Computer Systems Modelling

Simon Crosby

Twelve lectures (remainder of term) covering:

- Introduction to modelling: what is it, when is it useful ?
- Simple queueing theory: Stochastic processes, Markov chains.
- Birth/death processes, flow balance equations, queueing systems. The M/M/1 queue.
- Series and parallel stages, Multiple queue systems.

Computer Systems Modelling

- Operational analysis: balanced system bounds
- Queuing networks: Bottleneck analysis, improvements based on bottleneck analysis.
- Simulation modelling: algorithms, random variate generation, output analysis.

Computer Systems Modelling

- Measurement techniques: software and hardware monitors.
- Case studies.

Recommended books:

- Kleinrock: Queueing Systems - Vol 1: Theory, Wiley 1975
- Lazowska et al: Quantitative System Performance, Prentice Hall 1984
- Leung: Quantitative Analysis of Computer Systems, Wiley 1988

Why model?

1. A manufacturer may have a range of compatible systems with different performances. A model might be used to determine the best configuration for a particular application.
2. If an existing system is performing inadequately a model can be used to determine how to improve the system and give an indication of the performance of the resulting system.
3. Fundamental design decisions may affect performance of a system. A model can be used as part of the design process to avoid bad decisions.

Introduction

- Begin with simple example: a canning line
- Cans arrive at rate $r = 1$ can/s
- Filling takes 0.9 s/can
- Capacity $c = 1/0.9 = 1.1$ can/s
- What if $r \geq c$? CHAOS
- Mean capacity MUST exceed mean arrival

Introduction

- More general flow problems have random properties: arrival and service times
- This complicates the analysis considerably. Examples:
  - arrival of keystrokes
  - arrival of disk requests
  - arrival and service of network packets
- Single resource – difficult, network of resources seems impossible
- Simplifying assumptions and powerful techniques needed

A toy problem

Database on type A system runs like a drain.

Type B1 system available immediately, and a type B2 system in the future.

What should the company do?

system	CPU time	disk time	total
A	4.6	4.0	8.6
B1	5.1	1.9	7.0
B2	3.1	1.9	5.0

How can modelling help?

Typical performance questions we might ask include:

- How long will a database request wait before CPU service?
- What is the expected utilization of the resource (CPU, disk...)?
- How many requests will typically be queued ?
- What is the expected queue length at time t ?

These in general involve analysis using prob distributions of the time periods involved (arrival, service etc).

Analytical Techniques

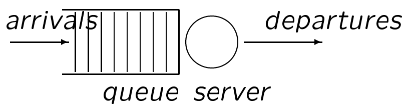
More than 1 way to skin a cat...

- Queueing theory – stochastic processes, queueing networks
- Operational analysis – weaker assumptions but computationally expensive
- Back of the envelope – efficient but gives limited insight

Simulation Techniques

- Construct a computer program which emulates system under study
- The simulation measures its own performance
- Useful for awkward modelling problems with no analytical solution
- Problems include complexity, output analysis and stopping rules.

A Simple Queueing System



- Arrival process
  $A(t) = \mathbb{P}($ inter-arrival time $\leq t)$
- Service process
  $B(x) = \mathbb{P}($ service time $\leq x)$

Queueing Systems Analysis

Various properties of service system are important:

- Storage capacity available for waiting customers
- The number of servers available
- The types of the arriving customers (different demands)
- Queueing discipline used: FCFS, LCFS, SIRO, priority
- Defections, jockeying, balking, bribing...

Queueing Systems Notation

The Kendall notation describes a single queueing system using the notation $A/B/m/k/l$ where:

- A is the inter-arrival time distribution of customers
- B is the service time distribution
- m is the number of parallel servers
- k is the limit on the queue length
- l is the the population size

Queueing Systems Notation

Typical values for A, B include:

- M – exponential distribution
- E_r – r stage Erlangian distribution
- H_r – r stage Hyperexponential distribution
- D – Deterministic
- G – General

Examples:
$M/M/1$ – exponential inter-arrival, exponential service, single server
$M/E_r/1$ – exponential inter-arrival, r stage Erlang service, single server

Probability Theory Refresher

We write the probability that an event will occur as $\mathbb{P}$(event)

$\mathbb{P}(X \leq 0.5)$ is the probability that a rv X is less than or equal to 0.5.

The **joint** probability of two events a, b is denoted $\mathbb{P}(a, b)$

The **conditional** probability of a given that event b has occurred is

$$\mathbb{P}(a|b) \triangleq \frac{\mathbb{P}(a, b)}{\mathbb{P}(b)}$$

To describe the **distribution** from which a rv X is drawn we use 2 functions: the CDF

$$F_X(x) \triangleq \mathbb{P}(X \leq x)$$

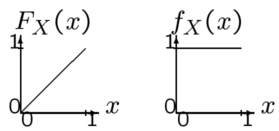
and the pdf

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

with

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy\,, \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(y)dy = 1.$$

Thus for a random variable X uniformly distributed from 0 to 1, the CDF and pdf would be



The **expected value** of X, $\mathbb{E}(X)$ (sometimes called the average, or mean), is just

$$\mathbb{E}(X) \triangleq \overline{X} \triangleq \int_{-\infty}^{\infty} x f_X(x) dx\,.$$

Note that $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Also called the first **moment** of the distribution. The n^{th} **moment** is defined as

$$\mathbb{E}(X^n) \triangleq \overline{X^n} \triangleq \int_{-\infty}^{\infty} x^n f_X(x) dx\,,$$

and the n^{th} **central moment** is

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

The second central moment is known as the **variance**,

$$\sigma_X^2 \triangleq \overline{(X-\overline{X})^2} = \overline{X^2} - (\overline{X})^2.$$

The square root of the variance is σ_X and is called **standard deviation**. A dimensionless quantity, the **coefficient of variation**, is obtained by dividing the standard deviation by the average:

$$C_X \triangleq \frac{\sigma_X}{\overline{X}}\,.$$

Similar definitions apply to the case of discrete random variables taking values x_i $(i \in I)$, say, such that $\sum_{i\in I} P[X = x_i] = 1$. The expected value of X is

$$\mathbb{E}(X) = \sum_{i\in I} x_i P[X = x_i]$$

and similarly with the other moments the integration becomes a summation.

The notion of joint distributions is useful. The joint CDF of two random variables X and Y is

$$F_{XY}(x,y) \triangleq P[X \le x, Y \le y]\,,$$

and the joint pdf is given by

$$f_{XY}(x,y) \triangleq \frac{d^2 F_{XY}(x,y)}{dxdy}\,.$$

The random variables X and Y are said to be **independent** if the the joint pdf factors

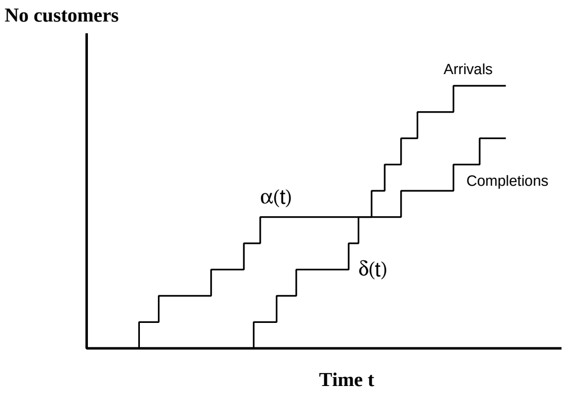
$$f_{XY}(x,y) = f_X(x) f_Y(y)\,.$$

Little's Result

Begin with a neat derivation of a very general result: Little's Result

What is the relationship between the (mean) number of customers in the queue, the arrival rate of customers and the time spent in the queue? Let:

$\alpha(t)$ = number of arrivals in $(0,t)$
$\delta(t)$ = number of departures in $(0,t)$
Then $N(t) = \alpha(t) - \delta(t)$ is the number in the system at t



The area $\gamma(t)$ between the curves $\alpha(t)$ and $\delta(t)$ represents the total time all customers have spent in system in $(0, t)$

Little's Result

Let

$$\lambda_t = \alpha(t)/t \tag{1}$$
$$T_t = \gamma(t)/\alpha(t) \tag{2}$$
$$\bar{N}_t = \gamma(t)/t \tag{3}$$

- the average arrival rate during $(0, t)$
- system time per customer averaged over all customers in $(0, t)$
- average number of customers in system during $(0, t)$

Little's Result

Easily from the above

$$\bar{N}_t = \lambda_t T_t.$$

Assume the following limits exist as $t \to \infty$

$$\lambda = \lim_{t\to\infty} \lambda_t \tag{4}$$
$$T = \lim_{t\to\infty} T_t \tag{5}$$

If these exist then so does

$$\bar{N} = \lambda T \tag{6}$$

In English: the average number in system = average arrival rate × average time in system

This is **Little's result**. The proof makes no assumptions about $A(t), B(t)$, number of servers, or queueing discipline.

First proved by Little in 1961.

Applications of Little's Result

We can re-state this for any boundary of our queueing system: if we apply this to just the queue then we find that

$$\bar{N}_q = \lambda W$$

Where $\bar{N}_q$ is the average number in queue and W is the average waiting time

Applied to the server only

$$\bar{N}_s = \lambda \bar{x}$$

Where $\bar{N}_s$ is average number in service, and $\bar{x} = \mathbb{E}(B(x))$

Note that $T = \bar{x} + W$

Utilization of the Server

For the single server clearly $\rho = \lambda \bar{x}$ is the average utilization of the server. For a m-server this simply becomes $\rho = \lambda \bar{x}/m$

ρ is the average number of busy servers, and we require that $0 \leq \rho < 1$ for system stability of $G/G/1$ queue

Let τ be a large interval. Then we expect with probability $= 1$ that the number of arrivals $\approx \lambda\tau$

Let $p_0 = \mathbb{P}$(server idle at some time). Then server is busy for $\tau(1-p_0)$ seconds and with probability 1 the number of customers served in τ is nearly $(\tau(1-p_0))/\bar{x}$. Thus

$$\lambda\tau \approx \frac{\tau - \tau p_0}{\bar{x}}$$

and for $\tau \to \infty, \lambda\bar{x} \to (1-p_0)$ giving

$$\rho = 1 - p_0 \tag{7}$$

Stochastic Processes

A stochastic process is a function $X(t)$ where t frequently denotes time and whose values are random variables.

$$\{X(t)\} \to \{(w,t)\}$$

where $w \in S$ and S is a sample space.

The PDF for the process $X(t)$ is given by

$$F_X(x,t) \stackrel{\triangle}{=} \mathbb{P}(X(t) \leq x)$$

The joint PDF, $F_{\underline{X}}(\underline{x},\underline{t}), \forall t \in T$ is given by

$$F_{X_1X_2\ldots X_n}(x_1x_2\ldots x_n, t_1t_2\ldots t_n) = \mathbb{P}(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n)$$

We classify stochastic processes by 3 parameters:

- the state space (finite, infinite)
- the index parameter (time)
  - countable $\to$ discrete time process (use X_n in place of $X(t)$)
  - infinite $\to$ continuous process
- statistical dependencies between values of $X(t)$ for various $t \in T$: examples follow

$X(t)$ is **stationary** iff $X(t)$ is invariant to time shifts. For any constant τ

$$F_{\underline{X}}(\underline{x}, \underline{t}+\tau) = F_{\underline{X}}(\underline{x}, \underline{t})$$

$X(t)$ is **wide sense stationary** if $\mathbb{E}(X(t))$ is independent of t and $\mathbb{E}(X(t)X(t+\tau))$ depends only on τ

Markov Processes

Stochastic process can have continuous or discrete state space
A discrete space Markov process $\{X_n = X_1, X_2, \ldots$ is called a Markov chain iff

$$\mathbb{P}(x_{n+1} = X_{n+1} | x_1 = X_1, \ldots, x_n = X_n) = \mathbb{P}(x_{n+1} = X_{n+1} | x_n = X_n) \tag{8}$$

In other words current state depends only on previous state

In continuous time state changes can occur at any time

This, together with (8) places strong conditions on description of time spent in a particular state

We **cannot know** how long process has been in a given state

Other types of Stochastic Process

- Birth Death Process: MC or MP in which transitions occur only between neighboring states
- Semi-Markov Process: relax memoryless condition for state residence time but Markov property holds at transition instants
- Random Walks: refer to Probability 1a notes
- Renewal Process: $X(t)$ counts state transitions in $(0, t)$

The Memoryless Property

We show that for a continuous time MP the state residence time is **exponentially** distributed. For a MC it is **geometrically** distributed.

Let τ be a rv for the time spent in a state E_i. Given $\tau > s$

$$P[\tau > s + t \mid \tau > s] = h(t)\,.$$

For some function $h(t)$ not depending on s. Now

$$\begin{aligned} P[\tau > s + t \mid \tau > s] &= \frac{P[\tau > s + t, \tau > s]}{P[\tau > s]} \\ &= \frac{P[\tau > s + t]}{P[\tau > s]} \\ &= h(t)\,. \end{aligned}$$

Therefore

$$P[\tau > s + t] = P[\tau > s]\, h(t)\,.$$

Now set $s = 0$, and since $P[\tau > 0] = 1$, we have

$$P[\tau > t] = h(t)$$

so that

$$P[\tau > s + t] = P[\tau > s]P[\tau > t]\,. \qquad (9)$$

Now the pdf of τ is given by

$$\begin{aligned} f_\tau(t) &= \frac{d}{dt}P[\tau \le t] \\ &= \frac{d}{dt}(1 - P[\tau > t]) \\ &= -\frac{d}{dt}P[\tau > t]\,. \qquad (10) \end{aligned}$$

Differentiating (9) with respect to s gives,

$$\begin{aligned} \frac{d}{ds}P[\tau > s + t] &= \frac{d}{ds}P[\tau > s]P[\tau > t] \\ &= -f_\tau(s)P[\tau > t]\,. \end{aligned}$$

Set $s = 0$ and divide by $P[\tau > t]$ to give,

$$\frac{dP[\tau > t]}{P[\tau > t]} = -f_\tau(0)ds\,.$$

Now integrate from 0 to t,

$$\begin{aligned} log_e P[\tau > t] &= -f_\tau(0)t \\ P[\tau > t] &= e^{-f_\tau(0)t}\,. \end{aligned}$$

From (10) above,

$$f_\tau(t) = f_\tau(0)e^{-f_\tau(0)t}$$

It is customary to use λ for $f_\tau(0)$, giving the familiar exponential distribution

$$f_\tau(t) = \lambda e^{-\lambda t}\,.$$

The mean of this distribution is

$$\begin{aligned} E[\tau] &= \int_0^\infty t\lambda e^{-\lambda t} dt \\ &= \frac{1}{\lambda} \end{aligned}$$

The variance is

$$\begin{aligned} E[\tau^2] &= \int_0^\infty t^2 \lambda e^{-\lambda t} dt \\ &= \frac{2}{\lambda^2} \end{aligned}$$

and the coefficient of variation is

$$\begin{aligned} C_\tau &\triangleq \frac{\sigma_\tau}{\overline{\tau}} \\ &= 1\,. \end{aligned}$$

We have shown that the **memoryless distribution** is necessarily the **exponential distribution**.

Exercise show for MC that the memoryless property requires the geometric distribution.

Note that the parameter λ can be interpreted as the rate of state transition.

The Poisson Process

- The memoryless arrivals (or inter-event) process is the well known Poisson process with parameter λ

- It can be defined as the dual to the exponential distribution: let $o(\Delta t)$ be a quantity having an order of magnitude less than Δt

- Suppose $\exists \lambda$ constant such that for any small time Δt,

  $\mathbb{P}(0 \text{ arrivals in} t, t + \Delta t) = 1 - \lambda \Delta t + o(\Delta t)$

- events in the interval $(t, t + \Delta t)$ are independent and

  $\mathbb{P}(1\text{arrival in } t, t + \Delta t) = \lambda \Delta t + o(\Delta t)$

- Let

  $A_c(t) = \mathbb{P}(\text{time of next arrival, } x > t_0 + t)$

Compute the inter-arrival time distribution $A(t)$ between two events thus:

$A_c(t + \Delta t) = \mathbb{P}(z > t + \Delta t)$ where $z = x - t_0$

or

$$\begin{aligned} A_c(t + \Delta t) &= \mathbb{P}(z > t)\,\mathbb{P}(\text{ no arrivals in } \Delta t) \\ &= A_c(t)[1 - \lambda \Delta t + o(\Delta t)] \end{aligned}$$

using the assumption of independent events in disjoint intervals.

Rearranging etc.

$$\lim_{\Delta t \to 0} \frac{A_c(t + \Delta t) - A_c(t)}{\Delta t} = -\lambda A_c(t)$$

Thus

$$A_c'(t) = -\lambda A_c(t)$$

Subject to boundary condition $A_c(0) = 1$ we have the following solution

$$A_c(t) = e^{-\lambda t}, t \geq 0$$

And hence

$$A(t) = 1 - e^{-\lambda t}$$

Demonstrating that Poisson arrivals $\rightarrow$ exponential inter-arrivals.

How many arrivals in t?

- Indicate briefly the pmf $f_i(t) (i = 0, 1, 2...)$ which specifies the probability of i arrivals in t
- Note that this probability is *independent* of where the interval begins
- Set $t = m\Delta t$
- Find probability of 1 arrival in i of the m intervals of length Δt and no arrival in the remaining $m - i$

$$\begin{aligned} f_i(t) &= \lim_{\Delta t \to 0} \binom{m}{i} [\lambda \Delta t + o(\Delta t)]^i \\ &\quad [1 - \lambda \Delta t + o(\Delta t)]^{m-i} \end{aligned}$$

where $m = \frac{t}{\Delta t}$

This can be reduced to

$$\begin{aligned} f_i(t) &= \frac{(\lambda t)^i}{i!} \lim_{\Delta m \to \infty} \frac{m(m-1)\dots(m-i+1)}{m^i} \\ &\quad \lim_{\Delta m \to \infty} \left[1 - \frac{\lambda t}{m}\right]^m \end{aligned}$$

From which we obtain

$$f_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t} \qquad (11)$$

Equation (11) defines the Poisson distribution. Note that the inter-arrival time distribution is given by $A(t) = 1 - f_0(t)$.

It is easily shown from Equation (11) that

$$f_i(\Delta t) = \begin{cases} 1 - \lambda \Delta t + o(\Delta t) & i = 0 \\ \lambda \Delta t + o(\Delta t) & i = 1 \\ o(\Delta t) & i > 1 \end{cases}$$

demonstrating the consistency of the definitions for the Poisson Process.

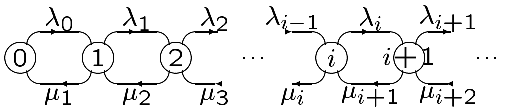
Birth Death Processes

The BDP is a special case of a CTMC in which we allow transitions only between neighboring states

The state space is the set of non-negative integers and we allow only a single birth or death per transition. Thus given $X_n = i$

$$X_{n+1} = \begin{cases} i+1 & \text{birth (or arrival)} \\ i-1 & \text{death (or departure)} \end{cases}$$

We use λ_i to represent the birth rate in state i, and μ_i to represent the death rate in state n



Time Dependent Solution of BDP

We denote by $P_i(t)$ the probability of being in state i at time t

Probability of a birth in an interval Δt when the system starts in state i is assumed to be $\lambda_i \Delta t + o(\Delta t)$

Probability of a death in Δt when the system starts in state i is $\mu_i \Delta t + o(\Delta t)$

Probability of > 1 event in Δt is $o(\Delta t)$

$o(\Delta t)$ denotes a quantity which becomes negligible when compared with Δt as $\Delta t \to 0$

To solve for $P_i(t)$ we write a set of difference equations called the **Chapman Kolmogorov** equations

Chapman Kolmogorov Equations

For $i \geq 1$

$$\begin{aligned} P_i(t+\Delta t) &= P_i(t)(1-\lambda_i \Delta t)(1-\mu_i \Delta t) \\ &+ P_{i+1}(t)(\mu_{i+1}\Delta t)(1-\lambda_{n+1}\Delta t) \\ &+ P_{i-1}(t)(\lambda_{i-1}\Delta t)(1-\mu_{i-1}\Delta t) \\ &+ o(\Delta t) \end{aligned}$$

For $i = 0$

$$\begin{aligned} P_0(t+\Delta t) &= P_0(t)(1-\lambda_0 \Delta t) \\ &+ P_1(t)(\mu_1 \Delta t)(1-\lambda_1 \Delta t) \\ &+ o(\Delta t) \end{aligned}$$

We derive differential-difference equations from these by dividing through by Δt and taking the limit as $\Delta t \to 0$:

$$\begin{aligned} \frac{dP_i(t)}{dt} &= -(\lambda_i + \mu_i)P_i(t) \\ &+ \mu_{i+1}P_{i+1}(t) \\ &+ \lambda_{i-1}P_{i-1}(t) \end{aligned}$$

for $i \geq 1$ and

$$\begin{aligned} \frac{dP_0(t)}{dt} &= -\lambda_0 P_0(t) \\ &+ \mu_1 P_1(t) \end{aligned} \tag{12}$$

The time dependent solution is **extremely difficult** for systems of interest, so we will study the stationary solution

Stationary Solution

We are interested in the long term probabilities after the system has reached an **equilibrium**.

These probabilities are independent of the initial conditions.

System reaches equilibrium if, for all i,

$$\lim_{t\to\infty} P_i(t) = p_i \qquad \text{exists}.$$

The transitions into state i are births in state $i-1$ and deaths in state $i+1$.

Now $dP_i(t)/dt$ is 0 and Equations (12) become

$$\begin{aligned} 0 &= -(\lambda_i + \mu_i)p_i + \mu_{i+1}p_{i+1} + \lambda_{i-1}p_{i-1} \\ 0 &= -\lambda_0 p_0 + \mu_1 p_1 \end{aligned} \tag{13}$$

Rewriting gives

$$\begin{aligned} p_{i+1} &= \frac{\lambda_i + \mu_i}{\mu_{i+1}}p_i - \frac{\lambda_{i-1}}{\mu_{i+1}}p_{i-1} \\ p_1 &= \frac{\lambda_0}{\mu_1}p_0 \end{aligned} \tag{14}$$

For $i \geq 1$.

These equations can be solved iteratively (using induction). Proof is left as an **exercise**

Stochastic Balance

Provides short cut method to obtaining Equation (14)

Require total flow into a state to equal total flow out of a state under steady-state conditions

Intuitively the total flow is the product of the steady state probabilities and the flow rates (λ, μ)

We enter state i at rate $p_{i-1}\lambda_{i-1} + p_{i+1}\mu_{i+1}$.

We exit state i at rate $p_i\lambda_i + p_i\mu_i$.

The equation

$$p_{i-1}\lambda_{i-1} + p_{i+1}\mu_{i+1} = p_i\lambda_i + p_i\mu_i \quad \text{for } i > 0\,.$$

describes flow into state i.

This is the so-called **global balance** equation

the two equations:

$$p_i\lambda_i = p_{i+1}\mu_{i+1} \quad \text{for } i \geq 0$$

describing flow from state i to state $i+1$, and

$$p_i\mu_i = p_{i-1}\lambda_{i-1}$$

describing flow from state i to state $i-1$ are the **detailed balance** equations

Rewriting gives

$$p_{i+1} = \frac{\lambda_i}{\mu_{i+1}} p_i$$

which gives rise to the product solution

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{for } k \geq 0$$

Since the sum of state probabilities must be unity,

$$\begin{aligned} p_0 + \sum_{k=1}^{\infty} p_k &= 1 \\ p_0 + \sum_{k=1}^{\infty} p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} &= 1 \\ p_0 \left[1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right] &= 1 \end{aligned}$$

so that

$$\begin{aligned} p_0 &= \left[1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right]^{-1} \\ p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \end{aligned}$$

which are known as the **general flow balance equations**. It may be shown that an equilibrium distribution p_k always exists if $\frac{\lambda_k}{\mu_k} < 1$ for all $k > K$ for some fixed K.

The M/M/1 Queue

The BDP maps well onto our domain of study – queueing systems

Births represent arrivals to queue, deaths represent departures as customers finish service

The M/M/1 queue is an infinite customer system, with infinite waiting room, and a state independent service rate

This means that $\lambda_i = \lambda$ and $\mu_i = \mu$ for all i and we can solve the balance equations:

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} \\ &= p_0 \left(\frac{\lambda}{\mu} \right)^k . \end{aligned}$$

Writing $\rho = \frac{\lambda}{\mu}$

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

Recall that these results are the same as we derived in the first lecture for the $G/G/1$ queue: p_0 is the probability that there is no customer in the system and $1 - p_0 = \rho$ is the **utilization** of the server

Performance measures

The number in the system is geometrically distributed

$$p_k = (1-\rho)\rho^k, \qquad k = 0, 1, 2, \ldots$$

If $\rho > 1$, i.e. if $\lambda > \mu$ the system will not reach equilibrium

What is the average number of customers, $\overline{N}$ in the system?

$$\begin{aligned} \overline{N} &= \sum_{k=0}^{\infty} k p_k \\ &= \sum_{k=0}^{\infty} k(1-\rho)\rho^k \\ &= (1-\rho)\rho \frac{\partial}{\partial \rho} \left(\sum_{k=0}^{\infty} \rho^k \right) \\ &= (1-\rho)\rho \frac{\partial}{\partial \rho} \left(\frac{1}{1-\rho} \right) \\ &= \frac{\rho}{1-\rho} \end{aligned}$$

An arriving customer will find, on average $\overline{N}$ in the system, and will spend a time, say $\overline{T}$, in the system. During $\overline{T}$ there will be, on average $\lambda\overline{T}$ arrivals, leaving $\overline{N}$ customers in the queue. Thus

$$\overline{N} = \lambda \overline{T}$$

which is Little's result restated. In our case

$$\begin{aligned} \overline{T} &= \frac{\overline{N}}{\lambda} \\ &= \frac{\frac{1}{\mu}}{1-\rho} \end{aligned}$$

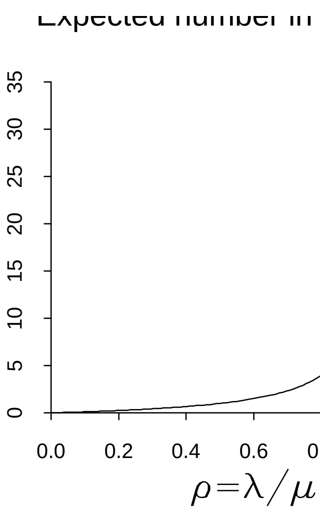
which is the M/M/1 average response time.

Note that

- μ is the service rate
- $\frac{1}{\mu}$ is the average service time
- ρ is the utilization

Performance at high load

At high utilizations ρ approaches one and the response time is unbounded.

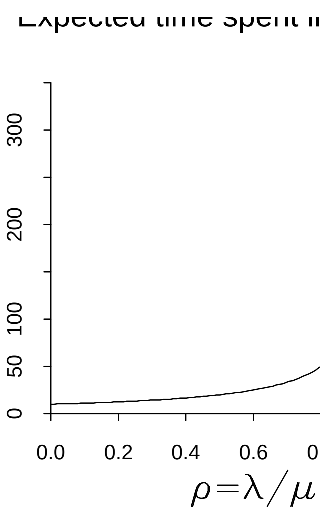


Response time

Consider the effect on response time by increasing the utilization by a constant load of 0.1.

utilization (ρ)			response time ($\overline{T}$)		
old	new	% $\uparrow$	old	new	% $\uparrow$
0.1	0.2	100.0	1.11 $\frac{1}{\mu}$	1.25 $\frac{1}{\mu}$	13
0.5	0.6	20.0	2 $\frac{1}{\mu}$	2.5 $\frac{1}{\mu}$	25
0.8	0.9	12.5	5 $\frac{1}{\mu}$	10 $\frac{1}{\mu}$	100

Predicting response times is very difficult at high loads.



Running systems at maximum utilization may please the **providers**, but it doesn't please the **users**.

M/M/1 – an example

H.R Cutt barber shop, no appointment needed!

Customers served FCFS

On Saturday mornings he is very busy and is considering hiring a helper

He measures the (Poisson) arrival rate of customers to be 5 per hour

Customers are prepared to wait, and he spends on average 10 min per cut

What are the average number of customers in the shop, the average number of customers waiting? What percentage of time can a customer receive a cut without waiting?

He has 4 seats in his waiting room. What is the probability that an arriving customer will find no seat and have to stand?

Example Solution

The average number of customers in the shop: $\lambda = 5$ per hour, $\mu = 6$ per hour So

$$\rho = 5/6, \text{ and hence } \bar{N} = 5$$

Average number of customers waiting:

$$\bar{N}_q = 4\frac{1}{6}$$

What percentage of time can a customer receive a cut without waiting?

$$p_0 = 1 - \rho = \frac{1}{6}$$

No seat

$$\begin{aligned} \mathbb{P}(\text{no seat}) &= \mathbb{P}(N \geq 5) \\ &= \rho^5 \\ &\approx 0.402 \end{aligned}$$

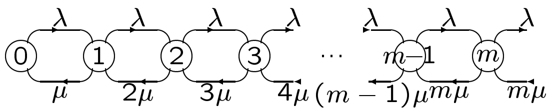
M/M/m — m servers

This is just like the M/M/1 system, except that there are m servers

For all $k, \lambda_k = \lambda$, but now the service rate is a function of the number of customers in the system

$$\mu_k = \begin{cases} k\mu & \text{if } 0 < k \leq m \\ m\mu & \text{if } k \geq m \end{cases}$$

For an equilibrium distribution to exist we require that $\frac{\lambda}{m\mu} < 1$.



M/M/m Performance Measures

Like the M/M/1 system, except that there are m servers

As before, for all $k, \lambda_k = \lambda$, but

$$\mu_k = \begin{cases} k\mu & \text{if } 0 < k \leq m \\ m\mu & \text{if } k \geq m \end{cases}$$

We require that $\frac{\lambda}{m\mu} < 1$.

The difference equations are given by:

$$\begin{aligned} 0 &= \mu p_1 - \lambda p_0 \\ 0 &= (k+1)\mu p_{k+1} - (\lambda + k\mu)p_k + \lambda p_{k-1} \quad \text{if } 0 < k \leq \\ 0 &= m\mu p_{k+1} - (\lambda + m\mu)p_k + \lambda p_{k-1} \quad \text{if } k \geq m \end{aligned}$$

From which we may easily solve iteratively for the p_k as follows:

$$\begin{aligned} p_1 &= \lambda/\mu p_0 \\ p_2 &= 1/2(\lambda/\mu)^2 p_0 \\ &\vdots \\ p_m &= 1/m!(\lambda/\mu)^m p_0 \\ p_{m+1} &= 1/(mm!)(\lambda/\mu)^{m+1} p_0 \\ &\vdots \end{aligned}$$

This gives us the general expression for p_k:

$$\begin{aligned} p_k &= (\lambda/\mu)^k \frac{p_0}{k!} \quad \text{if } k \leq m \\ &= (\lambda/\mu)^k \frac{p_0}{m!m^{k-m}} \quad \text{if } k \geq m \end{aligned}$$

We also require that $\sum_{k=0}^{\infty} p_k = 1$, we may solve for p_0:

$$\begin{aligned} p_0 &= \left[\sum_{k=0}^{m-1} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \frac{m^m}{m!} \sum_{k=m}^{\infty} \left(\frac{\lambda}{m\mu}\right)^k \right]^{-1} \\ &= \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{(1-\rho)m!} \right]^{-1}, \end{aligned}$$

making use of the fact that $\rho = \frac{\lambda}{m\mu} < 1$, whence the second term in the equation for p_0 converges:

$$
\begin{aligned}
\sum_{k=m}^{\infty} \rho^k &= \sum_{k=0}^{\infty} \rho^k - \sum_{k=0}^{m-1} \rho^k \\
&= \frac{1}{1-\rho} - \frac{1-\rho^m}{1-\rho} \\
&= \frac{\rho^m}{1-\rho}
\end{aligned}
$$

Easiest to start with the derivation of $\bar{N}_q$

$$
\begin{aligned}
\bar{N}_q &= \sum_{n=m+1}^{\infty} (n-m) p_n \\
&= \sum_{n=m+1}^{\infty} \frac{(n-m) m^m \rho^n p_0}{m!} \\
&= \frac{(m\rho)^m p_0}{m!} \sum_{n=m+1}^{\infty} (n-m) \rho^{n-m} \\
&= \frac{(m\rho)^m p_0}{m!} \rho \sum_{k=1}^{\infty} k \rho^{k-1} \\
&= \frac{(m\rho)^m p_0}{m!} \rho \frac{\partial}{\partial \rho} \left(\frac{1}{1-\rho} \right) \\
&= \frac{(m\rho)^m p_0 \rho}{m!(1-\rho)^2} \\
&= \frac{\rho p_m}{(1-\rho)^2}
\end{aligned}
$$

With $\bar{N}_q$, we can use Little's formula to solve: $\bar{N}_q = \bar{T}_q \lambda$, and thence obtain performance measures for the queue using $\bar{T} = \bar{T}_q + \frac{1}{\mu}$.

M/M/1/K — finite capacity

If there are K customers in the system, any further arrivals are lost.

Now for $k \geq K$, $\lambda_k = 0$ and for $k > K$, $p_k = 0$

Using the equations from the M/M/1 queueing system, but limiting the summation, and again writing $\rho = \frac{\lambda}{\mu}$,

$$p_k = p_0 \rho^k \quad \text{for } k \leq K$$

$$
\begin{aligned}
p_0 &= \frac{1}{1+\sum_{k=1}^{K}\rho^k} \\
&= \frac{1}{1+\frac{\rho-\rho^{K+1}}{1-\rho}} \\
&= \frac{1-\rho}{1-\rho^{K+1}}
\end{aligned}
$$

Note that p_0 is greater than in the M/M/1 case.

For this system with a finite state space an equilibrium distribution always exists whatever the arrival and departure rates.

M/M/1//N - finite population

Single server, unbounded queue capacity and a population of N customers.

We solve this system by modifying the λ_k to model the arrival rate.

Instead of having an arrival rate for the population as a whole, we assign an arrival rate to each customer, say λ.

If there are no customers in the system, then all of them are eligible to be born, so that

$$\lambda_0 = N\lambda\,.$$

As we have more customers in the system, we have less eligible to be born. So that,

$$\lambda_k = (N-k)\lambda, \quad \text{for } 0 \le k \le N\,.$$

With a single server the service rate is constant

$$\mu_k = \mu, \quad \text{for } k \ge 1\,.$$

M/M/1//N – finite population

If there are no customers in the system, then all of them are eligible to be born, so that

$$\lambda_0 = N\lambda\,.$$

As we have more customers in the system, we have less eligible to be born. So that,

$$\lambda_k = (N-k)\lambda, \quad \text{for } 0 \le k \le N\,.$$

With a single server the service rate is constant

$$\mu_k = \mu, \quad \text{for } k \ge 1\,.$$

Solving the difference equations for this queue we get

$$
\begin{aligned}
p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda(N-i)}{\mu} \text{ for } 0 \le k \le N \\
&= \begin{cases} p_0 \rho^k \dfrac{N!}{(N-k)!} & \text{if } 0 < k \le N \\ 0 & \text{if } k > N \end{cases}
\end{aligned}
$$

and

$$p_0 = \left[\sum_{k=0}^{N} \rho^k \frac{N!}{(N-k)!}\right]^{-1}$$

M/M/m/m - m server loss system

An application of this system is to model a link in a telephone network.

Such a link contains m circuits each of which carries a single call.

Suppose that calls arrive at the link according to a Poisson process of rate λ.

If there is a free circuit the call is connected and holds the circuit for an exponentially distributed length of time, with mean $\frac{1}{\mu}$.

At the end of this holding period the call terminates and the circuit is freed.

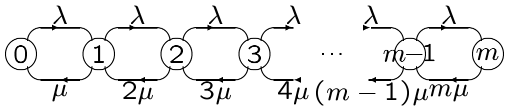
If there are no free circuits then the call is lost.

The birth and death rates are

$$\lambda_k = \begin{cases} \lambda & k < m \\ 0 & k \geq m \end{cases}$$

and

$$\mu_k = k\mu \quad \text{for } 1 \leq k \leq m\,.$$



The flow balance equations give for $k \leq m$

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \\ &= p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \\ &= p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \end{aligned}$$

and $p_k = 0$ for $k > m$.

Solving for p_0 gives

$$\begin{aligned} \sum_{k=0}^{m} p_k &= 1 \\ p_0 \sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} &= 1 \end{aligned}$$

so that

$$p_0 = \left[\sum_{k=0}^{m} \frac{\lambda}{\mu}^k \frac{1}{k!}\right]^{-1}.$$

The probability that an arriving call finds all circuits occupied, p_m, is called the *loss probability* for the telephone link. Thus,

$$\begin{aligned} p_m &= p_0 \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} \\ &= \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} \left[\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\right]^{-1}. \end{aligned}$$

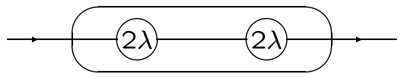
This expression for the loss probability is known as **Erlang's B formula**

Extensions to Birth Death Models

First we relax our constraints on the arrival process distribution.

We want to model systems in which the coefficient of variation of the interarrival time is less than one.

Consider a system in which a birth occurs in two stages.



Each stage has an exponentially distributed residence time.

If the desired birth rate is λ, then let each stage have a rate 2λ.

The average time to get through the combined birth process will be

$$\overline{\tau} = \frac{1}{2\lambda} + \frac{1}{2\lambda} = \frac{1}{\lambda}.$$

Since each stage has exponentially distributed residence times, the variance of each stage is

$$\sigma^2_{single} = \frac{1}{(2\lambda)^2}.$$

The two stages are independent, so the variance of τ, the time to get through both stages is

$$\sigma^2_{\tau} = \frac{1}{(2\lambda)^2} + \frac{1}{(2\lambda)^2} = \frac{1}{2\lambda^2}$$

So the coefficient of variation is

$$C_{\tau} = \frac{\sqrt{\frac{1}{2\lambda^2}}}{\frac{1}{\lambda}} = \frac{1}{\sqrt{2}}.$$

In general, if we use r stages each with rate $r\lambda$ we get an average time through all stages of $\frac{1}{\lambda}$ and a coefficient of variation of $\frac{1}{\sqrt{r}}$.

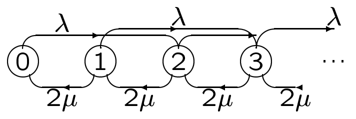
The distribution that describes this process is the Erlangian distribution denoted E_r.

Example, M/E$_2$/1

Allow the state of the process to represent the number of stages remaining to be served.

An arrival increases the number of stages remaining to be served by 2 and occurs at rate λ

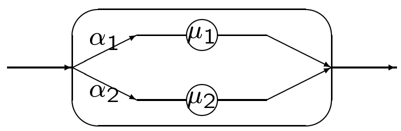
A departure from a stage reduces the number of stages to be served by 1 and occurs at rate 2μ.



Parallel Servers

Combining stages in series reduces the coefficient of variation.

If, instead, we combine them in parallel with a probability α_i of choosing the i^{th} parallel stage we get a service distribution with coefficient of variation larger than 1.



The coefficient of variation is given by

$$C_{\tau} = \frac{2\sum_{i=1}^{r} \frac{\alpha_i}{\mu_i^2}}{\left(\sum_{i=1}^{r} \frac{\alpha_i}{\mu_i}\right)^2} - 1 \geq 1.$$

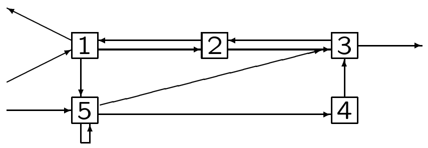
Queueing Networks

We have seen the solution of several queueing systems of interest

In general we want to solve a system of such queues representing a real world performance problem e.g. a distributed computing system

We represent the system under study as a network of connected queues

Customers move (instantaneously) between service centres where they receive service



Queueing Networks

Depending on whether we allow external arrivals and departures, the queueing network is either *open* or *closed*

An *open* network is called feed-forward if no customer can visit the same node twice

It is of key importance that we understand the output process from each queue – this forms the input to the next queue

In general the inter-departure time distribution from a queue is different from the service time distribution

The point process represented by departures is in general very complex, but a special case exists...

The Output of an M/M/m Queue

We have studied this queue in some detail – in particular we know that the utilization of the server is

$$\rho = \frac{\lambda}{m\mu}$$

Focus on the departure process during a small interval Δt, and compute the expected number of departures n:

- If the system is in state 1, the departure rate is $\mu \to n = \mu\Delta t$
- if the system is in state $1 < k \leq m \to n = k\mu\Delta t$
- if the system is in state $m \leq k, \to n = m\mu\Delta t$

Burke's Theorem

Averaging over the different states and their probabilities

$$\begin{aligned} n &= \mu\Delta t \left(\sum_{k=1}^{m} kp_k + \sum_{k>m} mp_k \right) \\ &= m\mu\rho\Delta t \\ &= \lambda\Delta t \end{aligned}$$

Further, because the service distribution is exponential, events in different intervals are independent

We thus conclude that the output process from an M/M/m queue is Poisson with rate λ

Consequences

Burke's Theorem implies that the input to each node in a feed-forward network is Poisson

This means that each queue can be solved in isolation

Although in general the number of customers at each node is independent, the waiting times of customers $\bar{T}_{q_i}$ are not

In spite of this, mean quantities may be obtained for the network by combining the individual means

Finally, terminal nodes may have general service time distributions

Open Queueing Networks

Allow customers to arrive from an external source, with rate γ

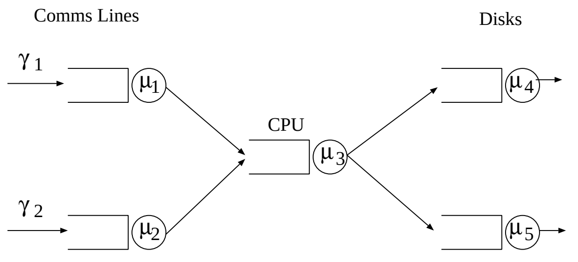
Customers leave the network on completion

Assume we have N nodes, each a M/M/m queue (our examples will use M/M/1) with infinite waiting room

Each server i has exponential service time with mean $1/\mu_i$, except possibly the terminal nodes

Customers arrive (from outside) as a Poisson stream at node i at rate γ_i

Example



Suppose that the service rates of all queues are exponential $\mu_i, 1 \leq i \leq 5$

The split between storage devices is equal

Determine the average response time for a message through the system

Example

From our earlier results, the mean response times at nodes 1 and 2 are

$$\bar{T}_1 = \frac{1}{(\mu_1 - \gamma_1)}$$

$$\bar{T}_2 = \frac{1}{(\mu_2 - \gamma_2)}$$

Similarly, given that the combined output from nodes 1 and 2 is Poisson with rate $\gamma_1 + \gamma_2$,

$$\bar{T}_3 = \frac{1}{(\mu_3 - (\gamma_1 + \gamma_2))}$$

Finally, the inputs to nodes 4 and 5 is Poisson at rate $0.5(\gamma_1 + \gamma_2)$ thus

$$\bar{T}_{4,5} = \frac{1}{(\mu_{4,5} - 0.5(\gamma_1 + \gamma_2))}$$

Example

Hence the mean response times for messages entering at nodes 1 and 2 respectively are

$$\bar{T}' = \bar{T}_1 + \bar{T}_3 + 0.5(\bar{T}_4 + \bar{T}_5)$$

$$\bar{T}'' = \bar{T}_2 + \bar{T}_3 + 0.5(\bar{T}_4 + \bar{T}_5)$$

Since the proportion of traffic entering at node $i = 1, 2$ is

$$\frac{\gamma_i}{(\gamma_1 + \gamma_2)}$$

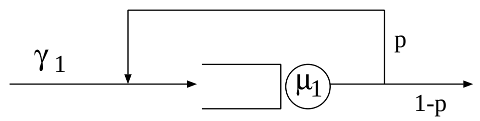
The required result is

$$\bar{T} = \frac{(\gamma_1 \bar{T}' + \gamma_2 \bar{T}'')}{(\gamma_1 + \gamma_2)}$$

General Topologies

We might be tempted to think that a more general topology in which feedback is allowed will also have the convenient Poisson arrival process at each node

Unfortunately this is not the case – feedback introduces dependence on the past



Consider the case when $p \approx 1$ and μ_1 is high

The memoryless property is violated

Luckily, it can be shown that each node does behave **as though** it were subject to a Poisson input

Jackson Networks

Define a more general topology network, called a Jackson Network:

A customer completing at node i moves to node j with probability q_{ij} for $(i, j = 1, 2, \ldots, N)$

Where

$$\sum_{j=1}^{N} q_{ij} \leq 1$$

A job leaves the network from node i with probability

$$q_{i0} = 1 - \sum_{j=1}^{N} q_{ij} \, .$$

The probabilities q_{ij} are called the

routing probabilities

Written as an $N \times N$ matrix this is called the routing matrix $Q = (q_{ij})$

The system state is $(k_1, k_2, \ldots, k_N)$, where k_i is the number of jobs present at node i

Jackson's Theorem

Essentially: In a Jackson Network each node behaves as if it were independent with Poisson input as far as the following are concerned:

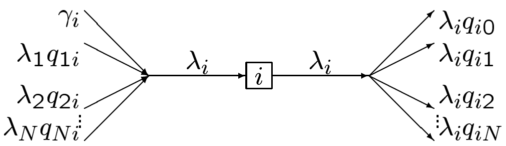
- the average number of customers at the node, and
- the average arrival rate to each node j (which is the sum of the external and internal arrival rates)

Let λ_j be the average number of arrivals to node j per unit time

On average the departure count per unit time is therefore λ_j

A function q_{ji} go to node i

The rate of traffic from node j to node i is thus $\lambda_j q_{ji}$



Solution of a Jackson Network

Adding together all contributions,

$$\lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j q_{ji} \quad i = 1, 2, \ldots, N\,.$$

These are known as the *traffic equations*

A necessary and sufficient condition for the existence of an equilibrium distribution of customers in the network is that

$$\rho_i \stackrel{\triangle}{=} \frac{\lambda_i}{\mu_i} < 1$$

where λ_i is the solution of the traffic equations.

Let $p(k_1, k_2, \ldots, k_N)$ be the steady state equilibrium distribution. Then *Jackson's theorem* states that

$$p(k_1, k_2, \ldots, k_N) = p_1(k_1) p_2(k_2) \cdots p_N(k_N)$$

where $p_i(k_i)$ is the equilibrium distribution that there are k_i jobs in an M/M/m queue with traffic intensity ρ_i.

Implications

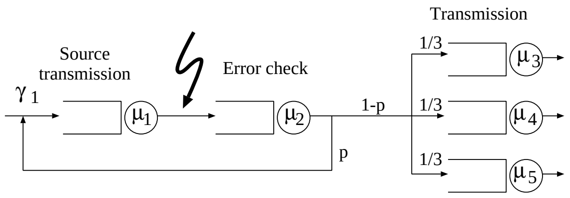
Jackson's theorem has some important implications.

1. The numbers of jobs at the various nodes are independent.

2. Node i behaves as if subjected to a Poisson stream with rate λ_i.

Jackson's theorem applies equally to the case where node i has n_i servers and so the nodes behave as independent M/M/n_i queues

Example – A Router of Sorts



Similar model – except that messages with errors are returned to the first node for correction, with probability p

Each of the 3 transmission lines has a different transmission rate μ_i and message sizes are exponentially distributed

Determine

- the average total delay for a packet
- the average total number of messages awaiting transmission

Solution

First determine input into each node using the traffic equations:

$$\begin{aligned}
\lambda_1 &= \gamma_1 + p\lambda_2 \\
\lambda_2 &= \lambda_1 \\
\lambda_3 &= \lambda_2(1-p)/3 \\
\lambda_4 &= \lambda_2(1-p)/3 \\
\lambda_5 &= \lambda_2(1-p)/3
\end{aligned}$$

In other words

$$\begin{aligned}
\lambda_1 &= \lambda_2 = \frac{\gamma}{1-p} \\
\lambda_3 &= \lambda_4 = \lambda_5 = \frac{\gamma}{3}
\end{aligned}$$

With the (obvious) check that

$$\gamma = \sum_{i=3}^{5} \lambda_i$$

From these we can calculate the response times

$$\begin{aligned}
\bar{T}_1 &= \frac{1}{(\mu_1 - \gamma/(1-p))} \\
\bar{T}_2 &= \frac{1}{(\mu_2 - \gamma/(1-p))} \\
\bar{T}_i &= \frac{1}{(\mu_i - \gamma/3)}
\end{aligned}$$

for $i = 3, 4, 5$

The number of times a message has to join the switching centre queue is geometrically distributed with

$\mathbb{P}$(message retransmitted n times) $= p^n(1-p)$ and with mean $1/(1-p)$

Thus the mean total delay through nodes 1 and 2 is

$$\frac{\bar{T}_1 + \bar{T}_2}{1-p}$$

The remaining average delay is $(\sum_{i=3}^{5} \bar{T}_i)/3$

Thus the mean total delay which we wanted is

$$\bar{T} = \frac{\bar{T}_1 + \bar{T}_2}{1-p} + \frac{\sum_{i=3}^{5} \bar{T}_i}{3}$$

Next, the mean number of messages in queue i is, by Little's law

$$\bar{N}_i = \lambda_i \bar{T}_i$$

Giving our final solution for the output queues of

$$\sum_{i=3}^{5} \bar{N}_i = \frac{\gamma}{3(1-p)} \sum_{i=3}^{5} \bar{T}_i$$

Closed Queueing Networks

Frequently used to model systems at high load or where a limited, constant number of jobs is admitted to service

No external arrivals or departures

Now the routing probabilities satisfy

$$\sum_{j=1}^{N} q_{ij} = 1\,, \quad i = 1, 2, \ldots, N$$

The number of jobs in the system is always a constant, denoted by K

The states of the system, described by the vector $(k_1, k_2, \ldots, k_N)$, thus satisfy the constraint,

$$\sum_{i=1}^{N} k_i = K$$

The state space, S, is then finite. The number of states is

$$\binom{K+N-1}{N-1}$$

The traffic equations become

$$\lambda_i = \sum_{j=1}^{N} \lambda_j q_{ji}\,, \quad i = 1, 2, \ldots, N$$

With a finite state space there always exists an equilibrium distribution, which is the solution to

$$p(k_1, k_2, \ldots, k_N) \sum_{i=1}^{N} \delta_{k_i - 1} \alpha_i(k_i) \mu_i =$$

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{k_j - 1} \alpha_i(k_i + 1) \mu_i q_{ij} p(\ldots, k_j - 1, .., k_i + 1, \ldots)$$

where δ_k is a unit step with value 1 if $k \geq 0$ and 0 elsewhere

and $\alpha_i(k_i)$ is the number of servers at node i which are busy when k_i customers are present

Analogous to Jackson's theorem for the open network case it may be shown that

$$p(k_1, k_2, \ldots, k_N) = \frac{1}{G} r_1(k_1) r_2(k_2) \cdots r_N(k_N)$$

where $r_i(k_i)$ is the probability that there are k_i jobs in an M/M/1 queue with traffic intensity given by a solution to the traffic equations

Open vs closed networks

The normalization constant G has to be determined by

$$G = \sum_{s \in S} r_1(k_1) r_2(k_2) \cdots r_N(k_N)$$

obtained by summing over all states $s = (k_1, k_2, \ldots, k_n) \in S$.

With closed networks need to compute the normalization constant G - a computationally hard problem

The constraint $\sum_{i=1}^{N} k_i = K$ means that the numbers of jobs in the various queues are no longer independent.

For instance, consider the extreme case where all K jobs are at one node

The M/G/1 Queue

Analysis of the M/M/1 queue was made possible by the assumption of exponential interarrival and service times

The process $X(t)$ describing the number in the system has the Markov property ie.

At t_0 the future behavior of $X(t)$ depends only on $X(t_0)$

Can we honestly convince ourselves that the transmission time for IP traffic at a router conforms to an exponential distribution?

Even the Markovian M/H_2/1 queue might be unable to model the service process in this case.

The M/G/1 Queue

In the M/G/1 queue we retain the Poisson assumption for the arrivals process

We are interested in the analysis of $X(t)$ assuming general service time distributions.

Now, if at t_0 a job is in execution, then the behavior of $X(t), t > 0$, will depend not only on $X(t_0)$ but on the amount of service already received.

Thus $X(t)$ is no longer a Markov process.

A solution

Redefine $X(t)$ so that its state space includes the values of a supplementary variable x that specifies the amount of service already received by the job in execution at t.

The new process $X'(t)$ is a pair (n, x), where

- n is the number in the system at time t
- x is a continuous parameter expressing the service time of the customer in service

The new process now has the Markov property

But the analysis is extremely difficult

A second approach

Key to solution is the recognition of the fact that although $X(t)$ does not have the Markov property at every point in time ...

there exist many embedded sequences of time points at which $X(t)$ does have the Markov property.

Define the time instants (or epochs) $t_1, t_2, \ldots$ so that t_i is the instant immediately after the i^{th}job execution

Since the service and interarrival intervals are mutually independent, $X(t_i)$ at the epochs t_i has the Markov property.

The discrete time process

$$X(t_i), i = 1, 2 \ldots,$$

constitutes a Markov chain which is said to be an *embedded* Markov chain with respect to $X(t)$.

The solution of the M/G/1 model is difficult

We simply present an intuitive derivation

M/G/1 Performance Measures

Let $\bar{s}$ denote the mean service time

Begin with the queueing time of a customer before it receives service, which we all know is given by

$$\bar{T}_q = \bar{N}_q \bar{s} + \bar{x},$$

where $\bar{x}$ is the remaining service time of the customer in service at the arrival of the given customer.

We also know from Little's law that

$$\bar{N}_q = \lambda \bar{T}_q$$

Substituting and solving for $\bar{T}_q$ gives

$$\bar{T}_q = \frac{\bar{x}}{(1 - \lambda \bar{s})}$$

It turns out (using arguments from renewal theory) that

$$\bar{x} = \frac{1}{2} \lambda \bar{s^2}$$

Thus

$$\begin{aligned} \bar{T}_q &= \frac{\lambda \bar{s^2}}{2(1 - \lambda \bar{s})} \\ &= \frac{\lambda \bar{s^2}}{2(1 - \rho)} \end{aligned}$$

where $\rho = \lambda \bar{s}$.

Writing C_s for the coefficient of variation of the service time gives us our solution:

$$\bar{T}_q = \frac{\bar{s} \rho (1 + C_s^2)}{2(1 - \rho)} \qquad (15)$$

Pollaczek-Khintchine Formula

From this we may derive directly the formula for $\bar{T}$

$$\bar{T} = \bar{s}\left[1 + \frac{\rho(1 + C_s^2)}{2(1-\rho)}\right] \qquad (16)$$

This is the *Pollaczek-Khintchine* or PK formula.

From Little's formula, equations for the other performance measures follow directly.

For example, the expected queue length is given by:

$$\bar{N} = \rho + \frac{\rho^2[1 + C_s^2]}{2(1-\rho)} \qquad (17)$$

Some Comments ...

Equation (17) is the mean number of customers that are left in the system by a departing customer.

Since we observe the system only at departure times, during the service time of the departing customer more than one new arrival may have occured.

You might like to construct the Markov chain for this model and obtain our earlier results for the M/M/1 queue.

Note particularly that $\bar{N}$ is determined not only by the mean interarrival and service times but also by $\overline{s^2}$.

Also, $\bar{N}$ increases linearly with $\overline{s^2}$

Considering constant service times, for example

$$\bar{N} = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{\rho(1-\rho/2)}{(1-\rho)}, \qquad (18)$$

compared with the expression $\rho/(1-\rho)$ for exponential service times

For small ρ the difference in the means is small, but as $\rho \to 1$ the two differ by a factor of 2

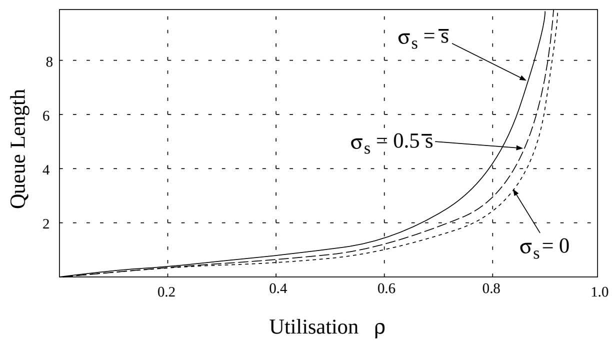
M/G/1 queue length

What values of $C_s = \sigma_s/\bar{s}$ is one likely to encounter?

- 0: This is the case of constant service times
- < 1: This ratio is better than the exponential case

  $\Rightarrow$ M/M/1 performance measures are conservative
- ratio ≈ 1: This corresponds to exponential service time. e.g. packet switched networks with variable size packets, batch customers on a mainframe etc.
- ratio > 1: You need to use an M/G/1 model. e.g. IP traffic

M/G/1 queue length



And non-Poisson arrivals ?

The same considerations apply to the arrival rate

- For Poisson arrivals the ratio $C_a \approx 1$
- If the observed ratio is < 1, arrivals tend to the evenly spaced
- If the ratio is > 1, arrivals tend to cluster

For the same λ, the performance of any queueing system in which arrivals are completely random will be worse than a system in which arrivals are completely predictable.

Leung gives the following rules of thumb

- $1 \leq C_a \leq 1.15$ – Poisson is a reasonable approximation
- $0.85 \leq C_a$ – Process becomes more constant
- $C_a \geq 1.15$ – Arrivals tend to be bunched.

The Poisson assumption over-estimates congestion for processes with $C_a \leq 1$, and under-estimates for $C_a > 1$.

Example

An IP service provider has a bank of dial-in modems

Each recieves messages according to a Poisson distribution with rate $\lambda_j = 10j$ packets per second, for $j = 1 \ldots 5$

Messages are multiplexed onto a single output link and transmitted in arrival order.

Calculate

- the mean datagram response time
- the mean number in the system

Given IP packets have a bimodal size distribution which gives them a mean service time of 5ms with a variance of 4ms

Solution

We recall that the sum of independent Poisson processes is again Poisson

We also immediately recognize that the M/M/1 performance measures, were we to use them, would be **conservative**

Here we get $\lambda = \sum_{i=1}^{5} 10 \times i = 150$ messages per second

Noticing that service times were given in ms, we convert to seconds thus $1/\bar{s} = 200$ messages/second

Immediately we see that $\rho = \lambda\bar{s} = 0.75$

Similarly, $C_s^2 = 4/25 = 0.16$

Using the PK formula for $\bar{T}$ gives,

$$\begin{aligned}\bar{T} &= 5\left[1 + \frac{0.75 \times (1 + 0.16)}{0.5}\right] \\ &= 13.7ms \\ \bar{N} &= 0.75 + \frac{0.75^2 \times (1 + 0.16)}{0.5} \\ &= 2.0551 \\ \bar{N}_q &= 1.011 - 0.75 \\ &= 0.1.305\end{aligned}$$

Applicability

The M/G/1 queueing model will apply to any system with a Poisson arrivals process and a service discipline which is :

- *work conserving* – work is neither created nor destroyed (customers cannot die/leave the queue, split in two etc),
- the service selection mechanism is not based on the *service time* of a particular customer, and
- scheduling of customers is non-preemptive.

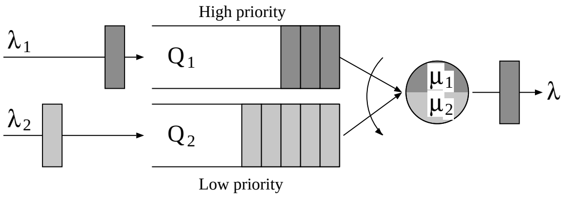
For any system which obeys these constraints, the PK formula applies.

Response time is invariant to the scheduling rule **BUT** only to the *mean value*, and not to the distribution or its other moments.

In the case of $\bar{N}$, however, the invariance applies to the distribution and the higher moments.

$\mathbf{M}_i/\mathbf{G}_i/\mathbf{1}$ Non-preemptive Priority

We will look at the case where a single server queue serves 2 priority classes



Analysis

The key is that the service is non-preemptive.

A low priority customer cannot be interrupted by the arrival of a high priority customer, and will complete service

Denote the two service queues by $Q_i, i = 1, 2$, where Q_1 is the high priority queue.

The arrivals processes to Q_i are Poisson distributed at rates λ_i

The service processes are of general distribution with means $\bar{s_i} = 1/\mu_i$ and variances $\sigma^2_{s_i}, i = 1, 2$.

The expected service time of the average customer, excluding the priority information is

$$\bar{s} = \frac{\lambda_1}{\lambda\mu_1} + \frac{\lambda_2}{\lambda\mu_2}$$

which follows directly from the definition.

Performance Measures

We commence with the derivation of the average time spent in the queue:

$$\bar{T}_q = \frac{\lambda_1}{\lambda}\bar{T}_{q_1} + \frac{\lambda_2}{\lambda}\bar{T}_{q_2}$$

where

$$\bar{T}_{q_1} = \mathbb{E}(Y_0) + \mathbb{E}(Y_1)$$

and

Y_0 = remaining service time of the customer in service

Y_1 = service time of the customers ahead of the arrival

An arriving customer sees the sum of the remaining service of the customer currently

receiving service plus the service time of the customers ahead of it

Clearly

$$\mathbb{E}(Y_1) = \frac{\bar{N}_{q_1}}{\mu_1} = \frac{\lambda_1\bar{T}_{q_1}}{\mu_1} = \rho_1\bar{T}_{q_1}$$

where we have defined $\rho_1 = \lambda_1/\mu_1$. Thus,

$$\bar{T}_{q_1} = \frac{\mathbb{E}(Y_0)}{1-\rho_1}.$$

Similarly, for arrivals to the second queue Q_2,

$$\bar{T}_{q_2} = \mathbb{E}(Y_0) + \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \mathbb{E}(Y_{1'})$$

where

Y_2 = service time of customers in Q_2

$Y_{1'}$ = service time of customers joining Q_1during the wait at Q_2

So

$$\mathbb{E}(Y_2) = \rho_2 \bar{T}_{q_2}$$

and

$$\begin{aligned}
\mathbb{E}(Y_{1'}) &= \mathbb{E}(\text{number of arrivals at } Q_1 \\
&\qquad \text{during service of } Q_2) \\
&\times \ \mathbb{E}(\text{service time of a customer in } Q_1) \\
&= \lambda_1 \bar{T}_{q_2} \times 1/\mu_1
\end{aligned}$$

Putting this all together gives

$$\begin{aligned}
\bar{T}_{q_2} &= \mathbb{E}(Y_0) + \rho_1 \bar{T}_{q_1} + \rho_2 \bar{T}_{q_2} + \rho_1 \bar{T}_{q_1} \\
&= \frac{\mathbb{E}(Y_0)}{(1-\rho_1)(1-\rho_1-\rho_2)}
\end{aligned}$$

where $\rho_1 + \rho_2 < 1$.

Similar to the derivation of the remaining time of a customer in service at the M/G/1 queue:

$$\mathbb{E}(Y_0) = \frac{1}{2}\sum_{i=1}^{2} \lambda_i \overline{s_i^2},$$

which permits the computation of $\bar{T}_q$.

Using Little's law:

$$\begin{aligned}
\bar{N}_{q_i} &= \lambda_i \bar{T}_{q_i} \\
\bar{N}_q &= \lambda \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{T}_{q_i} \\
&= \sum_{i=1}^{2} \bar{N}_{q_i} \\
\bar{T}_i &= \bar{T}_{q_i} + 1/\mu_i \\
\bar{N}_i &= \bar{N}_{q_i} + \rho_i \\
\bar{N} &= \sum_i \bar{N}_i
\end{aligned}$$

These results may be generalized for arbitrary numbers of queues Q_i – see Leung or Kleinrock for the full details.

Example

A second IP service provider who we are considering decides to queue the arriving traffic into two queues, on the basis of their sizes

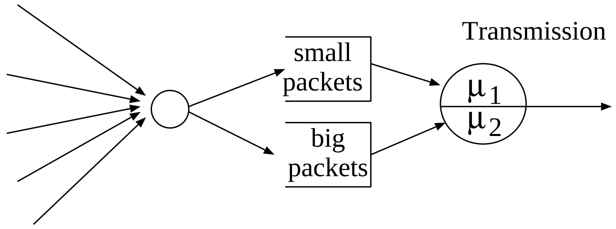
Small datagrams take 2ms to transmit, on average, with standard deviation 2ms, and are given priority

Large datagrams take an average of 17ms to transmit, with standard deviation 2ms, and are given low priority

The arrivals process is as described previously, and small packets make up 80% of the traffic

Compare the performance of this system to one in which the priorities are reversed. Should we use this IP service provider (and why)?

Solution



As before $\lambda = 150$ packets/second

The average service time is, as before

$$\bar{s} = 0.8 \times 2 + 0.2 \times 17 = 5ms$$

Now

$$\bar{T}_q = 0.8\bar{T}_{q_1} + 0.2\bar{T}_{q_2}$$

First we calculate $\mathbb{E}(Y_0)$:

$$\begin{aligned}
\mathbb{E}(Y_0) &= \frac{1}{2}\sum_{i=1}^{2} \lambda_i \bar{s_i^2} \\
&= \frac{1}{2}[0.8 \times 0.15 \times 8 + 0.2 \times 0.15 \times 293] \\
&= 4.875ms
\end{aligned}$$

Now

$$\begin{aligned}
\bar{T}_{q_1} &= \frac{\mathbb{E}(Y_0)}{(1-\rho_1)} \\
&= \frac{4.875}{1-0.24} \\
&= 6.414 \\
\bar{T}_{q_2} &= \frac{\mathbb{E}(Y_0)}{(1-\rho_1)(1-\rho_1-\rho_2)} \\
&= \frac{4.875}{(0.76)(0.76-0.51)} \\
&= 25.658
\end{aligned}$$

So, finally

$$\begin{aligned}
\bar{T}_q &= 0.8 \times 6.414 + 0.2 \times 25.658 \\
&= 10.263
\end{aligned}$$

Also

$$\begin{aligned}
\bar{N}_q &= \lambda \bar{T}_q \\
&= 1.539 \\
\bar{N} &= \bar{N}_q + \rho \\
&= \bar{N}_q + \left(\frac{\lambda_1}{\lambda\mu_1} + \frac{\lambda_2}{\lambda\mu_2}\right) \\
&= 1.539 + 0.75 \\
&= 2.289
\end{aligned}$$