

odi-mdsal-clustering

Moiz Raja

OpenDaylight Silicon Valley Meetup

04/28/2015

Topics

- Components
- Requirements
- Design
- Testing
- Monitoring
- Challenges
- Insights
- What's missing

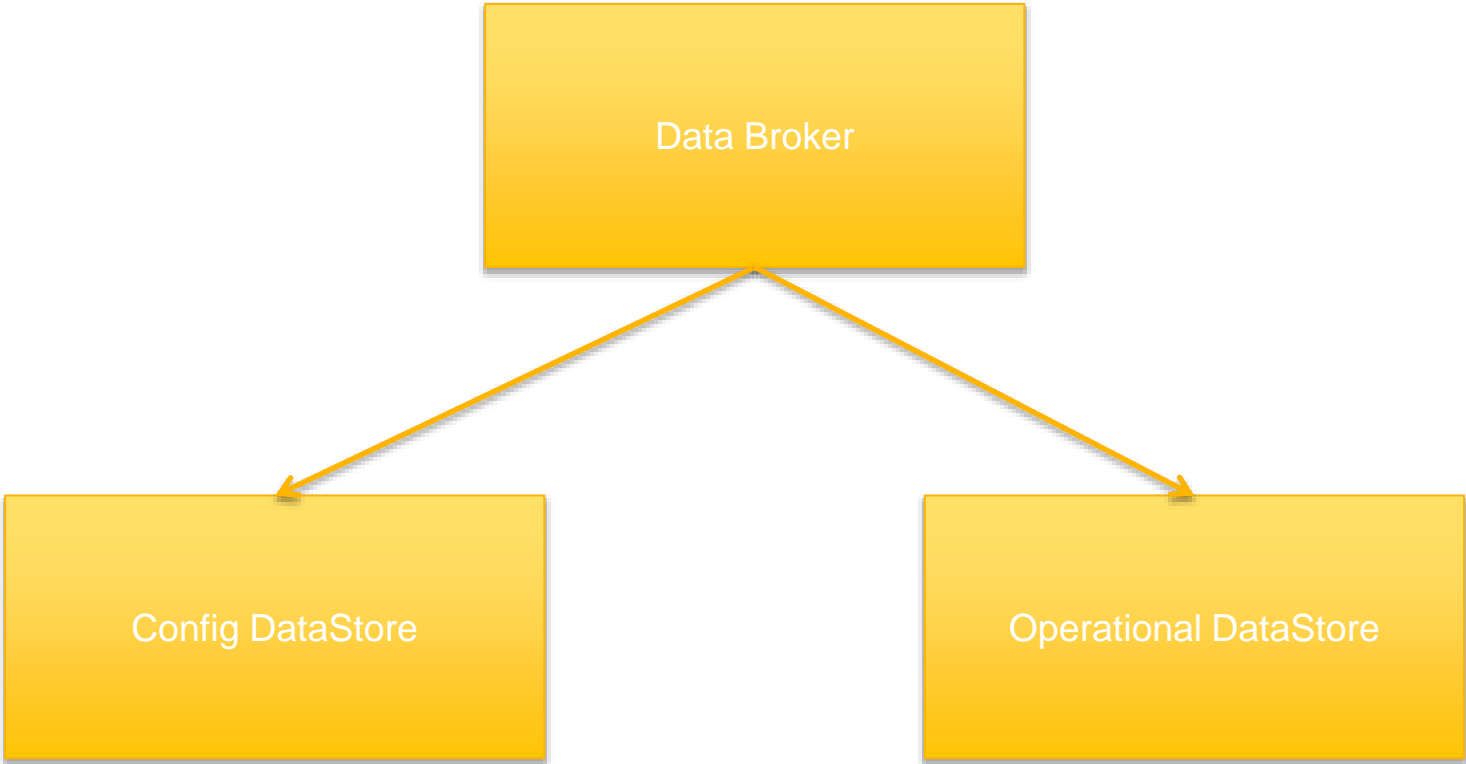
Components

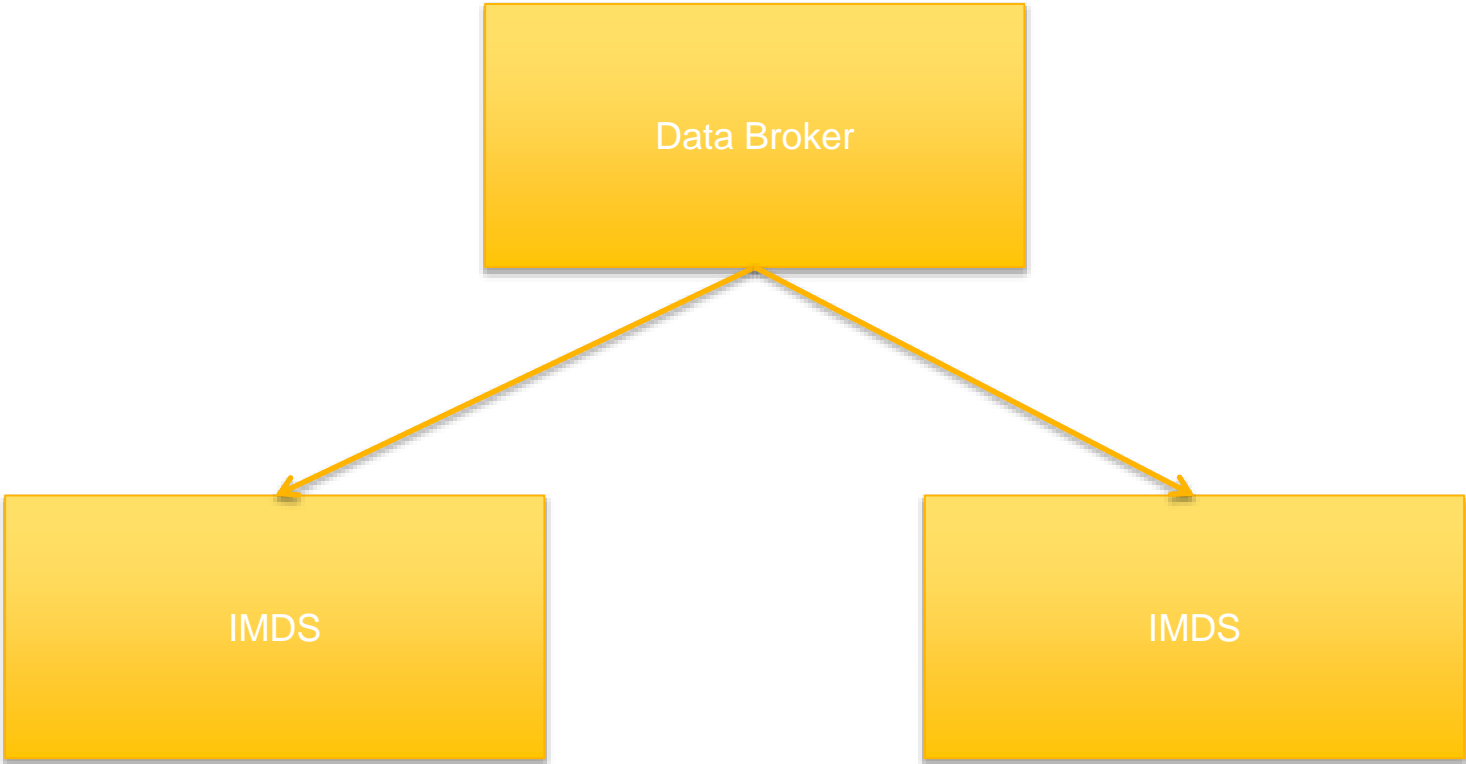
- Remote RPC
- **Distributed Data Store**
- Remote Notifications ???

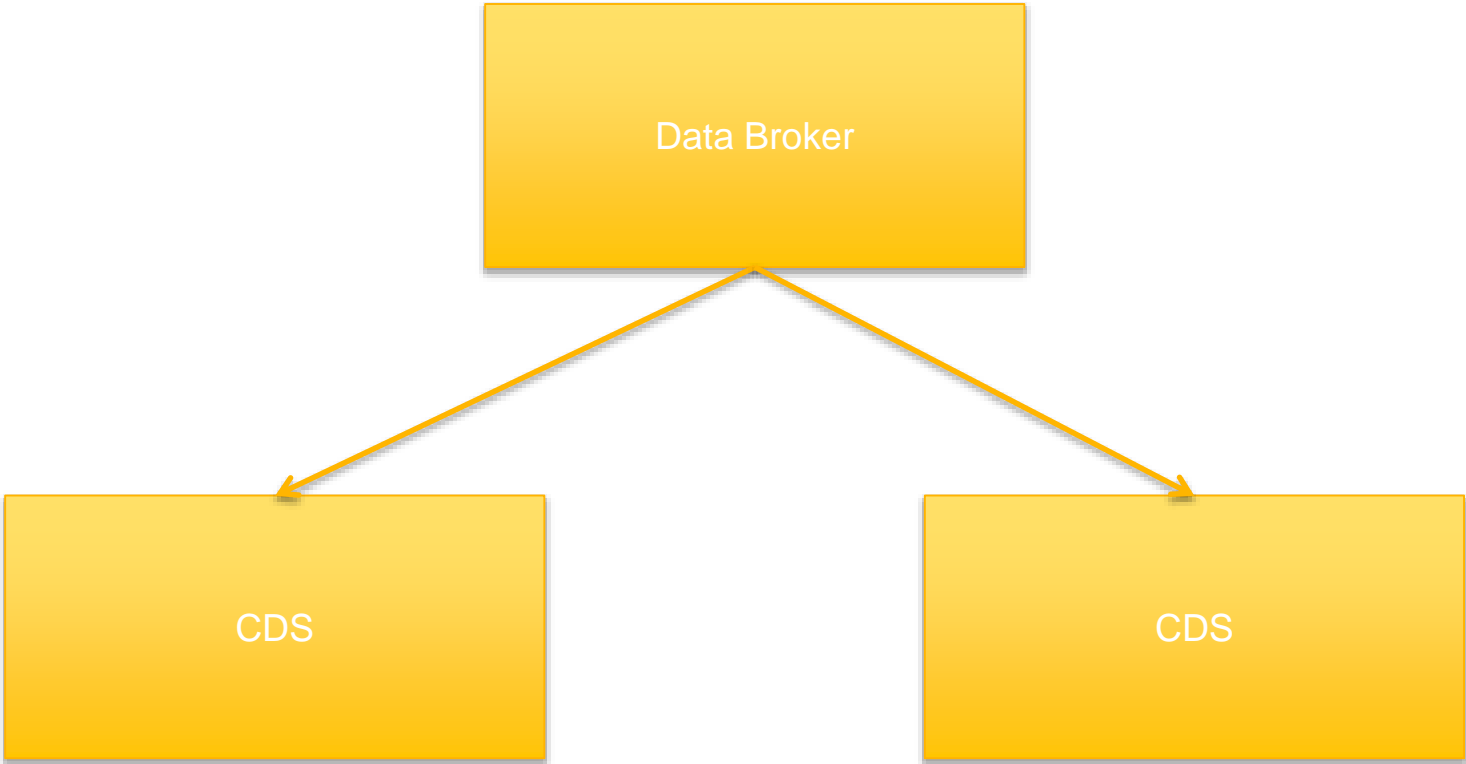
Requirements

- Location Transparency
- Drop in replacement for IMDS
- Persistence
- Strongly Consistency
- Data Change Notifications
- Static configuration
- Sharding capabilities

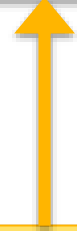
Design



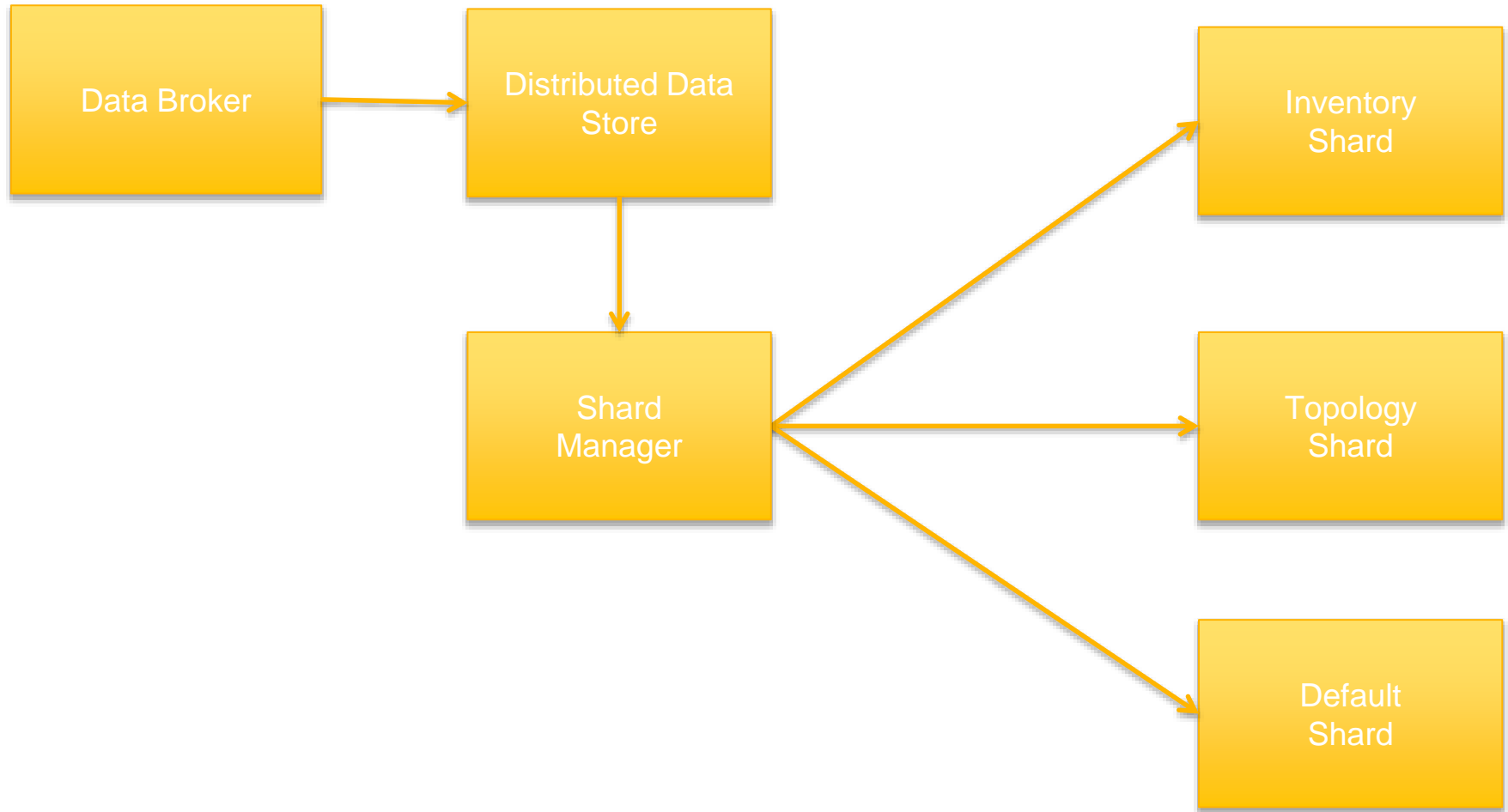




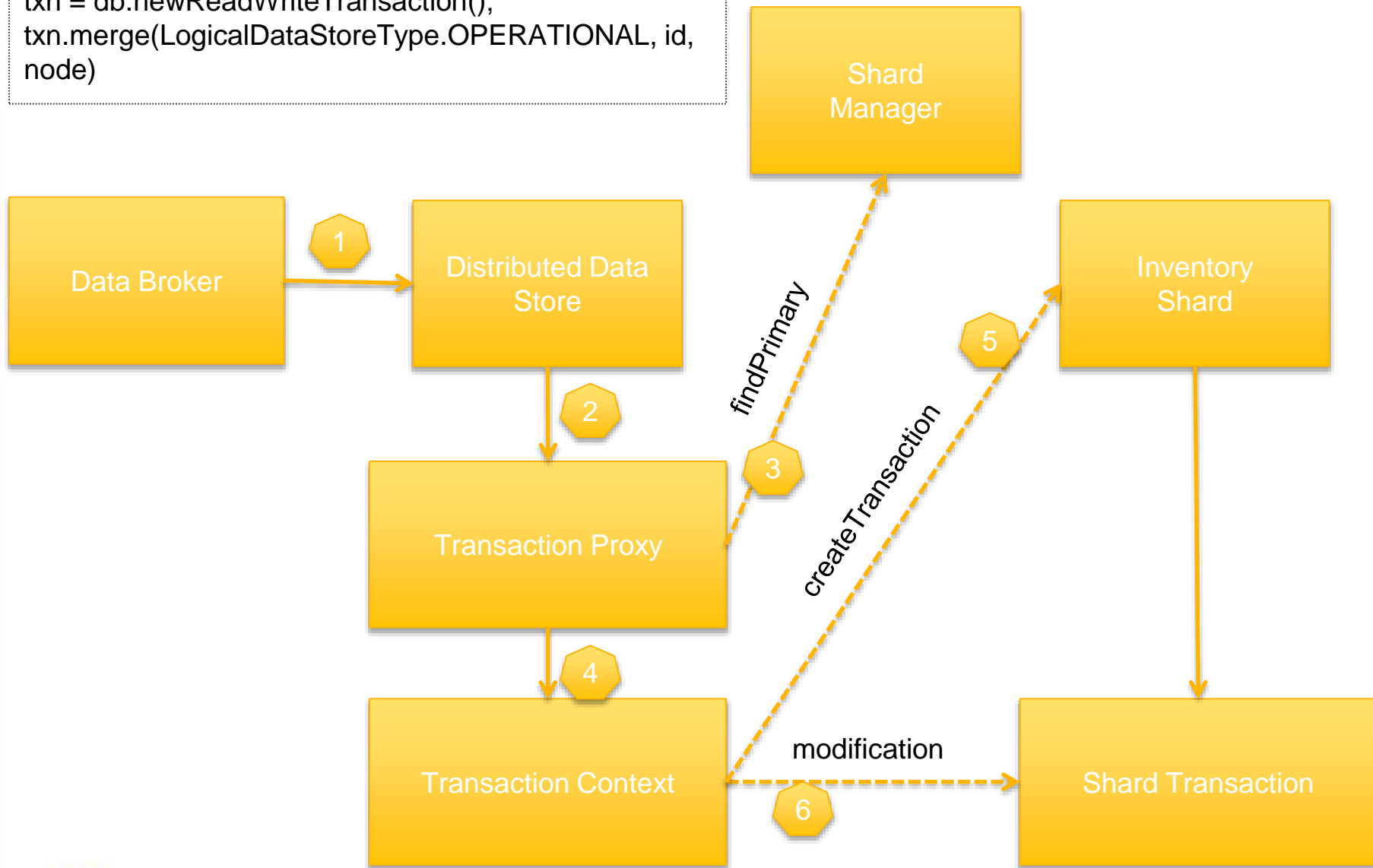
org.opendaylight.controller.sal.core.spi.data.DOMStore

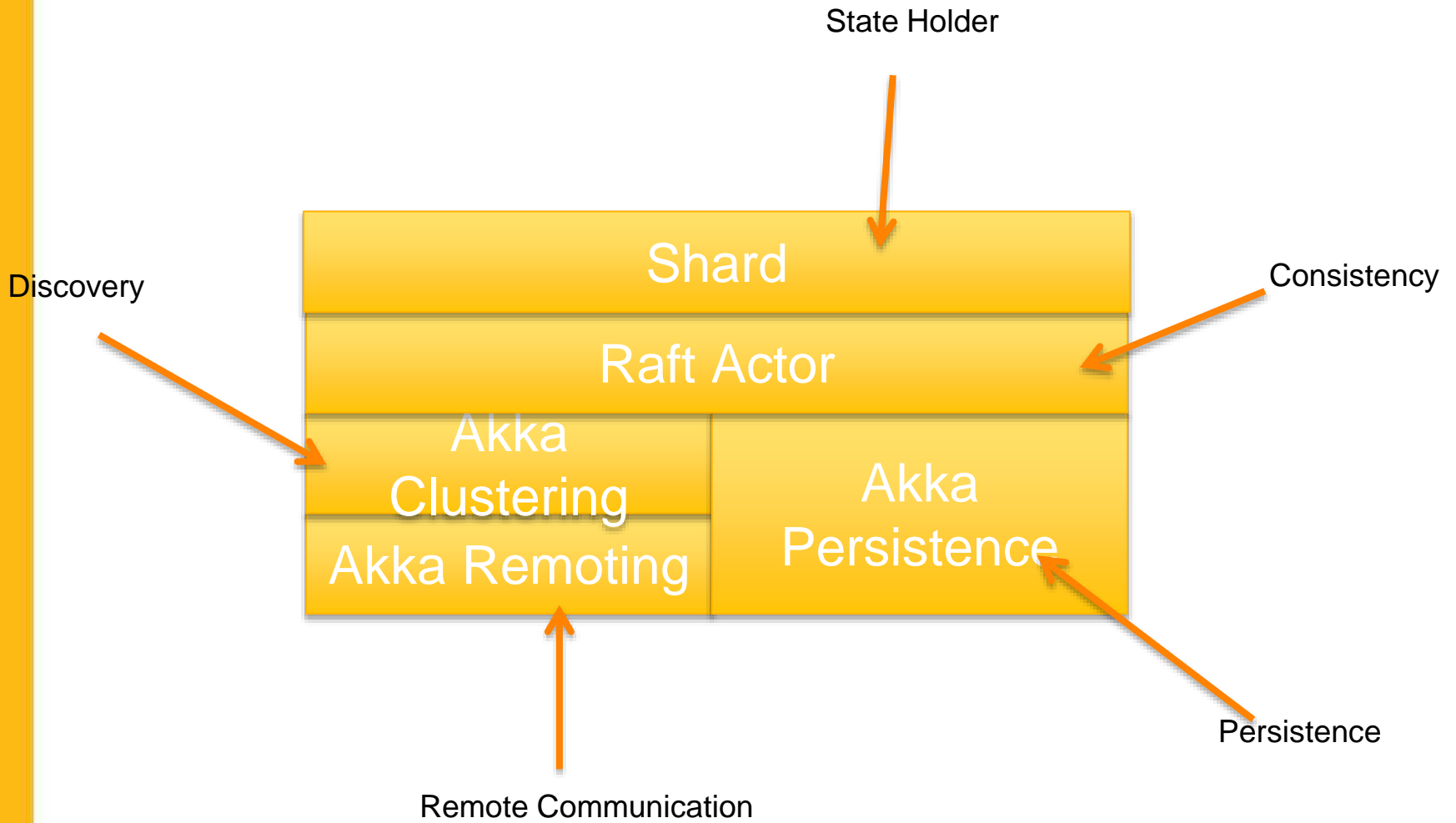


DistributedDataStore



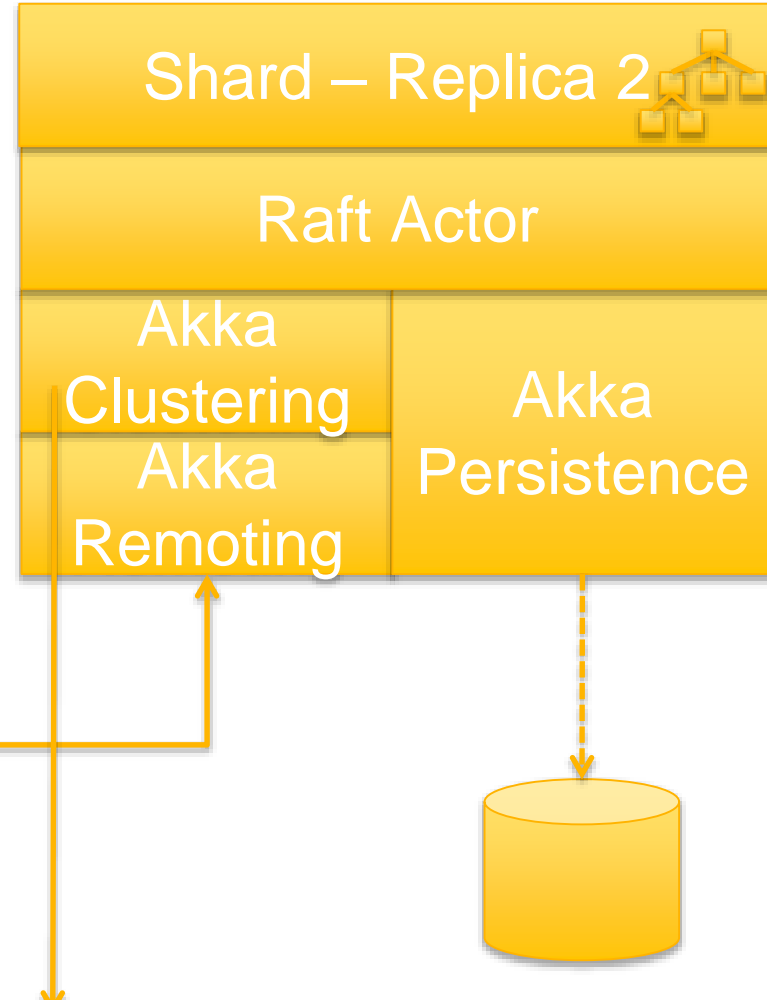
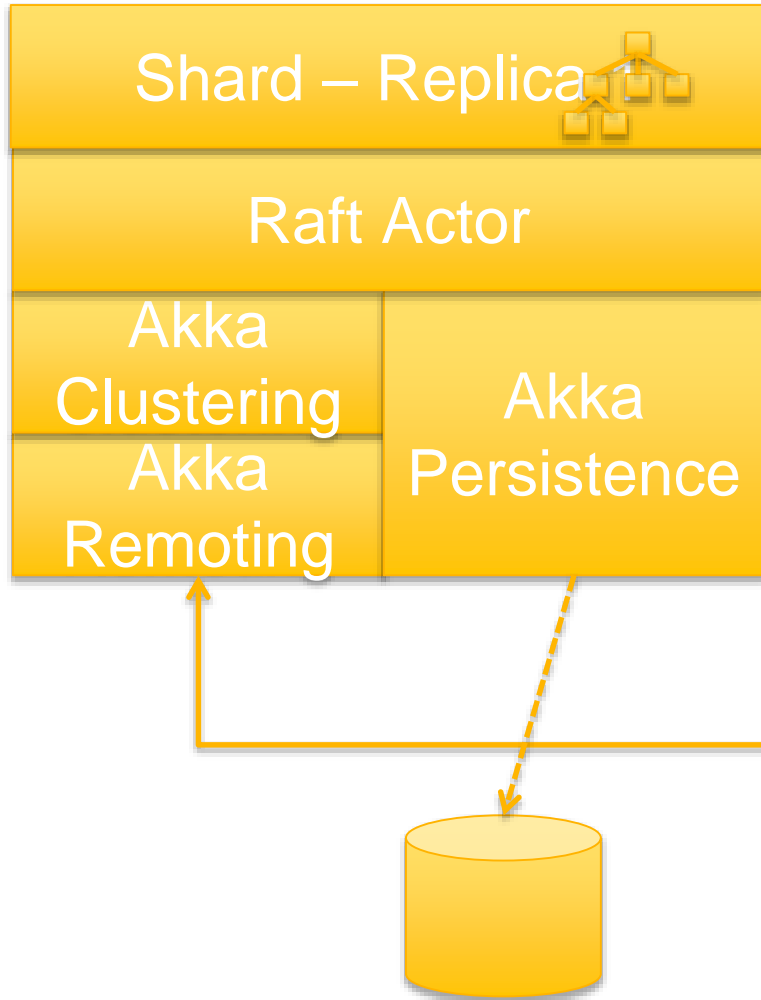
```
txn = db.newReadWriteTransaction();  
txn.merge(LogicalDataStoreType.OPERATIONAL, id,  
node)
```





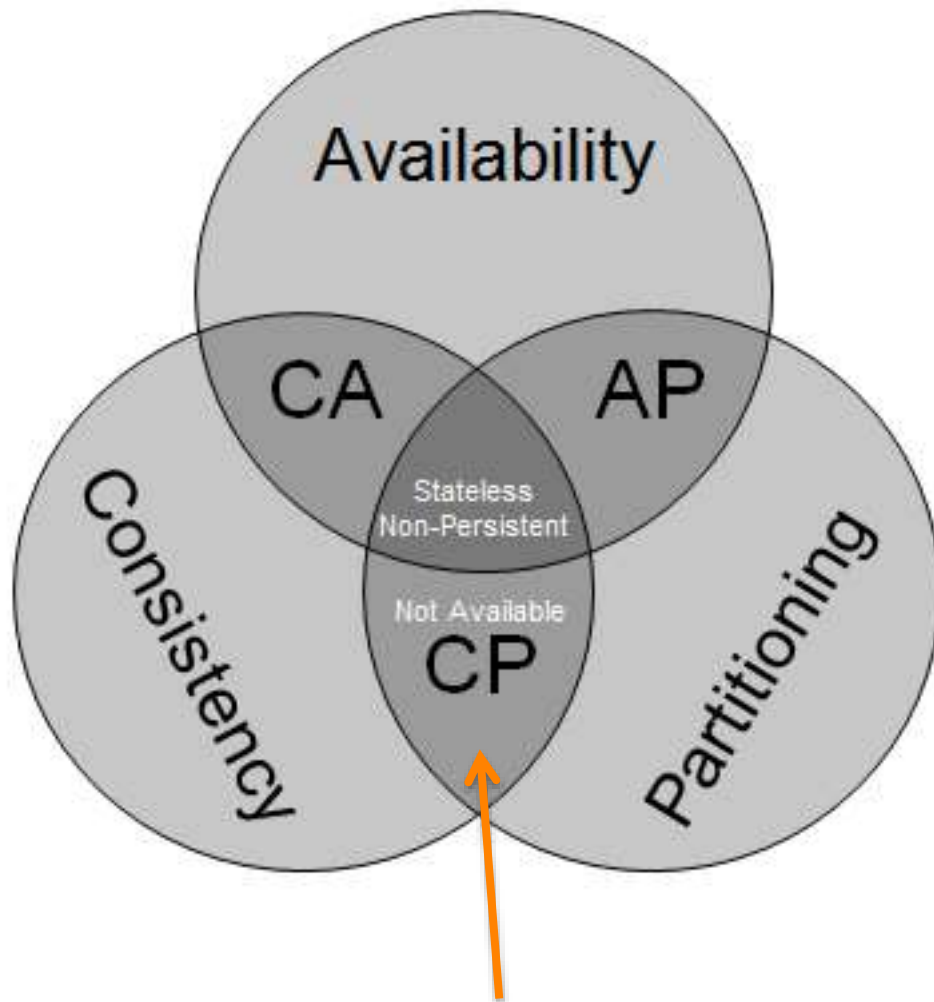
member-1 (10.194.126.242)

member-2 (10.194.126.243)



member-2 is on 10.194.126.243

member-1 is on 10.194.126.242



What about Availability?

- We are available as long as a majority of the replicas are connected
 - $(N / 2) + 1$ out of an N node cluster
 - 2 out of 3 nodes in a 3 node cluster
 - 3 out of 4 nodes in a 4 node cluster
 - 3 out of 5 nodes in a 5 node cluster
 - and so on

Configuration/RAFT determine availability

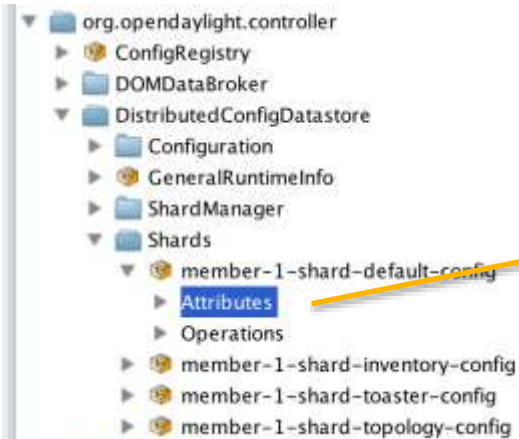
- If inventory is configured to be on,
 - member-1
 - member-2
 - member-3
 - then atleast 2 of those members need to be running for the inventory data cluster to be available
- If inventory in only configured to be on,
 - member-1
 - then just member-1 needs to be running for the inventory data cluster to be available

Testing

- Unit tests (> 80% code coverage)
- Integration test (aka car-people test) for testing HA/Failover in a real cluster
- dsBenchmark for performance testing
- Dummy Datastore for testing replication overhead
- Raft test driver for testing the Raft implementation on a single box
- Other Performance/Scale tests
 - BGP using exabgp and some other test scripts
 - PCEP using pcc-mock
 - Netconf using the netconf simulator
 - Cbench for openflow

Monitoring

- Mbeans
 - org.opendaylight.controller
 - DistributedConfigDataStore
 - DistributedOperationalDataStore
 - org.opendaylight.controller.actor.metric
 - org.opendaylight.controller.cluster.datastore



Attribute values

Name	Value
AbortTransactionsCount	0
CommitIndex	-1
CommittedTransactionsCount	0
CurrentNotificationMgrListener...	javax.management.openmbean.C...
CurrentTerm	14
DataStoreExecutorStats	
FailedReadTransactionsCount	0
FailedTransactionsCount	0
FollowerInfo	javax.management.openmbean.C...
FollowerInitialSyncStatus	true
InMemoryJournalDataSize	0
InMemoryJournalLogSize	0
LastApplied	-1
LastCommittedTransactionTime	1970-01-01 00:00:00.000
LastIndex	-1
LastLogIndex	-1
LastLogTerm	-1
LastTerm	-1
Leader	member-2-shard-default-config
MaxNotificationMgrListenerQue...	1000
NotificationMgrExecutorStats	javax.management.openmbean.C...
PeerAddresses	member-3-shard-default-config; ...
RaftState	Follower
ReadOnlyTransactionCount	0
ReadWriteTransactionCount	0
ReplicatedToAllIndex	-1
ShardName	member-1-shard-default-config
SnapshotCaptureInitiated	false
SnapshotIndex	-1
SnapshotTerm	-1
StatRetrievalError	
StatRetrievalTime	5.820 ms
VotedFor	



- ▶ /user/shardmanager-config.msg-rate.ActorInitialized
- ▶ /user/shardmanager-config.msg-rate.FindLocalShard
- ▶ /user/shardmanager-config.msg-rate.FindPrimary
- ▶ /user/shardmanager-config.msg-rate.FollowerInitialSyncUpStatus
- ▶ /user/shardmanager-config.msg-rate.LeaderStateChanged
- ▶ /user/shardmanager-config.msg-rate.MemberRemoved
- ▶ /user/shardmanager-config.msg-rate.MemberUp
- ▶ /user/shardmanager-config.msg-rate.RegisterRoleChangeListenerReply
- ▶ /user/shardmanager-config.msg-rate.RoleChangeNotification
- ▶ /user/shardmanager-config.msg-rate.UnreachableMember
- ▶ /user/shardmanager-config.msg-rate.UpdateSchemaContext
- ▶ /user/shardmanager-config.q-size
- ▶ /user/shardmanager-config/member-1-shard-default-config.msg-rate
- ▼ /user/shardmanager-config/member-1-shard-default-config.msg-rate.AppendEntries
 - ▶ **Attributes**
 - ▶ Operations
- ▶ /user/shardmanager-config/member-1-shard-default-config.msg-rate.AppendEntriesR
- ▶ /user/shardmanager-config/member-1-shard-default-config.msg-rate.ElectionTimeout
- ▶ /user/shardmanager-config/member-1-shard-default-config.msg-rate.FollowerInitialSyn
- ▶ /user/shardmanager-config/member-1-shard-default-config.msg-rate.GetOnDemandR

Attribute values

Name	Value
50thPercentile	0.0406925
75thPercentile	0.04588375
95thPercentile	0.06580455
98thPercentile	0.07828779999999992
999thPercentile	0.5315992490000001
99thPercentile	0.10381126000000003
Count	289345
DurationUnit	milliseconds
FifteenMinuteRate	1.579230387187583
FiveMinuteRate	1.6923527044010036
Max	0.5334209999999999
Mean	0.04150901264591439
MeanRate	1.8142818441782067
Min	0.0051909999999999994
OneMinuteRate	1.9184454614089654
RateUnit	events/second
StdDev	0.030344577956276358

- ▼ org.opendaylight.controller
 - ▶ ConfigRegistry
 - ▶ DOMDataBroker
 - ▼ DistributedConfigDatastore
 - ▶ Configuration
 - ▼ GeneralRuntimeInfo
 - ▶ **Attributes**
 - ▶ ShardManager
 - ▼ Shards
 - ▼ member-1-shard-default-config
 - ▶ Attributes
 - ▶ Operations
 - ▶ member-1-shard-inventory-config
 - ▶ member-1-shard-toaster-config
 - ▶ member-1-shard-topology-config

Attribute values	
Name	Value
TransactionCreationRateLimit	21.518857581727545

- ▼  org.opendaylight.controller.cluster.datastore
 - ▼  distributed-data-store.config.commit.rate
 - ▼ **Attributes**
 - Mean
 - StdDev
 - 50thPercentile
 - 75thPercentile

Attribute values	
Name	Value
50thPercentile	43.223379
75thPercentile	62.336458
95thPercentile	79.92874599999999
98thPercentile	79.92874599999999
999thPercentile	79.92874599999999
99thPercentile	79.92874599999999
Count	25
DurationUnit	milliseconds
FifteenMinuteRate	1.9136495194400216E-23
FiveMinuteRate	4.735725544968984E-64
Max	79.92874599999999
Mean	48.66985
MeanRate	1.565441541560738E-4
Min	31.579124
OneMinuteRate	7.6985803000629E-310
RateUnit	events/second
StdDev	17.34457046253307

Challenges

- API
- Serialization
- Memory
- Messaging and Context Switching
- Back Pressure
 - Operations
 - Transactions
- Remoting Latencies
- Persistence Latencies

Insights

- Bulk transactions work best
- Avoid multiple writers
- Try to write to only one shard in a transaction

What's missing?

- Remote Notifications
- Dynamic addition/removal of servers
- Fine grained sharding

The End!