[Revised] Compaction Perf Comparison Between 3.0 and 4.0 (STCS and LCS)

Setup

3 nodes cluster.
200+GB of data per node.
Single table.
Schema (for STCS):

```
CREATE KEYSPACE tlp_stress
WITH replication = {'class': 'NetworkTopologyStrategy', 'DC1': '3'}
AND durable_writes = true;

CREATE TABLE tlp_stress.keyvaluelargeblob_2 (
    key text,
    column_0 blob,
    column_1 blob,
    value blob,
    PRIMARY KEY (key, column_0, column_1)
) WITH CLUSTERING ORDER BY (column_0 ASC, column_1 ASC)
    AND additional_write_policy = '99p'
    AND bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND cdc = false
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactio
                      'max_threshold': '32', 'min_threshold': '4'}
    AND compression = {'chunk_length_in_kb': '16',
                       'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND crc_check_chance = 1.0
    AND default_time_to_live = 0
    AND extensions = {}
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair = 'BLOCKING'
    AND speculative_retry = '99p';
```

Schema (for LCS): only modify the compaction class to `org.apache.cassandra.db.compaction.LeveledCompactionStrategy`

Workload: WRITE : DELETE = 4 : 1

Result

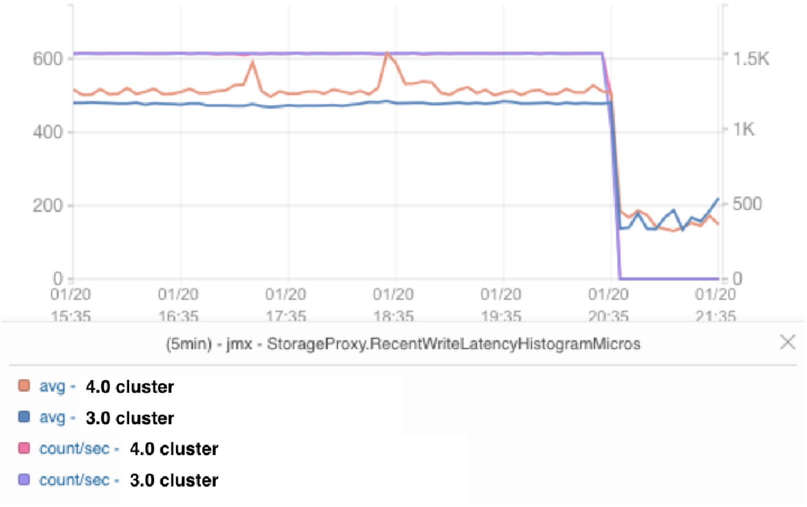
The compaction performance comparison mainly focused on the compaction related metrics, e.g. compaction throughput, pending tasks, the number of unleveled sstables (LCS only), etc. Query latency is **not** compared, since is can be affected by many other components that changed between 3.0 and 4.0.

Result of STCS

Under the same load, both 3.0 and 4.0 cluster show a similar compaction performance for STCS. The compaction in 4.0 runs slightly more actively.

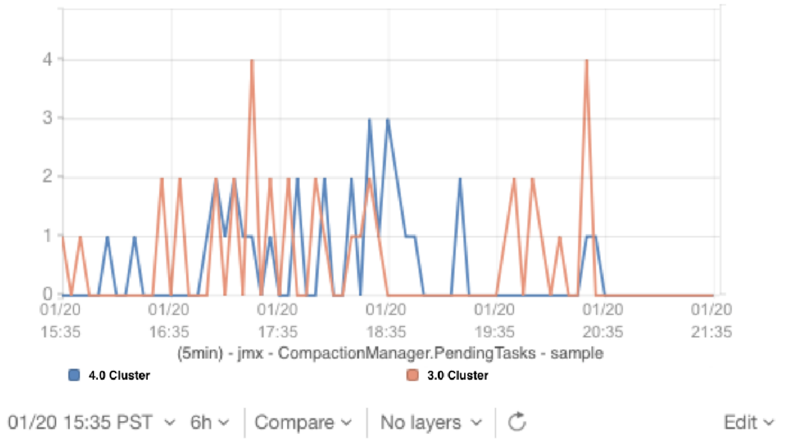
Write throughput

Write TPS (count/sec) is steady at 1.5k/sec throughout the test.



Pending tasks

Both 3.0 and 4.0 cluster have similar number of pending compaction tasks during the test.

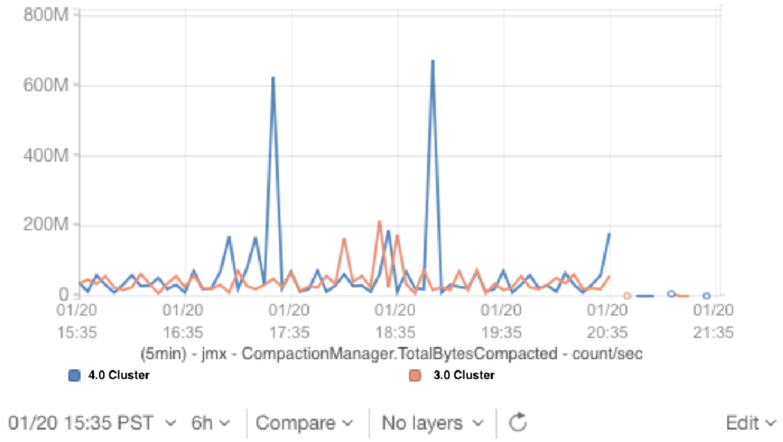


Compaction throughput

4.0 cluster shows higher peak compaction throughput than the 3.0 cluster. Other than the peak times, both 3.0 and 4.0 have similar compaction throughput.

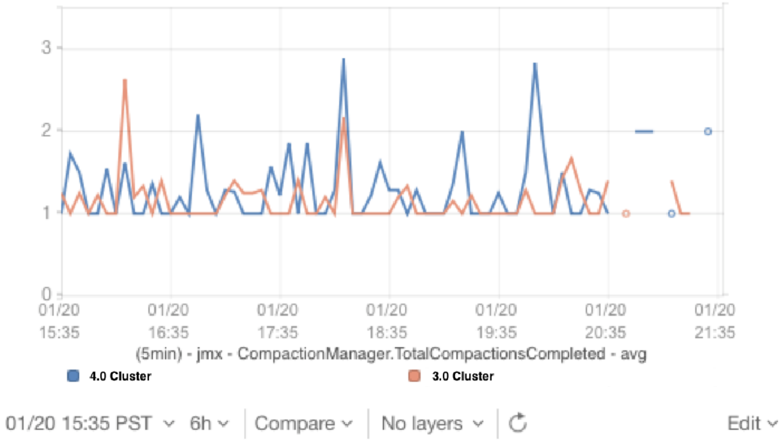
* The metric `TotalBytesCompacted -`

`count/sec` tracks the number of bytes compacted per seconds.



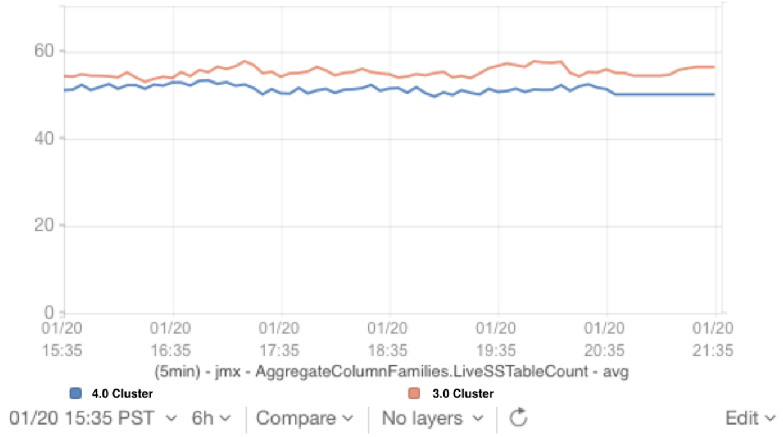
Compactions completed

The 4.0 cluster reports more number of compactions completed over time. It is more active on compaction.



Live sstables count

Both clusters have similar number of live sstables.

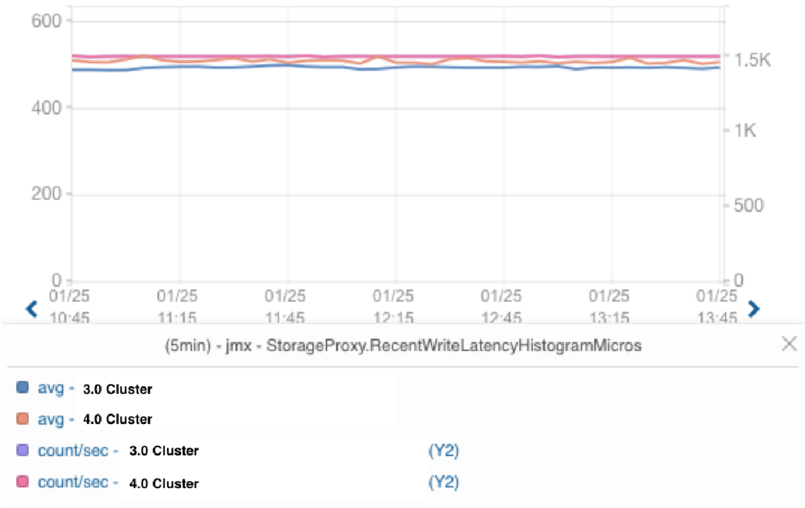


Result of LCS

Under the same load, the LCS compaction in 4.0 has better performance. The compaction in 3.0 cluster is lagging behind as the number of the unleveled sstables increases during the test. Meanwhile, the 4.0 cluster holds up to the load. The number of pending tasks, the compaction throughput and the unleveled sstables count remain steady throughout the test.

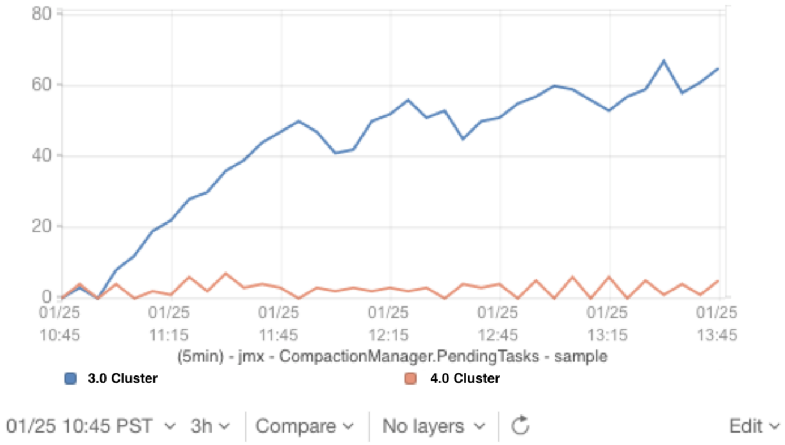
Write throughput

Write TPS is steady at 1.5k/s.



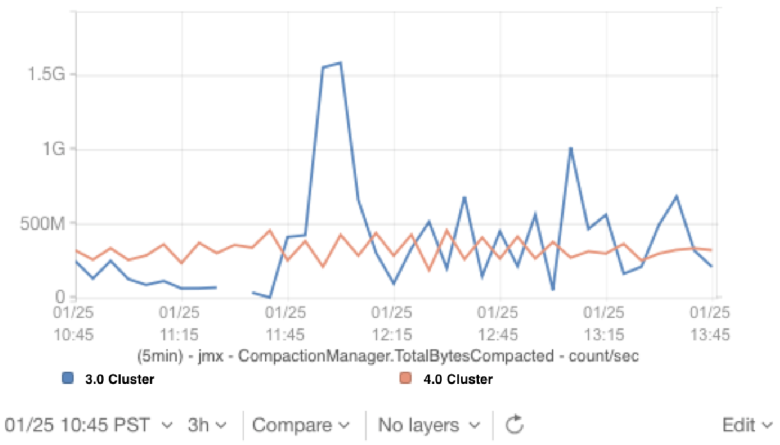
Pending tasks

The 3.0 cluster shows more pending tasks towards the end of the test. The pending tasks is increasing in the 3.0 cluster.
The number of the pending tasks is steady in the 4.0 cluster.



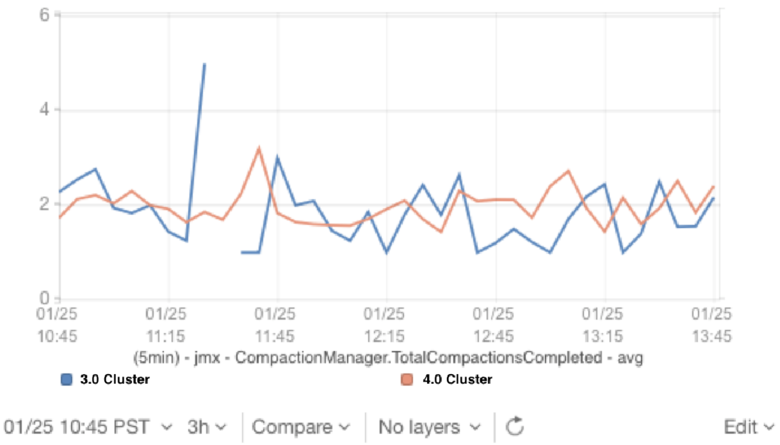
Compaction throughput

The 3.0 cluster has noticeably higher compaction throughput that is measured by compacted bytes per second. The large spike correlates with the spike in the unleveled sstables count chart below.
Meanwhile, the 4.0 cluster compaction throughput is steady.



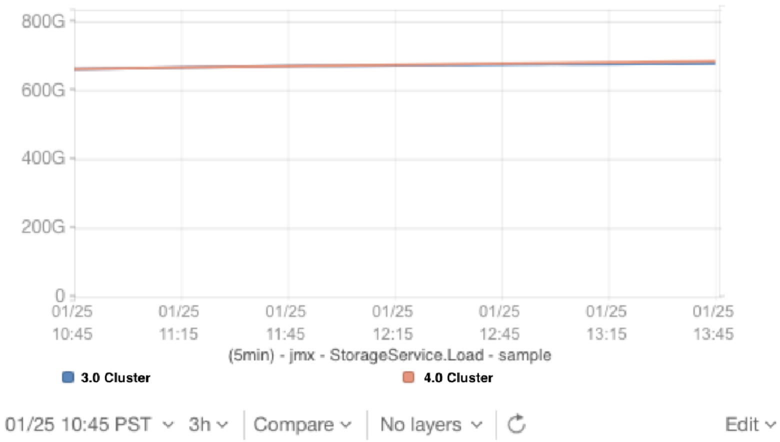
Compactions completed

Both 3.0 and 4.0 cluster indicate similar number of compactions completed.



Node load

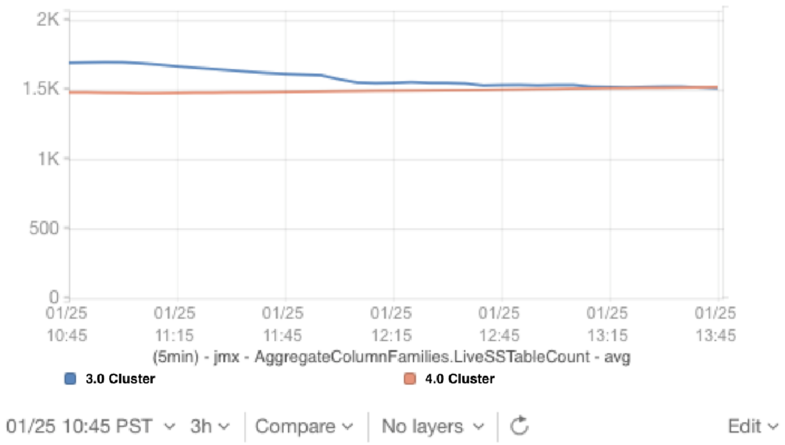
The data size on disk is almost identical for both 3.0 and 4.0 clusters.



Live sstables count

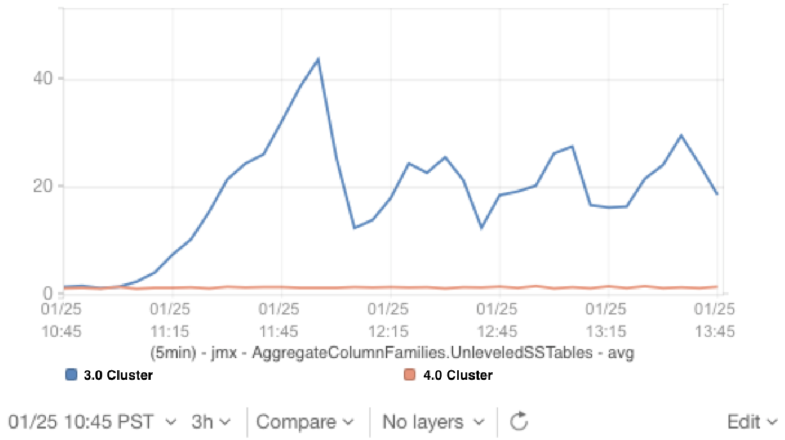
Both 3.0 and 4.0 clusters have similar

number of live sstables towards the end of the test.
(The 3.0 cluster has slightly more live sstables count just after the data population.)



Unleveled sstables count

Compaction keeps up with the load in the 4.0 cluster. The number of the unleveled sstables remains steady. However, in the 3.0 cluster, the compaction is lagging behind. The number of unleveled sstables increases during the test, and it stays around 20.


01/25 10:45
01/25 11:15
01/25 11:45
01/25 12:15
01/25 12:45
01/25 13:15
01/25 13:45
(5min) - jmx - AggregateColumnFamilies.UnleveledSSTables - avg
🟦 3.0 Cluster
🟧 4.0 Cluster
01/25 10:45 PST
3h
Compare
No layers
Edit