DES simulation assignment

Multiple queues and multiple servers

For this assignment we'll be using the following notation:
λ – the arrival rate into the system as a whole.
μ – the capacity of each of n equal servers.
ρ represents the system load. In a single server system, it will be: $\rho=\lambda/\mu$
In a multi-server system (one queue with n equal servers, each with capacity μ), it will be $\rho=\lambda/(n\mu)$.

Queuing theory tells us that for FIFO scheduling the average waiting times are shorter for an M/M/n queue and a system load ρ and processor capacity μ than for a single M/M/1 queue with the same load characteristics (and thus an n-fold lower arrival rate). Of course, ρ must be less than one, but the experiment only becomes interesting when ρ is not much less than one.

1) Look up and/or derive this theoretical result, at least for n=2. Describe how it is derived. Can you also give a non-mathematical explanation?
2) Write a DES program to verify this for n=1, n=2 and n=4. Make sure that your result has a high and known statistical significance. How does the number of measurements required to attain this depend on ρ?
3) Also compare the result to that for an M/M/1 queue with shortest job first scheduling, where you always give priority to the smallest jobs.
4) Now experiment with different service rate distributions. On the one hand try the M/D/1 and M/D/n queues, on the other hand try a long-tail distribution. For the latter you may e.g. use a distribution where 75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% an exponential distribution with an average service time of 5.0 (an example of a hyperexponential distribution).

Write your program using SimPy.