Advanced Systems Lab Report

Autumn Semester 2018

Name: Doruk Çetin

Legi: dcetin

Student Number: 18-947-382

Grading

Section	Points
1	
2	
3	
4	
5	
6	
7	
Total	

1 System Overview

We begin our introduction by first presenting the middleware system and its components (Section 1.1) and then discussing our methodology (Section 1.2) before moving on to the experiments and their results.

1.1 System Design

This section details the inner structure of the middleware. Aim of this middleware is to collect and forward queries from clients to servers, wait for the responses and reply the clients back — all under specified conditions and configurations. Aforementioned servers and clients are the single-threaded instances of Memcached [1] and of memtier_benchmark [2], respectively. The design of the middleware spans several classes and functions that implement different functionalities. The classes operate in a multithreaded fashion so the different functionalities can work together asynchronously. Middleware continuously collect and aggregate statistics so that we can interpret and characterize different workload conditions and experimental configurations.
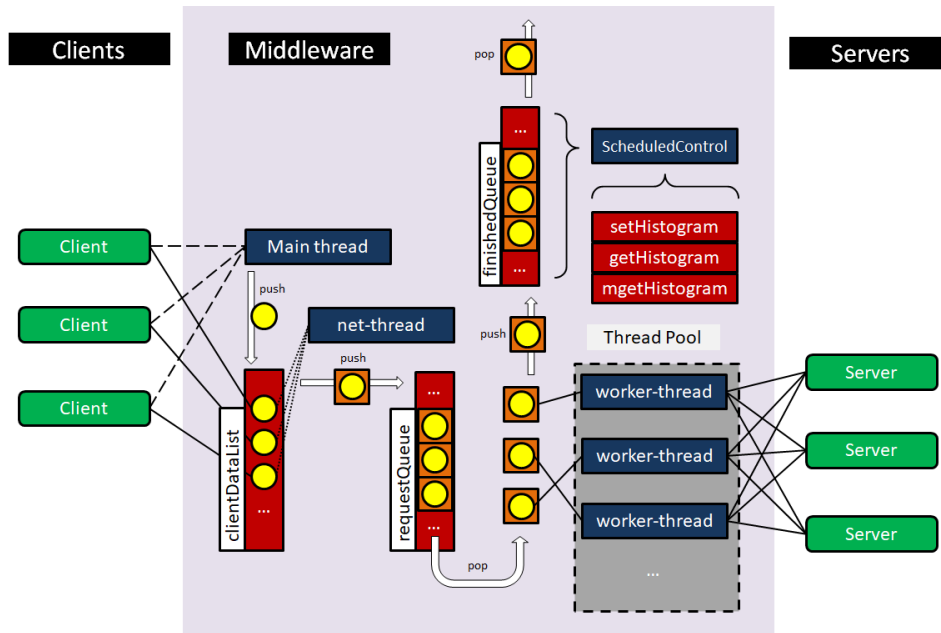

Figure 1: A simplified diagram of the system showing the inner structure of the system and the interaction of its components in an informal manner. Yellow circles represent the clientData objects and the orange squares represent the requestData objects, whose instances encapsulate the ones of the former. Dark blue rectangles represent the threads in the system, of which there are always three plus the size of the worker thread pool. Red rectangles represent the shared lists and queues which form the pipeline and help aggregation of the statistics. Green rounded rectangles illustrate the external systems, namely the clients and servers. Dashed line shows the communication with the welcomingSocket and the dotted line is net-thread checking if there is any request available. Solid lines represent the main information flow from clients to servers and again back to the clients.

1.1.1 Main thread

Main thread of the MyMiddleware class is responsible for setting up the working environment for the whole system. It first initializes the shared data structures that are to be used by other threads; namely, these are clientDataList (which stores information about each one of the connected clients), requestQueue (which is a first-in, first-out queue for queries awaiting execution) and finishedQueue (which temporarily stores the completed requests for the aggregation of their statistics). It then creates a clientHandler (so-called "net-thread", which is responsible for handling communication with clients) and the specified number of serverHandler threads (so-called "worker-threads", which are responsible with for handling communication with the servers). Throughout this report, names clientHandler and net-thread will be used interchangeably. Same also applies for worker-threads and serverHandlers. Lastly, the main thread creates the welcomingSocket and listens for incoming client connections in a loop. As soon as such a connection request is received, main thread creates the respective clientData object and pushes it into the clientDataList. welcomingSocket has a backlog queue for 512 connections, so it is expected to handle at most 512 concurrent incoming connections. This size is more than we will ever need with our experiments and provided as a safety measure.

clientData objects store the information relevant for communicating with the clients. They consist of the net socket, its respective reader and writer streams, an identifier number and lastly a flag that denotes if the client has already sent its request and awaiting for reply.

1.1.2 clientHandler (net-thread)

The clientHandler constantly iterates over the clientDataList, waiting for new requests from clients. It employs busy waiting as it non-blockingly check each client to see if there is an available request. As soon as it receives its first query from a client, clientHandler initializes the ScheduledControl thread, which is responsible for periodically aggregating statistics of the system. clientHandler checks for each client if it has not been waiting for a repy and data is available in its reader stream. That means that client has a new request so the clientHandler creates a new RequestData object and pushes it into the requestQueue for it to be handled by the worker-threads. clientHandler thread does not parse or modify the requests and directly transfer them to the worker-threads.

At the beginning of each iteration over the clientDataList, clientHandler compares the time with the timestamp of the last received query. If it is larger than some safe threshold, that means no other requests will arrive for our working conditions, so it moves on to close the system killing all threads. Before exiting clientHandler prints the response time histograms created by the ScheduledControl thread.

requestData objects contain the necessary relevant information about a request so that it can be executed correctly and its statistics can be aggregated with ease. Aside from information about its respective client, it stores an identifier number, the type of the request (set, get or multi-get), which server the request is sent to (if it is a get or multi-get request), which worker-thread handled the request, how many items were requested and received (if it is a get or multi-get request) and its timestamps.

Timestamps are all obtained in nanoseconds using Java's `System.nanoTime`. ns_netThreadReceived marks the time net-thread received the request from the client. ns_workerThreadReceived marks the time the request has exited the request queue to be handled by a worker-thread. The difference of those two timestamps give the **waiting time in the queue**. Lastly, there is ns_workerThreadFinished that marks the worker-thread received an answer from the servers (either successful or not) and replied to the querying client for the request. The difference of this last two timestamps (namely ns_workerThreadFinished and ns_workerThreadReceived)

gives the **service time of the memcached servers**. The terms queueing time and waiting time in the discussions that take place in the later sections refer respectively to waiting time in the queue and the service time of the memcached servers.

1.1.3   serverHandlers (worker-threads)

serverHandler threads initially set up dedicated connections to each server with their reader and writer streams. They also operate in an infinite loop, waiting for new request to take from the requestQueue in a blocking manner. It is blocking as a worker thread has no other jobs than handling requests. As soon as there is an available request in the queue the worker-thread wakes up, marks the time and unpacks the requestData structure. **Worker-threads are the ones that parse the messages**, so the certain specifics of the requests are unknown to the system until they exit the queue. After it has been handled accordingly to its type, the resulting requestData is pushed into the finishedQueue.

If it is a set query, the request is sent to all of the servers and reported as successful to the client only if all the responses indicate so. If not, one of the error messages are relayed to the sender client. If it is a get request there are two possibilities. If it is a multi-get request in a multi-server setting with `readSharded` option provided as true, then it is sharded into smaller requests. If the query requests a single object, there is a single server or `readSharded` is provided as false, then the request will be sent to only one server. In either case, middleware employs a simple load balancing scheme through a round-robin scheduling. It iterates over all servers while sending get requests or the sharded get requests. For both set and get requests middleware first sends the requests to servers and then collect the responses. Middleware outputs for the later experiments confirm that the workload is indeed distributed equally to multiple servers, as seen on Figure 2.
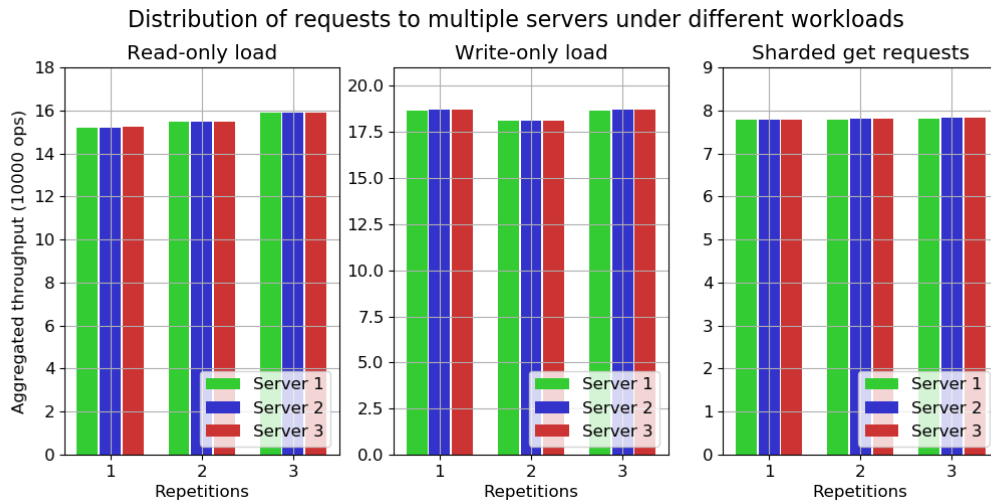

Figure 2: Distribution of the requests to three servers over three repetitions. Each of the plots illustrate a different workload condition. Results are taken from the experiments of later sections. First two plot use the results from 2K experiments 6 and the last plot use those from gets and multi gets experiments 5. Experiment with read-only and write-only loads inspect 1 middleware with 8 threads, whereas the sharded read experiment is plotted using aggreagted results for the get requests in one of the two middlewares, on 1:6 workload ratio.

1.1.4 ScheduledControl

ScheduledControl is started by the net-thread as it receives its first request. It periodically wakes up, aggregates the statistics over the most recent requests, then sleeps again. It goes over the entries in finishedQueue in its each run, collecting type-specific statistics. **For all our experiments, the aggregation period is set as 1 second**.

The thread aggregates the number of requests, queue waiting times and server waiting times separately for set, get and multi-get requests. For multi-get and get times, overall miss-rate is also aggregated together. Miss-rates are calculated over the total number of distinct get requests, breaking apart the multi-gets. For example, for two multi-gets with 5 keys if one item in one multi-get is a cache miss, the miss-rate is reported as 10%, not as 50%.

ScheduledControl also keeps track of the distribution of the response times for different types of requests by constructing a separate histogram for each type of request, namely sets, gets and multi-gets. While going over the list of requests the ScheduledControl assigns each of the requests to the appropriate histogram bin with respect to its response time and type. Lastly, length of the requestQueue is saved in each run of the ScheduledControl.

1.2 Methodology

Three repetitions were seem to be sufficient all the experiments as both our middleware and the external systems it communicated had shown performances that stayed fairly stable under repetitions. Similarly, trimming five seconds of warm-up and cool-down periods, respectively at the beginning and the end of experiments, was found as sufficient through the initial exploratory experiments performed. Clients can lag behind around at most one second per a three repetition run, but it is still negligible with the times cut for warm-up and cool-down periods. Baseline without middleware expriments (Section 2) were run for 100 seconds per each iteration, yielding a one and a half minute period after cutting the warm-up and cool-down times. Experiments involving middleware (Sections 3, 4, 5 and 6 consists of three repetitions of 70-second runs yielding a 60 second stable period for aggregation.

iPerf3 [4] is used to measure the maximum achievable bandwith in between different machines in our environment. The results can be seen in Table 1. For each machine in every experiment, we collect data through dstat [3]. dstat is run in its default mode, so it collects statistics cpu, disk, network, paging and system statistics. It collects data every second after an initial delay of one second — similar to the ScheduledControl in our middleware.

Machine	Bandwidth (Mbps)	Bandwidth (MB/s)	Effective bandwidth* (MB/s)
Client	200	25	~ 24
Middleware	800	100	~ 96
Server	100	12.5	~ 12

*: for read-only and write-only workloads

Table 1: Maximum achievable bandwidth values for all machines, obtained using iperf3. So-called "effective bandwidth" is the maximum bandwidth a machine can utilize only with the set queries or get replies. It is an informal measure that comes as useful while analyzing read-only and write-only workloads in later sections.

We illustrate the dispersion over different repetitions of individual experiments through standard variance. For every plot, error bars indicate a range of one standard variance calculated over the repetitions. At times such error bars may not be visible to the reader, indicating a fairly robust experimental setup. Most of the plots have their x-axis as the number of clients. It

is the effective number of different clients and is simply calculated by multiplying four numbers: number of virtual clients per thread, number of threads per memtier instance, number of memtier instances per virtual machine and the number of client machines. Using the abbreviations in Talbe 16 the formula could be written as $numClients = vcli \times tcli \times icli \times ncli$. The term latency is used interchangeably with the term response time. Plots specify the source of measurements as middlewares or clients in their subtitles, so it is easy to find for a plot whose perspective it reflects.

For every experiment, interactive law is verified by both calculating the throughput from the response times and vice versa. As it would approximately double all the plots we provide here to show that the interactive law indeed holds for all the experiments, we only provide the illustration of this guarantee in Section 2. Figures 3, 4, 5 and 6 show the actual and predicted values coincide perfectly. Reader should feel free to run the provided scripts for other experiments and see the interactive law in action in other experiments as well.

In experiments involving the get operations the servers are always populated beforehand so miss rates are either realistic and negligible. Server population is always done in the same manner: first 99% of the keys that are to be used in experiments were set by the sequential load option of the memtier. Therefore, for all the get requests we expect similar miss rates which are around 1%. Some actual miss rates, from the baseline with middleware experiments, can be viewed in Section 3.3.

2 Baseline without Middleware

In this section, we examine the characteristics of the clients and servers before the introduction of the middleware to the environment. For all different configurations of different experiments we show that the interactive law holds true by calculating the throughput values from the response times and vice versa. Values predicted through the interactive law are always illustrated alongside the actual values obtained by the memtier outputs.
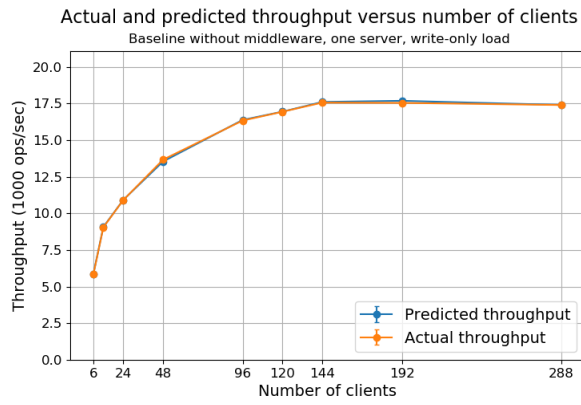
2.1 One Server

First setting contains three memtier clients connecting to one memcached server. Each of the client machines has one memtier instance with two client threads running in them. We vary the number of virtual clients and the type of workload (write-only and read-only) and observe the change in the performance of the system.

2.1.1 Explanation

We see the results for the one server configuration in Figures 3 and 4. Plots show that the interactive law holds true for both experiments.

For the write-only part, we can observe the systems starts as undersaturated and saturates around 144 clients by looking at the values of both plots. Throughput increases steadily until the saturation point, as additional clients are connected to the system. Afterwards we cannot observe even marginal increases in the throughput. We see a similar trend in the response time plot as it is slightly curved until around the saturation point but linear afterwards. After the saturation point, the systems cannot handle any more requests and the average response time increases linearly with respect to number of clients while there is no increase in throughput values as the system is already at full capacity.

The reason behind this saturation is meeting the network bandwidth capacity. dstat files show that each client can send at most 24MB per second on average. Looking at the iperf
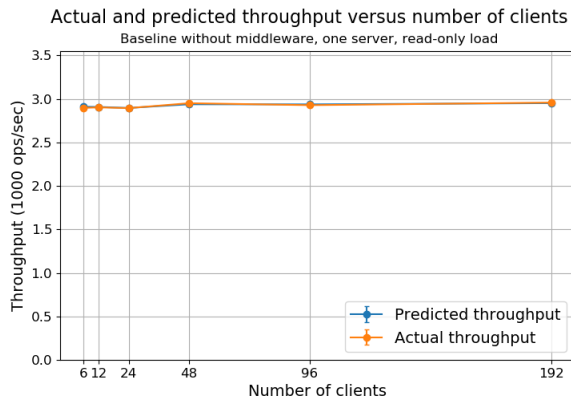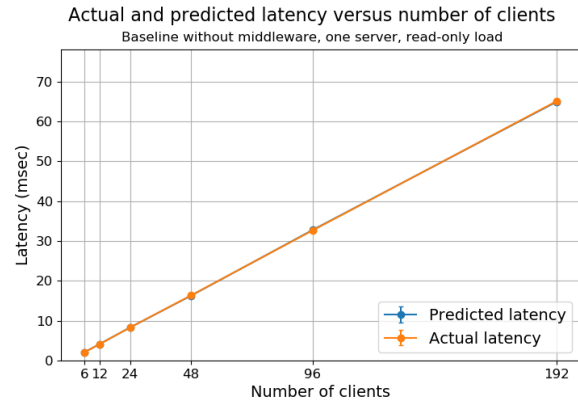
(a) Write-only throughput
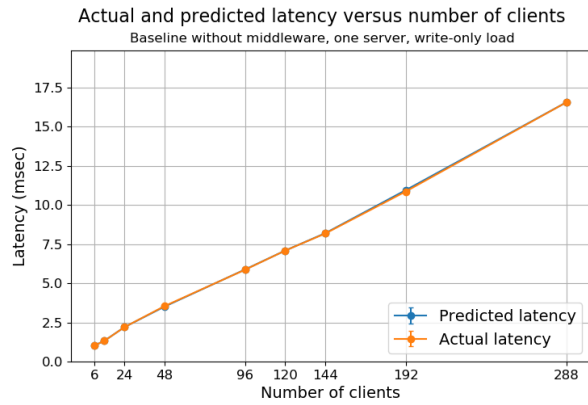
(b) Write-only latency

Figure 3: Throughput and latency values for the write-only workload for baseline without middleware, for one server case.

(a) Read-only throughput

(b) Read-only latency

Figure 4: Throughput and latency values for the read-only workload for baseline without middleware experiment with one server case.

results (Table 1) we can see that client to server communication is limited by 24MB per second for different configurations. As we send 4096B data that means we can send ~ 6000 requests per machine, ~ 18000 requests in total, and this corroborates our results.

In the read-only part, we see that the system begins operating in the overly-saturated region for the minimum number of clients we can achieve, which is 6. We conclude this as we cannot observe any noticeable change in throughput when we vary the number of clients and the response time is approximately a linear function of the number of clients everywhere. Although we can see the absolute maximum of the throughput plot is at 32 clients, it is clear that the system becomes highly inefficient with the addition of each additional client to the system after 6 clients. Only increases in the throughput are quite marginal whereas response times double when we double the number of clients in the system.

Again, dstat and iperf together confirm our assertions and indicates the reason for the saturation. iperf states the server to client connection is limited by 12MB per second (Table 1) and it is what we observe in our server for varying number of clients through all repetitions. This implies the server can reply to at most ~ 3000 requests per second and it is in line with our results.

is usage limited by correct, both here and below?

2.2 Two Servers

This configuration deals with one client machine sending requests to two servers. This time the client machine has two instances of memtier running (each connected to a different server) with one client thread per instance. We again change the workload (write-only and read-only) and change the total number of clients by varying the virtual number of clients and observe the changes in the performance.
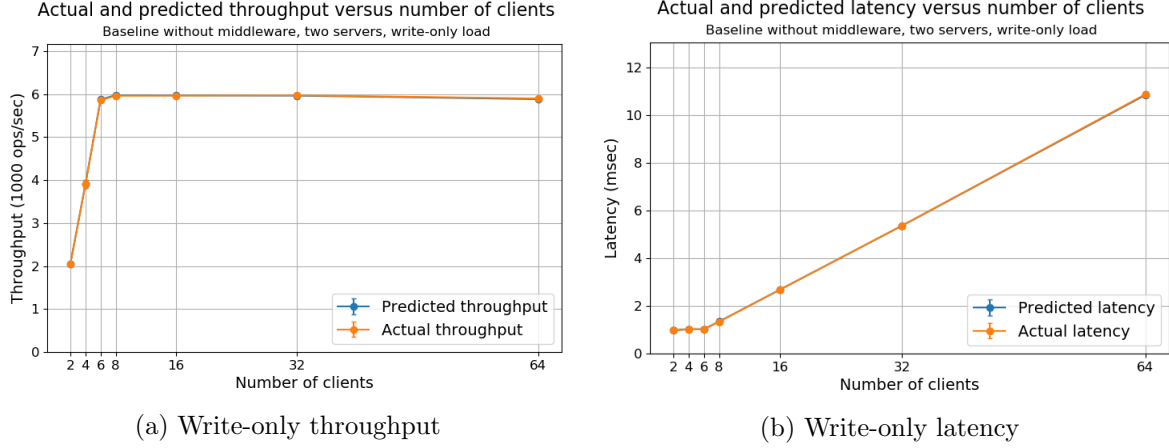

(a) Write-only throughput

(b) Write-only latency

Figure 5: Throughput and latency values for the write-only workload for baseline without middleware experiment with two servers.

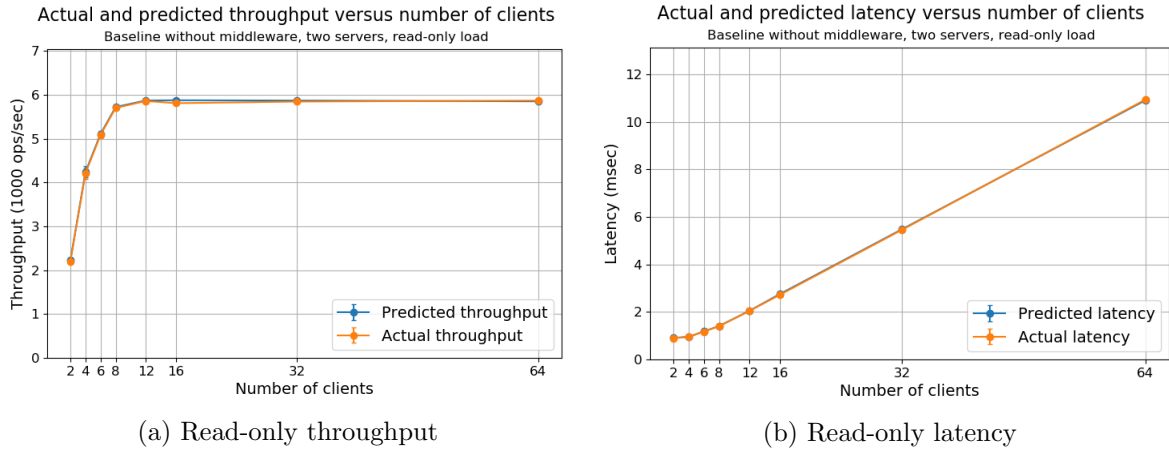(a) Read-only throughput

(b) Read-only latency

Figure 6: Throughput and latency values for the read-only workload for baseline without middleware experiment with two servers.

2.2.1 Explanation

Results are illustrated in Figures 5 and 6. Unlike the first part, we obtain similar plots for both the read-only and the write-only case. We say that the system saturates when there is 8 clients in the system for both cases, as it is elbow point for the throughput plots and the response time values continue increasing linearly after crossing 8 clients. We can clearly say that the system is under saturated before those points as the throughput rises quickly we only see marginal changes in the response times in both workloads.

The main reasons behind the saturation for both cases are the same: the network bandwith limitations. Although interestingly, the write-only and read-only cases have their bottlenecks in different parts of the network pipeline. Situtation is quite similar to the earlier results in Section 2.1. For the write-only case, dstat outputs for the saturated configurations show that the client sends at most 24MB per second to servers, which is its limit confirmed by iperf statistics (Table 1). This corresponds to a total of ~ 6000 requests distributed to two servers, corroborating our results. We also know through iperf that a server can send at most 12MB per second (Table 1) and and we see the same numbers in our dstat files for the read-only part. This equals to 24MB per second received by clients, again amounting to a total of ~ 6000 requests, this time as a result of another limiting factor. As a bottom line, both configuration saturating around 8 clients is a coincidental outcome of the bandwidth parameters since one configuration is limited by the client side whereas the other is limited by the server side.

2.3 Summary

Here, we comparatively analyse the different results for the previously discussed set of experiments. The results are explained in more detail in Sections 2.1.1 and 2.2.1, so we will not be going into the specifics of each experiment here.

Maximum throughput of different VMs.

	Read-only workload	Write-only workload	Configuration gives max. throughput
One memcached server	2897.9 ± 16.4	17556.5 ± 50.1	read-only: 6, write-only: 144 clients
One load generating VM	5698.5 ± 27.2	5960.9 ± 3.2	both workloads: 8 clients

Table 2: Summary of baseline without middleware experiments.

Table 2 reports the maximum throughput obtained in both experiments and the configurations that allow such maxima. For both experiments under either read-only or write-only workload conditions, we have concluded that the bottleneck of the system is always the network bandwidth. Intuitively enough, write-only requests have large queries and small responses as they send the data to be stored and only return with a confirmation or an error message. The situation is reversed for the read-only workload as get requests merely contain the keys but their replies carry the data itself (as we get a hit rather than a miss almost everywhere). Therefore, it is in line with our expectations that the throughput values for the experiments with write-only workload are limited by the client to server bandwidth and similarly, throughput values for the read-only counterparts are limited by the server to client bandwidth. These arguments are clearly evidenced by our results in Table 2.

Read-only throughput doubles as we introduce another server to our one server setting. Similarly for the write-only workload, by going from three load generating machines to a one client setting we can see expect the throughput to become one third of what it was before.

As a side note, we observe only the write-only workload with three client machines resulting a high percentage of CPU usage in servers, in the saturated region. The idle time of the server for this setting is under 20% most of the time. Aside from this specific case not one server or client experience any high CPU utilization, idle times are consistently above 90%.

It is also worth noting that for the read-only workload in one server setting (Section 2.1.1) we state the system is overly-saturated rather than saturated with the presence of 6 clients, even though we do not have any information regarding the fewer number of clients. It is the outcome of a comparison with the read-only experiments with two servers and one client, as we notice that with 6 clients on client machine can help achieve a throughput around 2000 requests.

Keeping in mind that the server performance is not a bottleneck anywhere in these experiments we can expect that three clients in one server hypothetically achieveing a throughput of 6000 requests whereas the actual value is only the half of it. Then we conclude that system saturates even earlier, whose effective number of clients we cannot observe with our experimental design.

3   Baseline with Middleware

In this section, we observe the changes in the overall system when the middleware is introduced to the environment, as well as we discuss the characteristics of the middleware that we can observe in the scope of this set of experiments. Interactive law for every experiment is verified using the measurements on clients. More discussion on interactive law can be found at the end of Section 3.3.

3.1   One Middleware

First experiment for this section is conducted on three client machines connected to one middleware, which in turn connected to one server. Apart from experimenting with two workloads (read-only and write-only) and varying the number of virtual clients we also change the number of worker threads in the middleware.
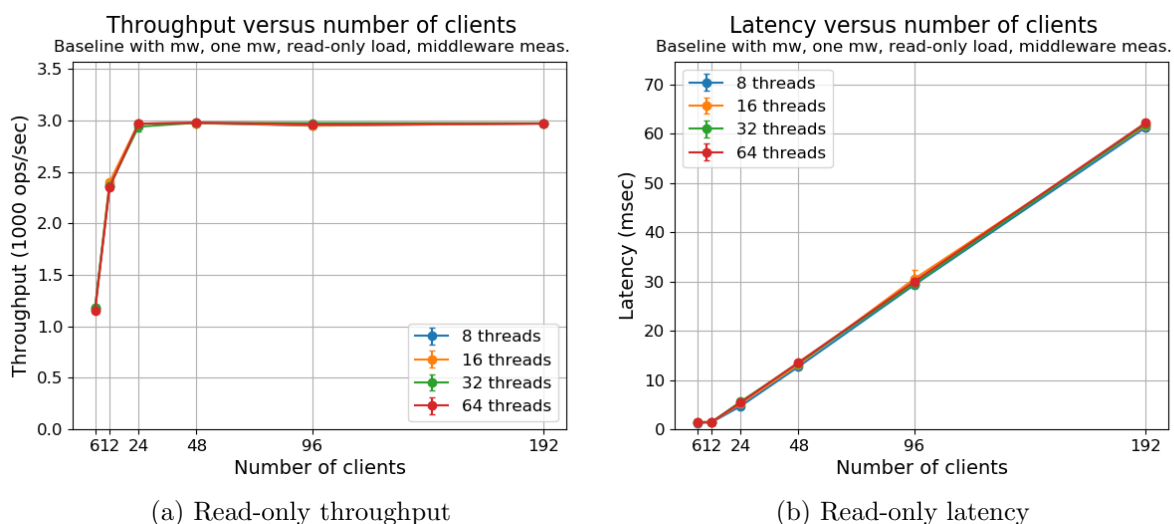

(a) Read-only throughput   (b) Read-only latency

Figure 7: Throughput and latency values for the read-only workload for baseline with middleware experiment with one middleware.

3.1.1   Explanation

Plots in Figures 7 and 9 illustrate the throughput and response time measured on the middleware, respectively for read-only and write-only loads. With the read-only load, we observe a trend quite similar to the read-only results in baselines without middleware. Until we reach around 24 clients we can observe quite an increase in the throughput which is only accompanied by relatively smaller increases in the response time, regardless of the size of the thread pool in middleware. Afterwards, the trend is again the same for different number of threads: an increase in the number of clients results in a linear increase in the response times and does not change the average throughput obtained. Therefore, we state that the system saturates when

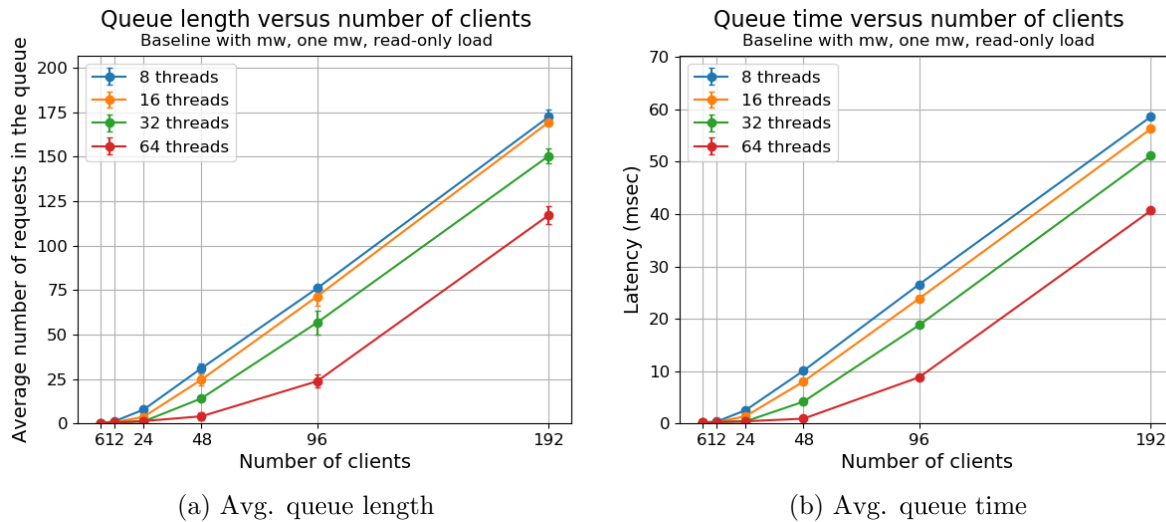(a) Avg. queue length   (b) Avg. queue time

Figure 8: Average queue lengths, queue times and waiting times for the read-only workload for baseline with middleware experiment with one middleware.

there are around 24 clients in the system, regardless of the size of the worker thread pool in the middleware.

The reason behind this saturation is familiar to us from the previous sections: server to middleware connection is limited by a bandwidth of 12MB per second (Table 1), so the server cannot send more replies because of the network limitations. This saturation happens so early in the system that we do not get to observe the effects of different threads in middleware. In other words, the middleware works efficiently even with 8 worker threads until the system saturates, so any number of worker threads could hypothetically handle more work if the server could be able respond more clients. It should also be noted that we cannot observe any network-wise limitations on top of the server-side send bottleneck and the CPU usage is fairly low for all the machines in the system (80% idle time at worst), as per dstat and iperf.

Figure 8a illustrates the average queue lengths for various configurations. As the system is network bound for this workload, the server can only reply a portion of the requests while the other thread wait to send their requests. We expect these number of waiting threads to be very small, if the number of clients are smaller than the number of middleware threads. For the range the number of clients are greater than the worker of threads every additional client should increase the average queue length linearly. The plot confirms both our intuitions as for every different number of worker threads the queue length stay close to zero until the number of clients exceed the number of worker threads and the difference in queue lengths for the linear region is proportional to the number of clients. Lastly, we can see by comparing the values in Figures 8a and 8b that the queue times and the queue lengths have an approximately linear correlation as expected. The reason behind such relation is that as the system is network bound it is not the middleware but the server side that affects the queue times in the saturated region so changing the number of worker threads does not make a difference.

In the write-only case, we can observe a trend which is completely different from the read-only case, as the network is no longer our bottleneck. The trend until reaching 24 clients is similar with the read-only case as differences in the number of worker threads do not correspond to a difference in terms of the performance of the system, except 8 thread configuration which saturates around 24 clients. This is because the system is under-saturated for 16, 32 and 64

marginal change for tmw's

this whole paragraph needs review
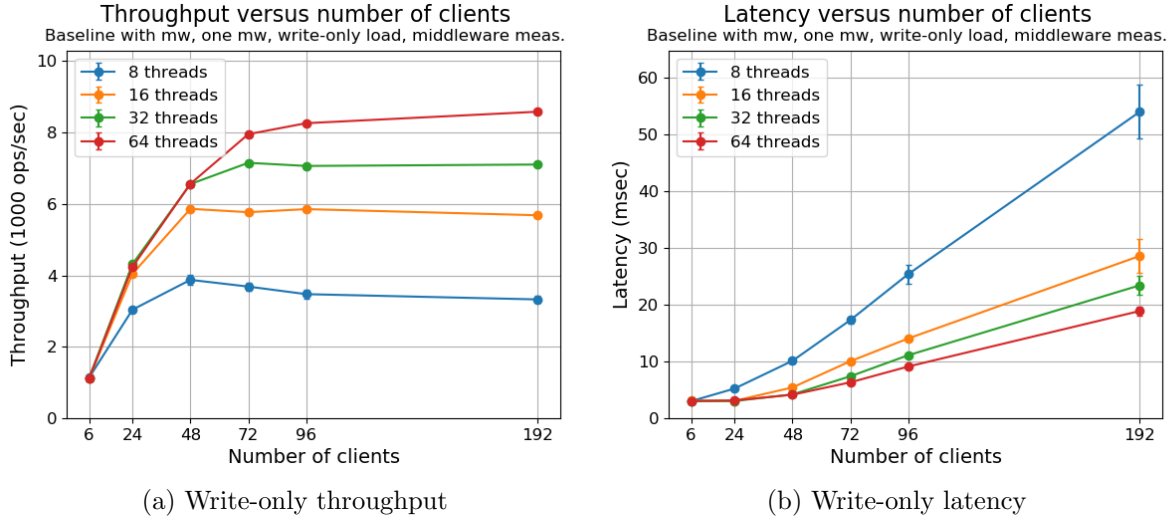
(a) Write-only throughput

(b) Write-only latency

Figure 9: Throughput and latency values for the write-only workload for baseline with middleware experiment with one middleware.

threads so we can expect their throughput and response time statistics to go hand in hand as the number of worker threads do not result in a noticeable difference in these metrics under light workload conditions. As we introduce more and more clients to the system we observe more divergence with respect to performance for different number of threads in the middleware.

We can see that there are saturation points with different values in both axes for different worker thread configurations. Due to the inherent limitations of the experiment, the number of clients were not explored in depth for us to pinpoint exact saturation points for all worker thread configurations. For example, we say that 96 clients is where the 64 thread setup saturates, but we believe a comparably fine-grained study might show that the system saturates with fewer number of clients than 96.

This time, the comparison of queue lengths and queue times (Figures 10a and 10b, respectively) yield a different interpretation. We can clearly see that, for the same queue lengths, different number of worker threads result in different scaling factors for the queue times. Moreover, queue length to queue time ratio increases with the number of worker threads. What that means is that the systems could process a longer queue for the same queueing times in average, since there will be more worker threads available to handle the requests.

After interpreting all the relevant plots we can conclude that for the write-only workload the middleware is the bottleneck. dstat measurements are also cross-checked and we do not report any network or CPU usage limit for the write-only configurations. We say that the number of worker threads is the limiting parameter for our experiments for the write-only load and more worker threads mean better system performances. As a manifestation of this interpretation, we could see that we cannot achieve the levels of throughput reached in the first part of the baseline without middleware, as the middleware introduces a considerable amount of overhead between the clients and the server.

3.2 Two Middlewares

Second experiment for this section again has an environment with three clients and one server, but two middlewares instead of one. Again, we vary the workloads, number of clients and number of worker threads in the middleware and discuss the results.
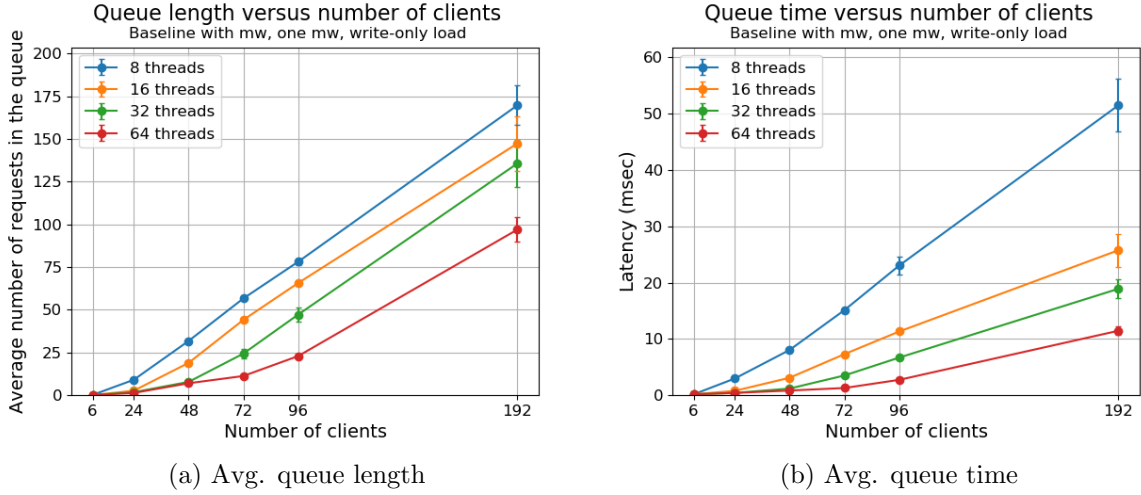
(a) Avg. queue length

(b) Avg. queue time

Figure 10: Average queue lengths, queue times and waiting times for the write-only workload for baseline with middleware experiment with one middleware.

3.2.1 Explanation

Plots in Figures 11 and 12 illustrate various statistics for the write-only load. We see results similar to the one middleware setup. System saturates at different number of clients for different number of worker threads. For the same average queue lengths, the queue times shorten as the number of worker threads increases. Again the number of working threads is our limiting factor for the system performance. Thus, we conclude the middleware being the bottleneck for the write-only case with two middlewares. Again, we find it intuitive that the statistics for the one middleware setup matches those of the two middleware setup with the half amount of worker threads, as it was the case with the read-only workload.

An intuitive correlation is as follows: statistics for the 8, 16 and 32 thread configurations for two middleware setup matches respectively with the 16, 32 and 64 thread settings for the one middleware. In other words, doubling the number of threads for one midleware has the same effects as introducing another middleware with the same number of threads on the system. Although, we can only say that for our experimental setup and configurations and this equivalence may not hold for factors much greater than two.

Results for the read-only workload is almost exactly the same as those from the setting with one middleware, as the bottleneck of the system is stays the same, which was detailed in the previous sections. Both the throughput and latency values (Figure 13) with respect to number of clients is quite similar to those of the configuration with one server, stemming from the same limitations.

Figure 14 illustrates the average queue length, queue time and waiting time (server service times) statistics for this experiment on read-only load. Here, we warn the reader to be careful when comparing the queue lengths for two and one middleware setups directly as the reported queue length is averaged over the middlewares. That is, to approximate the total number of requests waiting in the middleware queues one should multiply the averaged value by the number of middlewares in the environment. Lastly, we note the relation between queue lengths and queue times is the same as in the one middleware setup.
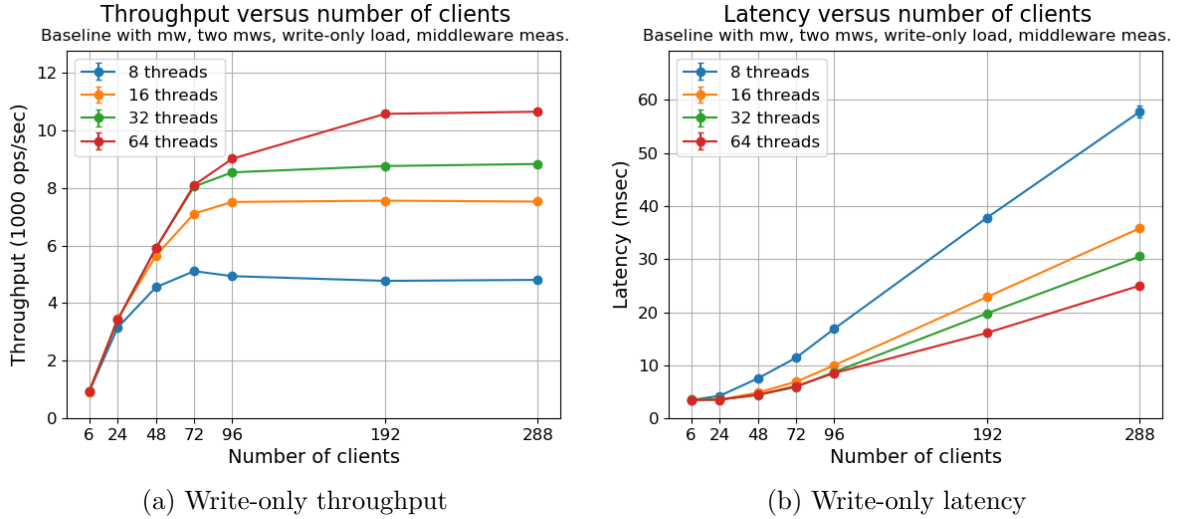
(a) Write-only throughput

(b) Write-only latency

Figure 11: Throughput and latency values for the write-only workload for baseline with middleware experiment with two middlewares.

3.3 Summary

Tables 3 and 4 reports the results for the middleware experiments, respectively the one middleware and the two middleware setups. All of the results shown in the tables are from the 64 worker thread configuration of the respective experiments.

Maximum throughput for one middleware.

	Throughput	Response time	Avg. queue time	Miss rate
Reads: Meas. on mw	2969.3 ± 1.5	5.413 ± 0.021	0.406 ± 0.011	0.0095404
Reads: Meas. on clients	2969.3 ± 1.8	8.083 ± 0.004	n/a	0.0095287
Writes: Meas. on mw	8264.1 ± 17.5	9.125 ± 0.017	2.740 ± 0.015	n/a
Writes: Meas. on clients	8263.5 ± 17.3	11.628 ± 0.026	n/a	n/a

Table 3: Summary of the baseline with one middleware.

Maximum throughput for two middlewares.

	Throughput	Response time	Avg. queue time	Miss rate
Reads: Meas. on mw	2964.3 ± 0.3	2.288 ± 0.061	0.181 ± 0.001	0.0093647
Reads: Meas. on clients	2964.4 ± 0.8	6.077 ± 0.002	n/a	0.0095260
Writes: Meas. on mw	10566.1 ± 59.6	16.121 ± 0.118	4.557 ± 0.039	n/a
Writes: Meas. on clients	10566.4 ± 59.6	18.206 ± 0.105	n/a	n/a

Table 4: Summary of the baseline with two middlewares.

First thing we can do is to compare the values aggregated by the clients with those obtained by the middlewares. Throughput and miss rate statistics show that they are approximately, but not absolutely, equal. As we aggregate data after cutting the warm-up and cool-down periods on both client and middleware outputs, this is a natural result. It is verified through additional experiments that the statistics match completely if the aggregations were to be done on the entirety of the data collected by both sides.

For response times, there is a significant but consistent difference between the averaged
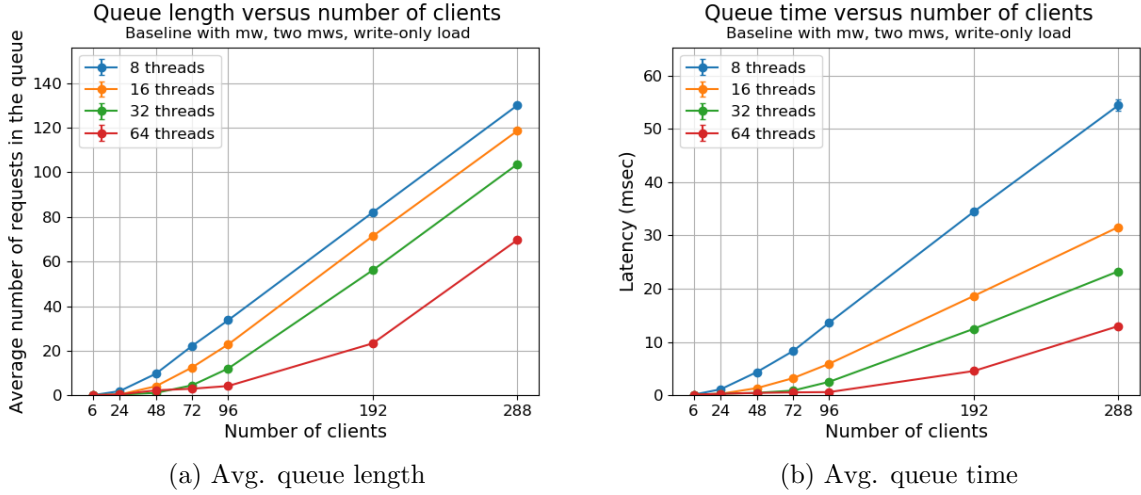
(a) Avg. queue length

(b) Avg. queue time

Figure 12: Average queue lengths, queue times and waiting times (server service times) for the write-only workload for baseline with middleware experiment with two middlewares.

values reported by the middlewares and the clients. Client response times also include the time a request requires to travel from a client to a middleware and its reply from the middleware to the client, similarly. The difference between the aforementioned response times are equal to the round-trip time (RTT) of the clients and serves for the respective experiments. We can also observe that this RTT is not always constant. It is again natural as the network characteristics may fluctuate in between (and even during) experiments and we also employ multiple clients on different physical machines, we expect not a constant difference but a RTT that samples a probability distribution.

As discussed in Section 1.2, we populate the server with the 99% of the keys we will use in the read-only experiments. This, in turn, (together with the fact that we do not allow any expiry) means that for each read-only experiment we should observe a miss rate around 1% when looked from both middleware and client perspectives. Looking at the reported miss rates we can see that the values observed is in line with our expectations and client results corroborate the middleware aggregations. The marginal difference between values reported by clients and middlewares are due to times cut, as it is the case with throughput values.

As explained in the preceding sections, maximum read-only throughput is just under 3000 requests per second, for both experiments. As the system is network bound under read-only load, changes in the number of middleware and the number of worker threads per middleware change neither where the system saturates, nor the maximum throughput achieved. Write-only experiments present us with a different picture. As the bottleneck of the system is middleware for both configurations, we see an increase in the maximum throughput obtained when we add a new middleware to the environment. This holds true not only for the 64 thread configuration which gives the best performance but for all sizes of the worker-thread pool. Looking once again at the Figures 9 and 11 we can see the saturation points for all worker-thread configurations go up with respect to clients. To illustrate, one middleware with 16 worker threads is already saturated with 48 clients but with two middlewares of 16 threads we see a continuous increase lasting between 72 and 96 clients. As discussed before, doubling the the number of middlewares has approximately the same effects on the system performance as doubling the size of the worker-thread pools in middlewares.

Lastly, we note that even with two middlewares of 64 threads we cannot reach the maximum
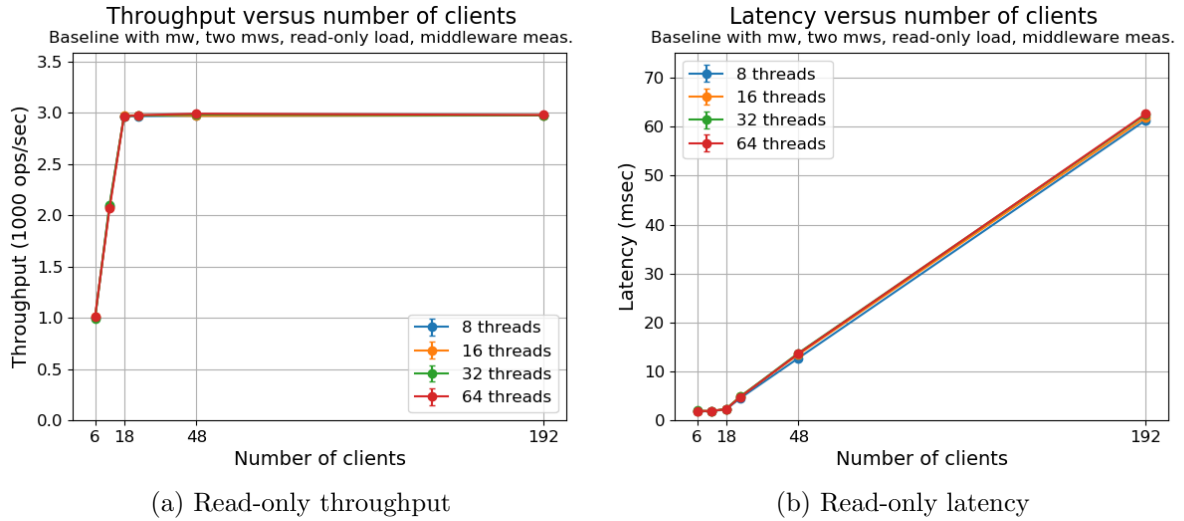
(a) Read-only throughput

(b) Read-only latency

Figure 13: Throughput and latency values for the read-only workload for baseline with middleware experiment with two middlewares.

throughput achieved without middleware in the environment. We say the parameter values explored through experiments of this section is not enough to obtain performances comparable to the ones in Section 2.1. For the same number of clients, additional network delays incurred and the overhead of the middlewares result in lower performances with the systems with middlewares.

Figure 15a provides an example illustration of the interactive law, reader should compare this plot with the actual values in Figure 13a. We only provide one exemplary plot here and the reader should feel free to use provided scripts and auxiliary plots to see the interactive law in action on other configurations that include middleware.

If we were to use the middleware measurements to predict the corresponding values, we would need to adapt the observed values. Our approach is to take the RTT between client and middleware machines as the think time of the clients when viewed from the perspective of the middleware. If we add the observed RTT times to the middleware response times for each experiment we would simply obtained the exact same results as ones with the client measurements. Instead, we also try modifying the middleware response times with an approximate RTT value for all expeirments and the results can be still satisfactory. Figure 15b illustrates such a case (actual values again in Figure 13a). Reader can see some discrepancies between actual and the predicted values, which indicates the end-to-end delays not being constant for the duration of one experimental configuration, even using the same physical setup. Quite intuitively, we cannot perfectly estimate the actual throughput and response time values from the middleware observations without knowing the network conditions for that specific experiment.

I should talk about queueing times and response times here possibly

4   Throughput for Writes

This section inspects the system with three clients, two middlewares and three servers, under write-only load. Number of worker-threads and the number of clients are varied in order to understand the characteristic of the system through different configurations.

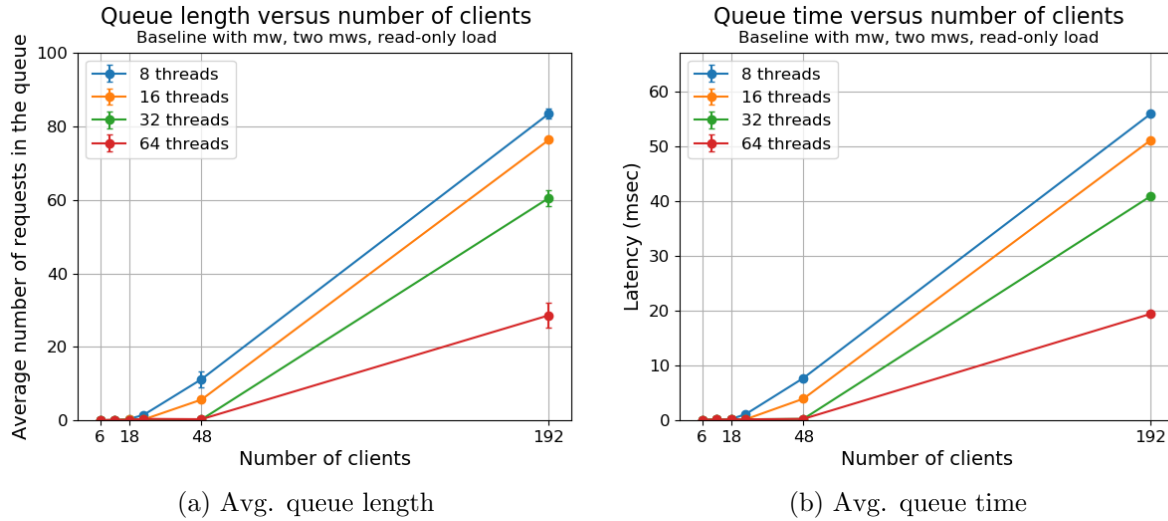(a) Avg. queue length   (b) Avg. queue time

Figure 14: Average queue lengths, queue times and waiting times (server service times) for the read-only workload for baseline with middleware experiment with two middlewares.

4.1 Full System

Figures 16a and 16b illustrate the average throughputs and response times of the system under different configurations. Each line represents a setup with a different number of threads for both plots. Results are similar to those analysed in Section 3.2, as the setup stays the same except the increase in the number of servers the system has. We see each configuration saturates at different points, since the middleware is the bottleneck of the system. Each of them conforms to the same trend: throughput gains become minimal and response times start to form a linear pattern as we move closer to the number of clients which saturates a given configuration.

For 8, 16, 32 and 64 worker-threads we state the saturation points are around 24, 36, 72 and 144 clients. With the exception of the 8 thread setup, we can see that the number of clients a saturated system can handle doubles as we double the worker-threads in the middleware. It is in line with our expectations as this ratio of number of worker-threads to the number of clients at the point of saturation should hold true for different number of worker threads with same configuration. Since middlewares are the bottlenecks of this setup, we except there should be enough requests in the systems waiting to be handled. This, in turn, means an increase in the number of worker-threads should amount to an approximately linear increase in the average number of requests that can be handled for a given period, which is indeed corroborated by the aforementioned results. Reader should note that the 8 worker thread setting has not the same level of granularity as the other settings and experiences a much sharper increase before leveling off. This points to the need for a finer grained search in the 6-24 client range, as the saturation of the 8 thread configuration is believed to be expected earlier than 24 clients, much likely around 18 clients considering the pattern described above.

Figures 17a and 17b respectively illustrate the average queue length and queueing time statistics for the throughput for writes experiment. As the experimental setup of this section is quite similar to the one in baseline experiments with two middlewares in Section 3.2, it is of no surprise that we obtain similar trends related to request queue. We can verify the saturation points for the different thread configurations as the points where lines in both plots become completely linear. We can again observe that corresponding lines in queue length and queue time plots have different slopes. For the system to reach a chosen queue length, we would need more clients with configurations with more threads. What that indicates is an increase in the
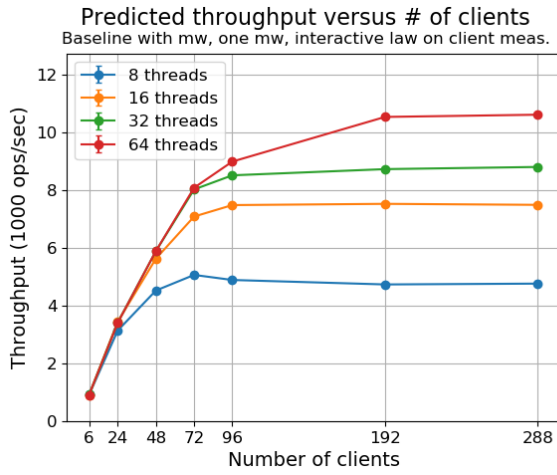
(a) Predictions on client measurements
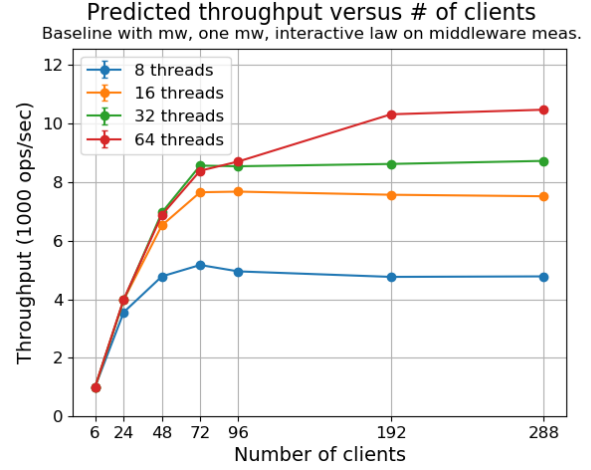
(b) Predictions on adapted middleware measurements

Figure 15: Application of the interactive law illustrated for predcitions made using client and midleware measurements. Each response time observed in middleware is increased by a constant amount to simulate the effects of RTT between clients and the middlewares. Results not being perfect indicate the RTT values are changing during experiments. These results show that RTT is not a constant value as in the ideal case, but it is a random variable.

number of threads in the middleware increases performance by reducing the queueing times, so the number of middleware threads is our limiting factor for this experiment. _____ waiting times

4.1.1 Explanation

4.2 Summary

see how the predicted values do not really match the acutal values

Maximum throughput for the full system				
	WT=8	WT=16	WT=32	WT=64
Throughput (Middleware)	3492.9 ± 49.1	4673.6 ± 23.5	6411.0 ± 5.0	8161.1 ± 24.9
Throughput (Derived from MW response time)	3459.6 ± 205.4	4698.3 ± 40.1	6246.1 ± 5.8	8124.1 ± 27.3
Throughput (Client)	3492.8 ± 49.1	4674.3 ± 23.4	6412.2 ± 6.2	8161.5 ± 25.2
Average time in queue	1.112 ± 0.062	0.622 ± 0.006	1.148 ± 0.010	1.779 ± 0.017
Average length of queue	2.0 ± 0.0	1.4 ± 0.1	4.0 ± 0.2	8.6 ± 1.2
Average time waiting for memcached	3.9 ± 0.4	5.0 ± 0.1	8.4 ± 0.0	13.9 ± 0.0

Table 5: Summary of the throughput for writes experiments.

5 Gets and Multi-gets

This experiment again deals with the full sytem having 3 clients, 2 middlewares and 3 servers. Each of the two middleware has 8 worker threads. Clients sen multi-get requests together with set requests and set to multi-get ratio for each experiment is $1 : X$ where X is the number of keys

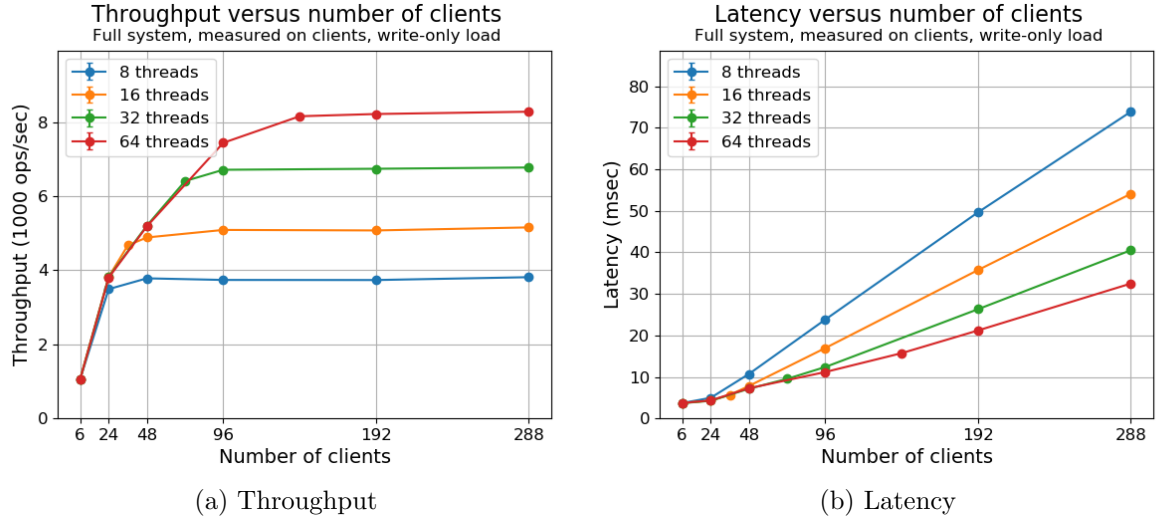|	
(a) Throughput	(b) Latency

Figure 16: Throughput and latency values for the throughput for writes experiment, write-only workload.

in the multi-get requests. We change this key size and see how the system reacts. A comparative analysis is made on effects of the sharded and non-sharded aproaches with multi-get handling.

5.1 Sharded Case

In sharded case we split the multi-get requests send all splits to respective servers and proceed to listen for replies in the order we sent the requests.

5.1.1 Explanation

5.2 Non-sharded Case

5.2.1 Explanation

5.3 Histogram

5.4 Summary

6 2K Analysis

In this section, we perform a 2^k factorial analysis on three parameters: number of middlewares, number of servers and the number of worker threads per middleware. We believe the effetcs of these parameters are unidirectional, so 2^k factorial analysis suits our interests best. We solve the linear equation systems for both throughput and response times on both read-only and write-only workloads. Our aim is to assess the impacts of different parameters for different configurations, as well as further investigating and whether an additive or a multiplicative model would capture the interactions and variations in our system best.

Table 6 shows the system parameters and their correspoing variables and values chosen for the analyses. For each configurationon each workload, we solve two linear systems, which are illustrated in Equations 1 and 2. The former uses the assumption that the effects of the factors are additive, whereas the latter one assumes multiplicative effects. Equation for the multiplicative model is only different from the additive one in the sense that the take the response variables as logarithms of the observed values. In the scope of this section e denotes
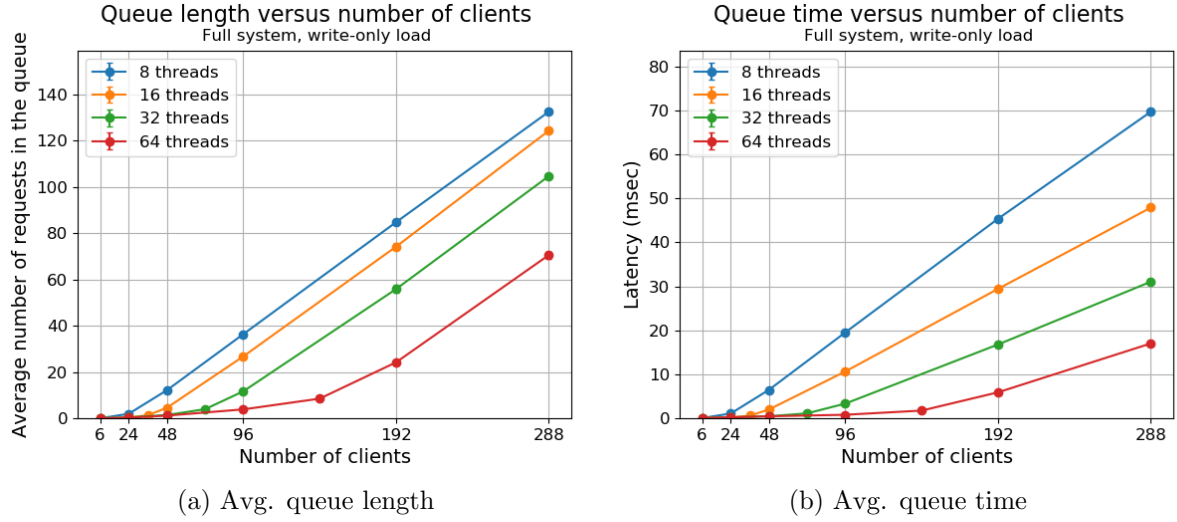
(a) Avg. queue length

(b) Avg. queue time

Figure 17: Average queue lengths, queue times and waiting times (server service times) for the throughput for writes experiment, write-only workload.

Parameter	Factor	Variable	Value at min. (-1)	Value at max. (+1)
Number of middlewares	A	xA	1	2
Number of servers	B	xB	1	3
Worker threads per middleware	C	xC	8	32

Table 6: 2^k experimental design used in this section, specifying the factors and treatments.

error whereas r denotes the number of repetitions.

$$y = q0 + qA*xA + qB*xB + qC*xC + qAB*xA*xB + qBC*xB*xC + qAC*xA*xC + qABC*xA*xB*xC + e \quad (1)$$

$$log(y) = q0 + qA*xA + qB*xB + qC*xC + qAB*xA*xB + qBC*xB*xC + qAC*xA*xC + qABC*xA*xB*xC + e \quad (2)$$

Before going into the results of the analysis, it is good to know how the linear equations are formed and solved. Table 7 shows an example of such equation systems, in this case the response variable being the throughput of the system on write-only load. Left part of the table is a sign table where -1's refer to the minimum treatment and $+1$'s refer to the maxiumum treatment for a variable. We always condust our factorial analyses with replications (so-called $2^k r$ design) so we solve our equations for the mean response (y_mean) and estimate the errors in response values as e1, e2 and e3 over three repetitions. After solving the equation system we obtain the effects of the variables (e.g. qA for variable xA or qBC for variable interaction xB*xC) and calculate allocation of variation for each factor through sum of squares values. Equations 3, 4 and 5 show how these calculations were done. The SSE accounts for the unexplained variation attributed to the experimental errors.

$$SSA = 2^2 qA^2, \text{ Fraction of variation explained (\%) by factor A} = SSA/SST \quad (3)$$

$$\text{Sum of Squares for Error: } SSE = \sum_{i=1}^{2^k} \sum_{j=1}^{r} e_{ij}^2, \text{ Fraction of unexplained variation} = SSE/SST \quad (4)$$

$$\text{Total Sum of Squares: } SST = SSA + SSB + SSC + SSAB + SSBC + SSAC + SSABC + SSE \quad (5)$$

Our principle reference for this chapter is Raj Jain's book [5] on system performace analysis. It is argued that the "knowledge about the system behavior should always take precedence over
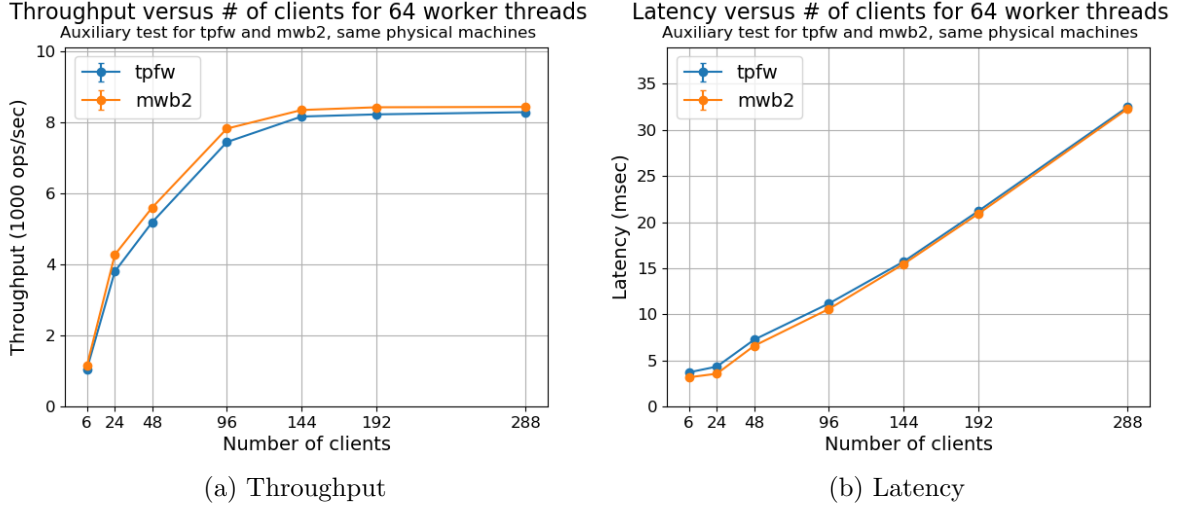
(a) Throughput

(b) Latency

Figure 18: Throughput and latency values for an auxiliary experiment, write-only workload with 64 worker threads on middlewares. tpfw stands for the throughput for writes experiment, whereas the mwb2 denotes the middleware baseline experiment with two baselines.

I	xA	xB	xC	xA*xB	xB*xC	xA*xC	xA*xB*xC	y_mean	e1	e2	e3
1	−1	−1	−1	+1	+1	+1	−1	3364.3	45.56	-15.33	-30.23
1	+1	−1	−1	−1	+1	−1	+1	4318.9	-19.85	-1.77	21.63
1	−1	+1	−1	−1	−1	+1	+1	2655.2	23.47	37.06	-60.54
1	+1	+1	−1	+1	−1	−1	−1	3729.6	9.09	-39.41	30.32
1	−1	−1	+1	+1	−1	−1	+1	5832.8	6.49	-17.43	10.93
1	+1	−1	+1	−1	−1	+1	−1	7321.3	17.68	-21.73	4.04
1	−1	+1	+1	−1	+1	−1	−1	4786.4	27.09	-18.71	-8.38
1	+1	+1	+1	+1	+1	+1	+1	6782.0	-16.40	-2.95	19.35

Table 7: An example system of equations for the 2^k factorial analyses. Specifically, this set of equations with the corresponding y values model the throughput for the write-only load, assuming additive variable effects. Variable xA, xB and xC denote respectively the effects of number of middlewares, number of servers and the number of worker threads in our model.

statistical considerations". As we are already familiar with the system through the different experimental setups in the previous sections, we do not view the system as a black box and rather choose the models that suits our parameter interactions best. For write-only workload we expect the factors xA (number of middlewares) and xC (number of threads per middleware) to be the important parameters, since our limiting factor is the total number of middleware threads in the system. As the total number of threads is the multiplication of these two factors and we expect the variation caused by the number of servers and the errors to be small, we choose multiplicative model for the write-only load. Similarly, with the read-only load we expect only the facor xB (number of servers) to be effective on the final results, so we choose the additive model as it also assumes the errors of the system are additive, not multiplicative.

Although we choose the models that fit the inner workings of our system, it is a good practice to model the system using both additive and multiplicative models (for both workloads and response variables) to see if there are any critical differences between two approaches. Respective results for the not chosen models are presented in the appendix part of this report, in Tables 12, 13, 14 and 15. We see that the differences between additive and multiplicative models for

(a) Latency

(b) Server waiting times

Figure 19: Throughput and response times of sharded and non-sharded reads in different key size experiments

all our analyses were not quite significant as both models essentially indicate the same results. There are several reasons behind this outcome. System is robust across repetitions and the residuals are low, thus modeling the errors one way or another does not affect the unexplained variation greatly. As it is also indicated in the book, "the logarithmic transformation is useful only if the ratio y_{max}/y_{min} is large. For a small range the log function is almost linear, and so the analysis with the multiplicative model will produce results similar to that with the additive model", which is exactly our case. Going from the additive to multiplicative model does not create a big difference in terms of the modeling power.

Results of the additive model on read-only load

Source	Mean Estimate	Variation Expl. (%)	Source	Mean Estimate	Variation Expl. (%)
qA	265.8600	0.009	qA	−0.0009	0.002
qB	2678.2889	0.944	qB	−0.0208	0.988
qC	263.8299	0.009	qC	−0.0009	0.002
qAB	262.0704	0.009	qAB	−0.0009	0.002
qBC	260.5430	0.009	qBC	−0.0009	0.002
qAC	−257.6810	0.009	qAC	0.0009	0.002
qABC	−260.9624	0.009	qABC	0.0009	0.002

Table 8: Throughput analysis   Table 9: Latency analysis

Now we turn our attention to the findings of the models. We present the mean estimate of the parameters and the percentage of variation in the response variable explained by the variation of those parameters for analyses on throughput and latency over read-only and write-only workloads. Tables 8 and 9 show the results for the read-only load. Clearly, the only important factor is the number of servers in the system while the other factors and factor interactions are equally unimportant. Analyses on both throughput and latency corroborate each other: number of servers have a stong positive influence on the throughput and a strong negative influence on the response time values. As we have discussed this issue on Secions 2 and 3, network bandwidth is the bottleneck for the read-only load most of the time. Our choice
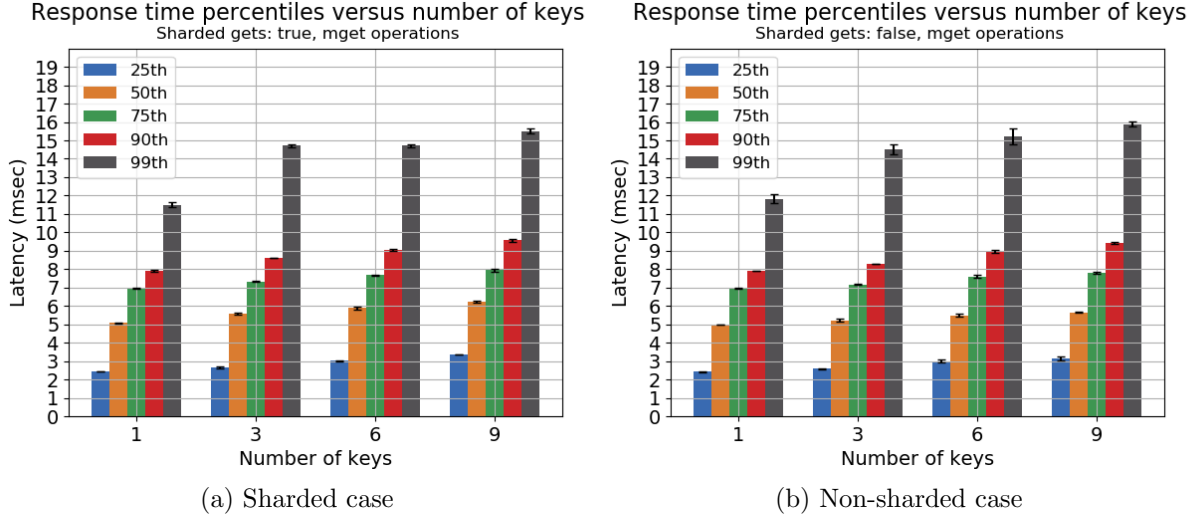
(a) Sharded case   (b) Non-sharded case

Figure 20: 25th, 50th, 75th, 90th and 99th percentiles of sharded and non-sharded reads, for 1, 3, 6 and 9 key cases. Instead of their respective sections each, plots are provided side by side together to help the reader when comparing two cases.

of minimum and maximum treatments for each factor (Table 6) results in number of servers being the bottleneck in every experiment except the 8 thread total with 3 servers case. This is in line with our expectations and findings in previous sections. Percentage of unexplained variation due to errors amount to 0.002 and 0.001, respectively for throughput and response time analyses. Therefore, we rule out the effects of sources other than the number of servers, including the error, as their variation does not really change values of the response variable.

Results of the multiplicative model on write-only load

Source	Mean Estimate	Variation Expl. (%)	Source	Mean Estimate	Variation Expl. (%)
qA	0.1457	0.194	qA	−0.1384	0.165
qB	−0.0822	0.062	qB	0.0963	0.080
qC	0.2832	0.733	qC	−0.2730	0.642
qAB	0.0264	0.006	qAB	−0.0192	0.003
qBC	0.0137	0.002	qBC	−0.0031	0.000
qAC	−0.0017	0.000	qAC	0.0124	0.001
qABC	0.0039	0.000	qABC	0.0065	0.000

Table 10: Throughput analysis   Table 11: Latency analysis

For the write-only workload we expect the total number of worker threads to have a strong effect on system performance such that more worker threads should correspond to higher throughphut and lower response time values. Similarly, we believe the number of servers for write-only workloads should hinder the system performance, perhaps not as noticeably as with the total number of threads. We first send the queries to all servers and then wait for their responses. That way we do not experience the RTT for all the servers, although the time a worker thread waits increases linearly with the number of servers. In the optimistic case, reply from the first server would arrive as we finish sending the query to last server.

In Tables 10 and 11 we can see the factorial analysis results for the write-only workload. Number of middlewares and the number of threads per middleware together account for the vast majority of the variation for both throughput and latency as expected. The reason that the

(a) Sharded case, client measurements   (b) Sharded case, middleware measurements

(c) Non-sharded case, client measurements   (d) Non-sharded case, middleware measurements
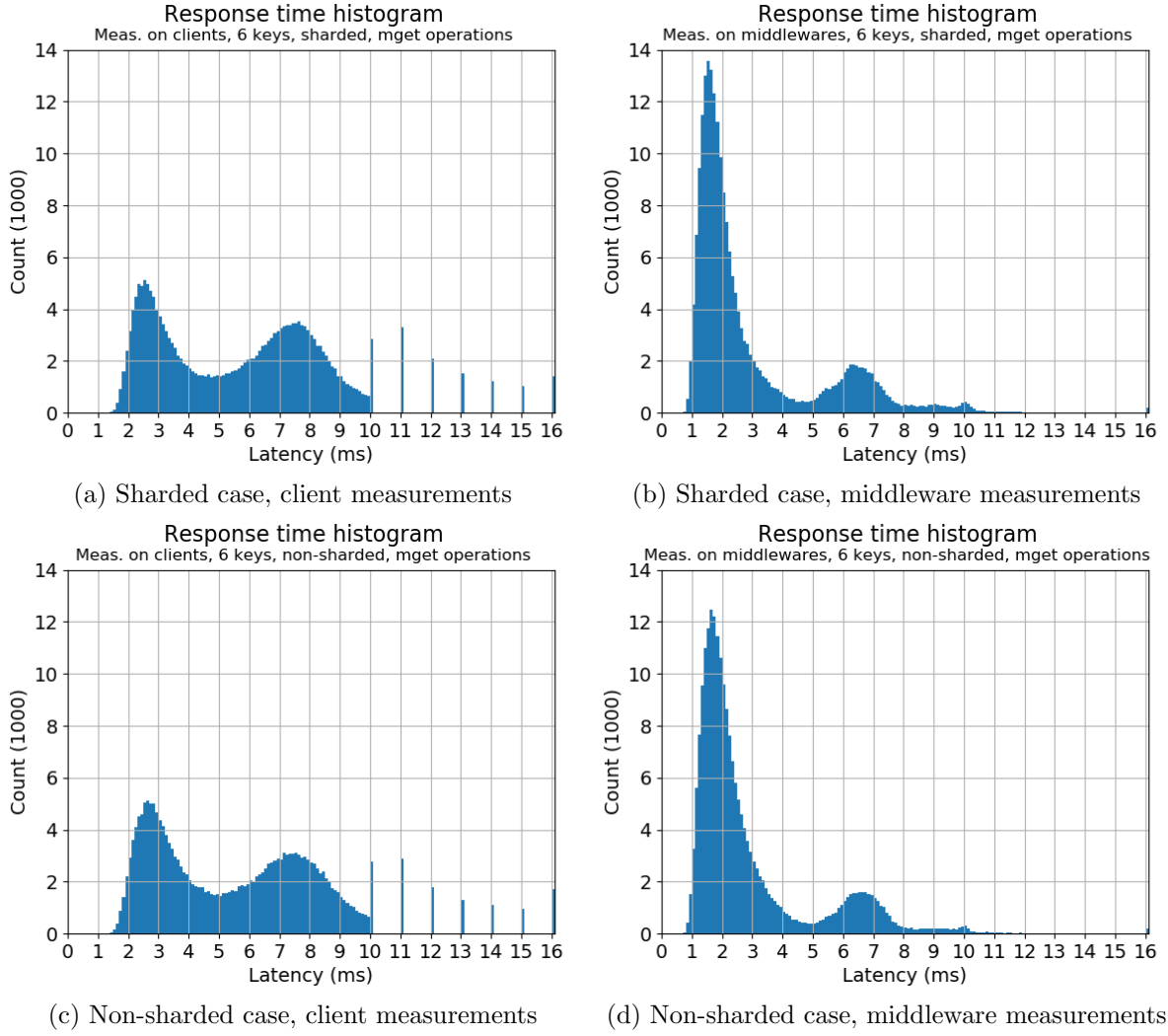
Figure 21: Response time distribution histograms for 6 key case. Bucket size is chosen as $100\mu s$ for all histograms, except for their last bins which contains the number of outlier response times that are greater than $16ms$. When choosing this cut-off value, it is ensured that less than 1% of the occurences are collected in the last bin for each individual result used in aggreagation.

variation due to number of threads per middleware (xC) is greater than that of the number of middlewares (xA) should not be surprising as the min-max values for xC covers a greater change than the xA, with respect to total number of threads (Table 6). We can also see xB having a slightly negative influence on throughput (vice versa for response times) since the number of servers only affect the performance marginally. Unexplained variations (0.003 and 0.108 for throughput and response times, respectively) for the write-only load is greater than those for the read-only load, since the system conditions (e.g. end-to-end delays between physical machine pairs) have an effect on the performance of the system in the observed value ranges. It was not the case with the read-only workload as the network bandwidth limited the performance such that changing the number of threads in our range did not change the bottleneck being the bandwidth.

7 Queuing Model

7.1 M/M/1

7.2 M/M/m

7.3 Network of Queues

You need to consider both 1 and 2 middleware configuration. You need to analyze both read only and write only workloads. You can set the number of worker threads to the constant that delivers highest throughput.

(a) Response times

(b) Server waiting times

(c) Sharded case
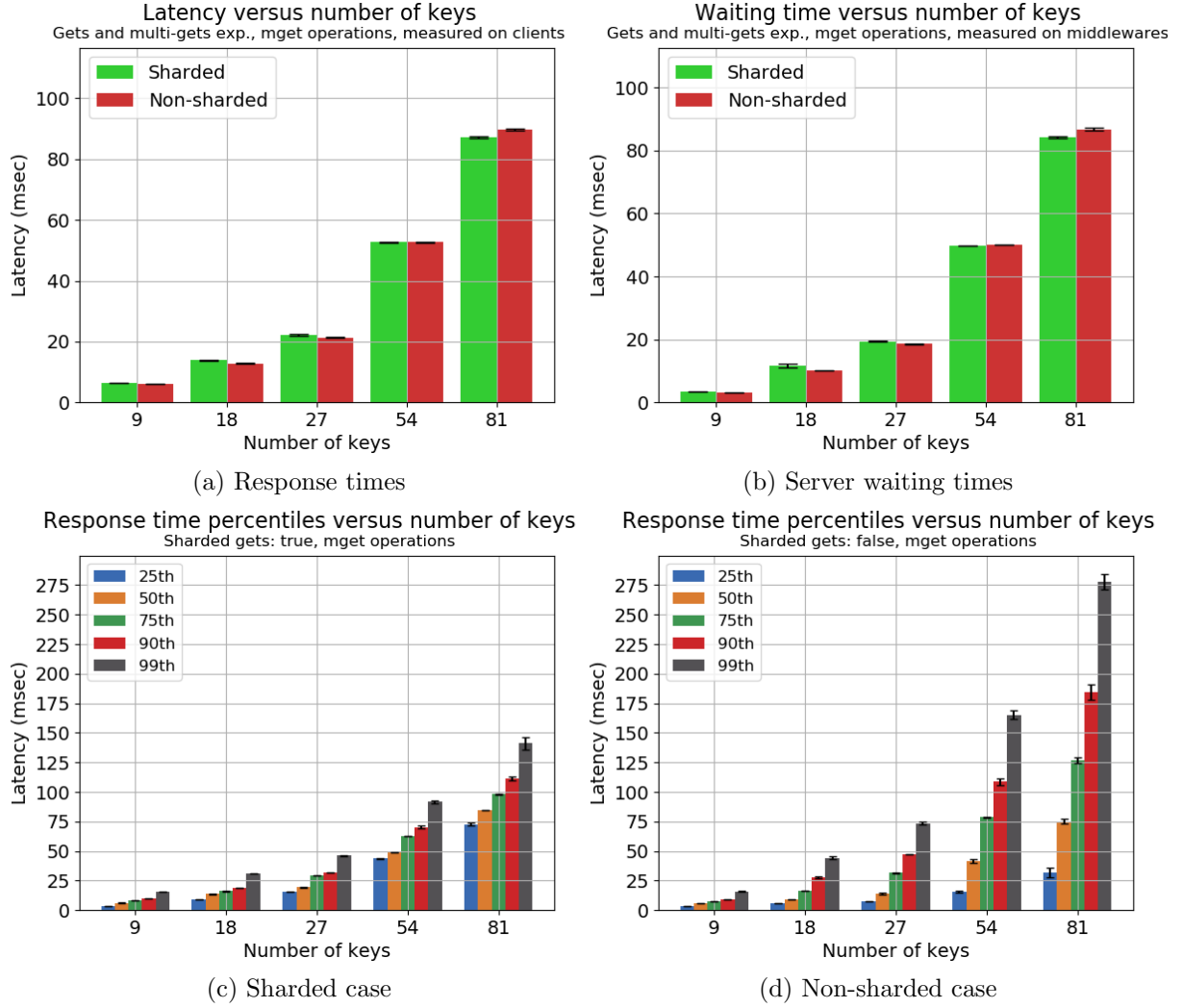
(d) Non-sharded case

Figure 22: An auxiliary experiment for comparing sharded and non-sharded performances on multi-get requests with bigger key sets. Figures 22a and 22b compare the average response time and server waiting time statistics for the sharded and non-sharded reads with 9, 18, 27, 54 and 81 key cases. Figures 22c and 22d illustrates 25th, 50th, 75th, 90th and 99th percentiles of sharded and non-sharded reads, for the experiments with aforementioned set of key sizes.

Appendix

2K Analysis: additional tables

Results of the additive model on write-only load

Source	Mean Estimate	Variation Expl. (%)
qA	689.1529	0.196
qB	−360.4905	0.054
qC	1331.8182	0.731
qAB	78.3606	0.003
qBC	−35.9031	0.001
qAC	181.8770	0.014
qABC	48.4125	0.001

Source	Mean Estimate	Variation Expl. (%)
qA	−0.0064	0.175
qB	0.0045	0.087
qC	−0.0120	0.620
qAB	−0.0015	0.010
qBC	−0.0013	0.008
qAC	0.0023	0.022
qABC	0.0008	0.003

Table 12: Throughput analysis   Table 13: Latency analysis

Results of the multiplicative model on read-only load

Source	Mean Estimate	Variation Expl. (%)
qA	0.0348	0.004
qB	0.5144	0.974
qC	0.0345	0.004
qAB	0.0335	0.004
qBC	0.0334	0.004
qAC	−0.0330	0.004
qABC	−0.0341	0.004

Source	Mean Estimate	Variation Expl. (%)
qA	−0.0351	0.005
qB	−0.5130	0.970
qC	−0.0353	0.005
qAB	−0.0348	0.004
qBC	−0.0346	0.004
qAC	0.0348	0.004
qABC	0.0356	0.005

Table 14: Throughput analysis   Table 15: Latency analysis

Running the code

Number of servers	nsvr		Multi-Get behavior	mgshrd
Number of client machines	ncli		Multi-Get size	mgsize
Instances of memtier per machine	icli		Number of middlewares	nmw
Threads per memtier instance	tcli		Worker threads per middleware	tmw
Virtual clients per thread	vcli		Repetitions	reps
Workload	wrkld		Test time	ttime

Table 16: Abbreviations for experiment parameters.

All the necessary commands for initializing the experimental setup and conducting the experiments are provided in a companion file called `commands.sh`. It also exemplifies the usage of auxiliary technologies used for further analysing the system and its environment as well as it contains some other helpful commands for copying and plotting the experiment results. `plot` folder contains all scripts (and their eventual outputs) necessary for plotting the results placed under the `res` folder. Every plot provided inside or alongside this report is reproducible with the provided data and the scripts. Lastly, `runner.sh` is the encapsulating script that runs the desired agent (i.e. client, middleware, server or dstat tool) for the configuration provided through the command line arguments. The arguments it expects share the same

Probably should detail which scripts corre- spond to which experi- ments

nomenclature we use in scripts to refer the experiment parameters. Here, in Table 16, we list these abbreviations. Result directory hierarchy also obeys the same nomenclature.

It is important to pay attention the order and timing of executing commands. Servers should be run first and the middlewares (if any) should start waiting connections. Then, both the clients and dstat processes should start simultaneously — we could do this by using Cluster SSH (also called as cssh) [6]. Clients, middlewares and dstat processes ran accordingly for three separate repetitions, afterwards they should be restarted by hand. Another important thing to keep in mind while using the project codebase is to remember that the codes do not cover erroneous cases. For example, middleware does not support any other operations than set or get. Similarly, plotting scripts may not work with data with erroneous content, filenames, directory hierarchy and such.

Internal parameters

The system has several options hardcoded as parameters of MyMiddleware class:

- `verboseLogs` specifies if the system should output the work logs or not. Such logs are printed when a request is received, relayed to servers, got answered and replied back to the clients. It also prints out details about which thread handled the request, which client sent the request, to which server the request was sent to and the completely parsed messages and replies. It is a useful option for understanding the inner workings and state transitions of the middleware system. It should be set as false if user is not debugging.

- `verboseAggr` controls the outputting of the aggregated statistics. It is always true for our experiments as it is our main tool for analysing the system.

- `timeoutSecs` specifies the amoun of time the system should wait until new queries before shutting itself down. Altough the middleware can work seamlessly without any interruption, systems needs to be restarted to operate in its most efficient manner in its current implementation. 3 seconds are proved to be enough for our experiments.

- `initDelaySecs` and `periodSecs` respectively specify the initial delay and the period of the ScheduledControl runs. Both are given as 1, that means the ScheduledControl thread begins collecting data after an initial delay of 1 seconds and it wakes itself at the end of every 1 second period after that.

- `sep` sets the column separator to be used by ScheduledControl when outputting aggregated statistics. We output the aggregation table inside the output file in a comma separated format, so we set sep as ','.

References

[1] Memcached: a distributed memory object caching system `https://memcached.org/`

[2] memtier_benchmark: a high-throughput benchmarking tool for Redis & Memcached `https://github.com/RedisLabs/memtier_benchmark`

[3] Dstat: a versatile resource statistics tool `http://dag.wiee.rs/home-made/dstat/`

[4] iPerf: the TCP, UDP and SCTP network bandwidth measurement tool `https://iperf.fr/`

[5] Jain, R. (1990). The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. John Wiley & Sons.

[6] Cluster SSH: cluster administration tool via SSH `https://github.com/duncs/clusterssh`