M/M/c models

Consider a server that executes jobs arriving according to a Poisson process with an average inter arrival time $I = 2$ *s*, and serves them with an average service time *D = 1.6* s.

Determine:

1. The utilization of the system
2. The probability of having exactly two jobs in the system
3. The probability of having less than 5 jobs in the system
4. The average queue length (without considering job in service)
5. The average response time
6. The probability that the response time is greater than 2 s.
7. The 95 percentile of the response time distribution

After 1 year, the load has increased to λ = 1 job/s, making the current solution no longer applicable. The system administrator adds a second server and a load balancer: jobs enqueues at the load balancer, and then are sent to the first available server. Considering the communication time between load balancers and servers to be negligible compared to the service times, we can model this system as an M/M/2 queue. Determine for this new configuration:

1. The total and average utilization of the system
2. The probability of having exactly two jobs in the system
3. The probability of having less than 5 jobs in the system
4. The average queue length (jobs not in service)
5. The average response time

The next year, the load has increased to λ = 4 job/s, making the current solution no longer applicable. Determine the minimum number of servers *c* required to make the system stable. Modelling this system as an M/M/c queue, determine for this new configuration:

1. The total and average utilization of the system
2. The probability of having exactly two jobs in the system
3. The probability of having less than 5 jobs in the system
4. The average queue length (jobs not in service)
5. The average response time

Finally, the user moves the service to the cloud, using an automatic scaling technique that can provide a parallelization level that might be considered infinite. Modelling this system as an M/M/∞ queue, and considering an arrival rate λ = 10 job/s, determine for this new configuration:

1. The total utilization of the system
2. The probability of having exactly two jobs in the system
3. The probability of having less than 5 jobs in the system
4. The average response time