The description of the solution:

To overcome the workload, the number of consumers is scaled by launching or terminating consumers according to the current workload. Generally, the scaling could be implemented either scheduled or dynamic. In our scenario, the dynamic scaling is implemented because of the workload is unpredictable (i.e. different from time to time). The dynamic scaling is based on criteria such as CPU utilization. In this simulation, the criteria are strongly based on the number of messages that the consumer can process it. Therefore, dividing the total number of published messages by the capability of consumer to process the messages is equaling to the required consumer instances. This allows running as few consumers as possible in compatible with the current workload.

The main steps of the scaling process are:

1. Calculating the required consumer instances according to the workload:

   criteria = maximum number of messages that the consumer can process it

   $$\text{required consumers} = \left\lceil \frac{\textbf{total number of messages (i.e.queue length)}}{\textbf{criteria}} \right\rceil$$

2. Comparing the required and the actual consumers to scale the consumer group (i.e. increase or decrease the number of consumers)
3. Partition the total number of messages into partitions according to the new consumer group scale.

The main rules between the queue and consumer group:

1. Each partition of messages is consuming by only one consumer.
2. The number of consumers should not be more than the total number of partitions.