Infrastructure - Story #8162

Replication tasks can contain stale potential target nodes

2017-08-23 18:25 - Rob Nahf

Status:	New	**Start date:**	2017-08-23
Priority:	Normal	**Due date:**	
Assignee:	Rob Nahf	**% Done:**	0%
Category:	d1_replication	**Estimated time:**	0.00 hour
Target version:	CCI-2.3.10		
Story Points:			

Description

When a replication task is created, a list of potential replication targets is added. Some time in the future, the task is executed (mn.replicate() requests sent).

If a MemberNode turns off replication in the time between task creation and task execution, the mn.replicate will still be sent. Even though the MN should protect itself from unwanted requests, it seems bad form for the CN to send the request after the MN told the CN not to.

ReplicationTaskQueue seems to be the class calling replicate. It should be easy enough to copy the original logic in ReplicationManager that checked the suitability of the potential member node so that the check can be repeated right before the call is made.

We should probably also recheck the systemMetadata, in case that's changed too.

Before doing anything, confirm that the tasks are potentially long-lived.

(Also, the NodeList is cached, so we already allow for 3 minutes of being out of date)

Related issues:

Related to Infrastructure - Story #8639: Replication performance is too slow ...	**New**	**2018-07-04**

History

#1 - 2017-08-23 18:42 - Rob Nahf

- Description updated

#2 - 2017-08-25 16:12 - Rob Nahf

IARC also wants to turn off replication.

#3 - 2017-08-25 17:41 - Rob Nahf

The check for potential target nodes happens in a private method under ReplicationManager.createAndQueueTasks(Identifier pid). It seems to happen near the time of the request, but, there is a shift from direct execution of mn.replicate to submitting replicate requests by targetMN. Tasks can be for Pids (make more replicas) and PID+targetNode (call mn.replicate),

in Replication repository:
pid tasks can be in : NEW or IN_PROCESS state
in the object's systemMetadata
pid+target tasks can be : QUEUED, REQUESTED, FAILED, COMPLETED

ReplicationEventListener.entryUpdated() / .entryAdded()
(a Hz systemetadata map listener)
- if the pid's authoritativeMN is listed as a replica with status COMPLETE
create new replica task in the task repository unless there already is one
(if there are more than one task for the pid, delete existing ones and submit a new one)

(triggered every two minutes, trigger set up in ReplicationManager
ReplicationTaskProcessor.run()
move a page of tasks from NEW to IN_PROCESS status
- markInProcess
- call createAndQueueTasks (pid)

createAndQueueTasks (pid)

(lockPid)
processPid
removeReplicationTasks if sysmeta disallows replication or number of replicas are sufficient
determine potential target nodes (from NodeList and sysmeta)
createAndQueueTasks(pid, potentialTargets,desiredNumber)
loop to create a few per-target-node tasks
cnReplication.updateReplicationMetadata
if success,

requeueReplicationTask(pid) #return the pid to the NEW state because we don't assume the requisite number of replicas have been created
ReplicationTaskQueue.processAllTasksForMN(targetMN)
(lock the targetNode)
get replication tasks in the QUEUED state for this target node
foreach task, requestReplication (ReplicationService.requestQueuedReplication(pid,target)
targetMn.replicate(pid)
(unlock the targetNode)
(unlockPid)

#4 - 2017-10-24 16:46 - Dave Vieglais

- Target version changed from CCI-2.3.5 to CCI-2.3.7

#5 - 2017-10-31 17:46 - Dave Vieglais

- Target version changed from CCI-2.3.7 to CCI-2.3.8

#6 - 2018-01-17 18:40 - Dave Vieglais

- Sprint set to Infrastructure backlog

#7 - 2018-01-17 18:40 - Dave Vieglais

- Sprint changed from Infrastructure backlog to CCI-2.3.8

#8 - 2018-03-02 21:47 - Dave Vieglais

- Target version changed from CCI-2.3.8 to CCI-2.3.10

#9 - 2018-07-04 11:18 - Dave Vieglais

- Related to Story #8639: Replication performance is too slow to service demand added