SOLUTION NOTES

Computer Systems Modelling 2001 Paper 7 Question 8 (TLH)

(a) What criteria would you consider when selecting between a model based on queueing theory and one based on simulation? When might you use both approaches?

- Stage of development: how complete is the system design? Has it been implemented? Detailed simulation requires a sufficiently detailed design.
- Time available: simulations may take a long time to execute and must be re-run for each different set of input parameters. Mathematical models may take time to derive, but can then give results over a range of parameterizations.
- Resources: people with the relevant skills, computers to run simulations.
- Accuracy: how valid are the simplifying assumptions needed to make a mathematical model tractable? Is the workload characterized sufficiently well for a simulation to be realistic?
- Saleability: will people have faith in the results?

Might use both to confirm that they give consistent answers – particularly if there are no data-points from a real system.

(b) Describe the structure of a discrete event simulator. What is the principal data structure involved?

The simulated system is divided into a number of *physical processes* which interact solely by sending and receiving time-stamped events.

The main data structure is the *event queue* which holds events which have been sent but not yet received. They are ordered by *virtual time*. The simulator proceeds by taking the earliest event from the queue and dispatching it to its target (which may in turn generate further events for insertion into the queue).

(c) An queueing network is characterised by a set of *visit counts*, V_i, and *per-visit service requirements*, S_i, for each of N devices. Derive upper bounds on the system throughput *(i)* when the load is very low and *(ii)* as the load tends to infinity.

Let the minimum residence time for a customer be $R_{min} = \sum_{i=1}^{N} V_i S_i$.

At low load there is no contention at the servers and hence with k customers k/R_{min} forms an initial upper bound on throughput.

At high load the system throughput is limited by the performance of a bottleneck device – i.e. one for which $V_i S_i$ is greatest. If this is device b then it can service one visit per S_b and hence one customer's complete set of visits every $V_b S_b$. The maximum throughput is therefore $1/V_b S_b$.

(d) In what situations may the bounds be particularly imprecise? What can be done to construct tighter bounds for the system throughput?

These bounds may be imprecise at mid-range loads for which there are some queueing delays but the bottleneck device(s) are not yet close to saturation.

Tighter bounds can be constructed by considering *balanced systems* corresponding to pessimistic distributions of the load over some number of devices.

Two possible systems that could be used to give upper bounds:

- Construct a balanced system in which the parameters for *each* device correspond to those of a bottleneck device in the original system.
- Construct a balanced system which preserves the total service demand – i.e. the sum of $V_i S_i$ – but does so using a smaller number of devices each corresponding to a bottleneck device in the original system.