JobManager High Availability

The problem: The JobManager (JM) is a single point of failure. When it crashes, TaskManagers (TM) fail all running jobs and try to reconnect to the same JM. A failed JM looses all state and can not resume the running jobs; even if it recovers and the TMs reconnect.

Solution: implement JM fault tolerance/high availability by having multiple JM instances running with one as leader and the other(s) in standby. The exact coordination and state update protocol between JM, TM, and clients is the topic of this document.

JIRA: FLINK-2287

Distributed Coordination with ZooKeeper (FLINK-2288)

Having standby JM instances requires distributed coordination between JM, TM, and clients. For this, we will use ZooKeeper (ZK).

Pros:

- Proven solution (other projects use it for this as well)
- Apache TLP with large community, docs, and library with required "recipies" like leader election (see below)

ZK dependency

- There is no separate ZK dependency for the client and the server
- Makes it easy to start ZK in "managed" mode (see below), because we will ship ZK with Flink

JM/TM/Client configuration

- Add HA mode for JMs (allowing non-HA and HA JM operation)
    - HA mode when a ZooKeeper quorum is configured
    - zookeeper.quorum: "host:port,[...,]host:port"
- TM, clients need to:
    - connect directly to configured JM (non-HA)
    - connect to JM leader; ZK coordinates the JM address (HA)
- Note: as of current stable ZK version (3.4.6) the ZK quorum needs to be statically configured (see ZK-107)

ZK client configuration in Flink

- ZK client configuration in Flink only requires the address of the ZK quorum to connect to
- Znode root and znode paths relative to this root, e.g. /flink (root), and /flink/leader

ZK server configuration

1. **ZK managed by Flink**: Flink provides script to start ZK servers
    a. Configuration via zoo.cfg
2. **Dedicated ZK cluster** (user sets up ZK)

Startup scripts

- Add an optional file `masters` to use for HA mode (contains all masters)
- No extra startup scripts, HA mode is activated if quorum is configured
    - scripts start JobManager on hosts configured in `masters`

Leader Election

- TaskManager and Client retrieve *leading* JobManager using ZooKeeper
- Every time they lose connection to the current JobManager, they try to reconnect to the current leader using ZooKeeper