Data Storage Institute

A Design of Metadata Server Cluster in Large Distributed Object-based Storage

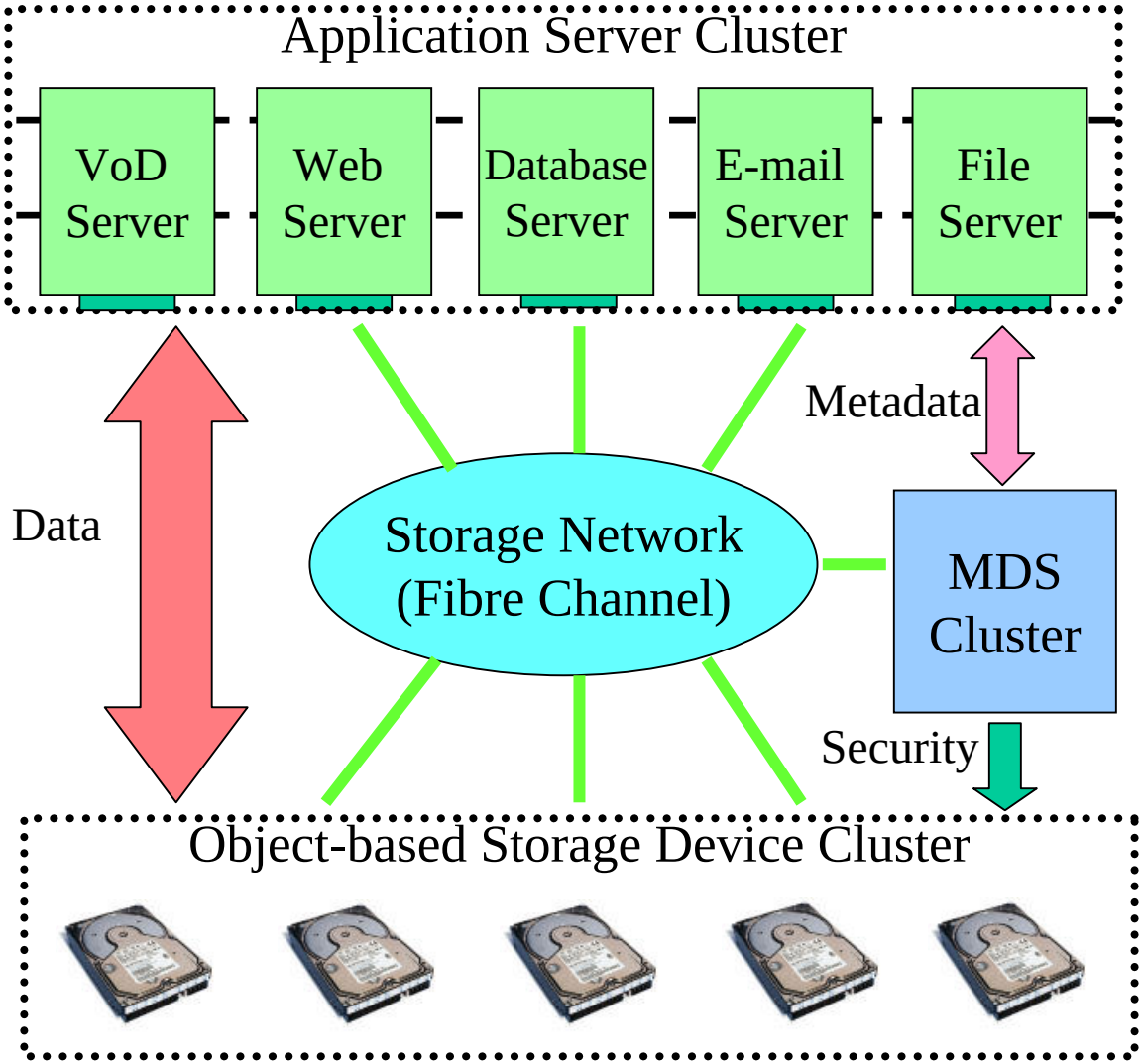
Yan Jie, Jeffrey

Tel:+0065-68748158

e_mail: yan_jie@dsi.a-star.edu.sg

NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

Context



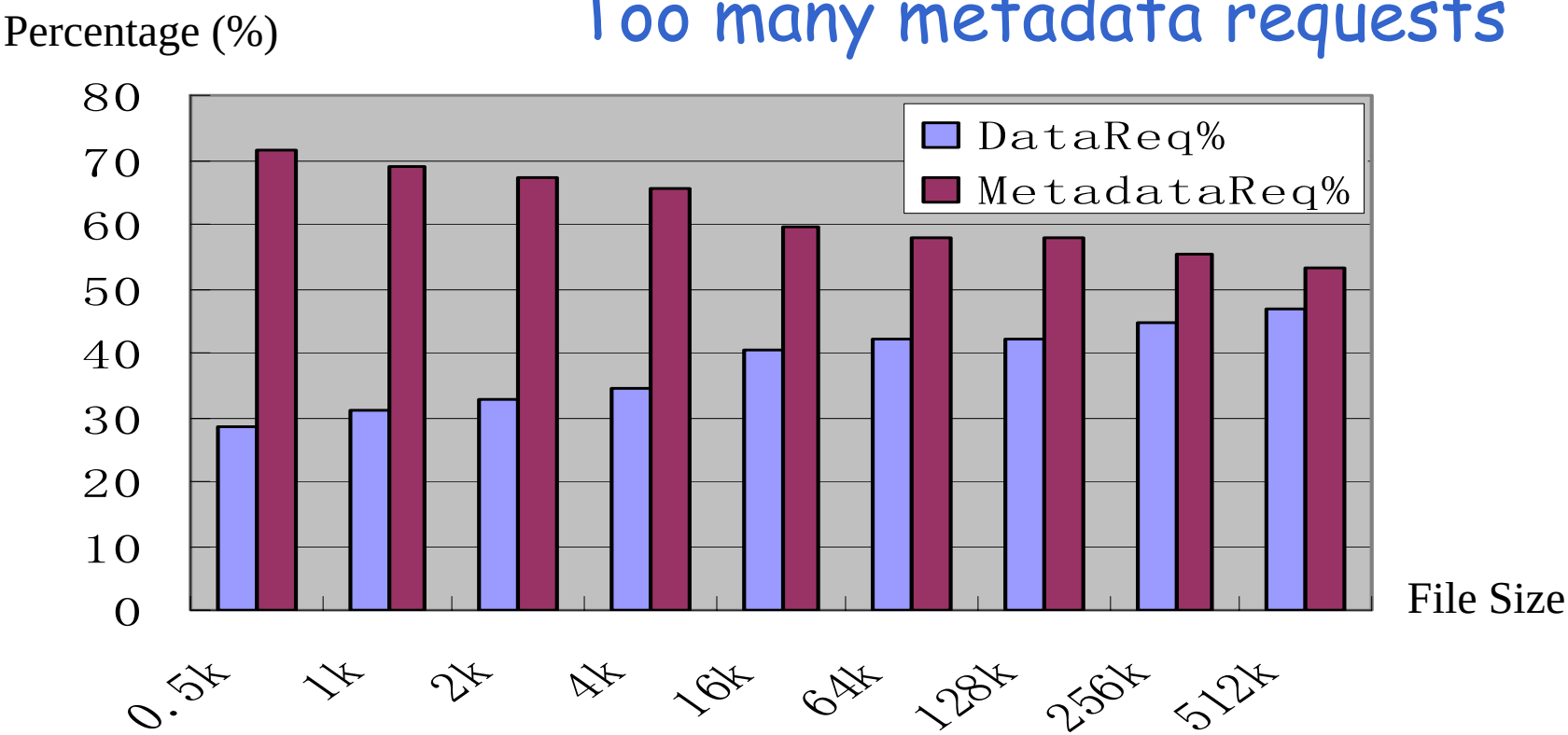
BrainStor

Problem

Data Storage Institute

A*STAR

Too many metadata requests

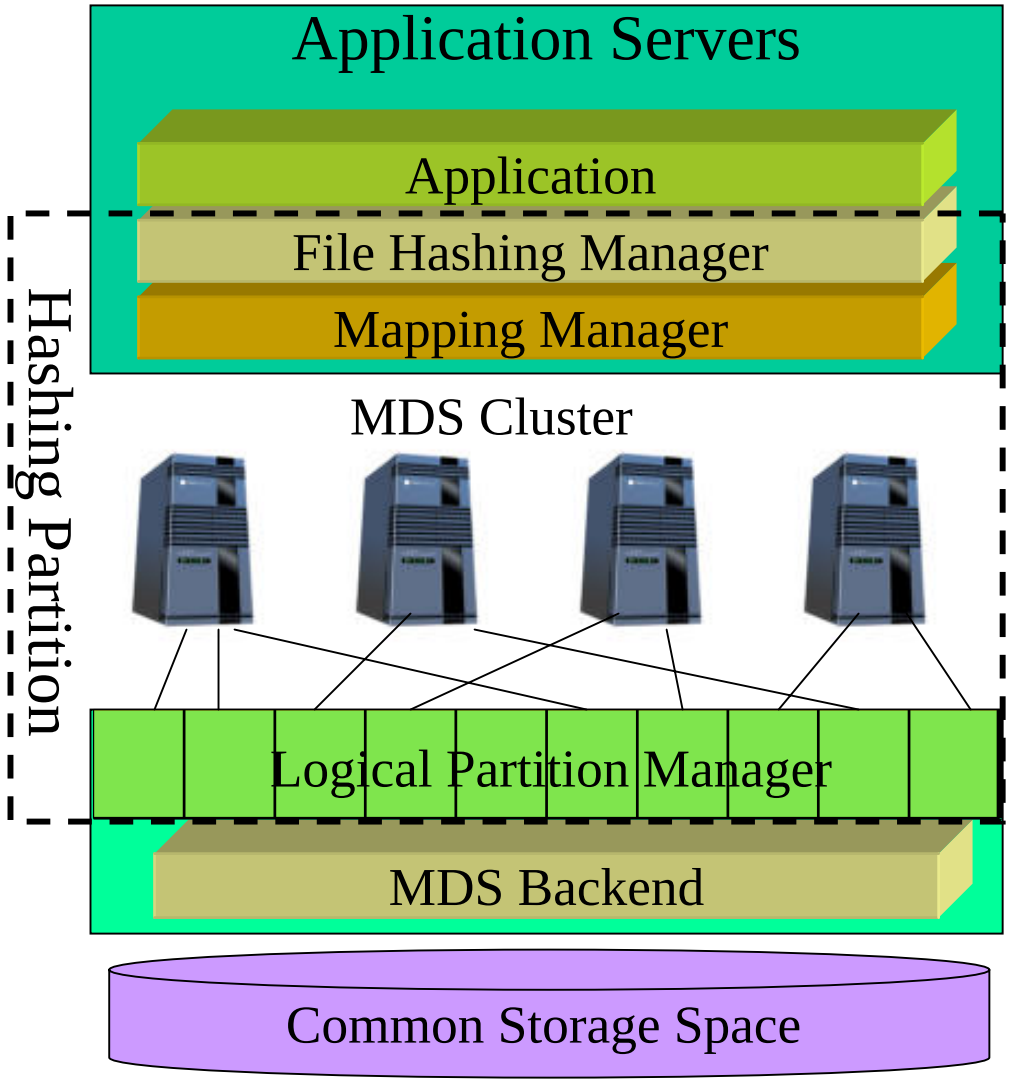


This figure shows the data request percent (DataReq%) and the metadata request percent (MetadataReq%) of the total requests. This test is based on our **BrainStor** prototype (one client, one MDS and one Object Storage Module) connected by 2G Fibre Channel, using Postmark (1000 files, 10 subdirectories, random access, 500 transactions).

Solution: Hashing Partition (HAP)

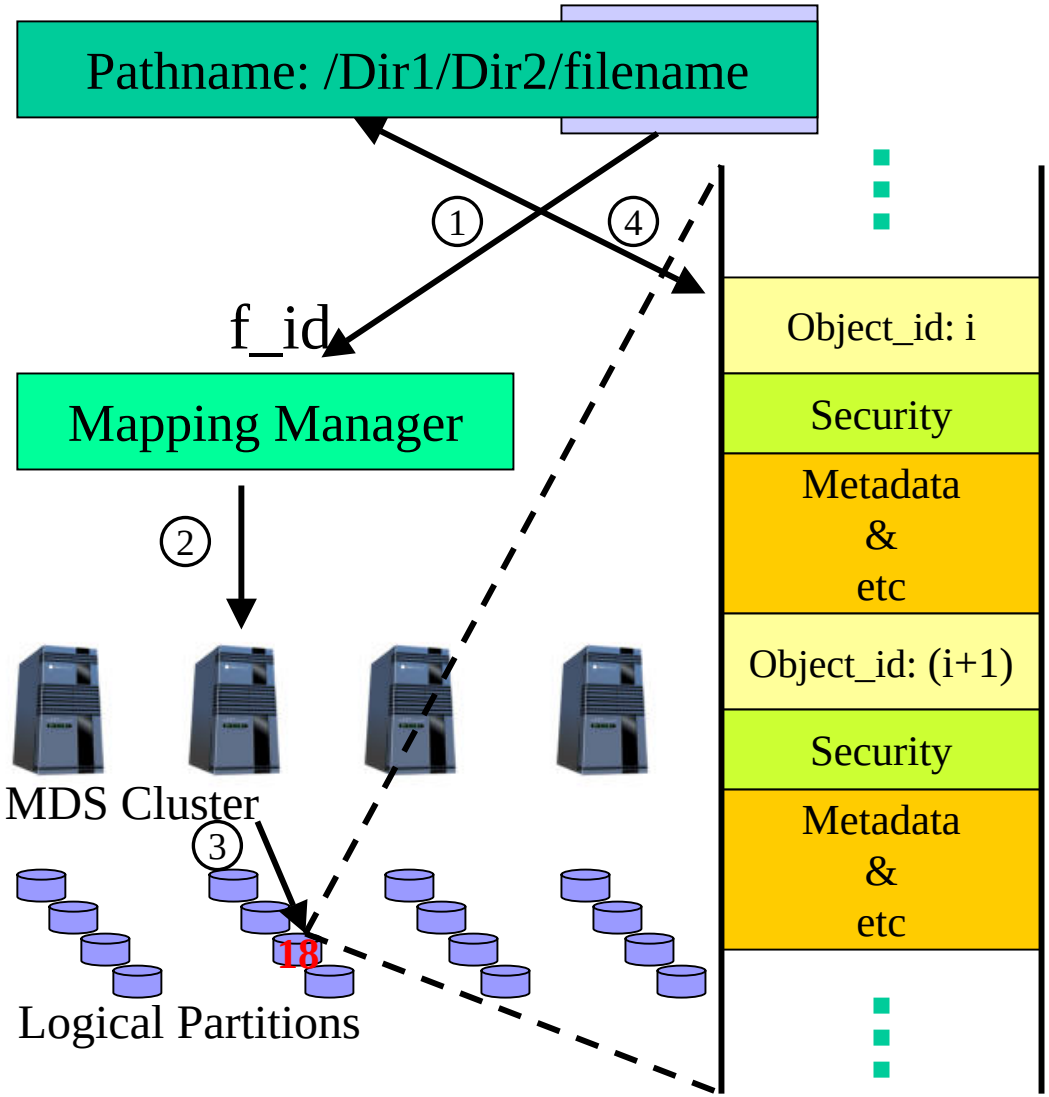
Data Storage Institute

- **Features:**
- 1. Hashing method instead of directory subtree management, to reduce the number of metadata requests
- 2. Common Storage Space (divided into logical partitions) to facility Load balancing, Scalability and Failover design



HAP for BrainStor

Data Storage Institute
A*STAR

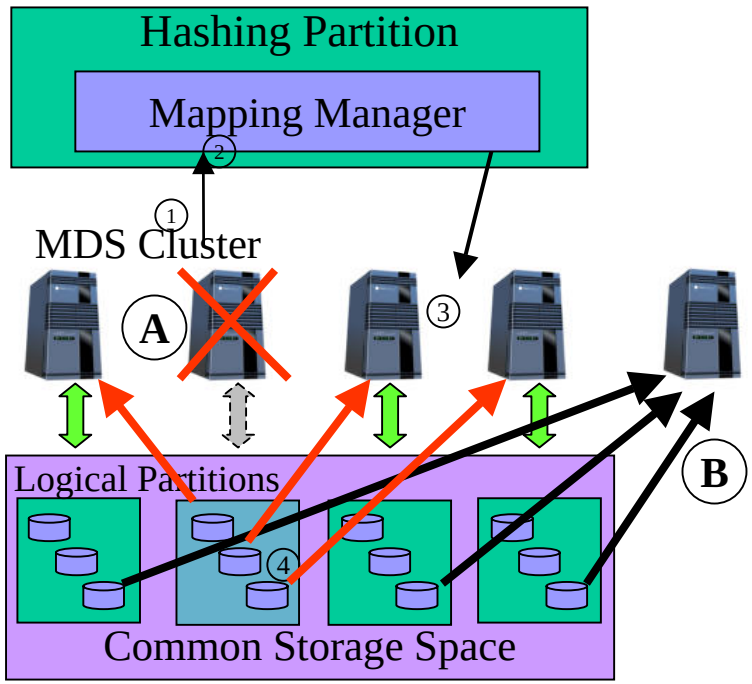


- 1. Hashing to f_id
- 2. Mapping from f_id to obj_id(s) to partition number. Then sending it to the dedicate MDS mounting that partition
- 3. Accessing metadata and checking the permission
- 4. Returning metadata to server or deny request

HAP

Load balancing, Scalability and Failover Design

Data Storage Institute



A MDS Failover

- 1. Detecting the MDS failure
- 2. Adjust the mapping relationship
- 3.Other MDSs take over logical partitions of the failure one
- 4.Journal recovery

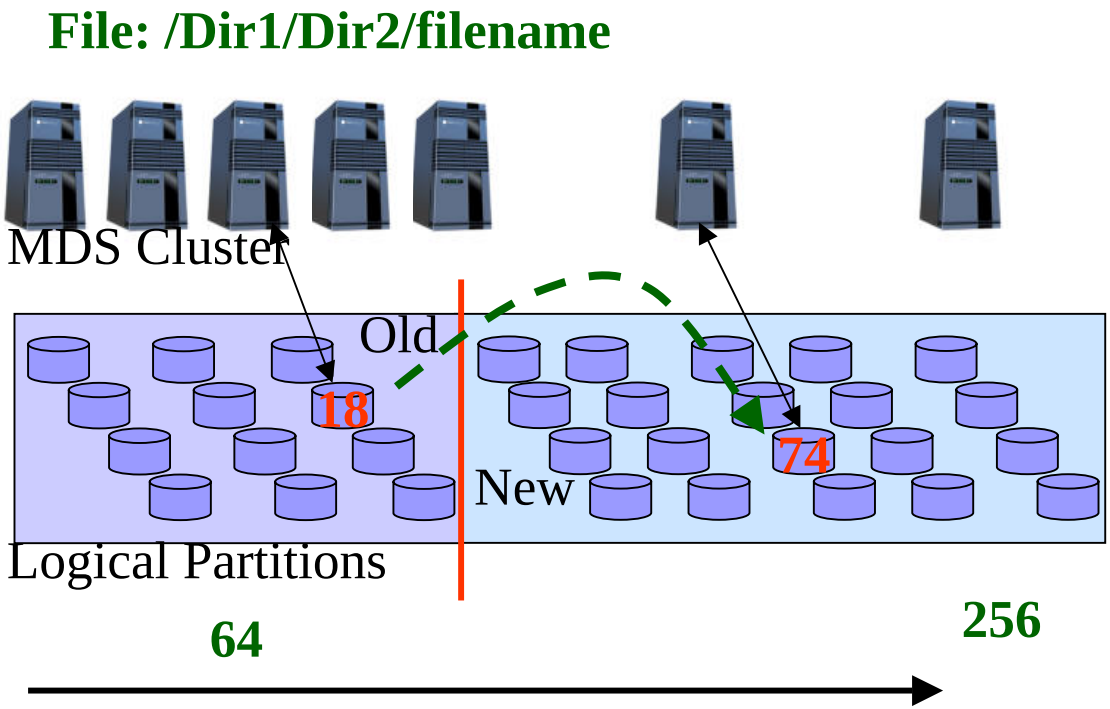
B MDS Cluster Scalability

Conclusion: **if the number of logical partitions is not changed,** Load balancing, Scalability and Failover can be simply and efficiently implemented just by some mount/umount operations on logical partitions.

HAP --- MDS Cluster Rebuild

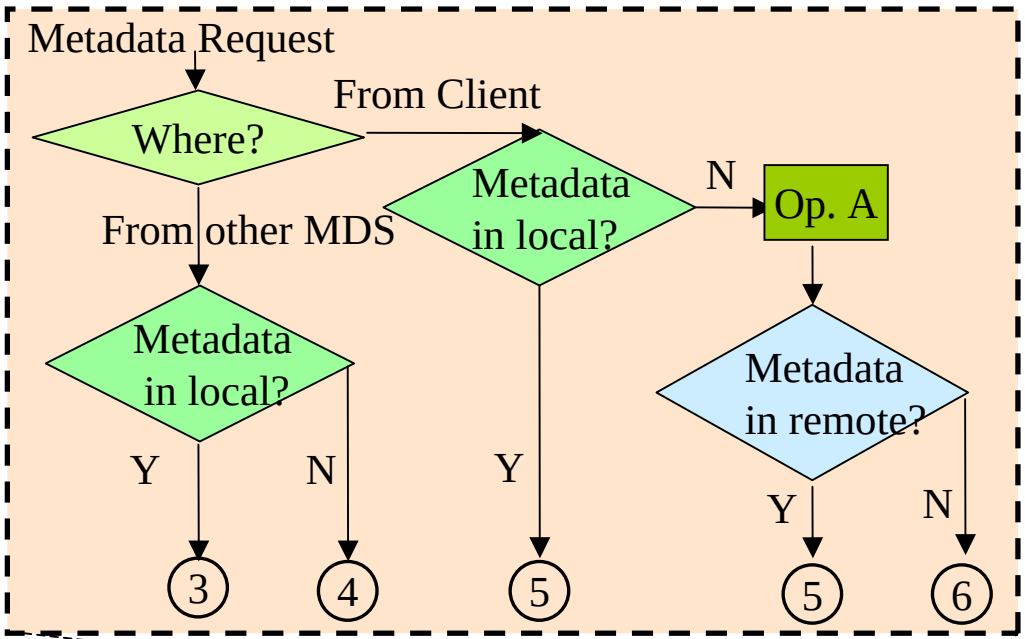
Data Storage Institute

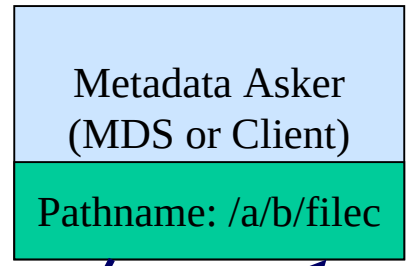
- But if the number of logical partitions is changed, … … ??

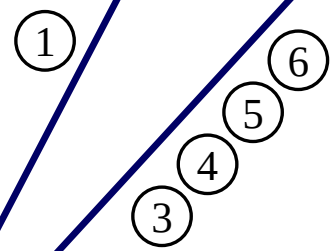


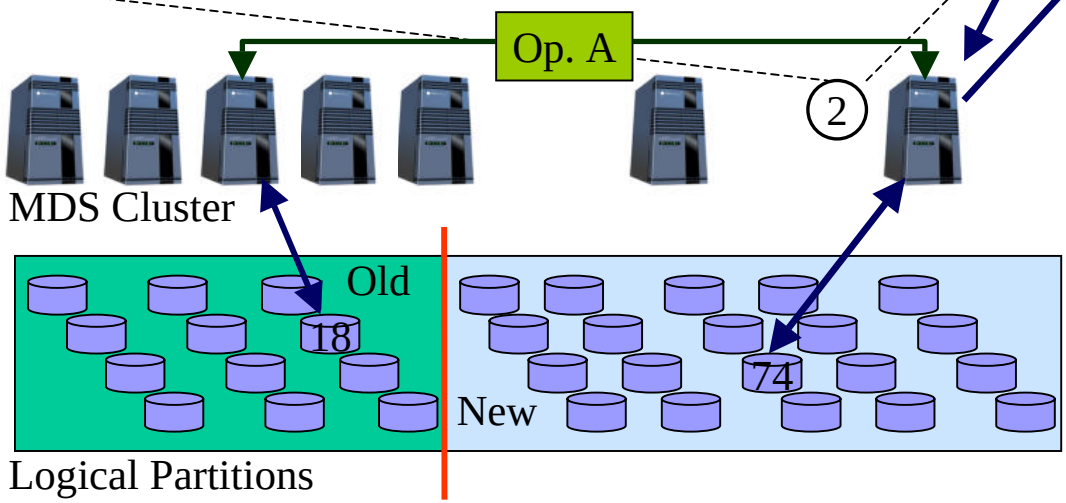
Existed metadata records need to be redistributed among logical partitions. This procedure is called **MDS Cluster Rebuild**.

HAP --- MDS Cluster Rebuild









Op. A

1. Computing old partition number based on the old *f*
2. Finding the MDS that mounting the old partition based on the new *MLT*
3. Issuing a request to get metadata from the MDS.

Conclusion

- **HAP** reduces the number of metadata requests based on the hashing method.
- **HAP** uses filename hashing policy to remove the overhead of multi-MDS communication.
- **HAP** provides efficient solutions for load balancing, failover and scalability of MDS cluster.