HBase:1.0 and Beyond

Ted Yu

yuzhihong@gmail.com

About myself

- **Graduated from TsingHua University**
- **Have been working on Hbase for over four years**
- **Have been Hbase PMC member since June 2011**
- **Senior MTS at Hortonworks**

Agenda

- HBase 1.0
- HydraBase:cross DC high availability
- Local Index support in Phoenix
- Per column family flush
- Q & A

Major Changes for 1.0

- Stability: Co-locate hbase:meta with Master
- Simplify, Improve region assignment reliability: Fewer components involved
- Master embeds a RegionServer, hosting only system tables
- Backup masters can be configured to host user tables
- Plumbing is all there, **OFF** by default

- http://issues.apache.org/jira/browse/HBASE-10569

Major Changes for 1.0 (contd)

- Availability: Region Replicas
- Multiple RegionServers host a Region

① One is “primary”, others are “replicas”

② Only primary accepts writes

- Baby step toward quorum reads, writes
- Plumbing is all there, **OFF** by default

- http://issues.apache.org/jira/browse/HBASE-10070
- http://issues.apache.org/jira/browse/HBASE-11183
- http://www.slideshare.net/HBaseCon/features-session-1

Major Changes for 1.0 (contd)

- Usability: Client API changes
- Improved self-consistency
- Simpler semantics
- @InterfaceAudience annotations

- http://s.apache.org/hbase-1.0-api
- https://github.com/ndimiduk/hbase-1.0-api-examples

Client API usage sample

```
Connection conn =
    ConnectionFactory.createConnection(job.getConfiguration());
try {
  UserProvider userProvider =
    UserProvider.instantiate(job.getConfiguration());
  TokenUtil.addTokenForJob(conn, userProvider.getCurrent(), job);
} finally {
  conn.close();
}
```

http://issues.apache.org/jira/browse/HBASE-12493

Major Changes for 1.0 (contd)

- Online config change: ported from 89-fb HBASE-12147
- Automatic tuning of global MemStore and BlockCache sizes
- BucketCache easier to configure
- Pluggable replication endpoint
- Greatly expanded hbase.apache.org/book.html
- Combining mvcc/seqid
- Sundry security, tags, labels improvements

Online / Wire Compatibility

- Direct migration from 0.94 supported

①Similar to upgrade from 0.94 to 0.96: requires downtime

②Not tested yet, will be before release

- RPC is backward-compatible to 0.96

① Enabled mixing clients and servers across versions

- Rolling upgrade "out of the box" from 0.98

① 0.96 cannot read HFileV3, the new default

Client Application Compatibility

- API is backward compatible to 0.96
  1. No code change required
  2. You’ll start getting new deprecation warnings
- ABI is **NOT** backward compatible
  1. Cannot drop current application jars onto new runtime
  2. Recompile your application vs. 1.0 jars
  3. similar to 0.96 to 0.98 upgrade

Hadoop / Java Versions

- Hadoop 1.x is NOT supported
  ① you’ll enjoy the performance benefits
- Hadoop 2.x only
  ① Most thoroughly tested on 2.4.x, 2.5.x
  ② less thoroughly tested on 2.2.x, 2.3.x

- JDK 6 is NOT supported!
- JDK 7 is the target runtime

- https://hbase.apache.org/book/ configuration.html#hadoop

HydraBase

- Goal: 4 9’s of availability in steady state
- No data loss at cluster level failures
- All failures should be quick to recover from
- distributed consensus shouldn’t affect write throughput
- each region will be hosted by a set of Region Servers

HydraBase: Replication Protocol

- There will be only one leader amongst the set of replicas
- Leader serves all the read and write requests to the client
- The election of the leader will be done using the RAFT protocol
- Each replica will have its own Write Ahead Log, stored locally
- Writes will be replicated synchronously by the leader to the replica set

HydraBase: RMAP

- RMap contains the quorum configuration information for each Region
- Based on the network latency to the client, each Data Center will have a rank number
- DC with the lowest network latency to the client, will have the highest rank
- Qualified quorum member with higher DC ranking is able to take over the leadership
- Replica with higher rank (DC-rank + machine-rank) will have a lower election timeout

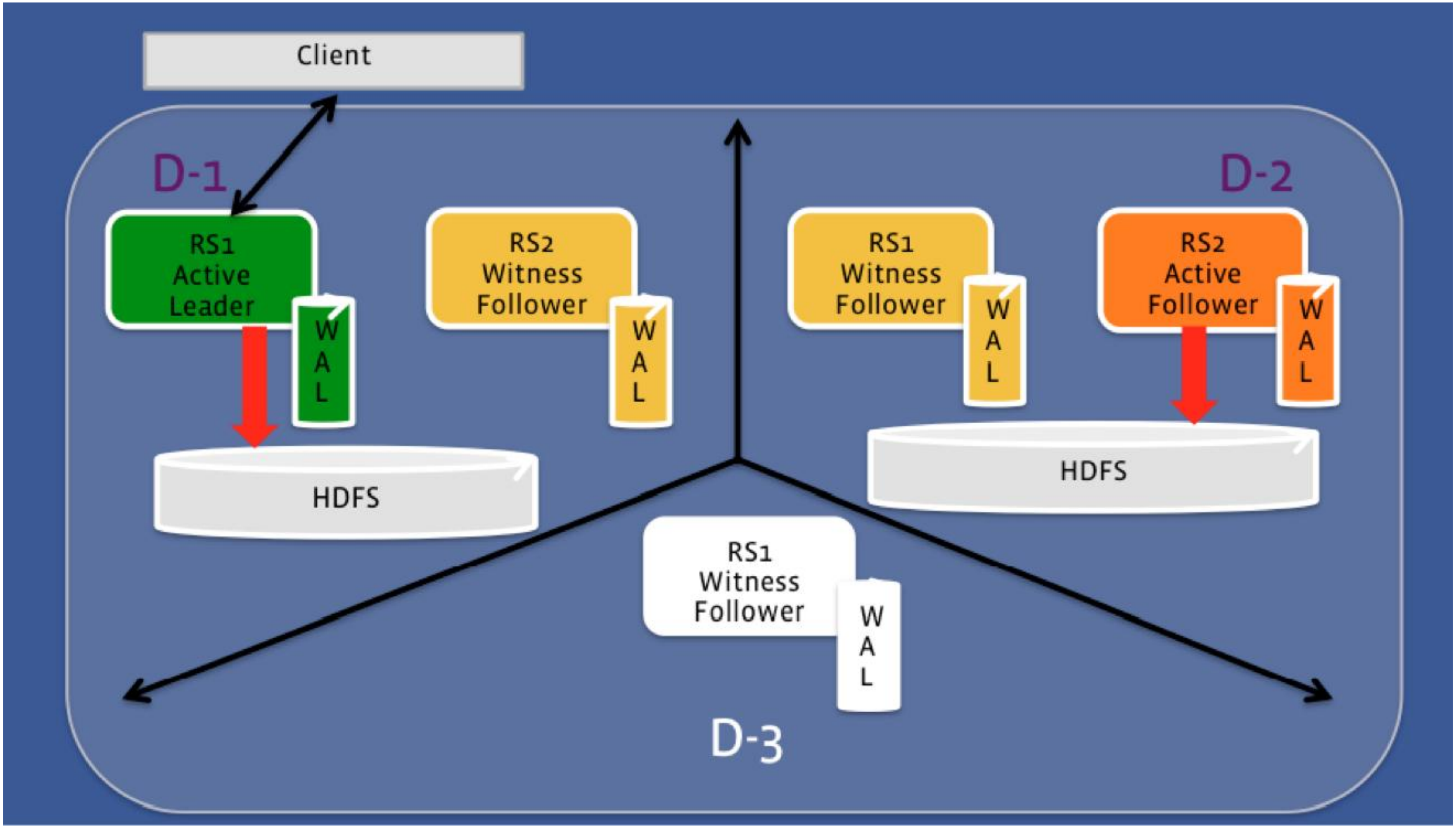
HydraBase: Types of Replicas

- **ACTIVE:** performs all the LSM operations in the RegionServers. This includes flushes and compactions. By default, the current leader is always ACTIVE

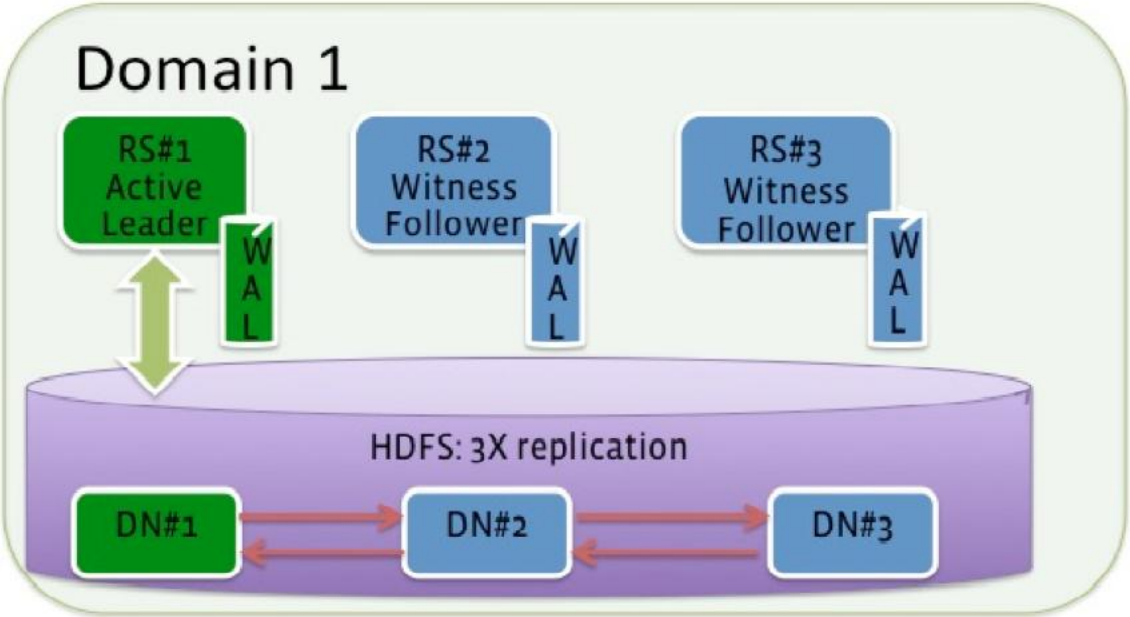
- **ACTIVE-WITNESS:** has a memstore associated with it, but is not performing any LSM operations
- There can be one or more active-witness replicas per HDFS cluster

- **SHADOW-WITNESS:** one who is only participating in the replication via the protocol

HydraBase: Multi Cluster Deployment Setup

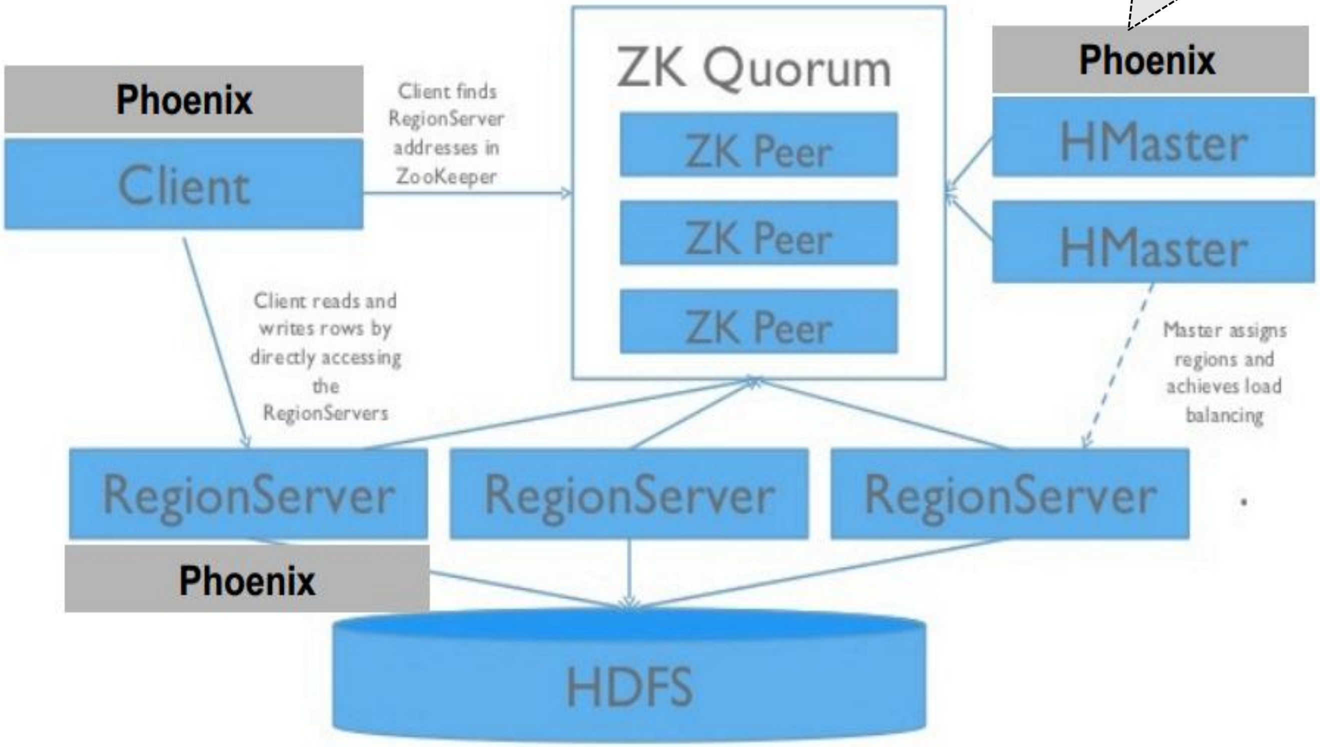


HydraBase: Single Cluster Deployment Setup

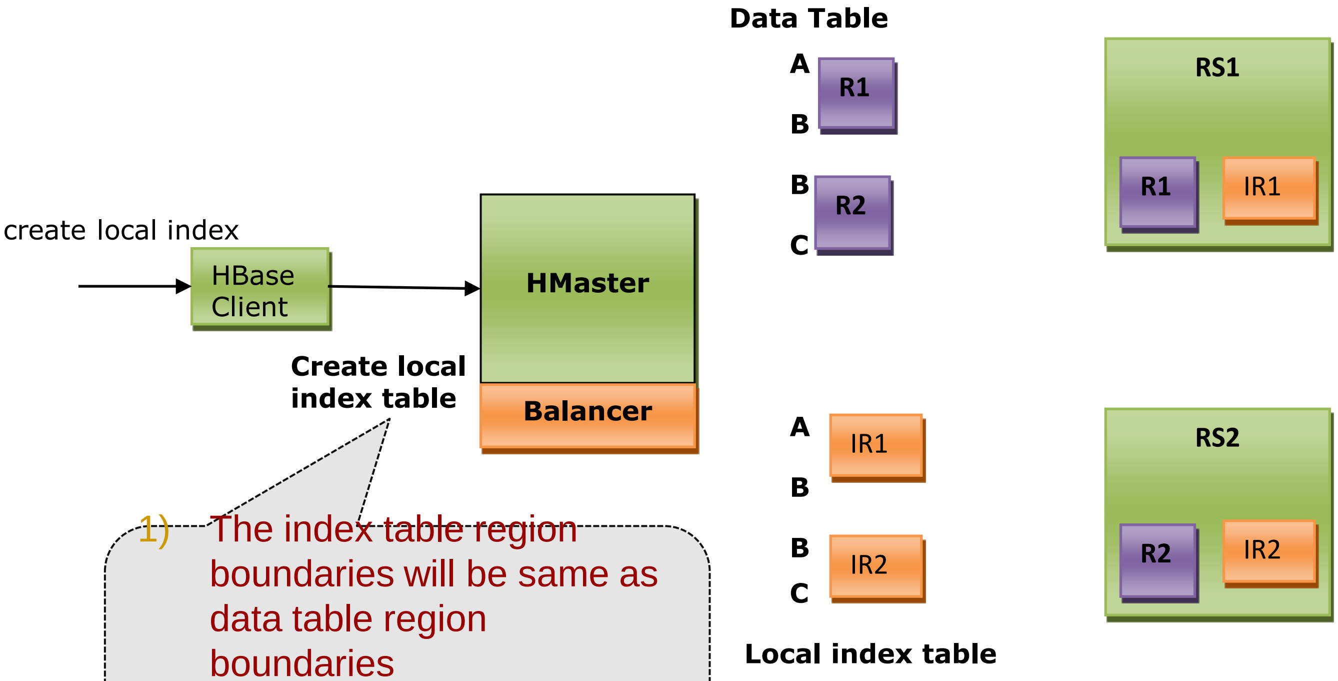


Phoenix Local Index Architecture

To ensure data and index regions co-location
1) Index master observer
2) Custom load balancer


Phoenix
Client
Client finds RegionServer addresses in ZooKeeper
ZK Quorum
ZK Peer
ZK Peer
ZK Peer
Phoenix
HMaster
HMaster
Client reads and writes rows by directly accessing the RegionServers
Master assigns regions and achieves load balancing
RegionServer
RegionServer
RegionServer
Phoenix
HDFS

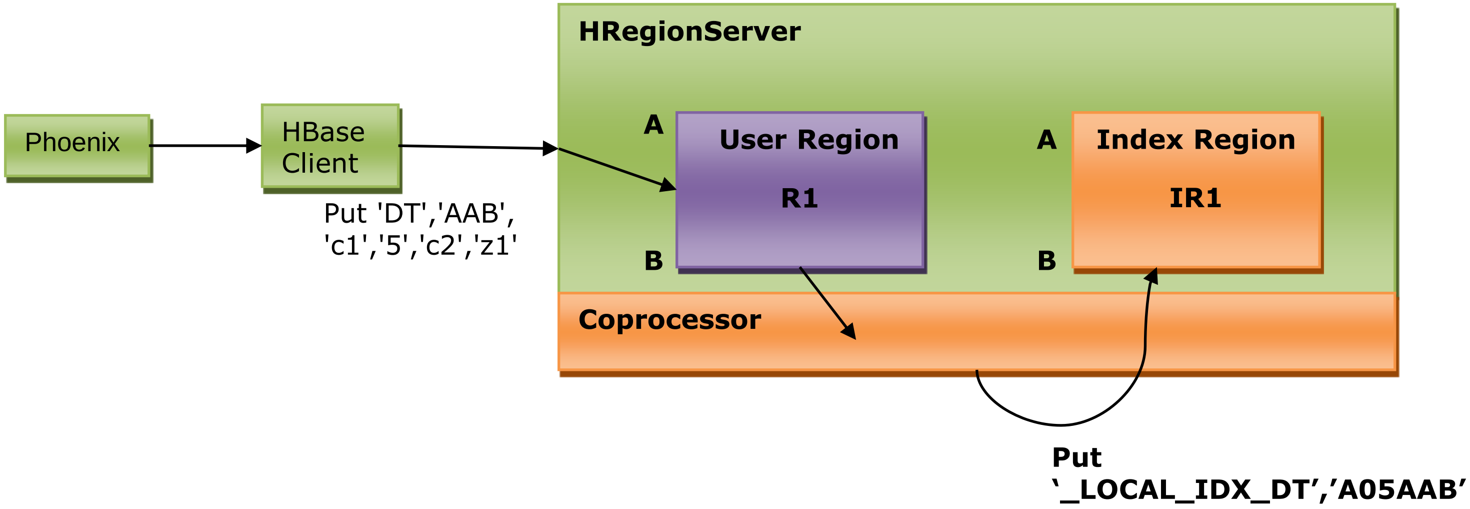
Regions Co-locate



1) The index table region boundaries will be same as data table region boundaries
2) The table will be shared for all local indexes on the data table

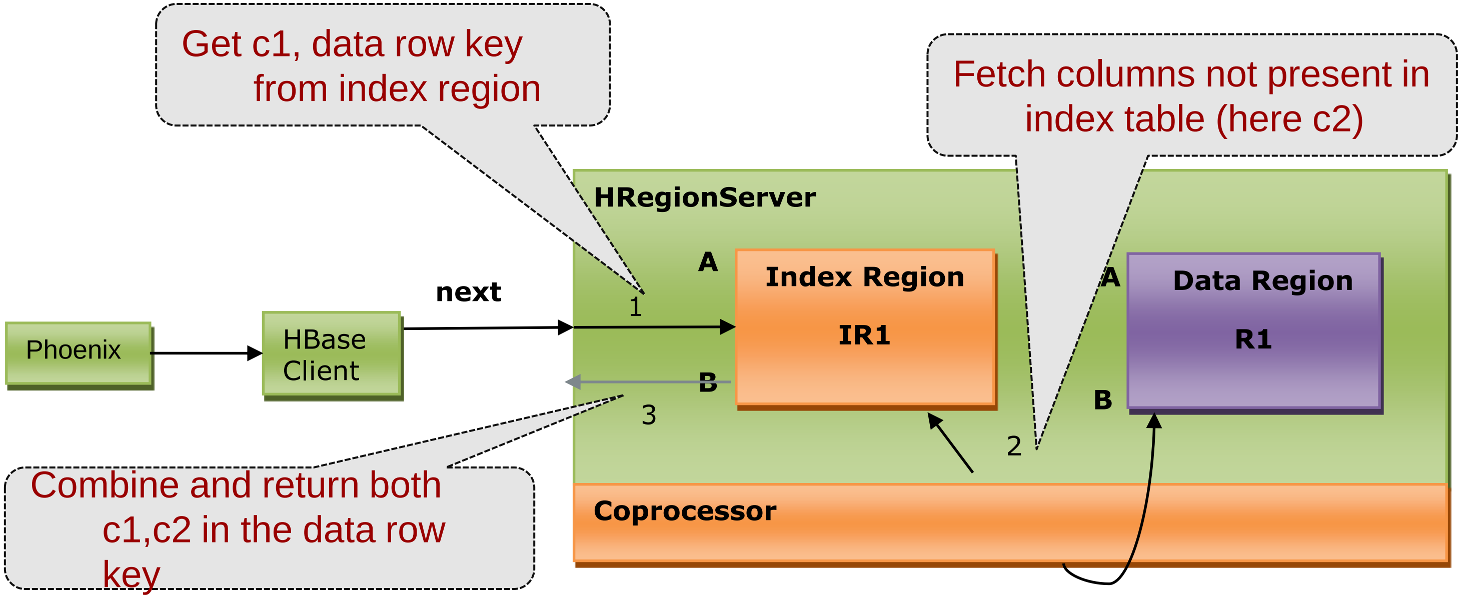
Write path

- Data table DT with columns pk,c1,c2
- Create local index LIDX on DT(c1)
- Local index table -> _LOCAL_IDX_DT



Read path

- **Select c1, c2 from DT where c1 = 5**



Get c1, data row key from index region

Fetch columns not present in index table (here c2)

Combine and return both c1,c2 in the data row key

Local Index Performance

- **Created table pre split (30 regions)**
- **Number of indexes: 4**
- **Data size : 500MB**

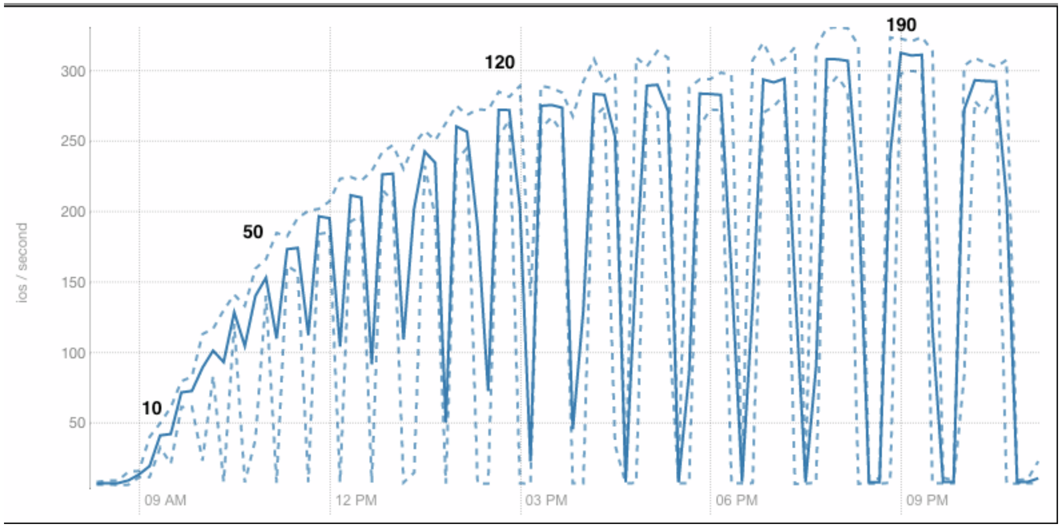
① **No index : 4955 sec**

② **Local mutable indexes : 1152**

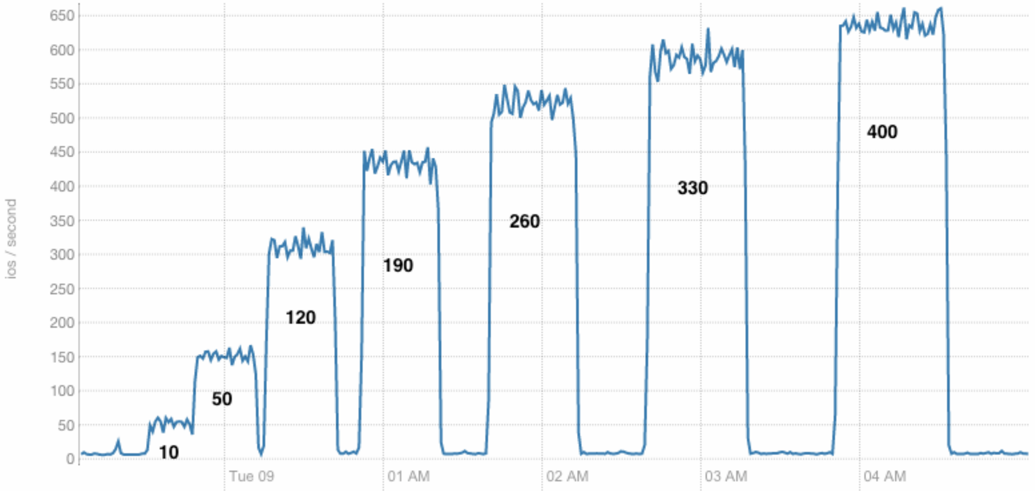
③ **Global mutable indexes : 1679 sec**

Multi-WAL HBASE-5699

- **multiwal-1_1_to_200_threads**



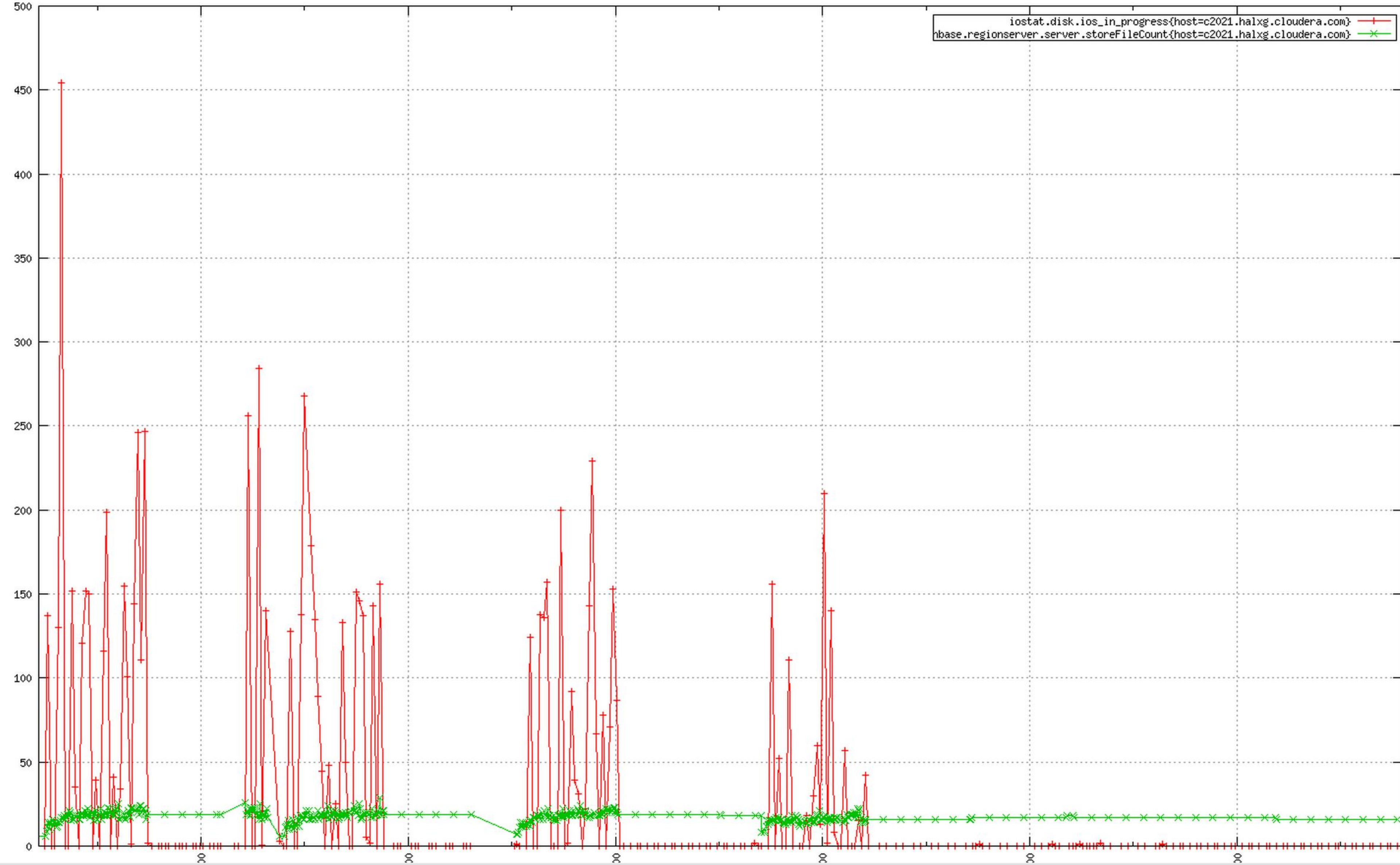
- **multiwal-2-10_to_400_threads**



Per column family flush

- HBASE-10201 Port from 0.89-fb branch
- Reduces write amplification by 10%
- Lower bound on flush size per column family
- FlushPolicy controls whether all stores are flushed
- Make use of sequence Id of each Store

Per column family flush – I/O reduction



Colors

Primary Colors

Hortonworks
Green
R- 105
G- 190
B- 40

Hortonworks
Black
R- 30
G- 30
B- 30

Secondary Colors

Hortonworks
Orange
R- 225
G- 112
B- 0

Hortonworks
Dusty Blue
R- 68
G- 105
B- 125

Hortonworks
Gray
R- 129
G- 138
B- 143

For PPT Only

Darker
Text Gray
R- 127
G- 127
B- 127

Simple Slide: Arial, 36pt, Left Justified

- **Bulleted Body Text:**
  Arial, 18pt, Bold
  - Second Level Sub Bullet: Arial 16pt
    - Third Level Sub Bullet: Arial 14 pt

- **Bulleted Body Text:**
  Arial, 18pt, Bold
  - Second Level Sub Bullet: Arial 16pt
    - Third Level Sub Bullet: Arial 14 pt

- **Bulleted Body Text:**
  Arial, 18pt, Bold
  - Second Level Sub Bullet: Arial 16pt
    - Third Level Sub Bullet: Arial 14 pt

Transition Slide: Arial 54pt

Transition Slide Sub Title: Arial 28pt

- **Bulleted Body Text:**
  Arial, 18pt, Bold
  - Second Level Sub Bullet: Arial 16pt
    - Third Level Sub Bullet: Arial 14 pt
- **Bulleted Body Text:**
  Arial, 18pt, Bold
  - Second Level Sub Bullet: Arial 16pt
    - Third Level Sub Bullet: Arial 14 pt
- **Bulleted Body Text:**
  Arial, 18pt, Bold
  - Second Level Sub Bullet: Arial 16pt
    - Third Level Sub Bullet: Arial 14 pt

Closing Slide: Arial, 54pt

Closing Sub Title: Arial, 28pt