SIMULATION AND MODELLING

Tony Field and **Giuliano Casale**

{ajf,gcasale}@doc.ic.ac.uk

Suggested Books

- Performance Modeling and Design of Computer Systems: Queueing Theory in Action
  Mor Harchol-Balter
  Cambridge University Press, 2013
- Probability, Markov Chains, Queues and Simulation
  William Stewart
  Princeton University Press, 2009
- Discrete-event System Simulation (5/e)
  J. Banks, J.S. Carson, B.L. Nelson and D.M. Nicol
  Prentice Hall International, 2010
- Simulation Modeling and Analysis
  A.M. Law and W.D. Kelton
  McGraw Hill, 2000

Syllabus

Part I (Tony)

1. Introduction
2. Operational laws
3. Poisson processes
4. Discrete-event simulation
5. Markov processes

Part II (Giuliano)

1. Markovian queues
2. Open queueing networks
3. Fork-join subsystems
4. Application examples
5. Parallel discrete-event simulation

Introduction

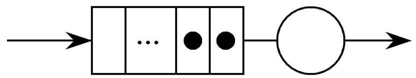
- This course is about using measurements and models to understand performance aspects of real-world systems
- We'll focus on computer systems but the principles are widely applicable
- Performance models capture the way jobs/customers/entities move around a system and compete for its resources
- It then becomes a tool for reasoning about the system's performance, e.g. in order to:
  - Understand the observable behaviour of an existing system
  - Guide changes, rewrites or upgrades to a system
  - Study new or imaginary systems

Applications

- There are many application areas, e.g.
  - Compute/web servers
  - Cloud computing/storage systems
  - Distributed systems
  - Mobile & sensor networks
  - Manufacturing
  - Transport & logistics
  - Healthcare provision
  - Military logistics & strategy
- We'll (try to!) balance theory and practice, so you'll understand how/why the techniques you'll be learning work

Motivation: Some Example Problems

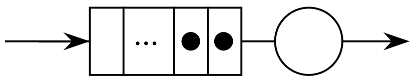
An in-memory TP system accepts and processes a stream of transactions, mediated through a (large) FIFO job queue:



- Transactions arrive "randomly" at some specified rate
- The service times are distributed exponentially, with some specified rate

Q: If both the arrival rate and service rate are doubled, what happens to the mean response time?

Consider an operating system scheduler where the job sizes (X) are highly variable

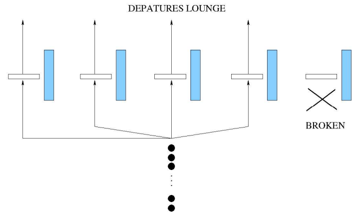


$$10 \quad \leq \quad C_X^2 = \frac{VAR(X)}{E(X)^2} \quad \leq \quad 50$$

Rank the following schedulers in order of mean processing time per job:

- Round Robin (preemptive)
- First-Come-First-Served
- Shortest Job First (non-preemptive)
- Shortest Remaining Processing Time (preemptive)

There are five security scanners between the check-in and departures area at Heathrow (T4); one of them is broken:



- Around 0.5 customers pass through the terminal each second and it takes just under 8 seconds on average to scan each passenger
- The average delay is about 30 minutes (1600 seconds)

Q: How long would it take on average if all 5 scanners were working?