etcd metrics
etcd 度量指标概览

Jingyi Hu 胡景懿 (Google)
Wenjia Zhang 张文嘉 (Google)

Speakers

- Jingyi Hu 胡景懿 (Google)
  - **etcd maintainer,** kubernetes member
  - github/jingyih
  - jingyih@google.com
- Wenjia Zhang 张文嘉 (Google)
  - etcd contributor, kubernetes member
  - github/wenjiaswe
  - wenjiazhang@google.com

Agenda

- etcd metrics port
- Documented metrics
- New metrics
- How to analyze etcd metrics

Wednesday, June 26 • 11:20 - 11:55

深入了解：etcd - Jingyi Hu，Google

Click here to remove from My Sched.

https://sched.co/Nrg3 Tweet Share

作为一个分布式键值存储，etcd 是 Kubernetes 控制平面中最关键的组件，为集群元数据提供了强大的一致性和持久性。 etcd 实施了 Raft 共识算法，以跨多个节点分发数据。所有数据复制都由 Raft 完成。 您是否知道，etcd Raft 软件包也被用于许多其他项目？CockroachDB 为其组成员协议分享 etcd Raft 实施。TiKV 将 etcd Raft 接入 Rust（最初在 Go 中编写），并将其用于实施分布式事务数据库。 本演讲将介绍 Raft 共识算法的基础知识、其实施细节以及未来的 Raft 软件包路线图。

Speakers

Jingyi Hu
Software Engineer, Google

Jingyi Hu is a Software Engineer for Google Cloud. He is a maintainer of etcd and an active contributor to Kubernetes.

etcd metrics port
Etcd 度量指标接口

etcd metrics port 监控指标接口

Each etcd server exports metrics under the /metrics path on its client port and optionally on locations given by --listen-metrics-urls.

- $ curl -L http://localhost:2379/metrics
- --listen-metrics-url http://localhost:9379
  - $ curl -L http://localhost:9379/metrics

Documented metrics

https://github.com/etcd-io/etcd/blob/master/Documentation/metrics.md

etcd_server_ 服务器状态指标

Name	Description	Type
has_leader	Whether or not a leader exists. 1 is existence, 0 is not.	Gauge
leader_changes_seen_total	The number of leader changes seen.	Counter
proposals_committed_total	The total number of consensus proposals committed.	Gauge
proposals_applied_total	The total number of consensus proposals applied.	Gauge
proposals_pending	The current number of pending proposals.	Gauge
proposals_failed_total	The total number of failed proposals seen.	Counter

etcd_disk_ 硬盘状态指标

Name	Description	Type
wal_fsync_duration_seconds	The latency distributions of fsync called by wal	Histogram
backend_commit_duration_seconds	The latency distributions of commit called by backend.	

etcd_network_ 网络状态指标

Name	Description	Type
peer_sent_bytes_total	The total number of bytes sent to the peer with ID `To`.	Counter(To)
peer_received_bytes_total	The total number of bytes received from the peer with ID `From`.	Counter(From)
peer_sent_failures_total	The total number of send failures from the peer with ID `To`.	Counter(To)
peer_received_failures_total	The total number of receive failures from the peer with ID `From`.	Counter(From)
peer_round_trip_time_seconds	Round-Trip-Time histogram between peers.	Histogram(To)
client_grpc_sent_bytes_total	The total number of bytes sent to grpc clients.	Counter
client_grpc_received_bytes_total	The total number of bytes received to grpc clients.	Counter

etcd_network_ 网络状态指标

Name	Description	Type
peer_sent_bytes_total	The total number of bytes sent to the peer with ID `To`.	Counter(To)
peer_received_bytes_total	The total number of bytes received from the peer with ID `From`.	Counter(From)
peer_sent_failures_total	The total number of send failures from the peer with ID `To`.	Counter(To)
peer_received_failures_total	The total number of receive failures from the peer with ID `From`.	Counter(From)
peer_round_trip_time_seconds	Round-Trip-Time histogram between peers.	Histogram(To)
client_grpc_sent_bytes_total	The total number of bytes sent to grpc clients.	Counter
client_grpc_received_bytes_total	The total number of bytes received to grpc clients.	Counter

New metrics
新指标

Version related

etcd_cluster_version

etcd_server_version (To replace Kubernetes etcd-version-monitor)

etcd_server_go_version

Snapshot metrics

To Monitor Snapshot Save Operations on local node

etcd_snap_db_fsync_duration_seconds_count

etcd_snap_db_save_total_duration_seconds_bucket

etcd_snap_fsync_duration_seconds

To Monitor Snapshot Operations between remote peers

etcd_network_snapshot_send_success

etcd_network_snapshot_send_failures

etcd_network_snapshot_send_total_duration_seconds

etcd_network_snapshot_receive_success

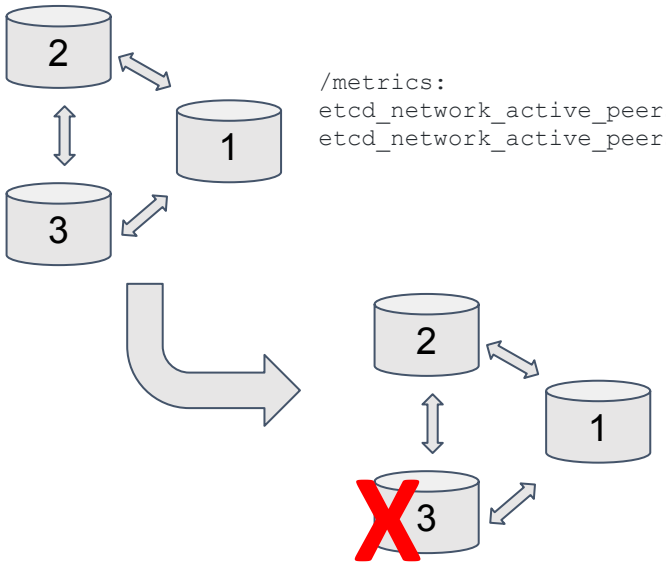
etcd_network_snapshot_receive_failures

etcd_network_snapshot_receive_total_duration_seconds

Peers healthiness

etcd_network_active_peers

etcd_network_disconnected_peers_total




```
/metrics:
etcd_network_active_peers{Local="1",Remote="2"} 1
etcd_network_active_peers{Local="1",Remote="3"} 1
```

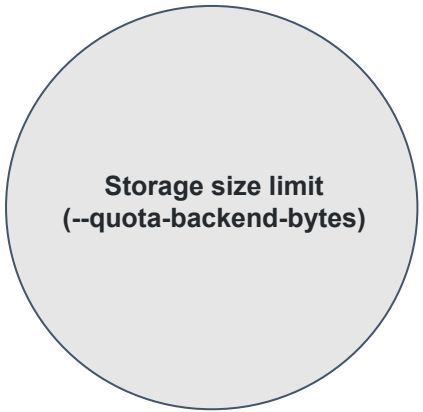
```
/metrics:
etcd_network_active_peers{Local="1",Remote="2"} 1
etcd_network_active_peers{Local="1",Remote="3"} 0
etcd_network_disconnected_peers_total{Local="1",Remote="3"} 1
```

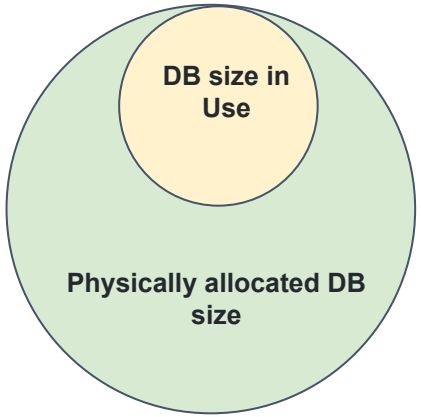
Database size metrics

etcd_server_quota_backend_bytes

etcd_mvcc_db_total_size_in_bytes

etcd_mvcc_db_total_size_in_use_in_bytes



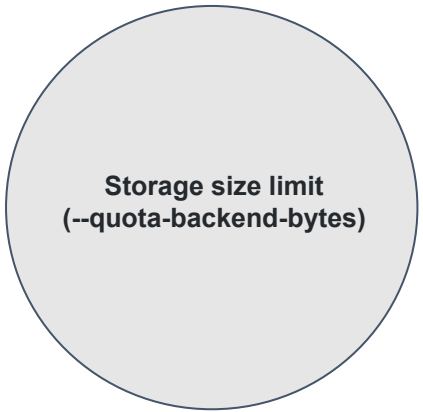


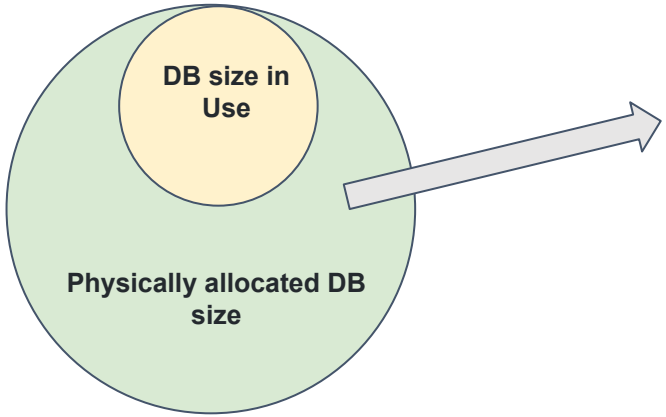
Database size metrics

etcd_server_quota_backend_bytes

etcd_mvcc_db_total_size_in_bytes

etcd_mvcc_db_total_size_in_use_in_bytes





Can be saved from defragmentation!

Storage layer metrics

etcd_server_heartbeat_send_failures_total

etcd_server_slow_apply_total

etcd_disk_backend_defrag_duration_seconds

etcd_mvcc_hash_duration_seconds

etcd_mvcc_hash_rev_duration_seconds

Indication of possible overloading of slow disk

Server side metrics

etcd_server_is_leader

etcd_server_id

etcd_server_health_success

etcd_server_health_failures

etcd_server_read_indexes_failed_total

etcd_server_slow_read_indexes_total

etcd learner metrics

etcd_server_is_learner

etcd_server_learner_promote_failures

etcd_server_learner_promote_successes

Ref:

etcd learner implementation: https://github.com/etcd-io/etcd/pull/10645

gRPC proxy expose endpoint metrics

Metrics and Health

The gRPC proxy exposes `/health` and Prometheus `/metrics` endpoints for the etcd members defined by `--endpoints`. An alternative define an additional URL that will respond to both the `/metrics` and `/health` endpoints with the `--metrics-addr` flag.

```
$ etcd grpc-proxy start \
  --endpoints https://localhost:2379 \
  --metrics-addr https://0.0.0.0:4443 \
  --listen-addr 127.0.0.1:23790 \
  --key client.key \
  --key-file proxy-server.key \
  --cert client.crt \
  --cert-file proxy-server.crt \
  --cacert ca.pem \
  --trusted-ca-file proxy-ca.pem
```

bboltdb transaction debugging

etcd_debugging_disk_backend_commit_rebalance_duration_seconds

etcd_debugging_disk_backend_commit_spill_duration_seconds

etcd_debugging_disk_backend_commit_write_duration_seconds

Note that any etcd_debugging_* metrics are experimental and subject to change.

etcd leases debugging

etcd_debugging_lease_granted_total

etcd_debugging_lease_revoked_total

etcd_debugging_lease_renewed_total

etcd_debugging_lease_ttl_total

Note that any etcd_debugging_* metrics are experimental and subject to change.

Example of how to use etcd metrics

如何分析etcd指标值

- Warning “Apply entry took too long”

  W | etcdserver: apply entries took too long [3.21342s for 1 entries]

  - Request too large
  - Slow disk: backend_commit_duration_seconds
  - CPU starvation, memory swapping

如何分析etcd指标值

- Client request timeout

  $ ETCDCTL_API=3 etcdctl put foo bar --endpoints “XXX”

  Error: context deadline exceeded

  - Can cluster make progress:

    etcd_server_has_leader, proposals_failed_total

  - Networking: peer_sent_failures_total, peer_round_trip_time_seconds
  - Slow apply: etcd_server_slow_apply_total

Thanks!

- Jingyi Hu 胡景懿 (Google)
  - github/jingyih, jingyih@google.com
- Wenjia Zhang 张文嘉 (Google)
  - github/wenjiaswe, wenjiazhang@google.com

Speakers

- Jingyi Hu 胡景懿 (Google)
  - **etcd maintainer,** kubernetes member
  - github/jingyiZh
  - jingyih@google.com
- Wenjia Zhang 张文嘉 (Google)
  - etcd contributor, kubernetes member
  - github/wenjiaswe
  - wenjiazhang@google.com

RAFT Consensus algorithm

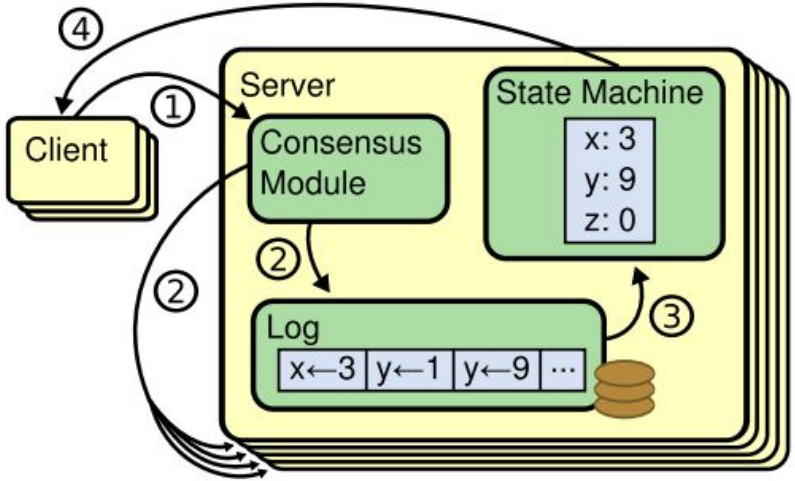


Figure 1: Replicated state machine architecture. The consensus algorithm manages a replicated log containing state machine commands from clients. The state machines process identical sequences of commands from the logs, so they produce the same outputs.

https://raft.github.io/raft.pdf

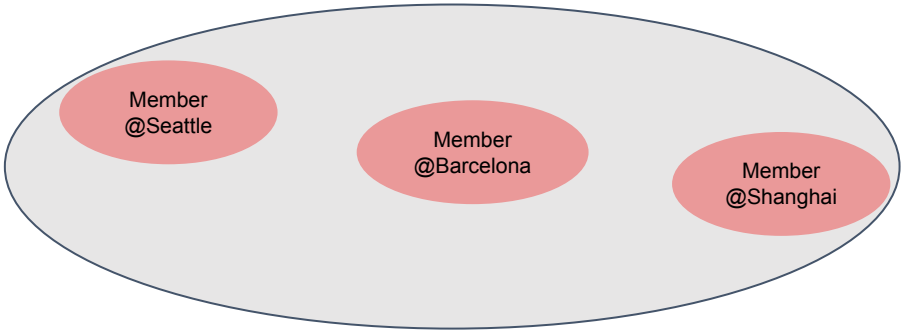
etcd_network_server_stream

etcd_network_server_stream_failures_total

The total number of stream failures from the local server.

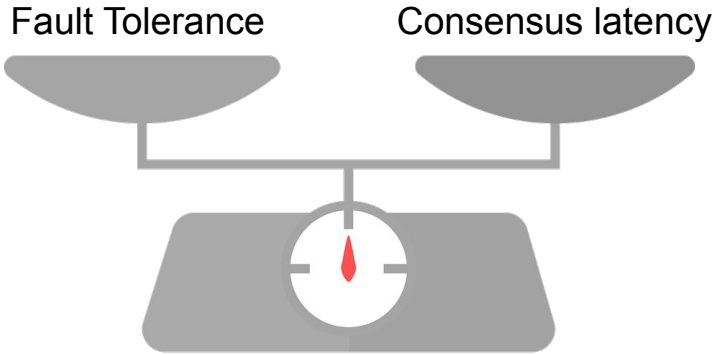
Example output:

```
etcd_network_server_stream_failures_total{API="lease-keepalive",Type="receive"} 1
etcd_network_server_stream_failures_total{API="watch",Type="receive"} 1
```



Tuning
heartbeat interval and election timeout setting

CPU | Disk | Networking




KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2019