All about ZooKeeper (and ClickHouse Keeper, too!)

Robert Hodges and Altinity Engineering
30 March 2022

Altinity

Let’s make some introductions

Robert Hodges

Database geek with 30+ years on DBMS systems. Day job: Altinity CEO

Altinity Engineering

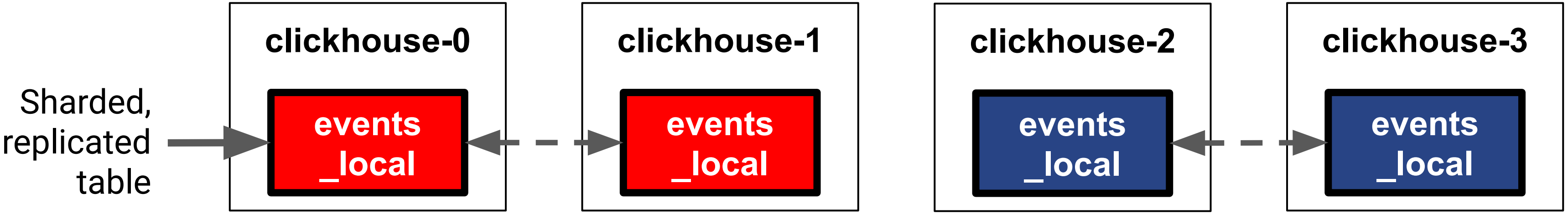
Database geeks with centuries of experience in DBMS and applications

Altinity

ClickHouse support and services including Altinity.Cloud
Authors of Altinity Kubernetes Operator for ClickHouse
and other open source projects

Why does ClickHouse need ZooKeeper?

Horizontal scaling is a key to ClickHouse performance



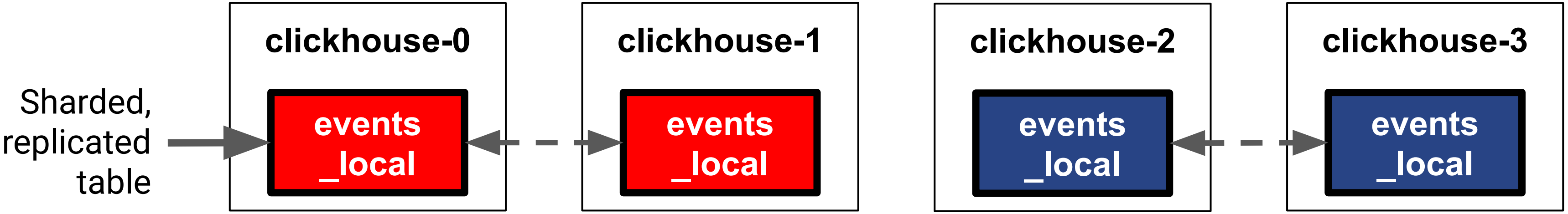
Let’s create the table and try it out!

```
CREATE TABLE IF NOT EXISTS `events_local` ON CLUSTER '{cluster}' (
  EventDate DateTime, CounterID UInt32, Value String
)
Engine=ReplicatedMergeTree(
'/clickhouse/{cluster}/tables/{shard}/{database}/events_local',
'{replica}')
PARTITION BY toYYYYMM(EventDate)
ORDER BY (CounterID, EventDate, intHash32(UserID))

INSERT INTO events_local(EventDate, EventID, Value) VALUES
    (now(), 1, 'In-Progress'), (now(), 2, 'OK')

. . .
```

What could possibly go wrong?



Sharded, replicated table

clickhouse-0

events _local

clickhouse-1

events _local

clickhouse-2

events _local

clickhouse-3

events _local

☠

ON CLUSTER command failed on one node!

☠

Two replicas merge overlapping parts!

☠

Node offline for maintenance; missed the memo!

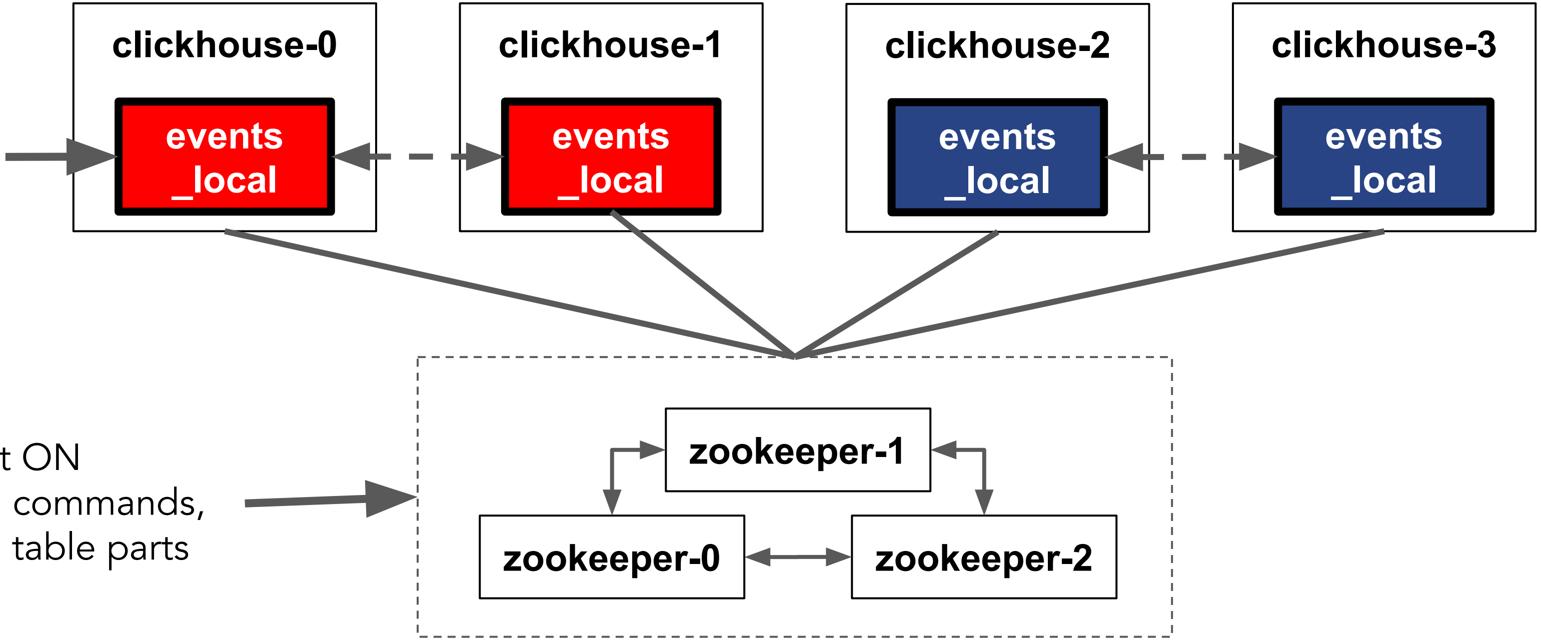
☠

Two replicas delete overlapping parts!

Altinity

ZooKeeper solves the distributed consistency problem

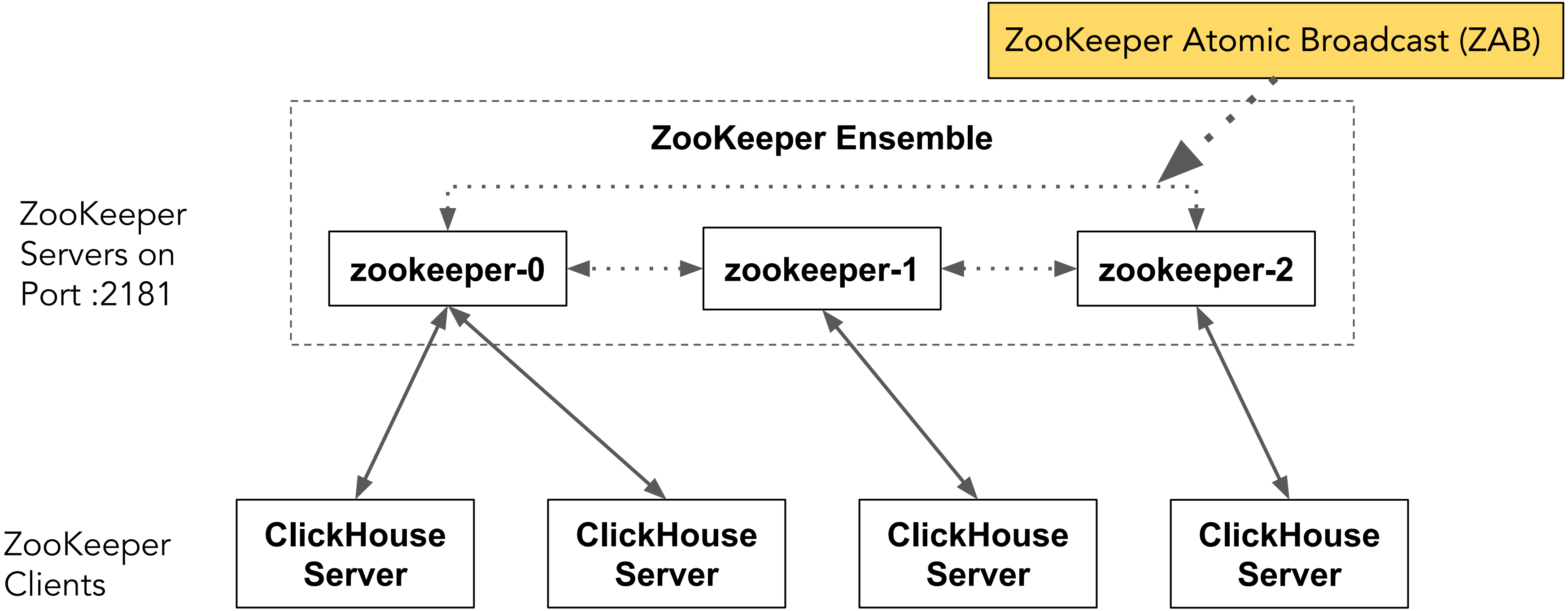
Sharded, replicated table



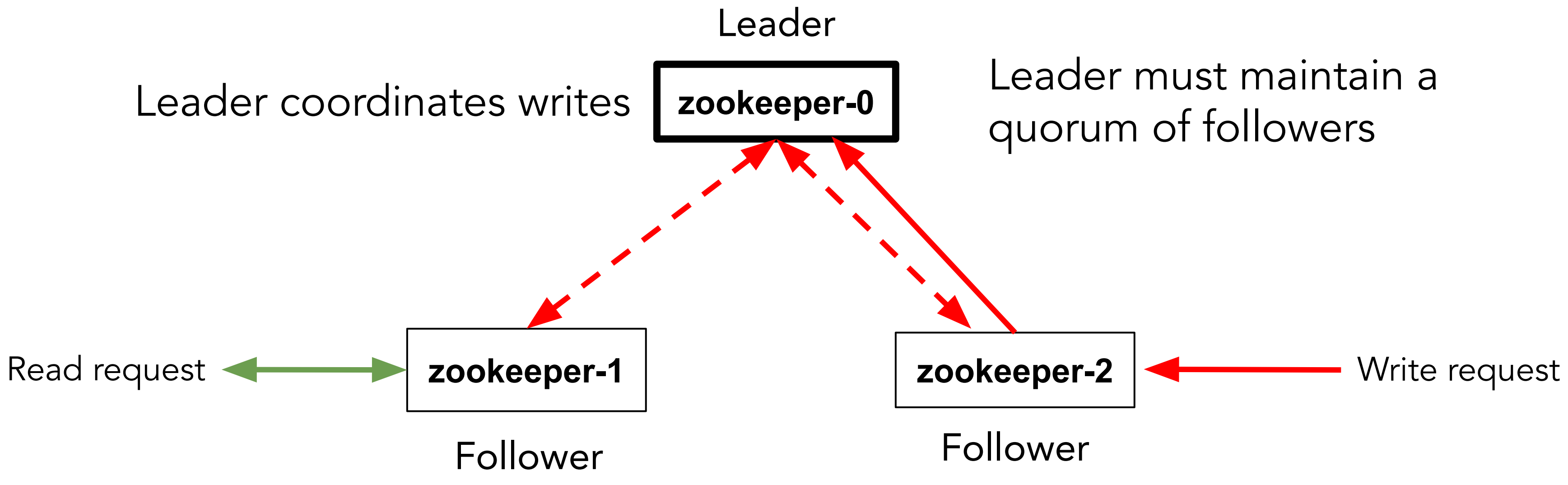
Consistent ON CLUSTER commands, replicated table parts

How ZooKeeper Works

ZooKeeper Architecture

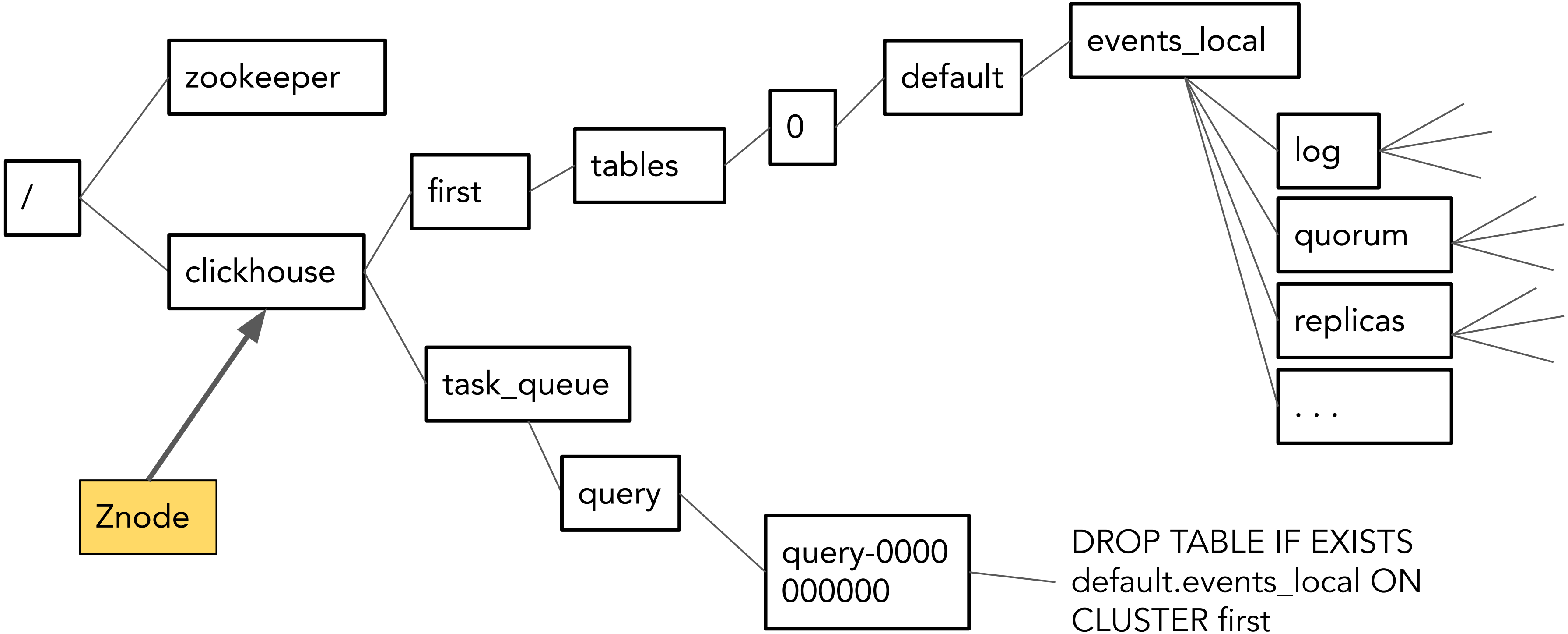


ZooKeeper leaders and followers



Followers handle reads and delegate writes to leader

ZooKeeper directory structure for ClickHouse



DROP TABLE IF EXISTS default.events_local ON CLUSTER first

What kind of ClickHouse information is stored in znodes?

- Tasks
  - Pending and completed ON CLUSTER DDL commands
- Table information
  - Schema information
  - Replicas
  - Leader elections used to control merges and mutations
  - Log of operations on the table (insert, merge, delete partition, etc.)
  - Parts contained in each replica
  - Last N blocks inserts so we can deduplicate data
  - Data to ensure quorum on writes

Installing and configuring ZooKeeper

Installing a ZooKeeper on Ubuntu

Install Zookeeper 3.4.9 or greater

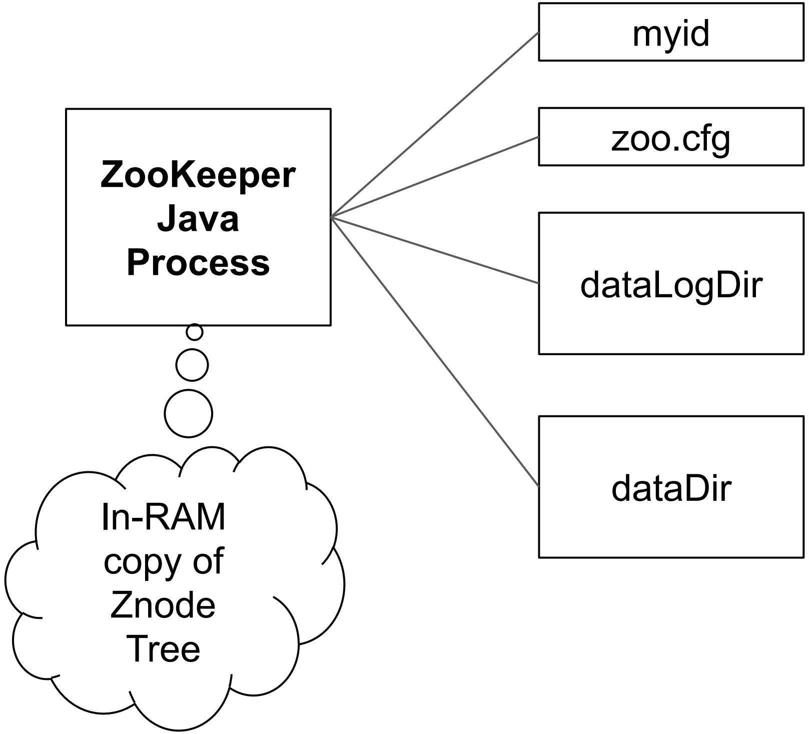
```
sudo apt update
sudo apt install zookeeper netcat
(edit /etc/sysconfig/config/zoo.cfg to set configuration)
```

Ensuring ZooKeeper maximum speed and availability

Host recommendations

- Dedicated host for ZooKeepers - don’t share with other applications
- Put ZooKeeper log on dedicated SSD
- Low network latency between ZooKeeper nodes
- At least 4GiB of RAM
- Disable swap (remove entry from /etc/fstab)
- Tune the Java heap to use as much RAM as possible
  - E.g., 3GiB out of 4GiB available RAM

ZooKeeper moving parts



ZooKeeper Java Process

In-RAM copy of Znode Tree

myid — Server instance ID (1-255)

zoo.cfg — Configuration file

dataLogDir — Logs with changes to data

dataDir — Snapshots of data

Editing important zoo.cfg settings

☠

```
...
autopurge.purgeInterval=1
autopurge.snapRetainCount=5
...
server.1=zookeeper1:2888:3888
server.2=zookeeper2:2888:3888
server.3=zookeeper3:2888:3888
...
dataDir=/var/lib/zookeeper
...
dataLogDir=/ssd/zookeeper/logs
```

Must be added; prevents snapshots from accumulating

Servers in ensemble; must be identical everywhere

Location for snapshots

Put logs on fast storage

Starting ZooKeeper and ensuring it's up

```
sudo -u zookeeper /usr/share/zookeeper/bin/zkServer.sh
ZooKeeper JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... STARTED
echo ruok | nc localhost 2181
imok
echo mntr | nc localhost 2181
zk_version      3.4.10-3--1, built on Sat, 03 Feb 2018 14:58:02 -0800
. . .
echo stat | nc localhost 2181
zk_version      3.4.10-3--1, built on Sat, 03 Feb 2018 14:58:02 -0800
. . .
```

Tell ClickHouse where ZooKeeper lives

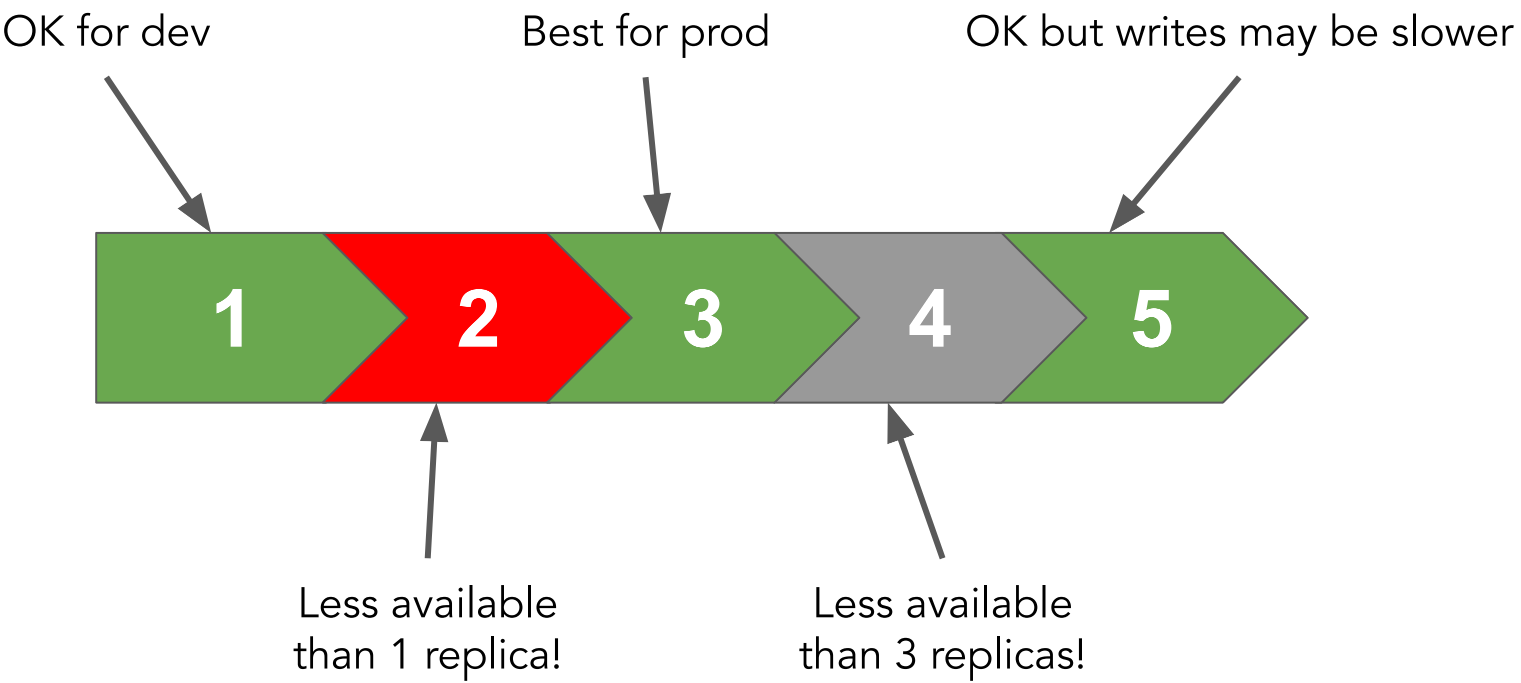
```
<yandex>
    <zookeeper>
        <node>
            <host>zookeeper.zoo1ns</host>
            <port>2181</port>
        </node>
    </zookeeper>
    <distributed_ddl>
        <path>/clickhouse/first/task_queue/ddl</path>
    </distributed_ddl>
</yandex>
```

Add macros so ON CLUSTER commands can run

```
<yandex>
    <macros>
        <installation>first</installation>
        <all-sharded-shard>0</all-sharded-shard>
        <cluster>first</cluster>
        <shard>0</shard>
        <replica>chi-first-first-0-0</replica>
    </macros>
</yandex>
```

Practical Administration Tips

How many ZooKeepers are enough?


OK for dev
Best for prod
OK but writes may be slower
1
2
3
4
5
Less available than 1 replica!
Less available than 3 replicas!

What’s in ZooKeeper? The system.zookeeper table knows!

```
SELECT * FROM system.zookeeper WHERE path = '/'
ORDER BY name FORMAT Vertical
```

```
Row 1:
──────
name:           clickhouse
value:
czxid:          4294967298
mzxid:          4294967298
ctime:          2021-12-08 01:54:50
mtime:          2021-12-08 01:54:50
. . .
path:           /
```

Path value is required!

If this query works, ClickHouse can see ZooKeeper!

Printing znode values from system.zookeeper

```
SELECT name, value FROM system.zookeeper
WHERE path = '/clickhouse/first/task_queue/ddl/'
FORMAT Vertical
```

Prints values for znodes under this path

```
Row 1:
──────
name:  query-0000000009
value: version: 1
query: CREATE TABLE IF NOT EXISTS default.events_local UUID
\'2a8ed83e-a6ef-48b4-aa8e-d83ea6efa8b4\' ON CLUSTER first (`EventDate`
DateTime, `EventID` UInt32, `Value` String) ENGINE =
ReplicatedMergeTree(\'/clickhouse/{cluster}/tables/{shard}/{database}/even
ts_local\', \'{replica}\') PARTITION BY toYYYYMM(EventDate) ORDER BY
(CounterID, EventDate, intHash32(UserID))
hosts:
. . .
```

Using the zkCli utility to talk to ZooKeeper directly

```
(Connect to ZooKeeper host)
$ zkCli.sh
Connecting to localhost:2181
. . .
[zk: localhost:2181(CONNECTED) 0] ls /
[clickhouse, zookeeper]
[zk: localhost:2181(CONNECTED) 1] get
/clickhouse/first/task_queue/ddl/query-0000000009
version: 1
query: CREATE TABLE IF NOT EXISTS default.events_local UUID
\'2a8ed83e-a6ef-48b4-aa8e-d83ea6efa8b4\' ON CLUSTER first . . .
```

ZooKeeper four letter word commands

Example: `echo ruok | nc localhost 2181` → `imok`

Command	What it does
ruok	Check server liveness
conf	Print server config
cons	Print connections
mntr	Dump monitoring information
srvr	Dump server information

There are more commands! Check the docs.

ZooKeeper Monitoring

Older approach for Nagios and Icinga[2]

- Use check_zookeeper.pl

Newer approach: Use Prometheus + AlertManager + Grafana

- ZooKeeper by Prometheus Dashboard for Grafana

The Altinity Knowledge Base has a page on ZooKeeper Monitoring

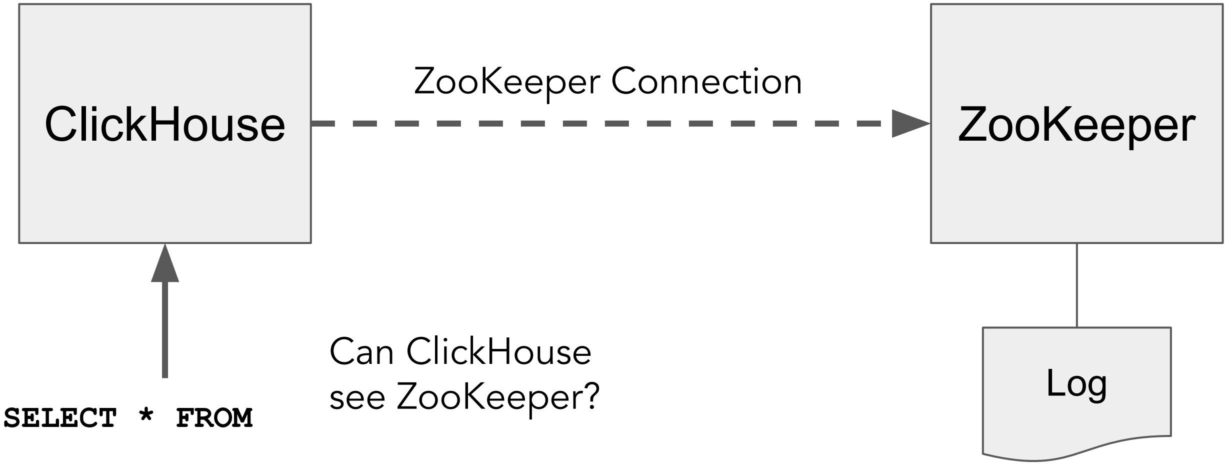
The dreaded read-only table error

```
INSERT INTO events2_local (EventDate, EventID, Value)
  VALUES (now(), 1, 'In-Progress'), (now(), 2, 'OK')
```

```
Received exception from server (version 21.8.10):
Code: 242. DB::Exception: Received from 34.83.194.130:9000.
DB::Exception: Table is in readonly mode (zookeeper path:
/clickhouse/first/tables/0/default/events2_local).
(TABLE_IS_READ_ONLY)
```

ZooKeeper is offline!

Steps to address read-only tables



Is ZooKeeper alive?

```
echo ruok | nc localhost 2181
```

```
SELECT * FROM
system.zookeeper
WHERE path = '/'
```

Can ClickHouse
see ZooKeeper?

Loss of quorum?
ZXID Rollover?
Other errors?

/var/log/zookeeper/zookeeper.log

ZooKeeper "Session Expired" errors

If ClickHouse loses its connection to ZooKeeper, pending INSERTs or ON CLUSTER commands may fail with a Session Expired error.

1. Occasional failure is normal in distributed systems. Retry the operation!!
2. If the problem happens commonly, you may have a ZooKeeper problem.
   a. Check ZooKeeper logs for errors
   b. This could be an ZXID overflow due to too many transactions on ZooKeeper. Check that only ClickHouse is using ZooKeeper!
   c. Too many parts in the table? (> 5000)
   d. Jute.maxbuffer seting on ZooKeeper is too low.

Recovering from failures

Loss of a single ZooKeeper node

1. Create fresh node with same ZooKeeper instance ID as lost node
2. Ensure new host name is correct in all zoo.cfg files
3. Start new node

Loss of entire ZooKeeper ensemble

1. Briefly consider taking an immediate vacation
2. Bring up new ZooKeeper ensemble
3. Use SYSTEM RESTORE REPLICA command to restore metadata from ClickHouse server(s)

ClickHouse Keeper

So…What is ClickHouse Keeper?

It’s a from-scratch reimplementation of ZooKeeper

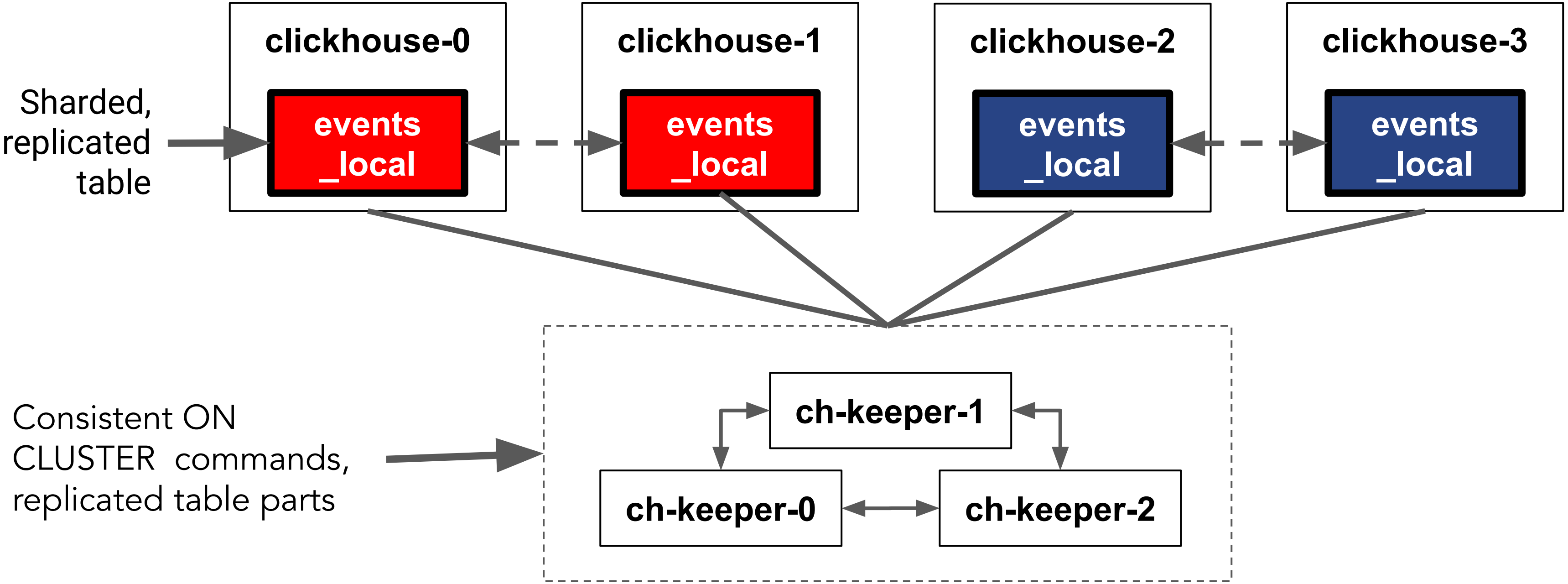
- Mimics ZooKeeper API and admin commands
- Uses Raft protocol instead of ZAB for consensus
- Is written in C++
- Is part of ClickHouse

No extra installation required!

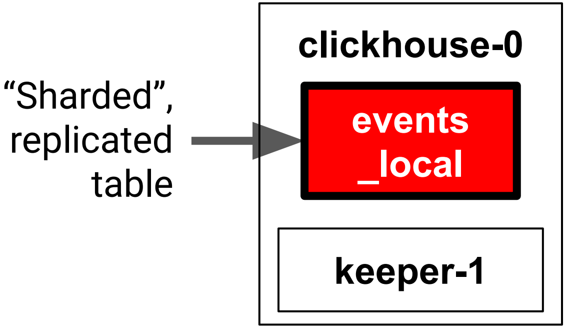
Why replace ZooKeeper?

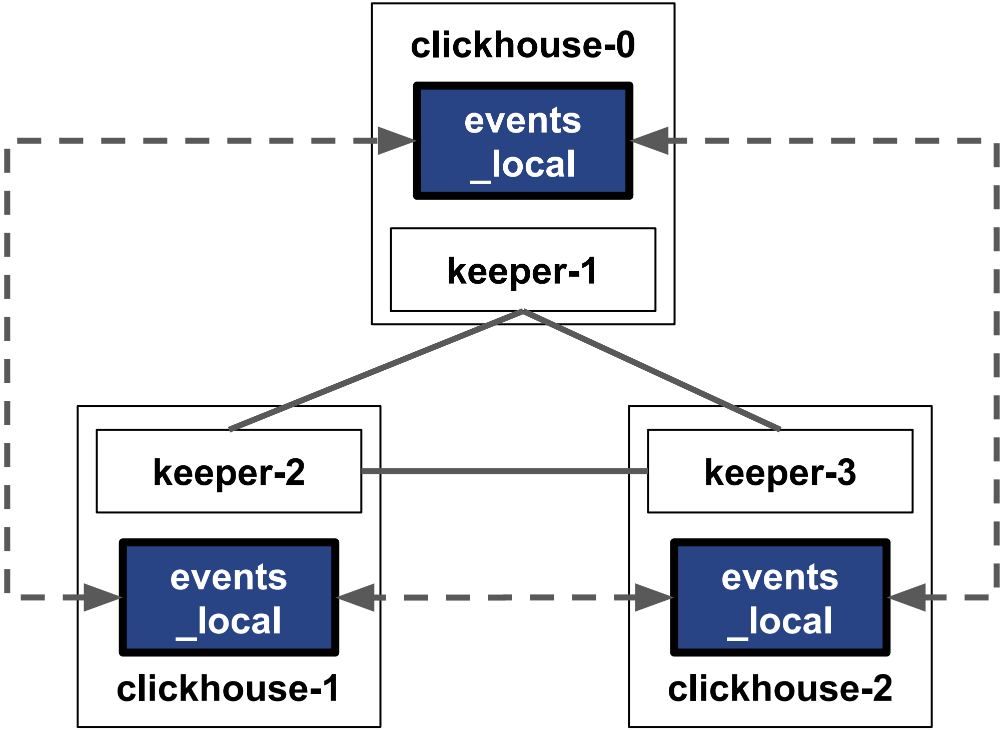
- ClickHouse should contain everything it needs to run
- Old, not very actively developed
- Java executable adds dependencies and requires tuning
- Many people find it hard to operate
- Problems like ZXID rollover, uncompressed logs, etc.

ClickHouse Keeper can be a drop-in ZK replacement…



Or it can run directly in ClickHouse itself!





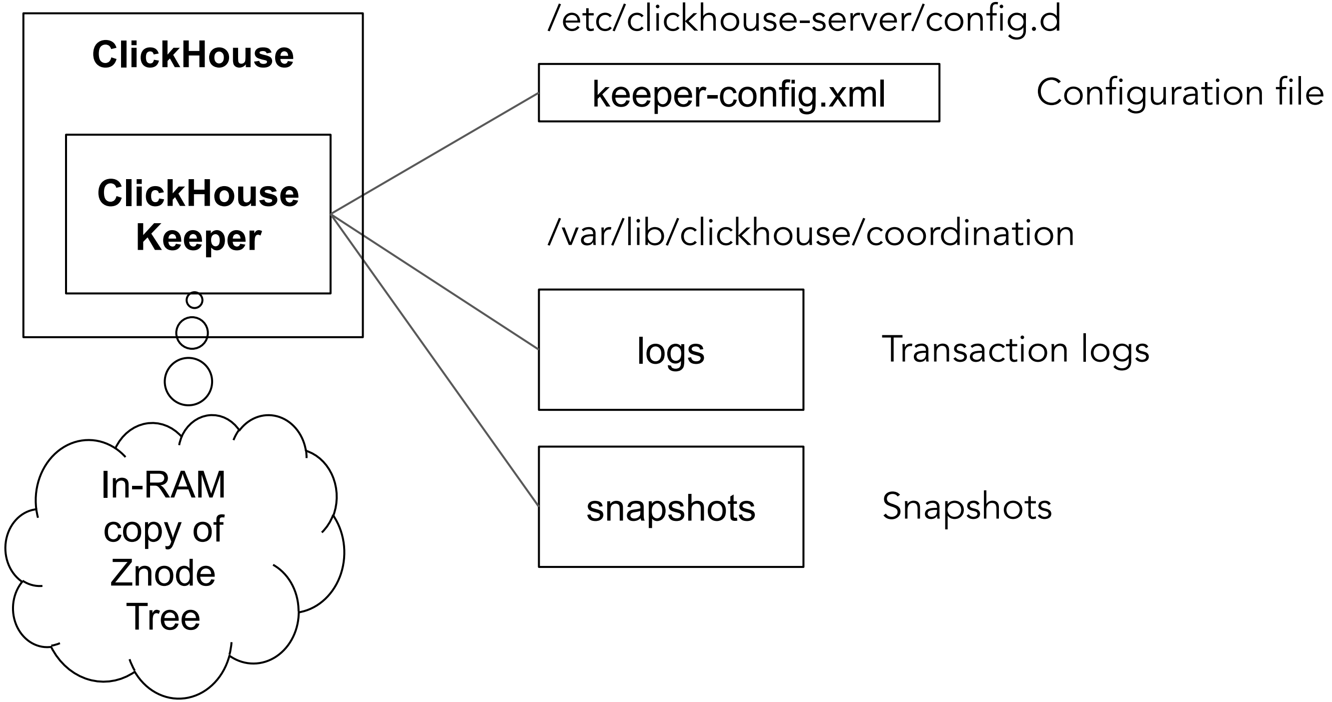
Altinity

© 2022 Altinity, Inc.

ClickHouse Keeper single node configuration

```
<yandex>
    <keeper_server incl="keeper_server">
        <server_id>1</server_id>
        <tcp_port>2181</tcp_port>
        <coordination_settings>
            <raft_logs_level>debug</raft_logs_level>
        </coordination_settings>
        <raft_configuration>
            <server>
                <id>1</id>
                <hostname>logos3</hostname><port>9444</port>
            </server>
        </raft_configuration>
</keeper_server> </yandex>
```

ClickHouse Keeper moving parts for single node install



ClickHouse Keeper “just works”

1. ON CLUSTER commands and replication work exactly as before
2. System.zookeeper table shows directory structure
3. ZooKeeper four letter commands work
4. You can use zkCli.sh (and other tools) to navigate the directory structure

How to tell you are using ClickHouse Keeper

```
$ echo srvr |netcat logos3 2181
ClickHouse Keeper version:
v22.3.2.1-prestable-92ab33f560e638d1989c5ca543021ab53d110f5c
Latency min/avg/max: 0/0/12
Received: 1456
Sent : 1457
Connections: 1
Outstanding: 0
Zxid: 405
Mode: standalone
Node count: 54
```

How do I migrate from ZooKeeper to ClickHouse Keeper?

Clickhouse-keeper-converter converts ZooKeeper logs and snapshots.

Procedure for migration:

1. Stop ZooKeeper ensemble.
2. Restart the ZooKeeper leader node to create a consistent snapshot.
3. Run clickhouse-keeper-converter
4. Copy to ClickHouse Keeper snapshot directory and start ClickHouse Keeper

Test the procedure carefully before applying to production systems.

Is ClickHouse Keeper ready for prime time?

It’s getting there.

ClickHouse Keeper is much more convenient for developers

It fixes a number of known problems like ZKID overflow

There will be glitches but our experience is ‘so far, so good’

ClickHouse Keeper is ready for prod use on 22.3

References

List of references for more information

ZooKeeper Docs: https://zookeeper.apache.org/

ClickHouse Docs: https://clickhouse.com/docs/

Altinity Knowledge Base: https://kb.altinity.com/

Altinity Docs: https://docs.altinity.com

Alexander Sapin ClickHouse Keeper talk:
https://www.slideshare.net/Altinity/clickhouse-keeper

Thank you!

Questions?

https://altinity.com

info@altinity.com