1. If a queuing system with one server has a workload of 1000 tasks arriving per second, and the average number of tasks waiting or getting service (i.e., tasks in the system) is 5, what is the average response time per task?

$$N = X \cdot R$$

$$\frac{5}{1000} = R$$

$$R = 0.005\,seconds \quad \text{or} \quad 5\,ms$$

2. Even when the request rate is below the server's service rate, bursty arrivals suffer queuing delay. Please use an example to explain why.

Suppose the server can process each task at 0.1 seconds.
If there's a bursty arrival of 20 tasks at the same time, then the response time of that is

$$\frac{(0.1+0.2+0.3+0.4+0.5+0.6+0.7+0.8+0.9+1.0+1.1+1.2+1.3+1.4+1.5+1.6+1.7+1.8+1.9+2.0)}{20}$$

$$\frac{21}{20} = 1.05\,seconds$$

The first arrival task gets done with no wait time, but every other task after it suffers a wait time.