Performance indices of an G/G/c queue

An authentication server receives jobs according to a Poisson process of rate $\lambda = 500\ j/s$. The duration of each job is distributed according to an Hypo-Exponential, of rate $\mu_1 = 1500\ j/s$ and $\mu_2 = 1000\ j/s$.

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 and stabilizes: now it can be considered distributed according to a 4 stage Erlang distribution, with $\lambda = 4000\ j/s$. To support this new scenario, a second authentication server is 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/2 queue), compute:

1. The average utilization of the system
2. The approximate average response time
3. The approximate average number of jobs in the system