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 (Redundant Array of Independent Disks)
    - 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 (**Redundant Array of Independent Disks**)
    - RAID-5: Tolerates one disk failure
    - RAID-6: Tolerates two disk failures

![Diagram showing RAID-5 and RAID-6 data reconstruction](image)
Time to Failure and MTTDL

Reliability Metric: **MTTDL** (Mean Time to Data Loss)

➢ Continuous Time Markov Chain Models

**RAID 5:**

- $\lambda$: \(1/\text{MTTF for disks}\)
- $\mu$: \(1/\text{MTTR}\)

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

[Paterson et al. 1988]

**RAID 6:**

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

[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)

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

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

Data loss owing to:
- **DF:** Disk Failure
- **UF:** Unrecoverable Failure

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

\[
h = (N - 2)C_dP_s + O(P_s^2)
\]

\[
q = \left(\frac{N - 1}{2}\right)C_dP_s^2 + O(P_s^3)
\]

$q \ll h$ 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}

\[ (1-h) \mu q \mu (1-q) \mu h \mu (N) \lambda (N-1) \lambda (N-2) \lambda \]

\[ P_s \ll q \quad \text{for} \quad P_s \ll \]
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

- **Non-Markov-based analysis**

General non-exponential failure and rebuild time distributions

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

---

Enhancing the Reliability of Large-Scale Data Storage Systems
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 \( D_1, D_2, \ldots, D_k \) is placed:

        ![Diagram showing clustered and declustered placement]

        - 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
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

\[
\text{MTTDL} \approx \sum_{i=1}^{m} E(T_i) \approx \frac{E(T)}{P_{DL}} \quad \text{EAFDL} = \frac{E(Q)}{E(T) \cdot U}
\]

MTTDL and EAFDL expressions obtained using non-Markov Analysis
Theoretical Results

- \( n \): number of storage devices
- \( c \): amount of data stored on each device
- \( r \): replication factor
- \( b \): reserved rebuild bandwidth per device
- \( 1/\lambda \): mean time to failure of a storage device

\[
\text{MMTDL} \approx \left\{ \frac{\left( b \right)^{r-1}}{\left( \lambda c \right)^{r-1}} \frac{1}{n \lambda}, \left( b \right)^{r-1} \left( \frac{\lambda c}{2} \right)^{r-1} \frac{1}{n \lambda} \right\} \prod_{e=1}^{r-2} \left( \frac{n-e}{r-e} \right) \left( \frac{r-e-1}{n} \right) \\
\text{EAFDL} \approx \left\{ \frac{\left( \frac{\lambda c}{b} \right)^{r-1}}{\left( \frac{2 \lambda c}{b} \right)^{r-1}} \frac{\lambda}{(r-1)!} \prod_{e=1}^{r-1} \left( \frac{r-e-1}{n-e} \right)^{r-e} \right\} \\
\text{Symmetric placement}
\]

4 to 64
12 TB
2, 3, 4
96 MB/s
10,000 h - Weibull distributions with shape parameters greater than one
- increasing failure rates over time
  - shape parameter = 1.5
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

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

\[c = 12 \text{ TB}\]
\[b = 96 \text{ MB/s}\]
\[\text{MTTR} = 35 \text{ h}\]
\[\text{MTTF} = \frac{1}{\lambda} = 50,000 \text{ h}\]
\[\frac{\text{MTTR}}{\text{MTTF}} = 0.0007\]
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


- 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

![Erasure Coded Schemes Diagram]

- 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 : \(s_{\text{eff}} = l / m\)

Facebook : Reed-Solomon (14,10) \(\rightarrow\) \(s_{\text{eff}} = 71\%\)
Windows Azure : Reed-Solomon (16,12) \(\rightarrow\) \(s_{\text{eff}} = 75\%\)
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


- \( n \): 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

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

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

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

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

- For fixed storage efficiency $s_{eff}$
  - $n$: Number of storage devices
  - $1/\lambda$: Mean Time to Failure (MTTF) for a device
  - $1/\mu$: Time to read the data of a device

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

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

- 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

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

$\lambda/\mu = \lambda c/b = 0.001$
Network Rebuild Bandwidth Constraints

Declustered Placement

\[ B_{\text{eff}} = \min(\tilde{k}b, B_{\text{max}}) \]

Clustered Placement

\[ B_{\text{eff}} = \min(lb, B_{\text{max}}) \]

Distributed rebuild from \( \tilde{k} \) devices

Rebuild from \( l \) devices

Reserved spare space

Spare device

A1  B1  C1  D1  E1  F1  G1  H1  I1  J1

A1  B1  C1  D1  E1  F1  G1  H1  I1  J1

B1  B1  B1  B1  B1  B1  B1  B1  B1  B1

C1  C1  C1  C1  C1  C1  C1  C1  C1  C1

D1  D1  D1  D1  D1  D1  D1  D1  D1  D1

E1  E1  E1  E1  E1  E1  E1  E1  E1  E1

F1  F1  F1  F1  F1  F1  F1  F1  F1  F1

G1  G1  G1  G1  G1  G1  G1  G1  G1  G1

H1  H1  H1  H1  H1  H1  H1  H1  H1  H1

I1  I1  I1  I1  I1  I1  I1  I1  I1  I1

J1  J1  J1  J1  J1  J1  J1  J1  J1  J1
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