

Copysets : Reducing the Frequency of Data Loss in Cloud Storage

Wei-Chih Chien
60347046S

NTNU CSIE LAB107

Author & Reference

Author:

A. Cidon S. Rumble R. Stutsman
S. Katti J. Ousterhout M. Rosenblum
(Stanford University)

Source:

2013 USENIX Annual Technical Conference
(Awarded Best Student Paper)

¹ www.usenix.org/conference/atc13/technical-sessions/presentation/cidon

Outline

- 1 Introduction
 - Random Replication
 - Copysets Replication
- 2 Intuition
 - Probability of data loss
 - The Trade-off
- 3 Design
- 4 Related Work
- 5 Conclusion

Random Replication

Widely used in data center storage systems to prevent data loss.

- Hadoop Distributed File System (HDFS)
- RAMCloud (<https://ramcloud.stanford.edu>)
- Google File System (GFS)
- Windows Azure

However, large-scale correlated failures such as **cluster power outages** handled poorly by random replication.^{[1][2][3][4]}

This stresses the **availability** of the system.

¹R. J. Chansler. Data Availability and Durability with the Hadoop Distributed File System.

²J. Dean. Evolution and future directions of large-scale storage and computation systems at Google.

³D. Ford et al. Availability in globally distributed storage systems.

⁴K. Shvachko et al. The hadoop distributed file system.

Copysets Replication

- Split node into **copysets**
- Replicas of single chunk can only be stored on **one copyset**.
- Data loss events occur only when all the nodes of some copyset fail **simultaneously**.
- **Decrease** the probability of data loss under power outages.

Outline

- 1 Introduction
 - Random Replication
 - Copysets Replication
- 2 **Intuition**
 - Probability of data loss
 - The Trade-off
- 3 Design
- 4 Related Work
- 5 Conclusion

Probability of data loss

- N : # nodes in the system
- R : # replicas of each chunk

$$\frac{\#copyset}{\binom{N}{R}}$$

Example :

$\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}$

$$N = 9$$

$$R = 3$$

$$\# \text{ copysets} = 3$$

Random Replication

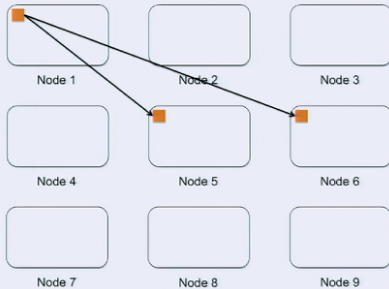


Figure : 1

Random Replication

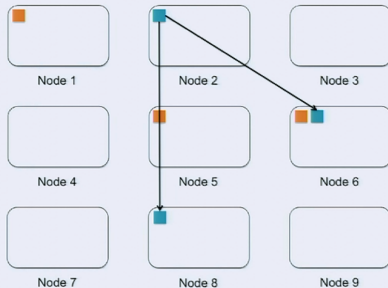
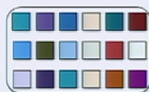


Figure : 2

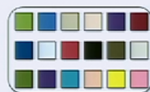
Random Replication



Node 1



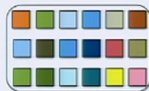
Node 2



Node 3



Node 4



Node 5



Node 6



Node 7



Node 8



Node 9

{1, 5, 6}

{2, 6, 8}

{3, 4, 5}

...

{5, 6, 9}

MinCopolysets



Node 1



Node 2



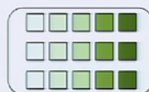
Node 3



Node 4



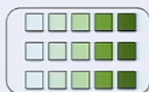
Node 5



Node 6



Node 7



Node 8



Node 9

{1, 5, 7}

{2, 4, 9}

{3, 6, 8}

Probability of data loss

- N : # nodes in the system
- R : # replicas of each chunk

$$\frac{\#copyset}{\binom{N}{R}}$$

Example :

$\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}$

$$N = 9$$

$$R = 3$$

$$\# \text{ copysets} = 3$$

The Trade-off

	MinCpysets	Random Replication
Mean time to Failure	625 years	1 year
Amount of Data Lost	1 TB	5.5 GB

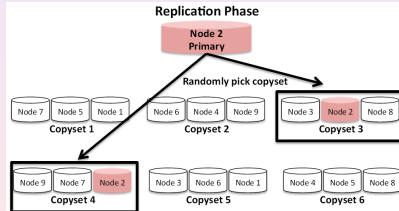
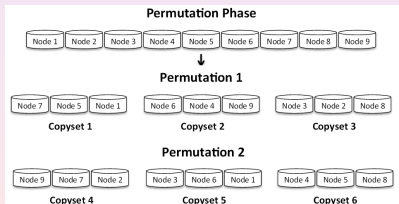
5000-node cluster

Outline

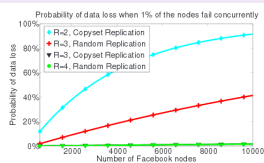
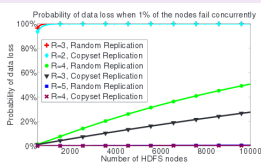
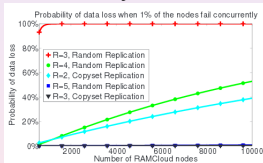
- 1 Introduction
 - Random Replication
 - Copysets Replication
- 2 Intuition
 - Probability of data loss
 - The Trade-off
- 3 **Design**
- 4 Related Work
- 5 Conclusion

Design

- 2 phases : **Permutation** & **Replication**.
- Scatter width : # nodes that store copies for each nodes data.



Data loss probability of random replication and Copyset Replication in different systems.



Outline

- 1 Introduction
 - Random Replication
 - Copysets Replication
- 2 Intuition
 - Probability of data loss
 - The Trade-off
- 3 Design
- 4 Related Work**
- 5 Conclusion

Related Work

- BIBD. (Balanced Incomplete Block Designs) [Fisher, '40]
- Power downs. [Harnik et al '09, Leverich et al '10, Thereska '11]
- Multi-fabric interconnects. [Mehra, '99]

Outline

- 1 Introduction
 - Random Replication
 - Copysets Replication
- 2 Intuition
 - Probability of data loss
 - The Trade-off
- 3 Design
- 4 Related Work
- 5 Conclusion**

Conclusion

- 1 Many Storage systems **randomly** spray their data across a large number of nodes.
- 2 Serious problem with **correlated failures**.
- 3 **Copyset Replication** is a better way of spraying data that **decreases the probability** of correlated failures.

Thank You for Your Listening

