IBM

Zurich Research Laboratory

Reliability Assessment of Erasure-Coded Storage Systems with Latent Errors

Ilias Iliadis
ili@zurich.ibm.com
April 18-22, 2021

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

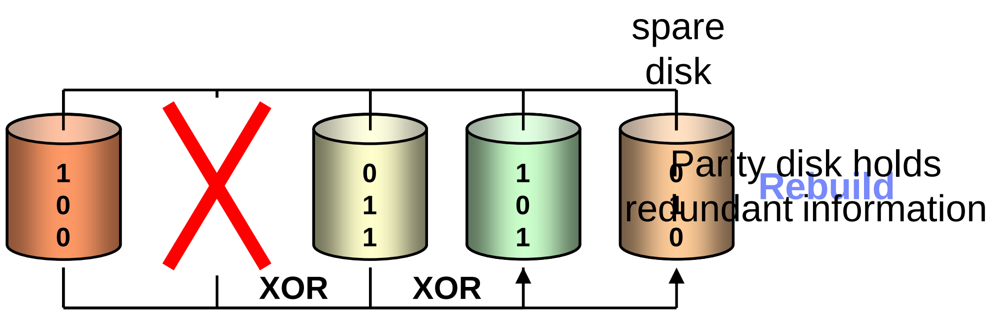
- Affiliations
  - IARIA Fellow
  - senior member of IEEE
  - IFIP Working Group 6.3

- 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)

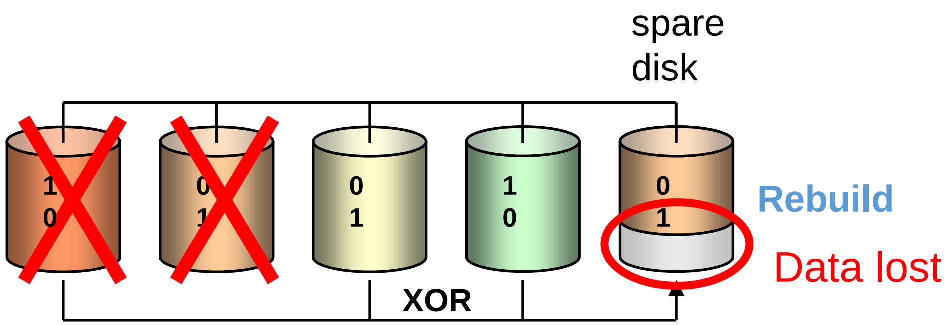


  - RAID-5: Tolerates one disk failure [Patterson *et al*. 1988]

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

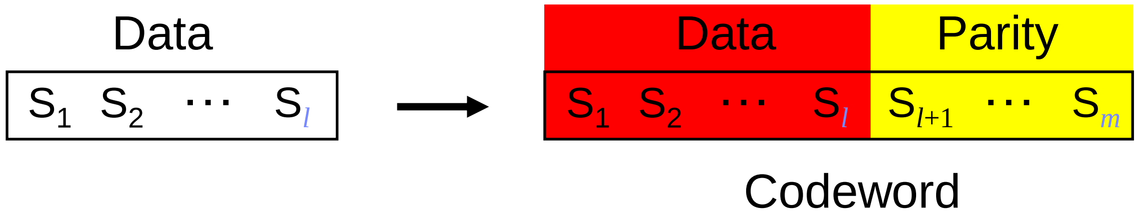
- Reliability enhanced by a large variety of redundancy and recovery schemes
  - RAID systems



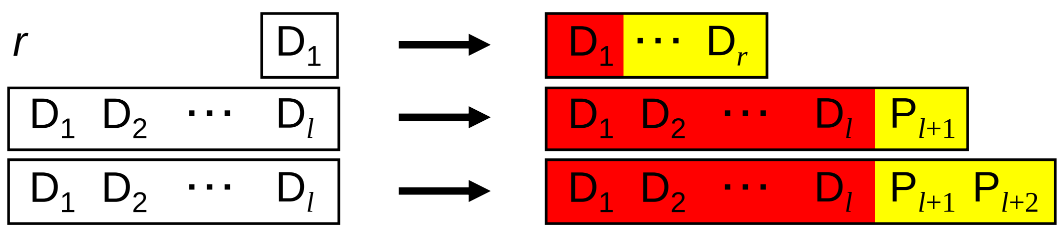
  - RAID-5: Tolerates one disk failure
  - RAID-6: Tolerates two disk failures

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

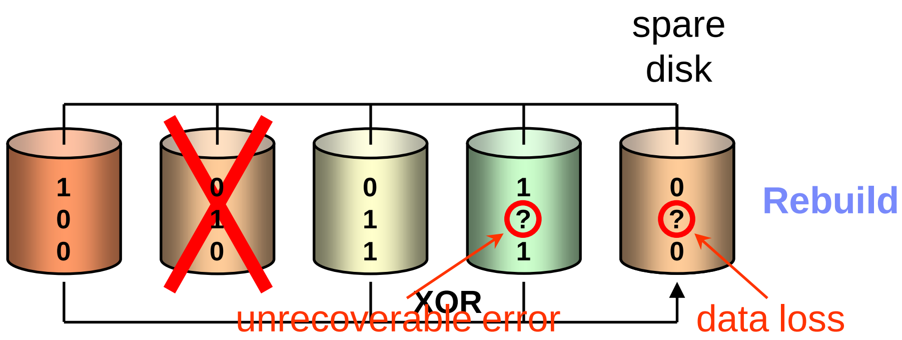
- Google : Three-way replication (3,1) → s_{eff} = 33% to Reed-Solomon (9,6) → s_{eff} = 66 %
- Facebook : Three-way replication (3,1) → s_{eff} = 33% to Reed-Solomon (14,10) → s_{eff} = 71 %
- Microsoft Azure : Three-way replication (3,1) → s_{eff} = 33% to LRC (16,12) → s_{eff} = 75 %

Media errors

- "bit rot" problem: the magnetism of a single bit or a few bits is flipped
  - ➢ This type of problem can often (but not always) be detected and corrected with low-level ECC embedded in the drive
- physical damage can occur on the media
  - ➢ head crash
  - ➢ media scratch

- Disk drives exhibit unrecoverable sector errors (*latent sector faults*)
  - a block or set of blocks are inaccessible
  - sectors are *corrupted silently* without the disk being able to detect it
  - a sector error is detected when the sector is accessed for storing or retrieving information
- Factors contributing to unrecoverable sector errors
  - increased areal density of disk drives
    - ➢ errors such as bit spillovers on adjacent tracks can corrupt more bits
  - increased use of cheap low-end desktop drives (Integrated Drive Electronics/Advanced Technology Attachment drives)
    - ➢ low cost, less tested, less machinery to handle disk errors
  - increased amount of software used on the storage stack
    - ➢ firmware on a desktop drive contains about 400 thousand lines of code
    - ➢ bugs are inevitable

Unrecoverable Errors and Data Loss



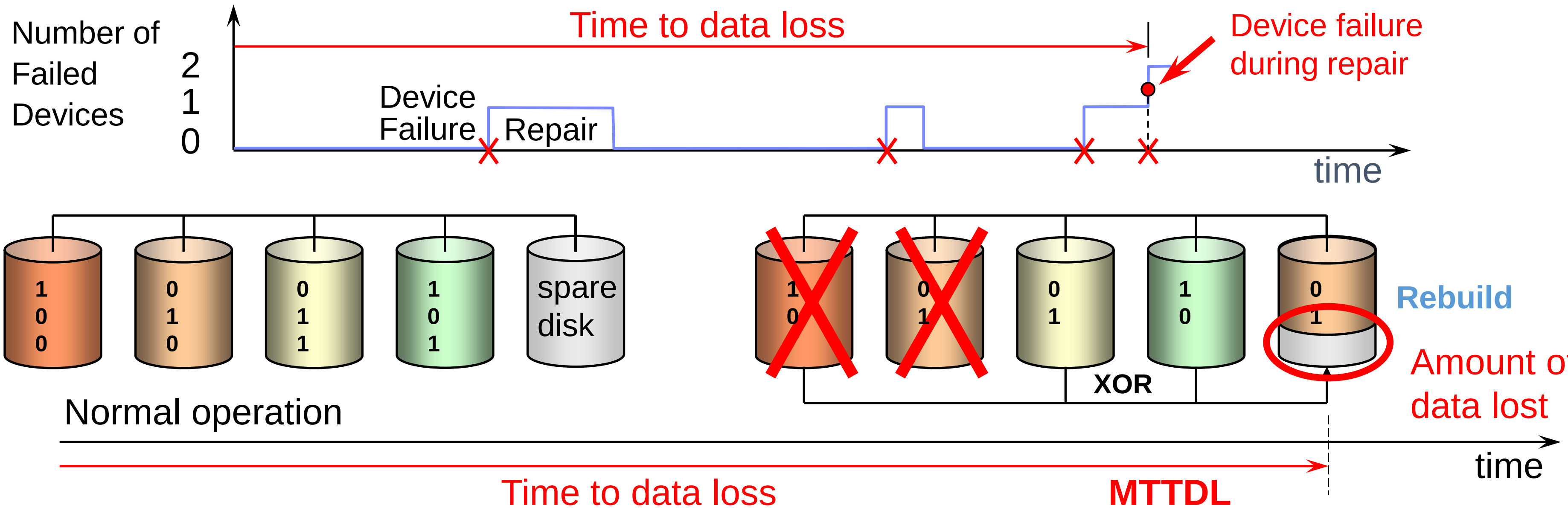
OBJECTIVE

To assess the extent of data loss due to disk failures and unrecoverable errors

RESULTS

- Theoretical assessment of the effect of latent errors 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

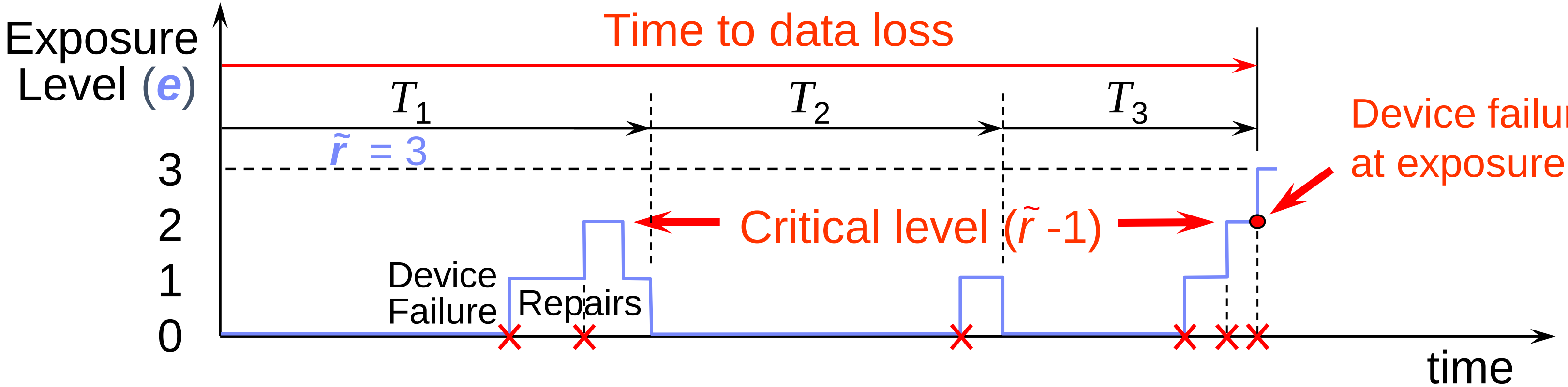
Reliability Metrics – MTTDL and EAFDL



- 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 ($\tilde{r} = m - l + 1$)
  - 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

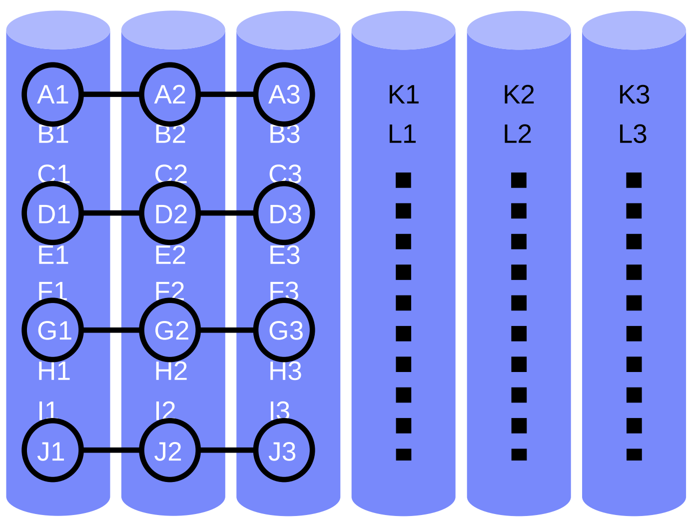
➢ $\text{MTTDL} = \sum_{i=1}^{m} E(T_i) = \frac{E(T)}{P_{DL}} \approx \frac{1}{n \lambda P_{DL}}$ $\qquad$ $\text{EAFDL} \approx \frac{n \lambda E(Q)}{U}$

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

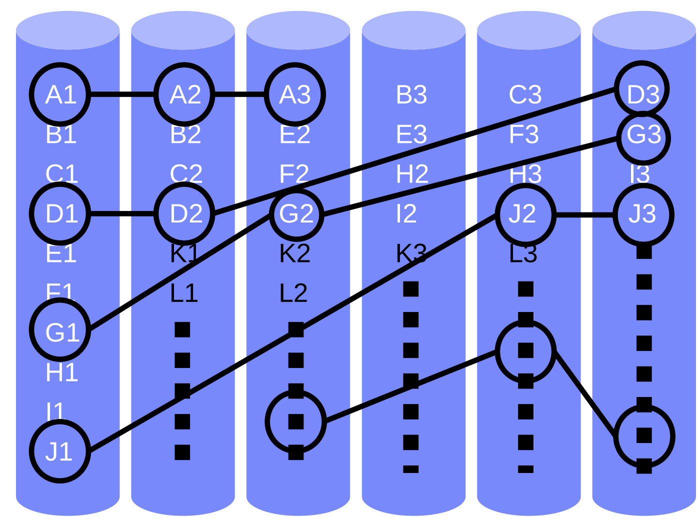
MTTDL and EAFDL expressions obtained using non-Markov analysis

Redundancy Placement

Erasure code with codeword length 3

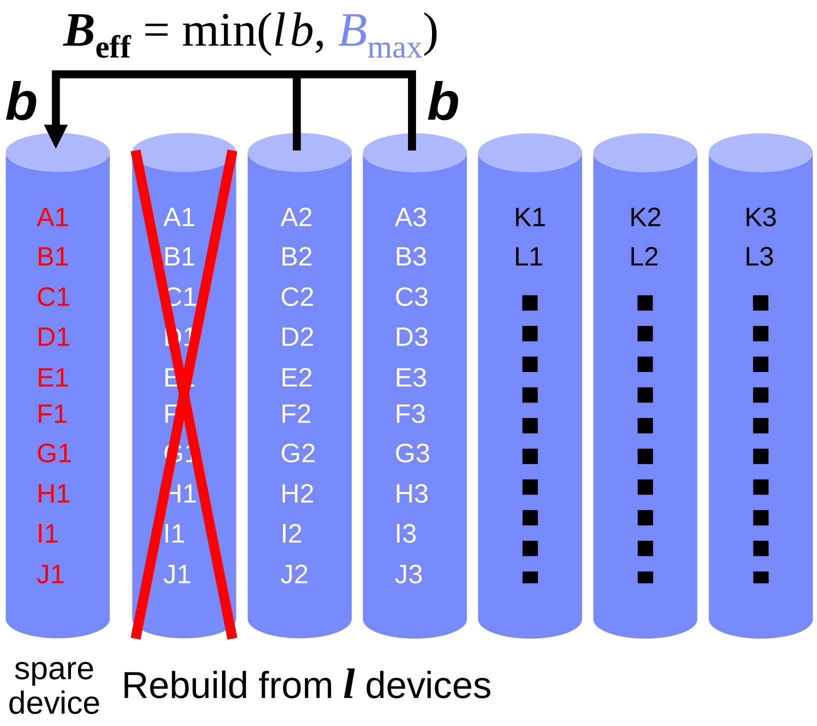


Clustered Placement

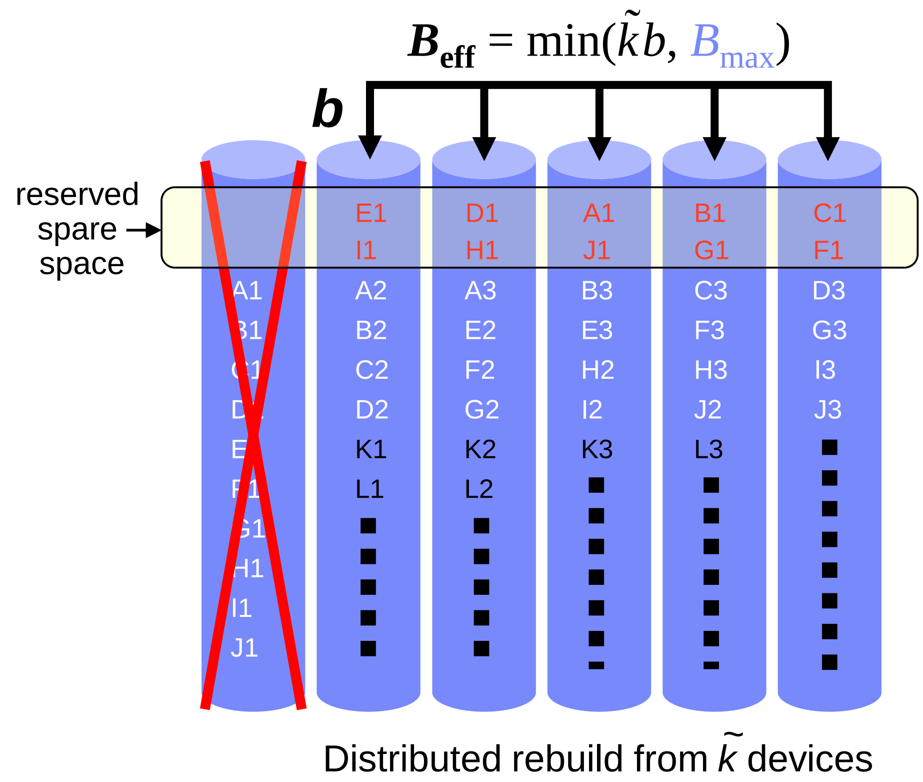


Declustered Placement

Device Failure and Rebuild Process

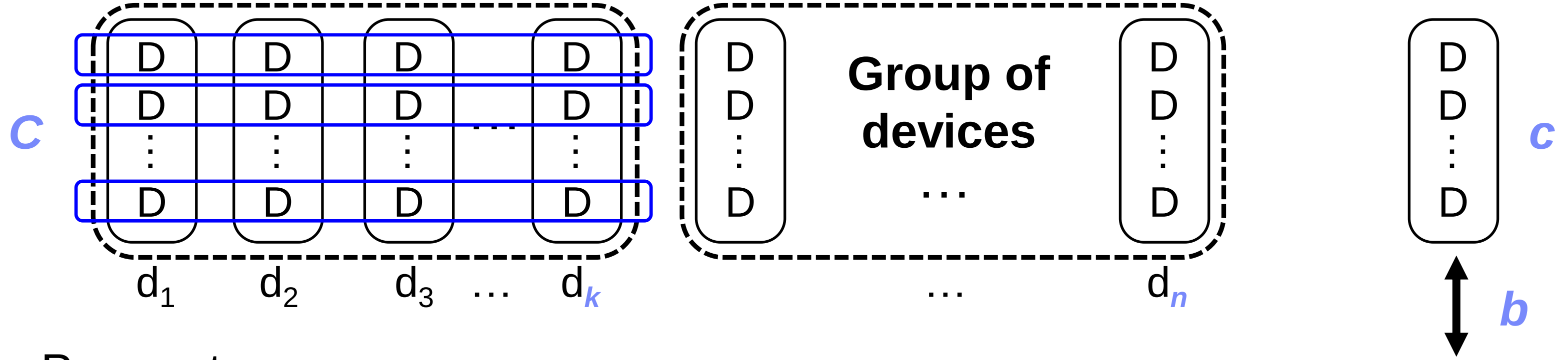


Clustered Placement



Declustered Placement

System Model

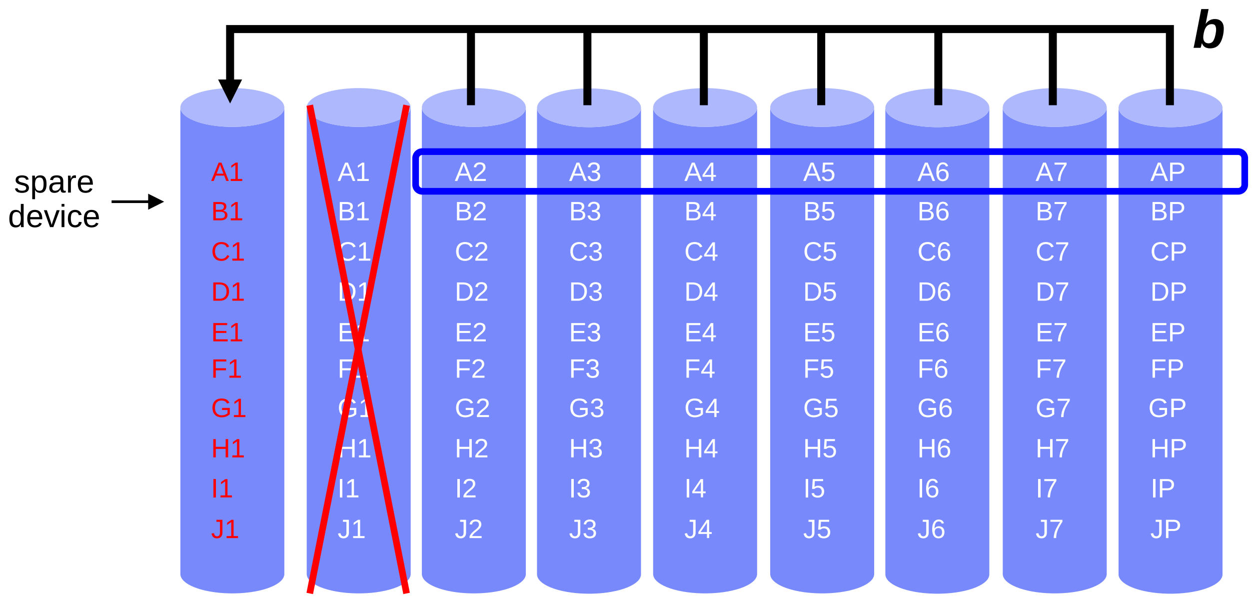


- Parameters
  - n : number of storage devices
  - k : number of devices in a group
  - c : amount of data stored on each device
  - C : number of codeword symbols stored in a device
  - b : average reserved rebuild bandwidth per device

  - $1/\lambda$: 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: $\lambda / \mu \ll 1$

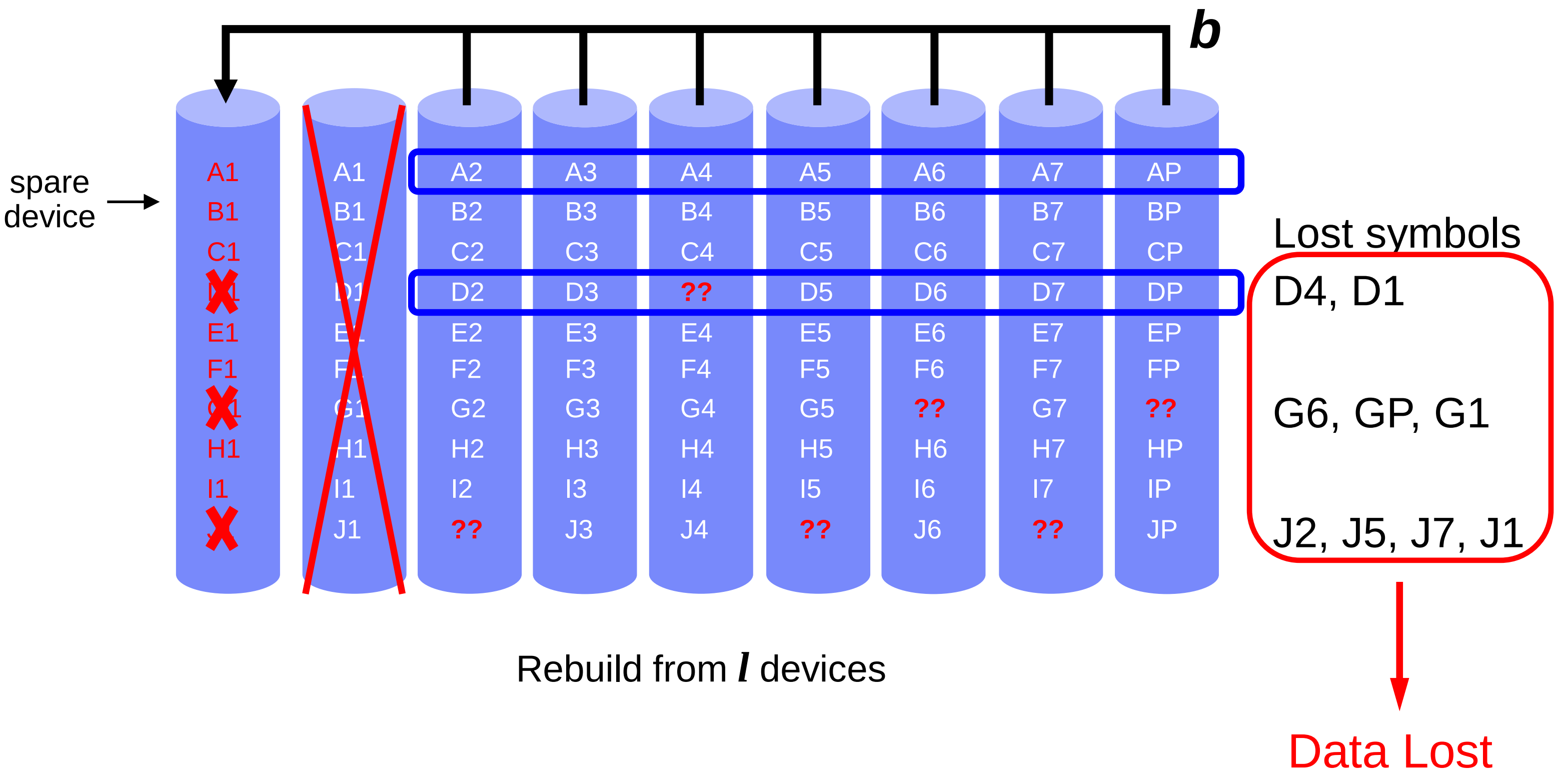
Device Failure and Rebuild Process



Rebuild from l devices

- No unrecoverable (latent) errors encountered during rebuild
  - Successful rebuild

Unrecoverble Failure during Rebuild Process



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
- 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

$$\text{MTTDL} \approx \frac{1}{n\,\lambda\,P_{\text{DL}}} \quad \text{and} \quad \text{EAFDL} \approx \frac{m\,\lambda\,E(Q)}{l\,c} \quad \text{where}$$

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

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

$$P_{\text{DF}} \approx (\lambda\,c)^{\tilde{r}-1}\,\frac{1}{(\tilde{r}-1)!}\,\frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}}\,\prod_{i=1}^{\tilde{r}-1}\frac{\tilde{n}_i}{b_i}\,V_i^{\tilde{r}-1-i}$$

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

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

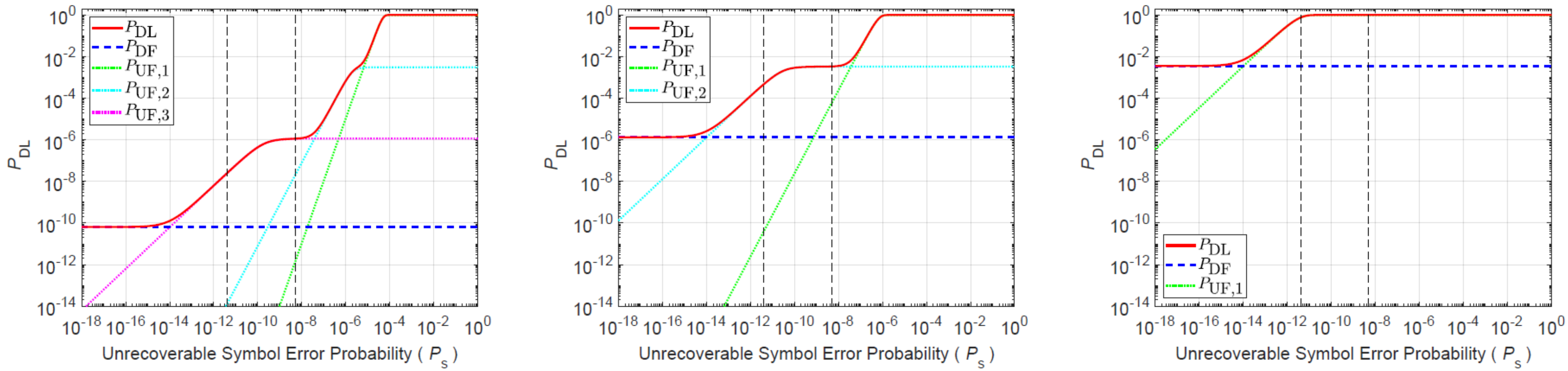
$$E(Q_{\text{DF}}) \approx c\,\frac{l}{m}\,(\lambda\,c)^{\tilde{r}-1}\,\frac{1}{(\tilde{r}-1)!}\,\frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}}\,\prod_{i=1}^{\tilde{r}-1}\frac{\tilde{n}_i}{b_i}\,V_i^{\tilde{r}-i}$$

Numerical Results

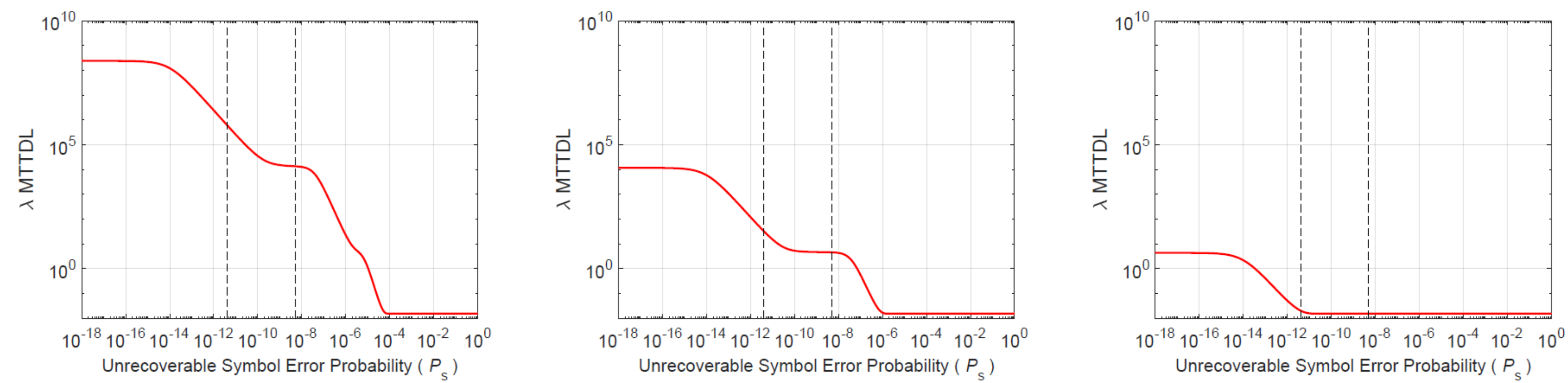
– n	$= 64$	:	number of storage devices
– c	$= 12$ TB	:	amount of data stored on each device
– s	$= 512$ B	:	sector size
– $1/\lambda$	$= 300{,}000$ h	:	MTTF
– b	$= 50$ MB/s	:	reserved rebuild bandwidth
➢ $1/\mu = c/b$	$= 66.7$ h	:	MTTR
➢ λ/μ	$= 0.0002 \ll 1$	:	MTTR to MTTF ratio
– m	$= 16$	:	number of 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 (declustered placement)

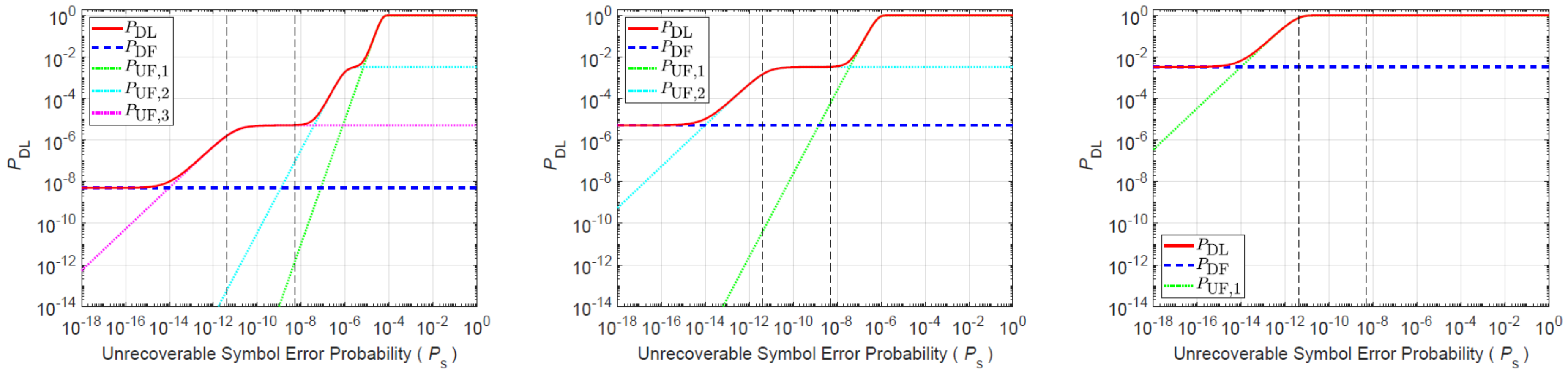


(a) $l = 13 \quad (\tilde{r} = 4)$ (b) $l = 14 \quad (\tilde{r} = 3)$ (c) $l = 15 \quad (\tilde{r} = 2)$

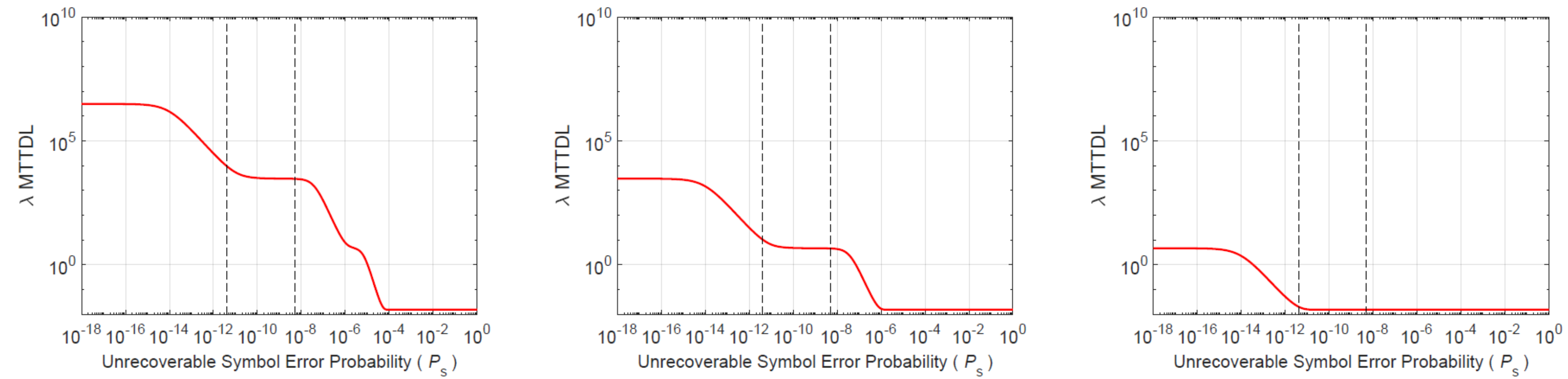


- MTTDL significantly degraded by the presence of latent errors
- Field measurements show P_s to be in the interval [4.096×10^{-11}, 5×10^{-9}]

Effect of Latent Errors on MTDDL (clustered placement)

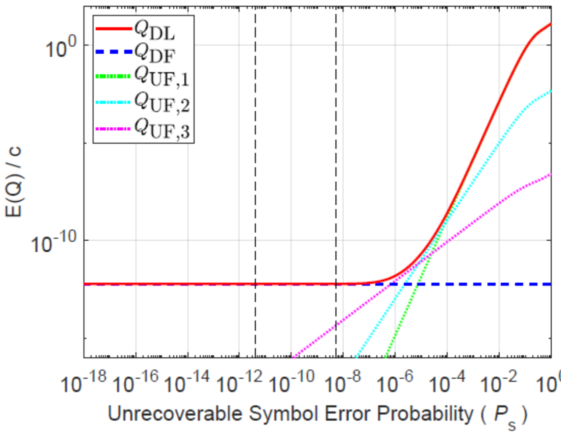


(a) $l = 13$ $(\tilde{r} = 4)$ (b) $l = 14$ $(\tilde{r} = 3)$, RAID-6 (c) $l = 15$ $(\tilde{r} = 2)$, RAID-5

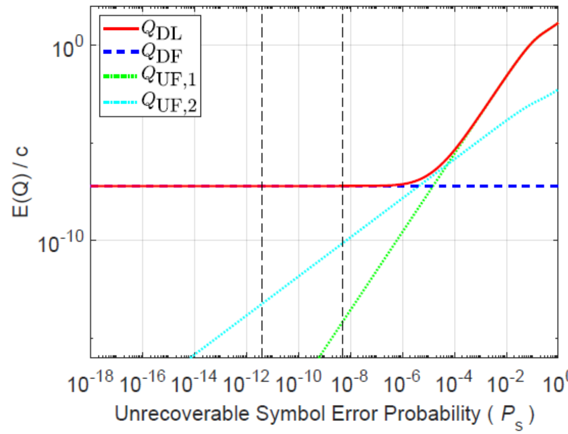


- MTTDL significantly degraded by the presence of latent errors
- Field measurements show P_s to be in the interval [4.096x10^-11^, 5x10^-9^]

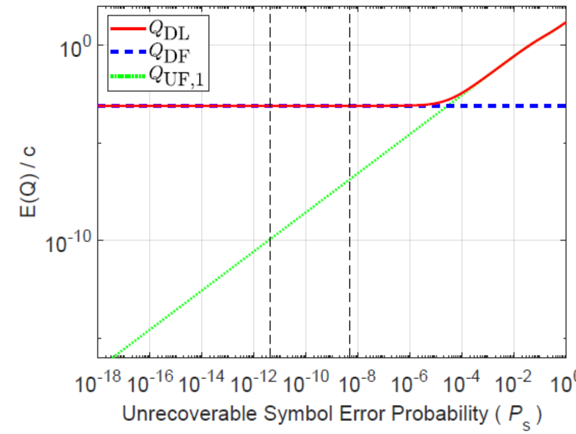
Effect of Latent Errors on EAFDL (declustered placement)



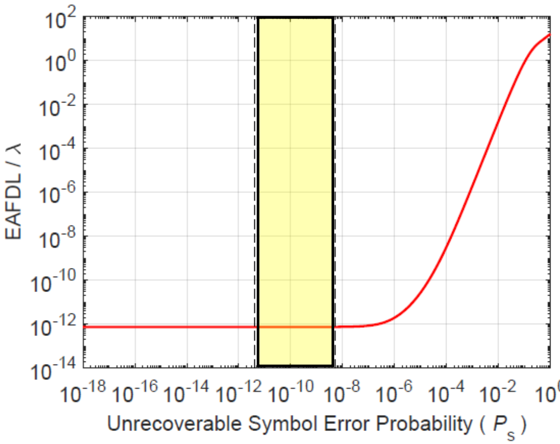
(a) $l = 13 \quad (\tilde{r} = 4)$

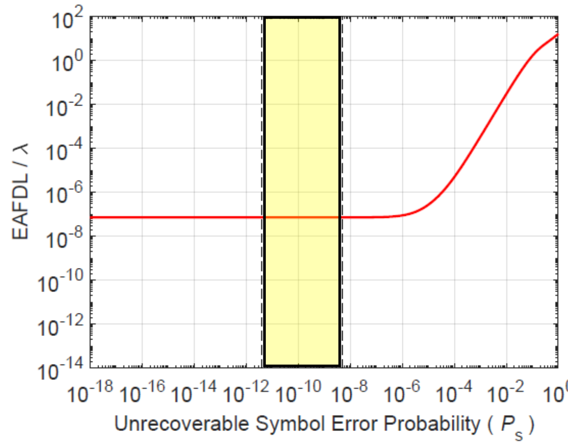


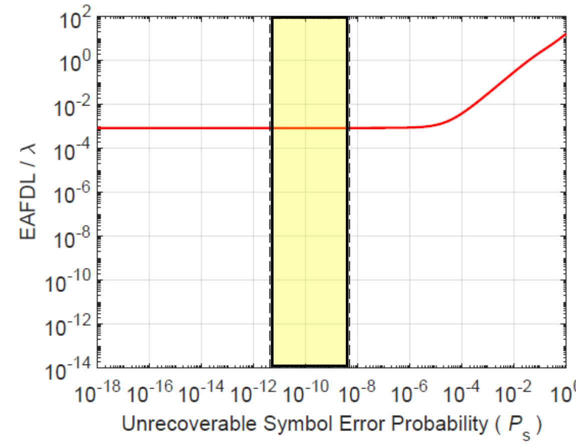
(b) $l = 14 \quad (\tilde{r} = 3)$



(c) $l = 15 \quad (\tilde{r} = 2)$

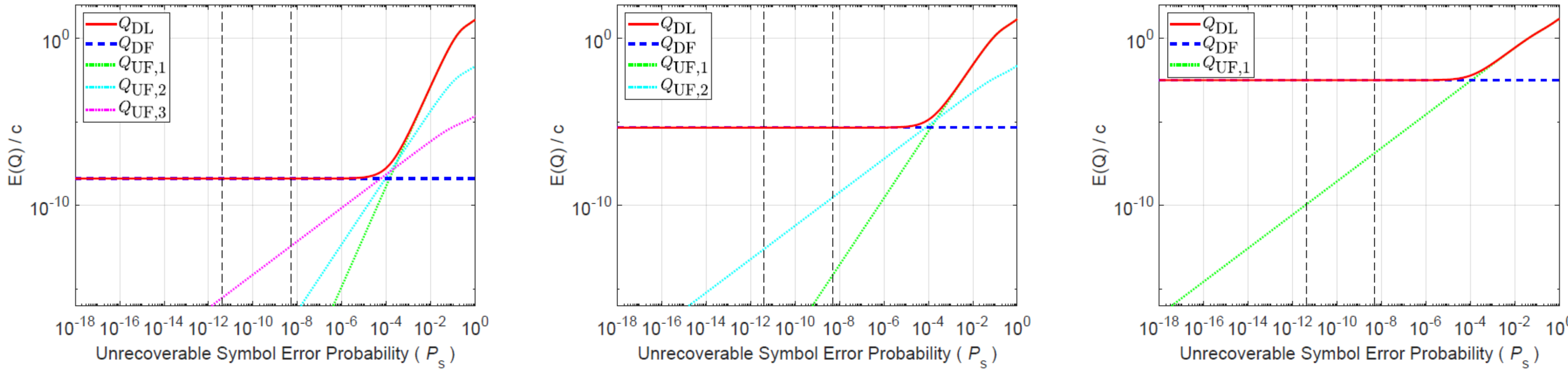




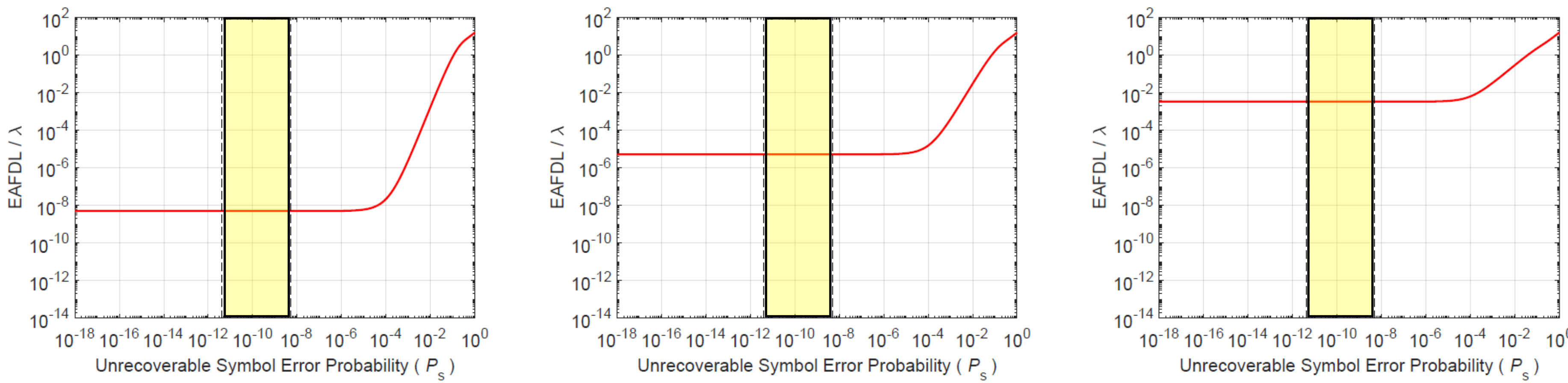


- EAFDL affected at high sector error probabilities
- EAFDL unaffected by the presence of latent errors in the region of practical interest

Effect of Latent Errors on EAFDL (clustered placement)



(a) $l = 13 \quad (\tilde{r} = 4)$ (b) $l = 14 \quad (\tilde{r} = 3)$, RAID-6 (c) $l = 15 \quad (\tilde{r} = 2)$, RAID-5



- EAFDL affected at high sector error probabilities
- EAFDL unaffected by the presence of latent errors in the region of practical interest

Summary

- Considered the reliability of erasure-coded storage systems in the presence of latent errors
- Assessed the MTTDL and EAFDL reliability metrics using a non-Markovian analysis
- Derived closed-form expressions for the MTTDL and EAFDL metrics
- 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.