hosted by Alibaba Group 阿里巴巴集团 APACHE HBASE

# HDFS optimization for Hbase At XiaoMi

Xie Gang

xiegang1@xiaomi.com

# Content

01 Latency & Availability monitoring

02 Latency optimization & practice

03 Detect the dead node in advance

# 01 Latency & Availability Monitoring

# Availability from SLA perspective

- Multiple replicas
- Namenode & Datanode

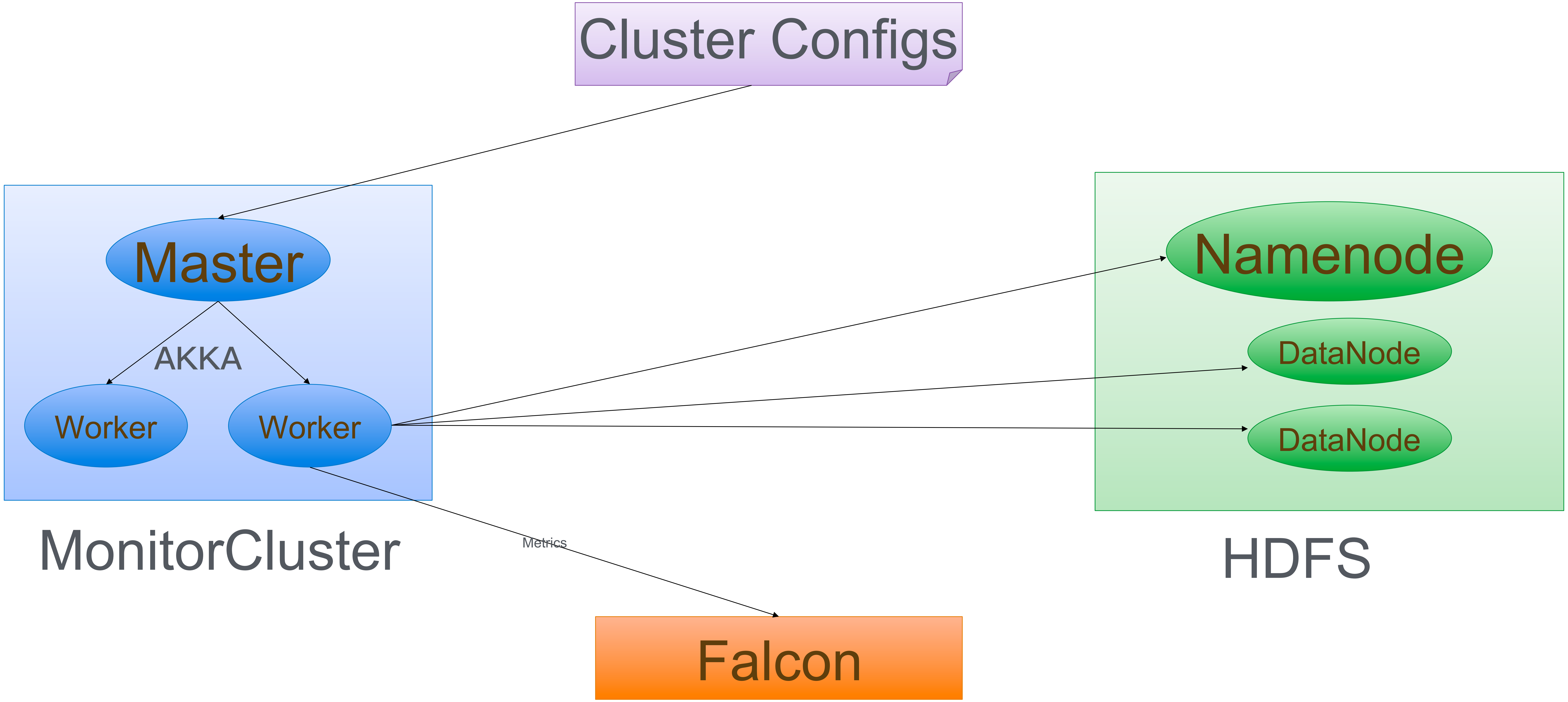
$$writeSLAAvailability = 1 - \sum_{i=1}^{r} C(i, k) \times C(r - i, N - k) / C(N, r)$$

$$readdSLAAvailability = 1 - k / N$$

$$availability = Min(namenodeAvailability, writeSLA, readSLA)$$

r=replication, factor k=fault datanodes, N=total datanodes

# Monitor the latency and availability

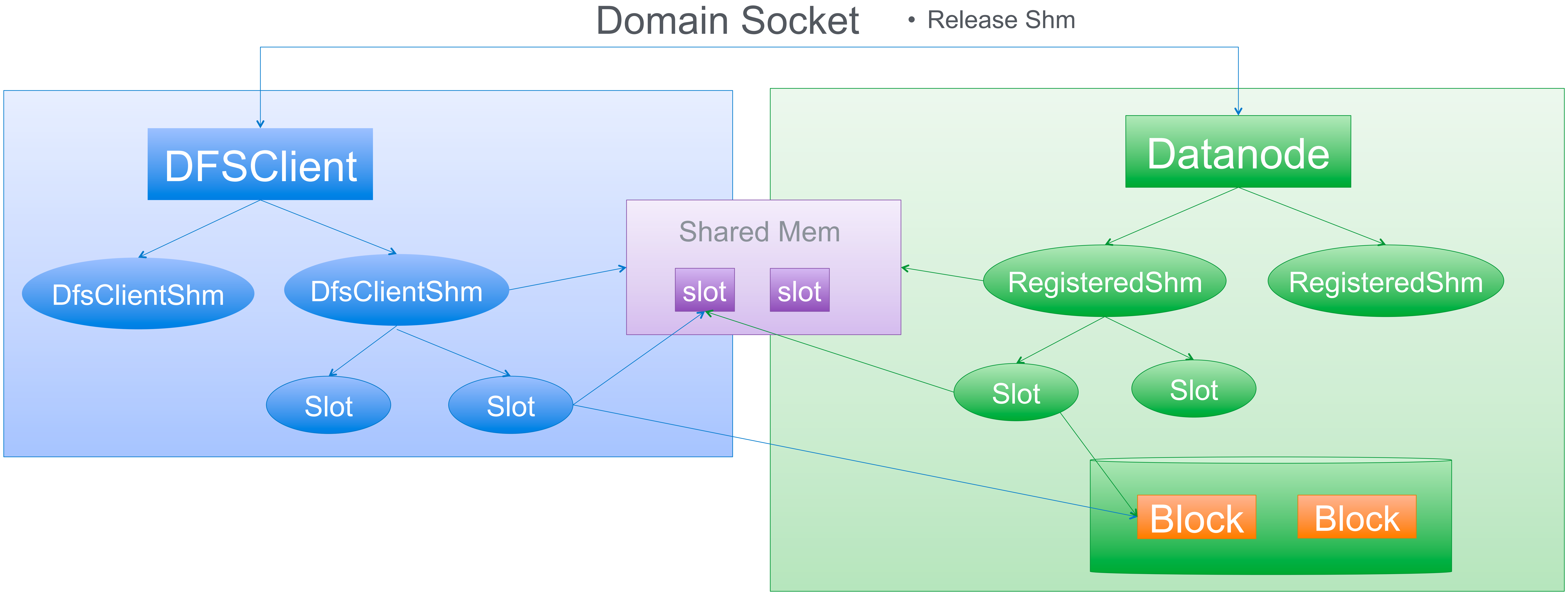


# 02 Latency optimization & practice

# Short Circuit Read optimization

## Architecture

- Allocate Shm
- Request slot & FDs
- Release Slot
- Release Shm



# Short Circuit Read optimization

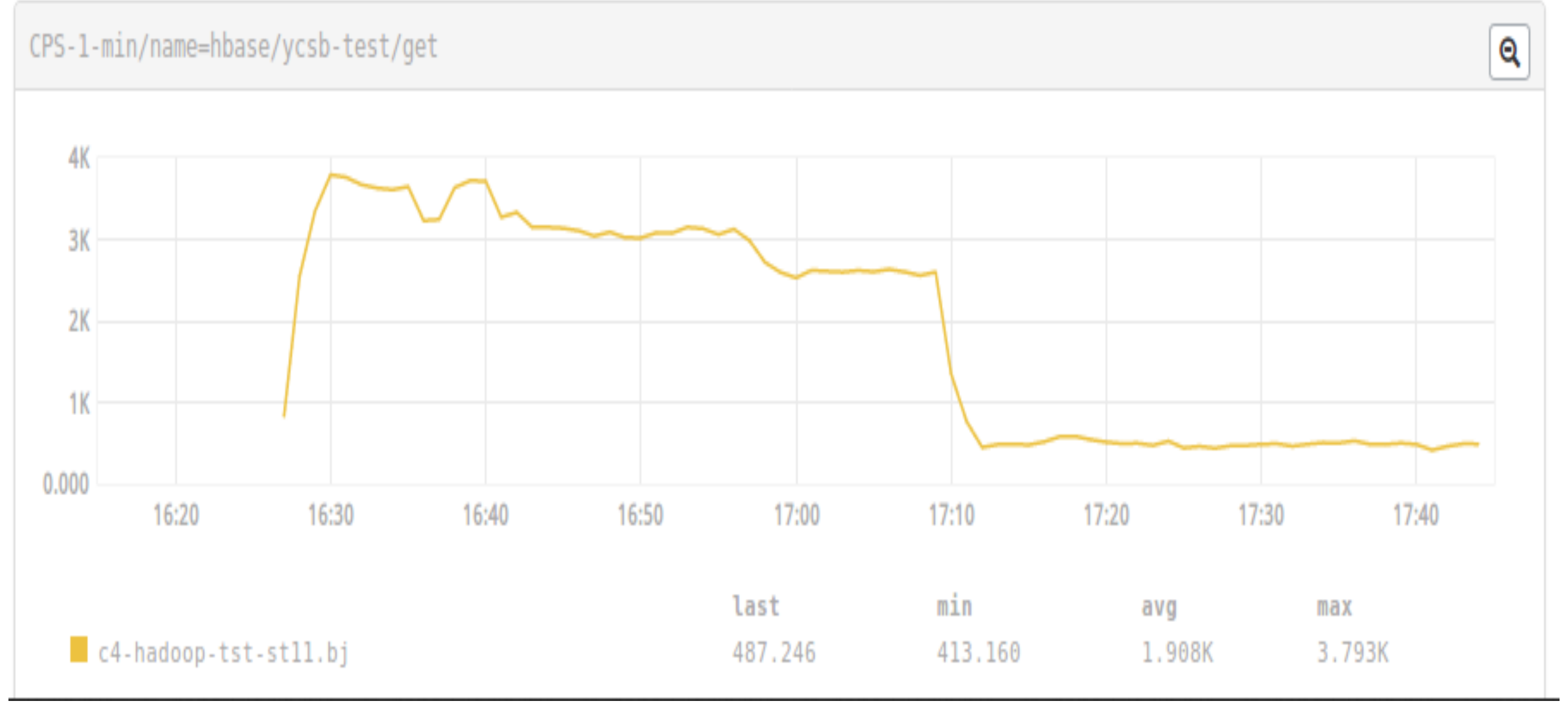
✧ Problem #1 : Slot Releaser is not fast enough

Scan on huge region (~2TB):

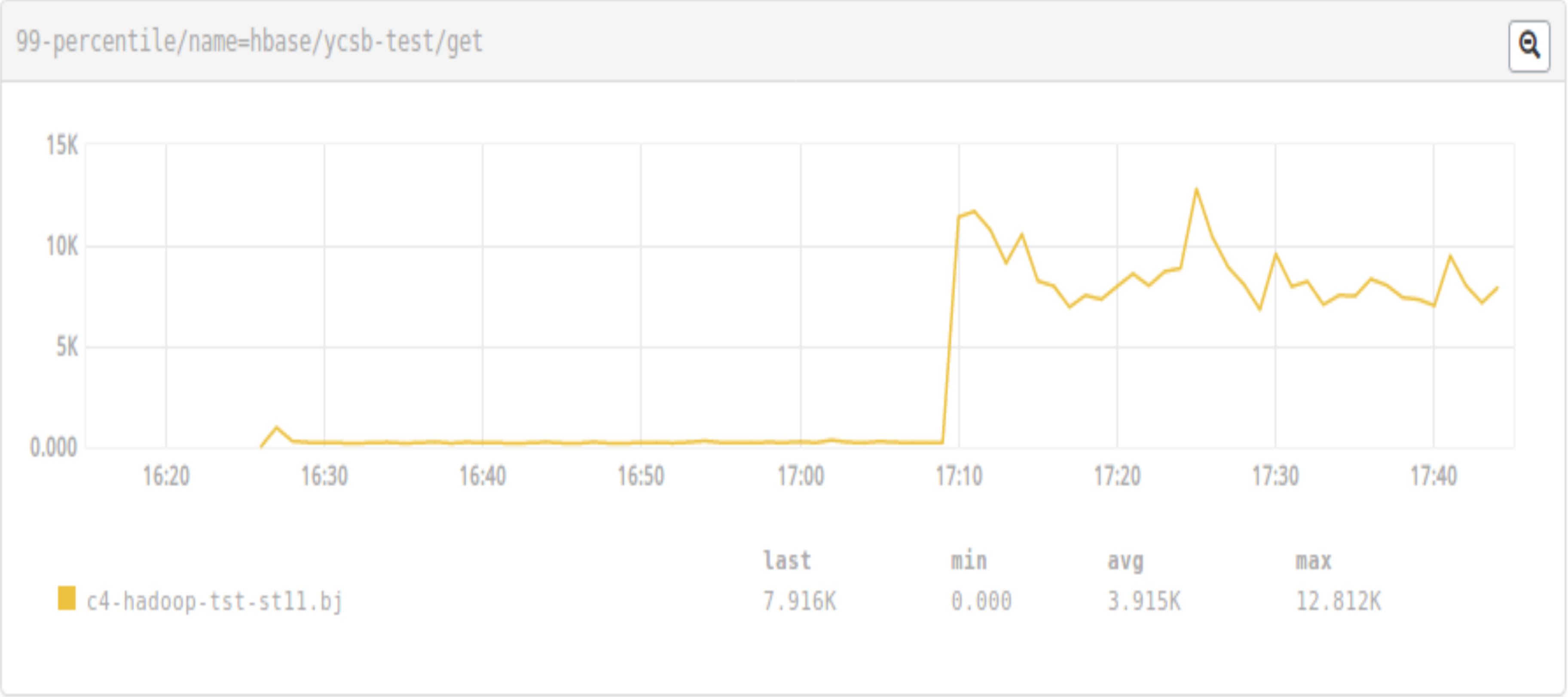
- Slot allocation QPS 1000 +
- Slot release QPS 1000 –
- Datanode Full GC caused by RegisteredShm

YCSB get:

- 3000+ QPS allction VS 1000+ QPS release



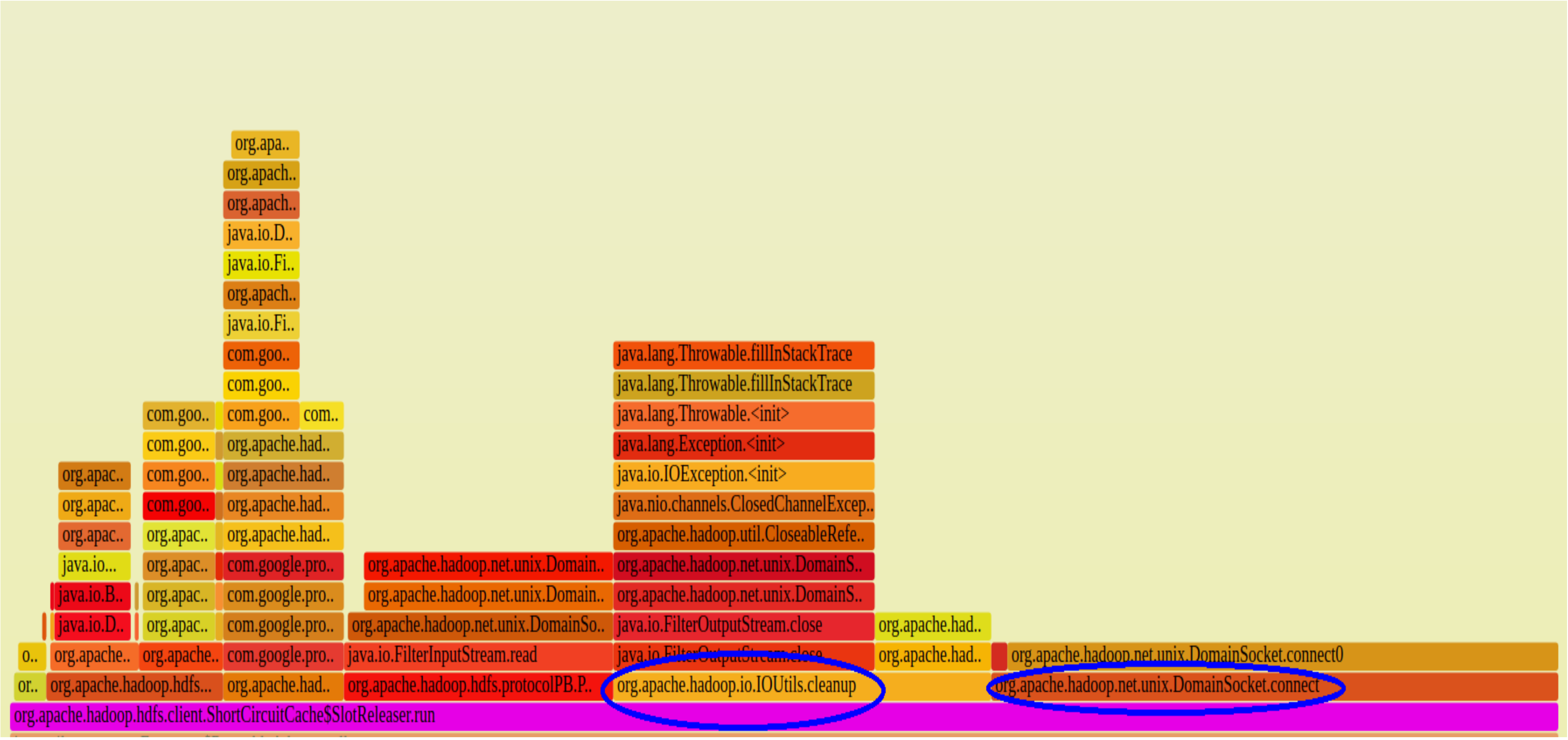
| | last | min | avg | max |
|---|---|---|---|---|
| c4-hadoop-tst-st11.bj | 487.246 | 413.160 | 1.908K | 3.793K |



| | last | min | avg | max |
|---|---|---|---|---|
| c4-hadoop-tst-st11.bj | 7.916K | 0.000 | 3.915K | 12.812K |

# Short Circuit Read optimization

✧ Domain socket connecting every time

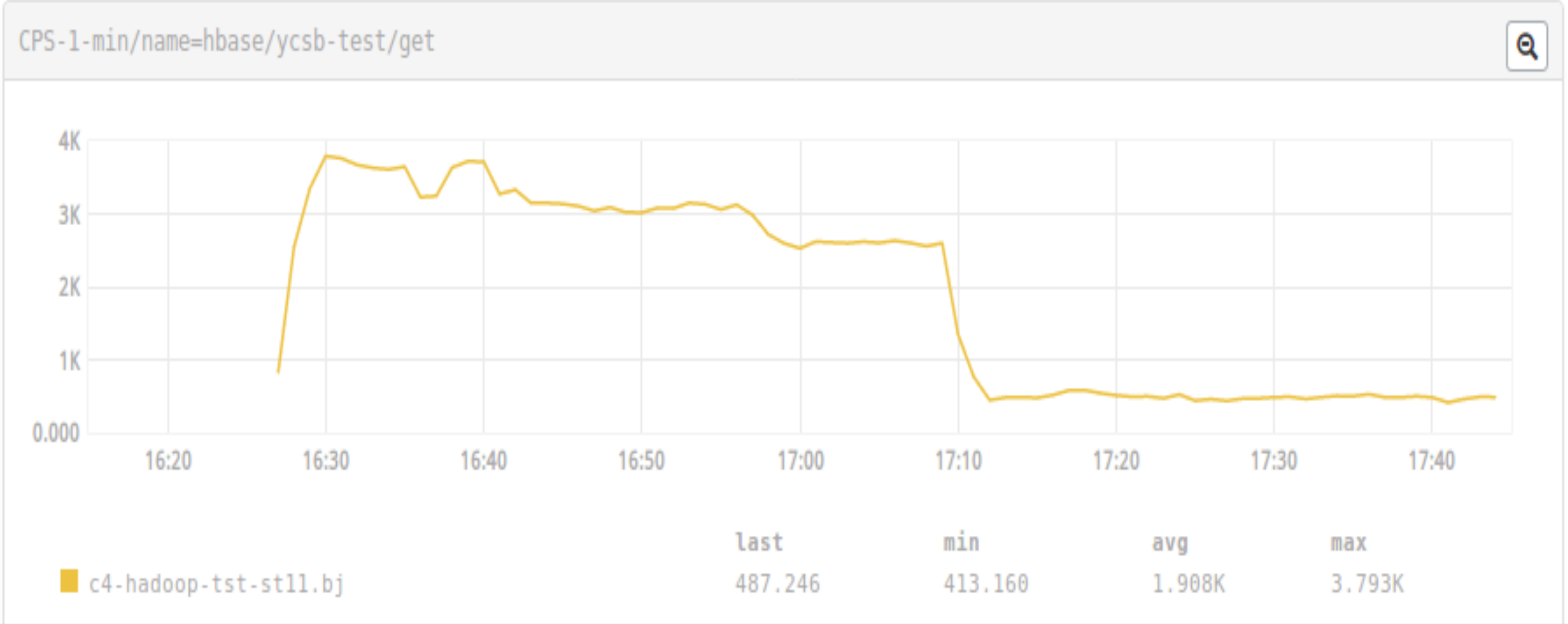


# Short Circuit Read optimization

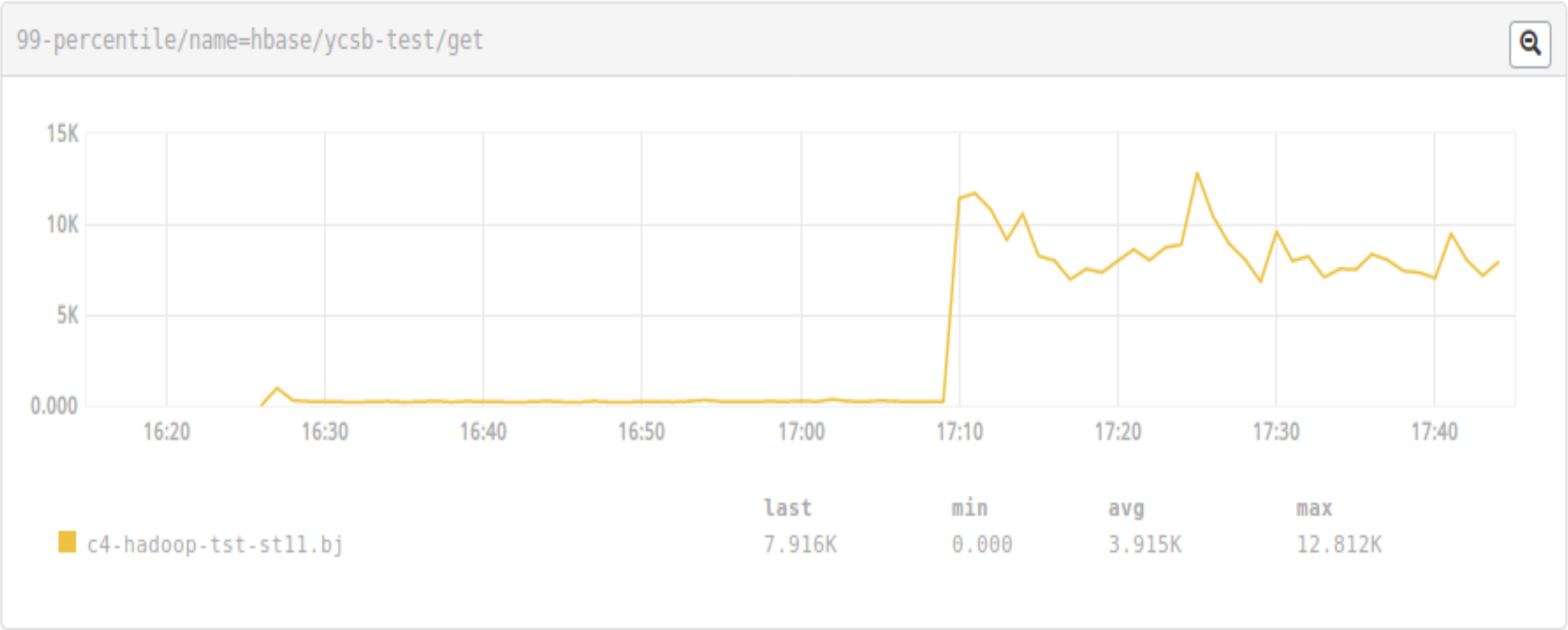
hosted by

✧ Reuse the connection of the domain socket

YCSB get ： 20% QPS incensement


CPS-1-min/name=hbase/ycsb-test/get

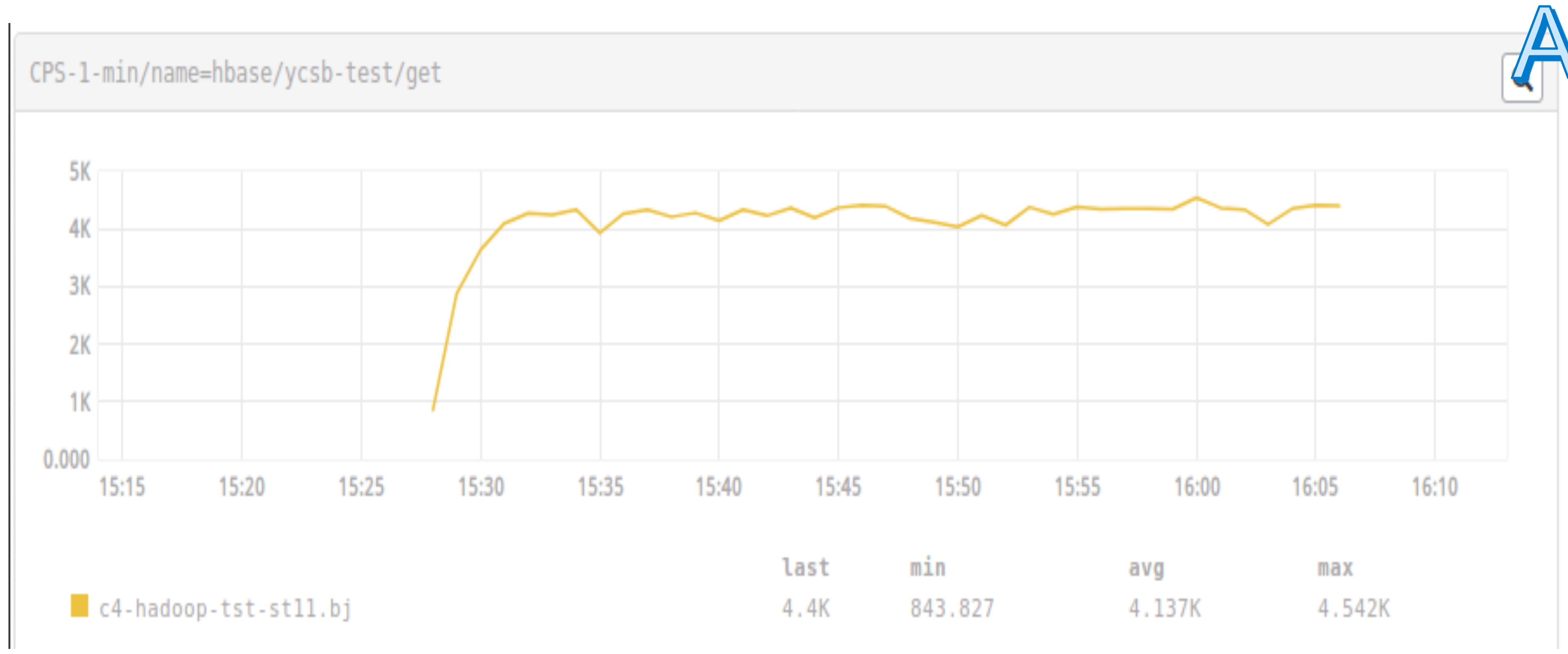
| | last | min | avg | max |
|---|---|---|---|---|
| c4-hadoop-tst-st11.bj | 487.246 | 413.160 | 1.908K | 3.793K |


99-percentile/name=hbase/ycsb-test/get

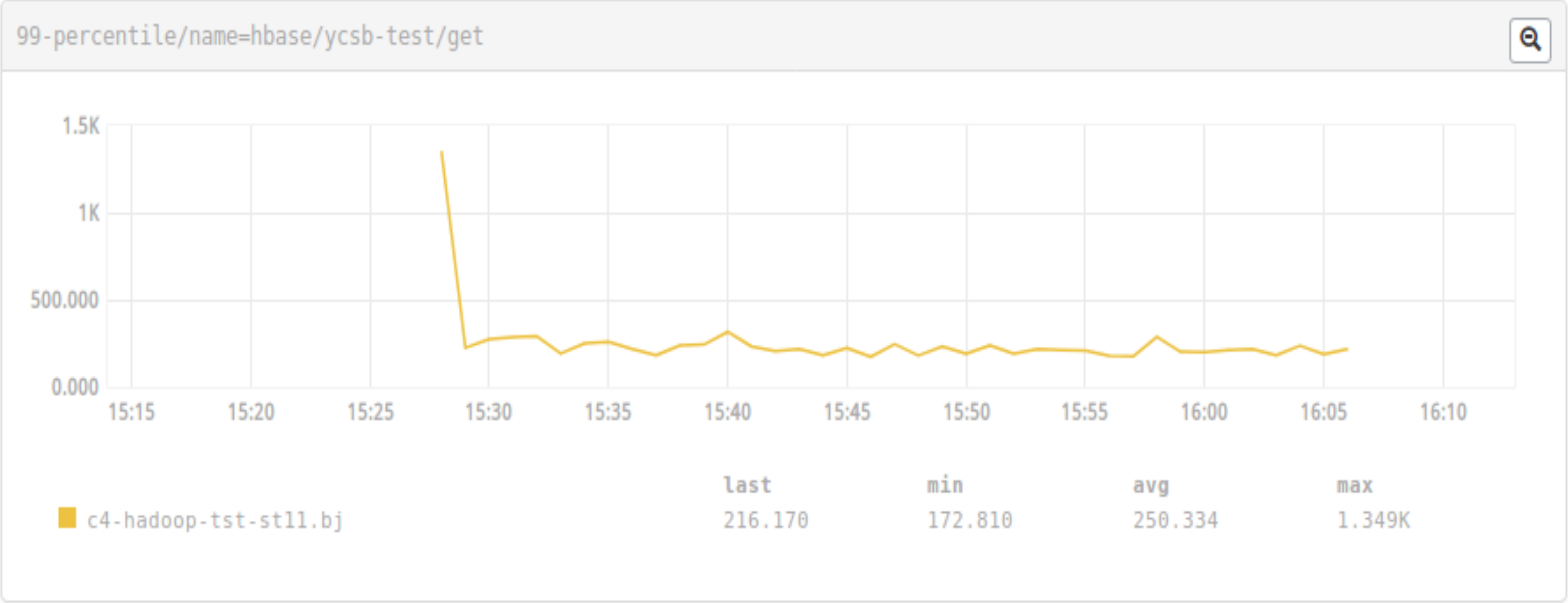
| | last | min | avg | max |
|---|---|---|---|---|
| c4-hadoop-tst-st11.bj | 7.916K | 0.000 | 3.915K | 12.812K |

Before

After


CPS-1-min/name=hbase/ycsb-test/get

| | last | min | avg | max |
|---|---|---|---|---|
| c4-hadoop-tst-st11.bj | 4.4K | 843.827 | 4.137K | 4.542K |


99-percentile/name=hbase/ycsb-test/get

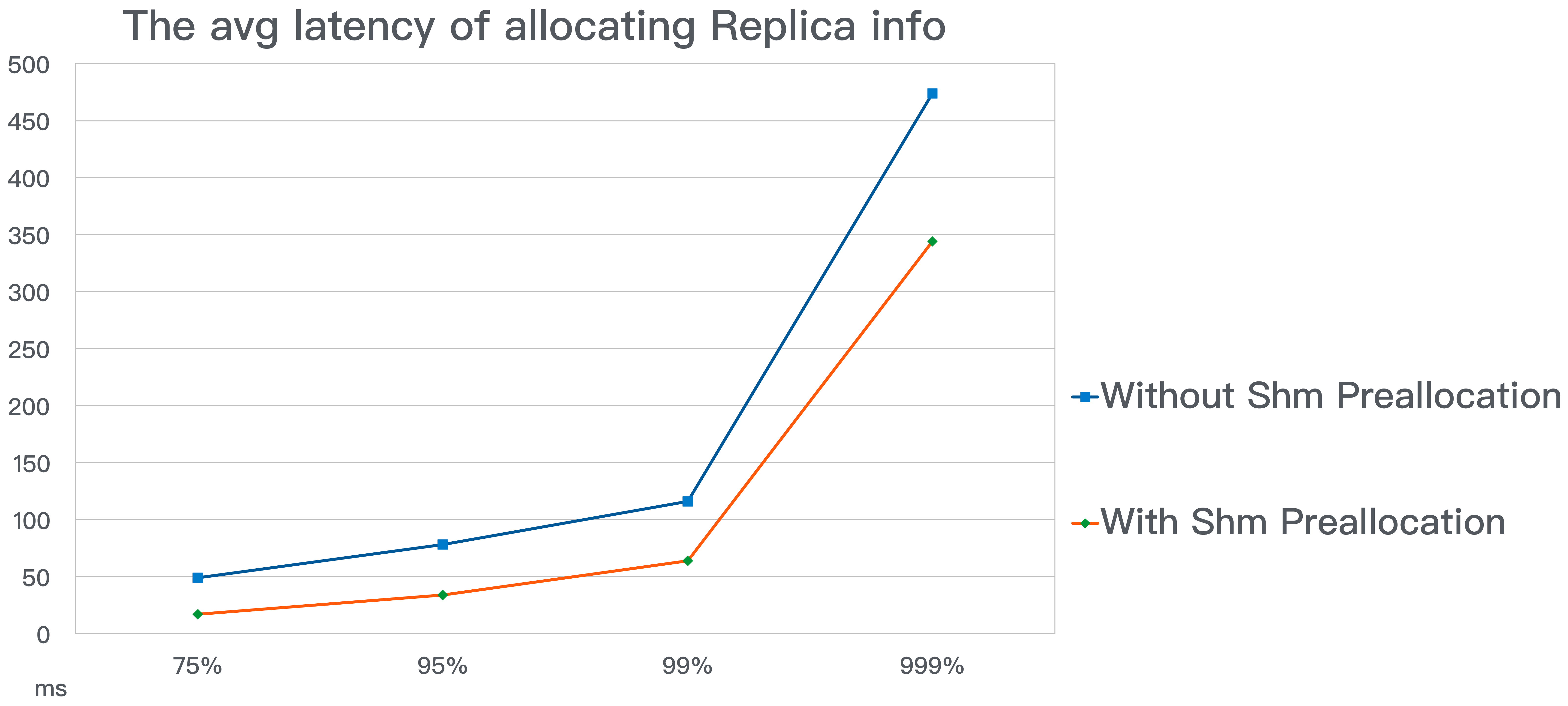
| | last | min | avg | max |
|---|---|---|---|---|
| c4-hadoop-tst-st11.bj | 216.170 | 172.810 | 250.334 | 1.349K |

# Short Circuit Read optimization

✧ Problem #2 : Shm allocation blocks all the slot allocations

Preallocate the Shm:

- 1000 files
- 60 threads
- seek read



# Short Circuit Read optimization

- ✧ Problem #3 : SCR is disabled on under–construction block
- ✧ HDFS–2757 : FileNotFound, block length mismatch
- ✧ Resolution:

1. User ensure read before flush
2. Handle the exception and fall back to remote read.

# HDFS Configuration Best Practice

- Listen drop on SSD cluster causes 3s delay

```
15:56:14.506610 IP x.x.x.x.62393 > y.y.y.y.29402: Flags [S], seq 167786998, win 14600, options [mss 1460,sackOK,TS val
1590620938 ecr 0,nop,wscale 7], length 0<<<--------timeout on first try

15:56:17.506172 IP x.x.x.x.62393 > y.y.y.y.29402: Flags [S], seq 167786998, win 14600, options [mss 1460,sackOK,TS val
1590623938 ecr 0,nop,wscale 7], length 0<<<--------retry

15:56:17.506211 IP y.y.y.y.29402 > x.x.x.x.62393: Flags [S.], seq 4109047318, ack 167786999, win 14480, options [mss
1460,sackOK,TS val 1589839920 ecr 1590623938,nop,wscale 7], length 0
```

- HDFS-9669

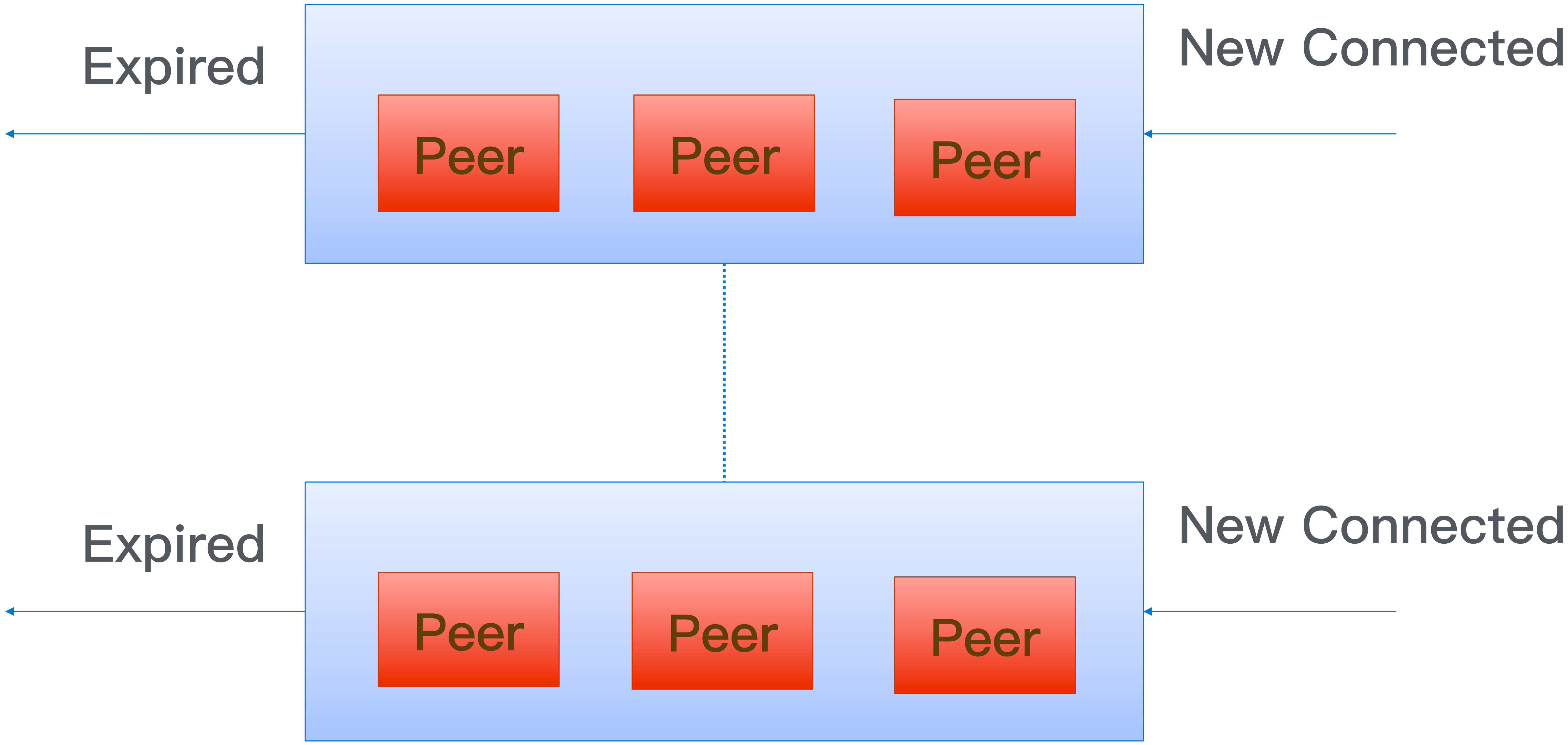
- After change backlog 128, 3s  delays reduced to ~1/10 on Hbase SSD cluster

Somaxconn=128    Default Datanode backlog=50

# HDFS Configuration Best Practice

✧ Peer cache bucket adjustment

Capacity=16 –>100+



# HDFS Configuration Best Practice

✧ Connection/Socket timeout of the DFSClient & Datanode

dfs.client.socket-timeout

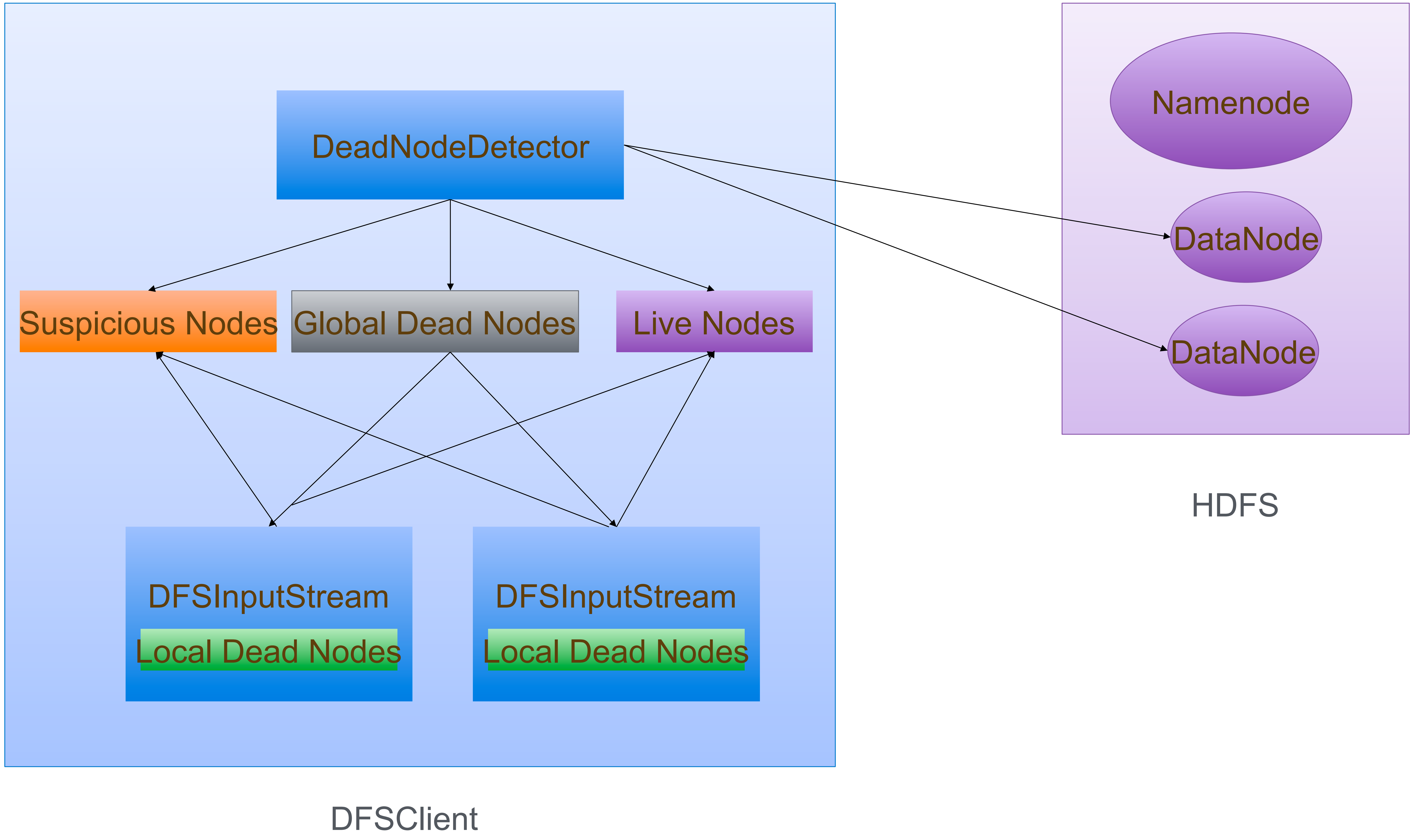
dfs.datanode.socket.write.timeout

- Reduce the timeout to 15s
- Avoid pipeline timeout, upgrade the DFSClient first

# 03 Detect a dead Datanode in advance

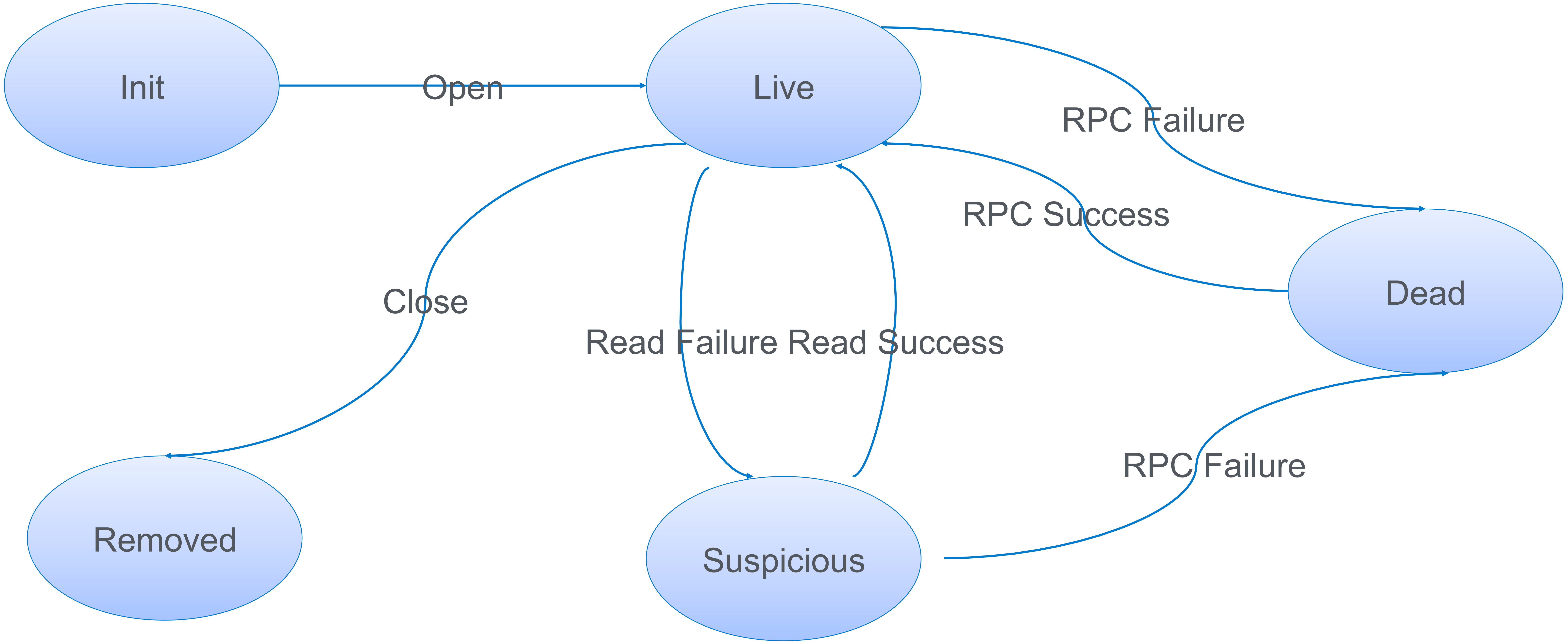
# Detect a dead Datanode in advance

- ✧ 60 seconds timeout if datanode dies
- ✧ Dead nodes not shared
- ✧ “Dead” node not actually dead



# Detect a dead Datanode in advance

- ✧ Node state machine



# Summary

Subtitle Text

## Locality

Maintain the data on local host as much as possible and reduce the over head of the local read

## Quick Response

Make sure the response to Hbase is returned as soon as possible even a failed one

## Less Minor GC

Minor GC from both Hbase and HDFS affects the latency.  Try to easy the one from HDFS on client side.

# Thanks