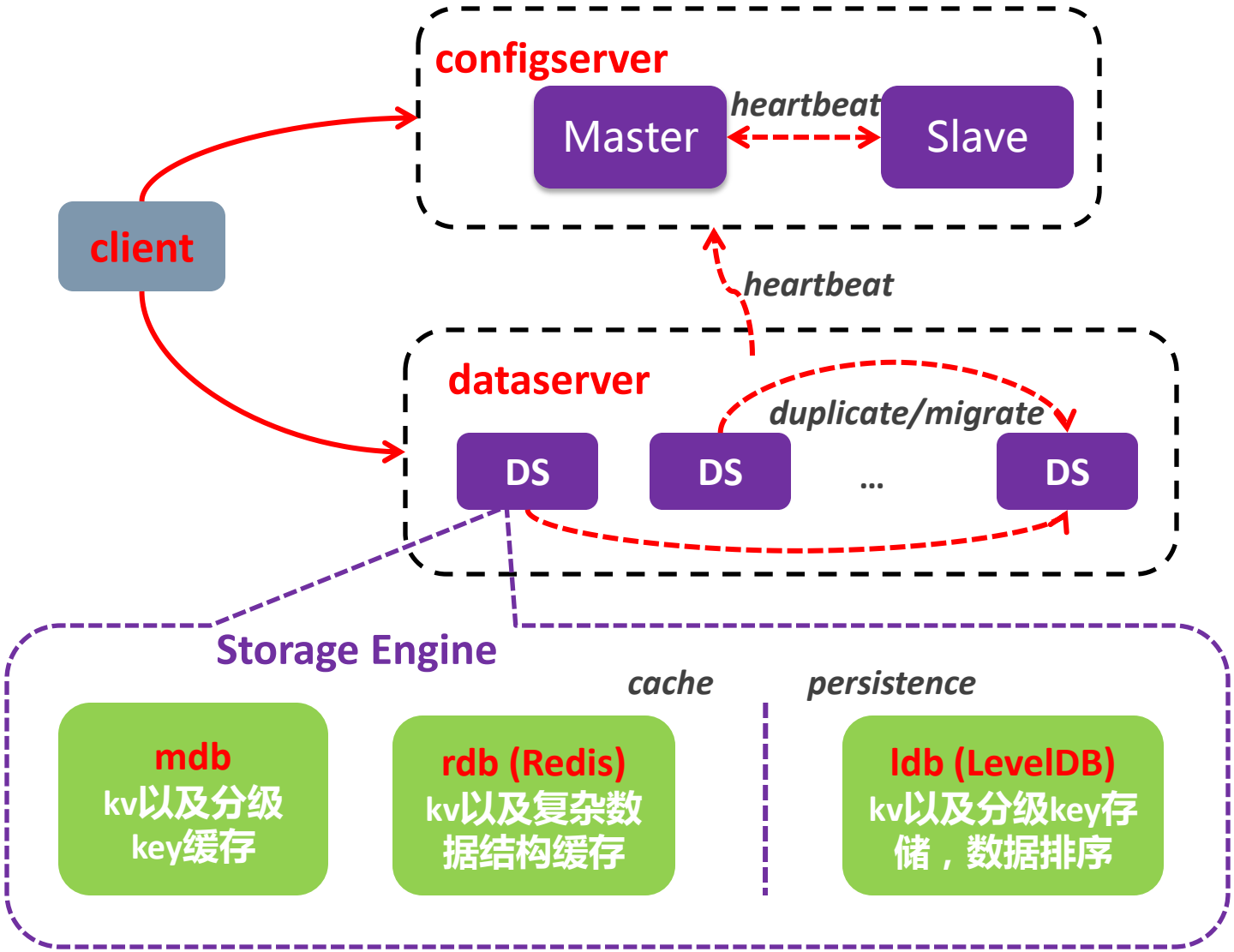
Tair 存储引擎之路

阿里巴巴集团-核心系统研发-那岩（王玉法）

目录

一．Tair概述

Tair概述

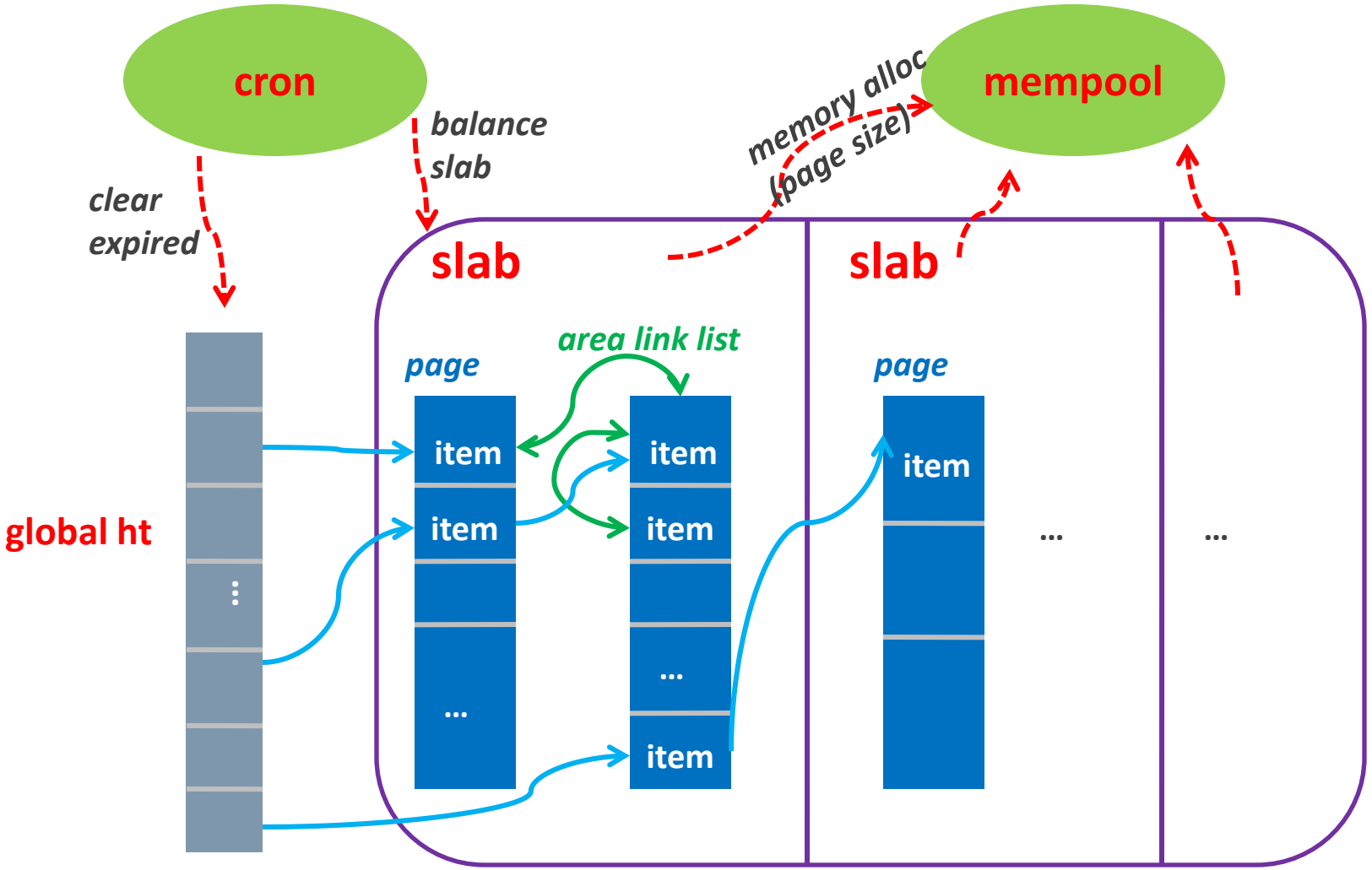

configserver
Master
heartbeat
Slave
client
heartbeat
dataserver
duplicate/migrate
DS
DS
...
DS
Storage Engine
cache
persistence
mdb
kv以及分级key缓存
rdb (Redis)
kv以及复杂数据结构缓存
ldb (LevelDB)
kv以及分级key存储，数据排序

二．mdb

mdb概述

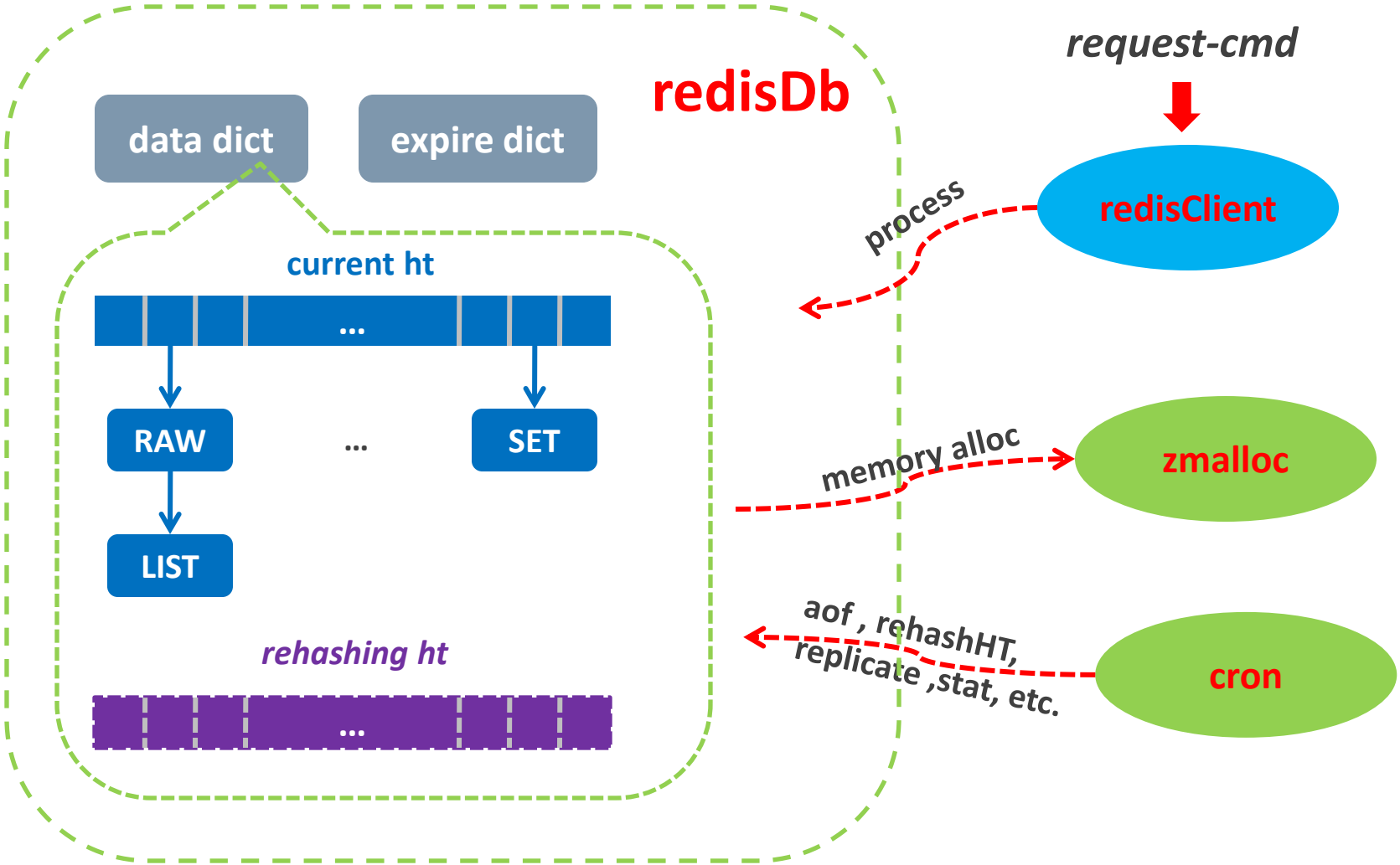
- **类memche，page/slab管理内存**
- **共享内存，重启数据不丢**
- **area逻辑管理**
  - area维度的quota与lru evict控制
  - 清理某个area数据
- **数据过期**
- **优化内存使用率，均衡slab**
- **详细statistics监控**

mdb流程

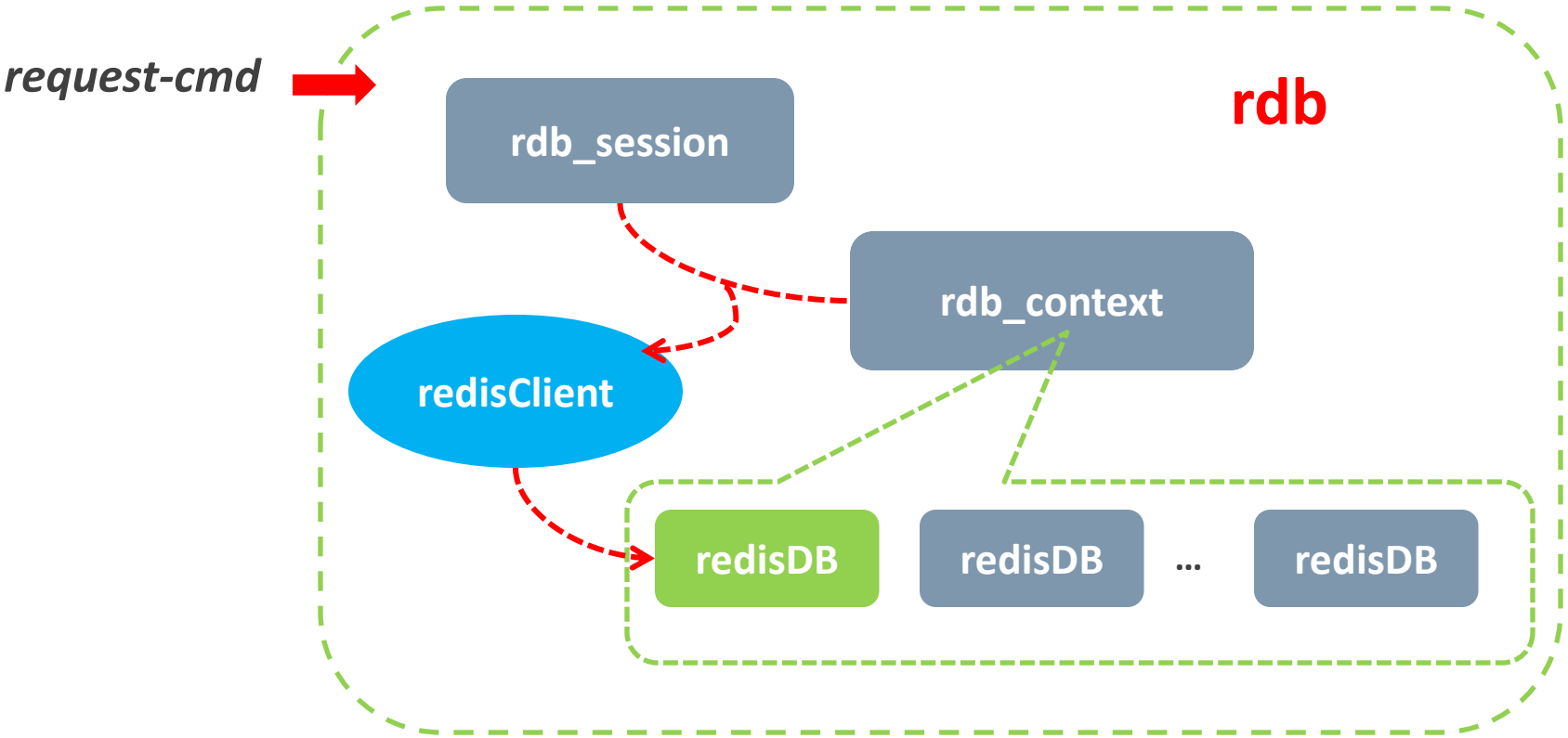

cron
balance
slab
clear
expired
memory alloc
(page size)
mempool
slab
slab
area link list
page
page
global ht
item
item
item
item
item
item
...
...
...
...

三．rdb（Redis）

Redis流程（存储）


redisDb
data dict
expire dict
current ht
...
RAW
...
SET
LIST
rehashing ht
...
request-cmd
redisClient
process
memory alloc
zmalloc
aof , rehashHT,
replicate ,stat, etc.
cron

rdb流程


request-cmd
rdb
rdb_session
rdb_context
redisClient
redisDB
redisDB
...
redisDB

rdb概述

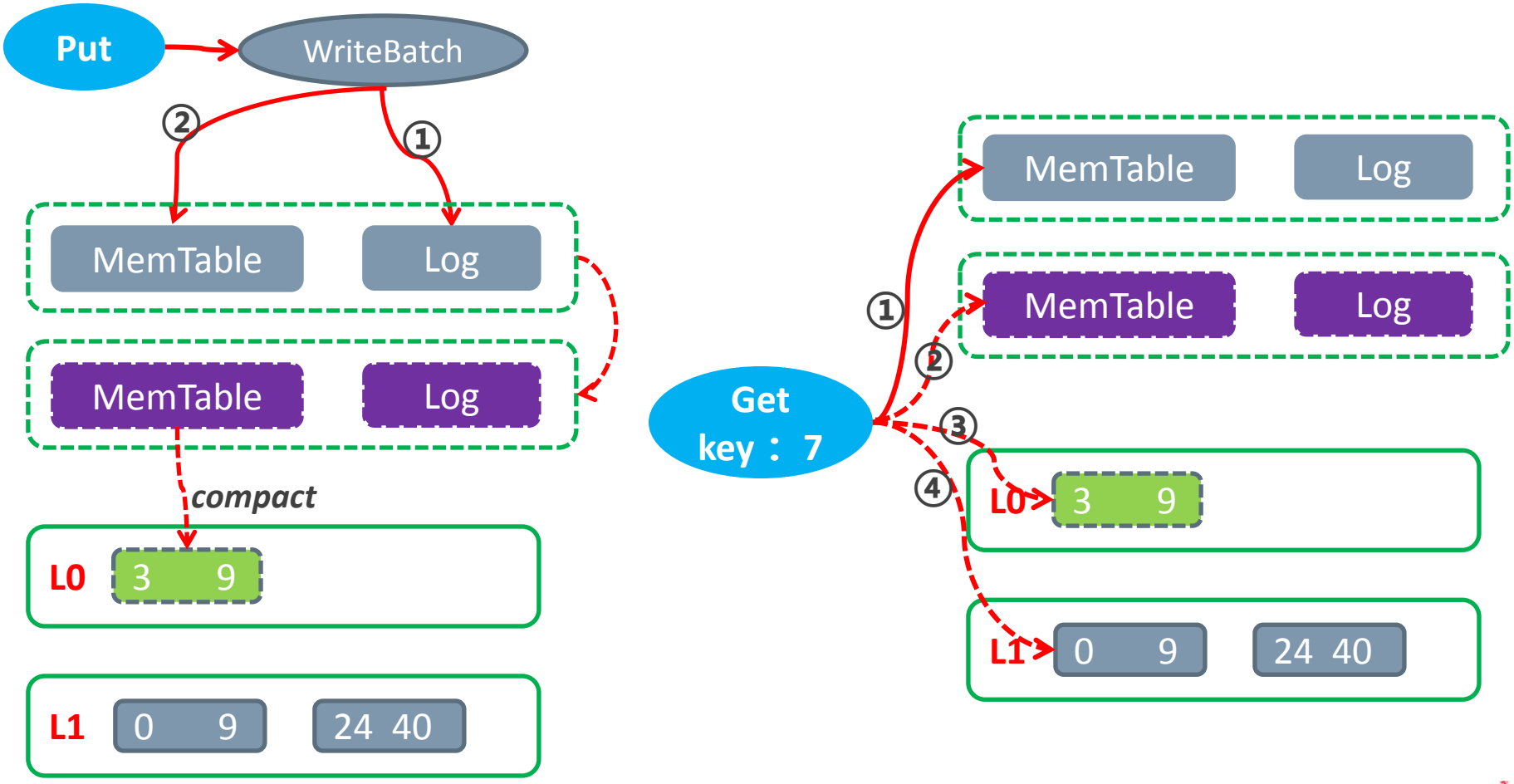
- **支持Redis所有数据结构**
- **设置限制**
  - 内存quota
- **logiclock**
  - lazy 清理db数据

rdb概述

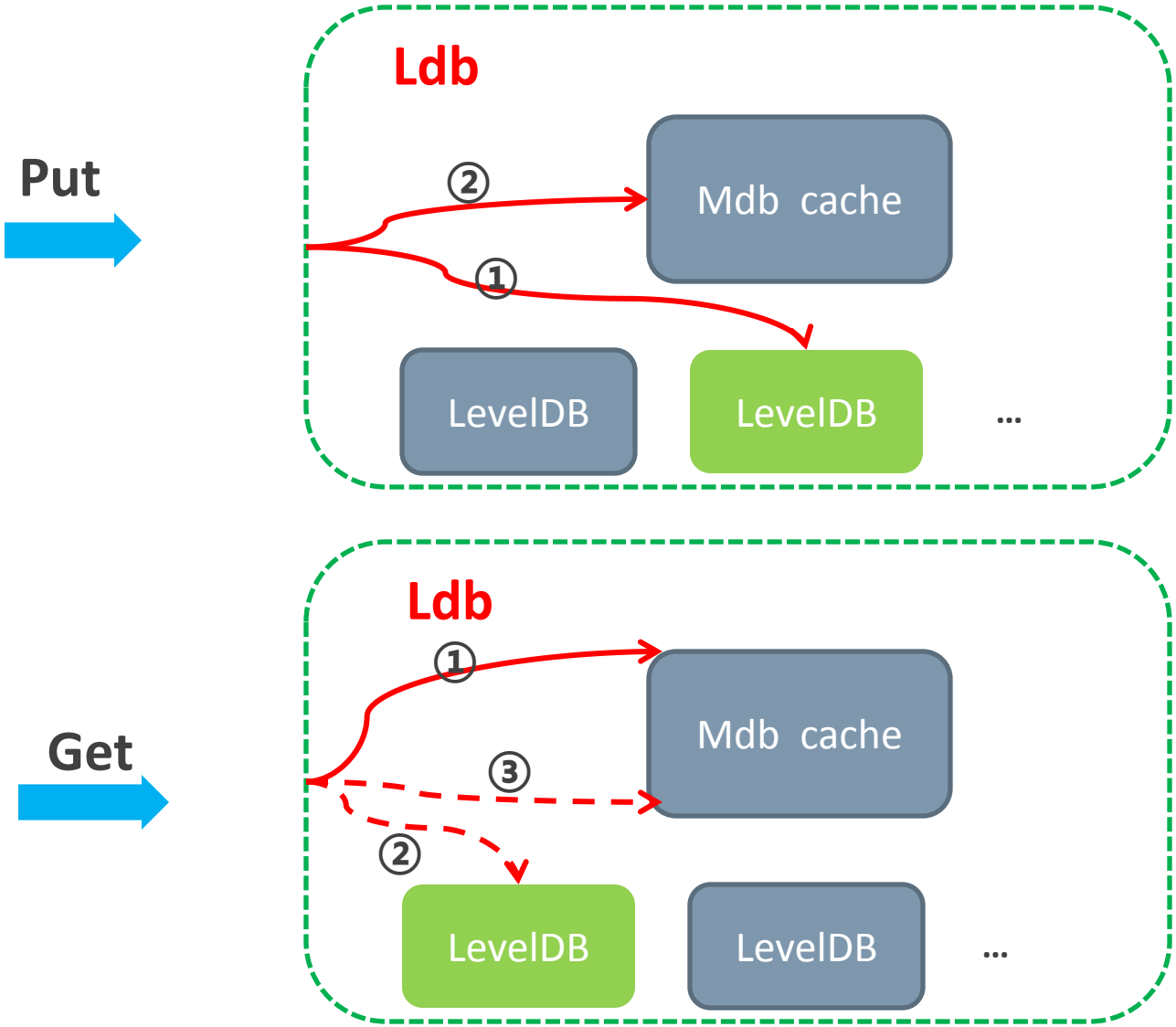
- **轻量化**
  - 去除aof/vm
  - 精简数据结构
- **Restful协议**
- **持久化**
  - 使用ldb作为rdb的持久化（TODO）

四．ldb（LevelDB）

LevelDB流程


Put
WriteBatch
②
①
MemTable
Log
MemTable
Log
compact
L0
3 9
L1
0 9
24 40
Get
key：7
①
②
③
④
MemTable
Log
MemTable
Log
L0
3 9
L1
0 9
24 40

ldb 流程


Put
Ldb
②
Mdb cache
①
LevelDB
LevelDB
...
Get
Ldb
①
Mdb cache
③
②
LevelDB
LevelDB
...

ldb概述 .配置使用

- **多实例配置使用，充分利用IO**
- **内嵌mdb作为KV级别cache**
- **配置灵活化，参数调优**
  - 适合大数据量的参数配置（ mmt/sst size，mmap限制，etc.）
  - 特定排序算法（字节，数字，etc.）

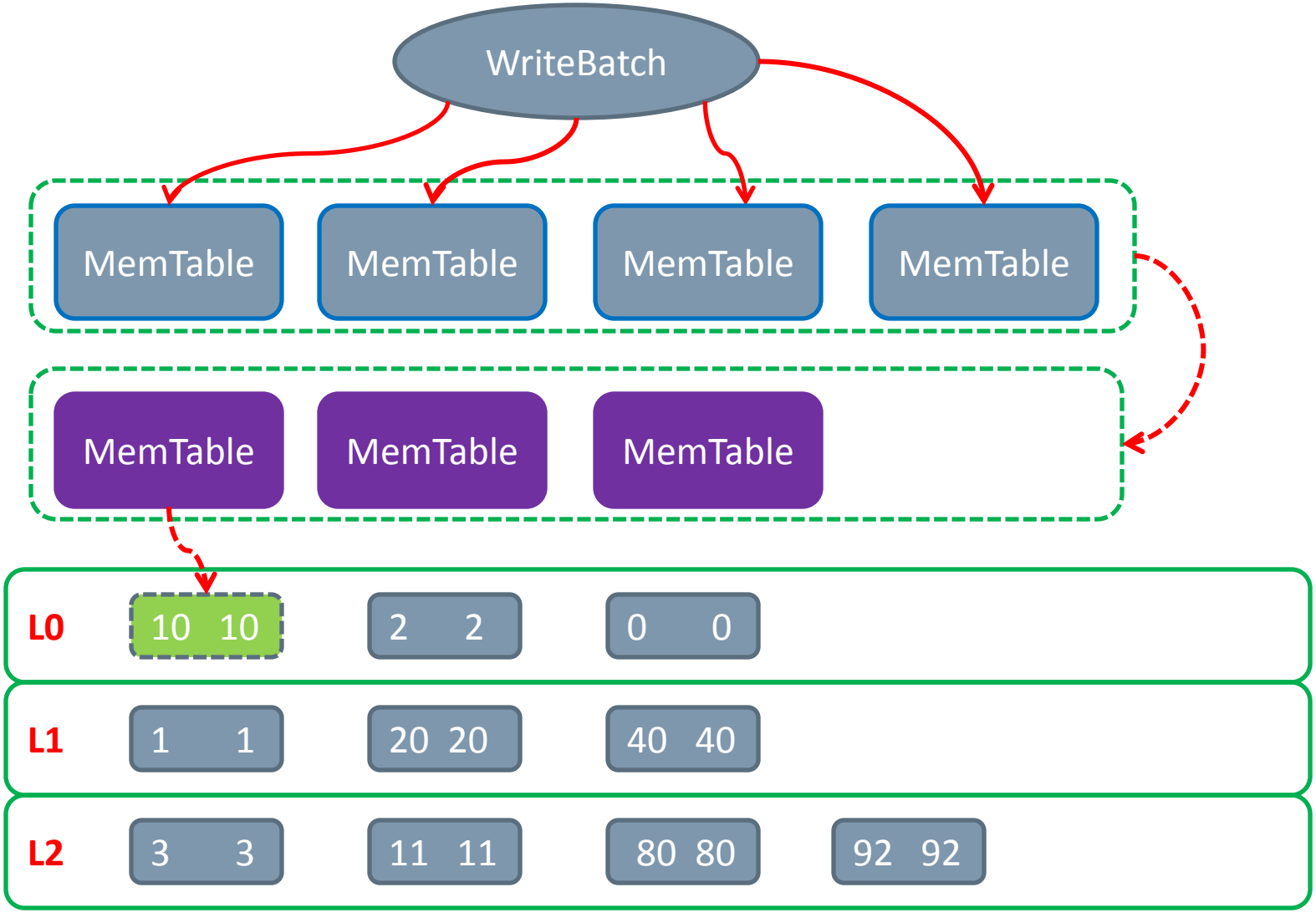
ldb概述.新功能

- **嵌入datafilter逻辑**
  - expiretime
  - 异步清理
- **主动触发compact**
  - Level-n => Level-n compact，清理垃圾数据
  - 高Level compact，加速range合并（TODO）
- **使用binlog做异步跨集群数据同步**

ldb概述.优化

- **主动限制compact**
  - 限制写入放大
  - 禁掉seek触发的compact
- **优化compact锁粒度**
  - 减弱数据量增长对读写的性能影响
- **大数据量导入FastDump**
  - 数据预排序，按桶分memtable

FastDump导入数据



相关资料

- **Tair开源页面**

  http://code.taobao.org/p/tair/src/

谢 谢 ！