APACHE SLING & FRIENDS TECH MEETUP

BERLIN, 26-28 SEPTEMBER 2012

CQ5 Cluster Deep Dive

Marcel Reutegger

Agenda

- Cluster basics
- Backup and restore
- Out-of-sync
- Application considerations
- Next release

Cluster basics

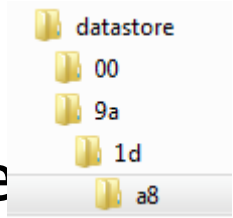
adaptTo()

Cluster basics

- CRX based on Apache Jackrabbit
  - CRX 2.2 -> Jackrabbit 2.2 / CRX 2.3 -> Jackrabbit 2.4
- Tar file based persistence manager and journal
- Shared and shared-nothing cluster deployment
- One master and *N* slaves
  - Slaves are eventual consistent
- Offers high availability & read scaling

Cluster basics – Data Store

- Contains binary properties >= 4k bytes
- Content addresseable storage
- File location calculated from SHA-1 hash
- Fixed height of directory tree: 3
  - 9a1da82c7cfde3294813803de0a15bd9e7 goes here:



- Adding items is atomic (tmp file & move)
  - File system must support move from datastore to destination directory

adaptTo()

Cluster basics – Data Store

- Default is non-shared
  - GC must be run on all cluster nodes
- Shared by cluster nodes
  - GC can be performed by any cluster node
- Shared by independent repositories
  - Mark only phase on all repositories
  - Manual delete
  - Described in knowledge base article: http://bit.ly/IzeJYY

Cluster basics – Data Store

- New in CRX 2.3
  - Improved reliability
  - Data Store items are broadcasted
  - Guaranteed consistency – Data Store updated first

Cluster basics – Journal

- Contains events, node type & namespace modifications, etc.
- Does not contain modified items
- Record has a monotonically increasing revision number
- maxAge defaults to 30 days
- TarJournal is not optmized even if logs may say so!

Cluster basics – Journal & Search Index

- Search index always non-shared
- Search index registered as synchronous EventListener
- Journal is transport channel for events from other cluster nodes
- Current (local) position in Journal identified by revision.log file
- Replaying events from Journal is independent from TarPM synchronization
  - Search index may see events for items that do not exist anymore
- Beware: cluster join without revision.log

Backup and restore

Backup and restore

- Per cluster node backup
  - Requires more space but simplifies restore
- Cluster node specific data
  - cluster_node.id, cluster.properties, clustered.txt
- Spare cluster node for backup
  - Temporarily stop cluster node and use any file system backup mechanism

Backup and restore

- Special handling of cluster node specific data may be needed
  - cluster_node.id, cluster.properties
- Backup does not retain executable flag
  - start/stop/serverctl scripts
- Automate and test!
  - Example: restore cluster node from 24h old backup
  - Measure re-sync time
  - With adobe.com usage pattern ~14 minutes (http://bit.ly/Pd24y5)

Backup and restore – Disaster recovery

- Which node to start first?
  - Delete clustered.txt on designated master
- Check size of data*.tar files
  - crx.default workspace
  - version store
  - tarJournal

Out-of-sync

Out-of-sync

- Slave nodes are eventually consistent
  - syncDelay: 5s
- Out-of-sync may happen when master dies suddenly
- Many clustering improvements after CQ 5.4 GA
  - Install recent CRX 2.2.0.x Hotfix

Out-of-sync

- Detection and log message
  - java.io.IOException: This cluster node and the master are out of sync. Operation stopped. Please ensure the repository is configured correctly.
    To continue anyway, please delete the index and data tar files on this cluster node and restart. Please note the Lucene index may still be out of sync unless it is also deleted.
- Reduce risk of out-of-sync
  - Stop slave nodes first
  - Write to slave node

Out-of-sync

- Recovery from out-of-sync
  - Restore from existing or new backup
  - Create cluster node from scratch (clone from master)
- Documentation on docs.day.com
  - http://bit.ly/RNcTf8
- New in CRX 2.3
  - Rollback to transaction: http://bit.ly/PzeZva

Out-of-sync

- Repository precaution
  - Remember slave role (clustered.txt file)
  - ‘Last call’ when master shuts down

Application considerations

adaptTo()

Application considerations

- Some things must only happen once in cluster
- Workflow and replication run on master only
- Dynamic repository descriptor ‘crx.cluster.master’
- com.day.cq.jcrclustersupport.ClusterAware
  - bindRepository(String repositoryId, String clusterId, boolean isMaster)
  - unbindRepository()
- InvalidItemStateException: ‘modified externally’
  - concurrent & conflicting writes

Next Release

- Performance improvements
- Enhanced JMX monitoring
- Auto-recovery from out-of-sync

CQ5 Cluster Deep Dive

Thank you