Performance Analysis

Peter Harrison & Giuliano Casale
Imperial College London

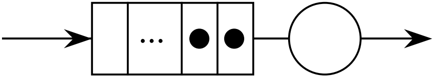
Spring term 2012

Course details

- Lectures: Wednesdays 9:00–11:00, in 144;
  Tutorials: Wednesdays 11:00–12:00, in 144.
- 18 lectures, 9 tutorials, 2 pieces of assessed coursework.
- Lecture notes on CATE: **you are responsible for printing them yourselves**
- Books:
  - *Performance modelling of communication networks and computer architectures* Peter G. Harrison, Naresh M. Patel, Addison-Wesley, 1992 ISBN 0201544199, 15 copies in library, but out of print.
  - *Probabilistic Modelling* I. Mitrani, Cambridge University Press ISBN 0521585309

Example 1: A simple transaction processing (TP) server

A transaction processing (TP) system accepts and processes a stream of transactions, mediated through a (large) buffer:

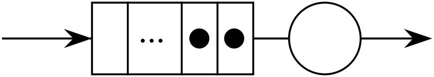


- Transactions arrive "randomly" at some specified rate
- The TP server is capable of servicing transactions at a given service *rate*

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

Example 2: A simple TP server

Consider the same system as above:

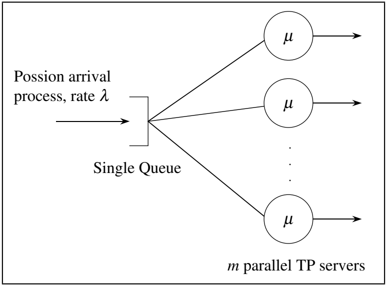


- The arrival rate is 15tps
- The mean service time per transaction is 58.37ms

Q: What happens to the mean response time if the arrival rate increases by 10%?

Example 3: A simple multiprocessor TP system

Consider our TP system but this time with multiple transaction processors

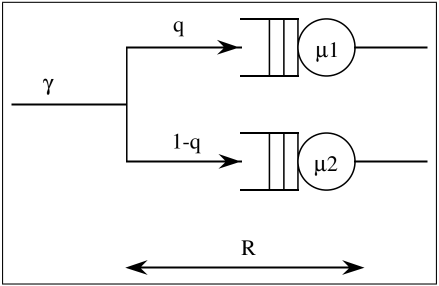


- The arrival rate is 16.5 tps
- The mean service time per transaction is 58.37ms

Q: By how much is the system response time reduced by adding one processor?

Example 4: File allocation

What is the best way to allocate disk blocks to a heterogenous disk I/O system?

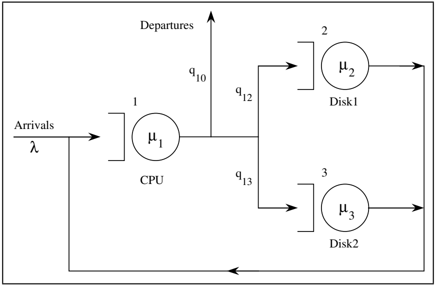


- Disk I/O requests are made at an average rate of 20 per second
- Disk blocks can be located on either disk and the mean disk access times are 30ms and 46ms respectively

Q: What is the optimal proportion of blocks to allocate to disk 1 to minimise average response time?

Example 5: A simple computer model

Consider an open uniprocessor CPU system with just disks

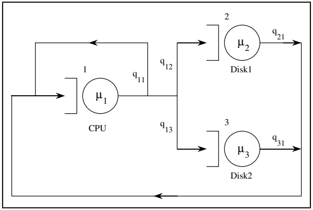


- Each submitted job makes 121 visits to the CPU, 70 to disk 1 and 50 to disk 2 *on average*
- The mean service times are 5ms for the CPU, 30ms for disk 1 and 37ms for disk 2

Q: What is the effect of replacing the CPU with one *twice* the speed?

Example 6: A Simple Batch Processor System

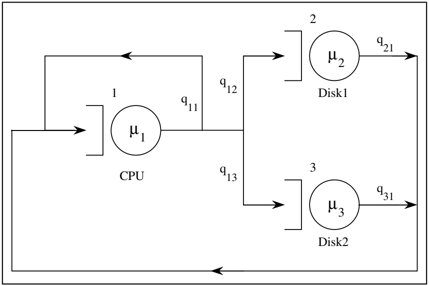
How does the above system perform in batch mode?



- Each batch job makes 121 visits to the CPU, 70 to disk 1, 50 to disk 2 *on average*
- The mean service times are 5ms for the CPU, 30ms for disk 1 and 37ms for disk 2

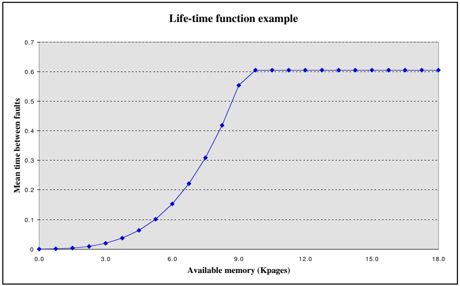
Q: How does the system throughput vary with the number of batch jobs and what is the effect of replacing Disk 1 with one (a) twice and (b) three times the speed?

Example 7: A multiprogramming system with virtual memory



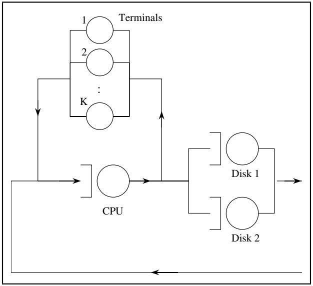
- Suppose now we add VM and make disk 1 a dedicated paging device
- Pages are 1Kbyte in size and the (usable) memory is equivalent to 64K pages

- Each job page faults at a rate determined by the following lifetime function:



Q: What number of batch jobs keeps the system throughput at its maximum and at what point does thrashing occur?

Example 8: A multiaccess multiprogramming system with virtual memory



- During the day the system runs in interactive mode with a number of terminal users
- The average think time of each user is 30 seconds

Q: How does the system response time and throughput vary with the number of terminals and how many terminals can be supported before the system starts to thrash?

Introduction

Computer systems are

1. **dynamic** – they can pass through a succession of states as time progresses
2. influenced by events which we consider here as **random phenomena**

We also see these characteristics in queues of customers in a bank or supermarket, or prices on the stock exchange.

Definition of a stochastic process

Definition

A **stochastic process S** is a family of random variables $\{X_t \in \Omega | t \in T\}$, each defined on some **sample space** Ω (the same for each) for a *parameter space* T.

- T and Ω may be either discrete or continuous
- T is normally regarded as time
  - real time: continuous
  - every month or after job completion: discrete
- Ω is the set of values each X_t may take
  - bank balance: discrete
  - number of active tasks: discrete
  - time delay in communication network: continuous

Example: The Poisson process

The Poisson process is a *renewal process* with *renewal period* (interarrival time) having cumulative distribution function F and probability density function (pdf) f

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$$
$$f(x) = F'(x) = \lambda e^{-\lambda x}$$

λ is the parameter or *rate* of the Poisson process.

Memoryless property of the (negative) exponential distribution

If S is an exponential random variable

$$P(S \leq t + s | S > t) = P(S \leq s) \quad \forall t, s \geq 0$$

(i.e. it doesn't matter what happened before time t)

Proof.

$$
\begin{aligned}
P(S \le t+s | S > t) &= \frac{P(t < S \le t+s)}{P(S > t)} \\
&= \frac{P(S \le t+s) - P(S \le t)}{1 - P(S \le t)} \\
&= \frac{1 - e^{-\lambda(t+s)} - (1 - e^{-\lambda t})}{e^{-\lambda t}} \\
&\qquad (\lambda \text{ is the rate of the exp. distribution}) \\
&= \frac{e^{-\lambda t} - e^{-\lambda(t+s)}}{e^{-\lambda t}} = 1 - e^{-\lambda s} \\
&= P(S \le s)
\end{aligned}
$$

$\square$

Residual life

- If you pick a random time point during a renewal process, what is the time remaining R to the next renewal instant (arrival)?
- e.g. when you get to a bus stop, how long will you have to wait for the next bus?
- If the renewal process is Poisson, R has the same distribution as S by the memoryless property
- This means *it doesn't matter when the last bus went!* (contrast this against constant interarrival times in a perfectly regular bus service)

"Infinitesimal definition" of the Poisson process

$$\begin{aligned} P(\text{arrival in } (t, t+h)) &= P(R \le h) = P(S \le h) \quad \forall t \\ &= 1 - e^{-\lambda h} \\ &= \lambda h + o(h) \end{aligned}$$

Therefore

1. Probability of an arrival in $(t, t+h)$ is $\lambda h + o(h)$ regardless of process history before t
2. Probability of more than one arrival in $(t, t+h)$ is $o(h)$ regardless of process history before t

- In fact we can take this result as an alternative *definition* of the Poisson process.
- From it we can derive the distribution function of the interarrival times (i.e. negative exponential) and the Poisson distribution for N_t (the *number of arrivals in time* t)

$$P(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

- *Assuming* this result, interarrival time distribution is

$$\begin{aligned} P(S \leq t) &= 1 - P(0 \text{ arrivals in } (0, t)) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

Derivation of the interarrival time distribution

$$\begin{aligned} P(S > t + h) &= P\big((S > t) \wedge (\text{no arrival in } (t, t+h])\big) \\ &= P(S > t)P(\text{no arrival in } (t, t+h]) \end{aligned}$$

by the memoryless property. Let $G(t) = P(S > t)$. Then:

$$\begin{aligned} G(t + h) &= G(t)P(\text{no arrival in } (t, t+h]) \\ &= (1 - h\lambda)G(t) + o(h) \end{aligned}$$

and so

$$\frac{G(t+h) - G(t)}{h} = -\lambda G(t) + o(1)$$

giving

$$\frac{\mathrm{d}G}{\mathrm{d}t} = -\lambda G(t) \quad \Rightarrow \quad G(t) = ke^{-\lambda t}$$

for constant k. Thus $F(t) = P(S \le t) = 1 - G(t) = 1 - ke^{-\lambda t}$ so $k = 1$ because we know $F(0) = 0$.

Superposition Property

If $A_1, \ldots, A_n$ are independent Poisson Processes with rates $\lambda_1, \ldots, \lambda_n$ respectively, let there be K_i arrivals in an interval of length t from process A_i $(1 \leq i \leq n)$. Then $K = K_1 + \cdots + K_n$ has Poisson distribution with parameter λt where $\lambda = \lambda_1 + \cdots + \lambda_n$. i.e. the superposition of PPs with rates λ_i is a PP with rate $\sum_i \lambda_i$.

Proof.

The distribution of K is the convolution of the distributions of the K_i which are Poisson. E.g. if $n = 2$

$$
\begin{aligned}
P(K = k) &= \sum_{i=0}^{k} \frac{(\lambda_1 t)^i}{i!} e^{-\lambda_1 t} \frac{(\lambda_2 t)^{k-i}}{(k-i)!} e^{-\lambda_2 t} \\
&= \frac{e^{-(\lambda_1+\lambda_2)t}}{k!} \sum_{i=0}^{k} \binom{k}{i} (\lambda_1 t)^i (\lambda_2 t)^{k-i} \\
&= e^{-(\lambda_1+\lambda_2)t} \frac{[(\lambda_1 + \lambda_2)t]^k}{k!}
\end{aligned}
$$

as required. The proof for arbitrary $n \geq 2$ is an easy induction on n. □

Decomposition Property

If a Poisson Process is decomposed into processes $B_1, \ldots, B_n$ by assigning each arrival of A to B_i with independent probability q_i ($\sum q_i = 1$), then $B_1, \ldots, B_n$ are independent Poisson Processes with rates $q_1\lambda, \ldots, q_n\lambda$.

Example: Two parallel processors, B_1 and B_2. Incoming jobs (Poisson arrivals A) are directed to B_1 with probability q_1 and to B_2 with probability q_2. Then each processor "sees" a Poisson arrival process for its incoming job stream.

Proof.

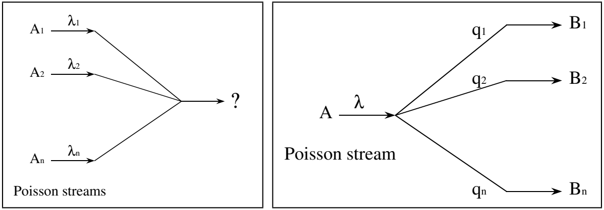
For $n = 2$, let K_i = number of arrivals to B_i in time t $(i = 1, 2)$, $K = K_1 + K_2$

$$\underbrace{P(K_1 = k_1, K_2 = k_2)}_{\text{joint probability}} = \underbrace{P(K_1 = k_1, K_2 = k_2 | K = k_1 + k_2)}_{\text{binomial}} \quad \underbrace{P(K = k_1 + k_2)}_{\text{Poisson}}$$

$$= \frac{(k_1 + k_2)!}{k_1! k_2!} q_1^{k_1} q_2^{k_2} \frac{(\lambda t)^{k_1 + k_2}}{(k_1 + k_2)!} e^{-\lambda t}$$

$$= \frac{(q_1 \lambda t)^{k_1}}{k_1!} e^{-q_1 \lambda t} \frac{(q_2 \lambda t)^{k_2}}{k_2!} e^{-q_2 \lambda t}$$

i.e. a product of two independent Poisson distributions ($n \geq 2$: easy induction)



Markov chains and Markov processes

- Special class of stochastic processes that satisfy the Markov property (MP) where given the state of the process at time t, its state at time $t + s$ has probability distribution which is a function of s only.
- i.e. the future behaviour after t is independent of the behaviour before t
- Often intuitively reasonable, yet sufficiently "special" to facilitate effective mathematical analysis
- We consider processes with discrete state (sample) space and:

  1. discrete parameter space (times $\{t_0, t_1, \ldots\}$), a *Markov chain* or *Discrete Time Markov Chain*
  2. continuous parameter space (times $t \geq 0, t \in R$), a *Markov process* or *Continuous Time Markov Chain*. E.g. number of arrivals in $(0, t)$ from a Poisson arrival process defines a Markov Process because of the memoryless property.

Markov chains

- Let $X = \{X_n | n = 0, 1, \dots\}$ be an integer valued Markov chain (MC), $X_i \geq 0, X_i \in Z, i \geq 0$. The Markov property states that:

$$P(X_{n+1} = j | X_0 = x_0, \dots, X_n = x_n) = P(X_{n+1} = j | X_n = x_n)$$

$$\text{for } j, n = 0, 1, \dots$$

- Evolution of an MC is completely described by its 1-step transition probabilities

$$q_{ij}(n) = P(X_{n+1} = j | X_n = i) \text{ for } i, j, n, \geq 0$$

Assumption: $q_{ij}(n) = q_{ij}$ is independent of time n (*time homogeneous*)

$$\sum_{j \in \Omega} q_{ij} = 1 \quad \forall i \in \Omega$$

- MC defines a transition probability matrix

$$Q = \begin{bmatrix} q_{00} & q_{01} & \cdots \\ q_{10} & q_{11} & \cdots \\ \vdots & \vdots & \ddots \\ q_{i0} & \cdots & q_{ii} \cdots \\ \vdots & & \end{bmatrix} \text{ in which all rows sum to 1}$$

- dimension $=$ # of states in Ω if finite, otherwise countably infinite
- conversely, any real matrix Q s.t. $q_{ij} \geq 0$, $\sum_j q_{ij} = 1$ (called a *stochastic matrix*) defines a MC

MC example 1: Telephone line

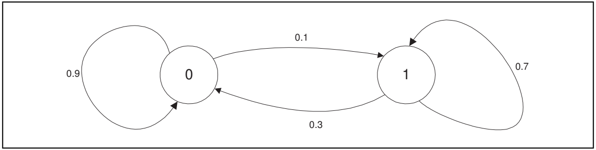
The line is either idle (state 0) or busy (state 1).

$$\begin{aligned}
\text{idle at time } n &\implies \text{idle at time } (n+1) \text{ w.p. } 0.9\\
\text{idle at time } n &\implies \text{busy at time } (n+1) \text{ w.p. } 0.1\\
\text{busy at time } n &\implies \text{idle at time } (n+1) \text{ w.p. } 0.3\\
\text{busy at time } n &\implies \text{busy at time } (n+1) \text{ w.p. } 0.7
\end{aligned}$$

so

$$Q = \begin{bmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{bmatrix}$$

may be represented by a state diagram:



MC example 2: Gambler

Bets £1 at a time on the toss of fair die. Loses if number is less than 5, wins £2 if 5 or 6. Stops when broke or holds £100. If X_0 is initial capital ($0 \leq X_0 \leq 100$) and X_n is the capital held after n tosses $\{X_n | n \leq 0, 1, 2, \dots\}$ is a MC with $q_{ii} = 1$ if $i = 0$ or $i \geq 100$, $q_{i,i-1} = \frac{2}{3}$, $q_{i,i+2} = \frac{1}{3}$ for $i = 1, 2, \dots, 99$ $q_{ij} = 0$ otherwise.

MC example 3: I/O buffer with capacity M records

New record added in any unit of time w.p. α (if not full). Buffer emptied in any unit of time w.p β. If both occur in same interval, insertion done first. Let X_n be the number of records in buffer at (discrete) time n. Then, assuming that insertions and emptying are independent of each other and of their own past histories, $\{X_n | n = 0, 1, \dots\}$ is a MC with state space $\{0, 1, \dots, M\}$ and state diagram:

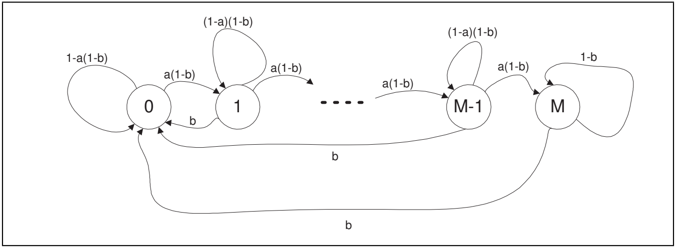


Figure: I/O buffer example

MC example 3: I/O buffer with capacity M records

The transition matrix follows immediately, e.g.

$$\begin{aligned} q_{12} &= \alpha(1-\beta) = q_{n,n+1} \qquad (0 \le n \le M-1) \\ q_{MM} &= 1-\beta \end{aligned}$$

Markov chain two-step transition probabilities

Let

$$
\begin{aligned}
q_{ij}^{(2)} &= P(X_{n+2} = j | X_n = i) \\
&= \sum_{k \in \Omega} P(X_{n+1} = k, X_{n+2} = j | X_n = i) \text{ from law of tot. prob.} \\
&= \sum_{k \in \Omega} P(X_{n+2} = j | X_n = i, X_{n+1} = k) P(X_{n+1} = k | X_n = i) \\
&= \sum_{k \in \Omega} P(X_{n+2} = j | X_{n+1} = k) P(X_{n+1} = k | X_n = i) \text{ by MP} \\
&= \sum_{k \in \Omega} q_{ik} q_{kj} \text{ by TH} \\
&= (Q^2)_{ij}
\end{aligned}
$$

Markov chain multi-step transition probabilities

Similarly, m-step transition probabilities :

$$\begin{aligned} q_{ij}^{(m)} &= P(X_{n+m} = j | X_n = i) \quad (m \geq 1) \\ &= (Q^m)_{ij} \end{aligned}$$

by induction on m. Therefore we can compute probabilistic behaviour of a MC over any finite period of time, in principle. E.g. average no. of records in buffer at time 50

$$E[X_{50} | X_0 = 0] = \sum_{j=1}^{\min(M,50)} j q_{0j}^{(50)}.$$

Long term behaviour

Markov chain multi-step transition probabilities can be computationally intractable. We wish to determine long-term behaviour: hope that asymptotically MC approaches a steady-state (probabilistically), i.e. that $\exists \{p_j | j = 0, 1, 2, \ldots\}$ s.t.

$$p_j = \lim_{n \to \infty} P(X_n = j | X_0 = i) = \lim_{n \to \infty} q_{ij}^{(n)}$$

independent of i.

Definitions and properties of MCs

Let v_{ij} = prob. that state j will eventually be entered, given that the MC has been in state i, i.e.

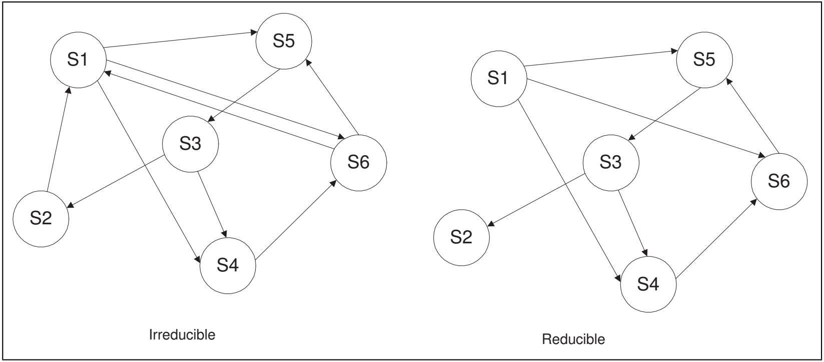
$$v_{ij} = P(X_n = j \text{ for some } n | X_0 = i)$$

- If $v_{ij} \neq 0$, then j is *reachable* from i
- If $v_{ij} = 0 \quad \forall j \neq i, v_{ii} = 1$, then i is an *absorbing* state

Example (1) & (3): all states reachable from each other (e.g. look at transition diagrams)
Example (2): all states in $\{1, 2, \ldots, 99\}$ reachable from each other, 0 & 100 absorbing

Closed and irreducible MCs

A subset of states, C is called *closed* if $j \notin C$, implies j cannot be reached from any $i \in C$. E.g. set of all states, Ω, is closed. If $\nexists$ proper subset $C \subset \Omega$ which is closed, the MC is called *irreducible*. A Markov Chain is irreducible if and only if every state is reachable from every other state. E.g. examples (1) & (3) and $\{0\}$, $\{100\}$ are closed in example (2).



Recurrent states

If $v_{ii} = 1$, state i is *recurrent* (once entered, guaranteed eventually to return). Thus we have:
i recurrent implies i is either not visited by the MC or is visited an infinite number of times – no visit can be the last (with non-zero probability).
If i is not recurrent it is called *transient*. This implies the number of visits to a transient state is finite w.p. 1 (has geometric distribution). E.g. if C is a closed set of states, $i \notin C$, $j \in C$ and $v_{ij} \neq 0$, then i is transient since MC will eventually enter C from i, never to return. E.g. in example (2), states 1 to 99 are transient, states 0 and 100 are recurrent (as is any absorbing state)

Proposition

If i is recurrent and j is reachable from i, then j is also recurrent.

Proof.

$v_{jj} \neq 1$ implies $v_{ij} \neq 1$ since if the MC visits i after j it will visit i repeatedly and with probability 1 will eventually visit j since $v_{ij} \neq 0$. This implies $v_{ii} \neq 0$ since with non-zero probability the MC can visit j after i but not return to i (w.p. $1 - v_{ji} > 0$). $\square$

Thus, in an irreducible MC, either all states are transient (a transient MC) or recurrent (a recurrent MC).

Further properties of MCs

Let X be an irreducible, recurrent MC and let $n_1^j, n_2^j, \ldots$ be the times of successive visits to state j, $m_j = E[n_{k+1}^j - n_k^j]$ $(k = 1, 2, \ldots)$, the mean interval between visits (independent of k by the MP).

<u>Intuition</u>:

$$p_j = \frac{1}{m_j}$$

Because the proportion of time spent, on average, in state j is $1/m_j$.

This is not necessarily true, because the chain may exhibit some *periodic* behaviour:

The state j is *periodic* with period $m > 1$ if $q_{ii}^{(k)} = 0$ for $k \neq rm$ for any $r \geq 1$ and $P(X_{n+rm} = j \text{ for some } r \geq 1 | X_n = j) = 1$.

Otherwise it is aperiodic, or has period 1. (Note that a periodic state is recurrent)

Periodicity in an irreducible MC

Proposition

In an irreducible MC, either all states are aperiodic or periodic with the same period. The MC is then called aperiodic or periodic respectively.

Proposition

If $\{X_n | n = 0, 1, \ldots\}$ *is an irreducible, aperiodic MC, then the limiting probabilities* $\{p_j | j = 0, 1, \ldots\}$ *exist and* $p_j = 1/m_j$.

Null and positive recurrence

If $m_j = \infty$ for state j then $p_j = 0$ and state j is *recurrent null*. Conversely, if $m_j < \infty$, then $p_j > 0$ and state j is *recurrent non-null* or *positive recurrent*.

Proposition

In an irreducible MC, either all states are positive recurrent or none are. In the former case, the MC is called positive recurrent (or recurrent non-null).

A state which is aperiodic and positive recurrent is often called *ergodic* and an irreducible MC whose states are all ergodic is also called ergodic. When the limiting probabilities $\{p_j | j = 0, 1, \dots\}$ do exist they form the

$$\left.\begin{array}{l}\text{steady state}\\ \text{equilibrium}\\ \text{long term}\end{array}\right\}\text{probability distribution (SSPD)}$$

of the MC. They are determined by the following theorem.

Steady state theorem for Markov chains

Theorem

An irreducible, aperiodic Markov Chain, X, with state space S and one-step transition probability matrix $Q = (q_{ij}|i, j \in S)$, is positive recurrent if and only if the system of equations

$$p_j = \sum_{i \in S} p_i q_{ij}$$

with $\sum_{i \in S} p_i = 1$ (normalisation) has a solution. If it exists, the solution is unique and is the SSPD of X.

Note

If S is finite, X is always positive recurrent, which implies the equations have a unique solution. The solution is then found by discarding a balance equation and replacing it with the normalising equation (the balance equations are dependent since the rows of homogeneous equations $\mathbf{p} - \mathbf{p}Q = \mathbf{0}$ all sum to zero).

MC example 3: Discrete time buffer

<u>$1 \le i \le M-1$</u>
From the steady state theorem:

$$p_i = \alpha(1-\beta)p_{i-1} + p_i(1-\alpha)(1-\beta)$$

i.e.

$$(\alpha(1-\beta)+\beta)p_i = \alpha(1-\beta)p_{i-1}$$

which is equivalent to the heuristic "flow balance" view (cf. CTMCs later):

Probability of leaving state i = Probability of entering state i

Therefore

$$p_i = kp_{i-1} = k^i p_0$$

where $k = \frac{\alpha(1-\beta)}{\alpha+\beta-\alpha\beta}$ $(1 \le i \le M-1)$.

Discrete time buffer state transition diagram

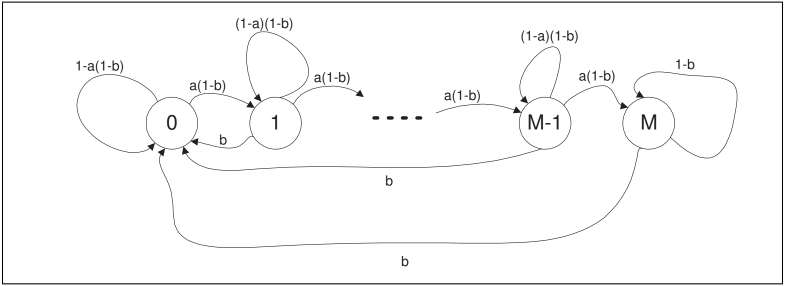


Figure: I/O buffer.

$\underline{i = M}$

$$\beta p_M = \alpha(1-\beta)p_{M-1}$$

therefore

$$p_M = \frac{\alpha(1-\beta)}{\beta} k^{M-1} p_0$$

$\underline{i = 0}$

Redundant equation (why?) – use as check

$$\begin{aligned}\alpha(1-\beta)p_0 = \sum_{i=1}^{M} \beta p_i &= \beta \sum_{i=1}^{M-1} k^i p_0 + k^{M-1}\alpha(1-\beta)p_0 \\ &= p_0\Big(\beta k \frac{1-k^{M-1}}{1-k} + \alpha(1-\beta)k^{M-1}\Big)\end{aligned}$$

rest is exercise

Markov processes

- Continuous time parameter space, discrete state space

$$X = \{X_t | t \geq 0\} \quad X_t \in \Omega$$

where Ω is a countable set.

- Markov property

$$P(X_{t+s} = j | X_u, u \leq t) = P(X_{t+s} = j | X_t)$$

- Markov process is *time homogeneous* if the r.h.s. of this equation does not depend on t
  - $q_{ij}(s) = P(X_{t+s} = j | X_t = i) = P(X_s = j | X_0 = i)$ $(i, j = 0, 1, \dots)$
  - *Transition probability functions* of the MP

Memoryless property

Markov Property and time homogeneity imply:
If at time t the process is in state j, the time remaining in state j is independent of the time already spent in state j: *memoryless property*

Proof.

$$
\begin{aligned}
P(S > t+s | S > t) &= P(X_{t+u} = j, 0 \le u \le s | X_u = j, 0 \le u \le t) \\
&\quad \text{where S} = \text{time spent in state } j, \\
&\quad \text{state } j \text{ entered at time } 0 \\
&= P(X_{t+u} = j, 0 \le u \le s | X_t = j) \text{ by MP} \\
&= P(X_u = j, 0 \le u \le s | X_0 = j) \text{ by T.H.} \\
&= P(S > s) \\
\implies P(S \le t+s | S > t) &= P(S \le s)
\end{aligned}
$$

$\square$

Time spent in state i is exponentially distributed, parameter μ_i.

Time homogeneous Markov Processes

The MP implies next state, j, after current state i depends only on i, j – state transition probability q_{ij}. This implies the Markov Process is uniquely determined by the products:

$$a_{ij} = \mu_i q_{ij}$$

where μ_i is the rate out of state i and q_{ij} is the probability of selecting state j next. The a_{ij} are the *generators* of the Markov Process.

- intuitively reasonable

Instantaneous transition rates

Consider now the (small) interval $(t, t+h)$

$$\begin{aligned} P(X_{t+h} = j | X_t = i) &= \mu_i h q_{ij} + o(h) \\ &= a_{ij} h + o(h) \end{aligned}$$

- a_{ij} is the *instantaneous transition rate* $i \rightarrow j$ i.e. the average number of transitions $i \rightarrow j$ per unit time spent in state i

MP example: a Poisson process

$$a_{ij} = \begin{cases} \lambda & \text{if } j = i+1 \\ 0 & \text{if } j \neq i, i+1 \\ & \text{not defined if } j = i \end{cases}$$

because

$$\underbrace{P(\text{arrival in } (t, t+h))}_{i \to i+1} = \underbrace{\lambda h}_{a_{i,i+1} h} + o(h)$$

MP example: The buffer problem

- Records arrive as a P.P. rate λ
- Buffer capacity is M
- Buffer cleared at times spaced by intervals which are exponentially distributed, parameter μ. Clearance times i.i.d. and independent of arrivals (i.e. clearances are an independent PP, rate μ)

$$a_{ij} = \begin{cases} \lambda \text{ if } j = i+1 \quad (0 \leq i \leq M-1) \\ \mu \text{ if } j = 0 \quad (i \neq 0) \\ 0 \text{ otherwise } (j \neq i) \end{cases}$$

[Probability of > 1 arrivals or clearances, or ≥ 1 arrivals <u>and</u> clearances in $(t, t+h)$ is $o(h)$]

MP example: The buffer problem state transition diagram

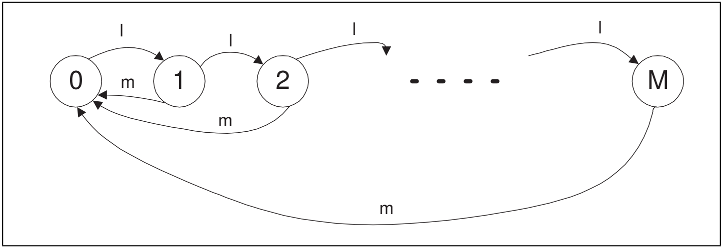
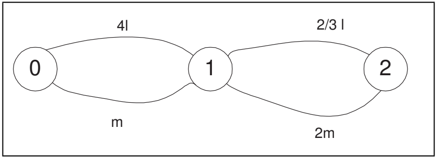


Figure: Buffer problem

MP example: Telephone exchange

Suppose there 4 subscribers and any two calls between different callers can be supported simultaneously. Calls are made by non-connected subscribers according to independent PPs, rate λ (callee chosen at random). Length of a call is (independent) exponentially distributed, parameter μ. Calls are lost if called subscriber is engaged. State is the number of calls in progress (0,1 or 2).

$$a_{01} = 4\lambda \quad \text{(all free)}$$
$$a_{12} = \frac{2}{3}\lambda \quad \text{(caller has } \frac{1}{3} \text{ probability of successful connection)}$$
$$a_{10} = \mu$$
$$a_{21} = 2\mu \quad \text{(either call may end)}$$



Exercise

The last equation follows because the distribution of the random variable $Z = \min(X, Y)$ where X, Y are independent exponential random variables with parameters λ, μ respectively is exponential with parameter $\lambda + \mu$. Prove this.

The generator matrix

$$\sum_{j \in \Omega, j \neq i} q_{ij} = 1 \implies \mu_i = \sum_{j \in \Omega, j \neq i} a_{ij}$$

(hence *instantaneous transition rate out of* i). Let $a_{ii} = -\mu_i$ (undefined so far)

$$(a_{ij}) = A = \begin{pmatrix} a_{00} & a_{01} & \dots & & \\ a_{10} & a_{11} & \dots & & \\ \dots & & & & \\ a_{i0} & a_{i1} & \dots & a_{ii} & \dots \\ \dots & & & & \end{pmatrix}$$

A is the generator matrix of the Markov Process. Rows of A sum to zero ($\sum_{j \in S} a_{ij} = 0$)

Transition probabilities and rates

- determined by A

$$q_{ij} = \frac{a_{ij}}{\mu_i} = -\frac{a_{ij}}{a_{ii}}$$
$$\mu_i = -a_{ii}$$

- hence A is all we need to determine the Markov process
- Markov processes are instances of state transition systems with the following properties
  - The state holding time are *exponentially* distributed
  - The probability of transiting from state i to state j depends only on i and j
- In order for the analysis to work we require that the process in **irreducible** – every state must be reachable from every other, e.g.

Steady state results

Theorem

(a) *if a Markov process is transient or recurrent null,* $p_j = 0 \quad \forall j \in S$ *and we say a SSPD does not exist*

(b) *If a Markov Process is positive recurrent, the limits* p_j *exist,* $p_j > 0, \sum_{j \in S} p_j = 1$ *and we say* $\{p_j | j \in S\}$ *constitute the SSPD of the Markov Process.*

Steady state theorem for Markov processes

Theorem

An irreducible Markov Process X with state space S and generator matrix $A = (a_{ij})$ $(i, j \in S)$ is positive recurrent if and only if

$$\sum_{i \in S} p_i a_{ij} = 0 \qquad \text{for } j \in S \text{ [Balance equations]}$$

$$\sum_{i \in S} p_i = 1 \qquad \text{[Normalising equation]}$$

have a solution. This solution is unique and is the SSPD.

Note

At equilibrium the fluxes also balance into and out of every closed contour *drawn around* any *collection of states.*

Justification of the balance equation

The rate going from state i to $j(\neq i)$ is a_{ij}, the fraction of time spent in i is p_i

$$\underbrace{\sum_{i \neq j} a_{ij} p_i}_{\text{Avg. no. of transitions } i \to j \text{ in unit time}} = \underbrace{\sum_{i \neq j} a_{ji} p_j}_{\text{Avg. no. of transitions } j \to i \text{ in unit time}}$$

$$\sum_{i \neq j} \text{flux}(i \to j) = \sum_{i \neq j} \text{flux}(j \to i) \quad \forall j$$

MP example: I/O buffer

By considering the flux into and out of each state $0, 1, \ldots, M$ we obtain the balance equations:

$$\begin{aligned} \lambda p_0 &= \mu(p_1 + \cdots + p_M) && \text{(State 0)} \\ (\lambda+\mu)p_i &= \lambda p_{i-1} && \text{(State } i\text{, for } 1 \le i \le M-1\text{)} \\ \mu p_M &= \lambda p_{M-1} && \text{(State } M\text{)} \end{aligned}$$

Normalising equation:

$$\begin{aligned} &p_0 + p_1 + \cdots + p_M = 1 \\ &\qquad \Longrightarrow \; p_j = \left(\frac{\lambda}{\lambda+\mu}\right)^j \frac{\mu}{\lambda+\mu} \quad \text{(for } 0 \le j \le M-1\text{)} \\ &\qquad\qquad p_M = \left(\frac{\lambda}{\lambda+\mu}\right)^M \end{aligned}$$

Thus (for example) mean number of records in the buffer in the steady state $= M\alpha^M + \sum_{j=0}^{M-1} j\alpha^j \frac{\mu}{\lambda+\mu}$ where $\alpha = \frac{\mu}{\lambda+\mu}$

MP example: Telephone network

$$4\lambda p_0 = \mu p_1 \qquad \text{(State 0)}$$

$$\left(\mu + \frac{2}{3}\lambda\right)p_1 = 4\lambda p_0 + 2\mu p_2 \qquad \text{(State 1)}$$

$$2\mu p_2 = \lambda \frac{2}{3} p_1 \qquad \text{(State 2)}$$

Thus $p_1 = \frac{4\lambda}{\mu}p_0$, $p_2 = \frac{\lambda}{3\mu}p_1 = \frac{4\lambda^2}{3\mu^2}p_0$ with

$$p_0 + p_1 + p_2 = 1 \implies p_0 = \left(1 + \frac{4\lambda}{\mu} + \frac{4\lambda^2}{3\mu^2}\right)^{-1}.$$

Average number of calls in progress in the steady state
$= 1.p_1 + 2.p_2 = \frac{\frac{4\lambda}{\mu}\left(1+\frac{2\lambda}{3\mu}\right)}{1+\frac{4\lambda}{\mu}+\frac{4\lambda^2}{3\mu^2}}$

Birth-death processes and the single server queue (SSQ)

A Markov process with state space $S = \{0, 1, \dots\}$ is called a birth-death process if the only non-zero transition probabilities are $a_{i,i+1}$ and $a_{i+1,i}$ ($i \geq 0$), representing births and deaths respectively. (In a population model, a_{00} would be 1 since 0 would be an absorbing state.) The SSQ model consists of

- a Poisson arrival process, rate λ
- a queue which arriving tasks join
- a server which processes tasks in the queue in FIFO (or other) order and has exponentially distributed service times, parameter μ (i.e. given a queue length > 0, service completions form a Poisson process, rate μ)

- The state is the queue length (including the task being served if any), i.e. the state space is $\{0, 1, \dots\}$
- SSQ model is a birth-death process
- λ, μ are in general functions of the queue length (i.e. state dependent) and we write $\lambda(n), \mu(n)$ for state n.

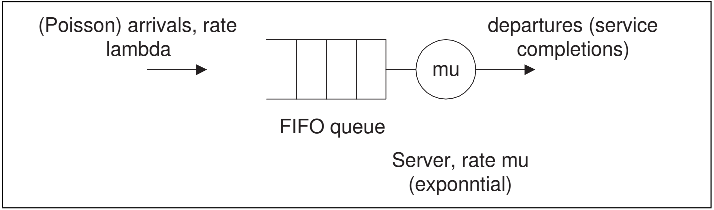


Figure: Single server queue.

Kendall's notation

Notation

The SSQ with Poisson arrivals and exponential service times is called an M/M/1 queue

- *the first M describes the arrival process as a Markov process (Poisson)*
- *the second M describes the service time distribution as a Markov process (exponential)*
- *the 1 refers to a single server (m parallel server would be denoted as M/M/m)*
- *Later we will consider M/G/1 queues, where the service time distribution is non-Markovian ("general")*

The memoryless property in the M/M/1 queue

SSQ therefore follows a Markov process and has the memoryless property that:

1. Probability of an arrival in $(t, t+h) = \lambda(i)h + o(h)$ in state i
2. Probability of a service completion in $(t, t+h) = \mu(i)h + o(h)$ in state $i > 0$ (0 if $i = 0$)
3. Probability of more than 1 arrival, more than one service completion or 1 arrival and 1 service completion in $(t, t+h) = o(h)$.
4. Form these properties we could derive a differential equation for the <u>transient</u> queue length probabilities – compare Poisson process.

State transition diagram for the SSQ

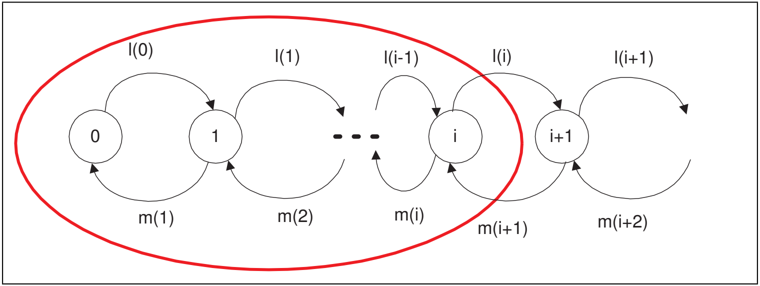


Figure: Single server queue state diagram.

- Consider the balance equation for states inside the red (thick) contour.
  - Outward flux (all from state i): $p_i\lambda(i) \quad (i \geq 0)$;
  - Inward flux (all from state $i+1$): $p_{i+1}\mu(i+1) \quad (i \geq 0)$.

Thus,

$$p_i\lambda(i) = p_{i+1}\mu(i+1)$$

so

$$p_{i+1} = \frac{\lambda(i)}{\mu(i+1)}p_i = \Big[\prod_{j=0}^{i}\rho(j)\Big]p_0$$

where $\rho(j) = \frac{\lambda(j)}{\mu(j+1)}$.

- Normalising equation implies

$$p_0\Big(1 + \sum_{i=0}^{\infty}\prod_{j=0}^{i}\rho(j)\Big) = 1$$

so

$$p_0 = \Big[\sum_{i=0}^{\infty}\prod_{j=0}^{i-1}\rho(j)\Big]^{-1}$$

where $\prod_{j=0}^{-1} = 1$ (the empty product). Therefore

$$p_i = \frac{\prod_{j=0}^{i-1}\rho(j)}{\sum_{k=0}^{\infty}\prod_{n=0}^{k-1}\rho(n)} \quad (i \geq 0).$$

- So, is there always a steady state?

- SSQ with constant arrival and service rates

$$\lambda(n) = \lambda, \quad \mu(n) = \mu, \quad \rho(n) = \rho = \lambda/\mu \quad \forall n \in S$$

implies

$$p_0 = \Big[\sum_{i=0}^{\infty} \rho^i\Big]^{-1} = 1 - \rho$$

$$p_i = (1-\rho)\rho^i \quad (i \geq 0)$$

- Mean queue length, L (including any task in service)

$$\begin{aligned} L &= \sum_{i=0}^{\infty} i p_i = \sum_{i=0}^{\infty} (1-\rho) i \rho^i \\ &= \rho(1-\rho)\frac{d}{d\rho}\Big\{\sum_{i=0}^{\infty} \rho^i\Big\} = \rho(1-\rho)\frac{d}{d\rho}\Big\{(1-\rho)^{-1}\Big\} \\ &= \frac{\rho}{1-\rho} \end{aligned}$$

- <u>Utilisation of server</u>

$$\begin{aligned} U &= 1 - P(\text{server idle}) = 1 - p_0 = 1 - (1-\rho) = \rho \\ &= \lambda/\mu. \end{aligned}$$

However, we could have deduced this without solving for p_0: In the steady state (assuming it exists),

$$\begin{aligned} \lambda &= \text{arrival rate} \\ &= \text{throughput} \\ &= P(\text{server busy}).\text{service rate} \\ &= U\mu \end{aligned}$$

This argument applies for any system in equilibrium – we didn't use the Markov property – see M/G/1 queue later.

Response times

To analyse response times, need to consider the state of the queue at the time of arrival of a task. We use the **Random Observer Property** of the Poisson process.
The state of a system at equilibrium seen by an arrival of a Poisson process has the same distribution as that seen by an observer at a random instant, i.e. if the state at time t is denoted by S_t,

$$P(S_{t_0^-} = i \mid \text{arrival at } t_0) = P(S_{t_0^-} = i)$$

If the queue length seen by an arriving task is j,

$$\begin{aligned} \text{response time} = {} & \text{residual service time of task in service (if } j > 0) \\ & + j \text{ i.i.d. service times} \end{aligned}$$

For exponential service times, residual service time has the same distribution as full service time, so in this case

$$\text{Response time} = \text{sum of } (j+1) \text{ i.i.d. service times.}$$

Therefore, the **mean response time**, W is

$$W = \sum p_j (j+1)\mu^{-1} = \Big(1 + \frac{\rho}{1-\rho}\Big)\mu^{-1} = \frac{1}{\mu - \lambda}$$

and **mean queueing time** W_Q (response time, excluding service time)

$$W_Q = W - \mu^{-1} = L\mu^{-1} = \frac{\rho}{\mu - \lambda}.$$

Distribution of the waiting time, $F_W(x)$

By the random observer property (and memoryless property)

$$F_W(x) = \sum_{j=0}^{\infty} p_j E_{j+1}(x)$$

where $E_{j+1}(x)$ is the convolution of $(j+1)$ exponential distributions, each with parameter μ — called the Erlang–$(j+1)$ distribution with parameter μ. Similarly, for density functions:

$$f_W(x) = \sum_{j=0}^{\infty} p_j e_{j+1}(x)$$

where $e_{j+1}(x)$ is the pdf corresponding to $E_{j+1}(x)$, i.e. $\frac{d}{dx} E_{j+1}(x) = e_{j+1}(x)$, defined by

$$e_1(x) = \mu e^{-\mu x}$$

$$e_{j+1}(x) = \underbrace{\mu \int_0^x e^{-\mu(x-u)} e_j(u) du}_{\text{convolution of Erlang–}j\text{ and exponential distributions}}$$

$$\implies e_j(x) = \mu \frac{(\mu x)^{j-1}}{(j-1)!} e^{-\mu x} \qquad \text{(Exercise)}$$

$$\implies f_W(x) = (\mu - \lambda) e^{-(\mu-\lambda)x} \qquad \text{(Exercise)}$$

These results can be obtained much more easily using Laplace transforms (which we will not detail here).

Example

Given m Poisson streams, each with rate λ and independent of the others, into a SSQ, service rate μ, what is the maximum value of m for which, in steady state, at least 95% of waiting times, $W \leq w$? (Relevant in "end-to-end" message delays in communication networks, for example.)
That is, we seek m such that $P(W \leq w) \geq 0.95$

$$\begin{aligned}
\Longrightarrow \qquad 1 - e^{-(\mu - m\lambda)w} &\geq 0.95 \\
\Longrightarrow \qquad e^{-(\mu - m\lambda)w} &\leq 0.05 \\
\Longrightarrow \qquad e^{(\mu - m\lambda)w} &\geq 20 \\
\Longrightarrow \qquad \mu - m\lambda &\geq \frac{\ln 20}{w} \\
\Longrightarrow \qquad m &\leq \frac{\mu}{\lambda} - \frac{\ln 20}{w\lambda}
\end{aligned}$$

Note

$m < \mu/\lambda$ is equivalent to the existence of a steady state.

Reversed processes

- The *reversed process* of a stochastic process is a dual process
  - with the same state space
  - in which the *direction of time is reversed*
  - cf. viewing a video film backwards.
- If the reversed process is *stochastically identical* to the original process, that process is called *reversible*

Detailed balance equations

- A reversible process satisfies — as a necessary and sufficient condition for reversibility — the *detailed balance equations*

$$\pi_i \, a_{ij} = \pi_j \, a_{ji} \quad \text{for all states } i \neq j \in S$$

  - $A = (a_{ij})$ is the process's generator matrix (transition rates $i \to j$)
  - $\boldsymbol{\pi} = (\pi_i \mid i \in S)$ is its equilibrium probability distribution vector
- Detailed balance equations simply say that the probability flux from state i to state j is equal to the probability flux from state j to state i for all states $i \neq j$

Example — the M/M/1 queue

Recall our derivation of steady state probabilities for the M/M/1 queue with state-dependent rates:

- Balancing probability flux into and out of the subset of states $\{0, 1, \ldots, i\}$ we found

$$\pi_i \; a_{i,i+1} = \pi_{i+1} \; a_{i+1,i}$$

- There are no other classes of directly connected states
- Therefore the M/M/1 queue is reversible – including M/M/m, a special case of state-dependent M/M/1

Burke's Theorem

Now consider the *departure process* of an M/M/1 queue

- It is precisely the arrival process in the reversed queue
  - remember, time is going backwards
  - so, state decrements (departures) become increments (arrivals) in the reversed process
- Since the reversed process is also an M/M/1 queue, its arrivals are Poisson and independent of the past behaviour of the queue
- Therefore the departure process of the (forward or reversed) M/M/1 queue is *Poisson and independent of the future state of the queue*
- Equivalently, the state of the queue at any time is independent of the departure process *before* that time

Reversed Processes

- Most Markov processes are not reversible but we can still define the *reversed process* X_{-t} of any Markov process X_t at equilibrium
- It is straightforward to find the reversed process of a Markov process if its steady state probabilities are known:
- The reversed process of a Markov process $\{X_t\}$ at equilibrium, with state space S, generator matrix A and steady state probabilities $\boldsymbol{\pi}$, is a Markov process with generator matrix A' defined by

  $$a'_{ij} = \pi_j a_{ji} / \pi_i \quad (i, j \in S)$$

  and with the same stationary probabilities $\boldsymbol{\pi}$
- So the flux from i to j in the *reversed process* is equal to the flux from j to i in the *forward process*, for all states $i \neq j$

Proof.

For $i \neq j$ and $h > 0$,

$$P(X_{t+h} = i)P(X_t = j | X_{t+h} = i) = P(X_t = j)P(X_{t+h} = i | X_t = j).$$

Thus,

$$P(X_t = j | X_{t+h} = i) = \frac{\pi_j}{\pi_i} P(X_{t+h} = i | X_t = j)$$

at equilibrium. Dividing by h and taking the limit $h \to 0$ yields the required equation for a'_{ij} when $i \neq j$. But, when $i = j$,

$$-a'_{ii} = \sum_{k \neq i} a'_{ik} = \sum_{k \neq i} \frac{\pi_k a_{ki}}{\pi_i} = \sum_{k \neq i} a_{ik} = -a_{ii}.$$

That $\boldsymbol{\pi}$ is also the stationary distribution of the reversed process now follows since

$$-\pi_i a'_{ii} = \pi_i \sum_{k \neq i} a_{ik} = \sum_{k \neq i} \pi_k a'_{ki}.$$

Why is this useful?

In an irreducible Markov process, we may:

- Choose a reference state 0 arbitrarily
- Find a sequence of directly connected states $0, \ldots, j$
- calculate

$$\pi_j = \pi_0 \prod_{i=0}^{j-1} \frac{a_{i,i+1}}{a'_{i+1,i}} = \pi_0 \prod_{i=0}^{j-1} \frac{a'_{i,i+1}}{a_{i+1,i}}$$

- So if we can determine the matrix Q' independently of the steady state probabilities $\boldsymbol{\pi}$, *there is no need to solve balance equations*.

RCAT, SPA and Networks

- A compositional result in Stochastic Process Algebra called RCAT (Reversed Compound Agent Theorem) finds many reversed processes and hence simple solutions for steady state probabilities
  - open and closed queueing networks
  - multiple classes of customers
  - ‘negative customers’ that ‘kill’ customers rather than add to them in a queue
  - batches and other synchronized processes
- Automatic or mechanised, unified implementation
- Hot research topic – see later in the course!

Multiple parallel servers — M/M/m queue

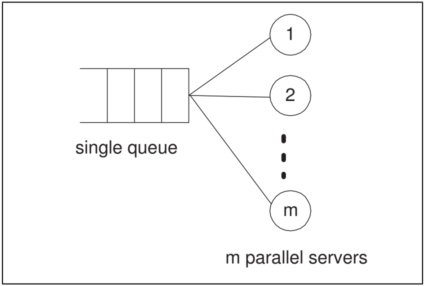


Figure: Multiple parallel servers.

M/M/m SSQ

SSQ representation:

$$\lambda(n) = \lambda \quad (n \geq 0)$$
$$\mu(n) = \begin{cases} n\mu & 1 \leq n \leq m \\ m\mu & n \geq m \end{cases}$$

The front of the queue is served by any available server.

By a general result for the M/M/1 queue:

$$p_j = p_0 \prod_{i=0}^{j-1} \frac{\lambda(i)}{\mu(i+1)} = \begin{cases} p_0 \frac{\rho^j}{j!} & 0 \leq j \leq m \\ p_0 \frac{\rho^j}{m! m^{j-m}} & j \geq m \end{cases}$$

so

$$p_0 = \frac{1}{\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \sum_{i=m}^{\infty} (\frac{\rho}{m})^{i-m}} = \frac{1}{\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{(m-1)!(m-\rho)}}.$$

Average number of busy servers, S

$$S = \sum_{k=1}^{m-1} k p_k + m \sum_{k=m}^{\infty} p_k = \cdots = \rho$$

Steady state argument

$$\begin{aligned} \text{arrival rate} &= \lambda \\ \text{average throughput} &= S\mu && (\mu \text{ per active server}) \\ \implies S\mu &= \lambda && (\text{in equilibrium}). \end{aligned}$$

Utilisation $U = S/m = \rho/m$, the average fraction of busy servers.

Waiting times

Waiting time is the same as service time if an arrival does not have to queue. Otherwise, the departure process is Poisson, rate $m\mu$, whilst the arrived task is queueing (all servers busy). This implies that the queueing time is the same as the queueing time in the M/M/1 queue with service rate $m\mu$.

- Let

$$\begin{aligned} q &= P(\text{arrival has to queue}) \\ &= P(\text{find} \geq m \text{ jobs on arrival}) \end{aligned}$$

- by R.O.P.

$$q = \sum_{j=m}^{\infty} p_j = p_0 \frac{\rho^m}{(m-1)!(m-\rho)} \quad \text{Erlang delay formula}$$

- Let Q be the queueing time random variable (excluding service)

$$
\begin{aligned}
F_Q(x) = P(Q \leq x) &= P(Q = 0) + P(Q \leq x | Q > 0) P(Q > 0) \\
&= (1 - q) + q(1 - e^{-(m\mu - \lambda)x}) \\
&= 1 - q e^{-(m\mu - \lambda)x}
\end{aligned}
$$

$P(Q \leq x | Q > 0)$ is given by the SSQ, rate $m\mu$, result for waiting time. (Why is this?)
Note that $F_Q(x)$ has a jump at the origin, $F_Q(0) \neq 0$.

- Mean queueing time

$$W_Q = \frac{q}{m\mu - \lambda}$$

- Mean waiting time

$$W = W_Q + 1/\mu = \frac{(q + m)\mu - \lambda}{\mu(m\mu - \lambda)}$$

- Exercise: What is the waiting time density function?

The infinite server

- In the M/M/m queue, let $m \to \infty$
  - "unlimited service capacity"
  - always sufficient free servers to process an arriving task - e.g. when the number tasks in the system is finite
  - no queueing $\implies$ infinite servers model delays in a task's processing

$$p_j = \frac{\rho^j}{j!} p_0 \implies p_0 = e^{-\rho}$$

  - balance equations have a solution for all λ, μ
  - this is not surprising since the server can never be overloaded
- This result is independent of the exponential assumption – a property of the queueing discipline (here there is no queueing – a pathological case)

M/M/m queues with finite state space

M/M/1/k (k is the max. queue length) queue

$$\mu(n) = n\mu \text{ for all queue lengths } n$$

$$\lambda(n) = \begin{cases} \lambda & 0 \leq n < k \\ 0 & n \geq k \end{cases}$$

Hence, if $\rho = \lambda/\mu$.

$$p_j = \begin{cases} p_0 \rho^j & j = 0, 1, \ldots, k \\ \frac{\prod_{i=0}^{j-1} \lambda(i)}{\prod_{i=0}^{j-1} \mu(i+1)} = 0 & j > k \text{ (as } \lambda(k) = 0) \end{cases}$$

Telephone network with maximum capacity of k calls

$$\lambda(n) = \begin{cases} \lambda & 0 \le n < k \\ 0 & n \ge k \end{cases} \quad \mu(n) = n\mu \text{ when } n \text{ calls are in progress}$$

so we have an M/M/k/k queue.

$$p_j = p_0 \frac{\rho^j}{j!} \quad j = 0, 1, \ldots, k \quad (\rho = \lambda/\mu)$$

Probability that a call is lost:

$$p_k = \frac{\rho^k/k!}{\sum_{j=0}^{k} \rho^j/j!} \quad \underline{\text{Erlang loss formula}}$$

$$\text{Throughput} = \lambda(1 - p_k) \quad \left[= \mu \sum_{j=1}^{k} j p_j \right]$$

Terminal system with parallel processing

N users logged on to a computer system with m parallel processors

- exponentially distributed think times, mean $1/\lambda$, before submitting next task
- each processor has exponential service times, mean $1/\mu$

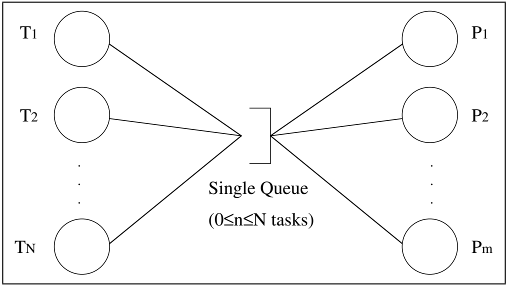


Figure: Terminals with parallel processors.

- tasks may use any free processor, or queue (FIFO) if there are none.
- state space $S = \{n | 0 \leq n \leq N\}$ where n is the queue length (for all processors)
- Poisson arrivals, rate $\lambda(n) = (N - n)\lambda$
- Service rate $\mu(n) = \begin{cases} n\mu & 1 \leq n \leq m \\ m\mu & m \leq n \leq N \end{cases}$

Steady state probabilities $\{p_i | 0 \le i \le N\}$

$$p_i = p_0 \frac{\prod \lambda(i-1)}{\prod \mu(i)} = \frac{N(N-1)\dots(N-i+1)}{i!}\rho^i p_0 \quad (0 \le i \le m)$$

$$\implies p_i = \begin{cases} \binom{N}{i}\rho^i p_0 & 0 \le i \le m \\ \frac{N!}{(N-i)!m!m^{i-m}}\rho^i p_0 & m \le i \le N \end{cases}$$

$$p_0 = \left\{ \sum_{i=0}^{m-1} \binom{N}{i} \rho^i + \sum_{i=m}^{N} \frac{N!}{(N-i)!m!m^{i-m}}\rho^i \right\}^{-1}$$

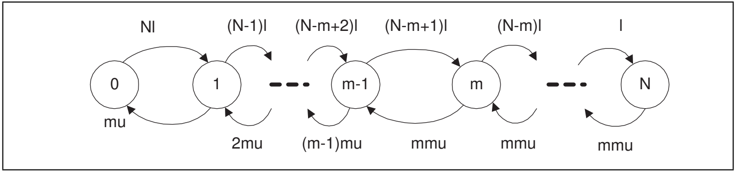


Figure: State transition diagram.

The throughput is either given by

$$\mu\Big\{\sum_{j=1}^{m-1} jp_j + m\sum_{j=m}^{N} p_j\Big\} \qquad \text{(mean departure rate from processors)}$$

or

$$\lambda\Big\{N - \sum_{j=1}^{N} jp_j\Big\} \qquad \text{(mean arrival rate)}$$

Other performance measures as previously

The case of "always sufficient processors" - $m \geq N$

- Here there is no case $m < i \leq N$

$$\implies p_i = \binom{N}{i} \rho^i p_0 \quad (0 \leq i \leq N)$$

$$\implies p_0 = \left\{ \sum_{i=0}^{N} \binom{N}{i} \rho^i \right\}^{-1} = (1+\rho)^{-N}$$

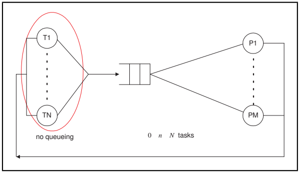
$$\text{Thus } p_i = \binom{N}{i} \Big(\frac{\rho}{1+\rho}\Big)^i \Big(\frac{1}{1+\rho}\Big)^{N-i}$$

(Binomial distribution)

- Probabilistic explanation
  - p_i is the probability of i "successes" in N Bernoulli (i.i.d.) trials
  - probability of success $= \frac{\rho}{1+\rho} = \frac{1/\mu}{1/\lambda+1/\mu} =$ fraction of time user is waiting for a task to complete in the steady state $=$ probability (randomly) observe a user in wait-mode.
  - probability of failure $=$ fraction of time user is in think mode in the steady state $=$ probability (randomly) observe a user in think-mode
  - random observations are i.i.d. in steady state $\implies$ trials are Bernoulli
  - hence Binomial distribution
- result is independent of distribution of think times.

Analogy with a queueing network

Regard the model as a 2-server, cyclic queueing network



$\Longrightarrow$

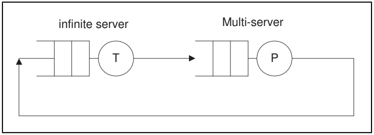


Figure: Original and equivalent network.

- As already observed, IS server is insensitive to its service time distribution as far as queue length distribution is concerned.
- Multi-server becomes IS if $M \geq N \implies$ 2 IS servers.

Little's result/formula/law (J.D.C. Little, 1961)

Suppose that a queueing system Q is in steady state (i.e. there are fixed, time independent probabilities of observing Q in each of its possible states at random times.) Let:

$$\begin{aligned} L &= \text{average number of tasks in } Q \text{ in steady state} \\ W &= \text{average time a task spends in } Q \text{ in steady state} \\ \lambda &= \text{arrival rate (i.e. average number of tasks entering} \\ &\quad \text{or leaving } Q \text{ in unit time)} \end{aligned}$$

Then

$$L = \lambda W.$$

Intuitive Justification

Suppose a charge is made on a task of £1 per unit time it spends in Q

- Total collected on average in unit time $= L$
- Average paid by one task $= W$
- If charges collected on arrival (or departure), average collected in unit time $= \lambda.W$

$$\implies L = \lambda W$$

- Example: M/M/1 queue

$$W = (L+1)/\mu \text{ by R.O.P.}$$

$$L = \lambda W \implies W = \frac{1}{\mu - \lambda}$$

Application of Little's law

Server utilisation U.

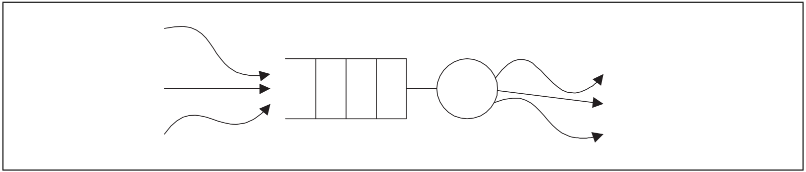


Figure: Little's law.

- Consider server, rate μ (constant), inside some queueing system: i.e. average time for one service $= 1/\mu$
- Let total throughput of server $= \lambda =$ total average arrival rate (since in steady state)
- Apply Little's result to the server only (without the queue)

$$\begin{aligned}
\text{mean queue length} &= 0.P(\text{idle}) + 1.P(\text{busy}) \\
&= U \\
\text{mean service time} &= \mu^{-1} \\
\implies U &= \frac{\lambda}{\mu}
\end{aligned}$$

- Not the easiest derivation! This is a simple work conservation law.

Single server queue with general service times: the M/G/1 queue

Assuming that arrivals are Poisson with constant rate λ, service time distribution has constant mean μ^{-1} (service rate μ) and that a steady state exists

$$\begin{aligned}\text{Utilisation, } U &= P(\text{queue length} > 0)\\ &= \frac{\mu^{-1}}{\lambda^{-1}} = \lambda/\mu = \text{“load.”}\end{aligned}$$

(For an alternate viewpoint, utilisation may be seen as the average amount of work arriving in unit time; we already know this, of course.)

Writing

$$\begin{aligned}
\text{Mean queue length } &= L \\
\text{Mean number in queue} &= L_Q \\
\text{Mean waiting time} &= W \\
\text{Mean queueing time} &= W_Q
\end{aligned}$$

then by Little's law,

$$\begin{aligned}
L &= \lambda W \\
L_Q &= \lambda W_Q
\end{aligned}$$

and by definition

$$W = W_Q + 1/\mu$$

So we have 3 equations and 4 unknowns.

The fourth equation

- By the random observer property, queue faced on arrival has mean length L_Q (excluding task in service, if any)
- By "residual life" result for renewal processes, average service time remaining for task in service (if any) is $\frac{\mu M_2}{2}$ where M_2 is the second moment of the service time distribution ($M_2 = \int_0^\infty x^2 f(x) dx$ where $f(x)$ is the pdf of service time)
- Thus, since $\rho = P(\exists$ a task being served at arrival instant)

$$W_Q = L_Q.1/\mu + \rho.\frac{\mu M_2}{2}$$

- Now

$$L_Q = \lambda W_Q \implies L_Q, W_Q, L, W$$

$$L = \rho + \frac{\lambda^2 M_2}{2(1-\rho)}$$

- Observe that if $\frac{\text{standard deviation}}{\text{mean}}$ (and hence the second moment) of service time distribution is large, L is also (not trivial as M_2 increases with μ^{-1} - but not difficult either!)

Embedded Markov chain

- State of the M/G/1 queue at time t is $X(t) \in S$ where the state space $S = \{n | n \geq 0\}$ as in M/M/1 queue.
- M/G/1 queue is *not* a Markov process
  - $P(X(t+s)|X(u), u \leq t) \neq P(X(t+s)|X(t)) \; \forall t, s$
  - e.g. if a service period does not begin at time t
  - no memoryless property
- Consider times $t_1, t_2, \ldots$ of successive departures and let $X_n = X(t_n^+)$

- Given X_i, $X(t)$ for $t > t_i$ is independent of $X(t')$ for $t' < t_i$ since at time t_i^+
  - time to next arrival is exponential with parameter λ because arrival process is Poisson
  - instantaneously, no task is in service, so time to next departure is a *complete* service time or that plus the time to next arrival (if queue empty)

  $\implies \{X_i | i \geq 1\}$ is a Markov Chain with state space S ($\{t_i | i \geq 1\}$ are called "Markov times")
- It can be shown that, in steady state of E.M.C., distribution of no. of jobs, j, "left behind" by a departure $=$ distribution of no. of jobs, j, seen by an arrival $= \lim_{n\to\infty} P(X_n = j)$ by R.O.P.
- Here we solve $p_j = \sum_{n=0}^{\infty} p_n q_{nj} \quad (j \geq 0)$ for appropriate one-step transition probabilities q_{ij}.

Balance equations for the M/G/1 queue

- Solution for $\{p_j\}$ exists iff $\rho = \frac{\lambda}{\mu} < 1$, equivalent to $p_0 > 0$ since $p_0 = 1 - U = 1 - \rho$ in the steady state.
- Valid one-step transitions are $i \to j$ for $j = i - 1, i, i + 1, i + 2, \ldots$ since we may have an arbitrary number of arrivals between successive departures.
- Let

  $$r_k = P(k \text{ arrivals in a service period})$$

  then

  $$\begin{aligned} q_{ij} &= P(X_{n+1} = j | X_n = i) && n \geq 0 \\ &= \underbrace{r_{j-i+1}}_{\text{because } i \to i-1+(j-i+1)} && i \geq 1, j \geq i - 1 \\ q_{0j} &= r_j && j \geq 0 \end{aligned}$$

[Eventually a job arrives, so $0 \to 1$, and then $1 \to j$ if there are j arrivals in its service time since then $1 \to 1 - 1 + j = j$]

$$\implies p_0 = 1 - \rho$$

$$p_j = p_0 r_j + \sum_{i=1}^{j+1} p_i r_{j-i+1} \quad (j \geq 0)$$

where $r_k = \int_0^\infty \frac{\lambda x)^k}{k!} e^{-\lambda x} f(x) dx$ if service time density function is $f(x)$ (known). This is because $P(k \text{ arrivals in service time}|\text{service time} = x) = \frac{\lambda x)^k}{k!} e^{-\lambda x}$ and $P(\text{service time} \in (x, x + dx)) = f(x) dx$

Solutions to the balance equations

- In principle we could solve the balance equations by “forward substitution”
  - p_0 is known
  - $j = 0$: p_0 allows us to find p_1
  - $j = 1$: p_0, p_1 allow us to find p_2

$$\vdots$$

  - $j = i$: $p_0, p_1, \ldots, p_i$ allow us to find p_{i+1}

  but computationally this is impractical in general

- Define generating functions

$$p(z) = \sum_{i=0}^{\infty} p_i z^i$$

$$r(z) = \sum_{i=0}^{\infty} r_i z^i$$

Recap:

$$p'(z) = \sum_{i=1}^{\infty} i p_i z^{i-1}$$

$$\implies p'(1) = \text{mean value of distribution } \{p_j | j \geq 0\}$$

$$p''(z) = \sum_{i=2}^{\infty} (i^2 - i) p_i z^{i-2}$$

$$\implies p''(1) = M_2 - M_1$$

- Multiply balance equations by z^j and sum:

$$
\begin{aligned}
p(z) &= p_0 r(z) + \sum_{j=0}^{\infty} \sum_{i=1}^{j+1} p_i r_{j-i+1} z^j \\
&= p_0 r(z) + z^{-1} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_{i+1} z^{i+1} r_{j-i} z^{j-i} \\
&\qquad \text{where } r_k = 0 \text{ for } k < 0 \\
&= p_0 r(z) + z^{-1} \sum_{i=0}^{\infty} p_{i+1} z^{i+1} \sum_{j=0}^{\infty} r_{j-i} z^{j-i} \\
&= p_0 r(z) + z^{-1} (p(z) - p_0) r(z)
\end{aligned}
$$

Solution for $p(z)$ and the Pollaczek-Khinchine result

$$
\begin{aligned}
p(z) &= \frac{p_0(1-z)r(z)}{r(z)-z} \\
\text{where } r(z) &= \int_0^\infty \sum_{k=0}^\infty \frac{(\lambda x z)^k}{k!} e^{-\lambda x} f(x) dx \\
&= \int_0^\infty e^{-\lambda x(1-z)} f(x) dx \\
&= f^*(\lambda - \lambda z)
\end{aligned}
$$

which is the Laplace transform of f at the point $\lambda - \lambda z$

Recap: Laplace transform f^* of f defined by

$$
\begin{aligned}
f^*(s) &= \int_0^\infty e^{-sx} f(x) dx \\
\text{so that } \frac{d^n}{ds^n} f^*(s) &= \int_0^\infty (-x)^n e^{-sx} f(x) dx \\
\implies \left.\frac{d^n f^*(s)}{ds^n}\right|_{s=0} &= (-1)^n M_n \quad n^{th} \text{ moment of } f(x)
\end{aligned}
$$

E.g. $f^*(0) = 1$, $-f^{*\prime}(0) =$ mean service time $= 1/\mu$. Thus,

$$
p(z) = \frac{(1-\rho)(1-z) f^*(\lambda - \lambda z)}{f^*(\lambda - \lambda z) - z}
$$

$p'(1) \implies$ P-K formula ...

Derivation of P-K formula

$$\frac{p'}{1-\rho} = \frac{(f^* - z)(-\lambda(1-z)f^{*\prime} - f^*) + (1-z)f^*(1+\lambda f^{*\prime})}{(f^* - z)^2}$$

where $f^* = f^*(\lambda - \lambda z)$ etc.

- When $z = 1$, both denominator and nominator vanish ($\implies f^*(\lambda - \lambda .1) = f^*(0) = 1$)
- L'Hopital rule $\implies$

$$\frac{p'(1)}{1-\rho} = \lim_{z\to 1}\left\{\frac{\lambda((1-2z)f^{*\prime} - \lambda z(1-z)f^{*\prime}) - \lambda f^{*\prime} + 2\lambda f^* f^{*\prime}}{-2(f^* - z)(1+\lambda f^{*\prime})}\right\}$$

- Again, when $z = 1$, both denominator and nominator vanish

- L'Hopital rule now gives

$$\frac{p'(1)}{1-\rho} = \frac{\lambda(-2f^{*\prime}(0) + \lambda f^{*\prime\prime}(0) - 2\lambda f^{*\prime}(0)^2)}{2(1+\lambda f^{*\prime}(0))^2}$$
$$\text{(since } f^*(0) = 1\text{)}$$
$$= \frac{\lambda^2 M_2 + 2(\lambda/\mu)(1-(\lambda/\mu))}{2(1-(\lambda/\mu))^2} \quad \text{P-K formula!}$$
$$\text{(since } f^{*\prime}(0) = 1/\mu\text{)}$$

- Hard work compared to previous derivation! But in principle we could obtain *any* moment ("well known" result for variance of queue length)

Waiting time distribution

- The tasks left in an M/G/1 queue on departure of a given task are precisely those which arrived during its waiting time

$$\implies p_j = \int_0^\infty \frac{(\lambda x)^j}{j!} e^{-\lambda x} h(x) dx$$

$$\text{because } P(j \text{ arrivals}|\text{waiting time} = x) = \frac{(\lambda x)^j}{j!} e^{-\lambda x}$$

$$P(\text{waiting time} \in (x, x + dx)) = h(x) dx$$

$$\implies p(z) = h^*(\lambda - \lambda z)$$

by the same reasoning as before.

- Laplace transform of waiting time distribution is therefore (let $z = \frac{\lambda - s}{\lambda}$)

$$h^*(s) = p\Big(\frac{\lambda - s}{\lambda}\Big) = \frac{(1-\rho)sf^*(s)}{\lambda f^*(s) - \lambda + s}$$

- Exercise: Verify Little's Law for the M/G/1 queue:

$$p'(1) = -\lambda h^{*\prime}(0).$$

Example: Disk access time model

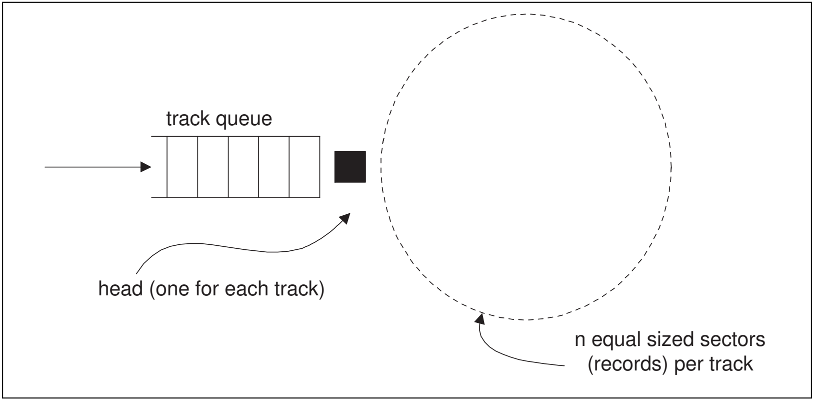


Figure: Fixed head disk.

Assumptions

- Tasks in the track queue require random sector
- arrivals to an empty queue always find the head at the beginning of a sector (as with the next task in the queue after a service completion)

$$
\Longrightarrow \text{ service times may be } \begin{array}{cr} \dfrac{1}{nR} & \text{(requires next sector)} \\ \dfrac{2}{nR} & \text{(next but one)} \\ \vdots & \\ \dfrac{1}{R} & \text{(the one just gone)} \end{array}
$$

(strictly, for arrivals to an empty queue, service times have continuous sample space $[1/nR, 1/nR + 1/R)$)

- Mean service time, $\mu^{-1} = \sum_{j=1}^{n} \frac{1}{n} \frac{j}{nR} = \frac{n+1}{2nR}$
- Second moment, $M_2 = \sum_{j=1}^{n} \frac{1}{n} (\frac{j}{nR})^2 = \frac{(n+1)(2n+1)}{6n^2R^2}$

Solution and asymptotic behaviour

- Load is $\rho = \lambda/\mu = \frac{\lambda(n+1)}{2nR} \implies \lambda < \frac{2nR}{n+1}$ if drum track is not to be saturated
- Mean queue length, $L = \rho + \frac{\lambda^2(n+1)(2n+1)}{12n^2R^2(1-\rho)}$
- As $n \to \infty$, i.e. many sectors on track
  - assumption about arrivals to an empty queue becomes exact (large $n \implies$ good approximation)
  - $\rho \to \frac{\lambda}{2R}$
  - $L \to \frac{\lambda}{2R} + \frac{\lambda^2}{3R(2R-\lambda)}$

Queueing Networks

- Collection of servers/queues interconnected according to some topology

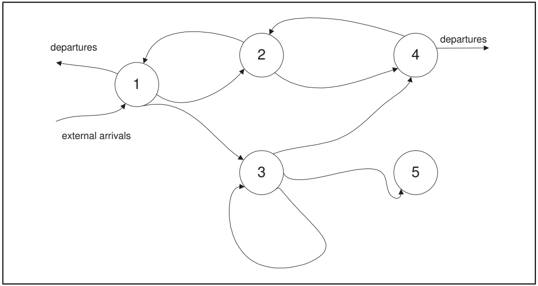


Figure: Network example

- Servers may be
  - processing elements in a computer, e.g. CPU I/O devices
  - stations/nodes in a communication network (may be widely separated geographically)
- Topology represents the possible routes taken by tasks through the system
- May be several different classes of tasks (multi-class network)
  - different service requirements at each node
  - different routing behaviours
  - more complex notation, but straightforward generalisation of the single-class network in principle
  - we will consider only the single class case

Types of network

- Open: at least one arc coming from the outside and at least one going out
  - must be at least one of each type or else the network would be saturated or null (empty in the steady state)
  - e.g. the example above
- Closed: no external (incoming or outgoing) arc
  - constant network population of tasks forever circulating
  - e.g. the example above with the external arcs removed from nodes 1 and 4
- Mixed: multi-class model which is open with respect to some classes and closed with respect to others – e.g. in the example above a class whose tasks only ever alternated between nodes 2 and 4 would be closed, whereas a class using all nodes would be open

Types of server

- Server defined by its service time distribution (we assume exponential but can be general for non-FCFS disciplines) and its queueing discipline (for each class)
  - FCFS (FIFO)
  - LCFS (LIFO)
  - IS (i.e. a *delay node*: no queueing)
  - PS (*Processor sharing*: service shared equally amongst all tasks in the queue)
- Similar results for queue length probabilities (in S.S.) for all

Open networks (single class)

- Notation:
  M servers, $1, 2, \ldots, M$ with FCFS discipline and exponential service times, mean $\frac{1}{\mu_i(n_i)}$, when queue length is n_i
  $(1 \leq i \leq M)$
  - *state dependent service rates* to a limited extent
  - $\mu_i(n_j)$ for $i \neq j \implies$ *blocking*: rate at one server depends on the state of another, e.g. rate $\rightarrow 0$ when finite queue at next is full
- External Poisson arrivals into node i, rate γ_i, $(1 \leq i \leq M)$
  $(= 0$ if no arrivals$)$

- Routing probability matrix $Q = (q_{ij} | 1 \leq i \leq M)$
  - q_{ij} = probability that on leaving node i a task goes to node j independently of past history
  - $q_{i0} = 1 - \sum_{j=1}^{M} q_{ij}$ = probability of external departure from node i
  - at least one $q_{i0} > 0$, i.e. at least one row of Q sums to less than 1
- State space of network $S = \{(n_1, \ldots, n_M)\} | n_i \geq 0\}$
  - queue length vector random variable is $(N_1, \ldots, N_M)$
  - $p(\underline{n}) = p(n_1, \ldots, n_M) = P(N_1 = n_1, \ldots, N_M = n_M)$

Traffic equations

Determine mean arrival rates λ_i to each node i in the network

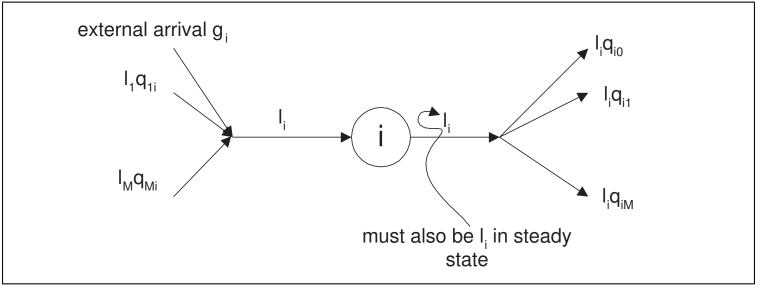


Figure: Traffic equations

In the steady state, $\lambda_i = \gamma_i + \sum_{j=1}^{M} \lambda_j q_{ji}$ for $(1 \leq i \leq M) \implies$ unique solution for $\{\lambda_i\}$ (because of properties of Q)

- independent of Poisson assumption since we are only considering mean numbers of arrivals in unit time
- assumes only the existence of a steady state

Example

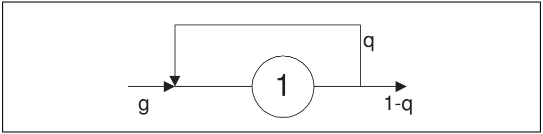


Figure: Traffic example.

$$\lambda_1 = \gamma + \lambda_1 q \implies \lambda_1 = \frac{\gamma}{1-q}$$

- Arrivals to a node are not in general Poisson, e.g. this simple example. If there is no feedback then all arrival processes are Poisson because
    1. departure process of M/M/1 queue is Poisson
    2. superposition of independent Poisson processes is Poisson
- Similarly, let R_i be the average interval between a task's arrival at server i and its departure from the network
    - R_i is the "average remaining sojourn time"

$$R_i = W_i + \sum_{j=1}^{M} q_{ij} R_j$$

Steady state queue length probabilities

Jackson's theorem

1. The number of tasks at any server is independent of the number of tasks at every other server in the steady state
2. Node i behaves *as if* it were subjected to Poisson arrivals, rate λ_i $(1 \leq i \leq M)$

- Thus, even though arrivals at each node are not, in general, Poisson, we can treat the system as if it were a collection of M independent M/M/1 queues

$$\implies p(n_1, \ldots, n_M) \propto \prod_{i=1}^{M} \left\{ \frac{\lambda_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \right\}$$

where $\mu_i(j)$ is the rate of server i when the queue length is j

- If service rates are constant, μ_i,

$$p(\underline{n}) \approx \prod_{i=1}^{M} \rho_i^{n_i} = \prod_{i=1}^{M} (1-\rho_i)\rho_i^{n_i}$$

  where $\rho_i = \frac{\lambda_i}{\mu_i}$
- $p(\underline{n}) \rightarrow$ usual performance measures such as mean queue lengths, utilisations, throughput by standard methods – mean waiting times by Little's result

Note

The normalising constant for $\{p(\mathbf{n}) | \mathbf{n} \in S\}$ is not shown for the general case: it is the product of those for the M/M/1 queues.

Mean Value analysis

- Can exploit Jackson's theorem directly, together with Little's result, if only <u>average</u> quantities are required
  - values for mean queue lengths L_i are those for isolated M/M/1 queues with arrival rates λ_i $(1 \leq i \leq M)$
  - assuming constant service rates μ_i

$$L_i = \frac{\rho_i}{1-\rho_i} \text{ for } 1 \leq i \leq M$$

    (average number of tasks at node i)
  - total average number of tasks in network

$$L = \sum_{i=1}^{M} L_i = \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}$$

- Waiting times
  - Average waiting time in the network, $W = L/\gamma$ by Little's result where $\gamma = \sum_{i=1}^{M} \gamma_i$ is the total arrival rate
  - Average time spent at node i during each visit

$$W_i = L_i/\lambda_i = \frac{1}{\mu_i(1-\rho_i)}$$

  - Average time spent queueing on a visit to node i

$$W_{Qi} = L_{Qi}/\lambda_i = \frac{\rho_i}{\mu_i(1-\rho_i)}$$

An alternative formulation

- Let v_i be the average number of visits made by a task to server i

$$\gamma v_i = \text{average number of arrivals to node } i \text{ in unit time} = \lambda_i$$

  where γ is the average number of arrivals to the whole network in unit time and v_i the average number of visits a given arrival makes to node i

$$v_i = \frac{\lambda_i}{\gamma} \quad (1 \leq i \leq M)$$

- Let D_i be the total average service demand on node i from one task

$$D_i = \frac{v_i}{\mu_i} = \frac{\rho_i}{\gamma}$$

  $\rho_i = \gamma D_i =$ average work for node i arriving from outside the network in unit time

- Often specify a queueing network directly in terms of $\{D_i | 1 \leq i \leq M\}$ and γ; then there is no need to solve the traffic equations
-
$$L_i = \frac{\rho_i}{1 - \rho_i} = \frac{\gamma D_i}{1 - \gamma D_i}$$
  L and W as before
- Let B_i = total average time a task spends at node i
$$B_i = v_i W_i = \frac{v_i}{\mu_i(1 - \rho_i)} = \frac{D_i}{1 - \gamma D_i}$$
  alternatively apply Little's result to node i with the external arrival process directly
$$B_i = \frac{L_i}{\gamma} = \frac{D_i}{1 - \gamma D_i}$$

- However, D_i and γ cannot be used to determine μ_i and hence neither W_i nor R_i
- Specification for delay nodes (IS) i
  - $L_i = \rho_i = \gamma D_i$
  - $B_i = D_i$ (no queueing)
  - $W_i = 1/\mu_i$
  - $D_i = v_i/\mu_i$ (as for any work conserving discipline also)
  - service time distribution arbitrary

Distribution of time delays

- Even though arrivals at each node are not, in general, Poisson, the Random Observer Property still holds: *waiting time distribution at node i is exponential, parameter* $\mu_i - \lambda_i$ again as expected from Jackson's theorem.
- Networks with no overtaking ("tree-like" networks) are easy to solve for time delay distributions:

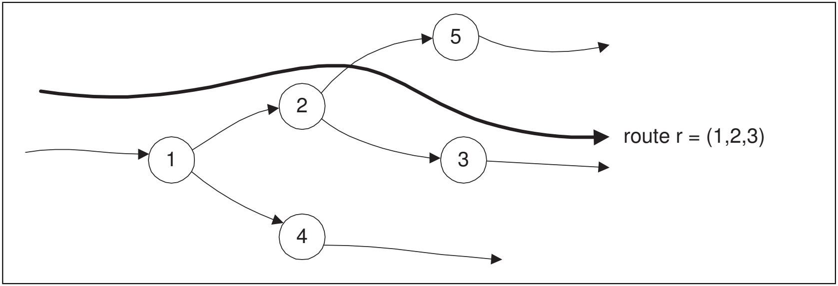


Figure: A network with no overtaking.

- sojourn time on any route is the sum of independent, exponential random variables
- this argument is independent of Jackson's theorem and one proof uses the idea of *reversibility* of the M/M/1 queue
- time delay distribution is a convolution of exponentials, e.g. $f_1 \star f_2 \star f_3$ for route r where $f_i(t) = (\mu_i - \lambda_i)e^{-(\mu_i - \lambda_i)t}$ for $i = 1, 2, 3$.

Time delays in general networks

- mean sojourn time for any route is always easy because the mean of a sum of random variables is equal to the sum of the means of those random variables, whether or not they are independent
- in networks with overtaking, the distribution of route sojourn times remains an open problem. For example, in the network

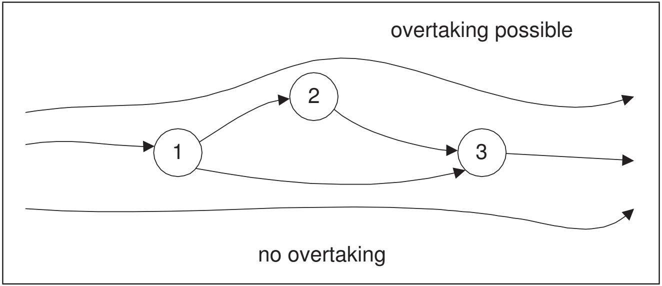


Figure: A network with overtaking.

- - sojourn time distribution on route $1 \to 3$ is the convolution of 2 exponentials
  - sojourn time distribution on route $1 \to 2 \to 3$ is *not* the convolution of 3 exponentials because the queue length faced at node 3 upon arrival depends on the number of departures from node 1 during the sojourn time at node 2.
  - The Random Observer Property is not applicable since the arrival to node 3 is not random when conditioned on the previous arrival at node 1.
- Jackson's theorem does not apply because it is concerned only with steady state probabilities, i.e. the asymptotic behaviour of $p_t(\underline{n})$ at the single point in time t as $t \to \infty$.
- Subject of many research papers.

Closed Queueing Networks

- No external arrivals or departures (no γ_i terms).
- Routing probabilities satisfy

$$\sum_{j=1}^{M} q_{ij} = 1 \text{ for } 1 \leq i \leq M$$

- State space $S = \{(n_1, \ldots, n_M) | n_i \geq 0, \sum_{j=1}^{M} n_j = K\}$ for population K
  - $|S| = \#$ of ways of putting K balls into M bags $= \binom{K+M-1}{M-1}$
  - finiteness of $S \rightarrow$ steady state always exists

- Traffic equations are

$$\lambda_i = \sum_{j=1}^{M} \lambda_j q_{ji} \text{ for } 1 \leq i \leq M$$

  - homogeneous linear equations with an infinity of solutions which differ by a multiplicative factor (because $|I - Q| = 0$ since rows all sum to zero)
  - let $(e_1, \ldots, e_M)$ be any non-zero solution (typically chosen by fixing one e_i to a *convenient* value, like 1)

therefore $e_i \propto$ arrival rate λ_i, i.e. $e_i = c\lambda_i$
$x_i = \frac{e_i}{\mu_i} \propto$ load $= \rho_i$, i.e. $x_i = c\rho_i$

Steady state probability distribution for S

- Jackson's theorem extends to closed networks which have a product form solution

$$p(n_1, \ldots, n_M) = \frac{1}{G} \prod_{i=1}^{M} \frac{e_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \text{ where } \sum_{i=1}^{M} n_i = K. \quad (1)$$

  where $\mu_i(j)$ is the service rate of the exponential server i when its queue length is j.

- G is the normalising constant of the network

$$G = \sum_{\boldsymbol{n} \in S} \prod_{i=1}^{M} \frac{e_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \quad (2)$$

  not easy to compute (see later)

- Prove the result by using the network's steady state balance equations:

$$\begin{aligned}
\text{Total flux out of state } \boldsymbol{n} &= p(\boldsymbol{n}) \sum_{i=1}^{M} \mu_i(n_i)\epsilon(n_i) \\
&= \text{Total flux into state } \boldsymbol{n} : \sum_{i,j} \boldsymbol{n}_i^j \to \boldsymbol{n} \\
&= \sum_{i=1}^{M} \sum_{j=1}^{M} p(\boldsymbol{n}_i^j)\epsilon(n_i)\mu_j((\boldsymbol{n}_i^j)_j)q_{ji}
\end{aligned}$$

$$\text{where } \epsilon(n) = \begin{cases} 0 & \text{if } n = 0 \\ 1 & \text{if } n > 0 \end{cases}$$

$$\boldsymbol{n}_i^j = \begin{cases} (n_1, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M) & \text{if } i \neq j \\ \boldsymbol{n}_i^i = \boldsymbol{n} & \text{otherwise} \end{cases}$$

note that $(\boldsymbol{n}_i^j)_j = \begin{cases} n_j + 1 & \text{if } i \neq j \\ n_i & \text{otherwise} \end{cases}$

- Try

$$p(n_1, \ldots, n_M) = \frac{1}{G} \prod_{i=1}^{M} \left(\frac{e_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \right) \qquad \text{where } \sum_{i=1} n_i = K.$$

Then require

$$\frac{1}{G}\prod_{i=1}^{M}\left(\frac{e_i^{n_i}}{\prod_{j=1}^{n_i}\mu_i(j)}\right)\sum_i \mu_i(n_i)\epsilon(n_i) = \frac{1}{G}\prod_{i=1}^{M}\left(\frac{e_i^{n_i}}{\prod_{j=1}^{n_i}\mu_i(j)}\right)\sum_{i,j:i=j}\mu_i(n_i)\epsilon(n_i)q_{ji}$$

$$+\frac{1}{G}\prod_{i=1}^{M}\frac{e_i^{n_i}}{\prod_{j=1}^{n_i}\mu_i(j)}\times$$

$$\sum_{i,j:i\neq j}\frac{e_j}{e_i}\frac{\mu_i(n_i)}{\mu_j(n_j+1)}\epsilon(n_i)\mu_j(n_j+1)q_{ji}$$

i.e.

$$\sum_i \mu_i(n_i)\epsilon(n_i) = \sum_i \mu_i(n_i)\epsilon(n_i)\left\{\frac{\sum_j e_j q_{ji}}{e_i}\right\}$$

which is satisfied if $\boldsymbol{e}$ satisfies the traffic equations.

- Note that if $\boldsymbol{e}' = c\boldsymbol{e}$ is another solution to the traffic equations, the corresponding probabilities $p'(\boldsymbol{n})$ and G' are

$$p'(\boldsymbol{n}) = \frac{1}{G'} G c^{\sum n_i} p(\boldsymbol{n}) = \frac{Gc^K}{G'} p(\boldsymbol{n})$$

$$G' = c^K G = \sum_{\boldsymbol{n} \in S} G c^{\sum n_i} p(\boldsymbol{n})$$

and therefore $p'(\boldsymbol{n}) = p(\boldsymbol{n})$. This confirms the arbitrariness of $\boldsymbol{e}$ up to a multiplicative factor.

Computation of the normalising constant

- We consider only the case of servers with constant service rates to get an efficient algorithm.
- There are also algorithms for networks with servers having state dependent service rates, e.g. the convolution algorithm.
- Less efficient but important since the alternative *MVA algorithm* also requires constant rates.

- Define $G = g(K, M)$ where

$$g(n, m) = \sum_{\underline{n} \in S(n,m)} \prod_{i=1}^{m} x_i^{n_i}$$

where $S(n, m) = \{(n_1, \ldots, n_m) | n_i \geq 0, \sum_{i=1}^{m} n_i = n\}$ and $x_i = \frac{e_i}{\mu_i}$ $(1 \leq i \leq m)$.

  - state space for subnetwork of nodes $1, 2, \ldots, m$ and population n.

- For $n, m > 0$

$$
\begin{aligned}
g(n,m) &= \sum_{\underline{n} \in S(n,m), n_m = 0} \prod_{i=1}^{m} x_i^{n_i} + \sum_{\underline{n} \in S(n,m), n_m > 0} \prod_{i=1}^{m} x_i^{n_i} \\
&= \sum_{\underline{n} \in S(n,m-1)} \prod_{i=1}^{m-1} x_i^{n_i} + x_m \sum_{\substack{k_i = n_i (i \neq m) \\ k_m = n_m - 1 \\ \underline{n} \in S(n,m)}} \prod_{i=1}^{m} x_i^{k_i} \\
&= g(n, m-1) + x_m g(n-1, m)
\end{aligned}
$$

because $\{\underline{k} | k_i \geq 0; \sum_{i=1}^{m} k_i = n - 1\} = S(n-1, m)$.

- Boundary conditions:

$$g(0, m) = 1 \text{ for } m > 0$$
$$g(n, 0) = 0 \text{ for } n \geq 0$$

- Note

$$g(0, m) = \sum_{\underline{n}=(0,\ldots,0)} \prod_{i=1}^{m} x_i^0 = 1$$

and

$$g(n, 1) = x_1 g(n-1, 1) = x_1^n$$

Marginal queue length probabilities and performance measures

- Although $p(\underline{k}) \propto \prod p_i(k_i)$ it is *not* the case that $P(N_i = k_i) = p_i(k_i)$, as in the open networks. The use of M/M/1 factors is just a convenient mathematical device, there is no probabilistic interpretation.
- Probability that a server is idle ($= 1 -$ utilisation)

$$P(N_M = 0) = \frac{1}{g(K,M)} \sum_{\underline{n} \in S(n,m), n_m = 0} \prod_{i=1}^{M-1} x_i^{n_i} = \frac{g(K, M-1)}{g(K,M)}.$$

In general

$$P(N_i = 0) = \frac{G_i(K)}{G(K)} \text{ for } 1 \le i \le M$$

where $G(K) = g(K, M)$ and $G_i(k)$ is the normalising constant for the same network with the server i removed and population k

$$G_i(k) = \sum_{\underline{n} \in S(k, M-1)} \prod_{j=1}^{M-1} y_j^{n_j}$$

where

$$y_j = \begin{cases} x_j & \text{for } 1 \leq j < i \\ x_{j+1} & \text{for } i \leq j \leq M-1 \end{cases}$$

- utilisation of node i

$$U_i(K) = 1 - \frac{G_i(K)}{G(K)}$$

Alternative Formulation: Cumulative Probabilities

- For $1 \le k \le K$ and $1 \le i \le M$

$$
\begin{aligned}
P(N_i \ge k) &= \sum_{\underline{n} \in S(K,M), n_i \ge k} \prod_{j=1}^{M} \frac{x_j^{n_j}}{G(K)} \\
&= \frac{x_i^k}{G(K)} \sum_{\substack{m_i = n_i - k \\ n_i \ge k \; m_j = n_j (j \ne i) \\ \underline{n} \in S(K,M)}} \prod_{j=1}^{M} x_j^{m_j} \\
&= \frac{x_i^k}{G(K)} \sum_{\underline{m} \in S(K-k,M)} \prod_{j=1}^{M} x_j^{m_j} \\
&= x_i^k \frac{G(K-k)}{G(K)}.
\end{aligned}
$$

Therefore the utilisation is given by

$$U_i = x_i \frac{G(K-1)}{G(K)}$$

- Equating two expressions for U_i yields the recurrence relation for $g(k, m)$ previously
- Throughput of server i,

$$T_i(k) = \mu_i U_i(k) = e_i \frac{G(K-1)}{G(K)}$$

proportional to visitation rate as expected.

- Queue length distribution at server i is $P(N_i = k) = p_i(k)$ for $0 \leq k \leq K$ and $1 \leq i \leq M$

$$\begin{aligned} p_i(k) &= P(N_i \geq k) - P(N_i \geq k+1) \\ &= x_i^k \frac{G(K-k) - x_i G(K-k-1)}{G(K)}. \end{aligned}$$

where $G(-1) = 0$ by definition.

- Notice that the previous formulation gives a more concise formulation for $p_i(k)$

$$\begin{aligned} p_i(k) &= \frac{1}{G(K)} \sum_{\underline{n} \in S(K,M), n_i = k} \prod_{j=1}^{M} x_j^{n_j} \\ &= \frac{x_i^k}{G(K)} \sum_{\underline{n} \in S(K-k,M), n_i = 0} \prod_{j=1}^{M} x_j^{n_j} \\ &= x_i^k \frac{G_i(K-k)}{G(K)} \end{aligned}$$

for $0 \leq k \leq K$ and $1 \leq i \leq M$. This is a generalisation of the result obtained for $U_i(k)$.

- Mean queue length at server i, $1 \leq i \leq M$, $L_i(k)$

$$
\begin{aligned}
L_i(K) &= \sum_{k=1}^{K} kP(N_i = k) \\
&= \sum_{k=1}^{K} P(N_i \geq k) \\
&= \frac{\sum_{k=1}^{K} x_i^k G(K-k)}{G(K)}.
\end{aligned}
$$

Equivalent open networks and the use of mean value analysis

- Consider an irreducible network – one in which every arc is traversed within finite time from any given time with probability 1

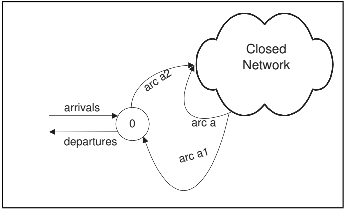


Figure: Equivalent open network.

- i.e. a network represented by an irreducible positive recurrent Markov process (finite state space)
- we introduce a node, 0, in one of the arcs and replace arc a by arc a_1, node 0 and arc a_2
  - Source of a_1 is source of a
  - destination of a_2 is destination of a
- Whenever a task arrives at node 0 (along arc a_1), it departs from the network and is immediately replaced by a stochastically identical task which leaves node 0 along arc a_2
- Define the network's throughput, T, to be the average rate at which tasks pass along arc a in the steady state.
  - i.e. T is mean number of tasks traversing a in unit time.
  - One can choose any arc in an irreducible network.

Visitation rates and application of Little's result

- Let the visitation rate be v_i and the average arrival rate be λ_i for node i, $1 \leq i \leq M$, then $\lambda_i = Tv_i$ where T is the external arrival rate.
- The set $\{v_i | 1 \leq i \leq M\}$ satisfies the traffic equations, as we could have derived directly by a similar argument.
- Suppose arc a connects node α to node β in the closed network $1 \leq \alpha, \beta \leq M$, then $v_0 = v_\alpha q_{\alpha\beta}$ because all traffic from α to β goes through node 0 in the open network.
- But every task enters node 0 exactly once, hence $v_\alpha = \frac{1}{q_{\alpha\beta}}$ since $v_0 = 1$. This determines $\{v_i | 1 \leq i \leq M\}$ uniquely.

- Little's result now yields

$$L_i = \lambda_i W_i = T v_i W_i$$

$$\sum_{i=1}^{M} L_i = K = T \sum_{i=1}^{M} v_i W_i$$

since the sum of all queue lengths is exactly K in the closed network

$$\implies \quad T = \frac{K}{\sum_{i=1}^{M} v_i W_i}$$

Mean waiting times

- $$W_i = \frac{1}{\mu_i}[1 + Y_i]$$

  where the first term is the arriving task's mean service time and Y_i is the mean number of tasks seen by new arrivals at server i.
- For an IS (delay) server $W_i = \frac{1}{\mu_i}$, otherwise ...
- Do *not* have the random observer property
  - number of tasks seen on arrival does not have the same steady state distribution as the queue length since K tasks are seen with probability 0. (arrival cannot "see itself")
  - do not have $Y_i = L - 1$ as in open networks

- Do have the analogous Task (or Job) Observer Property: The state of a closed queueing network in equilibrium seen by a new arrival at any node has the same distribution as that of the state of the same network in equilibrium with the arriving task removed (i.e. with population $K - 1$)
  - arriving task behaves as a random observer in a network with population reduced by one
  - intuitively appealing since the "one" is the task itself
  - but requires a lengthy proof (omitted)

Recurrence relations for throughput, mean waiting times and mean queue length

- Task observer property yields, for $1 \leq i \leq M, K > 0$,

$$Y_i(K) = L_i(K-1).$$

- Hence we obtain the recurrence relations

$$W_i(K) = \frac{1}{\mu_i}[1 + L_i(K-1)]$$

$$T(K) = \frac{K}{\sum_{i=1}^{M} v_i W_i(K)}$$

$$L_i(K) = v_i T(K) W_i(K)$$

  for $1 \leq i \leq M, K > 0$ and the initial condition $L_i(0) = 0$
- $T(K), W_i(K), L_i(K)$ easily computed by a simple iteration, calculating $2M+1$ quantities each time round the loop

$$\{L_i(K-1)\} \rightarrow \underbrace{\{W_i(K)\} \rightarrow T(K)}_{\rightarrow \{L_i(K)\}}$$

Alternative formulation

- Total average time spent at node i when population is K

$$B_i(K) = v_i W_i(K)$$

for $1 \leq i \leq M$

- Total average service time (demand) a task requires from node i

$$D_i = \frac{v_i}{\mu_i}$$

for $1 \leq i \leq M$ independent of population K.

- Therefore

$$B_i(K) = D_i[1 + L_i(K-1)]$$
$$T(K) = \frac{K}{\sum_{i=1}^{M} B_i(K)}$$
$$L_i(K) = T(K)B_i(K)$$

for $1 \leq i \leq M$

- Total average time in network spent by a task

$$W(K) = \sum_{i=1}^{M} v_i W_i(K) = \sum_{i=1}^{M} B_i(K) = \frac{K}{T(K)}$$

as expected by Little's law

- Utilisation of node i

$$U_i(K) = \frac{\lambda_i}{\mu_i} = T(K)\frac{v_i}{\mu_i} = T(K)D_i \leq 1$$

which implies

$$T(K) \leq \min_{1 \leq i \leq M} \frac{1}{D_i}$$

which implies the maximum throughput is dictated by a bottleneck server (or servers) - that (those) with maximum D_i

A faster approximate algorithm

- Algorithms given above need to evaluate $2M + 1$ quantities for each population between 1 and K
- Would be faster if we could relate $Y_i(K)$ to $L_i(K)$ rather than $L_i(K-1)$, which implies that we do not need to worry about populations less than K
- $$Y_i(K) = \frac{K-1}{K} L_i(K)$$

  is a good approximation, exact for $K = 1$ and correct asymptotic behaviour as $K \to \infty$ (exercise)

- then

$$B_i(K) = D_i\left[1 + \frac{K-1}{K} T(K) B_i(K)\right]$$

- $$B_i(K) = \frac{D_i}{1 - \frac{K-1}{K} D_i T(K)}$$

- Thus we obtain the fixed point equation

$$T(K) = f(T(K))$$

where

$$f(x) = \frac{K}{\sum_{i=1}^{M} \frac{D_i}{1-\frac{K-1}{K} D_i x}}$$

there are many numerical methods to solve such equations

Example: Multiprogramming computer system

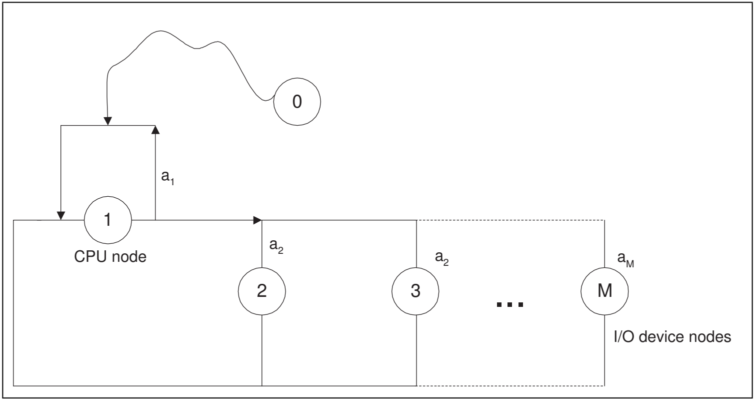


Figure: A multiprogrammed computer.

- Insert node 0 in route from CPU back to itself, *throughput* is the rate at which tasks are interrupted at CPU node, this is an arbitrary choice
- visitation rates

$$v_1 = a_1 v_1 + \sum_{j=2}^{M} v_j$$

$$v_i = a_i v_1 \text{ for } i \geq 2$$

$$v_1 = \frac{1}{a_1} \text{ with choice of position of node 0}$$

therefore

$$v_i = \frac{a_i}{a_1} \text{ for } i \geq 2$$

- Sanity check: is the first equation is satisfied?

$$D_i = \frac{a_i}{a_1 \mu_i} \text{ for } i \geq 2$$

$$D_1 = \frac{1}{a_1 \mu_1}$$

  therefore $T(K)$ follows by solving recurrence relations
- total CPU throughput $= \frac{T(K)}{a_1}$ (fraction a_1 recycles)

Application: A batch system with virtual memory

- Most medium-large computer systems use virtual memory
- It is well-known that above a certain degree of multiprogramming performance deteriorates rapidly
- Important questions are
  - How does the throughput and turnaround time vary with the degree of multiprogramming
  - What is the threshold below which to keep the degree of multiprogramming?

Representation of paging

- Suppose node 2 is the paging device (e.g. fast disk)
  - assume dedicated to paging
  - easy to include non-paging I/O also
- We aim to determine the service demand of a task at node 2, D_2, from tasks' average paging behaviour
  - use results from '70s working set theory
  - consider the page fault rate for varying amounts of main store allocated to a task
- Let the population be K

- paging behaviour will be represented by a *lifetime function*:
  - $h(x)$ = average CPU time between consecutive page faults when a task has x pages of memory
  - $h(0) \approx 0$ (the first instruction causes a fault)
  - $h(x) = C$, the total CPU time of a task for $x \geq$ some constant value m, where m is the "size" of the task
  - h is an *increasing function*: the more memory a task has the less frequent will be its page faults on average.
  - Two possible types of lifetime functions:

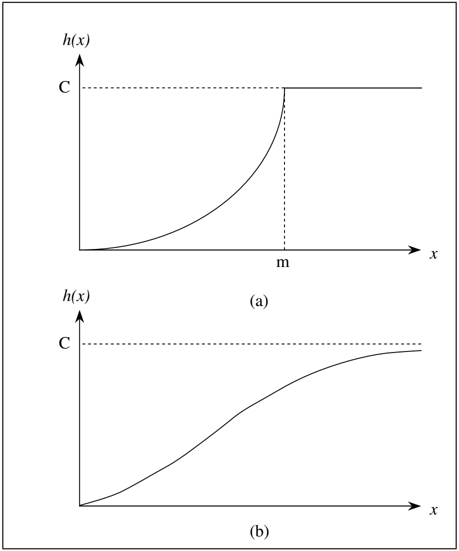


Figure: Lifetime functions

- For (a) $h(x) = \begin{cases} ax^b \text{ for } x \leq m \\ am^b = C \text{ for } x \geq m \end{cases}$, $b > 1$
- For (b) $h(x) = \frac{C}{1+(a/x)^2}$
- Note that for (b) there is no m s.t. $h(m) = C$
- Let each task have P/K pages of memory of size P pages
  - average CPU time between faults $= h(P/K)$
  - average number of faults during life of task $= \frac{D_1}{h(P/K)}$
  - average time at node 2 per fault $= \frac{1}{\mu_2}$

- Therefore average paging time of a task (from node 2)
  - $D_2 = H(K) = \frac{D_1}{\mu_2 h(P/K)}$
  - $D_2 = d_2 + H(K)$ if average non-paging requirement of a task from node 2 is d_2
- As $K \to \infty$, $T(K) \to 0$ since $T(K) \leq \min \frac{1}{D_i}$ and $D_2 \to \infty$

Solution

- We again use the alternative form of the MVA algorithm
- For population K and $1 \leq k \leq K$

$$
\begin{aligned}
B_i(k) &= D_i(K)[1 + L_i(k-1)] \quad (1 \leq i \leq M) \\
T(k) &= \frac{k}{\sum_{i=1}^{M} B_i(k)} \\
L_i(k) &= T(k)B_i(k) \quad (1 \leq i \leq M)
\end{aligned}
$$

- Notice that only D_2 is a function of K – D_1 and D_3 are constant

- Computation for a range of populations requires 2 nested loops:
  - outer loop for $K = 1, 2, \ldots, K_{max}$
  - inner loop for $k = 1, 2, \ldots, K$
  - need new value for $H(K)$ each time round the outer loop

Choice of lifetime function

We assume (arbitrarily) the lifetime function

$$h(x) = \begin{cases} D_1(\frac{x}{10000})^{3.5}, & x \leq 10000 \\ D_1 \text{ otherwise} \end{cases}$$

for all tasks, i.e.

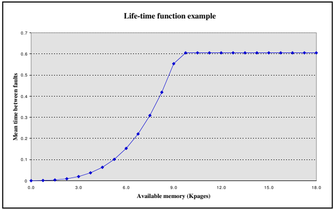


Figure: Lifetime function example.

Thrashing curves

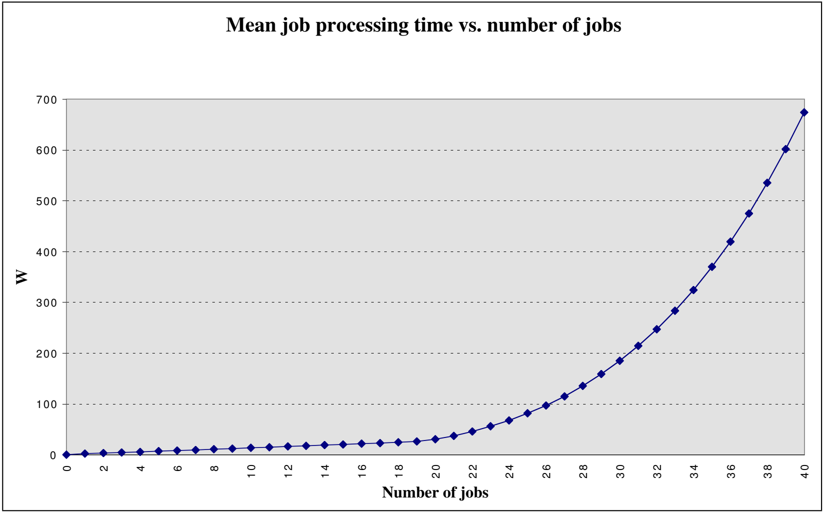


Figure: Mean job processing time vs. number of jobs.

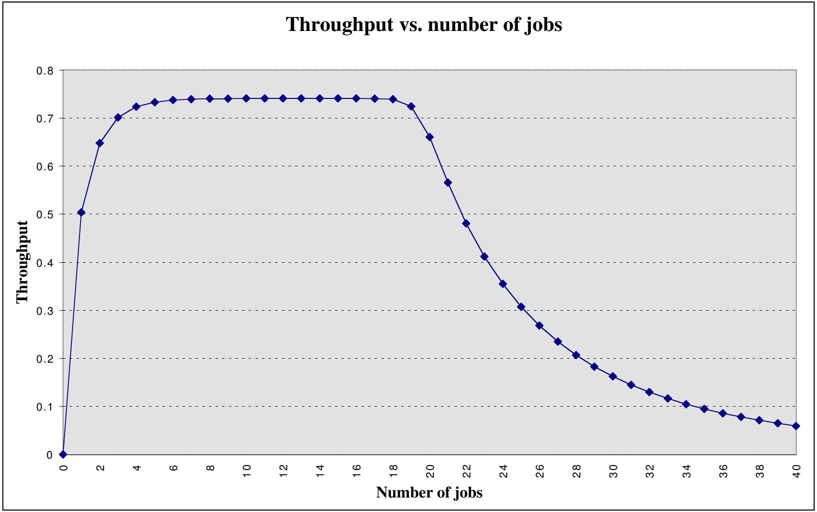


Figure: Throughput vs. number of Jobs.

Decomposition

- Separable queueing models are prone to break down in some situations:
  - There is usually a good model for small parts of any network
  - Composing the bits can be hard unless the criteria for separability are (reasonably) met
- Ideally, we need a method to:
  - Reduce complexity (breaking a large problem into smaller manageable ones)
  - Provide approximate solutions where the necessary assumptions do not hold

- The most powerful general method is *decomposition* (sometimes called *aggregation*)
- The idea is to replace a subnetwork of a (complex) system with a single *flow-equivalent server* (FES)

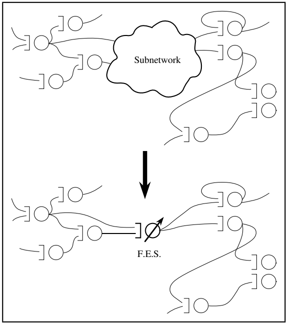


Figure: Flow equivalent server.

- We exploit the fact that servers in separable networks may have *state dependent* service rates, i.e. μ_n for queue length n
- To find the required μ_n, we compute the "throughput", τ_n, of the subnetwork when the population is n and set

$$\mu_n = \tau_n, \quad n = 0, 1, \ldots$$

- The throughput can be determined by *any* method, e.g. an explicit Markov model, standard queueing network algorithm (e.g. MVA), or even simulation
- The technique is analogous to Norton's Theorem in electrical circuit theory.

Method

1. Identify the subnetwork to be aggregated

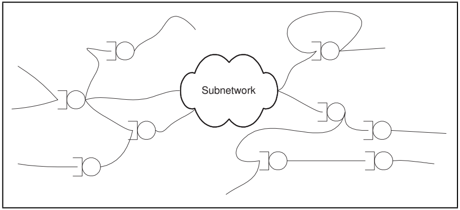


Figure: Flow equivalent server — step 1.

2. Short-circuit the rest of the network (i.e. use the same visitation rates as in the original subnetwork)

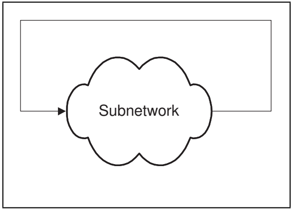


Figure: Flow equivalent server — step 2.

- Compute the throughput τ_n, on the short-circuit arc for all populations $0, 1, \ldots$

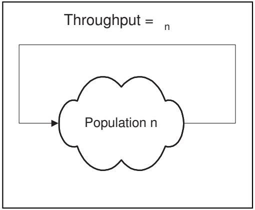


Figure: Flow equivalent server — step 3.

- Define a single FES with service rate $\mu_n = \tau_n, n = 0, 1, \ldots$
- Replace the subnetwork with the FES and re-solve

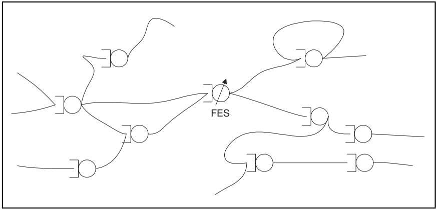


Figure: Flow equivalent server — step 5.

- If necessary repeat the process (i.e. hierarchically)

Note

We can either aggregate the tricky part or aggregate the easy part leaving a smaller tricky network behind!

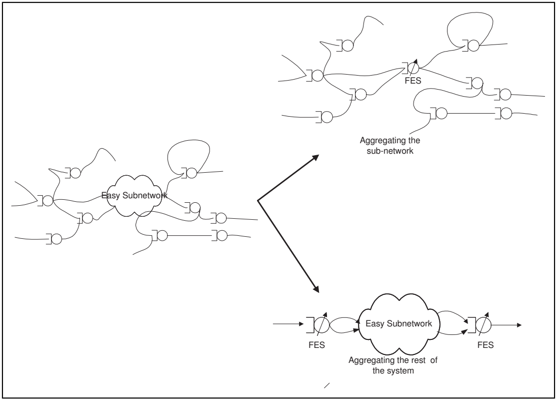


Figure: Flow equivalent server — trick.

Application: A multiprogramming system with virtual memory

- Let us now revisit the multiprogrammed virtual memory system, this time with terminals rather than batch processing:

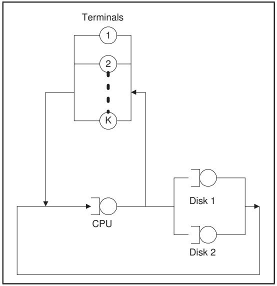


Figure: Multiprogramming system with virtual memory

- The average think time of each user is 30 seconds
- Q: How does the system throughput and response time vary with the number of terminals, K, and at what point does the system start to thrash?
- As users submit jobs, they increase the population at the computer system and consume memory
- The problem is that the performance of the computer subsystem depends on the number of jobs there
- But the number of jobs at the subsystem is governed by the number of terminals
- Specifically, the Disk 1 service rate depends on the population of the entire subnetwork: no simple (product-form) solution

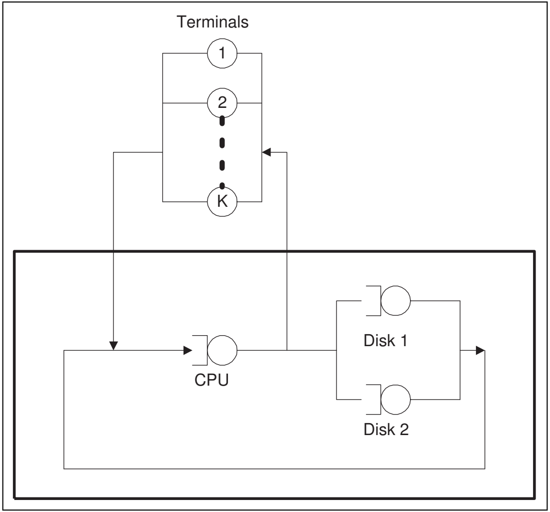


Figure: Multiprogramming system with virtual memory — step 1

- In the batch model we computed previously the throughput, τ_n, when there were n batch jobs being processed
- We can thus aggregate the server subsystem into a *flow-equivalent server* whose service rate at population n is

$$\mu_n = \tau_n$$

- That is

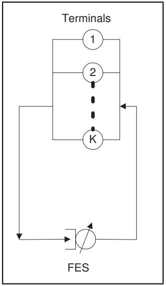


Figure: Multiprogramming system with virtual memory - FES

- approximation because of lack of product-form solution
- The resulting reduced network is a simple M/M/1 queue with state dependent arrival and service rates:

$$\lambda_n = (K - n)\lambda$$
$$\mu_n = \tau_n$$

  where $0 \leq n \leq K$ and where $\lambda = 1/30$ is the submission rate for a single terminal in the "think" state
- Let p_n be the equilibrium probability that there are n jobs at the subsystem (i.e. $K - n$ users in "think" mode)

- The magic formula for the M/M/1 queue gives

$$p_n = \frac{(K-n+1)\lambda}{\mu_n} \frac{(K-n)\lambda}{\mu_{n-1}} \dots \frac{K\lambda}{\mu_1} p_0$$
$$= \lambda^n \frac{K!}{(K-n)!} \prod_{i=1}^{n} \frac{1}{\mu_i}$$

  with the usual constraint that

$$p_0 + p_1 + \dots + p_k = 1$$

- The algorithm requires one loop to compute p_0 and a further loop to find the p_n, $1 \leq n \leq K$
- We are interested here in the throughput and mean response time for a given value of K

- The throughput is:

$$\tau = p_0\mu_0 + p_1\mu_1 + \cdots + p_K\mu_K$$

- We next need to find the mean population of the subsystem:

$$L = 0 \times p_0 + 1 \times p_1 + \cdots + K \times p_K$$

- The mean waiting time then follows from Little's law

$$W = L/\tau$$

- With the same paging model as the batch system earlier, this is what happens to τ as we increase K

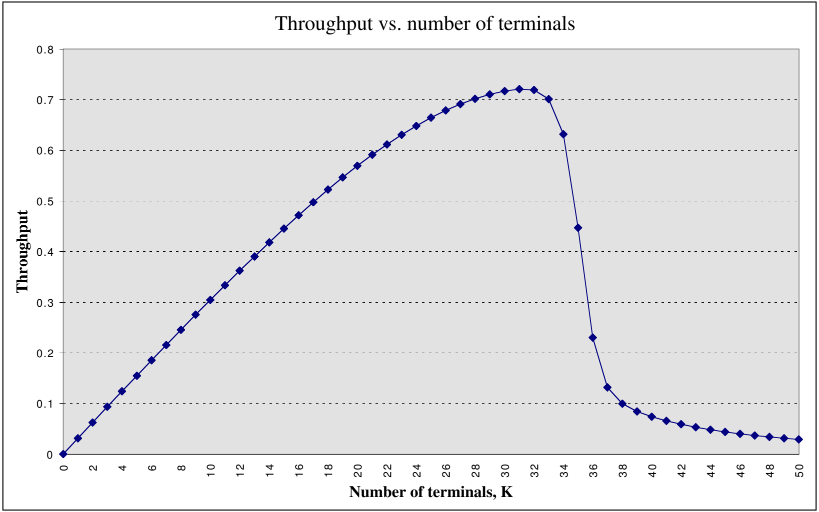


Figure: Throughput vs. number of terminals.

- The saturation point is clear! Beyond this the mean population of the computer subsystem is such that thrashing occurs
- Here is the behaviour of the mean waiting time as we increase W

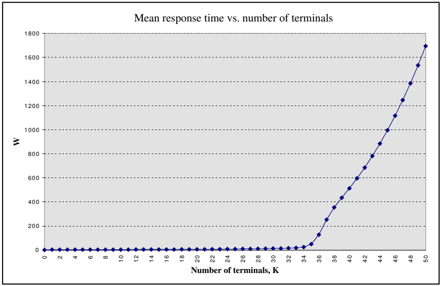


Figure: Mean response time vs. number of terminals.

- And finally (for the record), here is L:

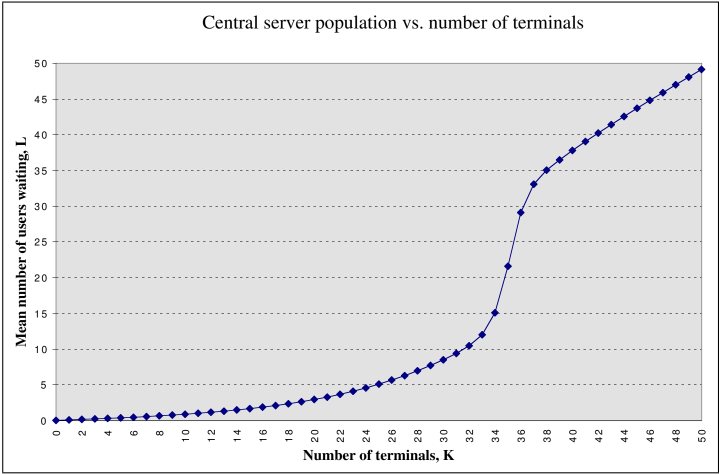


Figure: Central server population vs. number of terminals.

The Markov Modulated Poisson Process (MMPP)

- Poisson process is a good model for many arrival streams
  - more short interarrival times than long
  - superposition of many sparse independent renewal processes is asymptotically Poisson
  - right intuitive properties given minimal information (i.e. arrival *rate*)

- But cannot model many characteristics seen in modern networks
  - multiple traffic types with known switching pattern between types
  - correlated traffic, i.e. non-zero correlation between interarrival times
  - self-similarity

- MMPP accounts for some of these limitations
  - models stochastically changing modes with different rates
  - has non-zero autocorrelation which can be calculated
  - integrates well into queueing models – the MMPP/M/1 queue

Definition

- The MMPP is a Poisson process with rate that depends on the state (called *phase*) of an independent continuous time finite Markov chain $X(t)$
  - N is the number of phases
  - λ_k is the rate of the Poisson process in phase k
  - so the probability of an arrival in a small interval of length h in phase k is

$$\lambda_k h + o(h)$$

- The modulating Markov chain $X(t)$ is assumed irreducible and so has a steady state
    - let the generator matrix (instantaneous transition rate matrix) be Q
    - let the equilibrium probability of being in state k be

$$\pi_k = \lim_{t \to \infty} P(X(t) = k)$$

- Average arrival rate in steady state is then

$$\overline{\lambda} = \sum_{k=1}^{N} \pi_k \lambda_k$$

State at Arrival Instants

- Important for response time distributions
- Do *not* have the Random Observer Property
- In a long period of time T at equilibrium
  - expected time spent in state k is $\pi_k T$
  - expected number of arrivals that see state $k = \lambda_k \pi_k T$
  - total number of arrivals has mean $\bar{\lambda} T$
- So probability that an arrival sees state k in steady state is

$$\pi'_k = \frac{\lambda_k \pi_k}{\bar{\lambda}}$$

The MMPP/M/1 queue

- Suppose the queue has MMPP arrival process as defined above and server with exponential service times of constant mean μ^{-1}
- Simple MP
  - easy to write down the balance equations
  - apply steady state theorem
  - conditions harder to derive rigorously but expect that the queue has a steady state iff

$$\bar{\lambda} < \mu$$

- The MP has a 2-dimensional state space
  - phase of the MMPP horizontally (say)
  - queue length vertically
  - infinite lattice strip for unbounded queue
- Let the steady state probability for phase k and queue length j be denoted by π_{jk} and let the vector $\mathbf{v}_j = (\pi_{j1}, \pi_{j2}, \ldots, \pi_{jN})$
- Balance equations are, for state $(i, j+1)$

$$\pi_{j+1,i}\left(\lambda_i + \mu + \sum_{k \neq i} q_{ik}\right) = \sum_{k \neq i} \pi_{j+1,k} q_{ki} + \lambda_i \pi_{j,i} + \mu \pi_{j+2,i}$$

The MMPP/M/1 queue (continued)

- This may be written in matrix form as

$$\mathbf{v}_j\Lambda + \mathbf{v}_{j+1}(Q - \Lambda - \mu I) + \mu\mathbf{v}_{j+2}I = \mathbf{0}$$

  where I is the unit $N \times N$ matrix and Λ is the diagonal matrix with the arrival rates λ_k down the diagonal, i.e. $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$

- For states $(i, 0)$ we get

$$\mathbf{v}_0(Q - \Lambda) + \mu\mathbf{v}_1 I = \mathbf{0}$$

- The normalising equation is

$$\sum_{j=0}^{\infty} \mathbf{v}_j \cdot (1, 1, \ldots, 1) = 1$$

- These equations give the unique equilibrium probabilities if $\overline{\lambda} < \mu$

Exact solution methods

- We have a matrix recurrence relation of the form

$$\mathbf{v}_j Q_0 + \mathbf{v}_{j+1} Q_1 + \mathbf{v}_{j+2} Q_2 = \mathbf{0}$$

  for $j \geq 0$ and constant matrices Q_0, Q_1, Q_2

- **Spectral Analysis** (Chakka and Mitrani; see Mitrani's book, last chapter)

- find the eigenvalues ξ_k and eigenvectors $\mathbf{e}_k$ of a certain $N-$dimensional matrix given by Q and $(\lambda_1, \ldots, \lambda_N)$
- solution can then be shown to be, for a queue of unbounded capacity

  $$\pi = \sum_{k=1}^{N} a_k \xi^k \mathbf{e}_k$$

  where the a_k are constants determined by the equations at queue length 0 and normalisation
- more complex result for queues of finite capacity

- **Matrix Geometric Method**
  - Closely related to Spectral Analysis but a completely different approach
  - Try a solution of the form $\mathbf{v}_j = M^j \mathbf{u}$ for some matrix M and vector $\mathbf{u}$
  - Then solve
    $$Q_0 + MQ_1 + M^2 Q_2 = 0$$
    for M
  - $\mathbf{u}$ from the other equations

Approximate solution methods

- It often happens that phase changes are rare in comparison to arrivals and departures
  - rates in Q-matrix $<< \lambda_1, \dots, \lambda_N, \mu$
  - e.g. phase changes represent change of traffic type, such as data, video, voice
- Often then a good approximation to use (an approximate) decomposition, cf. queueing networks

- For each phase k, solve the M/M/1 queue assuming the arrival rate is constant, λ_k, giving equilibrium probability $p_j^{(k)}$ for queue length j, $1 \leq k \leq N$
  - i.e. solve all the 'columns' in the 2-D state space
  - can only work if every $\lambda_k < \mu$
- Then solve the phase process for the equilibrium probability r_k of being in phase k, i.e. solve

$$\mathbf{r} \cdot Q = \mathbf{0}$$

- Approximate the solution by

$$\pi_{jk} = p_j^{(k)} r_k$$

Fitting MMPPs

- MMPP can be fitted to observed data by matching:
  - average arrival rate $\overline{\lambda}$ (or mean interarrival time)
  - higher moments of interarrival time (variance etc.)
  - autocorrelation function (ACF)
  - Hurst parameter, a measure of self-similarity

- Matching moments is easy but often gives a poor match on correlation, the main reason for using the MMPP in the first place!

- Can use ACF to model correlated traffic directly
    - next slide
    - computationally difficult and expensive
    - possibly use spectral methods, cf. Fourier transforms
- Hurst parameter gives a good representation of 'how correlated' traffic is, but only contributes one parameter in the matching process
    - Better used to validate another matching process

Autocorrelation function of the MMPP

- The autocorrelation function at lag $k \geq 1$ for the above MMPP is

$$\rho_k = \frac{\pi^* R^{-2}\Lambda[\{I - \mathbf{e}\pi^* R^{-1}\Lambda\}\{R^{-1}\Lambda\}^{k-1}]R^{-2}\Lambda\mathbf{e}}{2\pi^* R^{-3}\Lambda\mathbf{e} - (\pi^* R^{-2}\Lambda\mathbf{e})^2}$$

  where the matrix $R = \Lambda - Q$, $\mathbf{e} = (1, 1, \ldots, 1)$ and $\pi^* = \frac{\pi\Lambda}{\pi.\lambda}$.

- Autocorrelation function is easily calculated by the above formula, *given* the parameters of the MMPP

- But the converse, to determine the parameters from the ACF, is hard!
    - gives non-linear equations in several variables (the sought-after parameters)
    - fixed point methods of solution tend to be unstable
    - research area

Derivation of the Poisson process

The proof of the claim is by induction on the integers. Let $P_n(t) = (P(N_t = n)$.

<u>Base case $(n = 0)$</u>

$$\begin{aligned}
P_0(t+h) &= P(N_t = 0 \;\&\; \text{no arrivals in } (t, t+h)) \\
&= P_0(t)(1 - \lambda h + o(h)) \qquad \text{(by independence)} \\
\implies \frac{P_0(t+h) - P_0(t)}{h} &= -\lambda P_0(t) + o(1) \\
P_0'(t) &= -\lambda P_0(t) \qquad \text{as } h \to 0 \\
\implies P_0(t) &= Ce^{-\lambda t} \\
P_0(0) &= P(N_0 = 0) = 1 \text{ therefore } C = 1
\end{aligned}$$

Inductive Step ($n > 0$)

$$\begin{aligned}
P_n(t+h) &= P_n(t)P_0(h) + P_{n-1}(h)P_1(h) + o(h) \\
&= (1-\lambda h)P_n(t) + \lambda h P_{n-1}(t) + o(h) \\
\implies P_n'(t) &= -\lambda P_n(t) + \lambda P_{n-1}(t) \\
\implies \frac{d}{dt}(e^{\lambda t}P_n(t)) &= \lambda e^{\lambda t} P_{n-1}(t)
\end{aligned}$$

Inductive hypothesis:

$$P_{n-1}(t) = e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}$$

This is OK for $n = 1$. Then

$$\frac{d}{dt}(e^{\lambda t} P_n(t)) = \frac{(\lambda)^n}{(n-1)!} t^{n-1}$$
$$\implies e^{\lambda t} P_n(t) = \frac{(\lambda t)^n}{(n)!} + D$$

For $t = 0$, $P_n(t) = 0$ for $n \geq 2$ and so $D = 0$. Thus

$$P_n(t) = e^{-\lambda t} \frac{\lambda t}{n!}$$

- Associated with each arc in a state transition diagram is a *rate*, e.g.:

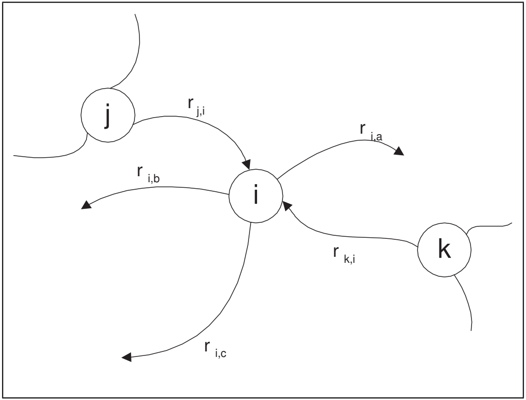


Figure: Transition rates in Markov chains

- The state holding time depends on the departure rates $(1/(r_{i,a} + r_{i,b} + r_{i,c})$ is the mean holding time for state i in the example).
- The *probability flux* from a state i to a state j is the average number of transitions from $i \rightarrow j$ per unit time at equilibrium:

$$\text{Probability flux}(i \rightarrow j) = r_{i,j} p_i$$

- At equilibrium (if it exists) the total flux into each state must balance the total flux out of it
- For example, in the example, for state i alone:

$$(r_{i,a} + r_{i,b} + r_{i,c}) p_i = r_{j,i} p_j + r_{k,i} p_k$$

- Treating all states the same way leads to a set of *linear* equations:

$$\sum_{j \neq i} a_{ij} p_i = \sum_{j \neq i} a_{ji} p_j = -a_{ii} p_i \text{ for all } j \in S$$

- Called the *balance equations* and subject to the *normalising equation*

$$\sum_{i \in S} p_i = 1$$

- The **steady-state theorem** tells us that if these equations have a solution then there is a steady-state and that the solution is unique
- This often provides a short-cut to the solution – see the M/M/1 queue later.