```
while(exit signal is off)

    temp_respond_time += get response time from URL

    if(check response time for 5 times)
        final_respond_time = temp_respond_time/5
        if(final_respond_time is not in range [0.3 - 1.2])
            if(final_respond_time is larger than upper limit for 3 times in last 5 samples)
                scale up
            if(final_respond_time is smaller than lower limit for 3 times in last 5 samples)
                scale down
        temp_respond_time = 0

    sleep for 1 second
```

The lower threshold is 0.3 seconds and higher threshold is 1.2 seconds.
When the server's response time is less than 0.3 seconds, we can assume that the server is a little bit over size for the current workload, such as three users while the server has a scale of three or more.
When the server's response time is more than 1.2 seconds, the client's actual response time would be 1.2 plus server work time, around 2.5 seconds. Therefore, in this case, the server may need to scale out to respond faster.

The final scale decision is made according to the last 5 sample times because during testing, we found that the final respond time may contains high peaks or low peaks.