Buffer size and Quality of service

David Allwright

Nodes in a communication system

At a node data arrives over fast (fibre-optic) links as discrete data jobs to be transmitted over a radio link.

The data arrives much faster than the radio link can transmit so there is a buffer at the node to store data that is awaiting transmission.

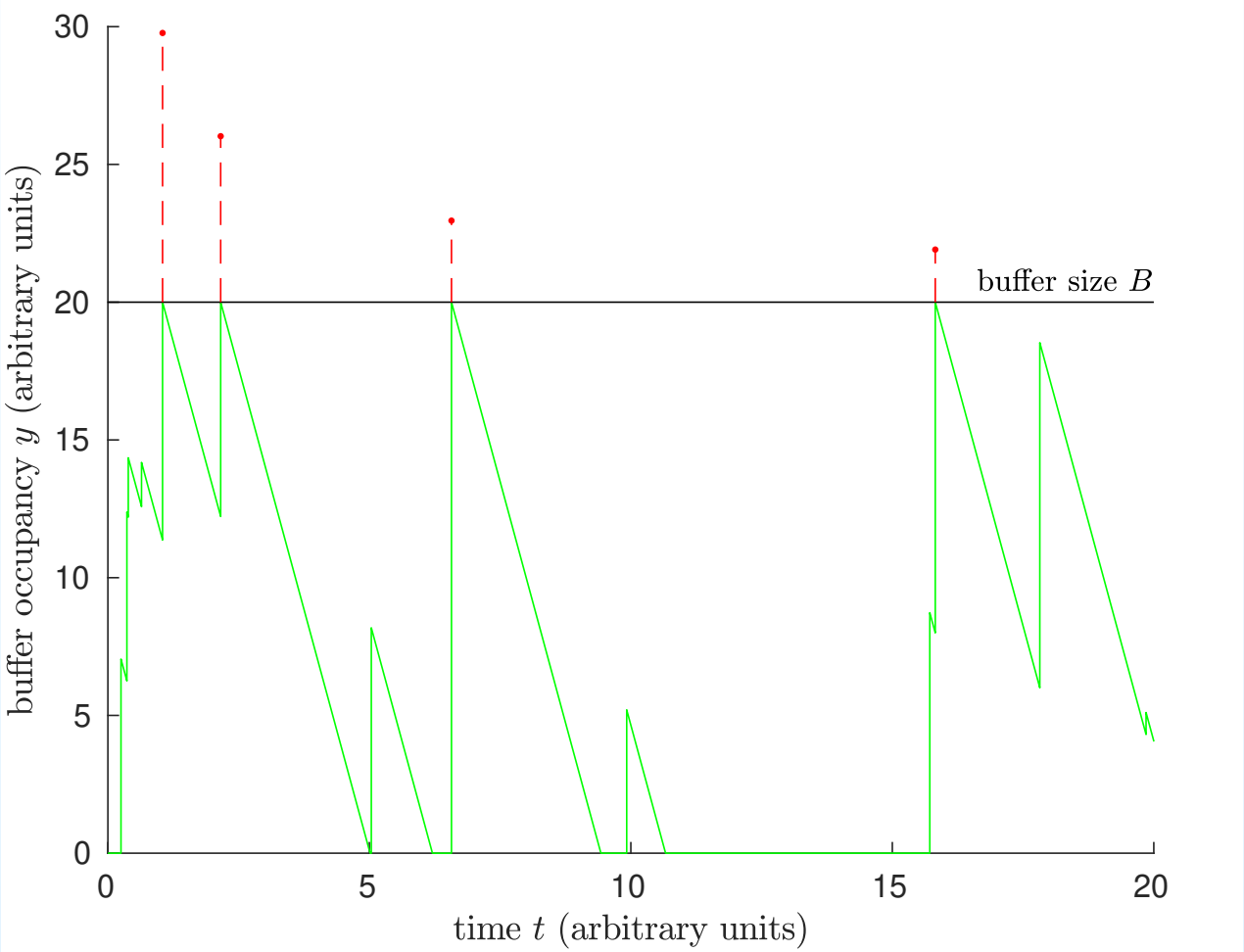
While there is stored data it is transmitted continuously.

The bigger the buffer at the node, the better the quality of service.

How do we quantify the relationship between buffer size and quality of service?

Details of the system

The buffer occupancy over time could behave something like this:



The dashed lines represent data that did not fit in the buffer on arrival at the node: it will have to be retransmitted later.

Quality of service measures

The service experienced by any particular user or job is unpredictable so the typical measures of quality of service (QoS) have to be specified as *averages*. For instance, the QoS requirements can include things like:

- the average proportion of data that has to be sent again is below some specified ϵ_{QoS};
- the average proportion of jobs that need some retransmission is below some specified δ_{QoS}.
- the average time a job has to wait before it begins transmission is below some specified t_{QoS}.

Job arrival process

Jobs arrive unpredictably, and in particular:

- There are “busy times” when a lot of data traffic comes through.
- There are different kinds of data jobs, that have very different typical sizes.
- The distribution of job sizes is heavy-tailed.

Questions to address

The communications engineers would like to have ways of answering questions like:

- Are there some “rules of thumb” they an use to get a rough idea of what buffer size to choose in a particular situation?
- Do those rules of thumb overestimate or underestimate the QoS?
- Are there mathematical models that are better than rules of thumb to use for this?
- Do they underestimate or overestimate?
- Are there situations where running a simulation over a long time and averaging the results is the best way?
- We have described the situation where there is just *one* server at the node. How can the modelling be extended to *multiple* servers?