Zurich Research Laboratory

Reliability of Data Storage Systems

Ilias Iliadis
April 20, 2015

Keynote NexComm 2015

www.zurich.ibm.com

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

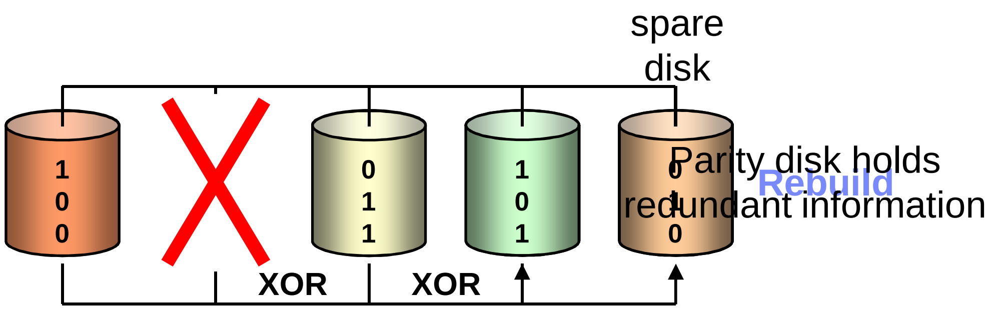
Storage

- 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)



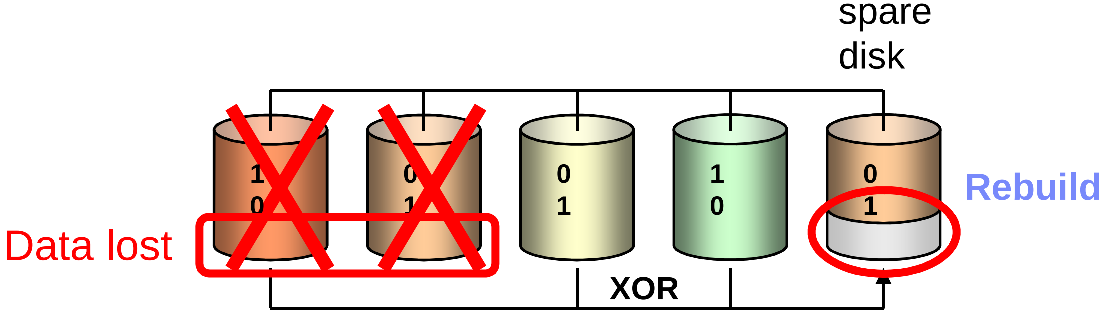
Parity disk holds redundant information

  - RAID-5: Tolerates one disk failure

Data Losses in Storage Systems

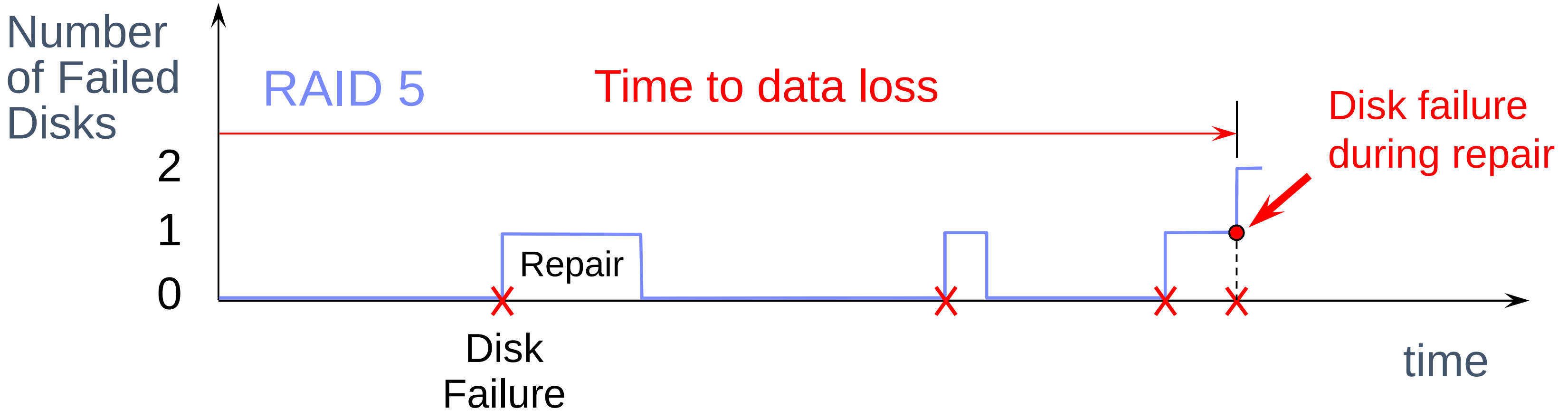
- Storage systems suffer from data losses due to
  - component failures
    - disk failures
    - node failures
  - media failures
    - unrecoverable and latent media errors

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

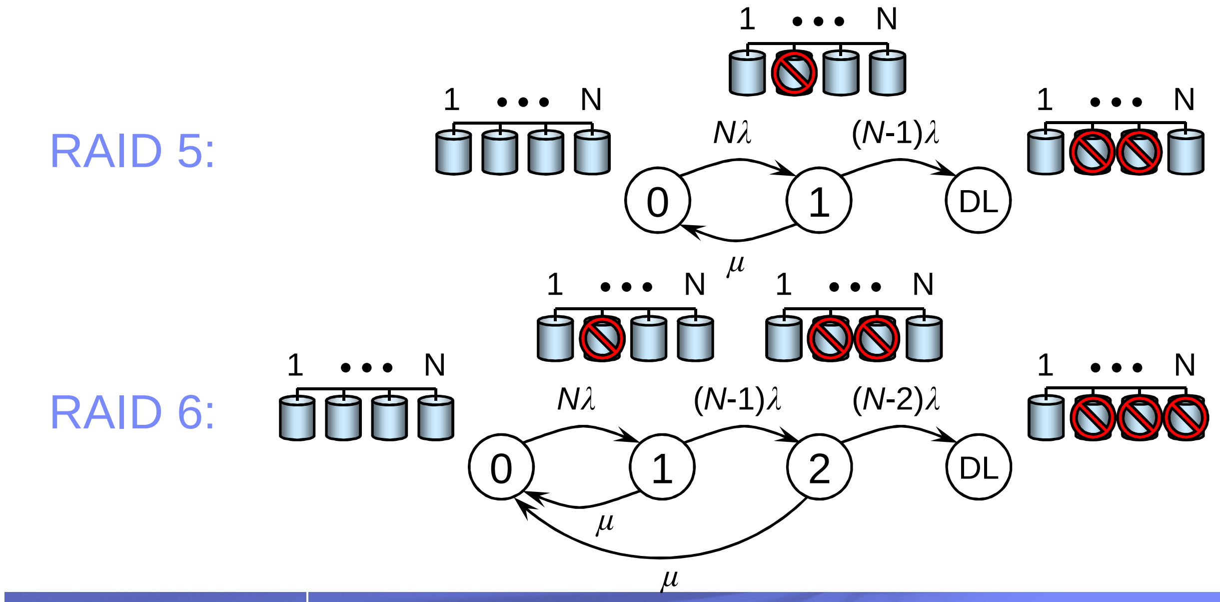


  - RAID-5: Tolerates one disk failure
  - RAID-6: Tolerates two disk failures

Time to Failure and MTTDL



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



- λ : 1/ MTTF for disks
- μ : 1/ MTTR

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

[Patterson *et al*. 1988]

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

[Chen *et al*. 1994]

original MTTDL equations

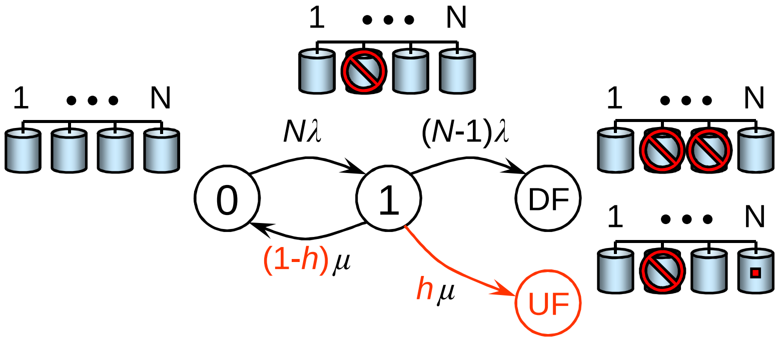
Markov Models for Unrecoverable Errors

- 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:

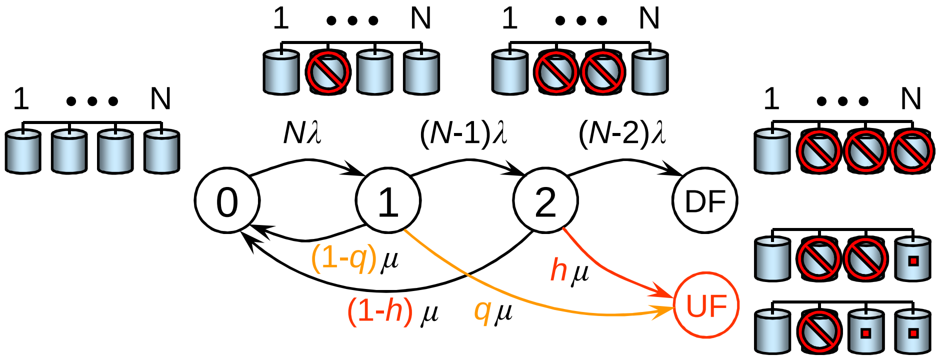


Data loss owing to:

- DF: Disk Failure
- UF: Unrecoverable Failure

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

RAID 6:



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

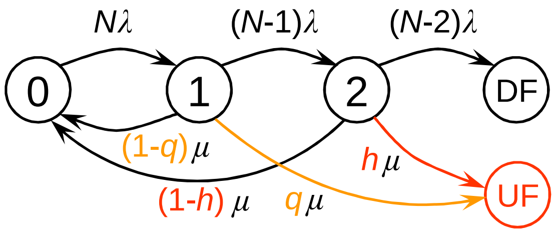
$$q = \binom{N-1}{2} C_d P_s^2 + O(P_s^3)$$

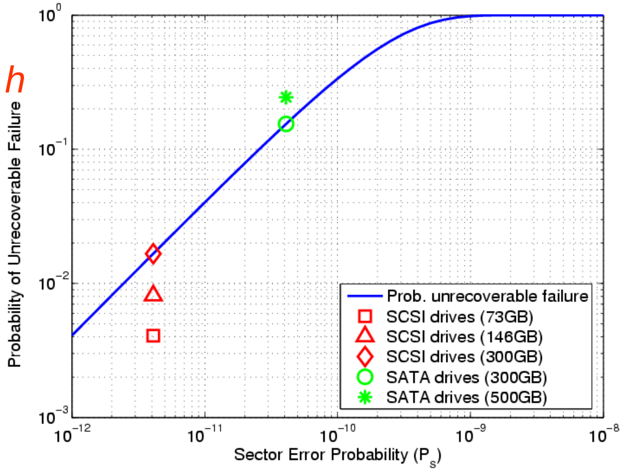
$q \ll h$ for $P_s \ll$

MTTDL for RAID 5 and RAID 6

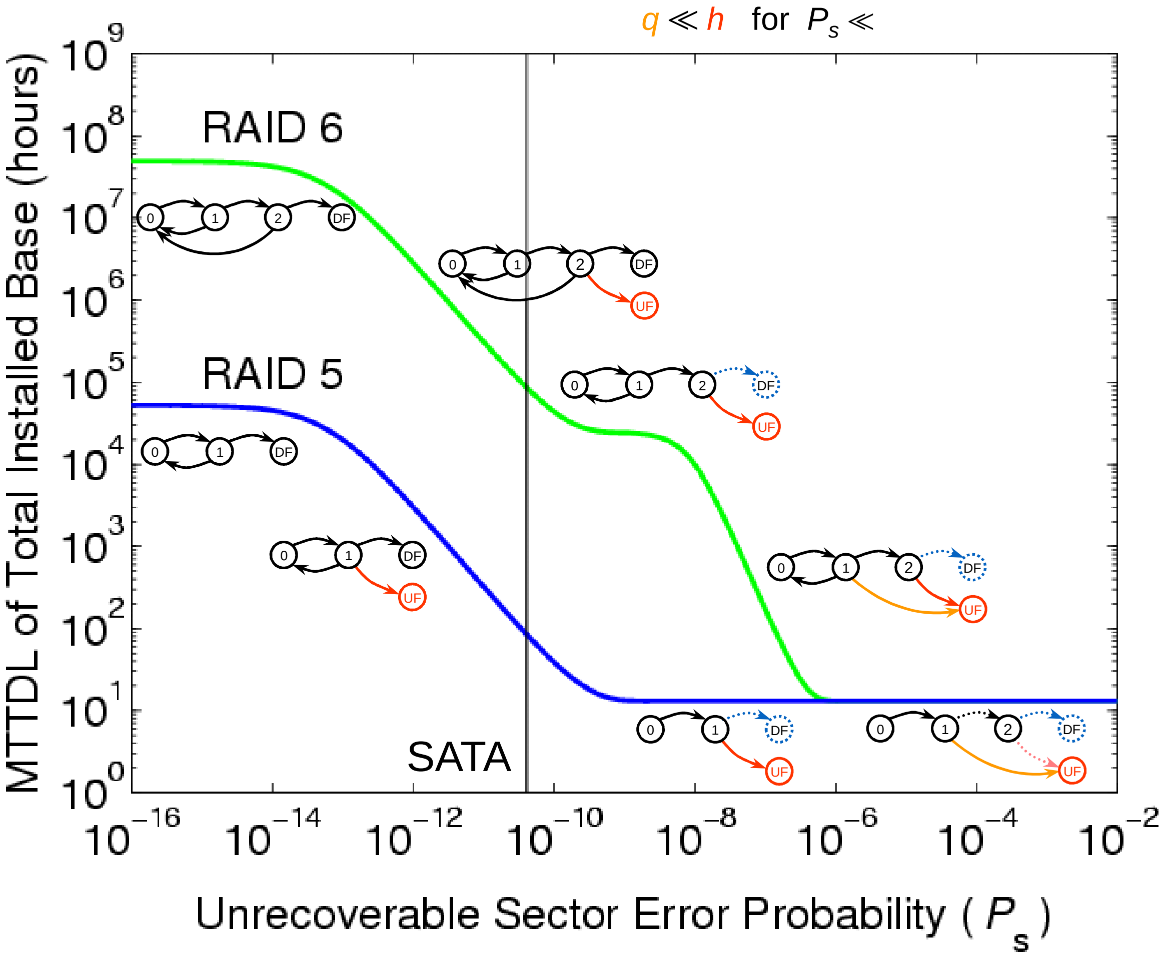
Assumptions:

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
$\Rightarrow P_s = 4096\times10^{-14} = 4.096\times10^{-11}$



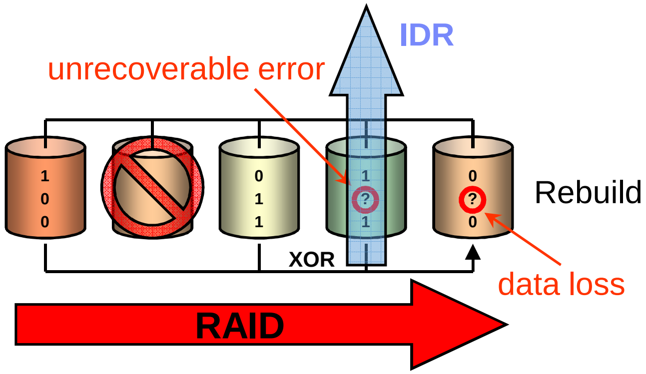


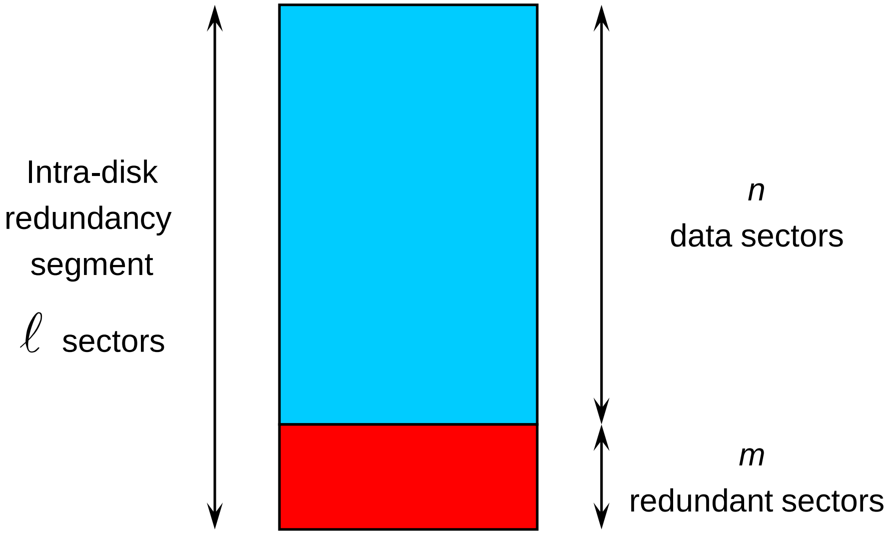
h : P(unrecoverable failure during rebuild in the critical mode)
q : P (unrecoverable failure during RAID 6 rebuild in the degraded mode)



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

Intra-Disk Redundancy (IDR) Scheme



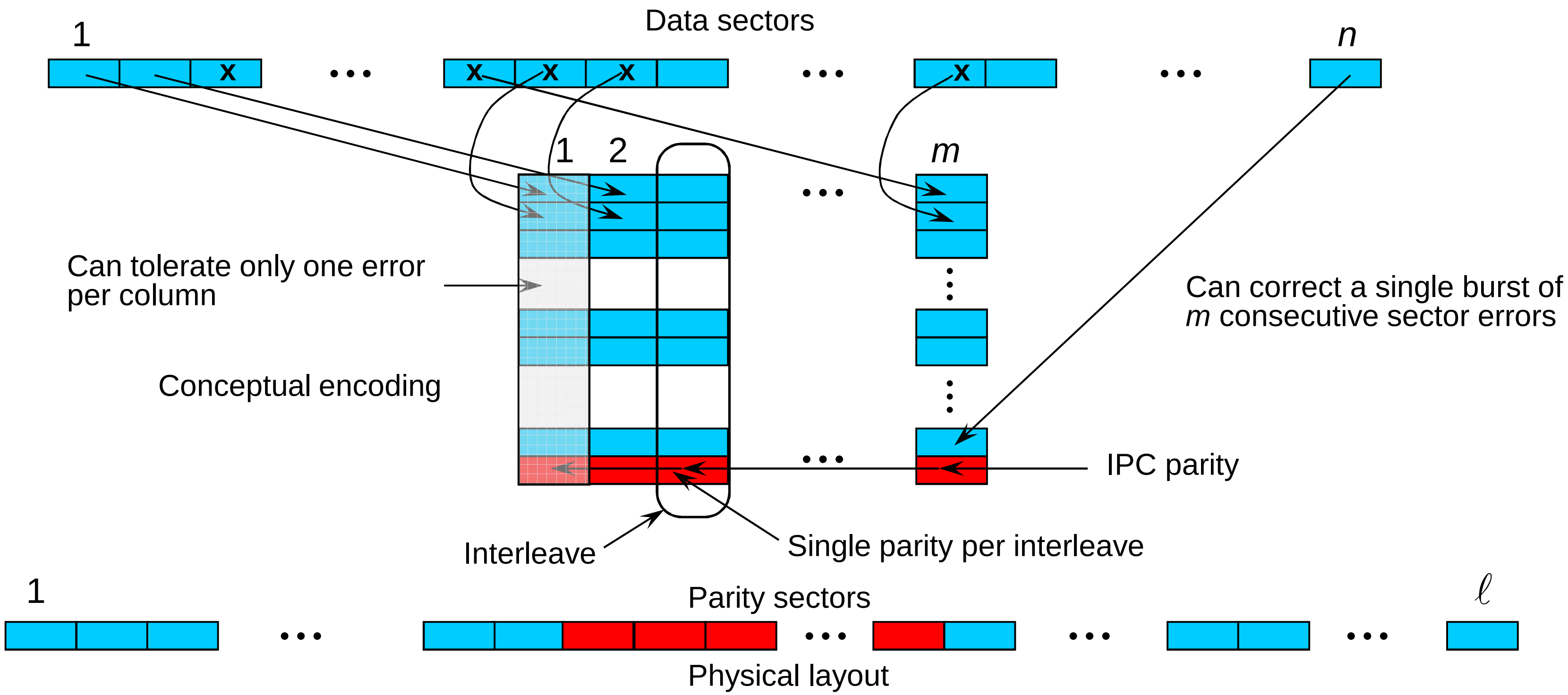


- Design concept:
  - For every ‘*n*’ data sectors, ‘*m*’ parity sectors are assigned
  - Redundant sectors are placed on the same disk drive as data
  - The ‘*m*’ parity sectors protect against uncorrectable media errors of any ‘*m*’ sectors in a group of ‘*n*’ sectors
- Intra-disk redundancy segment:
  - $\ell = n+m$ sectors
- Storage efficiency is $n/(n+m)$
- By choosing proper values of *n* and *m*, storage efficiency, performance and reliability can be optimized

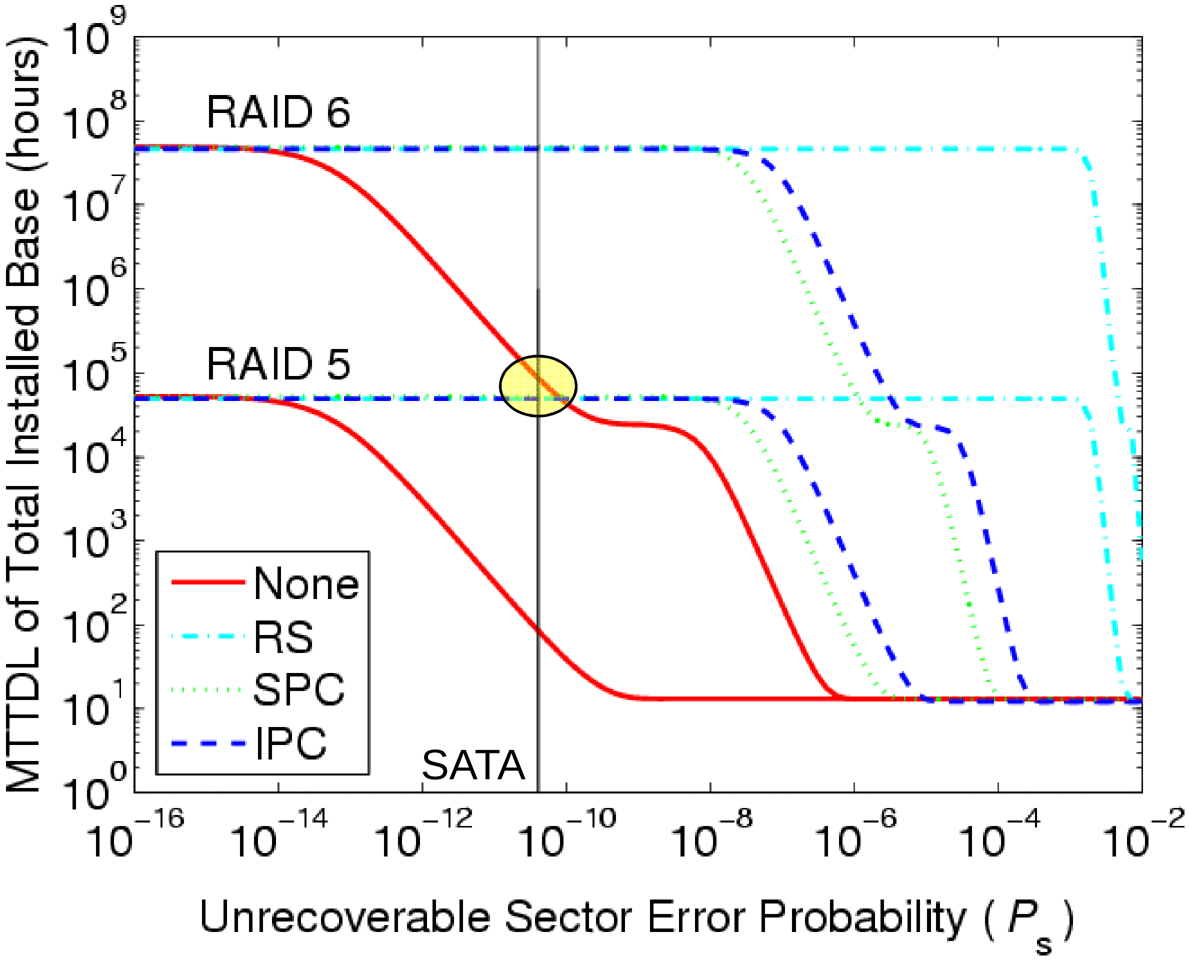
Dholakia *et al*., “A New Intra-disk Redundancy Scheme for High-Reliability RAID Storage Systems in the Presence of Unrecoverable Errors,” ACM Trans. Storage 2008

Interleaved Parity Check (IPC) Coding Scheme

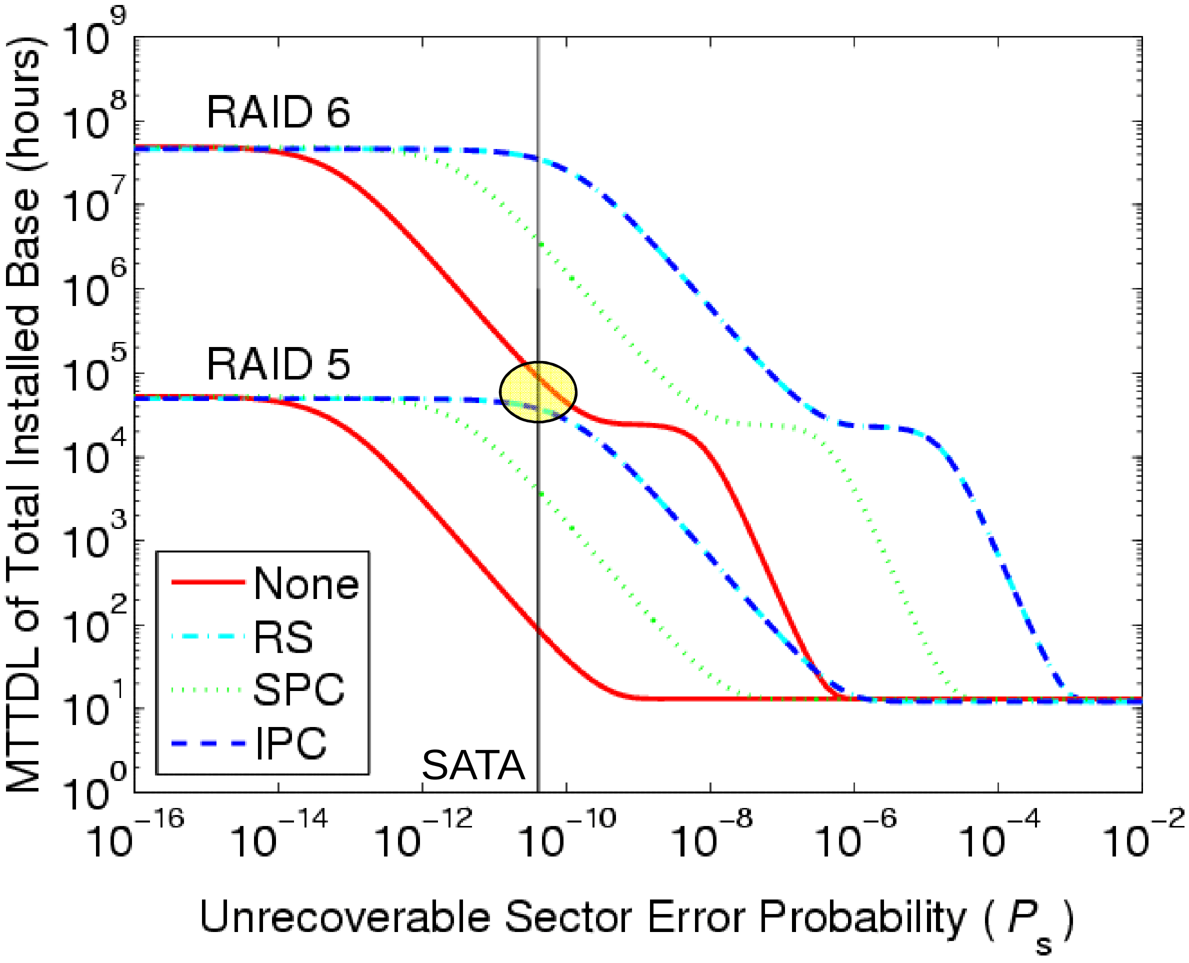
- Advantages
  - Easy to implement, using existing XOR engine
  - Flexible design parameters: segment size, efficiency
- Disadvantage
  - Not all erasure patterns can be corrected



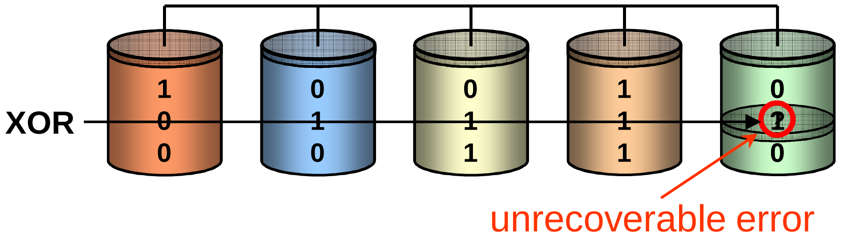
MTTDL for Independent Unrecoverable Sector Errors



MTTDL for Correlated Unrecoverable Sector Errors

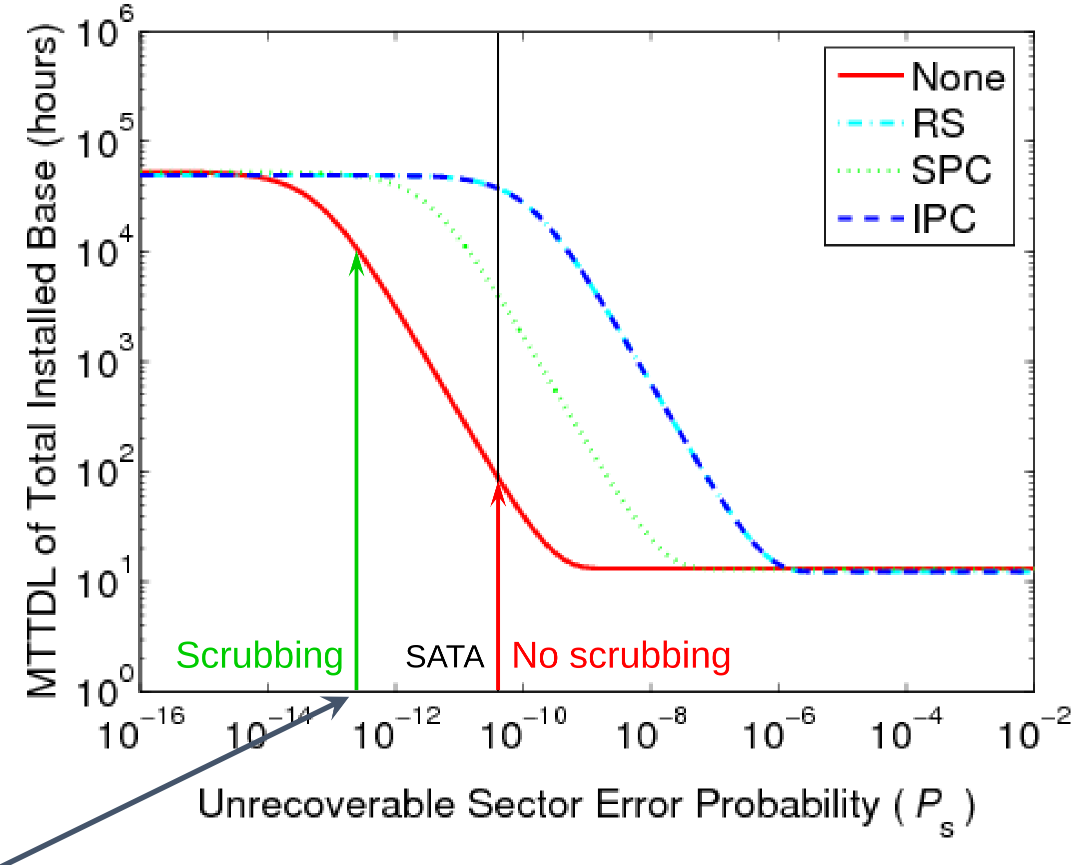


Disk Scrubbing



- Periodically accesses disk drives to detect unrecoverable errors
  - T_s : Scrubbing period = time required for a complete check of all sectors of a disk
- Identifies unrecoverable errors at an early stage
- Corrects the unrecoverable errors using the RAID capability
- Increases the workload because of additional read operations
- Sector write operations result in unrecoverable errors
  - $P_w = P$(sector-write operation results in an error)
    - ➢ Transition noise (media noise), “high-fly” write, off-track write
    - ➢ Contribution of thermal asperities and particle contamination ignored
- Disk-unrecoverable sector errors
  - are created by write operations and remain latent until read or successfully over-written
- Workload
  - h : load of a given data sector = rate at which sector is read/written
    - ➢ e.g. h=0.1 / day → 10% of the disk is read/written per day
  - r_w : ratio of write operations to read+write operations
    - ➢ typically 2/3

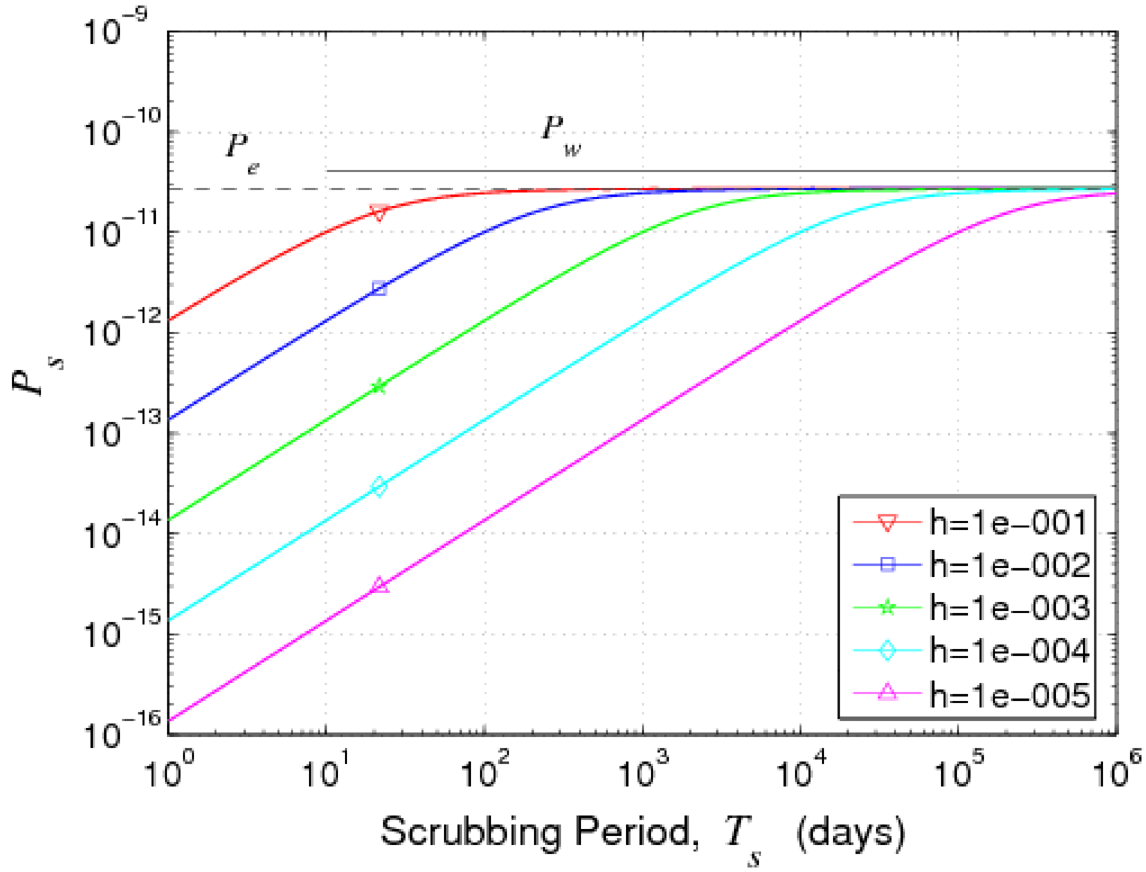
Modeling Approach



- derive P_s = P(sector error | scrubbing is used) = $f\,(T_s, P_w, h, r_w)$
- evaluate $MTTDL = f\,(P_s)$

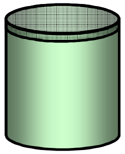
Analytical Results: Probability of Unrecoverable Sector Error

P_s : P(unrecoverable error on a tagged sector at an arbitrary time)



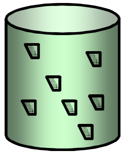
for SATA drives ($P_w = 4.096 \times 10^{-11}$)

- Without scrubbing: $P_s \rightarrow P_e = r_w P_w$
  - P_e depends on the ratio r_w of read/write operations, but not on the workload h
- Deterministic scrubbing scheme:



$$P_s = \left(1 - \frac{1 - e^{-hT_s}}{hT_s}\right) P_e$$

  - $P_s \leq P_e \leq P_w$
- Random scrubbing scheme:



$$P_s = \frac{hT_s}{1 + hT_s} P_e$$

  - P_s (deterministic) < P_s (random)
  - $hT_s \ll 1 \rightarrow$ P_s (deterministic) ≈ ½ P_s (random)

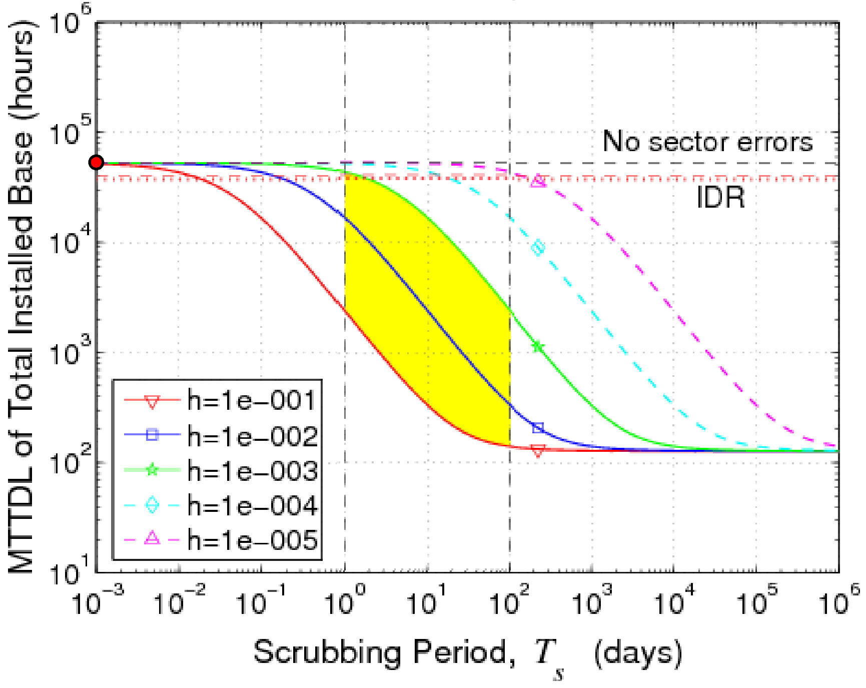
Iliadis *et al*., “Disk Scrubbing Versus Intradisk Redundancy for RAID Storage Systems,” ACM Trans. Storage 2011

Reliability Results for RAID-5 and RAID-6 Systems

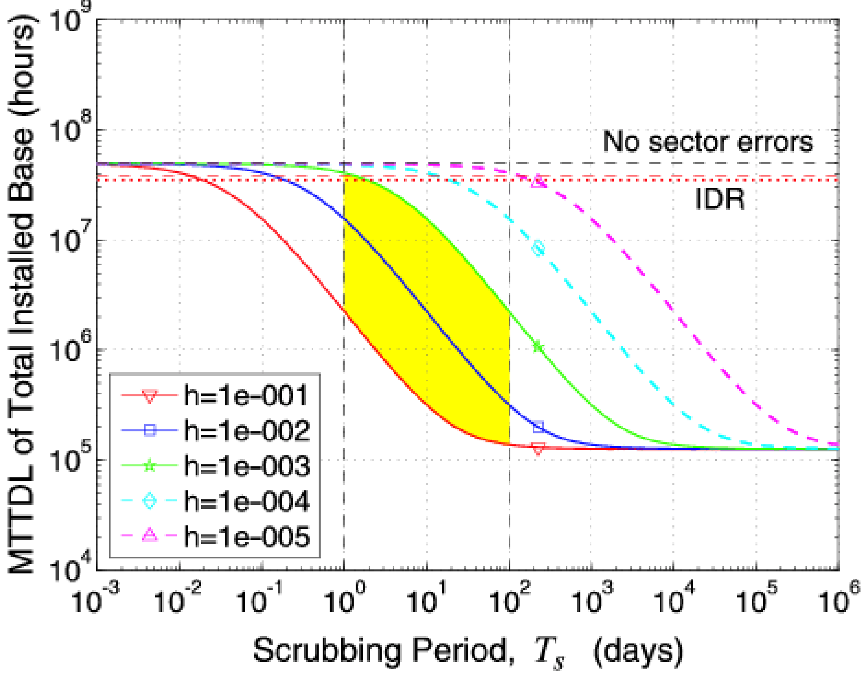
SATA disk drives: C_d = 300GB, MTTF_d = 500,000 h, MTTR=17.8 h, N=8 (RAID 5), N=16 (RAID 6)

MTTDL for an installed base of systems storing 10PB of user data

RAID 5



RAID 6

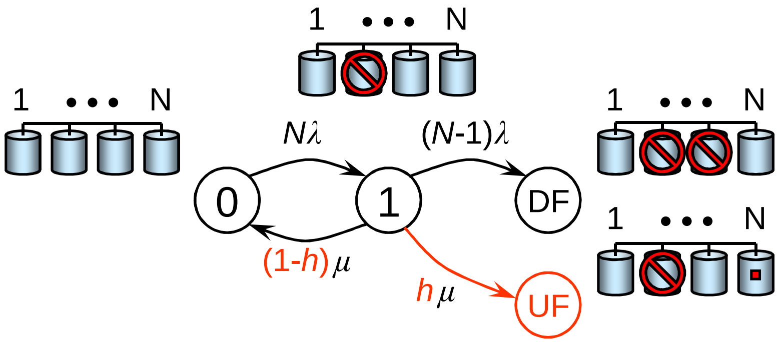


- The IDR scheme improves MTTDL by more than two orders of magnitude, which practically eliminates the negative impact of unrecoverable sector errors
- The scrubbing mechanism may not be able to reduce the number of unrecoverable sector errors sufficiently and reach the desired level of reliability

Enhanced MTTDL Equations for RAID Systems

- Latent or unrecoverable errors
    - ➢ P_s = P(sector error)

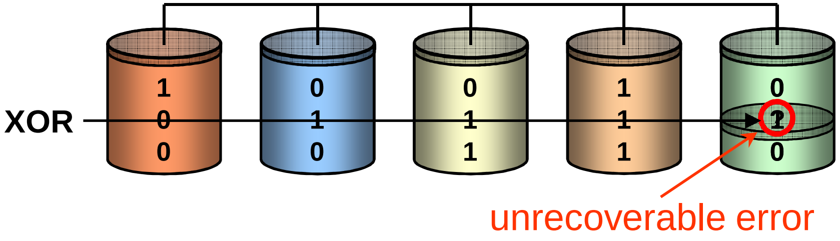
RAID 5:



Data loss owing to:

- DF: Disk Failure
- UF: Unrecoverable Failure

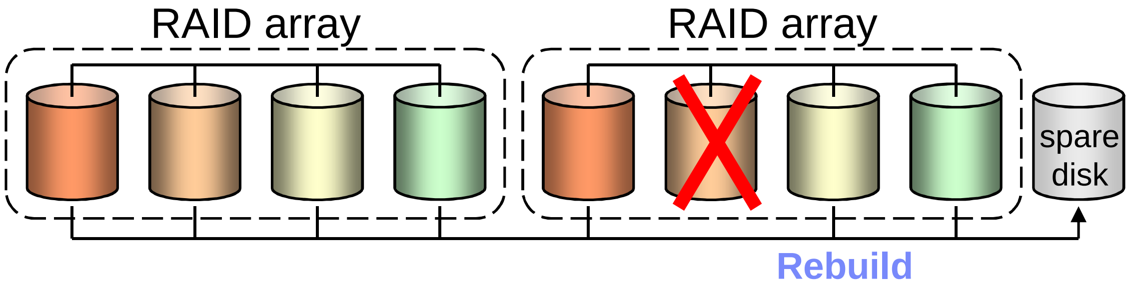
- Disk scrubbing
    - – Periodically accesses disk drives to detect unrecoverable errors



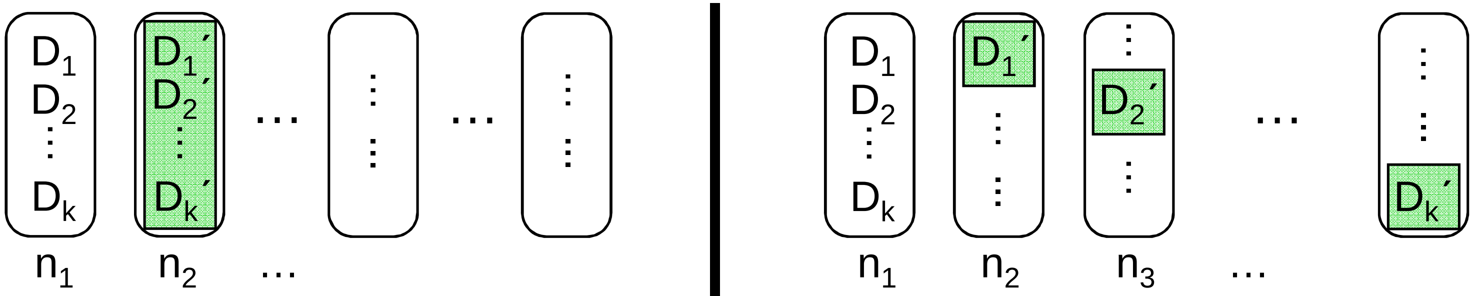
- – Identifies unrecoverable errors at an early stage
- – Corrects the unrecoverable errors using the RAID capability
    - ➢ P_s (equivalent) = P(sector error | scrubbing is used)

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

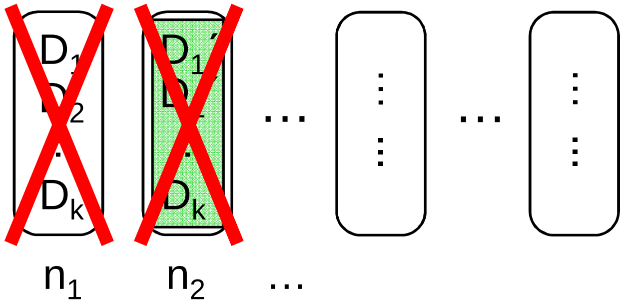
– V. Venkatesan et al. “A General Reliability Model for Data Storage Systems”, QEST 2012

General non-exponential failure and rebuild time distributions

- MTTDL is insensitive to the failure time distributions; it depends only on the mean value

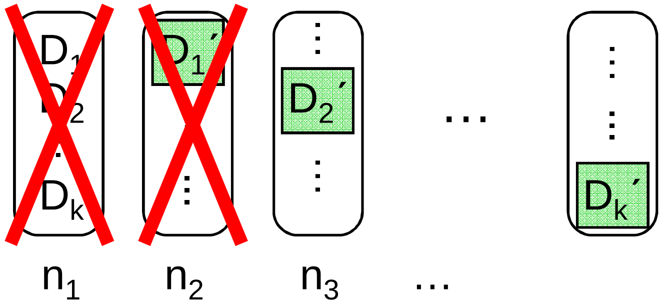
Time To Data Loss vs. Amount of Data Lost

- 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



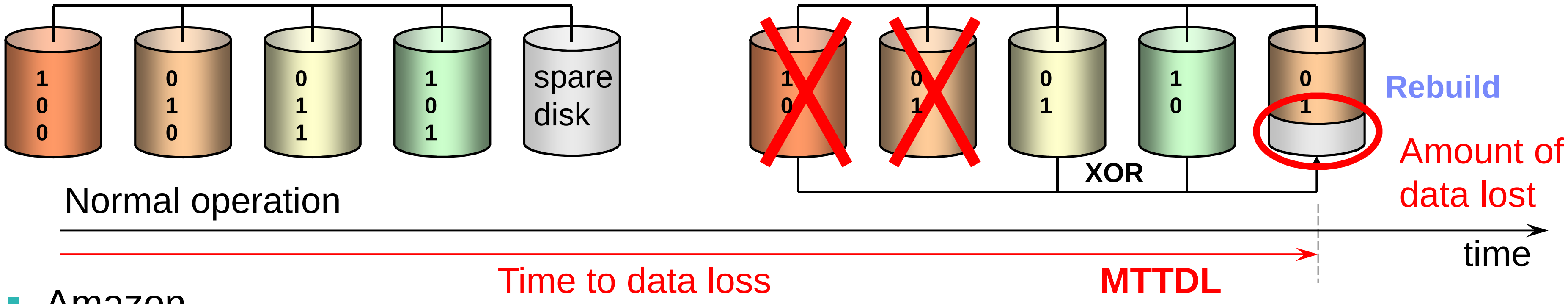
- on different nodes

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

Expected Annual Fraction of Data Loss (EAFDL)

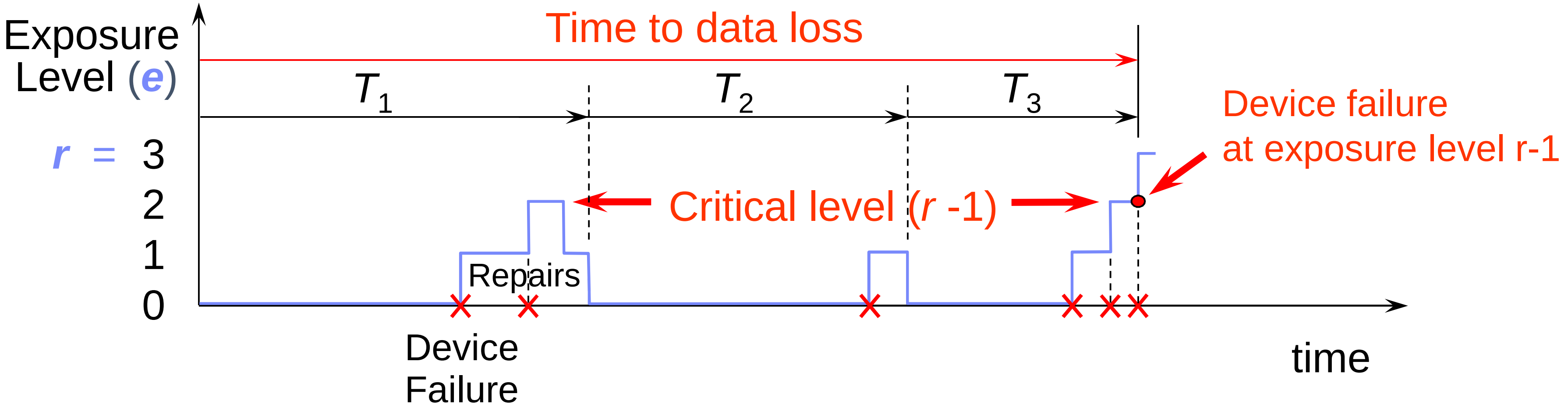


- Amazon
  - The Reduced Redundancy Storage option within Amazon S3 is designed to provide 99.999999999% durability of objects over a given year
    - ➢ average annual expected loss of a fraction of 10^{-11} of the data stored in the system
- Data loss events documented in practice by Yahoo!, LinkedIn, and Facebook
- Assess the implications of system design choices on the
  - frequency of data loss events
    - ➢ MTTDL
  - amount of data lost
    - ➢ **Expected annual fraction of data loss (EAFDL)**
      - Fraction of stored data that is expected to be lost by the system annually
    - ➢ EAFDL metric is meant to complement, not to replace MTTDL
  - These two metrics provide a useful profile of the magnitude and frequency of data losses
    - ➢ for storage systems with similar EAFDL
      - ✓ most preferable the one with the maximum MTTDL

Previous Work on Storage Reliability

Reliability Measure	Theory / Analysis	Simulation
MTTDL	▪ **Markov models** – Original RAID-5 and RAID-6 MTTDL equations – Enhanced MTTDL Equations ➢ Latent or unrecoverable errors ➢ Scrubbing operations ▪ **Non-Markov-based models** – General non-exponential failure and rebuild time distributions – Placement schemes – Network bandwidth, Latent errors, Erasure codes	Non-Markov-based MTTDL simulations
Other Metrics	I. Iliadis and V. Venkatesan, “Expected Annual Fraction of Data Loss as a Metric for Data Storage Reliability” IEEE MASCOTS September 2014 ?	▪ Normalized Magnitude of Data Loss (NOMDL) ▪ Fraction of Data Loss Per Year (FDLPY)* * equivalent to EAFDL

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

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

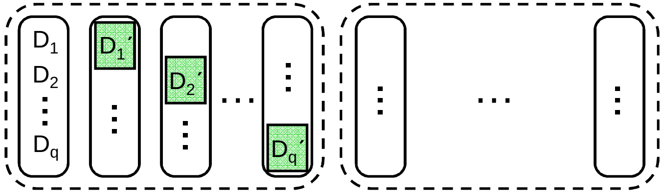
MTTDL / EAFDL equations obtained using non-Markov Analysis

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

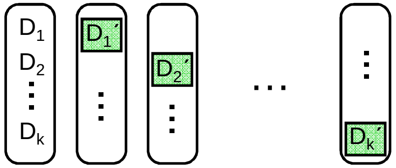
- ➢ increasing failure rates over time
  - shape parameter = 1.5

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

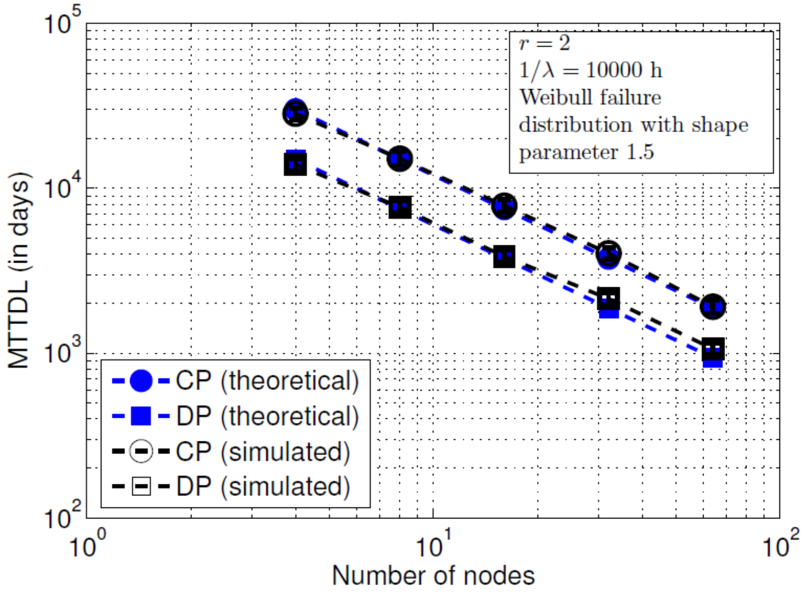


Symmetric placement

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



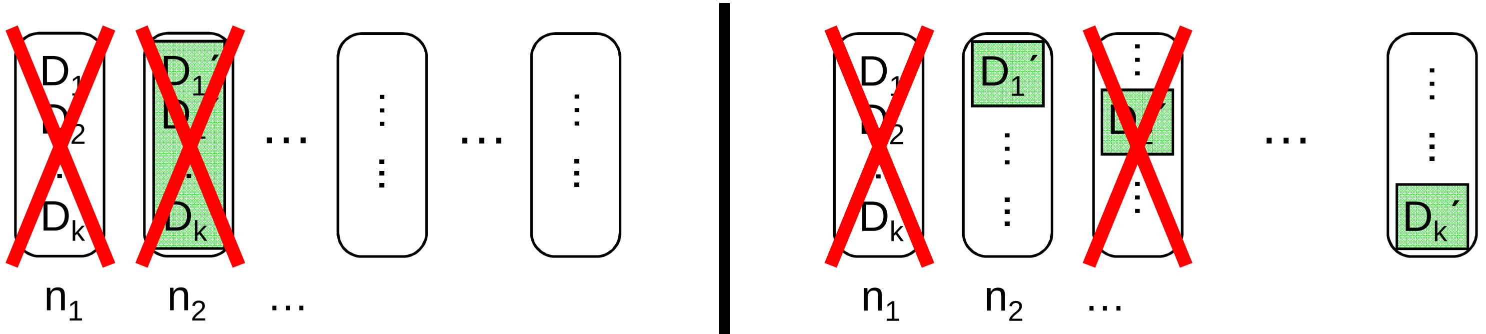
Reliability Results for Replication Factor of 2



- 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

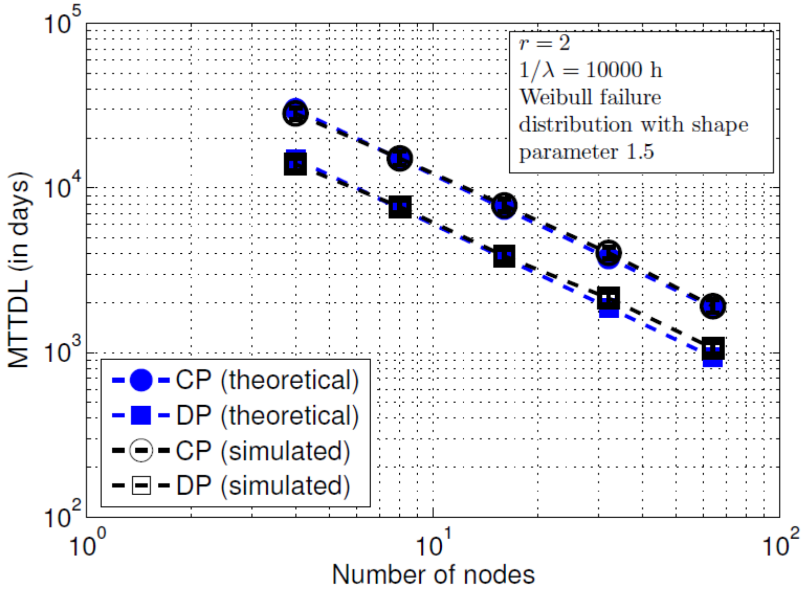
Clustered Placement

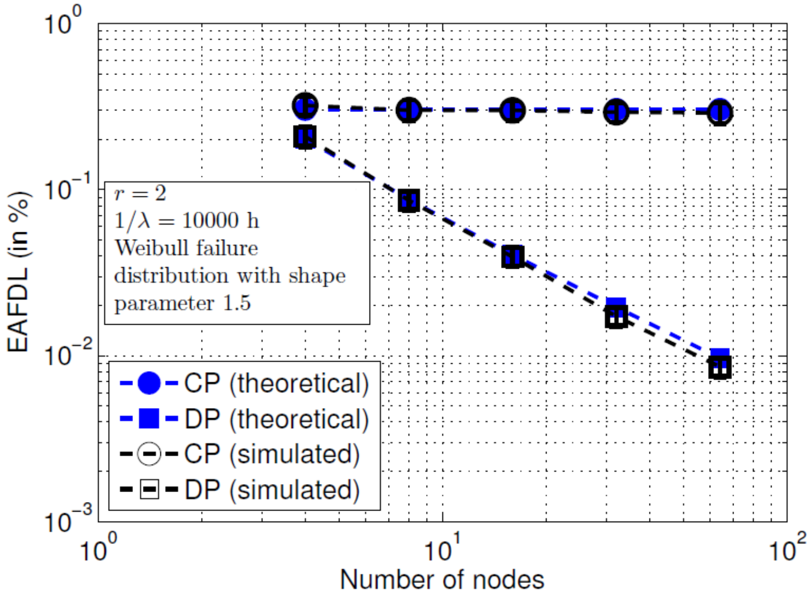
- on different nodes

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

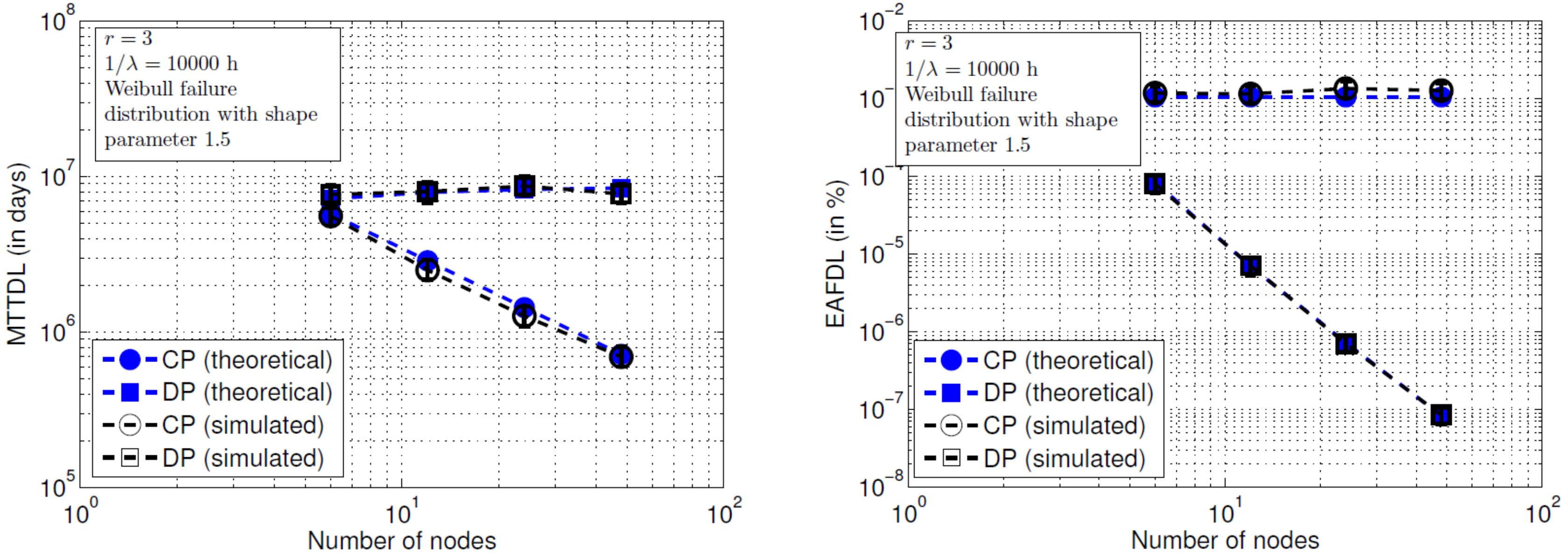




- MTTDL
  - Declustered placement not better than clustered one

- 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

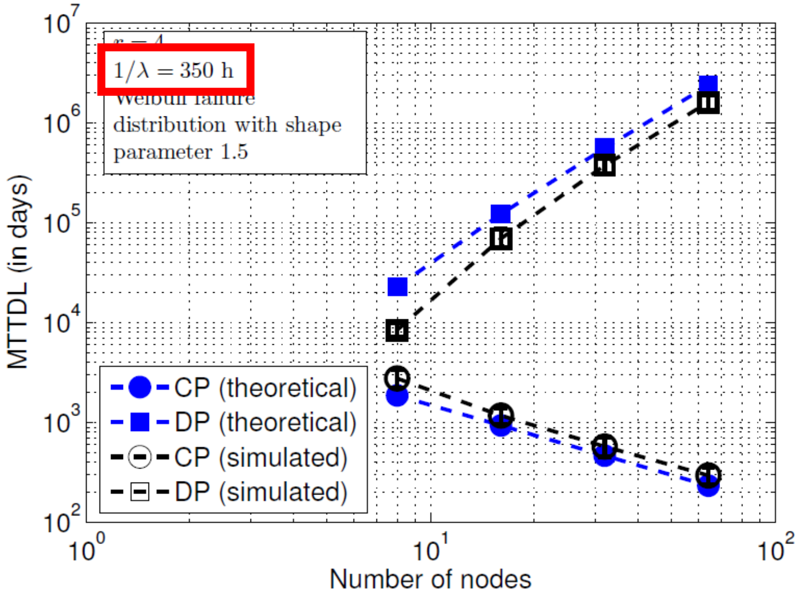
Reliability Results for Replication Factor of 3

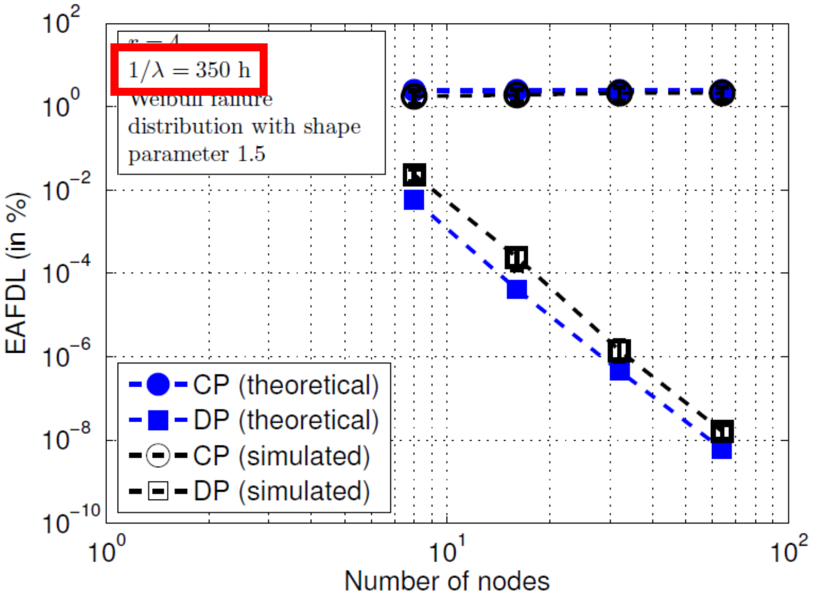


- MTTDL
  - Inversely proportional to the number of nodes for clustered placement
  - Independent of the number of nodes for declustered placement
    - ➢ Declustered placement better than clustered one

- 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

Reliability Results for Replication Factor of 4



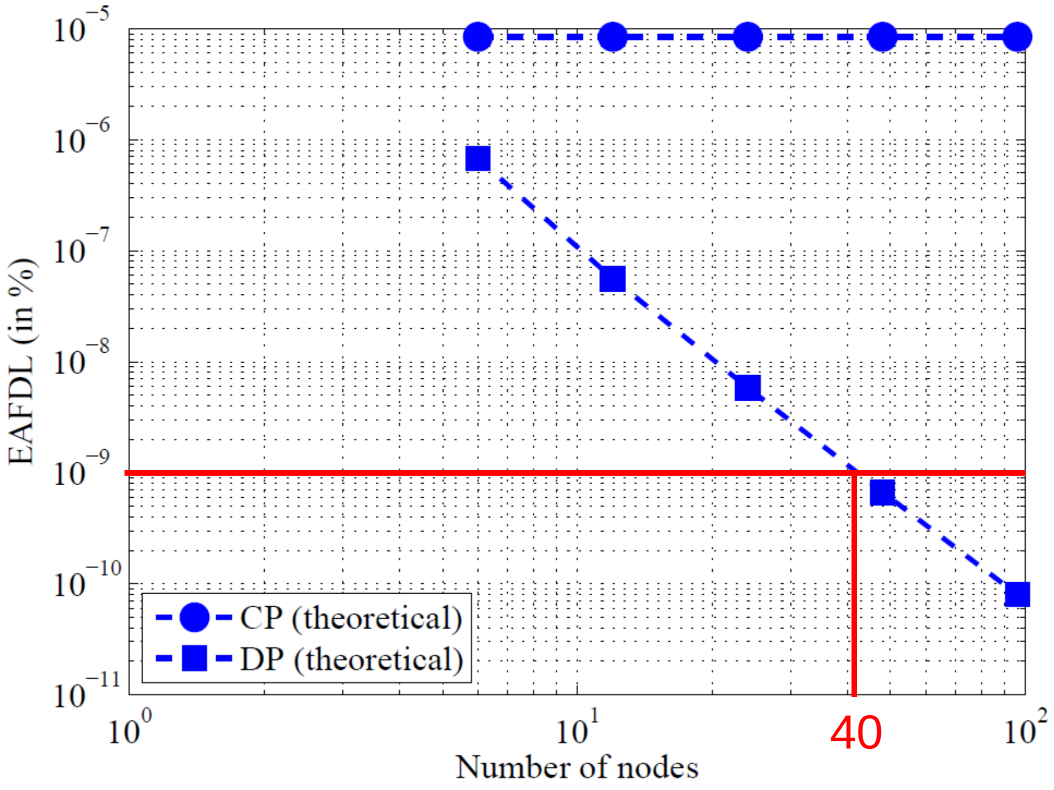


MTTR/MMTF ratio: 34.7/350 $\simeq$ 0.1 not very small $\Rightarrow$ Deviation between theory and simulation

- MTTDL
  - Proportional to the square of the number of nodes for declustered placement
    - Declustered placement far superior to the clustered one
- EAFDL
  - Inversely proportional to the sixth power of the number of nodes for declustered placement
    - Declustered placement far superior to the clustered one

Theoretical EAFDL Results for Replication Factor of 3

MTTF = 1/λ = 50,000 h



- 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

Summary

- Reviewed the widely used mean time to data loss (MTTDL) metric
- Demonstrated that unrecoverable errors are becoming a significant cause of user data loss
- Considered the expected annual fraction of data loss (EAFDL) metric
- Established that the EAFDL metric, together with the traditional MTTDL metric
  - provide a useful profile of the magnitude and frequency of data losses
  - can be jointly evaluated analytically in a general theoretical framework
- Derived the MTTDL/EAFDL in the case of replication-based storage systems that use clustered and declustered data placement schemes and for a
  - ➢ large class of failure time distributions
    - real-world distributions, such as Weibull and gamma
- Demonstrated the superiority of the declustered placement scheme

Future Work

- Apply the methodology developed to derive the reliability of systems using other redundancy schemes, such as erasure codes