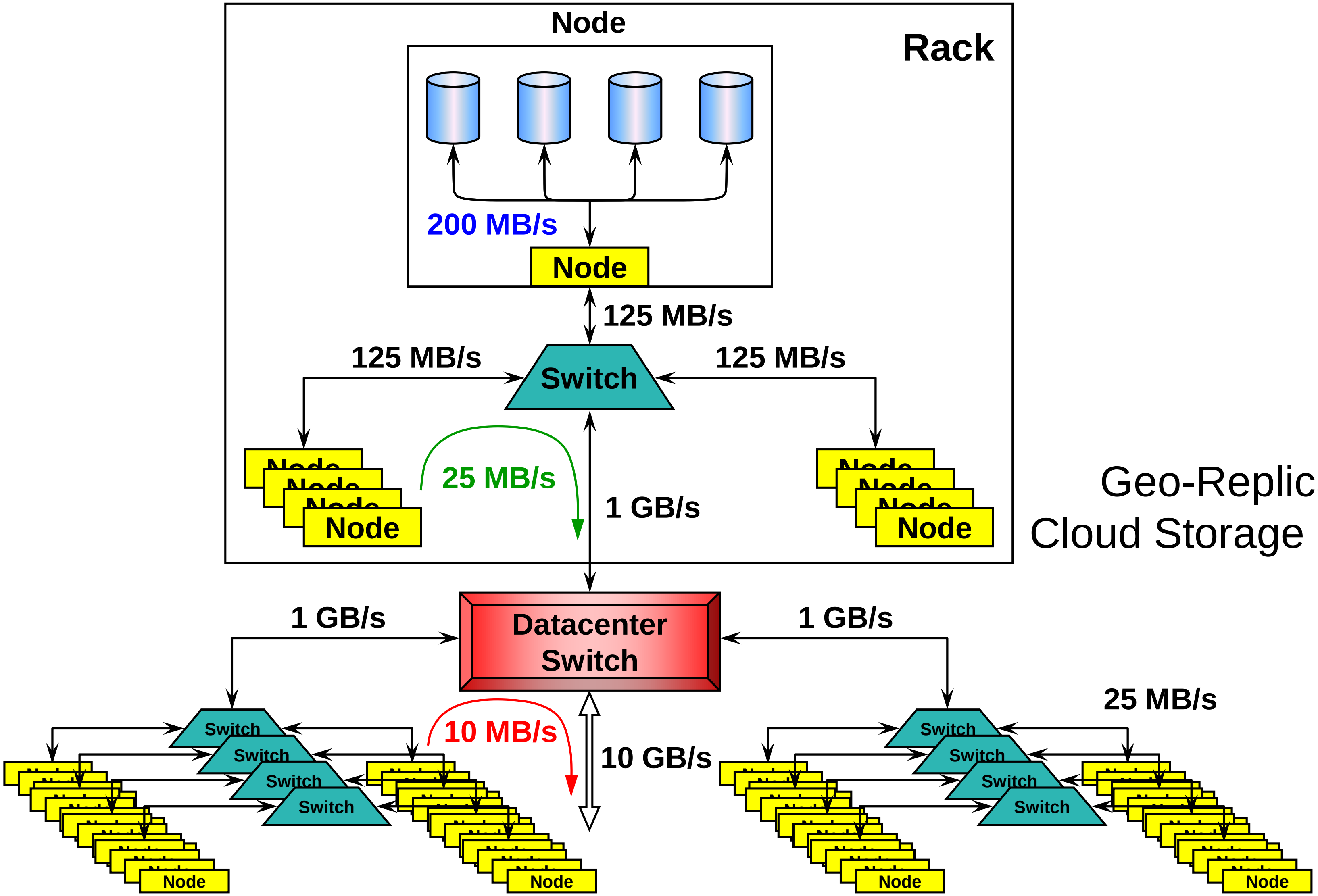
- Today's cloud storage systems are large
  - Exabytes of data stored in 1000s of storage components in 100s of data centers

- State-of-the-art data storage systems employ general erasure codes that affect
  - Reliability
  - Performance
  - Storage overhead
  - Reconstruction overhead of the system

- Various factors affect reliability
  - Placement of redundant data
  - Rebuild strategy / rebuild times
  - Spare space provided within each disk drive for rebuild
  - Component availability / failure
    - Hardware, disk drives, nodes, racks, clusters, data centers, networks

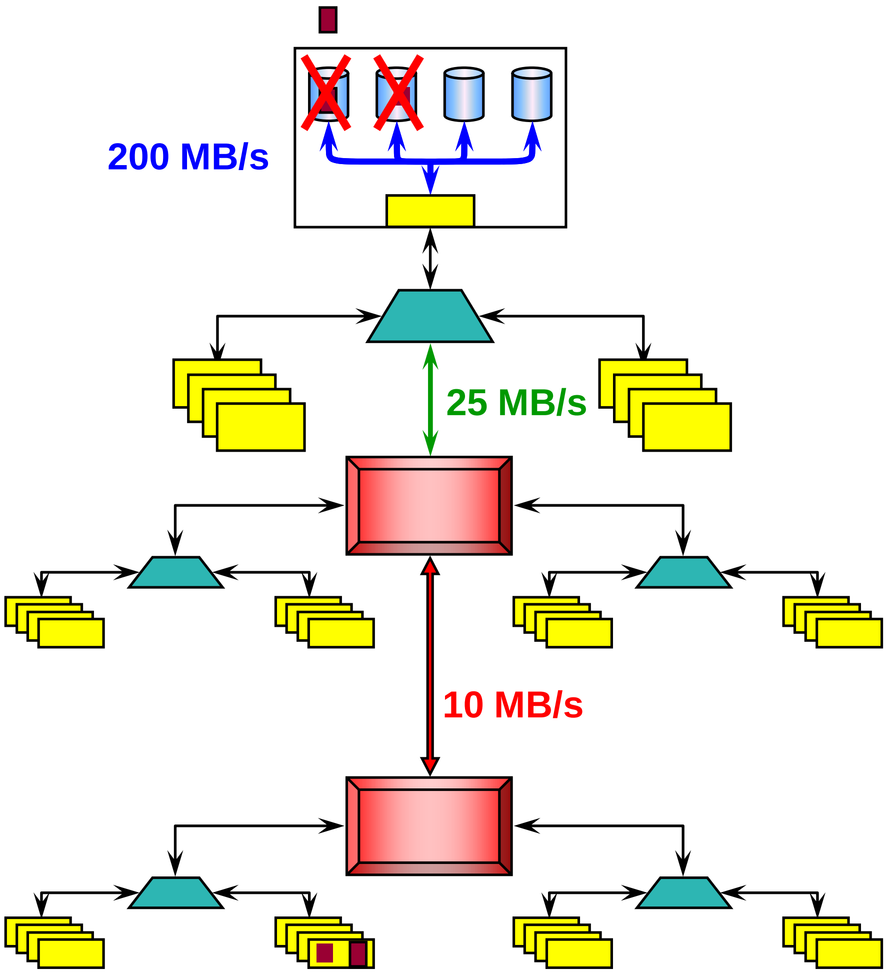
- Developed enhanced models and obtained reliability expressions
  - Disk/Node/Server failures
  - r-way replication
  - Erasure codes

Storage Hierarchy of a Data Center



Geo-Replicated Cloud Storage Systems

Reliability Issues in Geo-Replicated Cloud Storage Systems



Reliability improvement through data replication

- Replica placement
  - Within the same node
    - Fast rebuild at 200 MB/s (+)
    - Exposure due to disk failure correlation (-)
  - Across datacenters
    - No exposure due to correlated failures (+)
- Rebuild process
  - Direct rebuild to the affected node
    - Slow rebuild at 10 MB/s
      - Long vulnerability window (-)
  - Staged rebuild
    - First local rebuild
      - Fast rebuild at 200 MB/s
        - ✓ Short vulnerability window (+)
      - Same location
        - ✓ Exposure due to correlated failures (0)
    - Replica then migrated to the affected node
- Replication factor
  - How many replicas are required?

Tradeoffs among various placement and rebuild schemes

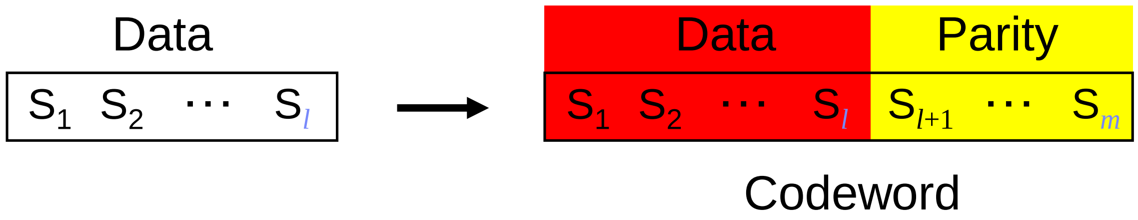
Geo-Replicated Cloud Storage Systems

I. Iliadis, et al., “Reliability of Geo-replicated Cloud Storage Systems”, *PRDC* 2014

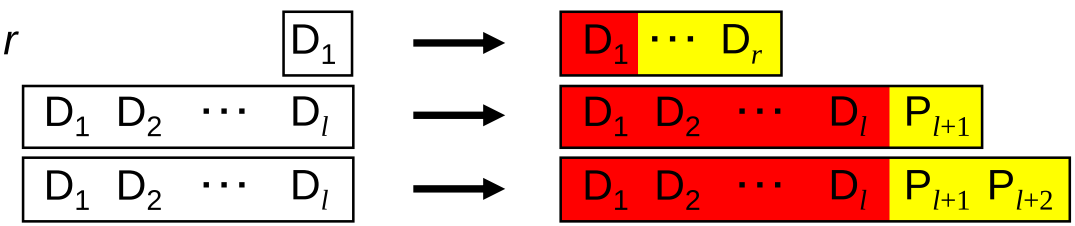
- First work to study the reliability of geo-replicated cloud storage systems under four different rebuild schemes: Direct, Direct+, Staged, and Staged+
- Closed-form expressions for the MTTDL were obtained and validated using simulations
  - In the absence of sector errors, staged rebuild was found to improve the MTTDL by one to three orders of magnitude
  - In the presence of sector errors, the improvement offered by staged rebuild is at most of one order of magnitude
  - Relative differences in reliability of the schemes considered are primarily influenced by the inter-, intra-site, and disk rebuild bandwidths
    - the one that is a bottleneck in the rebuild process determines the system reliability

Erasure Coded Schemes

- User data divided into blocks (symbols) of fixed size
  - Complemented with parity symbols
    - codewords
- (m,l) maximum distance separable (MDS) erasure codes



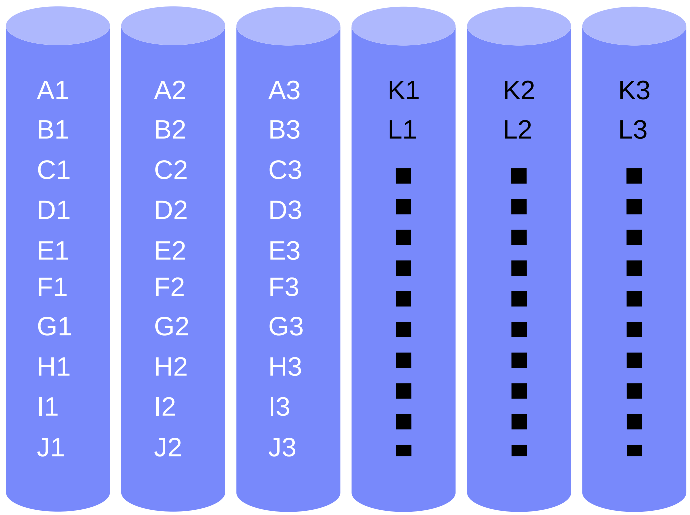
- Any subset of l symbols can be used to reconstruct the codeword
  - Replication : $l = 1$ and $m = r$
  - RAID-5 : $m = l + 1$
  - RAID-6 : $m = l + 2$



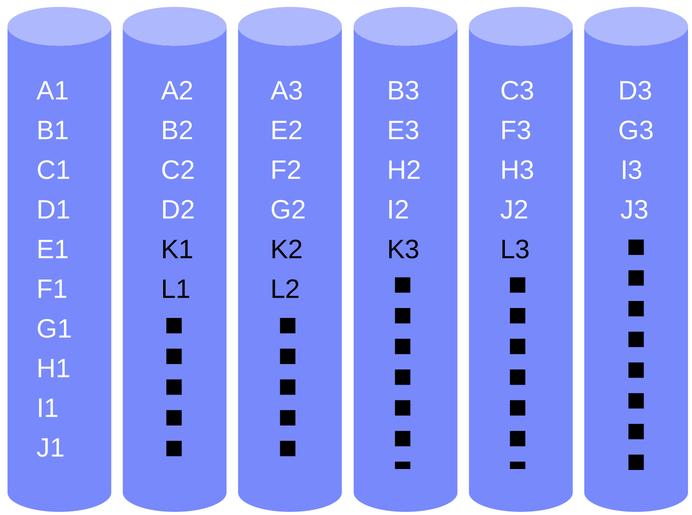
- Storage efficiency : $s_{eff} = l/m$
- Facebook : Reed-Solomon (14,10) → s_{eff} = 71 %
- Windows Azure : Reed-Solomon (16,12) → s_{eff} = 75 %

Redundancy Placement

Erasure code with codeword length 3

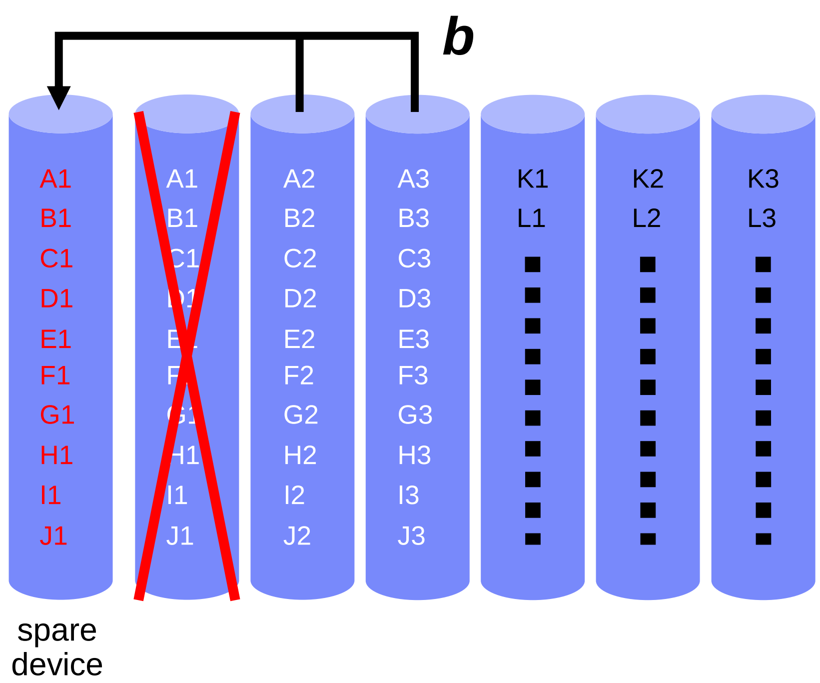


Clustered Placement

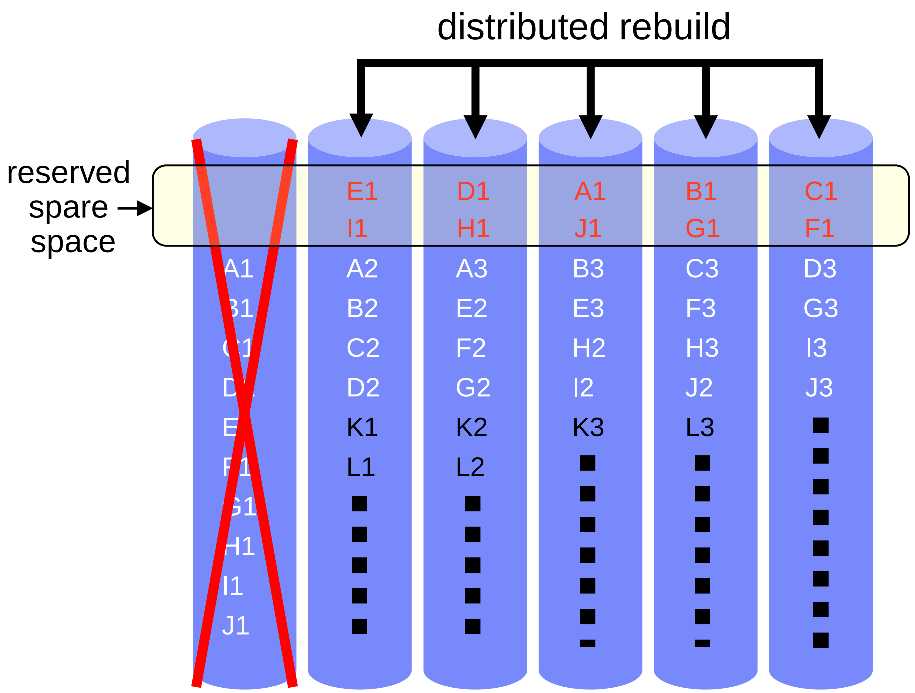


Declustered Placement

Device Failure and Rebuild Process

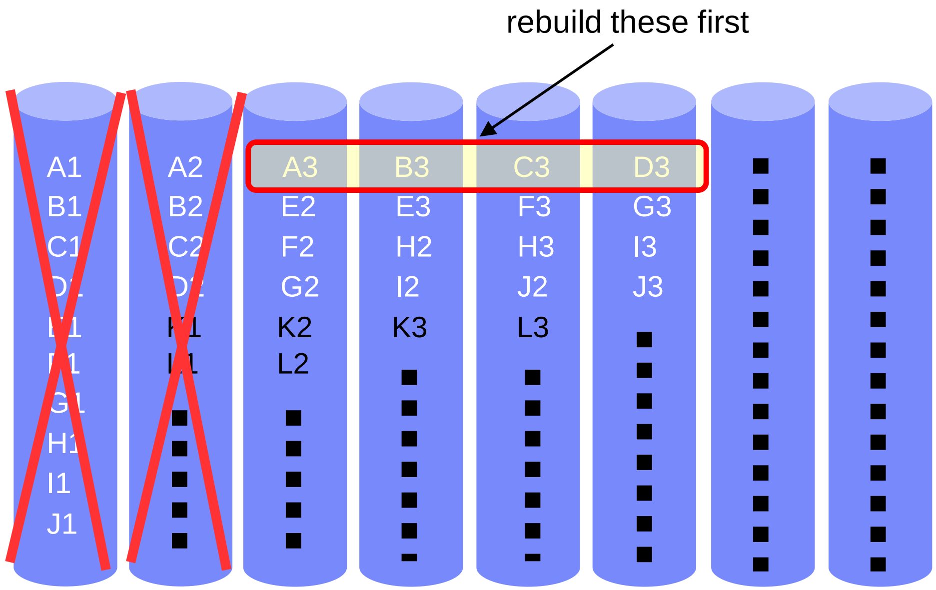


Clustered Placement



Declustered Placement

Rebuild Model



- Prioritized rebuilds
  - first rebuild the most-exposed data
    - data with the least number of surviving codeword symbols

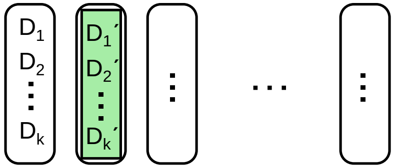
- For placement schemes that spread codeword symbols across many devices, e.g., declustered,
  - the amount of most-exposed data decreases combinatorially fast with each additional device failure
  - prioritizing the rebuilds of the most-exposed data
    - reduces the exposure time for this data
    - results in a substantial improvement of reliability

Reliability of Erasure Coded Systems

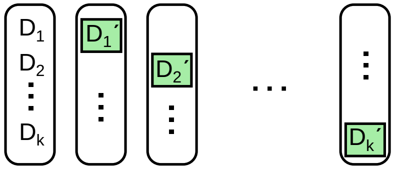
I. Iliadis and V. Venkatesan, “Reliability Assessment of Erasure Coded Systems”, CTRQ 2017

- *n* : number of storage devices
- *c* : amount of data stored on each device
- *(m,l)* : MDS erasure code
- *b* : reserved rebuild bandwidth per device
- $1/\lambda$: mean time to failure of a storage device

$$\text{MTTDL} \approx \begin{cases} \dfrac{1}{n\lambda}\left(\dfrac{b}{\lambda c}\right)^{m-l} \dfrac{1}{\binom{m-1}{l-1}}, & \text{for CP} \\ \dfrac{1}{n\lambda}\left[\dfrac{b}{(l+1)\lambda c}\right]^{m-l} (m-l)! \displaystyle\prod_{e=1}^{m-l}\left(\dfrac{n-e}{m-e}\right)^{m-l-e}, & \text{for DP} \end{cases}$$



$$\text{EAFDL} \approx \begin{cases} \lambda\left(\dfrac{\lambda c}{b}\right)^{m-l} \dbinom{m}{l-1}, & \text{for CP} \\ \left[\dfrac{(l+1)\lambda c}{b}\right]^{m-l} \dfrac{\lambda m}{(m-l+1)!} \displaystyle\prod_{e=1}^{m-l}\left(\dfrac{m-e}{n-e}\right)^{m-l+1-e}, & \text{for DP} \end{cases}$$

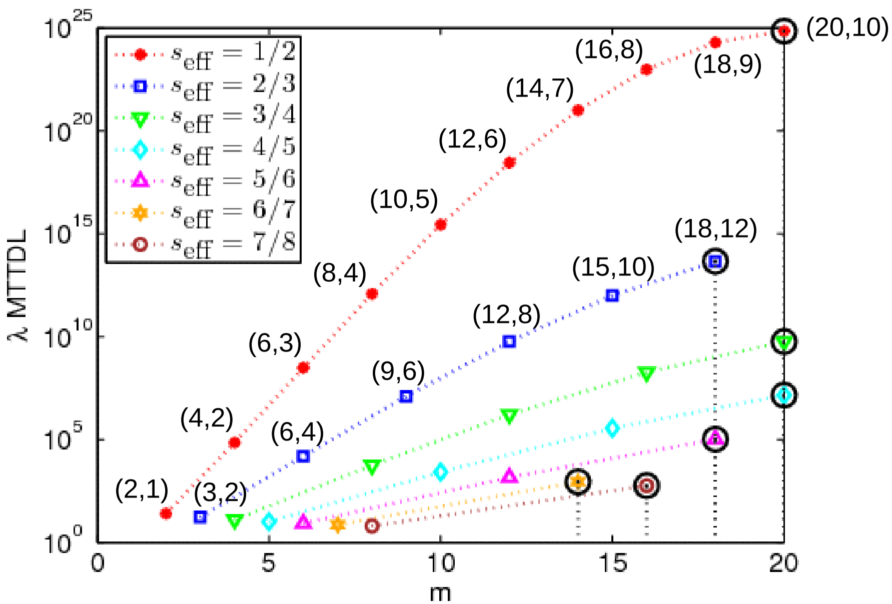


Reliability Comparison

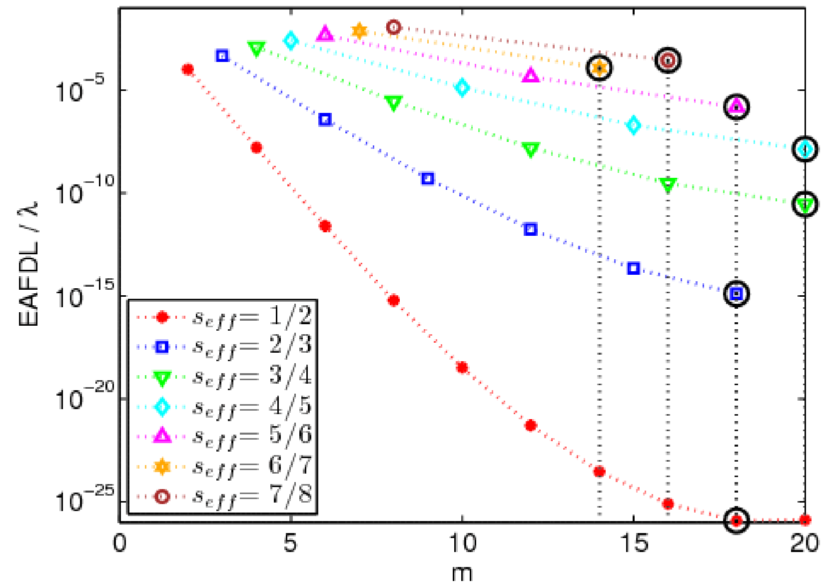
- Reliability of declustered placement under
  - fixed amount of user data, ***U***
  - fixed storage efficiency, $s_{eff} = l\,/\,m$
  - various codeword lengths, ***m***

- n : Number of storage devices
- $1/\lambda$: Mean Time to Failure (MTTF) for a device
- $1/\mu$: Time to read the data of a device

n = 20



$\lambda/\mu = \lambda c/b = 0.001$



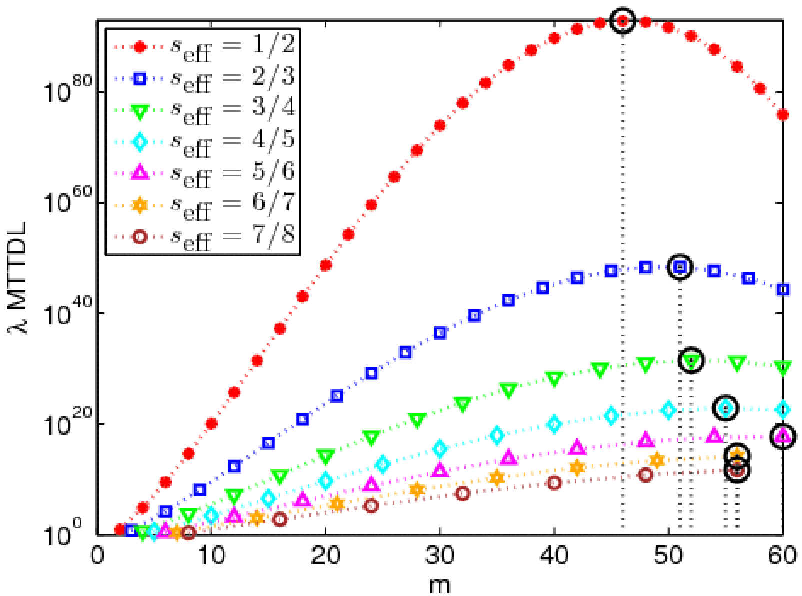
- For fixed storage efficiency s_{eff}
  - ➢ Reliability maximized for maximum codeword length ***m***
    - Large codewords can tolerate more device failures

Reliability Comparison

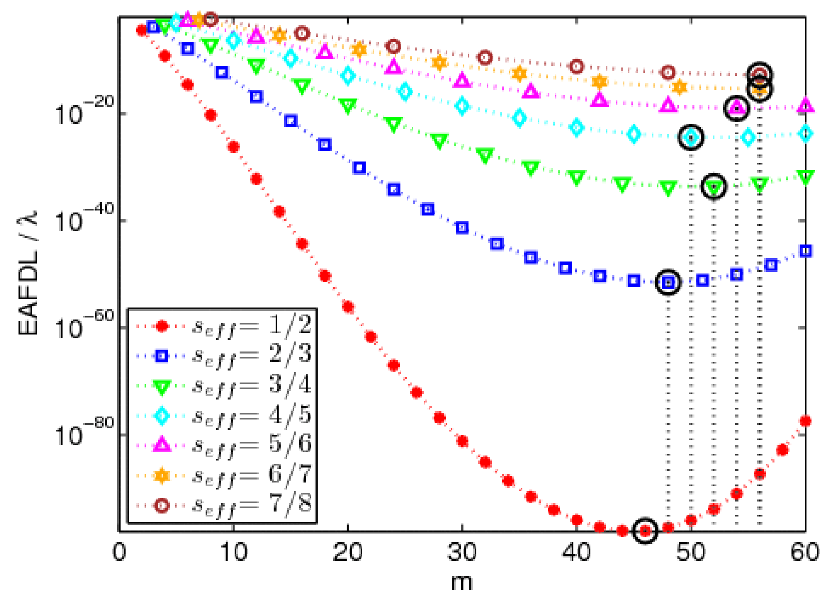
- Reliability of declustered placement under
  - fixed amount of user data, ***U***
  - fixed storage efficiency, $s_{eff} = l / m$
  - various codeword lengths, ***m***

- ***n*** : Number of storage devices
- $1/\lambda$: Mean Time to Failure (MTTF) for a device
- $1/\mu$: Time to read the data of a device

n = 60

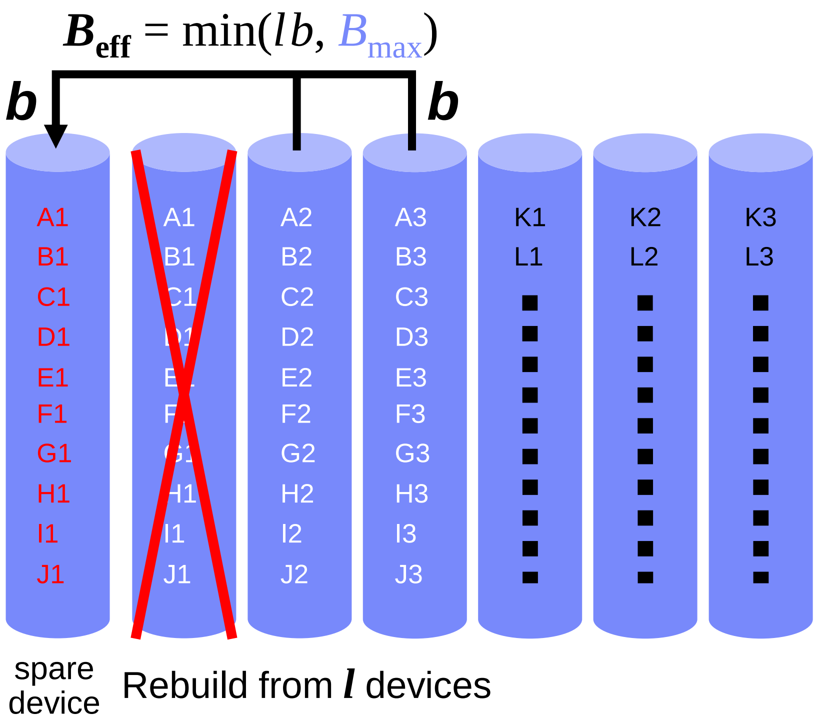


$\lambda/\mu = \lambda c/b = 0.001$

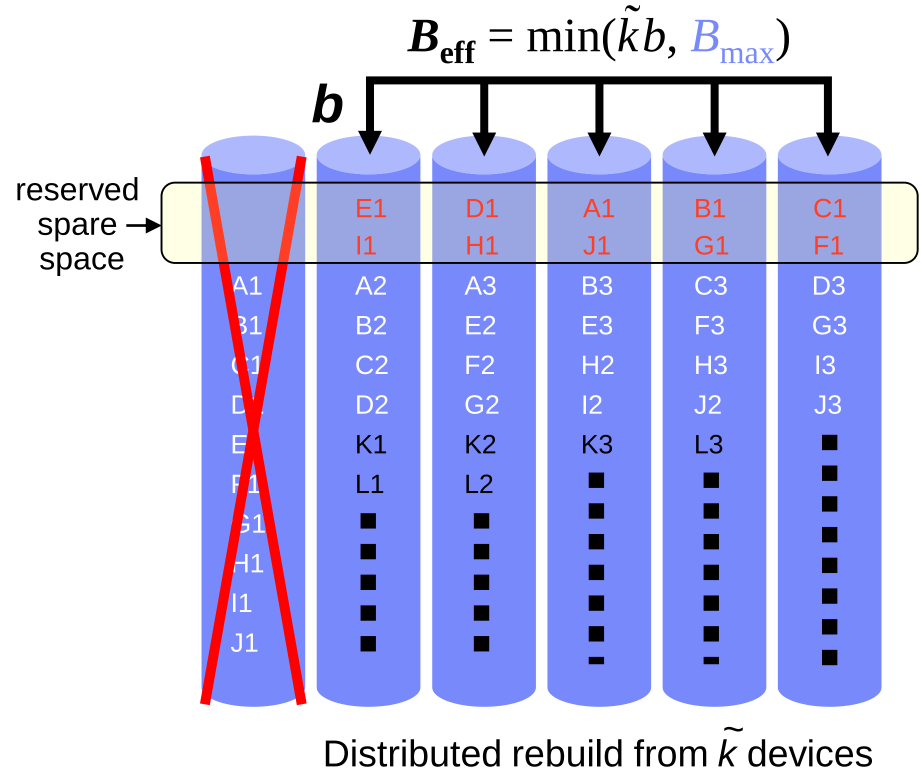


- For fixed storage efficiency s_{eff}
  - ➢ Reliability not maximized for maximum codeword length ***m***
    - Large codewords can tolerate more device failures
    - Large codewords spread across a larger # of devices - higher exposure degree to failure

Network Rebuild Bandwidth Constraints



Clustered Placement



Declustered Placement

Summary

- Considered the Mean Time to Data Loss (MTTDL) and the Expected Annual Fraction of Data Loss (EAFDL) reliability metrics
- Presented a methodology for assessing the two metrics analytically
  - Non-Markov analysis
    - large class of failure time distributions
      - real-world distributions, such as Weibull and gamma
- Derived closed-form expressions of MTTDL and EAFDL for various redundancy schemes
  - RAID-5, RAID-6, replication, erasure coding

  and for various placements schemes
  - Clustered
  - Declustered
    - Prioritized rebuilds
- Demonstrated the superiority of the declustered placement scheme
- Addressed reliability issues in Geo-Replicated Cloud Storage Systems

Future Work

- Reliability of erasure coded systems under bandwidth constraints
  - for arbitrary rebuild time distributions
  - in the presence of unrecoverable latent errors