# Examples class 2
# Computer Systems Modelling

November 6, 2000

**Operational analysis Q3** An queueing network described as *open* and *feed forward* is to be analyzed using operational laws.

The mean arrival rate into the system is known to be $\lambda$.

(*a*) Describe the significance of the two phrases in italics.

(*b*) The mean service times, visit counts and connectivity are known for each device. Why is it still impossible, in general, to calculate an upper bound on the residence time of a job in the network?

(*c*) Suppose, however, that there are never more than five jobs in the network at once. Outline how you could derive an upper bound on the residence time.

An *open* queueing network is one with external arrivals and departures: the number of jobs in the system can vary with time. In contrast a *closed* queueing network has neither external arrivals nor departures: the number of jobs remains fixed over time.

A *feed forward* network is one in which a particular job makes *at most* one visit to a particular node. In practise, a feed forward network would also be expected to be open.

Even with a known arrival rate of, say, $\lambda$ jobs per second it is generally impossible to bound the number of jobs in the network and consequently impossible to bound the residence time of jobs. To see why the number of jobs is unbounded notice that the rate $\lambda$ can be achieved by large bursts of work followed by idle periods. Then, to show that the residence time is also unbounded, consider a simple queueing network with a single first-come-first-served M/M/1 node: if the queue length is unbounded then an arriving customer must wait for all of those queued jobs to complete before it receives service.

Now suppose that an upper bound is known on the number of jobs in the system. If we can assume first-come-first-served queueing throughout the network then the maximum residence time can be bounded above by summing the maximum residence time at each node when the customer must queue behind each of the others.

**2000 Paper 9 Q7** A database system has a central processor and three (different) discs. Measurements are taken for 1000 transactions on a lightly loaded system and the following observations are made.

- The CPU scheduler initiated or resumed transaction processing 10,000 times. The total CPU usage was 25 seconds.
- Disc 1 made 5000 transfers with an average transfer time of 10 ms.
- Disc 2 made 2000 transfers with an average transfer time of 50 ms.
- Disc 3 made 2000 transfers with an average transfer time of 20 ms.

Derive the visit counts, service times and transaction service demands. What is the bottleneck device? What is the maximum throughput of the system measured in transactions per second? [6 marks]

Describe *two* balanced systems which bound the throughput of the system. What is the maximum throughput of these systems? [7 marks]

Recall that the throughput of a balanced system with $K$ devices, $N$ customers and service demand $D$ per device is

$$X(N) = \frac{N}{(N+K-1)} \times \frac{1}{D}$$

How many transactions do you expect to be in the system with a throughput of 7 transactions per second? [7 marks]

Recall that the *visit count* is the number of times that each job visits a particular node, that the *service time* is how long it spends there upon each visit and that the *service demand* is the total time it spends there while in the system.

Hence:

| | | | |
|---|---|---|---|
| CPU | $v_c = \frac{10000}{1000} = 10$ | $s_c = \frac{25s}{10000} = 2.5ms$ | $d_c = v_c s_c = 25ms$ |
| Disc 1 | $v_1 = \frac{5000}{1000} = 5$ | $s_1 = 10ms$ | $d_1 = v_1 s_1 = 50ms$ |
| Disc 2 | $v_2 = \frac{2000}{1000} = 2$ | $s_2 = 50ms$ | $d_2 = v_2 s_2 = 100ms$ |
| Disc 3 | $v_3 = \frac{2000}{1000} = 2$ | $s_3 = 20ms$ | $d_3 = v_3 s_3 = 40ms$ |

Furthermore, define $d_{max} = 100ms$, $d_{min} = 25ms$, $d_{av} = 53.8ms$, $d_{tot} = 215ms$ as the maximum, minimum, average and total service demands across the 4 devices.

The bottleneck device is the one with the largest service demand, $d_{max}$. In this case it is disc 2.

The maximum throughput is bounded above by the performance of the bottleneck device. Each transaction requires $d_2 = 100ms$ processing at disc 2, so the maximum throughput can be no greater than 10 transactions per second.

We can derive an upper bound on the system throughput by considering a balanced system in which the service demand $d_{tot}$ is spread evenly over the devices. In this case 4 devices each with a service demand of $d_{av} = 53.8ms$. The maximum of that system is 18.6 transactions per second.

We can derive a lower bound on the system throughput by considering a second balanced system in which the service demand $d_{tot}$ is distributed over a smaller number of devices, all of which have the load $d_{max}$. We require a total service demand of $215ms$ and hence require 3 devices at $100ms$. The maxmimum of that system remains 10 transactions per second.

We can now use these systems to bound the performance of the database being modelled.

$$
\begin{aligned}
X &= \frac{N}{(N+K-1)} \times \frac{1}{D} \\
DNX + DKX - DX &= N \\
DX(K-1) &= N(1-DX) \\
N &= \frac{DX(K-1)}{1-DX}
\end{aligned}
$$

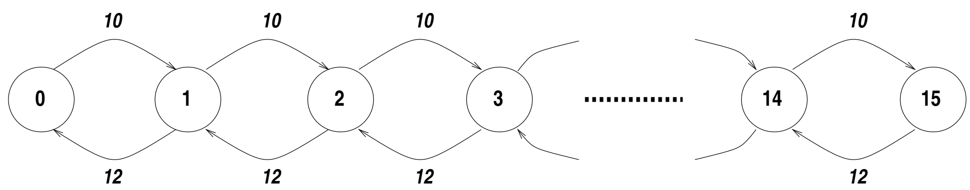
For $X = 7s^{-1}$, $D = D_{av} = 53.8ms$, $K = 4$ we get $N_{opt} = 1.81$.

For $X = 7s^{-1}$, $D = D_{max} = 100ms$, $K = 3$ we get $N_{pess} = 2.00$.

Hence $1.81 < N < 2.00$.

**Queueing theory Q4** Given an M/M/1/K queue with $\lambda = 10, \mu = 12, K = 15$ over what proportion of time are customers rejected from the queue? What is the effective arrival rate? What is the effective utilization of the server?

The $M/M/1/K$ queue has exponentially distributed inter-arrival and service times, a single server and a limit of $K$ customers in the system. We can represent the $K = 15$ case by a Markov chain in which the states $0 \ldots 15$ identify the number of customers present and the arrival/departure rates are encoded directly as the state transition rates.




Define $p_i$ to be the equilibrium probability that the system is in state $i$. (Note both that the arrivals rate $\lambda$ is less than the service rate $\mu$ and also that the state space is finite).

Proceeding in the usual way, consider the flow between states $i$ and $i-1$.

$$\begin{aligned}
\mu p_i &= \lambda p_{i-1} \qquad\qquad 1 \le i \le 15 \\
\Rightarrow p_i &= \frac{10}{12}\, p_{i-1} = \frac{5}{6}\, p_{i-1} \\
\Rightarrow p_i &= \left(\frac{5}{6}\right)^i p_0 \\
\text{also} \qquad \sum_{i=0}^{15} p_i &= 1 \\
\Rightarrow \sum_{i=0}^{15} \left(\frac{5}{6}\right)^i p_0 &= 1 \\
\Rightarrow p_0 \frac{(5/6)^{16} - 1}{5/6 - 1} &= 1 \\
\Rightarrow p_0 &\approx 0.176
\end{aligned}$$

Customers are rejected when the system is full. $p_{15} = 0.011$, i.e. 1.1% of the time.

The effective arrival rate is the rate at which customers are accepted into the system, i.e. $\lambda(1 - p_{15}) = 9.89s^{-1}$.

The effective utilization is simply the fraction of time that the server is busy, $1 - p_0 = 0.824$.

**1996 Paper 9 Q3** The Erlangian distribution $E_r$, with parameters $(r, \mu)$ is given by

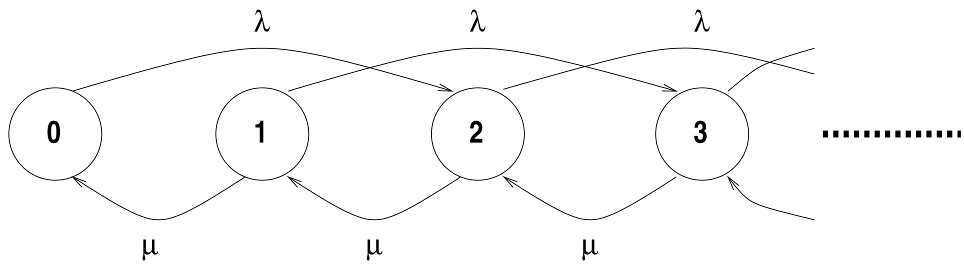
$$f_X(x) = \frac{\mu(\mu x)^{r-1}}{(r-1)!} \mathrm{e}^{-\mu x} \qquad x \ge 0$$

Given an $M/E_2/1$ queue, draw a Markov Chain describing the queueing system, and derive a formula for the variance of the service time distribution. Give an example of a queueing system in which the use of an Erlang service time distribution would be useful. [5 marks]

*(Other sections covered in the first examples class)*

In order to model the $M/E_2/1$ system as a Markov chain we need to express the server's operation in terms of exponentially-distributed times rather than the Erlangian distribution. To do this we'll use the fact that an $E_r$ server can be modelled as $r$ independent successive stages, each with an expoentially-distributed service time.

The states in the Markov chain will represent the number of these server stages that are left to complete: the arrivals at rate $\lambda$ will introduce two stages of work and the departures at rate $\mu$ will remove one stage of work.




There are two general techniques we might use for deriving the variance of the service time distribution. One option is to use the definition of the variance $\sigma^2$ as the second central moment:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \overline{X})^2 f_X(x)\, dx$$

However, in this case it is easier to use the second technique of expressing the required random variable in terms of others whose variance is more readily known. The variance of an exponentially distributed random variable with parameter $\mu$ is $\mu^{-2}$ so, because the stages are independent, the variance of an $E_2$ random variable is simple $2\mu^{-2}$.

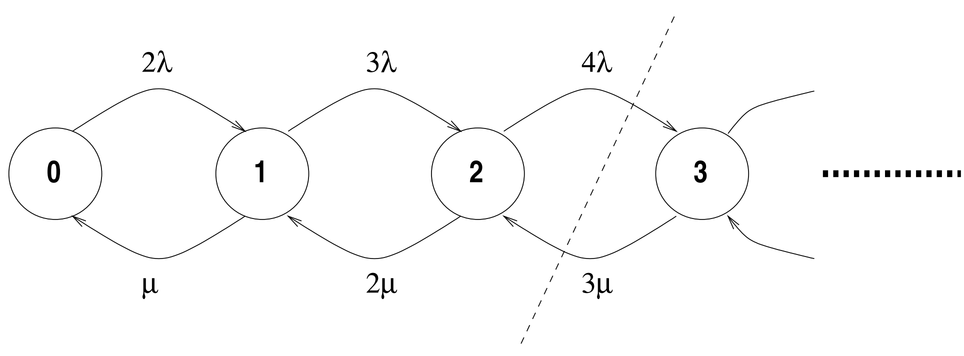
**1998 Paper 8 Q3** Define the term *Markov Chain*. Why is the Markov property useful in modelling queueing systems? [5 marks]

Consider a birth-death queueing system with the following birth and death coefficients in which the state index represents the number of customers in the system:

$$\begin{array}{ll} \lambda_k = (k+2)\lambda & k = 0, 1, 2 \ldots \\ \mu_k = k\mu & k = 1, 2 \ldots \end{array}$$

All other coefficients are zero. Solve for $p_k$, the set of equilibrium probabilities for all states $k$, for $k = 0, 1, 2 \ldots$ State how you would find the average number of customers in the system. [15 marks]

A set of random variables $X_n$ forms a Markov chain if the probability that the next state takes some particular value depends only on the current state. Consequently, in a continuous-time Markov chain, the state residence time must be exponentially distributed. The state space is discrete. The Markov property simplifies the mathematical model of the system.



Consider a cut between two parts of the Markov chain – shown above between states 2 and 3, but the generalization should be clear since the arrivals and departure rates are straightforward. For equilibrium we require $\lambda < \mu$ (even if it is only slightly less than $\mu$ note that this means that, from a certain point on the chain onwards, the arrival rate across the cut must be less than the departure rate).

Let $p_k$ be the probability of being in state $k$ when the system reaches equilibrium. Define $\rho = \lambda/\mu$. For $k \geq 1$:

$$
\begin{aligned}
k\mu p_k &= (k+1)\lambda p_{k-1} \\
p_k &= \frac{(k+1)\lambda}{k\mu} p_{k-1} \\
&= \frac{(k+1)!\,\lambda^k}{k!\,\mu^k} p_0 \\
&= (k+1)\rho^k p_0
\end{aligned}
$$

Note how, in fact, this result also holds for $k = 0$. As usual, eliminate $p_0$ by the fact that the probabilities sum to 1:

$$
\begin{aligned}
p_0 \sum_{i=0}^{\infty} (k+1)\rho^k &= 1 \\
p_0 \sum_{i=0}^{\infty} k\rho^{k-1} &= 1 \\
p_0 \sum_{i=0}^{\infty} \frac{\partial}{\partial\rho}\rho^k &= 1 \\
p_0 \frac{\partial}{\partial\rho} \sum_{i=0}^{\infty} \rho^k &= 1 \\
p_0 \frac{\partial}{\partial\rho}\left(\frac{1}{1-\rho}\right) &= 1 \\
p_0 &= (1-\rho)^2 \\
\Rightarrow \qquad p_k &= (k+1)\rho^k(1-\rho)^2
\end{aligned}
$$

The average number of customers in the system could be found as usual by

$$
\sum_{i=0}^{\infty} kp_k
$$