

# Advanced Systems Lab (Fall'16) – First Milestone

**Name:** *Georgios Touloupas*  
**Legi number:** *16-932-550*

## Grading

Section	Points
1.1	
1.2	
1.3	
1.4	
2.1	
2.2	
3.1	
3.2	
3.3	
Total	

# 1 System Description

## 1.1 Overall Architecture

The middleware implemented for the project consists of the following main parts: `NIOServer`, `WriteHandler` and `ReadHandler`

The `NIOServer`<sup>1</sup> class implements a NIO server accepting connections and requests from clients. When the thread receives a request, the memcached server that corresponds to that request's key is computed using consistent hashing by the `get` method of the `ConsistentHash`<sup>2</sup> class (presented in detail in Subsection 1.3). The request's message, as well as the `SocketChannel` it originated from are encapsulated in a `Request`<sup>3</sup> object. Using the `decode` method, the operation type (`set`, `delete` or `get`) of the request is decoded and the request is enqueued on the respective queue. For each memcache server, there is one `writeQueue` and one `readQueue`. These are blocking queues that allow concurrent access in a safe manner.

Each `WriteHandler`<sup>4</sup> thread handles `set` and `delete` requests for a specific primary memcached server. The thread also handles replication by forwarding the request to all of the corresponding memcached servers. The responses are received asynchronously and when the response of the last server is received the thread forwards to the client either the common response of the servers or an error message for different responses. The implementation of `WriteHandler` is presented in greater detail in Subsection 1.3.

Each `ReadHandler`<sup>5</sup> thread belongs to a thread pool corresponding to the memcached server that is the primary location of the request's key. The thread takes requests from the corresponding `readQueue` and forwards it to the memcached server. The response is received synchronously and is forwarded to the client. The implementation of `ReadHandler` is presented in greater detail in Subsection 1.4.

The middleware measures the following metrics for each request:

- **tMW**: Time spent between receiving a request from client and sending final response. A timestamp is taken by the `NIOServer` when the request is received. Another timestamp is taken after the memcached server response is forwarded back to the client by a `WriteHandler` or a `ReadHandler` thread. **tMW** is computed as the difference of these timestamps.
- **tQueue**: Time each request spends in a queue. A timestamp is taken by the `NIOServer` before the request is enqueued on the `writeQueue` or `readQueue`. Another timestamp is taken after the request is dequeued by a `WriteHandler` or a `ReadHandler` thread respectively. **tQueue** is computed as the difference of these timestamps.
- **tServer**: Processing time of the request (time between sending the request to the server and receiving the answer). A timestamp is taken by the `WriteHandler` or `ReadHandler` before the request is sent to the memcached server(s). Another timestamp is taken after the (last) memcached server response for a request is received by a `WriteHandler` or a `ReadHandler` thread. **tServer** is computed as the difference of these timestamps.

---

<sup>1</sup><https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/src/ch/ethz/gtouloup/NIOServer.java>

<sup>2</sup><https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/src/ch/ethz/gtouloup/ConsistentHash.java>

<sup>3</sup><https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/src/ch/ethz/gtouloup/Request.java>

<sup>4</sup><https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/src/ch/ethz/gtouloup/WriteHandler.java>

<sup>5</sup><https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/src/ch/ethz/gtouloup/ReadHandler.java>

- **fSuccess**: Flag encoding whether the request has been executed successfully (0 for fail, 1 for success). A **get** is considered successful if the response from memcached server ends with "END\r\n" and the flag is set by a **ReadHandler** thread. Regarding **set** and **delete** requests they are considered successful if the responses from memcached servers for a request are either "STORED\r\n" or "DELETED\r\n" and are all the same. In this case the flag is set by a **WriteHandler** thread.

## 1.2 Load Balancing and Hashing

The selection of the memcached server that corresponds to a request's key is performed by the **ConsistentHash** class. Consistent hashing maps the request keys to the same memcached server, as long as this server is available. If a memcached server is removed, its keys are handled by the next memcached server, that stores the key replicas if replication is enabled.

In the **ConsistentHash** constructor, using the **add** method each memcached server is associated with **virtualNodesNo** intervals, where the interval boundaries are determined by calculating the hash of the concatenation of the memcached server address with the interval id. If the memcached server is removed (**remove** method) its interval is taken over by a server associated with the next interval.

The **get** method of the class computes the memcached server that corresponds to a request's key by hashing the key and determining the memcached server associated with the interval that includes the hash.

The hash function used by the **hashFunction** method is MD5, which maps byte arrays on a 128-bit MD5 digest and gets the integer value of the digest. This means that each concatenation of the memcached server address with the interval id and each key are mapped to a range of  $2^{128}$  integers. Using a sufficiently large number of intervals for each memcached server (e.g. using 1000 for **virtualNodesNo**), the output space of the MD5 hash function tends to be uniformly partitioned among the memcached servers. Uniformly distributed keys will be uniformly mapped to the output space of the MD5 hash function. This means that keys will be uniformly distributed among the memcached servers, which leads to load balancing the requests.

## 1.3 Write Operations and Replication

Every write request is handled by the **WriteHandler** thread that corresponds to the memcached server that will be the primary location of the request's key. When a new request is dequeued from the **writeQueue** by the thread, it is forwarded to its primary memcached server as well as to the next **writeToCount - 1** servers, where **writeToCount** is the replication factor.

The **WriteHandler** thread is asynchronous, and can forward further write requests while receiving the responses from the memcached servers. When all **writeToCount** responses have been received, they are compared to each other. If they are the same, one of them is forwarded to the client, otherwise the error "SERVER\_ERROR replication error\r\n" is forwarded.

In a simple "write one" scenario, only one request would be forwarded to a single memcached server. When the response would be received, it would be forwarded directly to the client.

The implementation of **WriteHandler** does not leave the thread spinning between reading from the request queue and checking for responses. The thread blocks on a selector that awaits for server responses. When a new write request is enqueued for the thread, the thread is notified via the **notifyWriteHandler** method that wakes the selector, allowing the thread to dequeue the request. This prevents the thread from consuming empty CPU cycles, leading better performance.

When receiving a response from a server the thread must map it to the corresponding request. This is done by using a queue for each memcached server that stores the requests that are not yet answered by this server (**responseQueues**). Every server sends the responses in the

received order of the corresponding requests, therefore the request to which the next response will correspond will always be on the head of that queue.

For each write request sent to a memcached server the total latency consists of a latency for the transfer of the request and the response over the network, as well as the processing time of the request by the memcached server. For every write operation the request is sent to multiple servers for replication and the operation concludes when we have received the response from the last server. Therefore the latency of a write operation will be the maximum of the total latencies for the memcached servers. One factor that will limit the rate at which write operations can be carried out is the maximum network latency between the middleware and the memcached servers. Another factor is the volume of the requests to each server. If a memcached server is receiving a large volume of requests, then these requests will spend more time in the server's queue before being processed, which adds more latency and thus limits the rate at which reads can be carried out.

## 1.4 Read Operations and Thread Pool

Every read request is handled by a `ReadHandler` thread. This thread belongs to a thread pool corresponding to the memcached server that is the primary location of the request's key.

The `NIOServer` enqueues the request in `readQueue`, a blocking queue (`LinkedBlockingQueue`) on which the threads of the pool block. A blocking queue allows concurrent access in a safe manner.

When a new request is dequeued by a `ReadHandler`, it is forwarded to the memcached server. The thread blocks until a response is received, which in turn is forwarded to the client. Afterwards the thread blocks on the queue again for the next request.

Every `ReadHandler` thread has its own connection to its corresponding memcached server. The default maximum number of connections for memcached is 1024, so unless we want that many `ReadHandler` threads one connection per thread works. The alternative would be sharing one connection among the threads of pool that would have to be accessed in a safe concurrent manner. This approach was not chosen because it increases complexity and introduces delays due to the concurrent access to the pool's connection.

## 2 Memcached Baselines

In this experiment we measure the baseline performance of a single memcached server connected to one or multiple clients. We experiment with one, as well as multiple virtual clients increasing their number from 8 to 168 in steps of 8. Aside from the single client experiment where only one memaslap client is used, we use two memaslap clients with the same number of virtual clients (increasing in steps of 4).

We use three Basic\_A2 Azure machines for the memcached server and the two memaslap clients. Each experiment for a specific number of virtual clients runs for 30 seconds with the `smallvalue.cfg`<sup>6</sup> workload configuration (Key 16B, Value 128B, Writes 1%) and is repeated 5 times.

Number of servers	1
Number of client machines	1 to 2
Virtual clients / machine	1 to 84
Workload	Key 16B, Value 128B, Writes 1%
Middleware	Not present
Runtime x repetitions	30 sec x 5
Log files	baselinebench*

### 2.1 Throughput

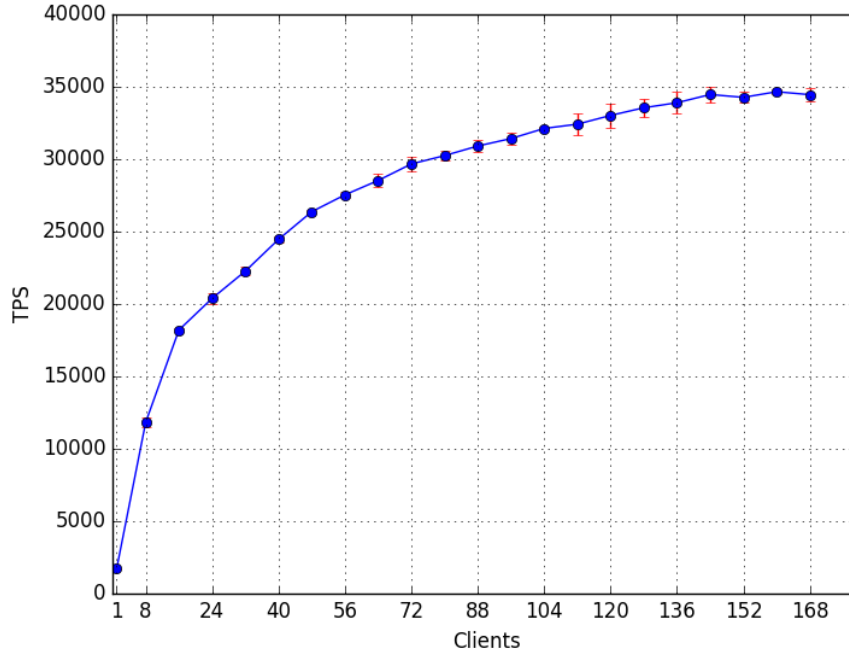


Figure 1: Baseline aggregate throughput

Figure 1 shows that the aggregate throughput is increasing as we increase the number of virtual clients, however the throughput per client is decreasing. This happens because more clients produce a larger volume of requests that have to be processed by the memcached server. Since the server has finite processing power, after a certain number of virtual clients it will start

<sup>6</sup><http://www.systems.ethz.ch/sites/default/files/file/as12016/memaslap-workloads.tar>

working on its maximum processing capacity, where the server is saturated. As we can see from the figure, when increasing the number of clients over 144 the throughput remains stable, therefore 144 virtual clients saturate a single memcached server. The maximum throughput of a single memcached server without the middleware is around 35000 TPS.

## 2.2 Response time

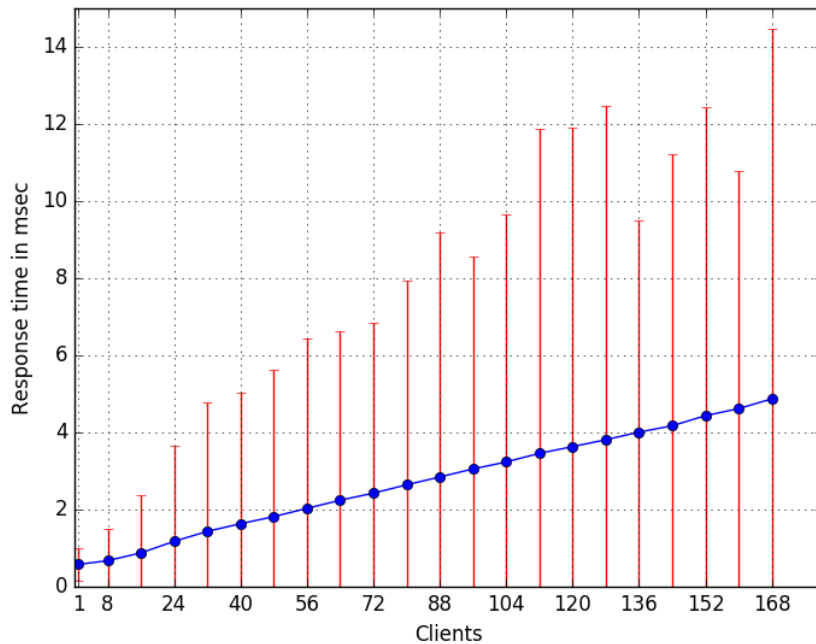


Figure 2: Baseline response time

Figure 2 shows that the average response time is increasing linearly as we increase the number of virtual clients. More clients produce a larger volume of requests that are queued for processing by the memcached server. This causes each request to take more time to be answered, since a larger volume of requests leads to increased time in the server's queue for the request and thus the response time increases.

Another thing worth noting is that the standard deviation of the response time is very high compared to the average response time.

### 3 Stability Trace

In this experiment we show that the middleware is functional and can handle a long-running workload without crashing or degrading in performance. We run the middleware with full replication for one hour connected to three memcached servers and three memaslap clients. During the experiment, the middleware runs with 64 threads per thread pool and each memaslap client runs with 64 virtual clients.

We use three Basic\_A2 Azure machines for the memcached servers, three Basic\_A2 Azure machines for the memaslap clients and one Basic\_A4 Azure machine for the middleware. For this experiment the `smallvalue.cfg`<sup>7</sup> workload configuration (Key 16B, Value 128B, Writes 1%) is used.

Number of servers	3
Number of client machines	3
Virtual clients / machine	64
Workload	Key 16B, Value 128B, Writes 1%
Middleware	Replicate to all (R=3)
Runtime x repetitions	1h x 1
Log files	stability_1, stability_2, stability_3, middleware_out, middleware_logger

#### 3.1 Throughput

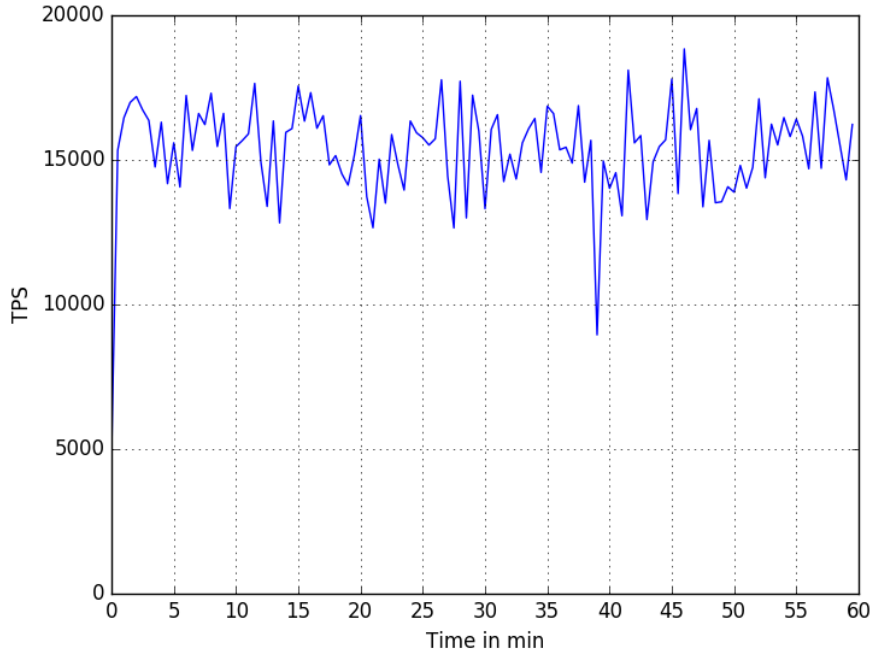


Figure 3: Stability trace aggregate throughput

Figure 3 shows that our system is stable during the one hour experiment. The sample rate used for the plot was once every 30 sec. In the first minutes of the experiment we can notice the warm up phase. We compute that during the steady state the system has a non degrading

<sup>7</sup><http://www.systems.ethz.ch/sites/default/files/file/asl2016/memaslap-workloads.tar>

aggregate throughput of 15350 TPS. The drop during minute 19 is probably due to garbage collection.

### 3.2 Response time

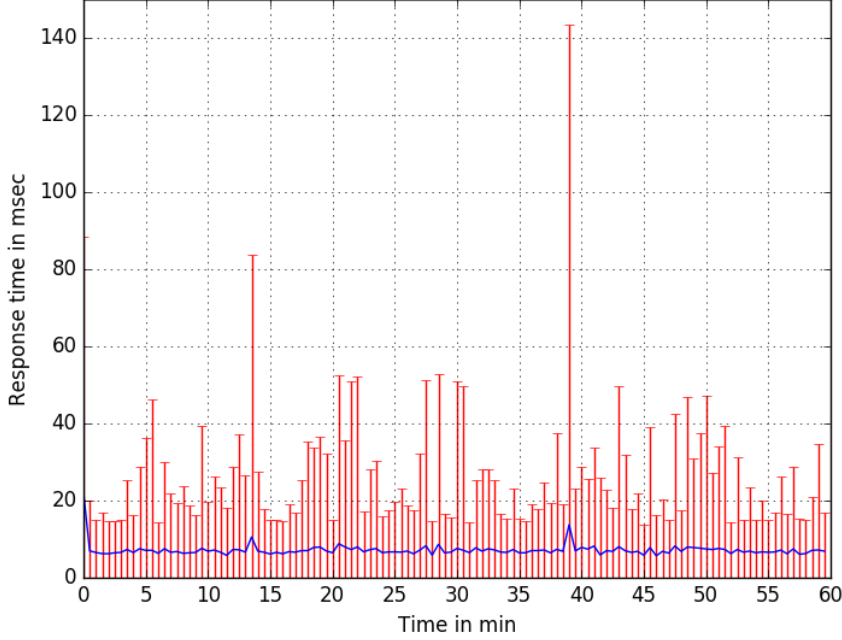


Figure 4: Stability trace response time

Figure 4 shows that the average response time is stable during the one hour experiment. The sample rate used for the plot was once every 30 sec. We compute that during the steady state the response time is 7.12 msec. The peaks during minute 13 and 39 are probably due to garbage collection.

We also note that the standard deviation of the response time is very high compared to the average response time, which is to be expected since the same happened without the middleware as we saw in Subsection 2.2.

### 3.3 Overhead of middleware

We compare the performance observed during the stability trace to the performance of the baseline experiment with 64 virtual clients, because during the stability trace we have 3 memcached servers and 3x64 virtual clients. Considering throughput, the average throughput during the stability trace is 15350 TPS, compared to 28530 TPS during baseline experiment. The average response time during the stability trace is 7.12 msec, compared to 2.24 msec during baseline experiment. As we can see the middleware introduces overheads in both throughput and response time. These overheads are caused by the write replication, the extra network latency during the communication with the memcached servers and the delays inside the middleware (processing the messages, queuing, garbage collection etc).

Metric	Without middleware	With middleware
Throughput (TPS)	28530	15350
Response time (msec)	2.24	7.12



## Logfile listing

Short name	Location
baselinebench_1.1.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_1_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_1_1.log</a>
baselinebench_1.2.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_2_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_2_1.log</a>
baselinebench_1.3.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_3_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_3_1.log</a>
baselinebench_1.4.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_4_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_4_1.log</a>
baselinebench_1.5.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_5_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_1_5_1.log</a>
baselinebench_8.1.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_1_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_1_1.log</a>
baselinebench_8.1.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_1_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_1_2.log</a>
baselinebench_8.2.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_2_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_2_1.log</a>
baselinebench_8.2.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_2_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_2_2.log</a>
baselinebench_8.3.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_3_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_3_1.log</a>
baselinebench_8.3.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_3_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_3_2.log</a>
baselinebench_8.4.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_4_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_4_1.log</a>
baselinebench_8.4.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_4_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_4_2.log</a>
baselinebench_8.5.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_5_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_5_1.log</a>
baselinebench_8.5.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_5_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_8_5_2.log</a>
baselinebench_16.1.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_1_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_1_1.log</a>
baselinebench_16.1.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_1_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_1_2.log</a>
baselinebench_16.2.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_2_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_2_1.log</a>
baselinebench_16.2.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_2_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_2_2.log</a>
baselinebench_16.3.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_3_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_3_1.log</a>
baselinebench_16.3.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_3_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_3_2.log</a>
baselinebench_16.4.1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_4_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_4_1.log</a>
baselinebench_16.4.2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_4_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/baseline/baselinebench_16_4_2.log</a>













[illegible]







stability_1	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/stability_1.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/stability_1.log</a>
stability_2	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/stability_2.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/stability_2.log</a>
stability_3	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/stability_3.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/stability_3.log</a>
middleware_out	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/middleware_out.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/middleware_out.log</a>
middleware_logger	<a href="https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/middleware_logger.log">https://gitlab.inf.ethz.ch/gtouloup/asl-fall16-project/blob/master/logfiles/stabilitytrace/middleware_logger.log</a>