1. For the first part of the homework we have set a fixed amount of threads for the wrk stress test @ 20 threads. A 7th node has been created from which all the benchmarks were made so that there would be no added latency and CPU bottlenecks on the nodes running the web-servers and on the node running the load balancer (node0). We can see that only the first scenario was able to finish the benchmark and handle 1000 connections. When we started increasing the Web-servers, the load balancer crushed at 400~500 connections and the benchmark stopped working. Although we have observed a significant increase in throughput when we increased the number of web servers, the successful requests were fewer than the actual requests since the load balancer was easily overloaded (~80% CPU usage @ 200 connections) and started dropping requests. We can also see that in terms of latency, running multiple web servers does decrease the 50th percentile latency experienced by the users but the 99th percentile is not affected much, except the case where we had 1 webserver. In that case we can see a sudden increase in latency which means that the load balancer was under heavy load and the requests were queued and then served. Another notable observation is the number of bad requests. While the number of web servers was increasing, the number of total requests was also increasing but also the number of bad requests; the requests that could not be served properly and returned HTTP error code 5xx or 4xx (HTTP Code 4xx means the site was not found, HTTP code 5xx means bad gateway). This is probably due to the high load on the load balancer and its inability to redirect the requests to the appropriate web servers. With the increase of web servers, the load balancer was easily overloaded and was hitting CPU utilization close to 100% at a low number of connections. In conclusion, having a load balancer can result at an increase in throughput, as we’ve seen, and in the decrease of latency but with a limited number of connections. But since the throughput is increased, then there will be fewer simultaneous connections on the load balancer since the requests will be served quicker meaning that in a real-world situation a load balancer would perform a lot better than in a synthetic benchmark. It should also be noted that the web servers were running under limited number of connections to better simulate queuing and request latency under heavy load.

1.2) For this part of the homework we’ve ran the hotel app monolith implementation. The scenarios stayed the same as the previous experiment with the only change in the number of connections each web server accepts since at 16 connections it was impossible to run the benchmark for more than 100 connections from wrk. The number of connections was increased to 500. The biggest improvement is seen in the throughput. As we increase the number of web servers the throughput is also increased. Latency is also improved but not by much, as connections increase the latency reaches a ceiling at around 1.5 Sec. Compared to the static site, the observed latency is larger due to the heavier website that must be loaded. Due to the increase of allowed connections to the web servers we have not observed any dropped or unserved requests.

1. For this part of the homework, we had to run the web search benchmark. There were 3 phases in this experiment, running a single index server, then 2 and lastly 4 Index servers. Running just 1 index server, we tested the web search instance using a high core count client. For each iteration of the test, the thread count of the client was doubled. We have observed an increase of the operations/sec as the threads were increasing but also an increase in latency. We can explain the increase of latency since the system (Index Server) was overloaded. As we have observed, when the clients (number of threads of the client) increased at around 16, the index server’s CPU utilization jumped @ around 60% and kept climbing as the thread count were increasing. Since the system was overloaded, the increase latency was expected.

2.2) For the second part, we have configured 2 index servers so that we could