Zurich Research Laboratory

Effect of Lazy Rebuild on Reliability of Erasure-Coded Storage Systems

Ilias Iliadis
ili@zurich.ibm.com
April 24-28, 2022

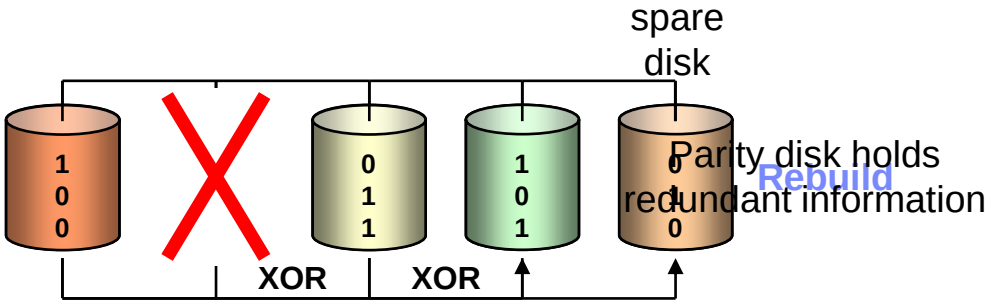
Short Résumé

- Position
  - IBM Research - Zurich Laboratory since 1988
- Research interests
  - performance evaluation
  - optimization and control of computer communication networks
  - reliability of storage systems
  - storage provisioning for Big Data
  - cloud infrastructures
  - switch architectures
  - stochastic systems
- Affiliations
  - IARIA Fellow
  - senior member of IEEE
  - IFIP Working Group 6.3
- Education
  - Ph.D. in Electrical Engineering from Columbia University, New York
  - M.S. in Electrical Engineering from Columbia University, New York
  - B.S. in Electrical Engineering from the National Technical University of Athens, Greece

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)

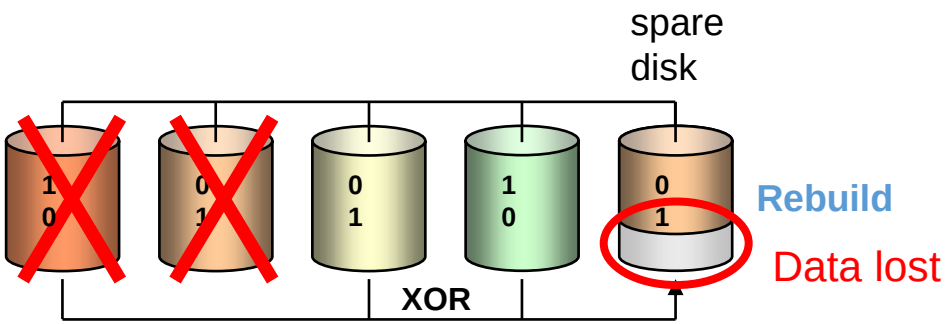


  - RAID-5: Tolerates one disk failure [Patterson *et al*. 1988]

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

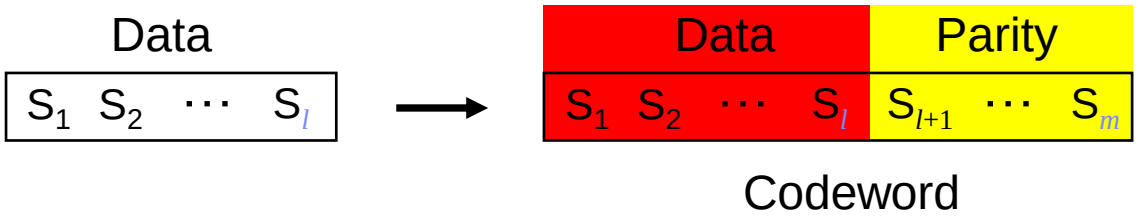
- Reliability enhanced by a large variety of redundancy and recovery schemes
  - RAID systems



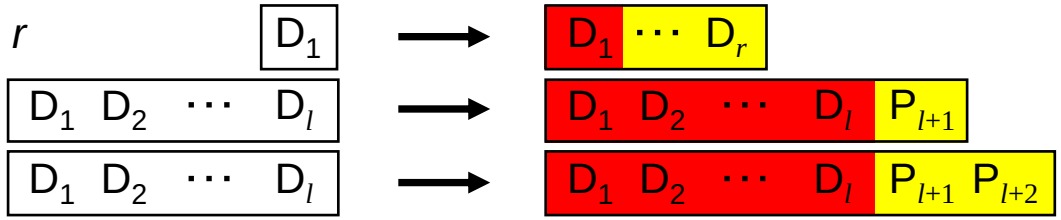
  - RAID-5: Tolerates one disk failure
  - RAID-6: Tolerates two disk failures

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 : $s_{eff} = l/m$ (Code rate)

- Google : Three-way replication (3,1) → s_{eff} = 33% to Reed-Solomon (9,6) → s_{eff} = 66 %
- Facebook : Three-way replication (3,1) → s_{eff} = 33% to Reed-Solomon (14,10) → s_{eff} = 71 %
- Microsoft Azure : Three-way replication (3,1) → s_{eff} = 33% to LRC (16,12) → s_{eff} = 75 %

Lazy Rebuild Scheme

- Erasure coding
  - reduction in storage overhead
  - improvement of reliability achieved

  but
  - repair problem
    - ➢ increased network traffic needed to repair data lost

    - ➢ Solution: **lazy rebuild**
      - rebuild process not triggered immediately upon first device failure
      - rebuild process delayed until additional device failures occur
        - ✓ reduces recovery bandwidth
        - ✓ keeps the impact on read performance and data durability low

      M. Silberstein et al. “Lazy means smart: Reducing repair bandwidth costs in erasure-coded distributed storage”, SYSTOR 2014

Reliability of Erasure Coded Systems

- Analytical closed-form expressions for the MTTDL and EAFDL of erasure coded systems in the presence of latent errors

  I. Iliadis, “Reliability Assessment of Erasure-Coded Storage Systems with Latent Errors”, CTRQ 2021

  - General method for obtaining the MTTDL and EAFDL
    - Most likely path that leads to data loss
      - direct path to data loss

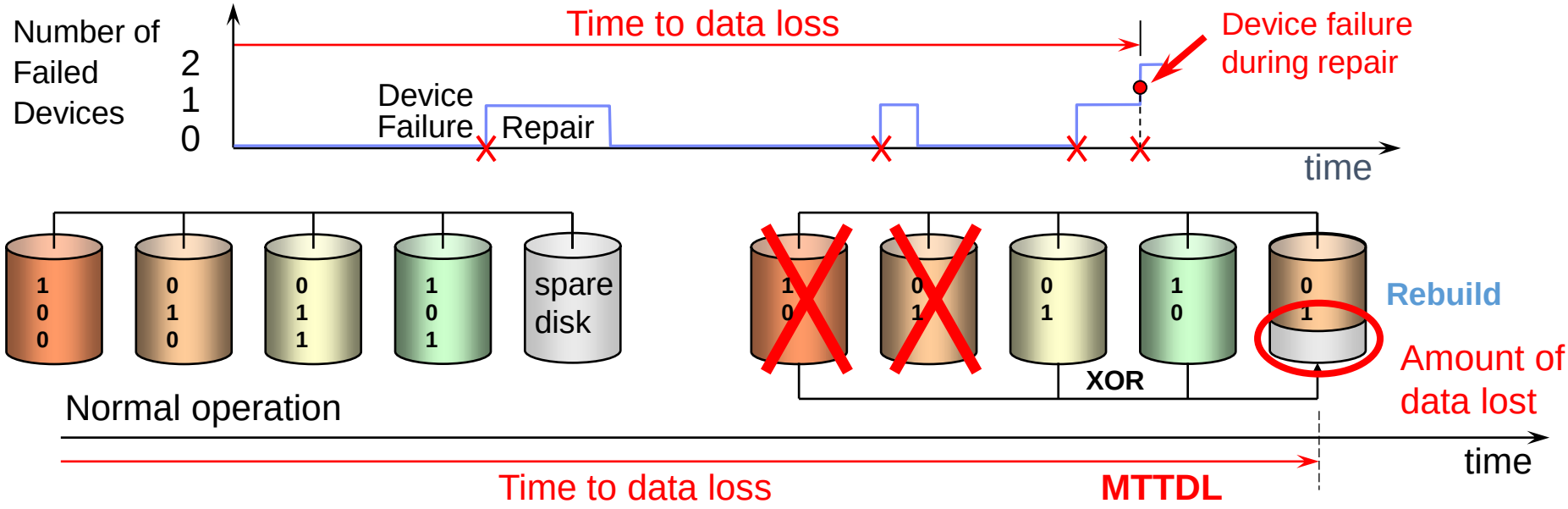
OBJECTIVE

To assess system reliability when the lazy rebuild scheme is employed

RESULTS

- Theoretical assessment of the effect of lazy rebuild on reliability
- Evaluation of MTTDL and EAFDL
  - Analytical approach that does not involve Markovian analysis
    - EAFDL and MTTDL tend to be insensitive to the failure time distributions
      - Real-world distributions, such as Weibull and gamma

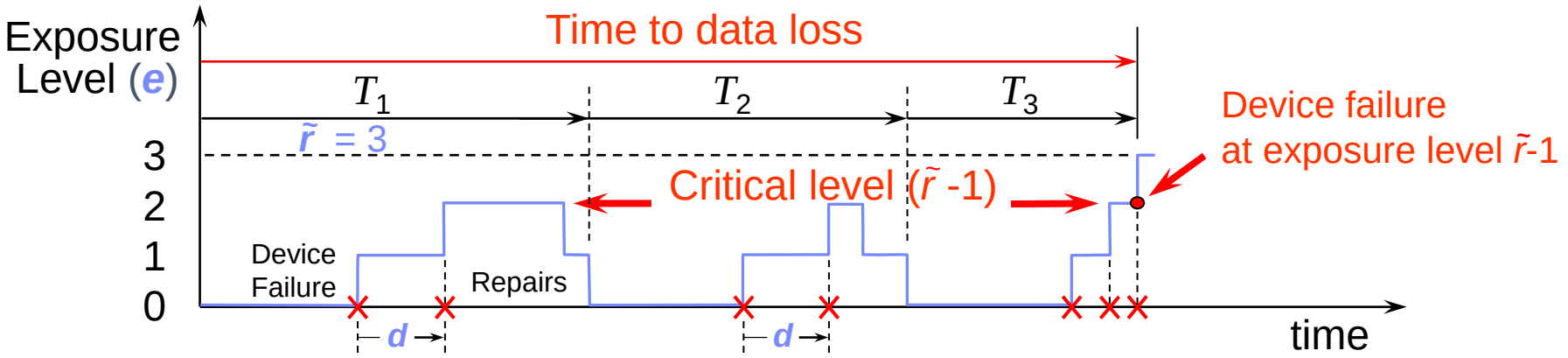
Reliability Metrics – MTTDL and EAFDL



- 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 MTTDL and EAFDL



- EAFDL evaluated in parallel with MTTDL
  - $\tilde{r}$: Minimum number of device failures that may lead to data loss ($\tilde{r} = m - l + 1$)
  - d : Lazy rebuild threshold ($0 \le d < m - l$)
  - e : Exposure Level: maximum number of symbols that any codeword has lost
  - T_i : Cycles (Fully Operational Periods / Repair Periods)
  - P_{DL} : Probability of data loss during repair period
  - Q : Amount of data lost upon a first-device failure
  - U : Amount of user data stored in a system comprised of n devices
  - $1/\lambda$: Mean Time to Failure (MTTF) of a device

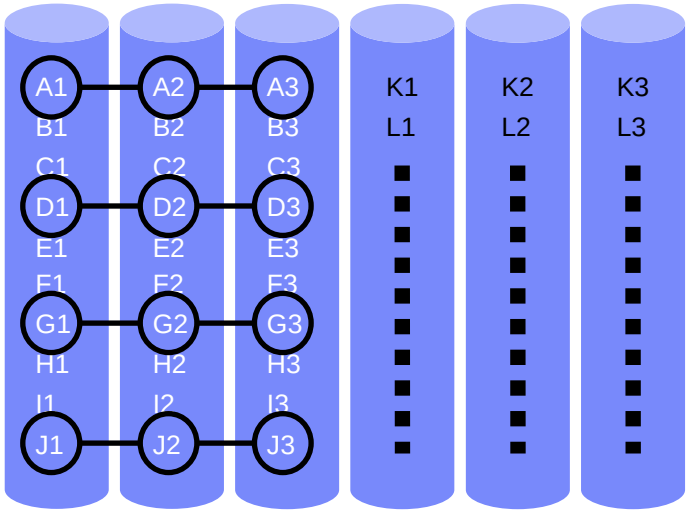
  - MTTDL = $\sum_i E(T_i) = \dfrac{E(T)}{P_{DL}}$ EAFDL $\approx \dfrac{E(Q)}{E(T)\,U}$

- System evolution does not depend only on the latest state, but on the entire path
  - underlying models are not semi-Markov

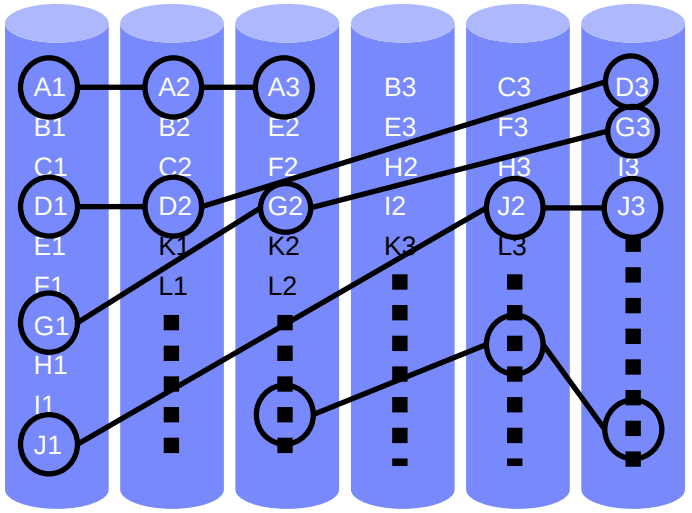
MTTDL and EAFDL expressions obtained using non-Markov analysis

Redundancy Placement

Erasure code with codeword length 3

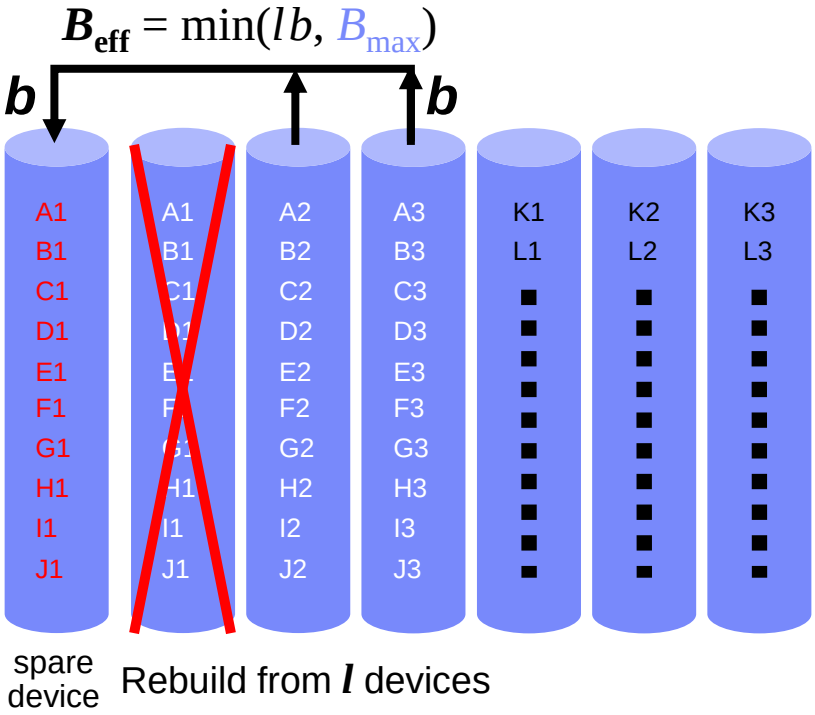


Clustered Placement

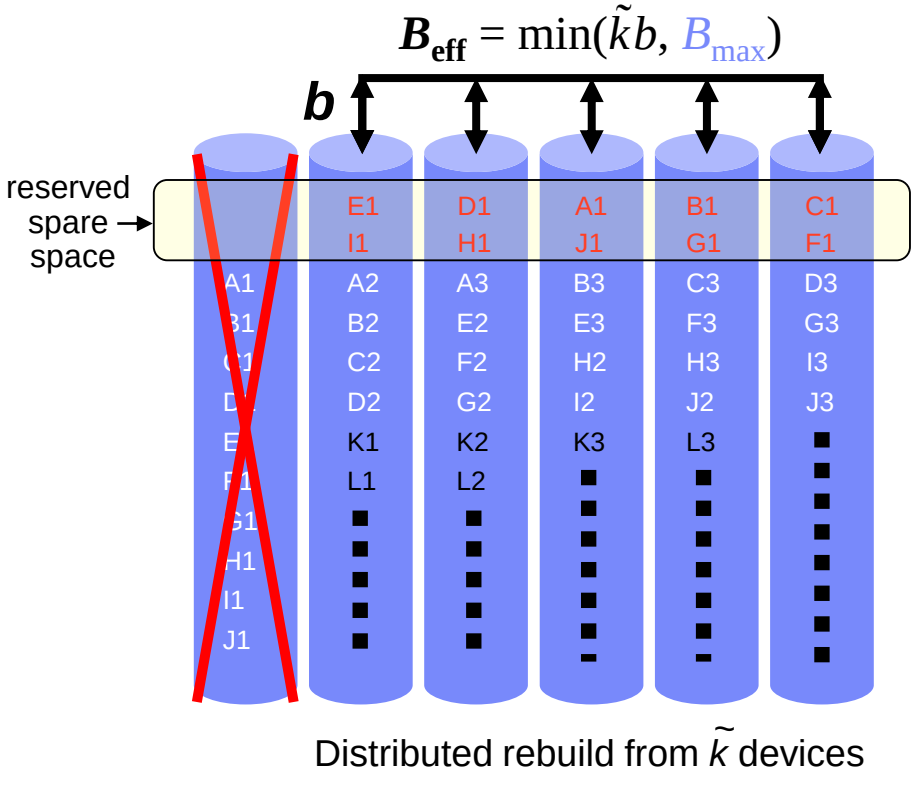


Declustered Placement

Device Failure and Rebuild Process

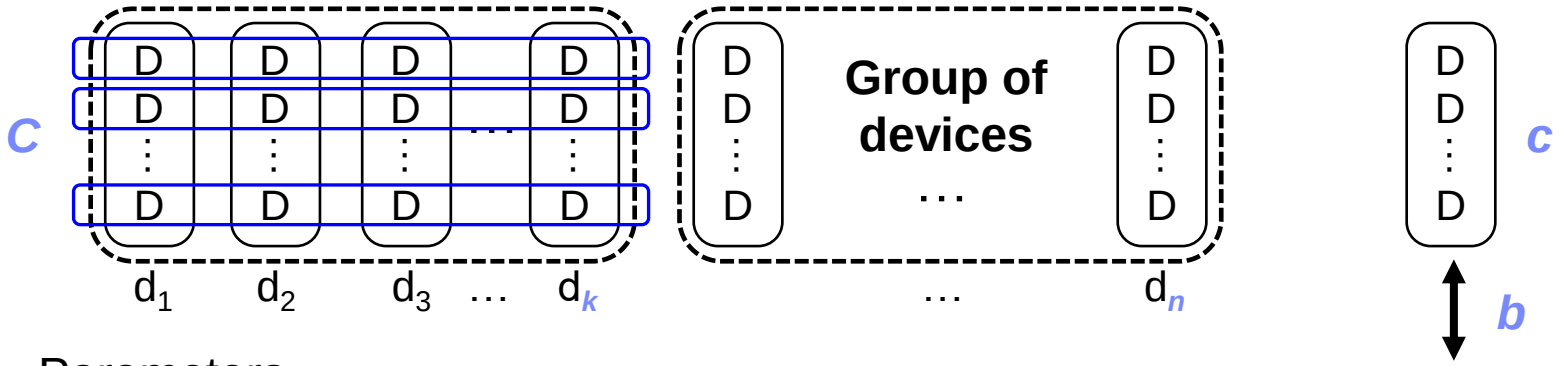


Clustered Placement



Declustered Placement

System Model



- Parameters
  - ***n*** : number of storage devices
  - ***k*** : number of devices in a group
  - ***c*** : amount of data stored on each device
  - ***C*** : number of codeword symbols stored in a device
  - ***b*** : average reserved rebuild bandwidth per device
  - $1/\lambda$: Mean Time to Failure (MTTF) of a device
    - General non-exponential failure distributions
  - $1/\mu$: Time to read (or write) an amount of ***c*** data at a rate ***b*** from (or to) a device
    - $1/\mu$ = ***c*** / ***b***
      - Highly reliable devices: λ / μ << 1

Theoretical Results

- n : number of storage devices
- k : group size (number of devices in a group)
- c : amount of data stored on each device
- (m,l) : MDS erasure code
- d : lazy rebuild threshold
- b : reserved rebuild bandwidth per device
- B_{max} : Maximum network rebuild bandwidth per group of devices
- $1/\lambda$: mean time to failure of a storage device
- P_s : probability of an unrecoverable sector (symbol) error

$$\mathrm{MTTDL} \approx \frac{E(T)}{P_{\mathrm{DL}}} \quad \text{and} \quad \mathrm{EAFDL} \approx \frac{E(Q)}{E(T)\cdot U} \quad \text{where}$$

$$P_{\mathrm{DL}} \approx P_{\mathrm{DF}} + \sum_{u=d+1}^{\tilde{r}-1} P_{\mathrm{UF}_u}$$

$$P_{\mathrm{UF}_u} \approx -\left(\lambda\, c \prod_{j=1}^{d} V_j\right)^{u-d-1} \frac{E(X^{u-d-1})}{[E(X)]^{u-d-1}} \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i} V_i^{u-1-i}\right) \log(\hat{q}_u)^{-(u-d-1)} \left(\hat{q}_u - \sum_{i=0}^{u-d-1} \frac{\log(\hat{q}_u)^i}{i!}\right)$$

$$P_{\mathrm{DF}} \approx \frac{(\lambda\, c \prod_{j=1}^{d} V_j)^{\tilde{r}-d-1}}{(\tilde{r}-d-1)!} \frac{E(X^{\tilde{r}-d-1})}{[E(X)]^{\tilde{r}-d-1}} \prod_{i=d+1}^{\tilde{r}-1} \frac{\tilde{n}_i}{b_i} V_i^{\tilde{r}-1-i}\,, \qquad E(T) = \left(\sum_{u=0}^{d} \frac{1}{\tilde{n}_u}\right) \Big/ \lambda$$

$$E(Q) \approx E(Q_{\mathrm{DF}}) + \sum_{u=d+1}^{\tilde{r}-1} E(Q_{\mathrm{UF}_u})$$

$$E(Q_{\mathrm{UF}_u}) \approx c\,\frac{l\,\tilde{r}}{m} \frac{\left(\lambda\, c \prod_{j=1}^{d} V_j\right)^{u-d-1}}{(u-d)!} \frac{E(X^{u-d-1})}{[E(X)]^{u-d-1}} \left(\prod_{j=1}^{d} V_j\right) \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i} V_i^{u-i}\right) \binom{m-u}{\tilde{r}-u} P_s^{\tilde{r}-u}$$

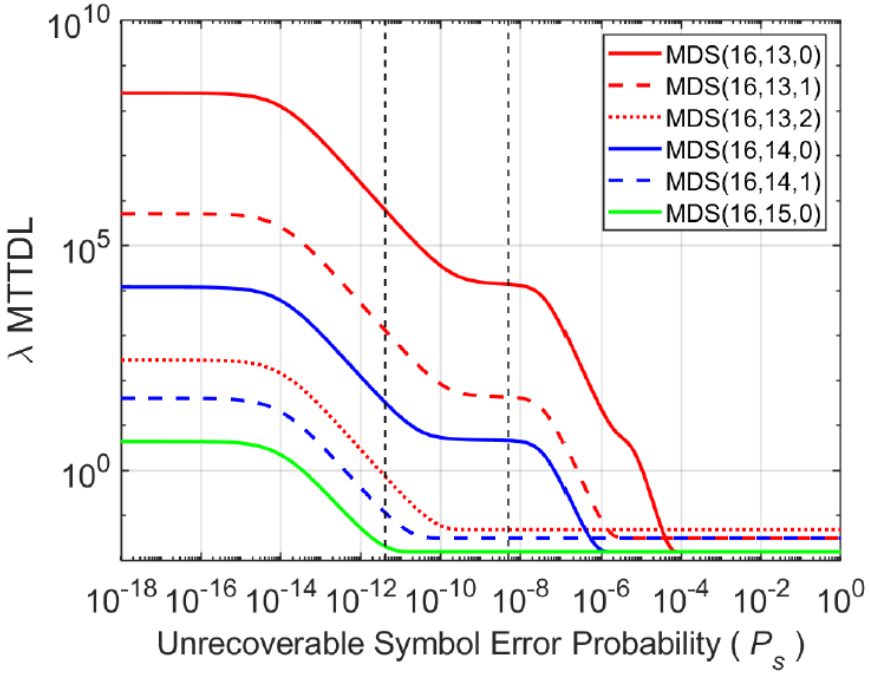
$$E(Q_{\mathrm{DF}}) \approx c\,\frac{l\,\tilde{r}}{m} \left(\lambda\, c \prod_{j=1}^{d} V_j\right)^{\tilde{r}-d-1} \frac{1}{(\tilde{r}-d)!} \frac{E(X^{\tilde{r}-d-1})}{[E(X)]^{\tilde{r}-d-1}} \left(\prod_{j=1}^{d} V_j\right) \left(\prod_{i=d+1}^{\tilde{r}-1} \frac{\tilde{n}_i}{b_i} V_i^{\tilde{r}-i}\right)$$

Numerical Results

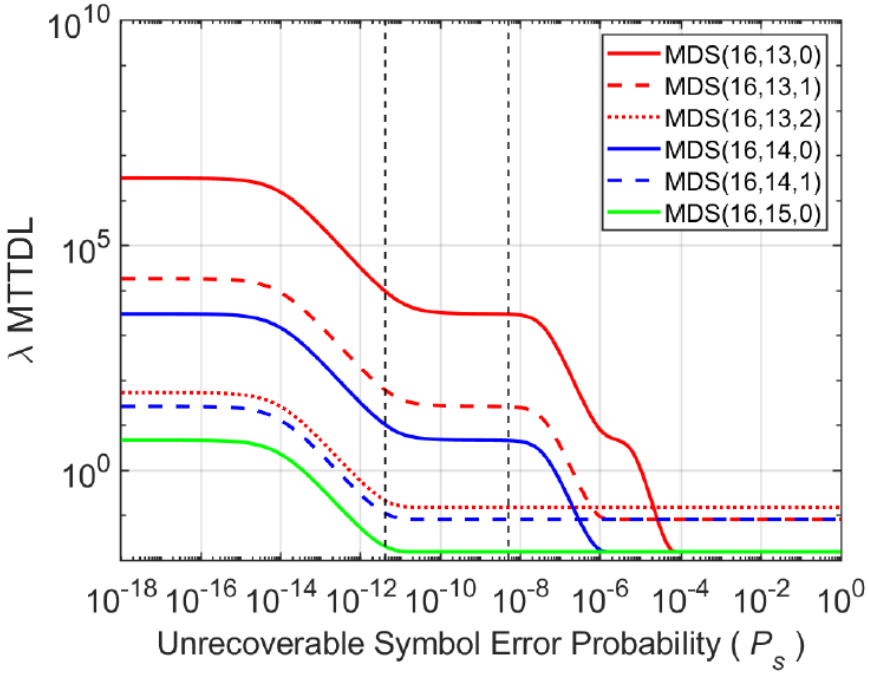
- n = 64 : number of storage devices
- c = 12 TB : amount of data stored on each device
- s = 512 B : sector size
- $1/\lambda$ = 300,000 h : MTTF
- b = 50 MB/s : reserved rebuild bandwidth
  - $1/\mu = c/b$ = 66.7 h : MTTR
  - λ/μ = 0.0002 $\ll 1$: MTTR to MTTF ratio
- m = 16 : number of symbols per codeword
- P_s : P(unrecoverable sector error)

- Numerical results for two system configurations
  - Declustered placement
    - $k = n = 64$
  - Clustered placement
    - $k = 16$
      - System comprises 4 clustered groups

Effect of Latent Errors on MTDDL



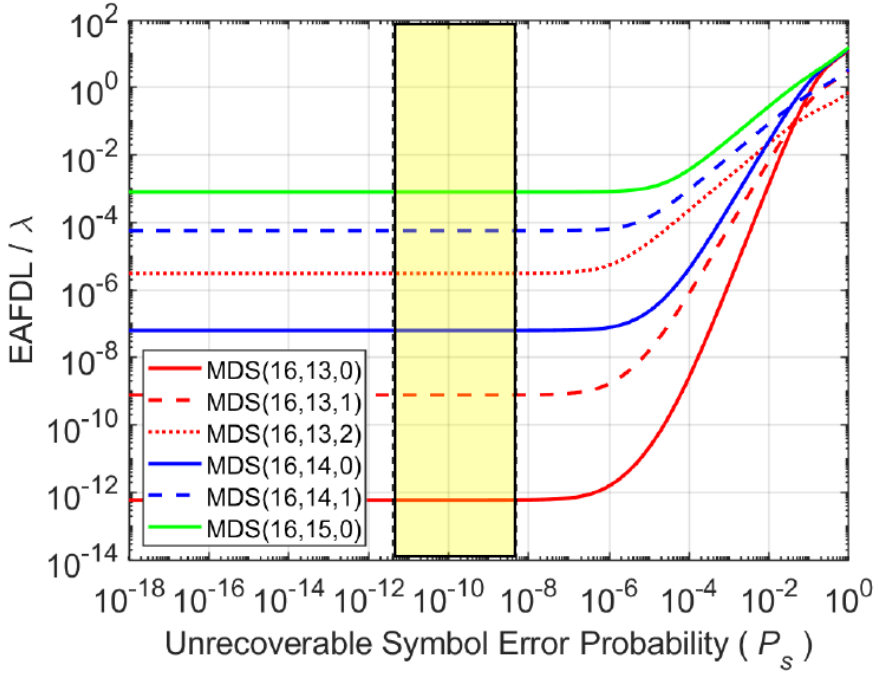
(a) $k = 64$ (declustered data placement scheme)



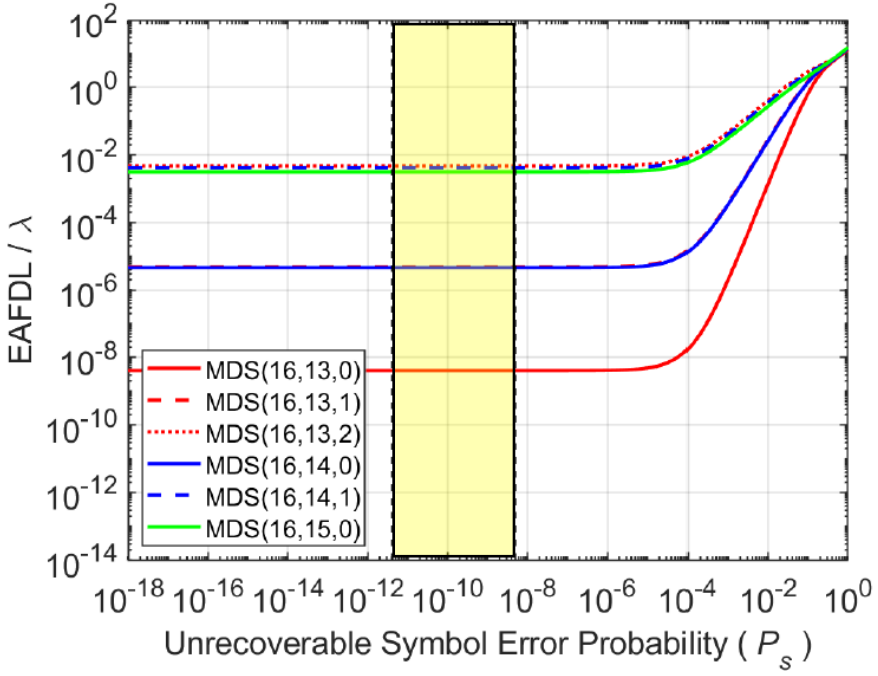
(b) $k = 16$ (clustered data placement scheme)

- MTTDL decreases monotonically with Ps and exhibits $m - l - d$ plateaus
- Field measurements show P_s to be in the interval [4.096x10^{-11}, 5x10^{-9}]
  - MTTDL significantly degraded by the presence of latent errors
- Increasing the number of parities (reducing l) improves reliability by orders of magnitude
- Employing lazy rebuild degrades reliability by orders of magnitude
- The declustered placement scheme achieves a significantly higher MTTDL than the clustered one

Effect of Latent Errors on EAFDL



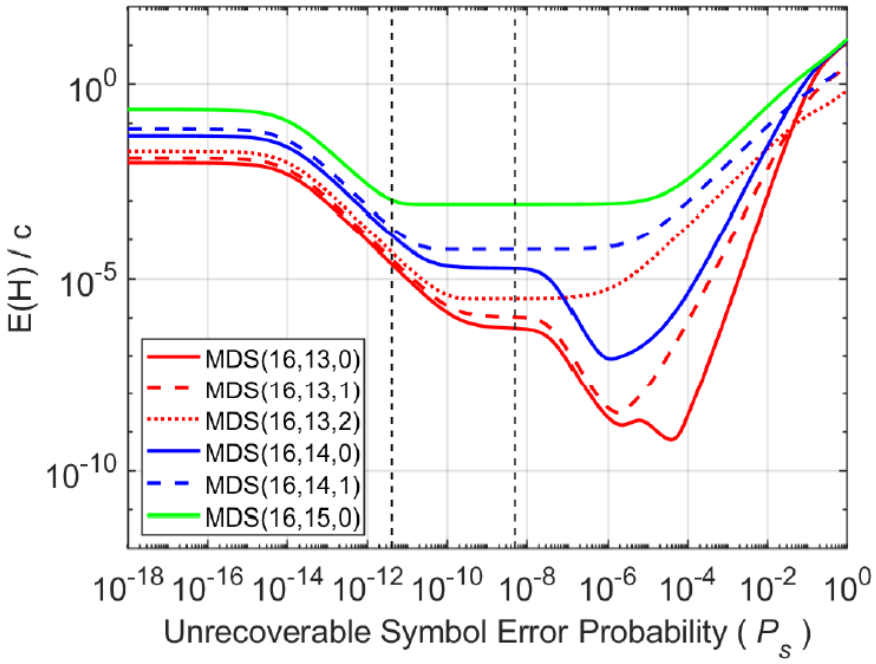
(a) $k = 64$ (declustered data placement scheme)



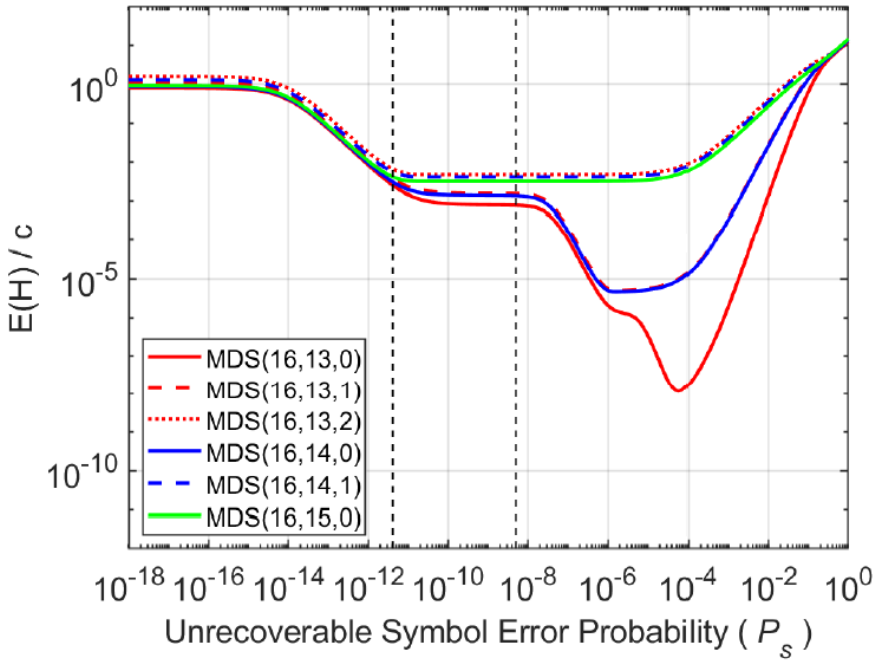
(b) $k = 16$ (clustered data placement scheme)

- EAFDL affected at high sector error probabilities
- EAFDL unaffected by the presence of latent errors in the region of practical interest
- Increasing the number of parities (reducing l) improves reliability by orders of magnitude
- Employing lazy rebuild degrades reliability by orders of magnitude
- The declustered placement scheme achieves a significantly lower EAFDL than the clustered one

Effect of Latent Errors on E(H)



(a) $k = 64$ (declustered data placement scheme)



(b) $k = 16$ (clustered data placement scheme)

- In the interval [4.096×10^{-11}, 5×10^{-9}] of practical importance for P_s
  - E(H) significantly degraded by the presence of latent errors
  - E(H) not significantly affected by the employment of lazy rebuild

Summary

- Considered effect of the lazy rebuild scheme on the reliability of erasure-coded data storage systems
- Assessed the MTTDL and EAFDL reliability metrics using a non-Markovian analysis
- Derived closed-form expressions for the MTTDL and EAFDL metrics
- Demonstrated that system reliability is significantly degraded by the employment of the lazy rebuild scheme
- Established that the declustered placement scheme offers superior reliability in terms of both metrics
- Demonstrated that for practical values of unrecoverable sector error probabilities
  - MTTDL is adversely affected by the presence of latent errors
  - EAFDL is practically unaffected by the presence of latent errors

Future Work

- The reliability evaluation of erasure-coded systems when device failures, as well as unrecoverable latent errors are correlated