Performance indices of a Deep Lerning training service

A transactional system, performs training of Deep Learning models, using both a CPU and GPU. Each job performs a training iteration: first it undergo a preprocessing stage on CPU, then it runs on the GPU, and finally it requires another run on the CPU to finalize the computation. Both CPU and GPU works in FCFS, and can be considered single server. Computation times can be considered Erlang distributed, with a coefficient of variation of 0.3333, and the following averages:

- CPU – first pass: 6 sec.
- GPU – 10 sec.
- CPU – second pass: 4 sec.

The think time can be considered exponentially distributed, with an average of Z = 20 sec.

Considering a variable population of *N = 1 to 20* users, compute using JMT:

1. The system throughput
2. The utilization of the CPU and of the GPU
3. The average system response time