Assumptions:
Number of services: 100
Number of APIs per service: 30
Average request timeout per API: 1 minute
Average request size = 1 KB

Questions:

1. How much memory do you need for request allowance in the Oracle?

Assuming each API needs a timer wheel, we shall need buckets proportional to the request timeout per API.

This is 1 minute = 60 seconds = 60 buckets.

Assume an average of 10 requests per second per API.

In total, we have we have (10 * number of apis) requests.

= (10 * number of apis per service * number of services)
= (10 * 30 * 100)
= (30,000)

Every second, we have 30,000 requests. Since we don't need to store the entire request in the oracle, we store only the request ID.

An 8 byte ID can uniquely identify each request. We have a requirement of 8B * 30,000.
= 240 KB.

Since the timer wheel has 60 buckets, we may need 240KB * 60.

~ 250KB * 60
= 1MB * 15
= 15 MB

This should easily fit in memory. However, for consistency and fault tolerance, it would be nice to have these records in a DB. (Also note that the oracle is a distributed service)

2) An API has a timeout of 10 seconds. Each request takes at least 1 second to process. What is the maximum queue wait time?

The request-response flow has three parts =

a) Time taken for request to arrive at service.
b) Time taken to process the request.
c) Time taken for response to arrive at client.

A) Request sent to Server ->
B) Request waiting in Queue ->
C) Request being processed ->
D) Response sent to client

The processing time includes the wait time in the queue.

Assume the travel time to and from the client is 100 ms. That means we have 10 seconds - 200ms = 9800 ms to process each request.

Removing raw processing time, we have 8800 ms. This is the maximum wait time.

3. In the above scenario, what should be the queue size to allow 1000 requests per second?

We have 1000 requests per second. Assume average length of request is 1 KB. That means we have 1000 * 1KB per second = 1MB per second.

As the maximum wait time is 8.8 seconds, we need 8.8 s * 1 MB/s = **8.8 MB queue size**.