CSCI 5817: Project 1
Spring 18

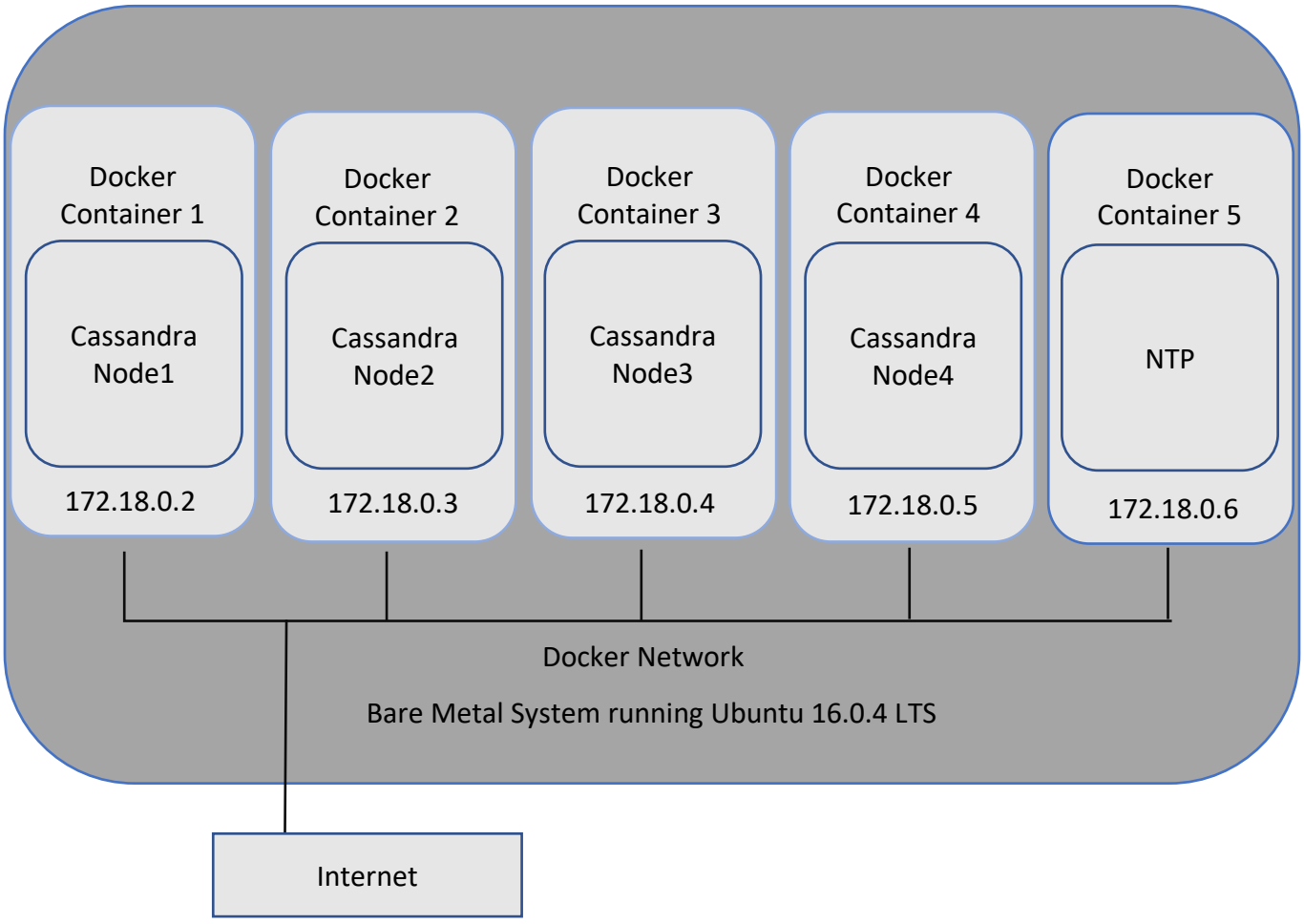
Measuring Latencies and Clock Skew in a Cassandra Cluster

By
Ilamvazhuthy Subbiah
Saikrishna Jaliparthy

Introduction

The objective of this project is to study how the performance of a Cassandra cluster varies with respect to different topologies. Three different topologies are constructed with varying physical separation between the nodes which bring with them varying levels of network delay. The Cassandra nodes are run inside a Docker container in all the topologies and all the nodes are synchronized to a single NTP server. An Overlay network is created to connect the Docker containers whenever needed. Read, write and mixed latencies have been measured using the Cassandra-stress tool. The time delay and offset between two nodes were measured by running the python scripts on the nodes.

Topology 1



Topology 1 consists of a Bare metal system running an Ubuntu VM. There are five Docker containers running inside the VM, four of them running a Cassandra node inside them and the fifth container runner an NTP server. The four Cassandra nodes are synchronized to the NTP server running on 172.18.0.6 and the synchronization is verified by using the commands *ntpstat* and *ntpdate*.

Measurements - Clock Skew

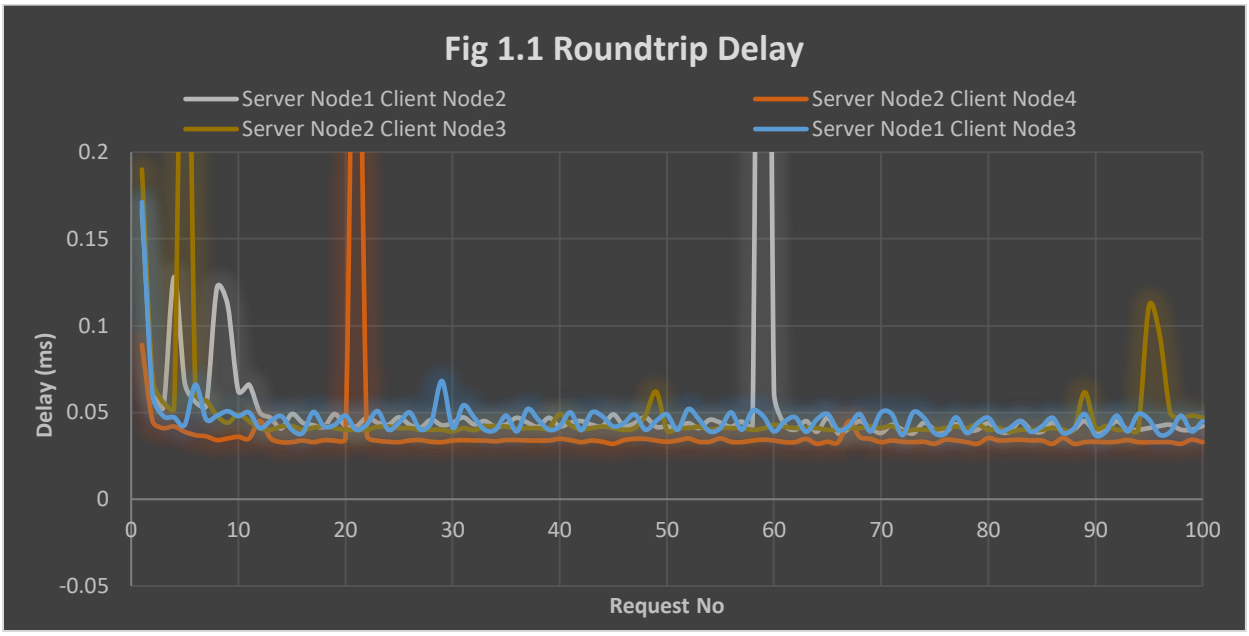
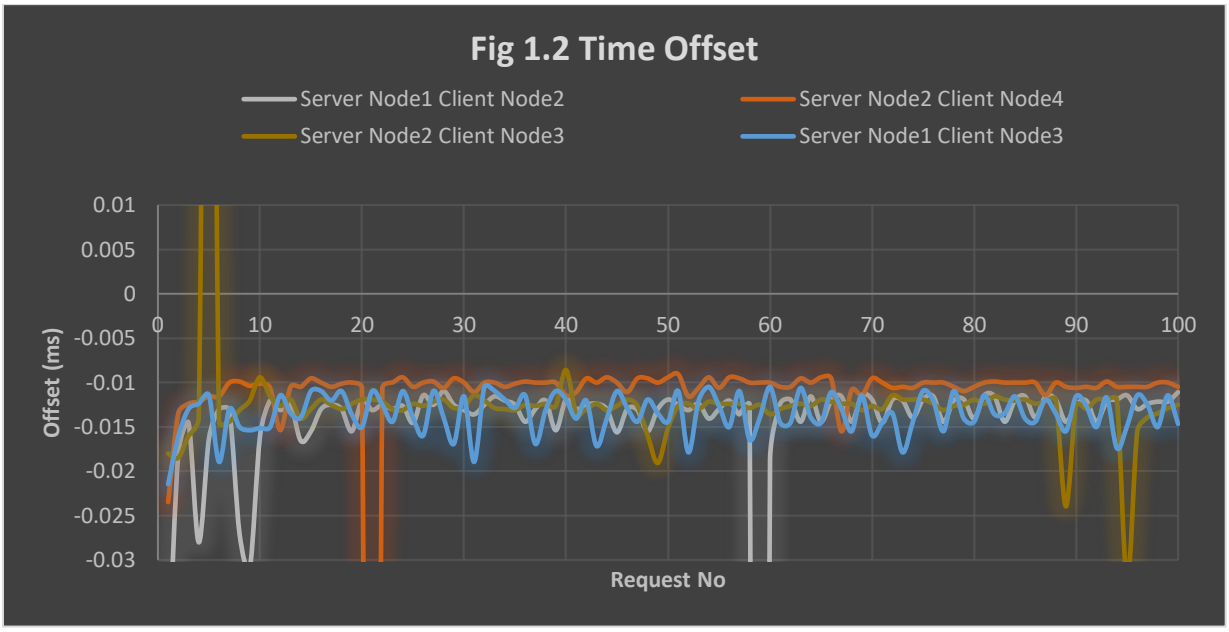
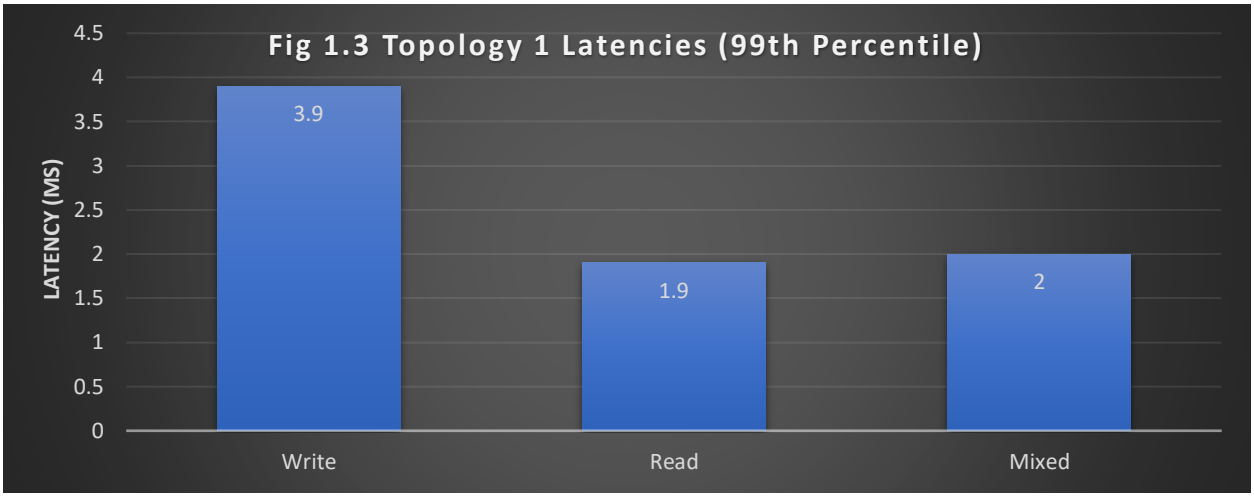


Fig 1.1 and Fig 1.2 show the roundtrip delay and the time offset between different nodes in the topology, respectively

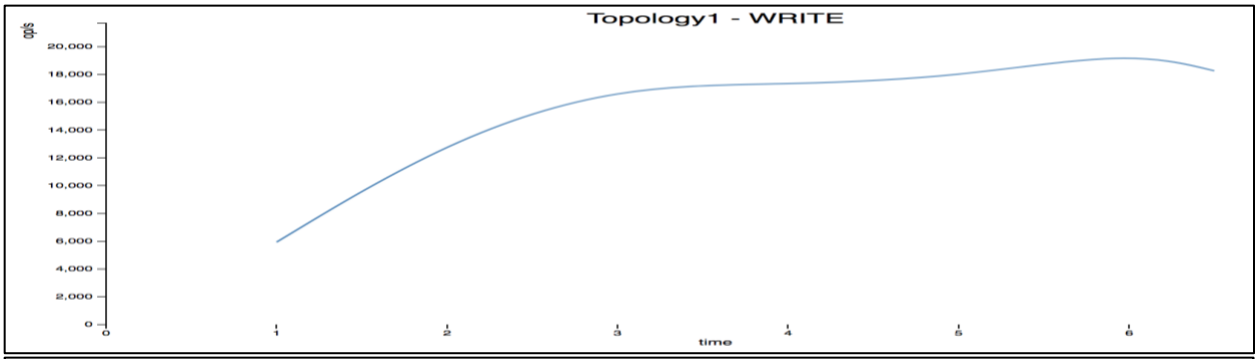


Latencies

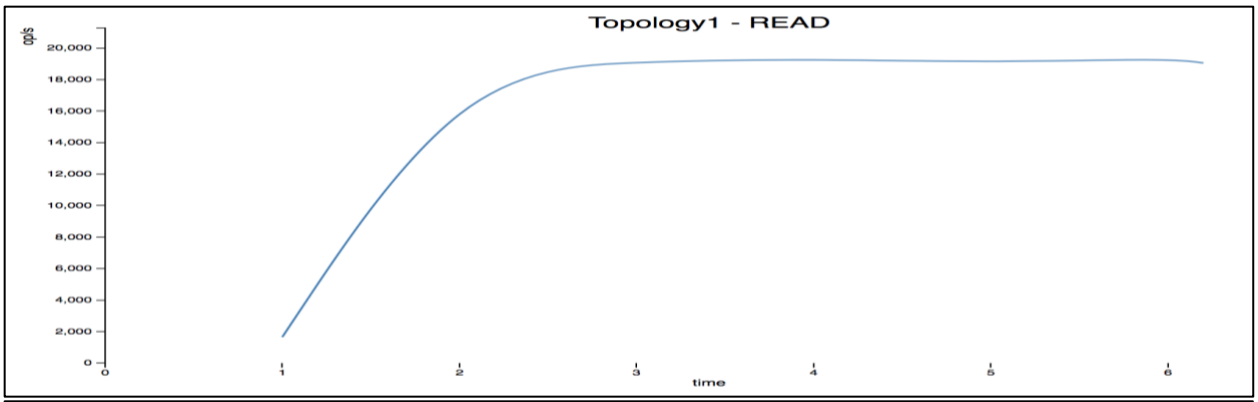
The Figure below shows the Write, read and mixed latencies for Topology 1. The write latencies are higher than the read and mixed latencies.



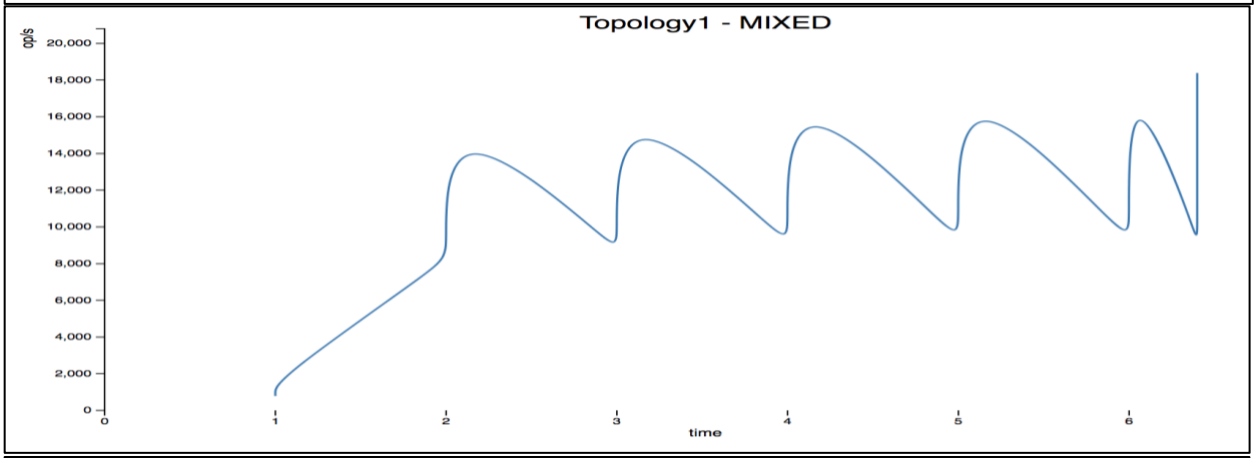
The Three figures below show the write, read and mixed latencies for 100000 operations run on 10 threads.




```
benchmark-0
----------------------------------------
op rate                   : 15,439 op/s [WRITE: 15,439 op/s]
partition rate            : 15,439 pk/s [WRITE: 15,439 pk/s]
row rate                  : 15,439 row/s [WRITE: 15,439 row/s]
latency mean              : 0.6 ms [WRITE: 0.6 ms]
latency median            : 0.4 ms [WRITE: 0.4 ms]
latency 95th percentile   : 1.5 ms [WRITE: 1.5 ms]
latency 99th percentile   : 3.9 ms [WRITE: 3.9 ms]
latency 99.9th percentile : 12.4 ms [WRITE: 12.4 ms]
latency max               : 65.6 ms [WRITE: 65.6 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:00:06
cmd: write n=100000 -rate threads=10 -graph file=Topology1.html title=Topology1 revision=benchmark-0
```

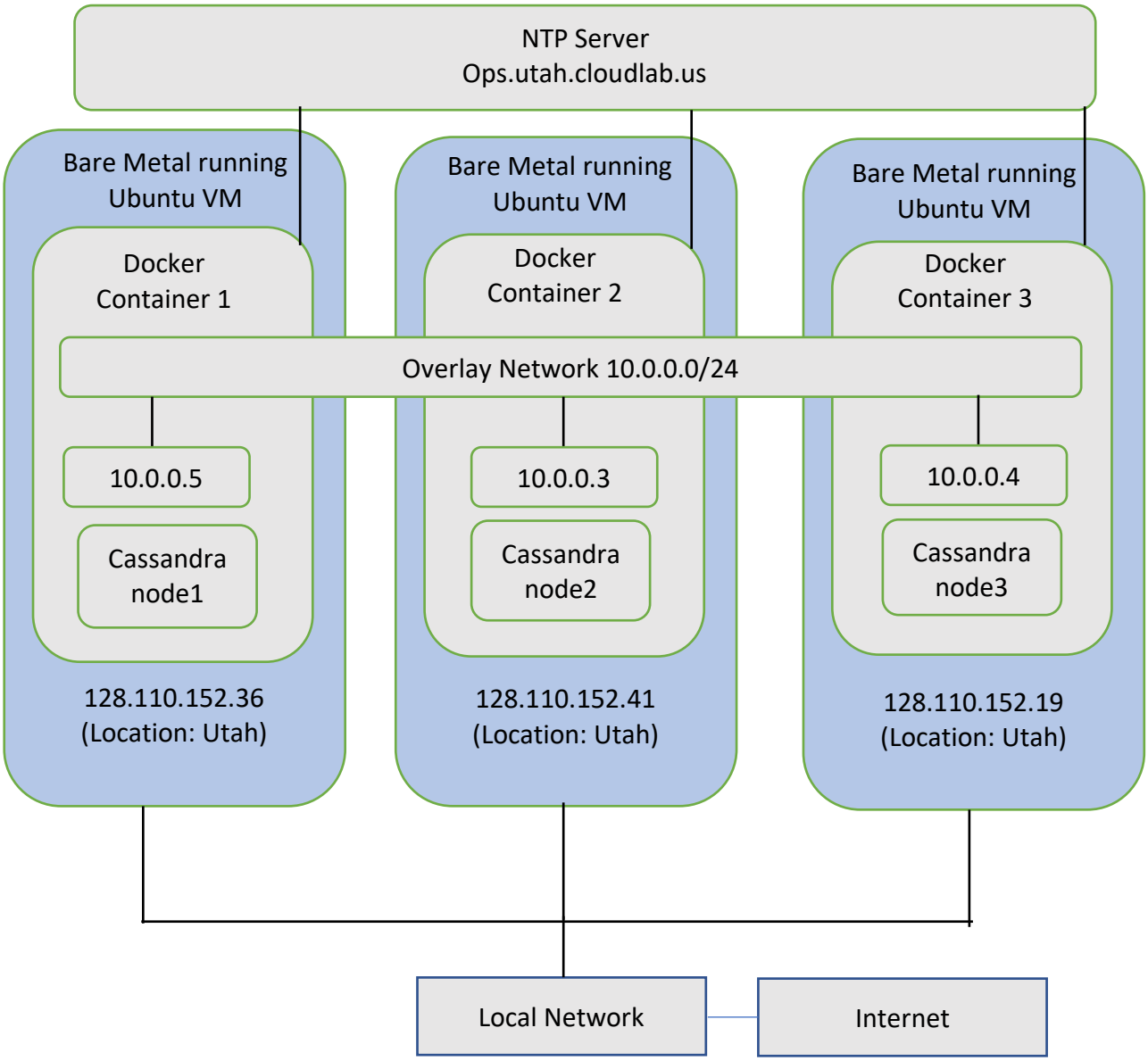



```
benchmark-0
----------------------------------------
op rate                   : 16,247 op/s [READ: 16,247 op/s]
partition rate            : 16,247 pk/s [READ: 16,247 pk/s]
row rate                  : 16,247 row/s [READ: 16,247 row/s]
latency mean              : 0.5 ms [READ: 0.5 ms]
latency median            : 0.4 ms [READ: 0.4 ms]
latency 95th percentile   : 1.0 ms [READ: 1.0 ms]
latency 99th percentile   : 1.9 ms [READ: 1.9 ms]
latency 99.9th percentile : 3.7 ms [READ: 3.7 ms]
latency max               : 27.2 ms [READ: 27.2 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:00:06
cmd: read n=100000 -rate threads=10 -graph file=Topology1.html title=Topology1 revision=benchmark-0
```




```
benchmark-0
----------------------------------------
op rate                   : 15,574 op/s [READ: 7,686 op/s, WRITE: 7,887 op/s]
partition rate            : 15,574 pk/s [READ: 7,686 pk/s, WRITE: 7,887 pk/s]
row rate                  : 15,574 row/s [READ: 7,686 row/s, WRITE: 7,887 row/s]
latency mean              : 0.5 ms [READ: 0.5 ms, WRITE: 0.5 ms]
latency median            : 0.4 ms [READ: 0.4 ms, WRITE: 0.4 ms]
latency 95th percentile   : 1.2 ms [READ: 1.2 ms, WRITE: 1.1 ms]
latency 99th percentile   : 2.0 ms [READ: 2.0 ms, WRITE: 2.0 ms]
latency 99.9th percentile : 7.0 ms [READ: 7.2 ms, WRITE: 6.7 ms]
latency max               : 48.2 ms [READ: 48.1 ms, WRITE: 48.2 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:00:06
cmd: mixed n=100000 -rate threads=10 -graph file=Topology1.html title=Topology1 revision=benchmark-0
```

Topology 2



Topology 2 consists of 3 bare metal systems, each running an Ubuntu VM. All 3 systems are located in Utah and are connected to the local network which in turn is connected to the internet. Inside each VM there is a Docker container running and inside each Docker container there is a Cassandra node. An Overlay network was created in Docker swarm mode to connect the Docker containers.

For time synchronization, we synchronized all the Docker containers to the NTP server ops.utah.cloudlab.us and the synchronization was verified using *ntpstat* and *ntpdate*. We were not able to run our own NTP server in the Docker swarm as the Docker *service* command did not allow us to use certain flags that we need for the NTP server to work properly.

Measurements

Clock Skew

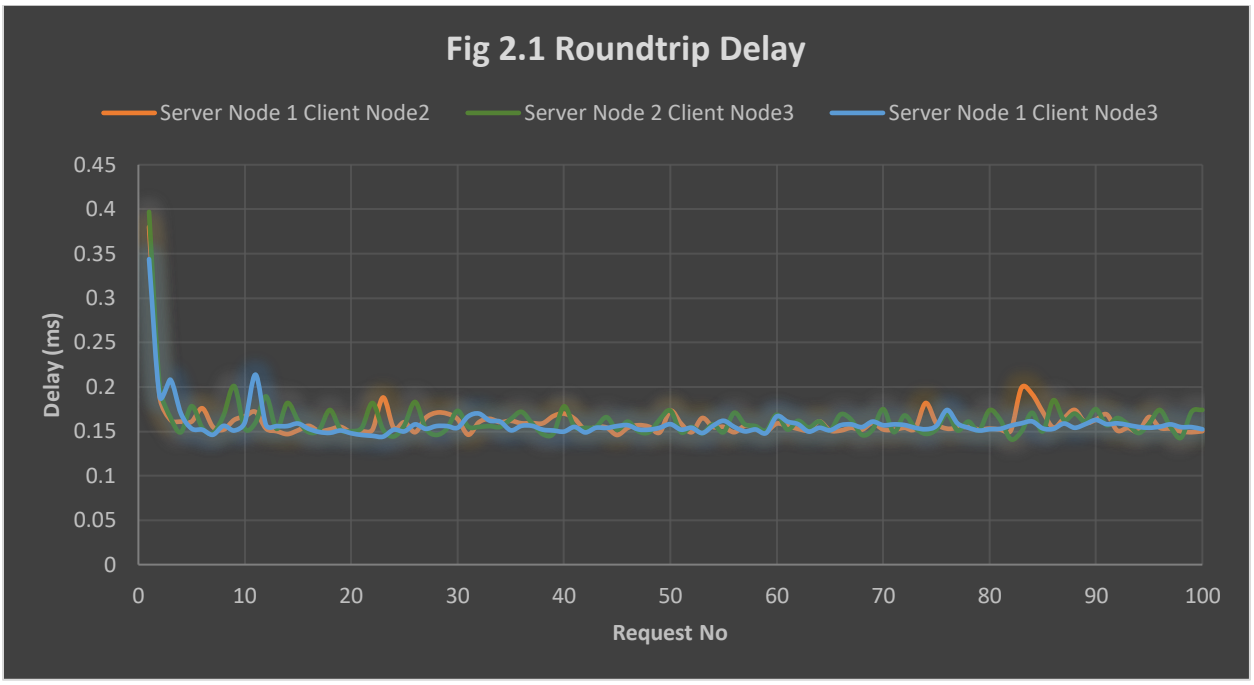
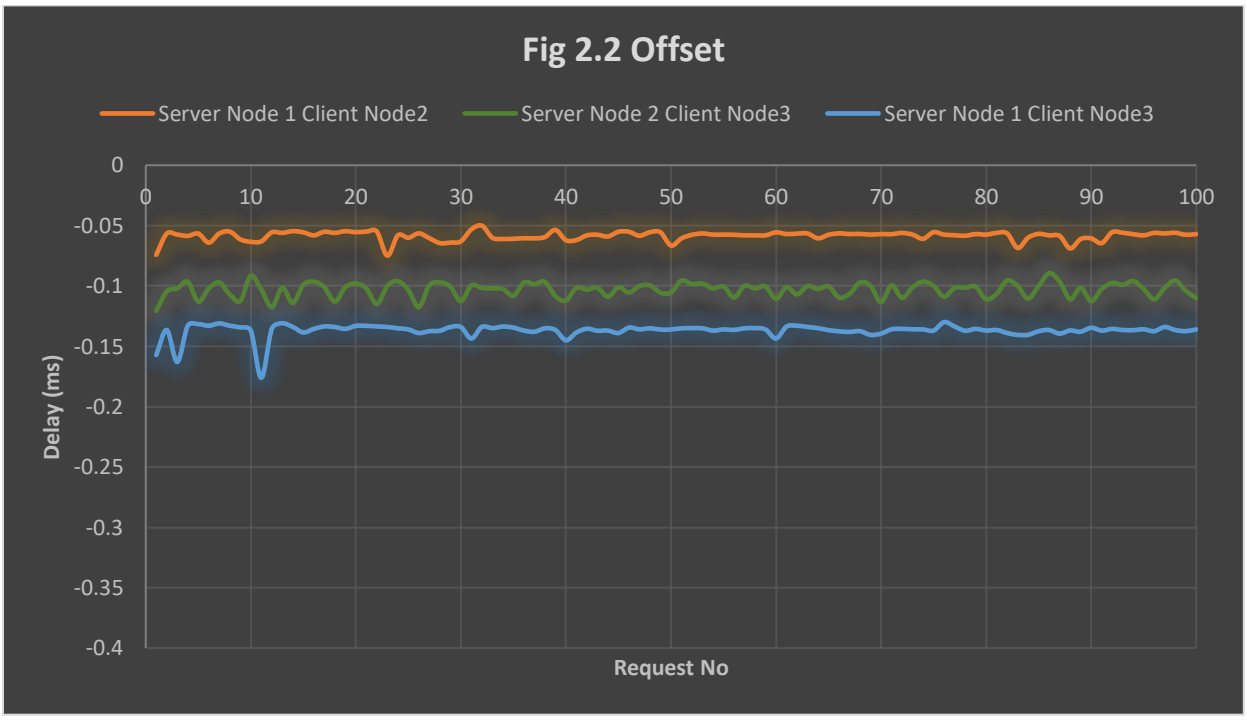
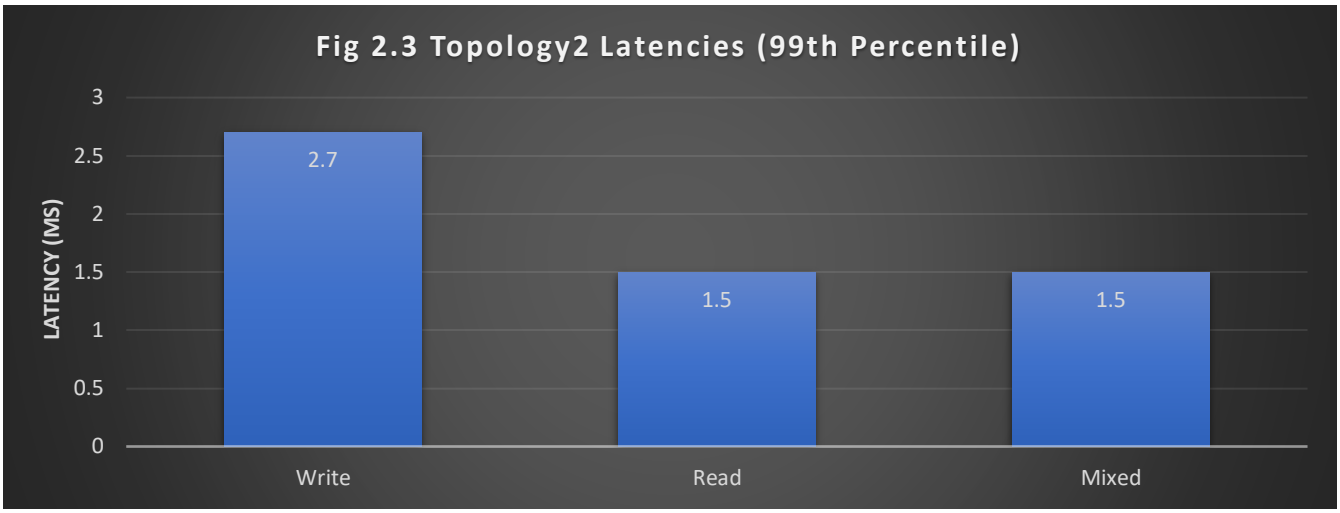


Fig 2.1 and Fig 2.2 show the roundtrip delay and the time offset between different nodes in the topology, respectively

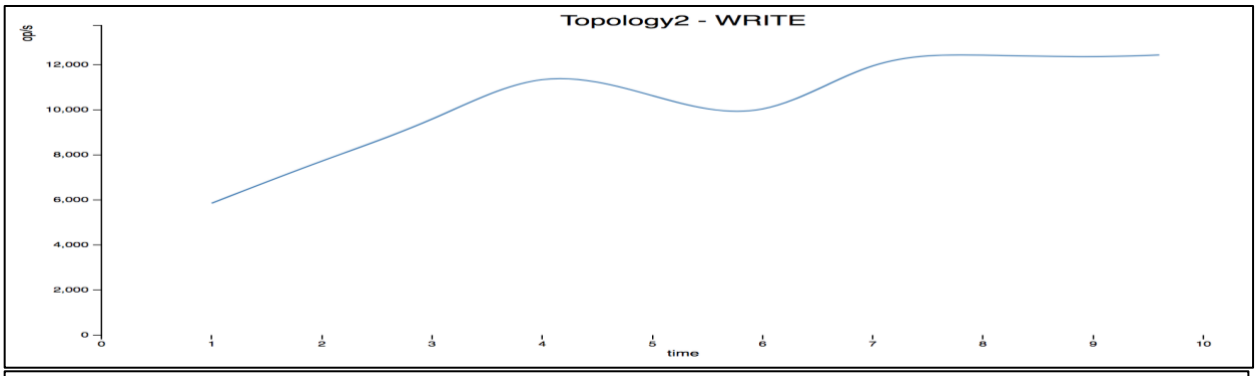


Latencies

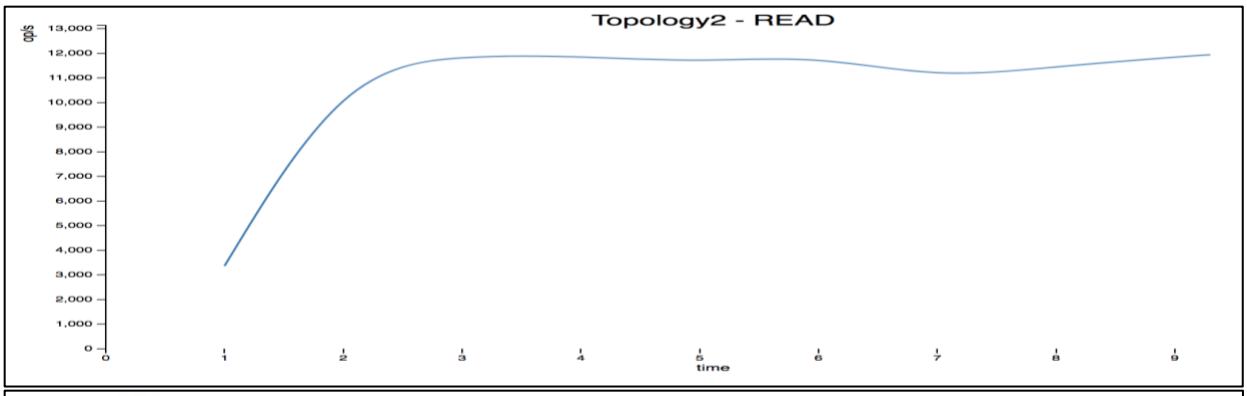
The Figure 2.3 below shows the Write, read and mixed latencies for Topology 2. Similar to Topology 1, the write latencies are higher than the read and mixed latencies.



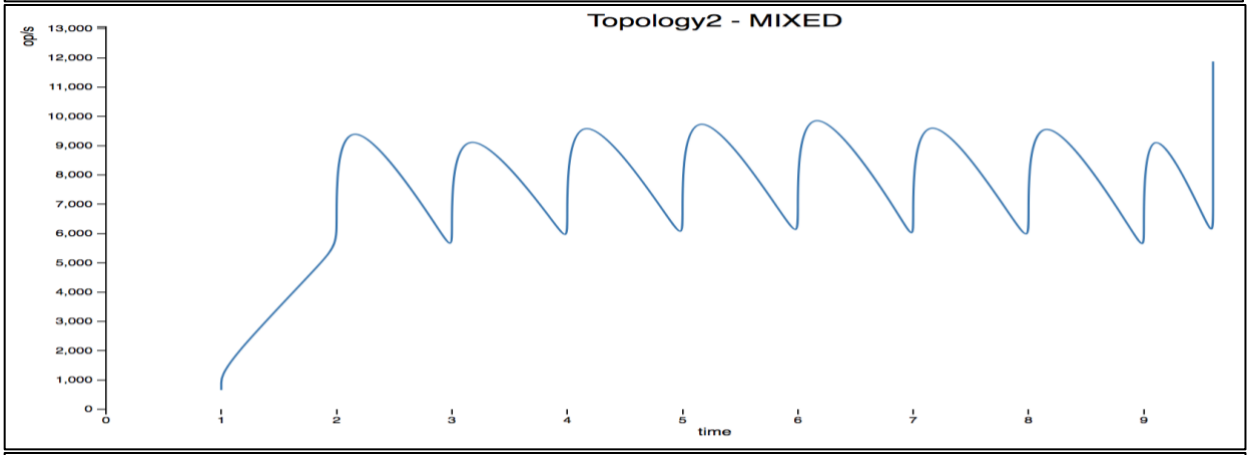
The Three figures below show the write, read and mixed latencies for 100000 operations run on 10 threads.




```
benchmark-0
----------------------------------------
op rate                   : 10,386 op/s [WRITE: 10,386 op/s]
partition rate            : 10,386 pk/s [WRITE: 10,386 pk/s]
row rate                  : 10,386 row/s [WRITE: 10,386 row/s]
latency mean              : 0.9 ms [WRITE: 0.9 ms]
latency median            : 0.8 ms [WRITE: 0.8 ms]
latency 95th percentile   : 1.4 ms [WRITE: 1.4 ms]
latency 99th percentile   : 2.7 ms [WRITE: 2.7 ms]
latency 99.9th percentile : 12.0 ms [WRITE: 12.0 ms]
latency max               : 129.0 ms [WRITE: 129.0 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:00:09
cmd: write n=100000 -node 10.0.0.3,10.0.0.4,10.0.0.5 -rate threads=10 -graph file=Topo2.html
```

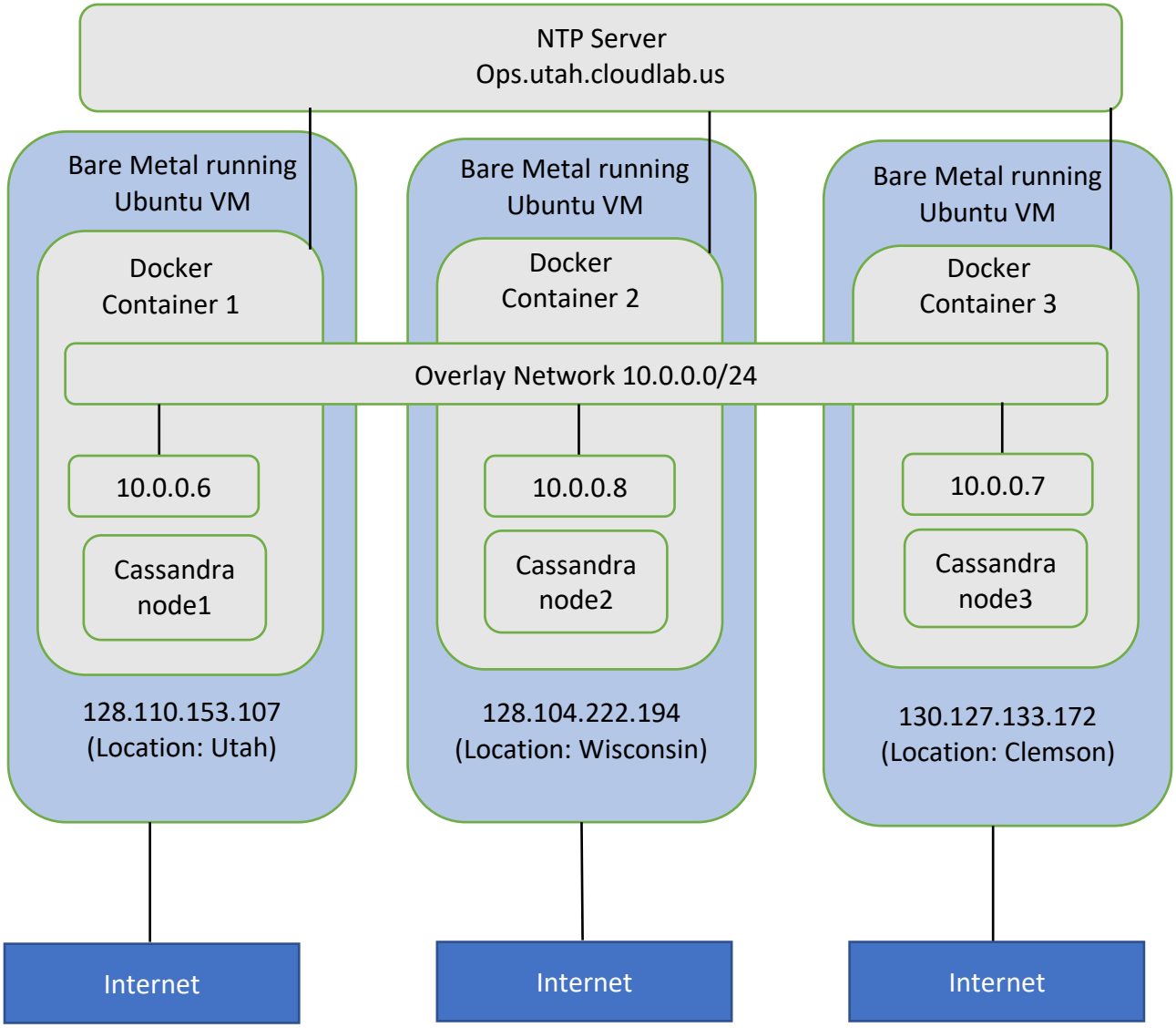



```
benchmark-0
----------------------------------------
op rate                   : 10,734 op/s [READ: 10,734 op/s]
partition rate            : 10,734 pk/s [READ: 10,734 pk/s]
row rate                  : 10,734 row/s [READ: 10,734 row/s]
latency mean              : 0.8 ms [READ: 0.8 ms]
latency median            : 0.8 ms [READ: 0.8 ms]
latency 95th percentile   : 1.2 ms [READ: 1.2 ms]
latency 99th percentile   : 1.5 ms [READ: 1.5 ms]
latency 99.9th percentile : 6.2 ms [READ: 6.2 ms]
latency max               : 74.6 ms [READ: 74.6 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:00:09
cmd: read n=100000 -node 10.0.0.3,10.0.0.4,10.0.0.5 -rate threads=10 -graph file=Topo2.html
```




```
benchmark-0
----------------------------------------
op rate                   : 10,381 op/s [READ: 5,191 op/s, WRITE: 5,190 op/s]
partition rate            : 10,381 pk/s [READ: 5,191 pk/s, WRITE: 5,190 pk/s]
row rate                  : 10,381 row/s [READ: 5,191 row/s, WRITE: 5,190 row/s]
latency mean              : 0.8 ms [READ: 0.9 ms, WRITE: 0.8 ms]
latency median            : 0.8 ms [READ: 0.8 ms, WRITE: 0.8 ms]
latency 95th percentile   : 1.2 ms [READ: 1.2 ms, WRITE: 1.2 ms]
latency 99th percentile   : 1.5 ms [READ: 1.5 ms, WRITE: 1.5 ms]
latency 99.9th percentile : 8.1 ms [READ: 7.7 ms, WRITE: 8.1 ms]
latency max               : 83.0 ms [READ: 83.0 ms, WRITE: 82.6 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:00:09
cmd: mixed n=100000 -node 10.0.0.3,10.0.0.4,10.0.0.5 -rate threads=10 -graph file=Topo2.html
```

Topology 3



Topology 3 consists of 3 bare metal systems, each running an Ubuntu VM. The 3 machines are located in Utah, Wisconsin and Clemson respectively and are connected to the internet. Inside each VM there is a Docker container running and inside each Docker container there is a Cassandra node. An Overlay network was created in Docker swarm mode to connect the Docker containers.

For time synchronization, we synchronized all the Docker containers to the NTP server ops.utah.cloudlab.us and the synchronization was verified using *ntpstat* and *ntpdate*. We were not able to run our own NTP server in the Docker swarm as the Docker *service* command did not allow us to use certain flags that we need for the NTP server to work properly.

Measurements

Clock Skew

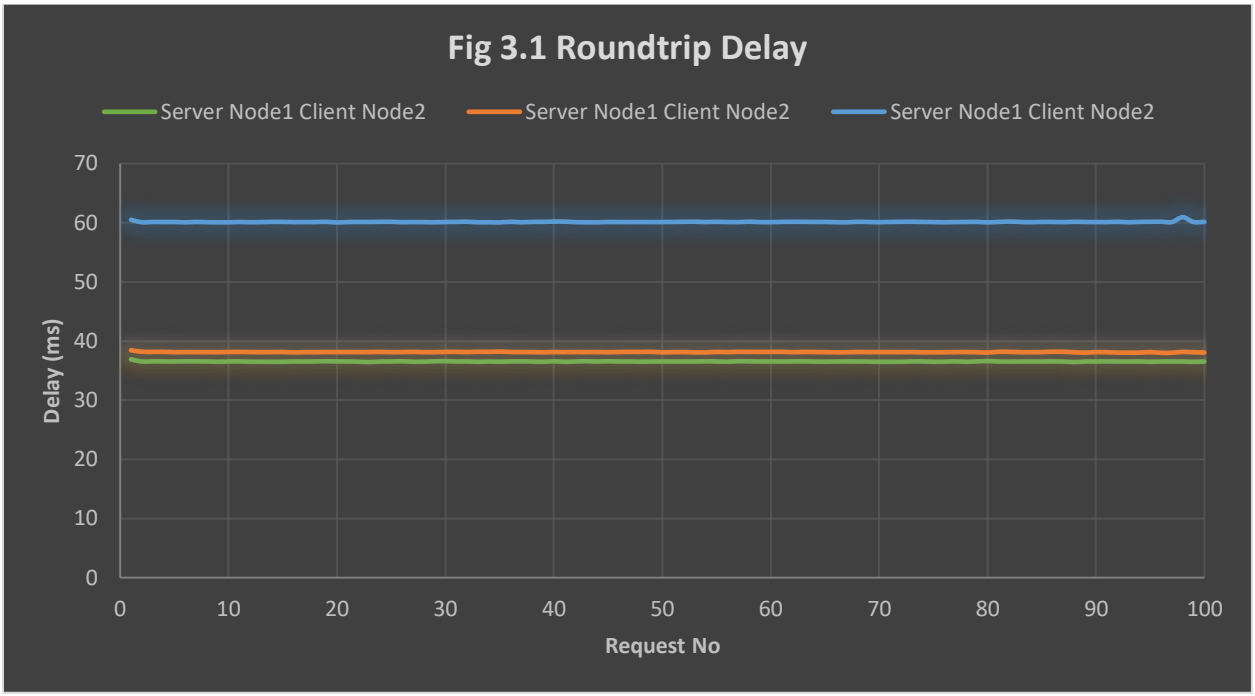
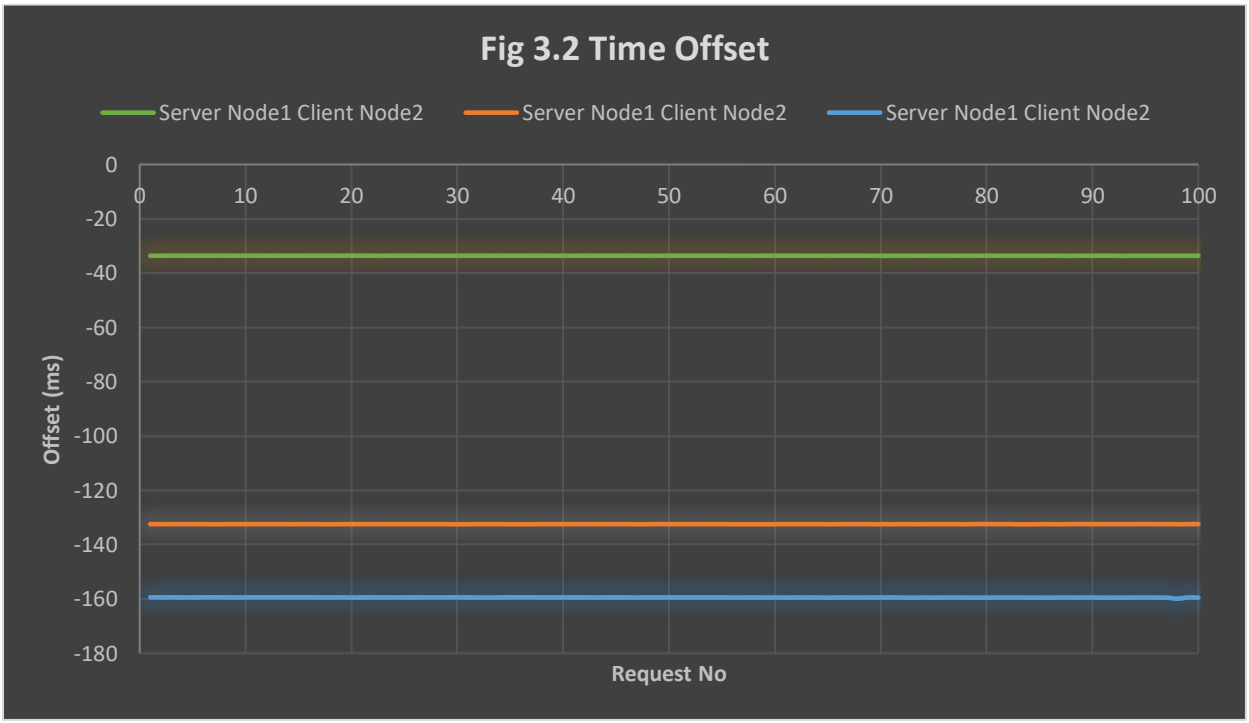
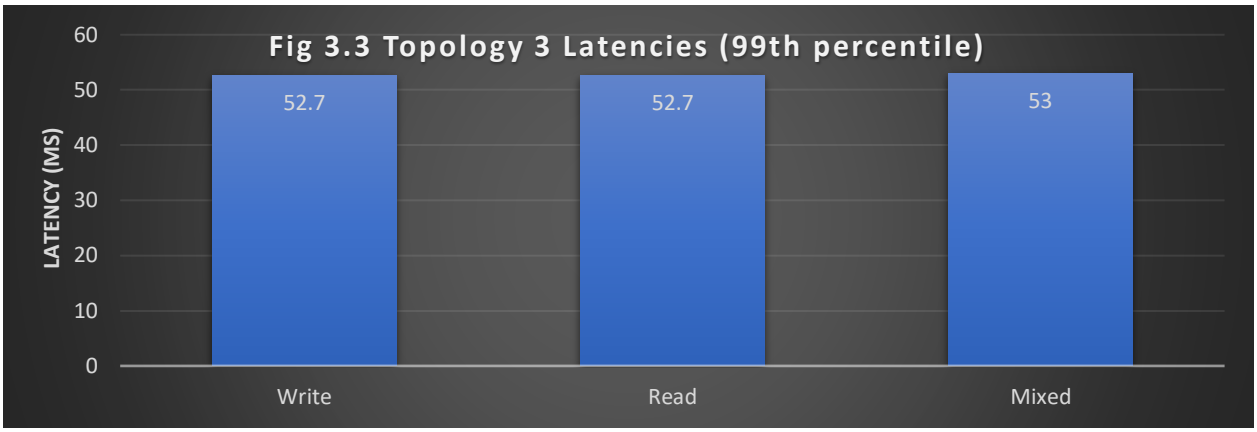


Fig 1.1 and Fig 1.2 show the roundtrip delay and the time offset between different nodes in the topology, respectively

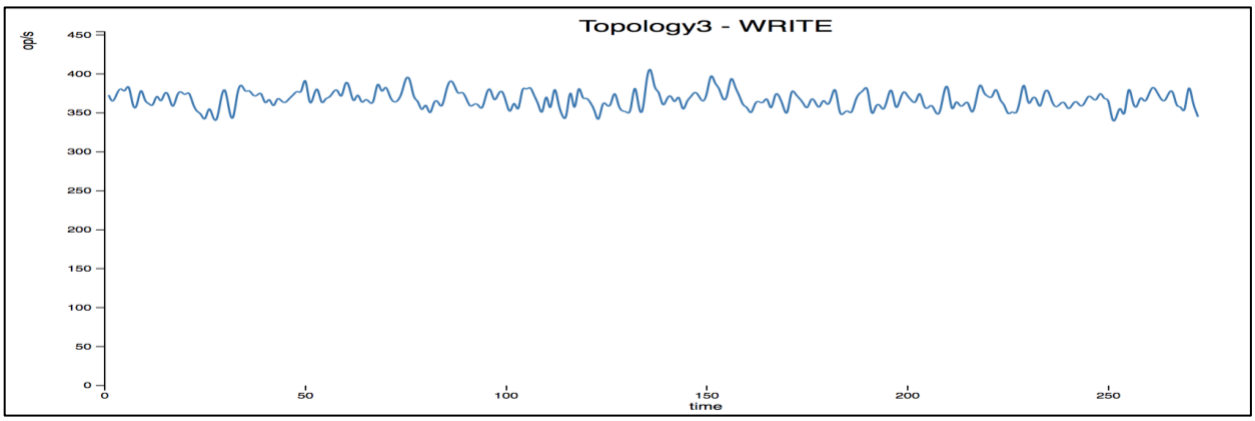


Latencies

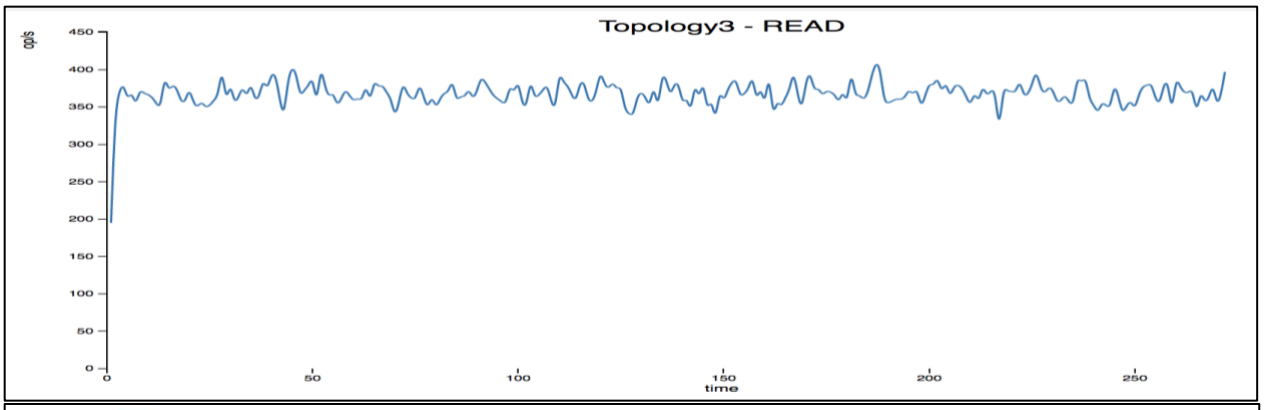
The Figure 3.3 below shows the Write, read and mixed latencies for Topology 3.



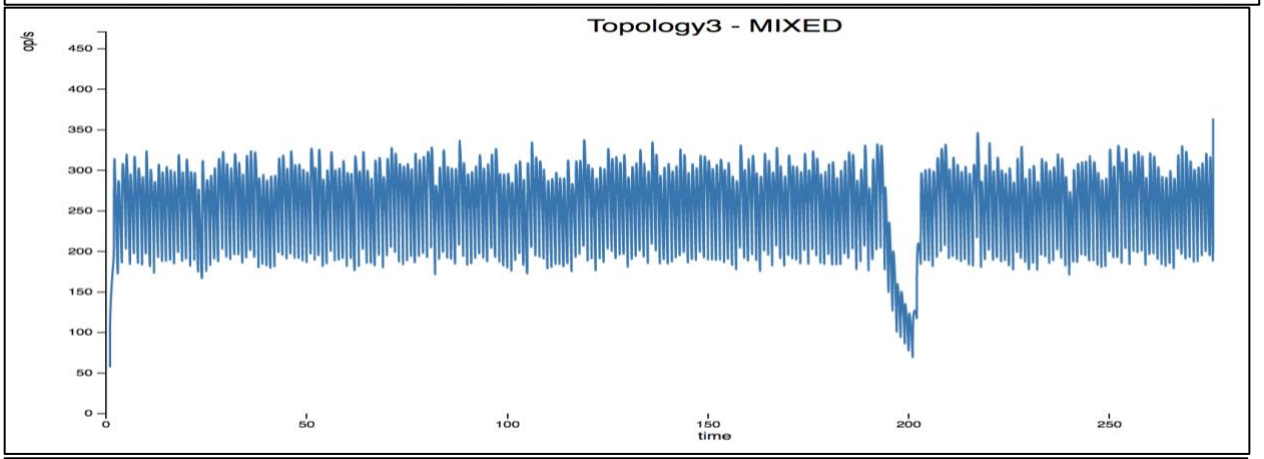
The Three figures below show the write, read and mixed latencies for 100000 operations run on 10 threads.




```
benchmark-0
---------------------------------------
op rate                   : 367 op/s [WRITE: 367 op/s]
partition rate            : 367 pk/s [WRITE: 367 pk/s]
row rate                  : 367 row/s [WRITE: 367 row/s]
latency mean              : 27.2 ms [WRITE: 27.2 ms]
latency median            : 26.3 ms [WRITE: 26.3 ms]
latency 95th percentile   : 52.5 ms [WRITE: 52.5 ms]
latency 99th percentile   : 52.7 ms [WRITE: 52.7 ms]
latency 99.9th percentile : 53.1 ms [WRITE: 53.1 ms]
latency max               : 110.0 ms [WRITE: 110.0 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:04:32
cmd: write n=100000 -node 10.0.0.6,10.0.0.7,10.0.0.8 -rate threads=10 -graph file=Topo3.html
```

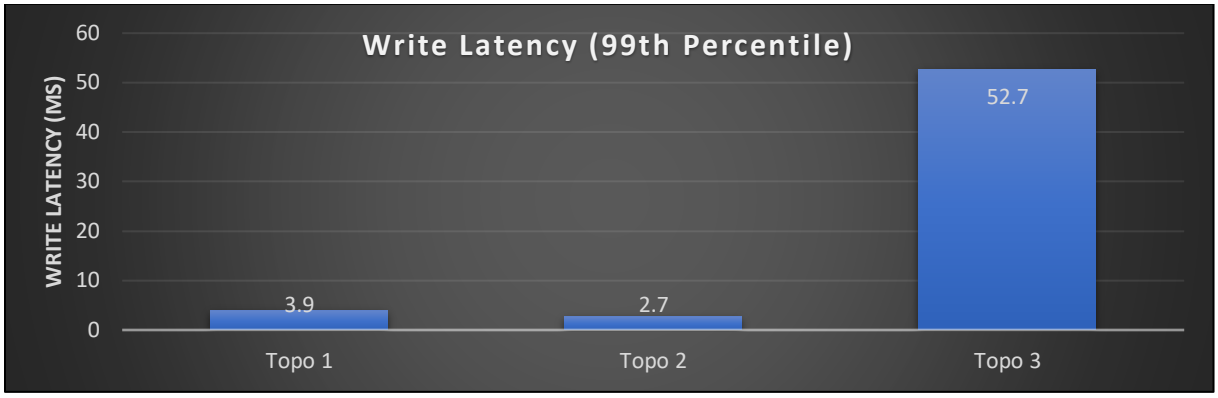


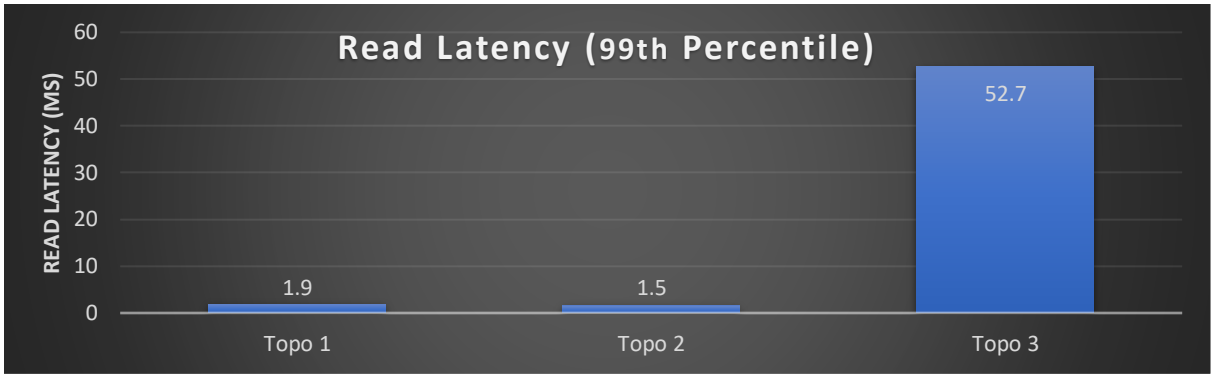

```
benchmark-0
---------------------------------------
op rate                   : 368 op/s [READ: 368 op/s]
partition rate            : 368 pk/s [READ: 368 pk/s]
row rate                  : 368 row/s [READ: 368 row/s]
latency mean              : 27.1 ms [READ: 27.1 ms]
latency median            : 26.3 ms [READ: 26.3 ms]
latency 95th percentile   : 52.5 ms [READ: 52.5 ms]
latency 99th percentile   : 52.7 ms [READ: 52.7 ms]
latency 99.9th percentile : 53.1 ms [READ: 53.1 ms]
latency max               : 84.5 ms [READ: 84.5 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:04:31
cmd: read n=100000 -node 10.0.0.6,10.0.0.7,10.0.0.8 -rate threads=10 -graph file=Topo3.html
```

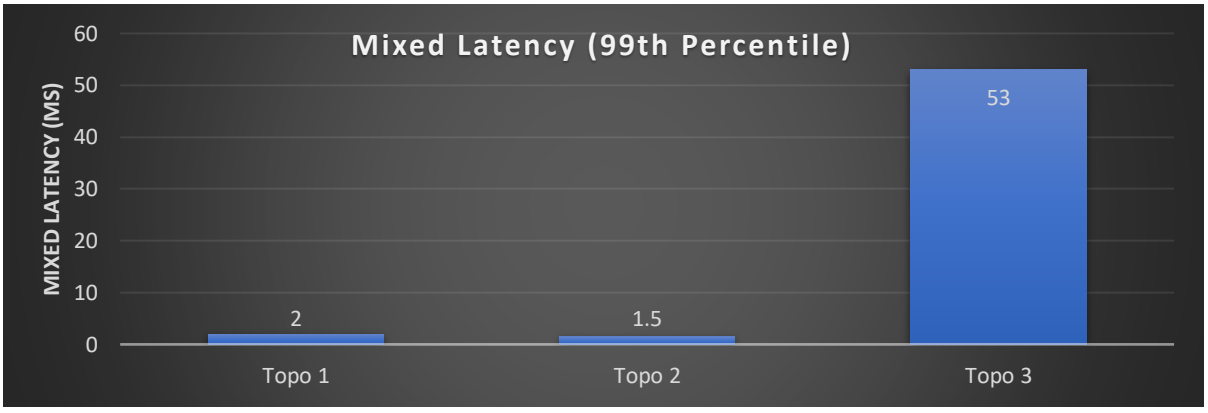



```
benchmark-0
---------------------------------------
op rate                   : 363 op/s [READ: 182 op/s, WRITE: 180 op/s]
partition rate            : 363 pk/s [READ: 182 pk/s, WRITE: 180 pk/s]
row rate                  : 363 row/s [READ: 182 row/s, WRITE: 180 row/s]
latency mean              : 27.5 ms [READ: 27.5 ms, WRITE: 27.4 ms]
latency median            : 26.3 ms [READ: 26.3 ms, WRITE: 26.3 ms]
latency 95th percentile   : 52.5 ms [READ: 52.5 ms, WRITE: 52.5 ms]
latency 99th percentile   : 53.0 ms [READ: 53.1 ms, WRITE: 53.0 ms]
latency 99.9th percentile : 111.7 ms [READ: 112.7 ms, WRITE: 111.3 ms]
latency max               : 131.1 ms [READ: 130.8 ms, WRITE: 131.1 ms]
total gc count            : 0
total gc memory           : 0.000 KiB
total gc time             : 0.0 seconds
avg gc time               : NaN ms
stddev gc time            : 0.0 ms
Total operation time      : 00:04:35
cmd: mixed n=100000 -node 10.0.0.6,10.0.0.7,10.0.0.8 -rate threads=10 -graph file=Topo3.html
```

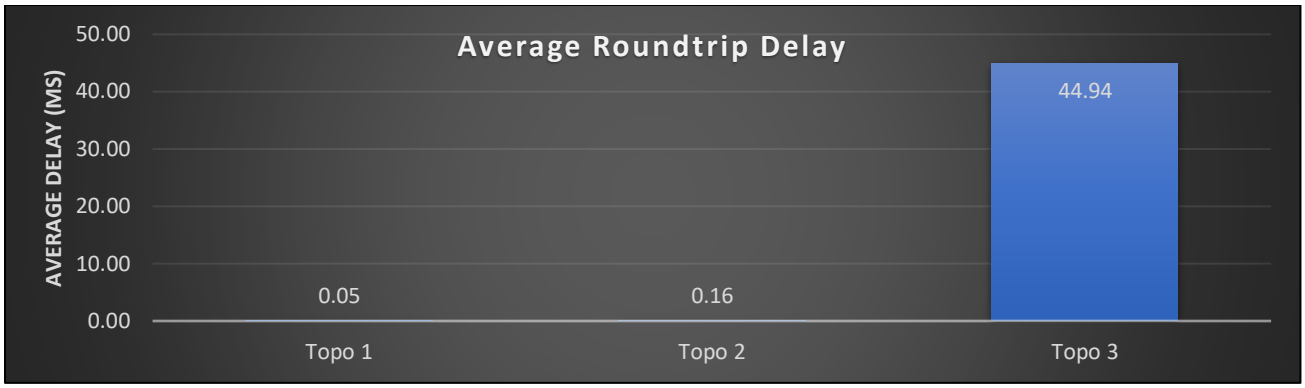
Comparison and Conclusion

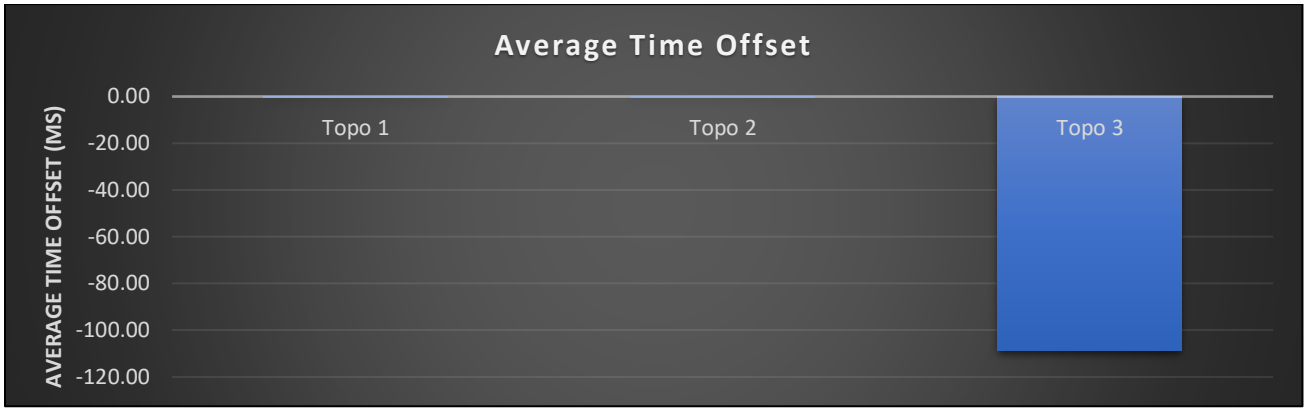






The latencies are almost the same for Topology 1 and 2, whereas for Topology 3 the latencies are significantly higher. This is expected given the fact that the nodes in Topology 3 are geographically separated.
The latencies of Topology 1 are slightly higher than that of Topology 2, which is against our intuition. The possible reason could be that Topology 1 has 4 nodes whereas Topology 2 has only 3 nodes.





The Roundtrip Delay and the Time Offset increase as we go from Topology 1 to 3. The difference in the clock skew between Topology 3 and the other topologies is significant. This is in in-line with our expectation given the physical separation in Topology 3.