分布式系统作业六

数据科学与计算机学院　17大数据与人工智能

17341015　陈鸿峥

问题 1. 对以下每个应用程序，你认为最多一次语义和最少一次语义哪个更好？

(a) 从文件服务器上读写文件；

(b) 编译一个程序；

(c) 远程银行

解答. (a) 最少一次(at least once)最好，因为可以重复多次尝试读写文件，失败后继续尝试

(b) 最少一次语义。同样可以重复多次尝试编译程序。

(c) 最多一次语义。为避免银行资金出现紊乱，服务器保证操作至多执行一次，当出现故障时最好进行人工干预（比如到银行柜台办理手续）。

问题 2. 简述*Flooding*、 *PAXOS*、 *RAFT*、 *PBFT*、 *PoW*的使用场景？

解答.

- Flooding：能够准确检测到失效，但是通讯量非常大，因为两两之间都要进行消息传递，泛洪共识在现实生活中已经用得很少了，更多是采用下面的共识协议 [1]。
- PAXOS：能够最终检测到失效，但是非常复杂，非常难实施；Google的NewSQL数据库Spanner [2]就是基于PAXOS搭的。
- RAFT：非拜占庭故障下达成共识的强一致性协议，也能最终检测到实现，但比PAXOS要好理解及好实现；在各种数据库及分布式系统中被广泛使用 [3]。
- PBFT：可以实现拜占庭容错，在区块链中会使用。其中的Leader选举是采用一种轮询的方式 [4]。
- PoW：现在的区块链用得最多，通过工作量/算力达成共识。PoW即确认工作端做过一定量工作的证明。比如在比特币系统 [5]中，大约每10分钟进行一轮算力比赛，获胜者将获得一次记账的权力，并向其他节点同步新增账本信息。

问题 3. 任意选择一种编程语言实现的*RAFT*程序，运行该程序并测试记录结果。

- *https://github.com/logcabin/logcabin*
- *https://github.com/streed/simpleRaft*

解答. 我使用了LogCabin进行测试。 LogCabin是一个基于raft的分布式存储系统，提供可靠、高度冗余、一致性的存储。

按照官方教程，需要先在Ubuntu系统安装scons、protobuf和crypto++，然后调用scons进行编译。

编译完成后，开启3个服务器端组成一个集群，服务器配置如下：

```
// File logcabin-1.conf
serverId = 1
listenAddresses = 127.0.0.1:5254

// File logcabin-2.conf
serverId = 2
listenAddresses = 127.0.0.1:5255

// File logcabin-3.conf
serverId = 3
listenAddresses = 127.0.0.1:5256
```

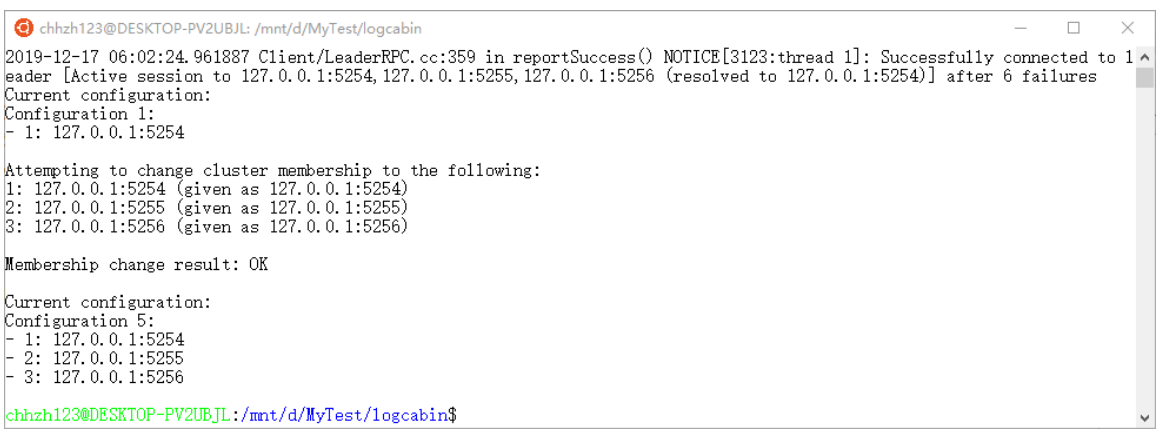
然后通过以下指令，开启服务器工作（注意要开启3个不同的命令行）

```
build/LogCabin --config logcabin-1.conf --bootstrap
build/LogCabin --config logcabin-1.conf # first terminal
build/LogCabin --config logcabin-2.conf # second terminal
build/LogCabin --config logcabin-3.conf # third terminal
```

然后开启第4个命令行，执行下列指令，将3个服务器捆绑为一个集群(cluster)。

```
ALLSERVERS=127.0.0.1:5254,127.0.0.1:5255,127.0.0.1:5256
build/Examples/Reconfigure --cluster=$ALLSERVERS set 127.0.0.1:5254 127.0.0.1:5255
    ↪  127.0.0.1:5256
```

会得到如下配置成功信息



同时在服务器端上也会显示新成员加入信息

```
chhzh123@DESKTOP-PV2UBJL: /mnt/d/MyTest/logcabin
  servers {
    server_id: 3
    addresses: "127.0.0.1:5256"
  }
}

2019-12-17 06:02:25.078198 Server/RaftConsensus.cc:318 in startThread() NOTICE[2:RaftService(12)]: Starting peer thread
for server 3
2019-12-17 06:02:25.079992 Server/RaftConsensus.cc:2074 in peerThreadMain() NOTICE[2:Peer(3)]: Peer thread for server 3
started
2019-12-17 06:02:25.100517 Server/RaftConsensus.cc:585 in setConfiguration() NOTICE[2:RaftService(12)]: Activating confi
guration 5:
prev_configuration {
  servers {
    server_id: 1
    addresses: "127.0.0.1:5254"
  }
  servers {
    server_id: 2
    addresses: "127.0.0.1:5255"
  }
  servers {
    server_id: 3
    addresses: "127.0.0.1:5256"
  }
}

2019-12-17 06:02:25.114322 Server/StateMachine.cc:608 in shouldTakeSnapshot() NOTICE[2:StateMachine]: Have applied 80% o
f the 5 total log entries
```

执行过程截图如下所示，有1个Leader，其余2个为Follower。

chhzh123@DESKTOP-PV2UBJL: /mnt/d/MyTest/logcabin

```
2019-12-17 06:05:18.969058 Server/ServerStats.cc:148 in dumpToDebugLog() NOTICE[1:StatsDumper]: ServerStats:
server_id: 1
addresses: "127.0.0.1:5254"
start_at: 1576562718968698400
end_at: 1576562718968801900
raft {
  current_term: 2
  state: LEADER
  commit_index: 10
  last_log_index: 10
  leader_id: 1
  voted_for: 1
  start_election_at: 9223372036854775807
  withhold_votes_until: 9223372036854775807
  cluster_time: 239645393600
  cluster_time_epoch: 191465517000
  last_snapshot_index: 0
  last_snapshot_bytes: 0
  log_start_index: 1
  log_bytes: 613
  last_snapshot_term: 0
  last_snapshot_cluster_time: 0
  num_entries_truncated: 0
  peer {
    server_id: 3
    addresses: "127.0.0.1:5256"
    old_member: true
    new_member: false
    staging_member: false
    request_vote_done: false
    have_vote: false
    suppress_bulk_data: false
    next_index: 11
    last_agree_index: 10
    is_caught_up: true
    next_heartbeat_at: 1576562719068954900
    backoff_until: -7646816969873255808
  }
  peer {
    server_id: 1
    addresses: "127.0.0.1:5254"
    old_member: true
    new_member: false
    staging_member: false
    last_synced_index: 10
  }
  peer {
    server_id: 2
    addresses: "127.0.0.1:5255"
    old_member: true
    new_member: false
    staging_member: false
    request_vote_done: false
    have_vote: false
    suppress_bulk_data: false
    next_index: 11
    last_agree_index: 10
    is_caught_up: true
    next_heartbeat_at: 1576562719068775800
    backoff_until: -7646816969873255808
  }
}
storage {
  num_segments: 2
  open_segment_bytes: 557
  metadata_version: 8
  metadata_write_nanos {
    average: 6070450
    count: 4
    ewma2: 6068387.5
    ewma4: 5989570.3125
    exceptional_count: 0
    last: 7022400
    min: 3198100
    max: 8169900
    sum: 24281800
    stddev: 1843661.6100846706
  }
  filesystem_ops_nanos {
    average: 454450
    count: 8
    ewma2: 231677.34375
    ewma4: 367363.76342773438
    exceptional_count: 0
```

```
chhzh123@DESKTOP-PV2UBJL: /mnt/d/MyTest/logcabin
2019-12-17 06:06:02.187851 Server/ServerStats.cc:148 in dumpToDebugLog() NOTICE[3:StatsDumper]: ServerStats:
server_id: 3
addresses: "127.0.0.1:5256"
start_at: 1576562762187657700
end_at: 1576562762187717100
raft {
  current_term: 2
  state: FOLLOWER
  commit_index: 10
  last_log_index: 10
  leader_id: 1
  voted_for: 0
  start_election_at: 1576562762829770460
  withhold_votes_until: 1576562762590390100
  cluster_time: 282863787400
  cluster_time_epoch: 191465517000
  last_snapshot_index: 0
  last_snapshot_bytes: 0
  log_start_index: 1
  log_bytes: 612
  last_snapshot_term: 0
  last_snapshot_cluster_time: 0
  num_entries_truncated: 0
  peer {
    server_id: 2
    addresses: "127.0.0.1:5255"
    old_member: true
    new_member: false
    staging_member: false
  }
  peer {
    server_id: 3
    addresses: "127.0.0.1:5256"
    old_member: true
    new_member: false
    staging_member: false
  }
  peer {
    server_id: 1
    addresses: "127.0.0.1:5254"
    old_member: true
    new_member: false
    staging_member: false
  }
}
storage {
  num_segments: 1
  open_segment_bytes: 612
  metadata_version: 4
  metadata_write_nanos {
    average: 3530875
    count: 4
    ewma2: 4337900
    ewma4: 3590190.625
    exceptional_count: 0
    last: 5964100
    min: 2535900
    max: 5964100
    sum: 14123500
    stddev: 1412069.9812243725
  }
  filesystem_ops_nanos {
    average: 368600
    count: 8
    ewma2: 178987.5
    ewma4: 290173.71826171875
    exceptional_count: 0
    last: 113100
    min: 113100
    max: 1641500
    sum: 2948800
    stddev: 485863.52250400523
  }
}
state_machine {
  snapshotting: false
  last_applied: 10
  num_sessions: 0
  num_snapshots_attempted: 0
  num_snapshots_failed: 0
  num_redundant_advance_version_entries: 0
  num_rejected_advance_version_entries: 0
  num_successful_advance_version_entries: 1
  num_total_advance_version_entries: 1
  min_supported_version: 1
```

接下来测试其文件功能，采用下述执行进行简单的文件操作。

```
build/Examples/TreeOps --cluster=$ALLSERVERS mkdir /test
echo -n hello | build/Examples/TreeOps --cluster=$ALLSERVERS write /world
build/Examples/TreeOps --cluster=$ALLSERVERS dump
```

最终可以看到test目录和world文件都成功生成。

```
chhzh123@DESKTOP-PV2UBJL: /mnt/d/MyTest/logcabin
chhzh123@DESKTOP-PV2UBJL:/mnt/d/MyTest/logcabin$ echo -n hello | build/Examples/TreeOps --cluster=$ALLSERVERS write /wor
ld
2019-12-17 06:36:55.131342 Client/ClientImpl.cc:455 in ClientImpl() NOTICE[3218:thread 1]: Configuration settings:
# begin config
# end config
2019-12-17 06:36:55.133437 Client/ClientImpl.cc:481 in initDerived() NOTICE[3218:thread 1]: Using server list: 127.0.0.1
:5254,127.0.0.1:5255,127.0.0.1:5256
2019-12-17 06:36:55.137154 Client/LeaderRPC.cc:338 in reportRedirect() NOTICE[3218:thread 1]: Server [Active session to
127.0.0.1:5254,127.0.0.1:5255,127.0.0.1:5256 (resolved to 127.0.0.1:5256)] is not leader, will try suggested 127.0.0.1:5
254 next (there have been 1 failed attempts during this outage)
2019-12-17 06:36:55.141903 Client/LeaderRPC.cc:359 in reportSuccess() NOTICE[3218:thread 1]: Successfully connected to l
eader [Active session to 127.0.0.1:5254 (resolved to 127.0.0.1:5254)] after 1 failures
chhzh123@DESKTOP-PV2UBJL:/mnt/d/MyTest/logcabin$ build/Examples/TreeOps --cluster=$ALLSERVERS dump
2019-12-17 06:37:03.434059 Client/ClientImpl.cc:455 in ClientImpl() NOTICE[3221:thread 1]: Configuration settings:
# begin config
# end config
2019-12-17 06:37:03.436196 Client/ClientImpl.cc:481 in initDerived() NOTICE[3221:thread 1]: Using server list: 127.0.0.1
:5254,127.0.0.1:5255,127.0.0.1:5256
/
/test/
/world:
    hello
chhzh123@DESKTOP-PV2UBJL:/mnt/d/MyTest/logcabin$
```

最后关闭其中一个服务器端，可以从下图看到其重新选举Leader的过程。

```
2019-12-17 07:04:17.404136 Server/RaftConsensus.cc:1555 in handleRequestVote() NOTICE[3:RaftService(12)]: Received RequestVote request from server 2 in term 3 (this server's term was 2)
2019-12-17 07:04:17.409996 Storage/SegmentedLog.cc:778 in updateMetadata() NOTICE[3:RaftService(12)]: Writing new storage metadata (version 5) to metadata1
2019-12-17 07:04:17.428433 Server/RaftConsensus.cc:2854 in printElectionState() NOTICE[3:RaftService(12)]: server=3, term=3, state=FOLLOWER,  leader=0, vote=0
2019-12-17 07:04:17.431166 Server/RaftConsensus.cc:1567 in handleRequestVote() NOTICE[3:RaftService(12)]: Voting for 2 in term 3
2019-12-17 07:04:17.433062 Storage/SegmentedLog.cc:778 in updateMetadata() NOTICE[3:RaftService(12)]: Writing new storage metadata (version 6) to metadata2
2019-12-17 07:04:17.440070 Server/RaftConsensus.cc:2854 in printElectionState() NOTICE[3:RaftService(12)]: server=3, term=3, state=FOLLOWER,  leader=0, vote=2
2019-12-17 07:04:17.457467 Server/RaftConsensus.cc:1313 in handleAppendEntries() NOTICE[3:RaftService(12)]: All hail leader 2 for term 3
2019-12-17 07:04:17.478314 Server/RaftConsensus.cc:2854 in printElectionState() NOTICE[3:RaftService(12)]: server=3, term=3, state=FOLLOWER,  leader=2, vote=2
2019-12-17 07:04:20.586784 Server/ServerStats.cc:148 in dumpToDebugLog() NOTICE[3:StatsDumper]: ServerStats:
```

从而通过此实验验证了raft共识协议的有效性，并且清晰地观察到其工作原理。

参考文献

[1] Flooding Consensus, http://fileadmin.cs.lth.se/cs/Personal/Amr_Ergawy/dist-algos-slides/eigth-presentation.pdf

[2] Google NewSQL Spanner, https://en.wikipedia.org/wiki/Spanner_(database)

[3] RAFT Implementations, https://raft.github.io/#implementations

[4] PBFT, https://blockonomi.com/practical-byzantine-fault-tolerance/

[5] Bitcoin, https://blockonomi.com/bitcoin-mining-software/