Performance indices of an G/G/c queue

A web server receives jobs according to a Poisson process of rate λ = 20 j/s. The duration of each job is distributed according to an Erlang distribution, of rate λ_e = 100 j/s and k = 4.

Compute:
1. The utilization of the system
2. The (exact) average response time
3. The (exact) average number of jobs in the system

After a year, the traffic increases in rate and variability: now it can be considered distributed according to an Hyper-Exponential distribution, with λ_1 = 40 j/s , λ_1 = 240 j/s , p_1 = 80% . To support this new scenario, several new web servers are added, together with a load-balancer that holds request in a single queue, and dispatches them to the first available server. Assuming the time required by the load balancer to be negligible (i.e., the system can be modelled with a G/G/c queue),

Compute:
1. The minimum number of servers *c* for which the considered system is stable
2. The average utilization of the system
3. The approximate average response time
4. The approximate average number of jobs in the system