Question 1:

Let us consider a simple DC topology with one web server and 200 identical backend servers. The web server scatters requests independently to all the backend servers in a batch in parallel and gathers the responses before sending the aggregated response to the client. The backend servers respond in 5 milliseconds for 99% requests and 200 milliseconds for the rest.

a) What is the probability of waiting for more than 200 milliseconds before the web server has responses from all the backend servers?

b) Say, the response time is 5 milliseconds for 99.9% of the requests. What is the probability now?

Question 2:

In the same setting as Q1 above, say each backend server prepares the response, generates a random number \( x \) between 0 and 1000, waits for \( x \) microseconds and then sends the response back to the web server, instead of sending it immediately.

a) What does this scheme want to achieve?

b) What is the main drawback of this method?

Question 3:

To deploy a cloud application that occupies multiple machines, we need to make a decision where to place it - to use the least possible number of racks (“rack local”) or to spread it evenly across the data center. What are the arguments for/against rack local deployments?
Question 4:

Does low link utilization always imply less processing at the switch? At a fixed line-rate and utilization, what is the impact of different packet size distributions on processing load?

We are happy to give individual feedback in person on request.