

**HUAWEI ENTERPRISE A BETTER WAY**

# Cache Service In Distributed FileSystem

**enterprise.huawei.com**  
HUAWEI TECHNOLOGIES CO., LTD.



SAN Array

Unified  
Storage

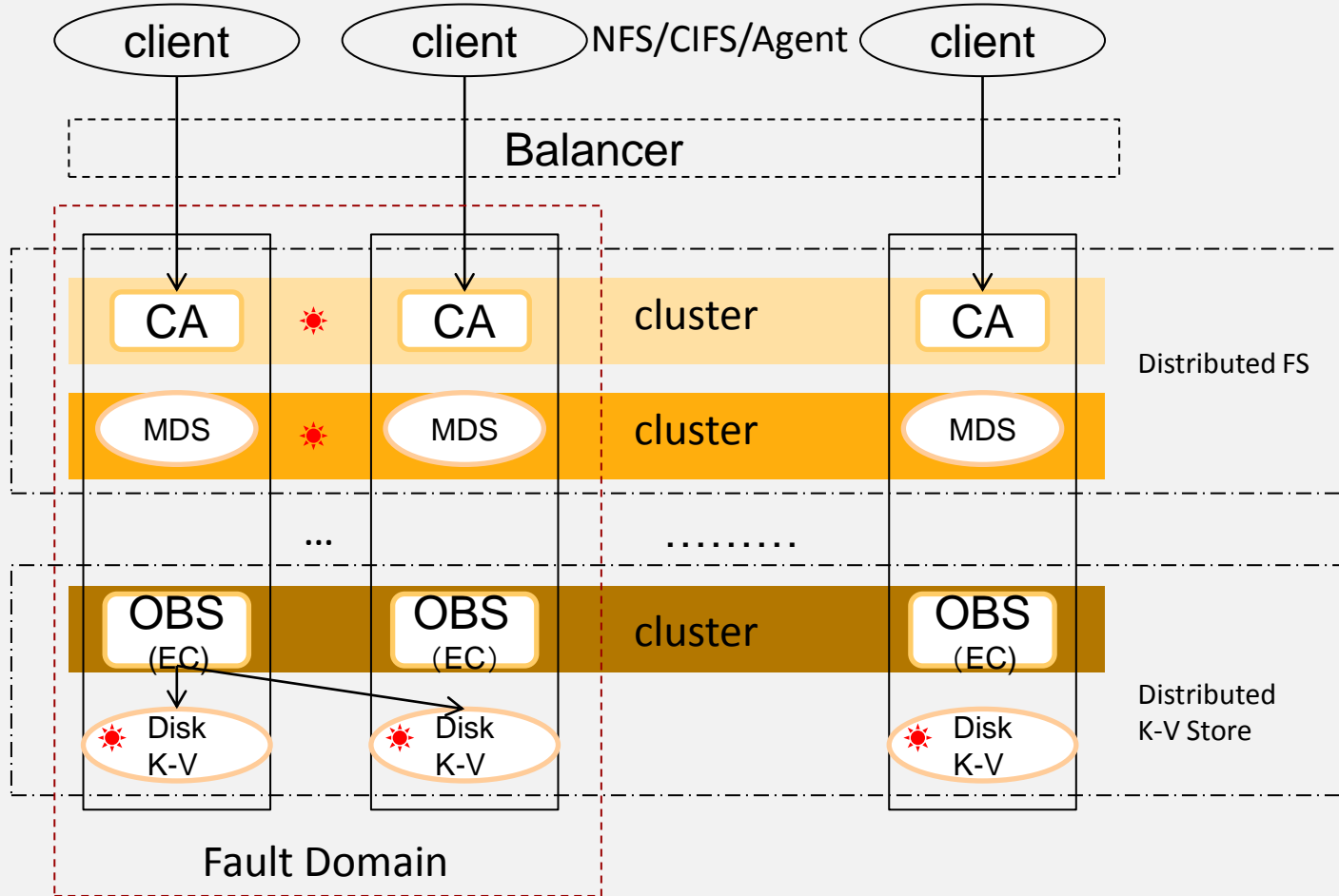
# HuaWei Storage

OceanStor  
9000

DSware

OceanStor  
ONE

# OceanStor 9000—Distributed FS



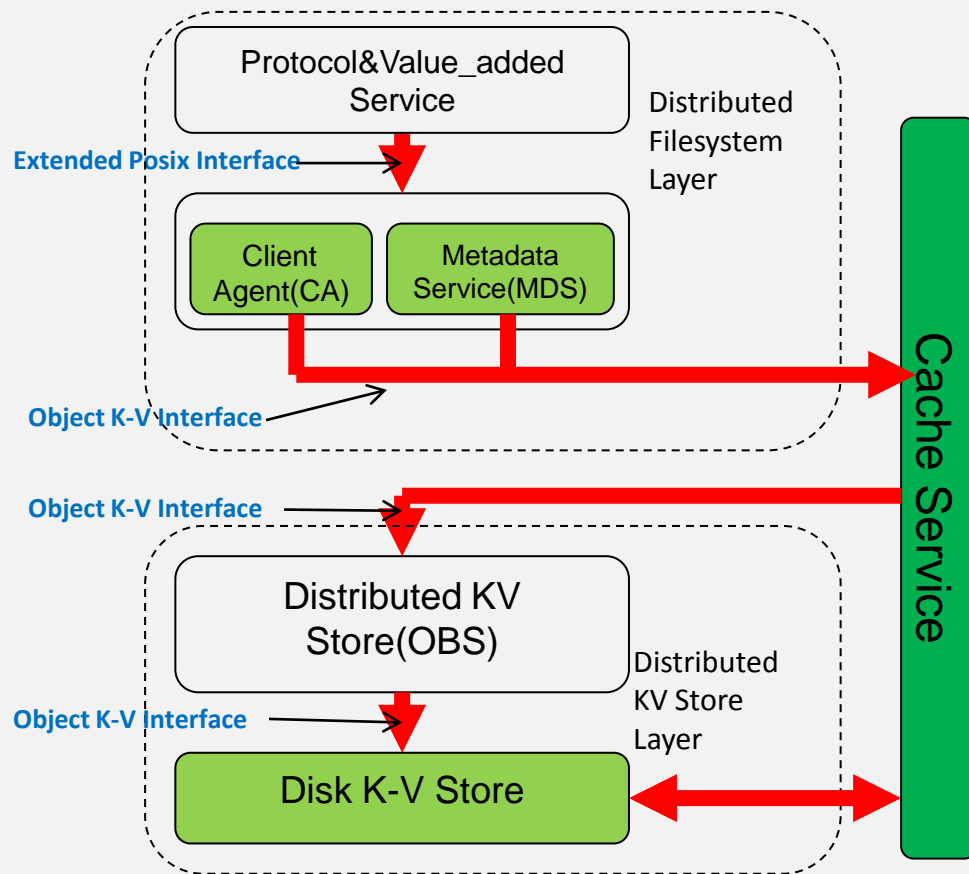
TODO:

Distributed Cache Service

# Problems & Cache Targets

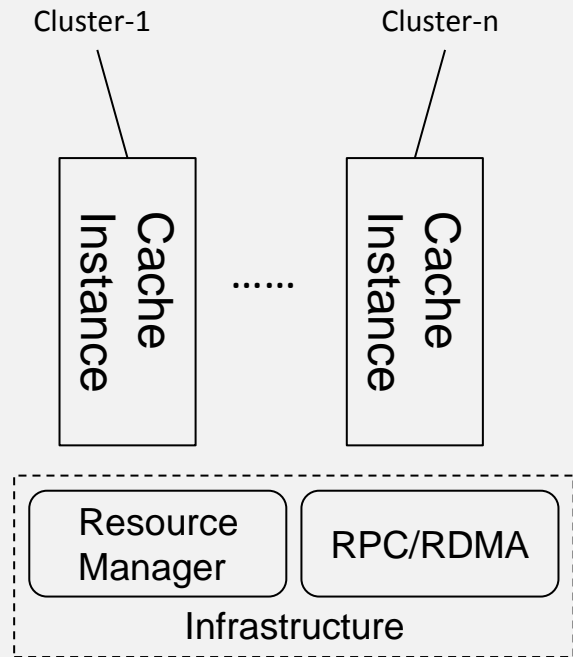
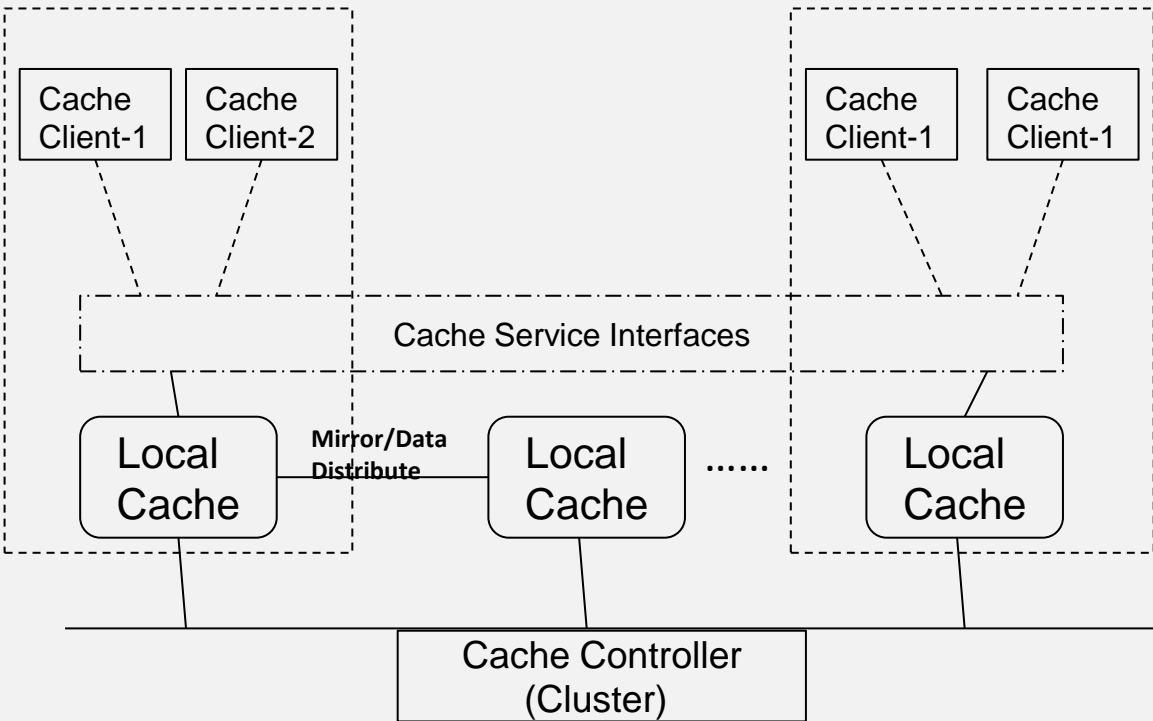
- **Problems**
  - OBS: EC write-amplification
  - MDS: costly to create file (6ms one file)
  - DISK K-V: data aggregation
  - CA: read is costly
- **Cache Targets**
  - Reduce/Eliminate EC write-amplification
  - Speed up metadata transactions
  - Write buffer cache
  - Data Prefetch and Caching
- **Challenges**
  - Multiple Cache Clients
  - Different data organization
  - Try to REUSE

# Logical Location

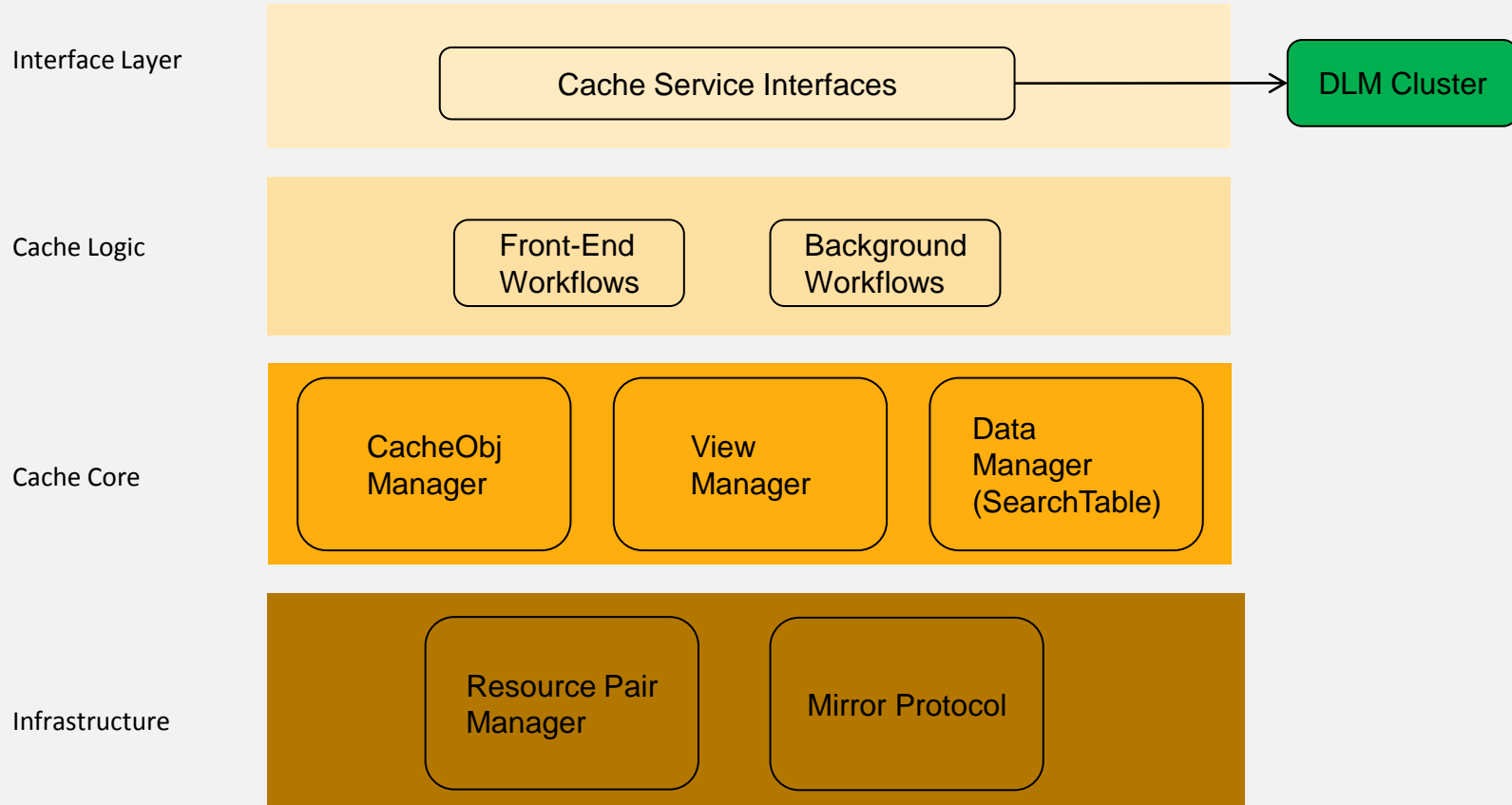


# Architecture

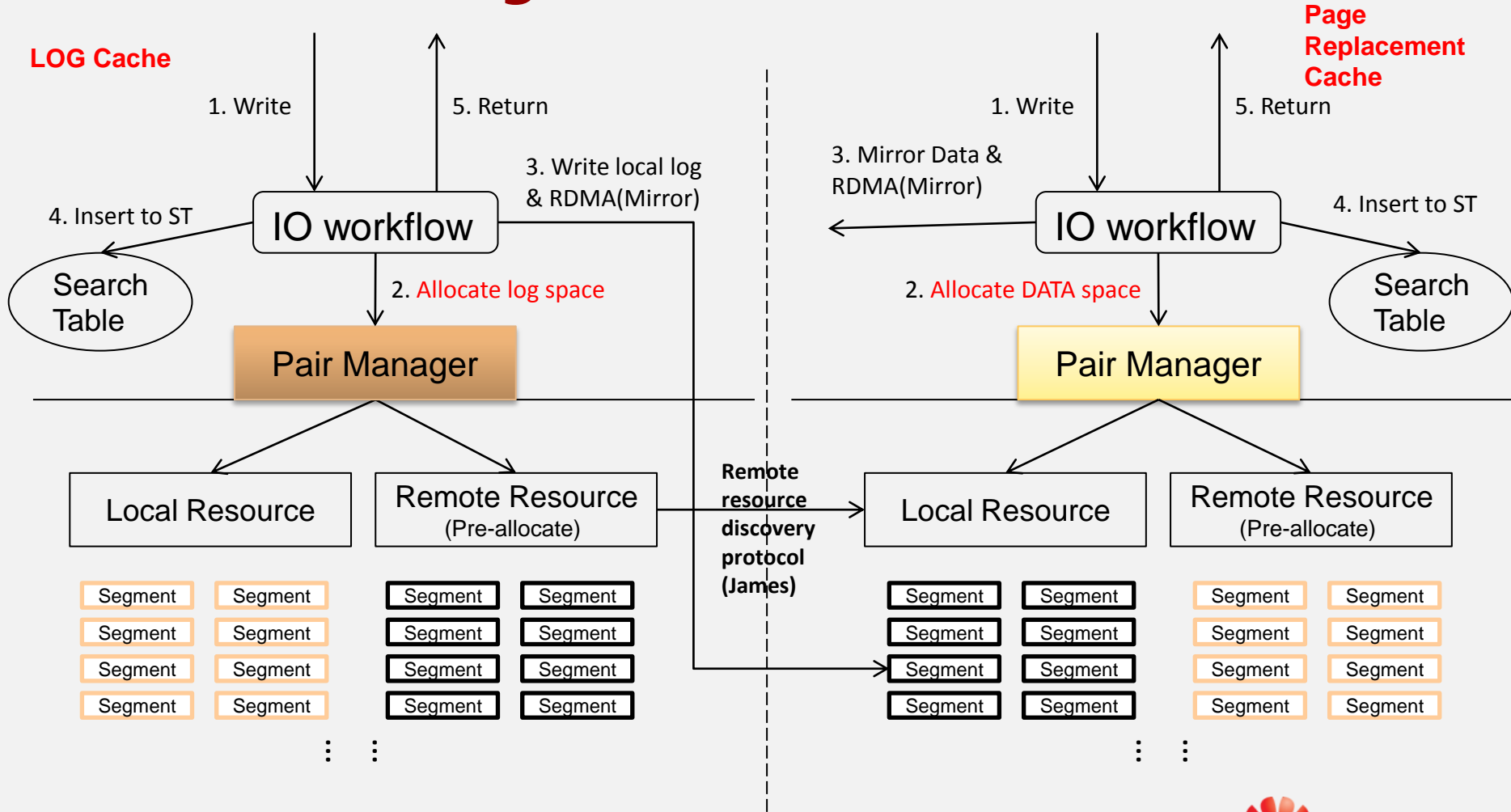
## Fault Domain



# Cache Instance

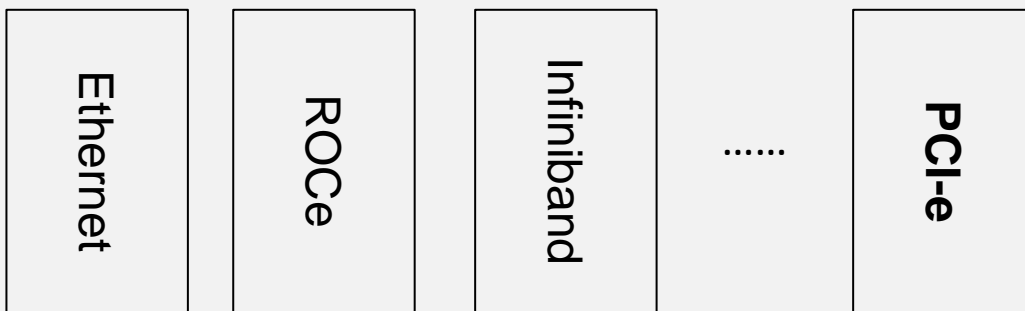


# Resource Manager





## Unified Communication Interfaces (RPC/RDMA)

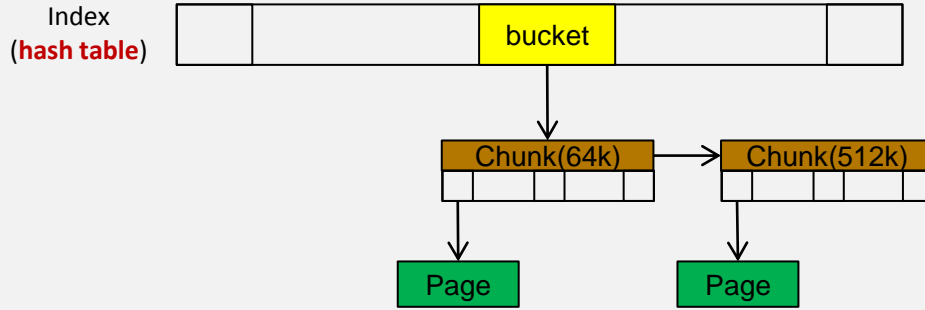


# Write

- **Write-back**
  - Write a log in mirrored NVDIMM(persistent resource)
  - Insert data to SearchTable
- **Write-through**
  - Push the overlapped data to flush
  - Write a "CMD" log to denote the corresponding logs are invalidated
  - Write current data to back-end storage

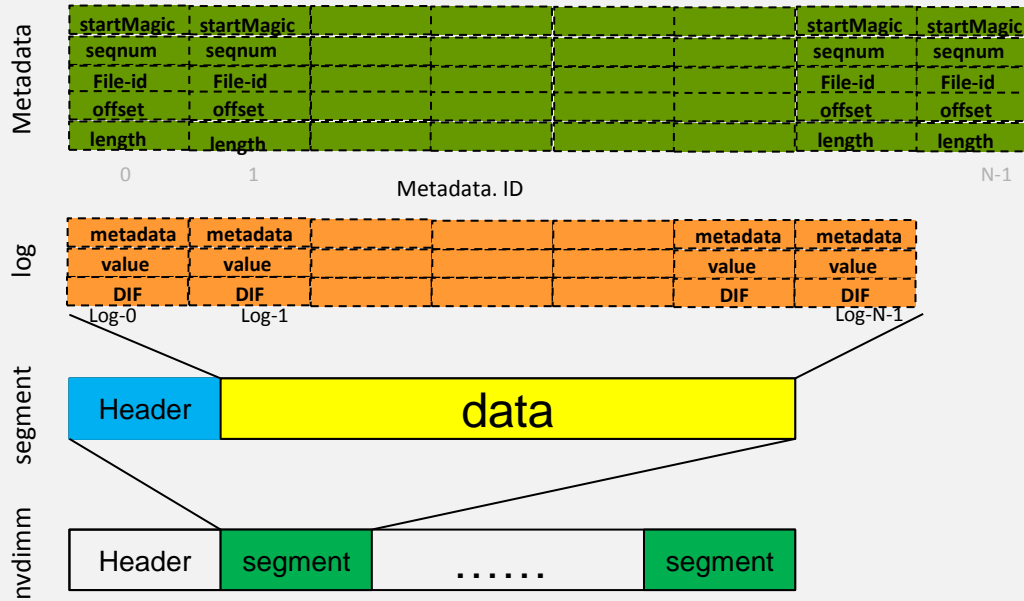
# Log Cache—NVDIMM+RAM

DRAM



- No direct information connect DRAM and NVDIMM
- No way to reference log through DRAM
- Only reference “chunk” from NVDIMM log through “KEY”
- NVDIMM space recycle process work with the “flush data” process
- There are special “command log” to handle other “flush” process

NVDIMM



# Read

- **Normal read**
  - Acquire “read lock” from DLM server successfully
  - Read data from local
  - If missed, fetch the data from back-end storage
- **Forwarding read**
  - Acquire “read lock” failed but get a write node
  - Read data from remote node

# Flush

- **DLM Lock recall**
  - One RECALL one "CMD" log
- **Write-Through**
  - One IO one "CMD" log
- **Log Space Recycle**
  - One segment one "CMD" log

## Notes:

- Avoid "dead lock" between resource "release" and "allocation"
- Write-Through does not count on any other Cache  
"workflow"&"resource"



**HUAWEI ENTERPRISE A BETTER WAY**

**Copyright©2012 Huawei Technologies Co., Ltd. All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.