Databus-lite performance benchmarking in AWS setup using S3

Setup details:
--------------
No. of EC2 instances: 2
EC2 instance types: m1-large (us-east-1 region)
Cluster setup: Node1 running producer + 1 scribe agent, Node2 running 1 scribe collector writing to S3.
Message size: 3000

Test 1) Find the maximum producer/agent throughput
----------------------------------------------------
In this test, we gradually increase the amount of data written by producer to local agent, until agent starts getting TRY_LATER msgs consistently (back-pressure) from downstream scribe collector; which will in turn cause data spooling at scribe agent.
In all tests, total number of messages written per producer thread: 100 K.

| Number of producer threads | Total produced Msg/sec | Total produced MB/s | Whether spooling at agent | Whether msg loss at producer |
|---|---|---|---|---|
| 3 | 1460 | 4.27 | No | No |
| 4 | 1930 | 5.65 | Yes | No |
| 6 | 2915 | 8.54 | Yes | No |
| 8 | 3740 | 11 | Yes | No |
| 10 | 4250 | 12.45 | Yes | No |

Conclusion:
-----------

1)       The scribe agent is able to write at least 4.27 MB/s without any spooling. Spooling starts happening consistently around 5.65 MB/s. Hence we can estimate the maximum acceptable write throughput of scribe agent that ensures steady state is ~ **4.5-5 MB/s**.
2) In all tests, no loss is observed at producer end. Hence the max acceptable write throughput for producer can be safely considered to be same as that for scribe agent.

Test 2) Find the maximum **collector** throughput
--------------------------------------------------

In this test, we gradually increase the amount of data written by producer to local agent. However spooling is now disabled at agent (by not configuring the secondary store). This is done to ensure that agent does not waste time in writing messages to disk/reading back from disk, and thus simulates writes as fast as possible to collector. In each run, the collector write throughput is calculated by looking at the S3 file sizes created during the run. The aim is to find the point when collector throughput maximizes.

In all tests, total number of messages written per producer thread: 100 K.

| Number of producer threads | Total produced Msg/sec | Total produced MB/s | Collector write throughput MB/S |
|---|---|---|---|
| 4 | 1930 | 5.65 | 9.1 |
| 6 | 2915 | 8.54 | 11.05 |
| 8 | 3740 | 11 | 11.2 |
| 10 | 4250 | 12.45 | 7.32 |
| 12 | 4966 | 14.56 | 6.34 |

Conclusion:
------------

1) The collector write throughput to S3 initially increases with load, peaks at 11.2 MB/s and thereafter gradually decreases. However there is no spooling happening at collector. This means the incoming rate at collector is decreasing beyond 8 producer threads, due to which it may not be able to exhaust S3.

Hence we can estimate the maximum collector write throughput to S3 is ~11 MB/s.

## Test 3:

Tried to run the test by changing the retry interval of agent to 10 secs and got the same collector throughput of 11.5 MB/s.
Number of agents : 2
Number of producer threads were : 16 (8 threads publishing messages to each agent)
Message Size : 3000
Message Per Sec : 500