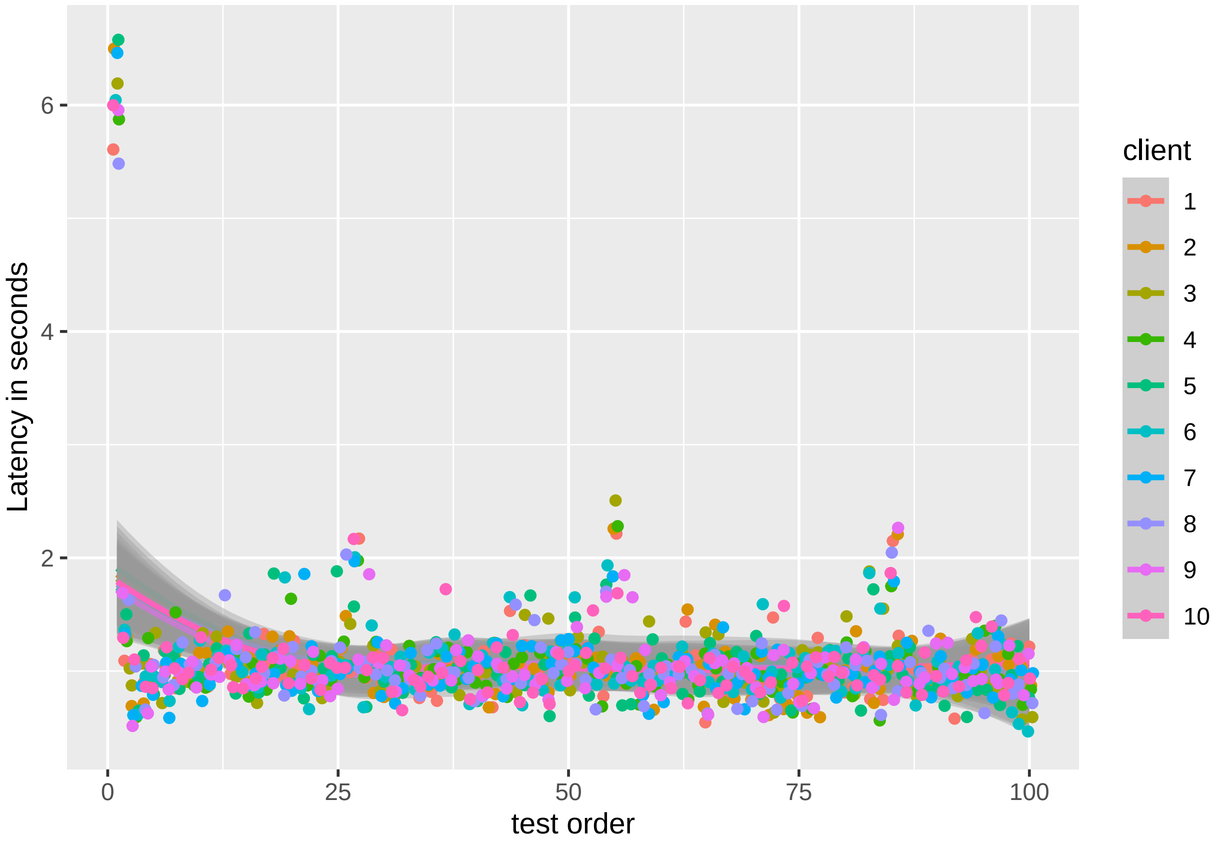
Snowflake performance

Inroduction

This document shows some results of the performance testing of the HR FHIR service running against `Showflake` backend. Snowflake backend was running on `X-Small` hardware at the time of the test. In order to limit our cost only a limited number of tests have been done as of May 18-21. All tests run a sequence of `GET Patient/<pid>` calls where patient ID (`pid`) was selected from a list of valid staging patient IDs at random.

10 concurrent clients running 100 request each

Here is a graph that shows change in latencies of 10 clients running concurrently against the HR FHIR server.



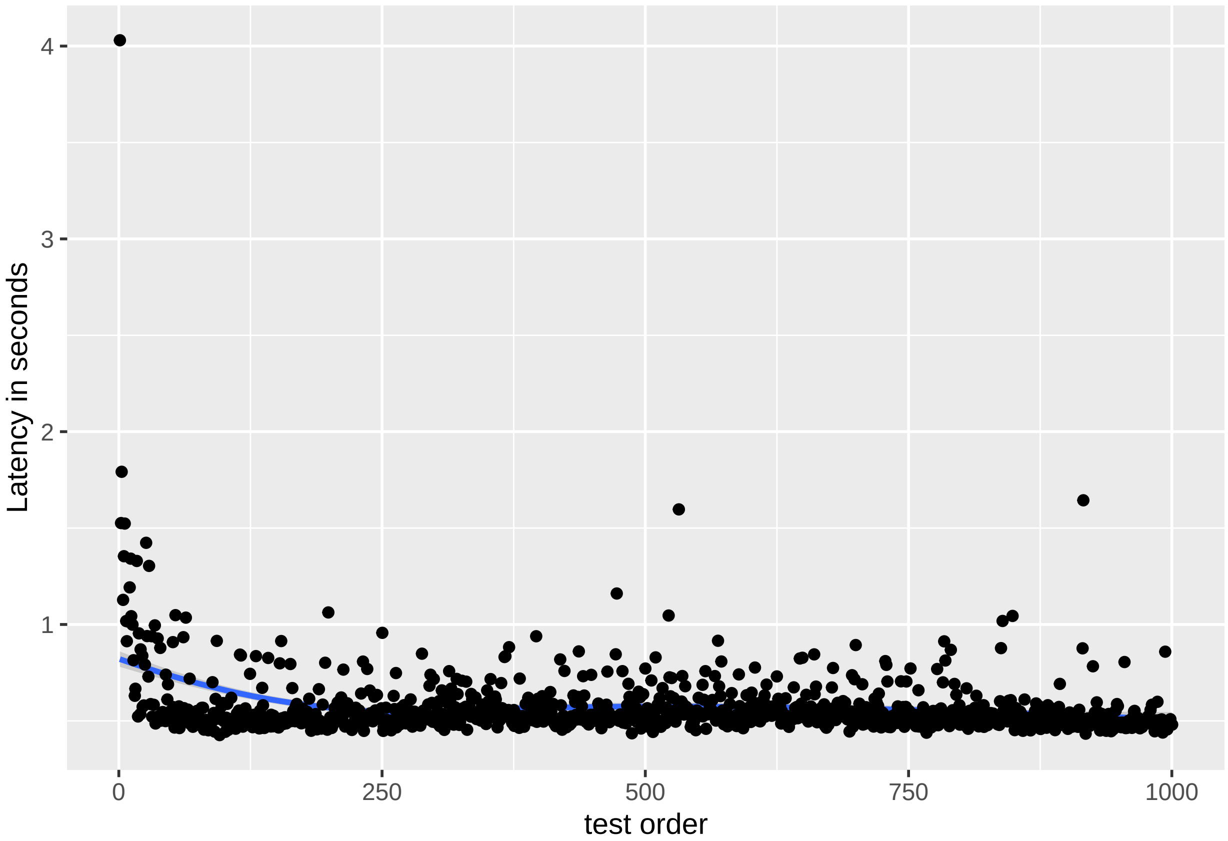
What we can see there:

- Initial calls can take multiple seconds to execute. Latencies as high as 30 seconds where observed in ad hoc tests
- After the warm up time has passed latencies hover around 1 seconds with occasional spikes in the range of a few seconds.
- At the end of the 100 tests run there was a consistent rise in latencies

If we ignore the warm up period and the slowdown at the tail, the average latencies across all 10 clients are 1.04 and median time is 0.99.

Single threaded test

Here is a graph that shows change in latencies of 1 client running against the HR FHIR server.



There are some similarities to what we see on the plot of the higher concurrency test results and some things are different:

- Initial calls can take multiple seconds to execute
- After the warm up time has passed latencies hover around 0.5 seconds with occasional spikes in the range of a few seconds.
- There is no rise in latencies at the end of the test or after 100 call execution

If we ignore the warm up period, the average latencies in this test are 0.56 and median time is 0.53.

Conclusions and TODOs

It appears that the best latencies we can get from a service running against Snowfloke current production setup can not be any better that 0.5 seconds on average. It also appears that 10 concurrent clients performance is worse with latencies going up to 1 second on average. One second average service latency does not seems acceptable. The best case of a half a second latency is barely acceptable as well.

TODO

In order to improve the service performance these are the performance improvements options that we can consider and implement. Performance tests must be re-run after the changes have been made to estimate the impact of the change.

- Use Warehouse running on more powerful hardware
- Add cacheing to the HR FHIR service
- Switch backend to GC database