SOLUTION MANUAL FOR SCHEDULING AND CONTROL OF QUEUEING NETWORKS

GIDEON WEISS
The University of Haifa
Draft, Do Not Circulate

Last updated October 2, 2021

Contents

Foreword

An essential part of a textbook are the exercises that accompany each chapter. It is impossible to master the material of a new topic without consulting some of the problems. The student, scholar, researcher may just glance at them to see how they tie up with the text, or figure out an approach, or see if she knows the solution, or sketch a proof, or solve some problem completely – all according to her previous background knowledge and degree of interest in the particular topic.

While writing "Scheduling and Control of Queueing Netowrks" I have solved most of the exercises, as an integral part of writing the text, and in this solution manual I have included solutions to all of them. The purpose of this manual is not to teach the solutions, but rather to serve as an aid for the reader in solving the problems on his own. The reader may find that the style and level of the solutions is somewhat uneven – I solved some easy problems at great length and am more terse on some other, possibly harder, problems. Also, I was less careful in the writing, the student may find many typos and slight errors, which I hope she will ignore. Furthermore, the reader may find that I am completely wrong on some of the problems, or even better, find a new shorter, easier, or more insightful solution to some. I will be most happy to incorporate and acknowledge such contributions in revising this manual.

Many of the exercises are extensions of the material in the text. For exercises which require more work, in particular when an exercise summarizes the results of an entire research paper, a reference is given at the end of the problem formulation in the text. For a few of these exercises I do not provide the solution here, and the student will then have to consult the original research paper.

Enjoy!

Gideon Weiss,

Haifa and San Mateo, September 2021

Part I

The Single Queue

1

Queues and their Simulations, Birth and Death Queues

Exercises

1.1 Use Excel to simulate the following 2 queueing systems, using the same pattern as in example (1.1):

(i) A single queue with two servers. Arrivals are Poisson rate 0.2, service is exponential with mean $m = 8$.

(ii) Two servers in tandem. Arrivals are Poisson rate 0.25, each arrival visits server 1 and then server 2. Service requirements are exponential, with average $m = 3$

1.2 Find an analog to Lindley's equation for M servers, under FCFS service.

Solution:

Define a vector $\mathbf{V}_n = (V_{n,1}, \ldots, V_{n,M})$ to be the vector of remaining workloads on the M servers just before the arrival time of customer n at time A_n, ordered so that $V_{n,1} \leq \ldots \leq V_{n,M}$. This means that all the M servers will complete the work that arrived before customer n, under FCFS policy, at the times $(V_{n,1} + A_n, V_{n,2} + A_n, \ldots, V_{n,M} + A_n)$. Notice that the workload vector is ordered from small to large, so $V_{n,k}$ is not necessarily on machine k.

If customer n requires service for a duration X_n, and customer $n + 1$ arrives after interarrival time $T_{n+1} = A_{n+1} - A_n$, then:

$$\mathbf{V}_{n+1} = \mathcal{R}(\mathbf{V}_n + \mathbf{e}X_n - \mathbf{i}T_{n+1}).$$

where: $\mathbf{e} = (1, 0, \ldots, 0)$, $\mathbf{i} = (1, 1, \ldots, 1)$ and $\mathcal{R}(\cdot)$ is the operator that returns its arguments sorted from smallest to largest.

1.3 Derive the waiting and sojourn times for customers in a stationary M/M/s queue.

Solution:

Let arrival rate be λ, each server has processing rate μ, and denote $\rho = \frac{\lambda}{s\mu}$.

$$p_i = \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} p_0, \quad i = 0, \ldots, s-1,$$

$$p_{s+k} = \left(\frac{\lambda}{s\mu}\right)^k p_s = \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} \left(\frac{\lambda}{s\mu}\right)^k p_0, \quad k = 0, 1, \ldots,$$

and ne needs to calculate numerically the value of p_0,

$$p_0 = \left(\sum_{j=1}^{s-1} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} + \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} \frac{s\mu}{s\mu - \lambda} \right)^{-1}.$$

Then one has, expected number of busy servers:

$$B = \sum_{i=1}^{\infty} \min(i, s) p_i = \sum_{i=1}^{\infty} \frac{\lambda}{\mu} p_{i-1} = \frac{\lambda}{\mu},$$

where the second equality holds term by term, as is easily checked.
Probability that a customer has to wait:

$$\Pi_W = \sum_{k=0}^{\infty} p_{s+k} = p_s \frac{s\mu}{s\mu - \lambda} = \frac{p_s}{1-\rho},$$

expected number of waiting customers:

$$L_q = \sum_{k=0}^{\infty} k p_{s+k} = \frac{p_s \rho}{(1-\rho)^2} = \Pi_W \frac{\rho}{1-\rho},$$

and expected waiting time (by Little's law), number in system, and sojourn time:

$$W_q = L_q/\lambda, \quad L = L_q + B, \quad W = L/\lambda = W_q + 1/\mu.$$

1.4 Calculate the average number of busy servers in an M/M/K/K queue.
Solution:
Using similar calculations to Exercise 1.3 one obtains

$$B = (1 - p_K)\frac{\lambda}{\mu}.$$

1.5 A taxi rank is modeled as follows:
- The rank has space for waiting taxis and for waiting passengers
- Maximal number of taxis in the rank: 2
- Maximal number of passengers in the rank: 3
- Taxis arrive in a Poisson stream of rate $\lambda = 1/6$ cabs per minute.
- Passengers arrive in a Poisson stream of rate $\mu = 1/8$ passengers per minute.
- A passenger that arrives when there are cabs in the rank leaves immediately with a cab. If there is no cab he joins the queue of waiting customers if there is space for waiting, otherwise he leaves without service.
- A cab that arrives when there are passengers at the rank leaves immediately with a passenger. If there are no passengers the cab joins the queue of other cabs, if there is space in the rank, otherwise It leaves without a passenger.

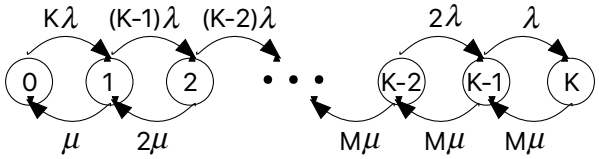
Suggest a Birth and Death description of this system, and calculate the following:

(i) Make a diagram of states and of the transition rates.
(ii) Calculate the stationary distribution of the state of the system. In particular, the probability that the rank is empty, has 1 or 2 or 3 passengers, or has 1 or 2 cabs.
(iii) What fraction of the passengers depart with a cab immediately.
(iv) What fraction of the passengers do not receive service.
(v) What is the distribution of the waiting time of passengers that receive service.

Solution:
Define state of the system as $Q(t) = x$, $x = -2, -1, 0, 1, 2, 3$ where $x > 0$ is number of waiting passengers, $x < 0$ is number of waiting taxi cabs, and $x = 0$ is empty rank. $Q(t)$ is a birth and death process.

1.6 *A K machines, M repairmen queueing system:* A workshop has a total of K machines, with operating time between failures (MTBF) exponential with rate λ. When a machine fails it needs service by a repairman for exponential rate μ repair time. There are $M \leq K$ repairmen. Describe this system as a birth and death queue, illustrate its state and transition rates diagram, and derive its stationary behavior. Performance measures for this system are the fraction of time that machines are operational, and the fraction of time that each repairman is busy. For $\mu/\lambda = 4$, prepare a table of these two performance measures for $K = 3, \ldots, 7$ and $M = 1, \ldots, K$.

Solution:
The K machine M repairmen states and transition rates for $Q(t)$, the number of machines that are down, are illustrated by:



Birth and death rates for K-machines M-repairmen

$$\lambda_n = (K-n)\lambda,$$
$$\mu_n = \begin{cases} n\mu, & n \leq M, \\ M\mu, & n > M, \end{cases}$$
$$n = 0, 1, 2, \ldots, K.$$

These can be used to calculate the stationary distribution of $Q(t)$, that is $\pi_k = \mathbb{P}(Q(t) = k)$.
Fraction of time that machine is operational is $K - \mathbb{E}(Q(t))/K$.
The number of busy repairman is $R(t) = \max(M, Q(t))$, and the fraction of time that each repairman is busy is $\mathbb{E}(R(t))/M$.

For example, if $K = 5$ and $M = 2$, the probabilities for k machines down are:

k	0	1	2	3	4	5
$P(k)$	0.31	0.39	0.20	0.074	0.018	0.0023

Table of machine and repairmen utilization:

K	M	1	2	3	4	5	6	7
3	utilization K	0.732	0.797	0.800	0	0	0	0
	utilization M	0.549	0.299	0.200	0	0	0	0
4	utilization K	0.689	0.791	0.800	0.800	0	0	0
	utilization M	0.689	0.395	0.267	0.200	0	0	0
5	utilization K	0.641	0.781	0.798	0.800	0.800	0	0
	utilization M	0.801	0.488	0.333	0.250	0.200	0	0
6	utilization K	0.589	0.769	0.796	0.800	0.800	0.800	0
	utilization M	0.883	0.576	0.398	0.300	0.240	0.200	0
7	utilization K	0.536	0.753	0.793	0.799	0.800	0.800	0.800
	utilization M	0.937	0.658	0.463	0.350	0.280	0.233	0.200

1.7 A gas station is modeled as follows:

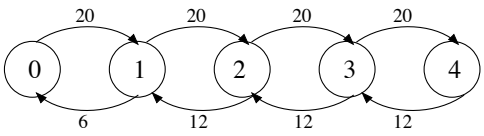
- There are two pumps, and space for a total of 4 cars.
- Cars arrive in a Poisson stream of rate $\lambda = 20$ cars per Hour.
- Time to fill up at the pump is exponential with mean 10 minutes.
- A car that finds the station full leaves without service

Suggest a Birth and Death description of this system, and calculate the following:

(i) Make a diagram of states and of the transition rates.
(ii) Calculate the stationary distribution of the state of the system.
(iii) What fraction of the cars are lost ?
(iv) What is the distribution of the sojourn time of cars that receive service, and what is its average.

Solution:

(i) The states are $k = 0, 1, 2, 3, 4$ the number of cars in the station

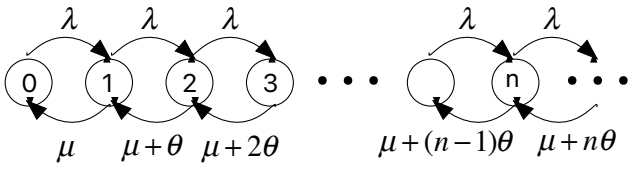


(ii) $p_1 = \frac{27}{272},\ p_2 = \frac{45}{272},\ p_3 = \frac{75}{272},\ p_4 = \frac{125}{272}$.

(iii) Lost are $p_4 = \frac{125}{272} \approx 1/3$.
(iv) Waiting time is 0 with probability $p_1 + p_2 = \frac{72}{272}$, Exp(6) with probability p_3 and Erlang(2,6) with probability p_4. Average wait is: $0 + p_3 * 5 + p_4 * 10 = \frac{1625}{272} = 6$ minutes.

1.8 For the gas station of the previous example, simulate the system (you can use excel or any other code) under the following alternative sets of conditions:
(i) The fill up time is exponential with mean 10.
(ii) The fill up time is distributed uniformly with mean $\sim U(8, 12)$.
In either case simulate the system for 1100 cars, discard the first 100, and present the following analysis of the results:
(i) What is the fraction of cars that abandon without service.
(ii) What is the mean and standard deviation of the sojourn time of cars that get service.
(iii) Plot a histogram of the sojourn times.

1.9 *A queue with abandonments:* In a single server queue customers arrive at rate λ, service is at rate μ, but customers have patience with mean $1/\theta$ and leave the system without service (abandon, or renege) if their waiting time exceeds their patience. Assume Poisson arrivals, and exponential service and patience times. Present this as a Markovian birth and death queue, and calculate its stationary distribution, and the average waiting time of customers that get served.
Solution:



The stationary distribution is:

$$\pi_n = \pi_0 \prod_{j=1}^{n} \frac{\lambda}{\mu + (j-1)\theta}.$$

The average queue length:

$$\mathbb{E}(Q(\infty)) = \pi_0 \sum_{n=0}^{\infty} \prod_{j=1}^{n} \frac{\lambda}{\mu + (j-1)\theta}$$

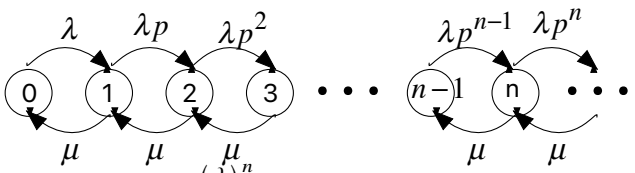
and by Little's law,

$$\mathbb{E}(\text{Sojourn}) = \mathbb{E}(Q(\infty)) \Big/ \lambda$$

1.10 *A queue with balking:* In a single server queue customers arrive at rate λ, service is at rate μ, but a customer that sees a queue of n customers in the

system only joins the queue with probability p^n (otherwise he balks). Assume Poisson arrivals, and exponential service times. Present this as a Markovian birth and death queue, and calculate its stationary distribution, the fraction of customers that balk, and the average waiting time of customers that join the queue.

Solution:



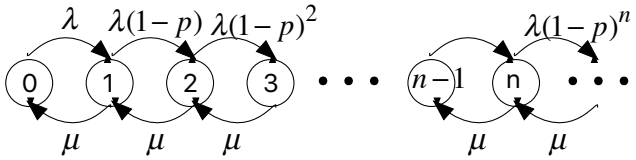
$$\pi_n = \pi_0 \left(\frac{\lambda}{\mu}\right)^n p^{n(n-1)/2},$$

$$\mathbb{P}(\text{balk}) = \pi_0 \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n p^{n(n-1)/2}\left(1 - p^n\right).$$

1.11 *The Israeli queue:* This describes a typical scenario in a queue for cinema tickets in Israel on the weekend in the 1950's. In a small country any two people know each other with probability p. When a person arrives at an M/M/1 queue he scans the customers waiting in queue from first to last and joins with the longest waiting customer, to be served jointly with him (at no extra service time), or if he does not find one he joins the end of the queue. Assume Poisson arrivals at rate λ, exponential service at rate μ. Present this as a Markovian birth and death queue, and calculate the stationary distribution of the queue length, the percentage of customers that find an acquaintance to join, and the distribution of the waiting time.

Solution:

The transition rates for this model are:



From these we obtain the stationary distribution:

$$\pi_n = \pi_0 \left(\frac{\lambda}{\mu}\right)^n (1-p)^{n(n-1)/2}.$$

Note that it decays much faster the geometric.

To calculate the waiting time we recall that by PASTA arrivals see time average, so an arriving customer will see n people in the queue with probability

π_n. If the queue is empty his sojourn is $Exp(\mu)$ If there are n in the queue on his arrival he will join in place k with probability $(1-p)^{k-1}p$ and his sojourn will be $Exp(\mu)^{*k}$ (convolution of k exponentials), i.e. Erlang(μ, k). Conditional on n in the system:

$$f_{W|n}(x) = \sum_{k=1}^{n} p(1-p^{k-1}\frac{\mu^k x^{k-1}}{(k-1)!}e^{-\mu x},$$

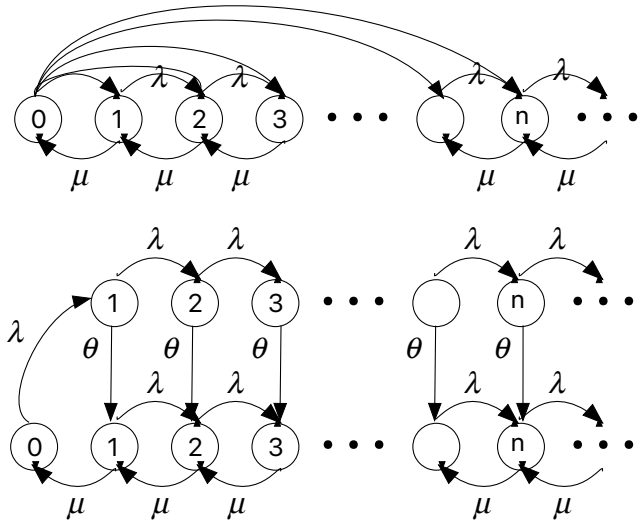
and

$$f_W(x) = \pi_0 \mu e^{-\mu x} + \sum_{n=1}^{\infty} \pi_n f_{W|n}(x).$$

1.12 (∗) *A queue with vacations:* In an M/M/1 queue the server goes on a vacation of duration ∼ Exp(θ) whenever the queue becomes empty. Describe the queue length for this system as a Markov chain, and find its stationary distribution [Servi and Finn (2002)].

Solution:

This is not a birth and death queue. The states and transitions can be represented in two ways:



Transition probabilities from the beginning to the end of a vacation are

$$p_{0,n} = \frac{\theta}{\lambda+\theta}\left(\frac{\lambda}{\lambda+\theta}\right)^n.$$

In the two row transition diagram, the top line is the state of the queue during vacation time.

It can be shown (see Servi and Finn (2002)) that the stationary distribution is:

$$\mathbb{P}(Q(\infty) = n) = \gamma\Big(1 - \frac{\lambda}{\mu}\Big)\Big(\frac{\lambda}{\mu}\Big)^n + (1-\gamma)\Big(1 - \frac{\lambda}{\lambda+\theta}\Big)\Big(\frac{\lambda}{\lambda+\theta}\Big)^n$$

where $\gamma = \dfrac{\theta}{\lambda + \theta - \mu}$.

Note that when $\theta > \mu - \lambda$ then $\gamma > 1$ and $(1-\gamma) < 0$, and when $\theta < \mu - \lambda$ then $\gamma < 0$ and $(1-\gamma) > 1$. When $\theta = \mu - \lambda$, the stationary distribution is just that of M/M/1: $\Big(1 - \frac{\lambda}{\mu}\Big)\Big(\frac{\lambda}{\mu}\Big)^n$.

In the derivation one obtains the generating function of the stationary distribution:

$$P(z) = \frac{1 - \frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}z} \, \frac{1 - \frac{\lambda}{\lambda+\theta}}{1 - \frac{\lambda}{\lambda+\theta}z}$$

which is product of the generating functions of two geometric distributions, so the queue length itself is the sum of two independent geometric random variables. The first is the queue length of the M/M/1 without vacations, and the second is the number of arrivals during the vacation.

This is a special case of the decomposition property for queues with vacations, that occurs for various more general single server models with vacations, see Doshi (1986).

2

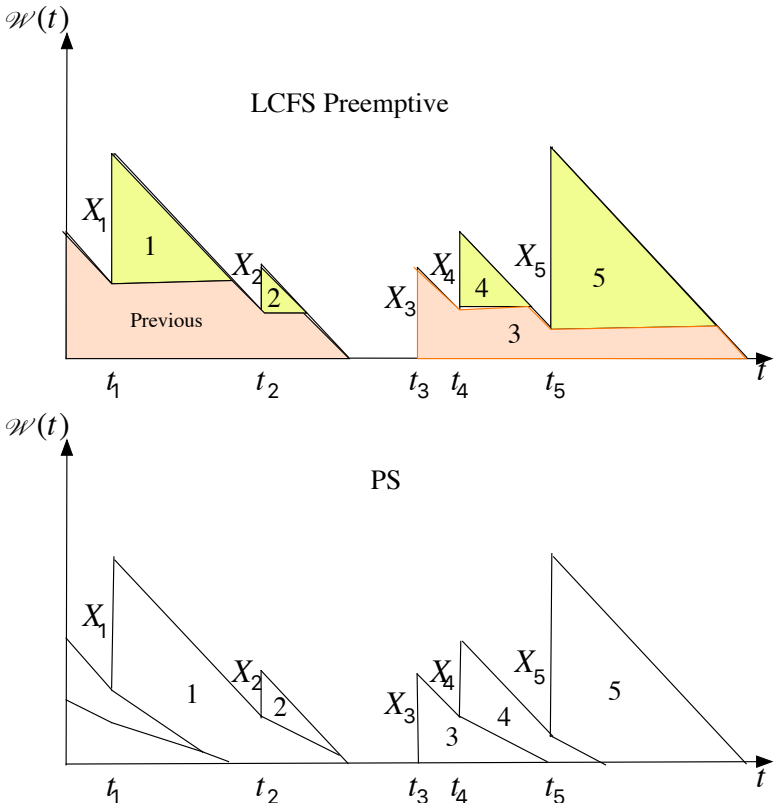
The M/G/1 Queue

Exercises

2.1 Take Figure 2.4 and, similar to Figure 2.5, plot the contributions to workload of each job under preemptive LCFS, and under PS (for the latter assume that $Q(0) = 2$, with equal processing times for both.

Solution

The figure for LCFS-preemptive and for PS are:



2.2 Prove that the renewal function satisfies $m(t) < \infty$ for all $t > 0$, that $\mathcal{A}(t) < \infty$ for all $t > 0$, and that $m(t) \to \infty$, as well as $\mathcal{A}(t) \to \infty$ almost surely, as $t \to \infty$.

Solution

By $\mathbb{P}(T_n = 0) < 1$ there is an x_0 and p_0 such that $\mathbb{P}(T_n \geq x_0) \geq p_0$. Define

$$\tilde{T}_n = \begin{cases} 0 & \text{if } T_n < x_0 \\ x_0 & \text{if } T_n \geq x_0 \end{cases},$$

and let $\tilde{\mathcal{A}}(t)$ be the renewal process formed by $\tilde{T}_n$, with renewal function $\tilde{m}(t)$. Then $\tilde{T}_n < T_n$, $n = 1, 2, \ldots$, and $\tilde{\mathcal{A}}(t) \geq \mathcal{A}(t)$, $t \geq 0$. But $\tilde{m}(t) = \frac{1}{p_0} \lfloor \frac{t}{x_0} + 1 \rfloor < \infty$, hence $m(t) < \infty$ and so also $\mathbb{P}(\mathcal{A}(t) < \infty) = 1$.

If $\lim_{t\to\infty} \mathcal{A}(t, \omega) = K < \infty$ then there exists $n \leq K$ such that $T_n = \infty$, which has probability 0. Hence $\mathbb{P}(\lim_{t\to\infty} \mathcal{A}(t, \omega) = \infty) = 1$, which also implies $\lim_{t\to\infty} m(t) = \infty$.

2.3 Derive the following steps towards the proof of the elementary renewal theorem, by the use of Wald's equation:

(i) Show that $\mathcal{A}(t) + 1$ is a stopping time for the sequence T_n, $n = 1, 2, \ldots$, while $\mathcal{A}(t)$ is not.

(ii) Use Wald's equation to calculate $\mathbb{E}(A_{\mathcal{A}(t)+1})$ and obtain a lower bound on $m(t)/t$.

(iii) Obtain an upper bound for bounded T_n, and prove the theorem for bounded T_n.

(iv) Consider the renewal process with truncated inter-event times $\tilde{T}_n = \min(T_n, a)$, to upper bound $m(t)/t$, and complete the proof.

Solution

(i) $\mathcal{A}(t) + 1$ is a stopping time for $T_1, T_2, \ldots$: once we see that $A_{n-1} \leq t < A_n$ we know that $\mathcal{A}(t) = n - 1$ and so $A(t) + 1 = n$. Hence, $A(t) + 1$ is determined by $T_1, \ldots, T_{\mathcal{A}(t)+1}$ as required.

$\mathcal{A}(t)$ is not a stopping time: if we observe $A_n < t$ all we know is that $\mathcal{A}(t) \geq n$, but we need to observe more than A_n to determine $\mathcal{A}(t) = n$.

(ii) By Wald's identity,

$$\mathbb{E}(A_{\mathcal{A}(t)+1}) = \mathbb{E}\left(\sum_{n=1}^{\mathcal{A}(t)+1} T_n\right) = \mathbb{E}(\mathcal{A}(t) + 1)\mathbb{E}(T_n) = \frac{m(t) + 1}{\mu},$$

but: $A_{\mathcal{A}(t)} \leq t < A_{\mathcal{A}(t)+1}$, and we can write $A_{\mathcal{A}(t)+1} = t + \mathcal{Y}(t)$, where $\mathcal{Y}(t)$ is the remaining time of $T_{\mathcal{A}(t)+1}$ from t to $A_{\mathcal{A}(t)+1}$ (excess time). So

$$(m(t) + 1)/\mu, = t + \mathbb{E}(\mathcal{Y}(t)),$$

or:

$$\frac{m(t)}{t} = \mu + \frac{\mu\mathbb{E}(\mathcal{Y}(t))}{t} - \frac{1}{t}.$$

Because $\mathcal{Y}(t) \geq 0$, we have $\frac{m(t)}{t} \geq \mu - \frac{1}{t}$, and so we have alower bound:

$$\lim_{t\to\infty} \frac{m(t)}{t} \geq \mu.$$

(iii) If T_n are bounded, say $T_n < b$ then we obtain also and upper bound:

$$\frac{m(t)}{t} \leq \mu + \frac{\mu b}{t} - \frac{1}{t},$$

and in the limit:

$$\lim_{t\to\infty} \frac{m(t)}{t} \leq \mu.$$

Which completes the proof for bounded T_n.
(iv) For general T_n, consider the renewal process $\tilde{\mathcal{A}}(t)$ formed by $\tilde{T}_n = \min(b, T_n)$, with renewal function $\tilde{m}(t)$. Then $\tilde{T}_n \leq T_n$, so $\mathcal{A}(t) \leq \tilde{\mathcal{A}}(t)$, and we get:

$$\mu \leq \lim_{t\to\infty} \frac{m(t)}{t} \leq \lim_{t\to\infty} \frac{\tilde{m}(t)}{t} = \frac{1}{\mathbb{E}(\min(b, T_n))},$$

which holds for any b. Letting $b \to \infty$ completes the proof.

2.4 Derive the following steps towards the proof of the elementary renewal theorem, by showing that $\mathcal{A}(t)/t$, $t \geq 0$ are uniformly integrable:

(i) Obtain a random variable $\tilde{T}_n$ with values 0 or a that is less or equal to T_n.
(ii) Calculate the renewal function of the renewal process $\tilde{\mathcal{A}}(t)$ of $\tilde{T}_n$, to show it is bounded above by $c_1 t + c_2 t^2$, for some constant c_1, c_2.
(iii) Use Markov's (Chebyshev's) inequality to show that for all $t > 1$, $\mathbb{P}(\tilde{\mathcal{A}}(t)/t > x) \leq \frac{c_1+c_2}{x^2}$ and therefore $\mathcal{A}(t)/t$, $t \geq 0$ are uniformly integrable to complete the proof.

Solution
(i) Choose a such that $0 < F(a) < 1$, define a sequence of truncated, smaller, random variables $\tilde{T}_n = a\mathbb{I}(T_n > a)$, and let $\tilde{\mathcal{A}}(t)$ be the corresponding renewal process, with renewal function $\tilde{m}(t)$.
(ii) $\tilde{\mathcal{A}}(t)$ has batch arrivals occurring at times which are multiples of a, and the batches are i.i.d. of sizes $K \sim$Geometric, with parameter $p = 1 - F(a)$.

$$\mathcal{A}(t) \leq \tilde{\mathcal{A}}(t) = \sum_{j=1}^{\lfloor t/a \rfloor} K_j.$$

We estimate $\tilde{\mathcal{A}}(t)$, sum of i.i.d. geometric random variables.

$$\mathbb{E}(\tilde{\mathcal{A}}(t)) = \lfloor t/a \rfloor \mathbb{E}(K), \qquad \mathbb{V}\mathrm{ar}(\tilde{\mathcal{A}}(t)) = \lfloor t/a \rfloor \mathbb{V}\mathrm{ar}(K),$$

$$\mathbb{E}(\tilde{\mathcal{A}}(t)^2) = \left(\lfloor t/a \rfloor \mathbb{E}(K)\right)^2 + \lfloor t/a \rfloor \mathbb{V}\mathrm{ar}(K) = c_1 t + c_2 t^2.$$

(iii) We now use Markov's (Chebyshev's) inequality:

$$\mathbb{P}(\mathcal{A}(t)/t > x) \le \mathbb{P}(\tilde{\mathcal{A}}(t)/t > x) = \mathbb{P}(\left(\tilde{\mathcal{A}}(t)/t\right)^2 > x^2)$$
$$\le \frac{\mathbb{E}(\tilde{\mathcal{A}}(t)^2)}{t^2x^2} \le \frac{c}{t^2x^2}, \quad c = c_1 + c_2, \quad t \ge 1,$$

which shows that $\mathcal{A}(t)$, $t \ge 0$ are uniformly integrable. Therefore

$$\lim_{t\to\infty} m(t)/t = \lim_{t\to\infty} \mathbb{E}(\mathcal{A}(t)/t) = \mathbb{E}(\lim_{t\to\infty} \mathcal{A}(t)/t) = \mu.$$

2.5 Prove the renewal reward theorem.

Solution

(i) We write:

$$\frac{C(t)}{t} = \frac{\mathcal{A}(t)}{t} \frac{\sum_{n=1}^{\mathcal{A}(t)} C_n}{\mathcal{A}(t)} \to \frac{\mathbb{E}(C_n)}{\mathbb{E}(T_n)} \text{ a.s.}$$

by the elementary renewal theorem and by the SLLN.

(ii) Consider $C(t)/t \le \mathcal{Y}(t) = \frac{1}{t}\sum_{n=1}^{\mathcal{A}(t)+1} |C_n|$. Recall that $\mathcal{A}(t)+1$ is a stopping time for the sequence $(T_n, C_n)_{n=1,2\ldots}$.

$$\mathbb{E}(\mathcal{Y}(t)) = \frac{1}{t}\mathbb{E}(\mathcal{A}(t)+1)\mathbb{E}(|C_n|) \to \frac{\mathbb{E}(|C_n|)}{\mathbb{E}(T_n)},$$

by Wald's equation and the elementary renewal theorem. In particular, this implies that $\mathcal{Y}(t)$, $t \ge 0$ are uniformly integrable, and this implies also that $C(t)/t$, $t \ge 0$ are uniformly integrable, and so

$$\lim_{t\to\infty} \frac{c(t)}{t} = \lim_{t\to\infty} \frac{\mathbb{E}(C(t))}{t} = \mathbb{E} \lim_{t\to\infty} \frac{C(t)}{t} = \frac{\mathbb{E}(C_n)}{\mathbb{E}(T_n)}.$$

2.6 Obtain the conditional probability $\mathbb{P}(\text{Excess} > y \mid \text{Age} = x)$, and use it to derive the joint distribution of age and excess.

Solution

Theorem: *In the stationary renewal process with interval distribution F, p.d.f. f and expected value $1/\mu$, the joint p.d.f. of age $A(t)$ and excess $Y(t)$ is: $f_{A,Y}(x,y) = \mu f(x+y)$.*

Proof We calculate first $\mathbb{P}(Y(t) > y \mid A(t) = x)$:

$$\mathbb{P}(Y(t) > y \mid A(t) = x) =$$
$$= \mathbb{P}(\text{ no renewals in } (x, x+y) \mid A(t) = x)$$
$$= \mathbb{P}(\text{ no renewals in } (x, x+y) \mid \text{ no renewals in } (0,x))$$
$$= \mathbb{P}(\text{ Interval exceeds } x+y)/\mathbb{P}(\text{ interval exceeds } x)$$
$$= \frac{1-F(x+y)}{1-F(x)}.$$

But for a stationary renewal process, the p.d.f. of $A(t)$ is $f_A(x) = \mu(1-F(x))$.

Hence the joint probability/density that $Y(t) > y$ and $A(t) = x$ is

$$\frac{1 - F(x+y)}{1 - F(x)} \times \mu(1 - F(x)) = \mu(1 - F(x+y)),$$

and taking derivative with respect to y and reversing the sign we have:

$$-\frac{d}{dx}\mu(1 - F(x+y)) = \mu f(x+y).$$

Note – Sanity check: integrating w.r.t. y we get the age p.d.f. and integrating w.r.t. x we get the excess p.d.f. □

2.7 Derive the joint distribution of age and excess directly from the renewal reward theorem, by considering the length of time that excess $> x$ and also age $> y$, within an interval of length T.
Solution

$$\begin{aligned}
&\mathbb{P}(T_{Fwd}(t) > x,\, T_{Bwd}(t) > y) \\
&= \lim_{t\to\infty} \frac{\text{measure of } r \in (0,t) \text{ that have forward recurrence} > x \text{ and age} > y}{t} \\
&= \frac{\mathbb{E}(T - x - y)^+}{\mathbb{E}(T)} = \mu \int_{x+y}^{\infty} (z - x - y) dF(z) \\
&= \mu \int_{x+y}^{\infty} (1 - F(z)) dz.
\end{aligned}$$

Taking derivative w.r.t. x:

$$\mathbb{P}(T_{Bwd}(t) > y,\, T_{Fwd}(t) = x)dx = \mu(1 - F(x+y)),$$

and taking derivative w.r.t. y and reversing the sign:

$$f_{Fwd,Bwd}(x,y) = \mu f(x+y).$$

2.8 Prove that when you observe a stationary renewal process at an arbitrary time t, t is uniformly distributed along the interval that includes t.
Solution
For the stationary distributions of age and excess, we calculate the conditional density:

$$f(Age = x|\text{Age+Excess} = z) = \frac{f(\text{Age} = x \cap \text{Excess} = z - x)}{f(\text{Interval} = z)} = \frac{\mu f(z)}{\mu z f(z)} = \frac{1}{z},$$

so $Age \sim$ Uniform(interval).

2.9 A bus arrives at your station according to a stationary renewal process, with interarrival time distributions $T \sim F$. You arrive at some arbitrary time. Calculate the distributions of: (1) the length of the interval in which you arrived, (2) the time since the last bus arrived, (3) your waiting time for the next bus, in the following cases:

(i) T is uniform, $T \sim U(a,b)$, i.e. $f_T(t) = \frac{1}{b-a}, \quad 0 < t < b.$
(ii) T is exponential, $T \sim \exp(\lambda)$, i.e. $f_T(t) = \lambda e^{-\lambda t}, \quad t > 0.$
(iii) T is Erlang 2, $T \sim \gamma(2,\lambda)$, i.e. $f_T(t) = \lambda^2 t e^{-\lambda t}, \quad t > 0.$

2.10 Use (2.11) to obtain the average number of customers in the queue, the average number of customers waiting for service, and the average number of customers at the server, for an M/G/1 queue.

Solution

Use Little's law: Average number at the server is $\rho = \frac{\lambda}{\mu}$, average number waiting for service: $\lambda \bar{V} = \frac{\rho^2}{1-\rho}\frac{1+c_s^2}{2}$, average number in queue is their sum.

2.11 Calculate the average waiting time for an M/G/1 system with given ρ, when the service time is distributed as (i) Exp(1), (ii) Erlang(k,k), (iii) Deterministic =1 (iv) $f_X(x) = \frac{1}{1+a}\frac{1}{a}e^{-x/a} + \frac{a}{1+a}ae^{-ax}$.
Note: the last distribution is a called a hyper-exponential distribution and has a large c.o.v.

Solution:

Recall the Pollaczek–Khinchine formula: $\overline{V} = m\,\frac{\rho}{1-\rho}\frac{1+c_s^2}{2}$. The only part that depends on the service time distribution is it c.o.v. Hence
(i) Exp(1): $c_s^2 = 1$, (ii) Erlang(k,k): $c_s^2 = 1/k$ (iii) Deterministic $c_s^2 = 0$
(iv) Hyper-exponential: The expected value is 1. The variance is $2a-3+2/a$, hence $c_s^2 = 2a - 3 + 2/a$, which is always ≥ 1 with the minimum for $a = 1$, i.e. exponential.

2.12 Consider a stationary M/M/1 queue at an arbitrary time t, and calculate the remaining length of the busy period. This will be the waiting time of a so called *standby customer*, who only starts service when the queue is empty.

Solution:

A standby arrival at an arbitrary time t will see the stationary workload, $w = \mathcal{W}(t)$ and he will then need to wait for the length of an EFSBP starting with w, so the conditional time for his start will be $\frac{w}{1-\rho}$. Unconditioning by Pollaczek-Khinchine formula will give:

$$\bar{V}_{\text{standby}} = \frac{\bar{\mathcal{W}}}{1-\rho} = m\frac{\rho}{(1-\rho)^2}\frac{1+c_s^2}{2}.$$

In particular for M/M/1 we get: $\bar{V}_{\text{standby}} = m\frac{\rho}{(1-\rho)^2}$.

2.13 Derive the following equation for the Laplace transform of the distribution of the length of a busy period of the M/G/1 queue:

$$BP^*(s) = G^*\big(s + \lambda - \lambda BP^*(s)\big),$$

where $G^*(s)$, $BP^*(s)$ are the Laplace transforms of the service time distribution and the length of busy period distribution [Abate et al. (1995); Kendall (1964)].

Solution:

Let X_1 be the processing time of the first customer, and $N(X_1)$ the number

of arrivals during his service, and let BP_j be the length of busy period of the j'th arrival during that service. Then:

$$\begin{aligned}
BP^*(s) &= \mathbb{E}(e^{-sBP}) = \mathbb{E}(e^{-s(X_1+\sum_{j=1}^{N(X_1)} BP_j)}) \\
&= \mathbb{E}_{X_1}\left(e^{-sX_1}\mathbb{E}_{N(X_1)}\left(e^{-s\sum_{j=1}^{N(X_1)} BP_j}\right)\right) \\
&= \mathbb{E}_{X_1}\left(e^{-sX_1}\mathbb{E}_{N(X_1)}\left(\mathbb{E}\left(e^{-sBP_j}\right)^{N(X_1)}\right)\right) \\
&= \mathbb{E}_{X_1}\left(e^{-sX_1}\mathbb{E}_{N(X_1)}\left(BP^*(s)^{N(X_1)}\right)\right) \\
&= \mathbb{E}_{X_1}\left(e^{-sX_1}e^{-\lambda X_1(1-BP^*(s))}\right) \\
&= G^*(s+\lambda-\lambda BP^*(s)),
\end{aligned}$$

where we use the non-standard notation $\mathbb{E}_X(\cdot) = \mathbb{E}(\mathbb{E}(\cdot|X))$.

2.14 Derive the variance and the second moment of the length of an M/G/1 busy period [Cohen (1982), Section II-2.2].

Solution:

Method 1

Take derivatives of the equation for the Laplace transform and solve.

Method 2

Let m, σ^2 be the mean and variance of the service times, and ρ the traffic intensity. Denote by U^2 the variance of a busy period.

Condition on first customer of length x. The conditional expectation of the busy period is $\frac{x}{1-\rho}$. We calculate the conditional variance, which we denote by U_x^2: $U_x^2 = \mathbb{V}\text{ar}(x + \sum_{j=1}^{N(x)} BP_j)$ where $N(x)$ is the number of arrivals in x which is Poisson(λx), and the BP_j are the busy periods started by the arrivals, which are independent. Hence

$$U_x^2 = \mathbb{V}\text{ar}(x + \sum_{j=1}^{N(x)} BP_j) = \lambda x\Big(U^2 + \frac{m^2}{(1-\rho)^2}\Big).$$

Hence:

$$U^2 = \mathbb{E}(U_X^2) + \mathbb{V}\text{ar}(\frac{X}{1-\rho}) = \rho\Big(U^2 + \frac{m^2}{(1-\rho^2)}\Big) + \frac{\sigma^2}{(1-\rho^2)},$$

from which we get:

$$\mathbb{V}\text{ar}(BP) = U^2 = \frac{\rho m^2 + \sigma^2}{(1-\rho)^3}.$$

$$\mathbb{E}(BP^2) = \frac{m^2+\sigma^2}{(1-\rho)^3} = \frac{\mathbb{E}(X^2)}{(1-\rho)^3}.$$

2.15 Derive the Laplace transform of the length of a busy period for an M/M/1 queue. Use it to obtain the p.d.f. of the length of a busy period of an M/M/1 queue.

Solution:

We need to solve:

$$BP^*(s) = G^*\big(s + \lambda - \lambda BP^*(s)\big),$$

where $G^*(s) = \frac{\mu}{\mu+s}$:

$$BP^*(s) = \frac{\mu}{\mu + s + \lambda - \lambda BP^*(s)},$$

so $PB^*(s)$ is a root of the equation

$$(\lambda + \mu + s)x = \mu + \lambda x^2.$$

We need the root that is < 1, so

$$BP^*(s) = \frac{1}{2\lambda}\left(\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}\right).$$

The probability density function of the busy period is obtained by inverting the Laplace transform:

$$f_{BP}(t) = \frac{1}{t\sqrt{\rho}} e^{-(\lambda+\mu)t} I_1(2t\sqrt{\lambda\mu}),$$

where I_1 is the modified Bessel function of order one:

$$I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!}.$$

The main lesson from this is that although the mean busy period equals in length to the mean sojourn time, the distributions are very different. One can obtain closed form expressions most performance measures of M/M/1, but they are not always simple expressions.

2.16 Verify the derivation for the generating function of the M/G/1 embedded Markov chain, given by equation (2.16).

Solution:

We have:

$$\pi_i = \pi_0 k_i + \sum_{j=1}^{i+1} \pi_j k_{i+j-1}, \quad i = 0, 1, \ldots,$$

so we have:

$$\Pi(z) = \sum_{i=0}^{\infty} \pi_i z^i = \sum_{i=0}^{\infty} \pi_0 k_i z^i + \sum_{i=0}^{\infty} \sum_{j=1}^{i+1} \pi_j k_{i-j+1} z^i$$
$$= \pi_0 K(z) + \frac{1}{z} \sum_{i=0}^{\infty} \sum_{j=1}^{i+1} \pi_j z^j k_{i-j+1} z^{i-j+1}$$

$$= \pi_0 K(z) + \frac{1}{z} \sum_{j=1}^{\infty} \pi_j z^j \sum_{i=j-1}^{\infty} k_{i-j+1} z^{i-j+1}$$
$$= \pi_0 K(z) + \frac{1}{z} \sum_{j=1}^{\infty} \pi_j z^j \sum_{l=0}^{\infty} k_l z^l$$
$$= \pi_0 K(z) + \frac{1}{z} (\Pi(z) - \pi_0) K(z).$$

from which we get:

$$\Pi(z) = \pi_0 \frac{K(z)(1-z)}{K(z) - z} = \frac{(1-\rho)K(z)(1-z)}{K(z) - z}.$$

2.17 Prove that the equation $\alpha = \mathbb{E}(\alpha^{B_n})$ has a unique solution in $(0, 1)$ if and only if $1 > \pi_0 > 0$ and $\mathbb{E}(B_n) > 1$.

Solution:

$f(z) = \mathbb{E}(z^j) = \sum_{j=0}^{\infty} \pi_j z^j$ satisfies $f(0) = \pi_0 > 0$, and $f(1) = 1$. Also, if $\pi_0 < 1$ both $f(z)$ and $f'(z)$ are strictly increasing, with $f'(0) = \pi_1 < 1$, and finally, we have $f'(1) = \mathbb{E}(B_n)$.

If $\mathbb{E}(B_n) > 1$ then $f(x)$, $0 < x < 1$ starts above 0, with slope < 1 and ends at 1, with slope > 1, so it must cross the line $g(x) = x$ at least once, and because $f'(x)$ is increasing it will cross no more than once, so there is a unique solution. On the other hand, if $\mathbb{E}(B_n) \leq 1$, then it starts at $f(0) = 0$, ends at $f(1) = 1$ and has derivative ≤ 1 throughout, so it cannot cross the line $g(x) = x$, and there is no solution.

2.18 Use equations (2.17), (2.19), to show that M/G/1 queue has the resource pooling property: when service is speeded up s fold, and arrivals are speeded up s fold, the queue length distribution remains exactly the same, while waiting and sojourn times are improved by a factor of s.

Solution:

Consider (2.17), let $G^{(n)}(x)$ be the distribution function of service time under n speedup. Then $G^{(n)}(x) = G(nx)$, and we have for the Laplace transform:

$$G^{(n)*}(n\lambda(1-z)) = \int_0^{\infty} e^{n\lambda(1-z)x} dG^{(n)}(x) = \int_0^{\infty} e^{\lambda(1-z)nx} dG(nx)$$
$$= G^*(\lambda(1-z)).$$

Hence $\Pi(z)$ is unchanged, i.e. the queue length distribution of the speeded up system equals that of the original system. This immediately implies that the waiting times are shorter by a factor of n.

2.19 Explain why the sojourn time of a job under preemptive LCFS equals the length of a busy period. In particular, use this to explain why in the M/M/1 queue, the expected busy period equals the expected sojourn time of a customer under any non-predictive work conserving policy.

Solution:

Clearly, the sojourn time of a customer under LCFS preemptive equals the length of a busy period: he starts service immediately on arrival, but will not depart until all customers arriving after him are all gone. On the other hand, for any non-predictive policy, in the M/M/1 system, the departure process while the server is busy is always Poisson with rate μ, so looking at departures minus arrivals in a busy period, the distribution of the total is independent of the policy. So the expected sojourn times for M/M/1 are indeed equal to the expected length of the busy period. Note: this is only true for M/M/1. Note also: the sum of the sojourns is independent of the policy, but not the individual processing times, so while expected sojourn time equals expected busy period, the distribution of sojourn time depends on the policy.

2.20 Derive the expected length of a busy period for an M/G/1 queue with vacations, where the vacation time has distribution H.

Solution:

At the end of a vacation, there will be 0, 1, or more customers waiting, and (assume FCFS) their processing will constitute an exceptional first service. We include the case of 0, with a busy period of length 0 in the calculation, i.e. we define BP as time batween two vacations. So the exceptional first service will consist of $Y = \sum_{j=1}^{N} X_j$ where N is the number of arrivals during the vacation, and X_j their processing times. Conditional on Y the BP will be expected to last $\frac{Y}{1-\rho}$, so $\bar{BP} = \frac{\mathbb{E}(Y)}{1-\rho}$ (the remainder of BP is independent of the exceptional first service). But, by Wald's equation, $\mathbb{E}(Y) = \mathbb{E}(N)m$. Furthermore, conditional on the vacation length being Z, $\mathbb{E}(N \mid Z) = \lambda Z$ (Poisson arrivals), and (again by Wald), $\mathbb{E}(N) = \lambda h$ where h is the expected length of the vacation. So: $\bar{BP} = \frac{\lambda h m}{1-\rho} = \frac{\rho h}{1-\rho}$.

3

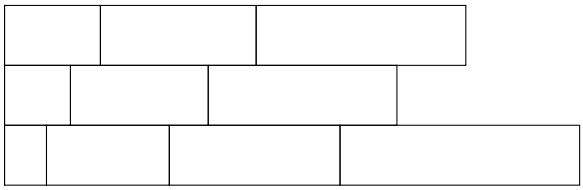
Scheduling

Exercises

3.1 Prove that SPT minimizes the flowtime of a batch of N jobs scheduled on M parallel identical machines [Lawler et al. (1993))].

Solution:

On the M parallel machines we can put up to M jobs as last on their machines, so that delay only themselves, up to M jobs as before last, that delay 2 jobs each and so on. By the same argument as for a single machine, we want the longest jobs to go last on their machines, the next longest to go second to last and so on. But this is exactly using FCFS, starting with the shortest job, and putting the k shortest job on the first machine that is free after scheduling the first $k-1$ jobs. See the following figure to convince yourself that jobs are ordered from shortest to longest and are then assigned to the machines in order $1, 2, \ldots, M$ repeatedly, until all are assigned.



3.2 What is the non-preemptive schedule that will minimize the flowtime for a batch of N jobs, on parallel machines that are working at different speeds, $s_1, \ldots, s_M$ so that job j with processing requirement X_j, performed on machine i will need processing for a duration X_j/s_i [Lawler et al. (1993))].

Solution:

Construct the following schedule: Take the values of k/s_i for $k = 1, \ldots, N$, $i = 1, \ldots, M$ as weights, and sort them from smallest to largest. Match the smallest weights to the longest jobs. Then put each job on the machine it was matched to in the kth to last position.

A more general problem is when job j on machine i requires processing time $X_{j,i}$. Then one can formulate the problem of scheduling the N jobs on M machines as the following integer programming problem (it is actually an

assignment problem): Find values for $u_{j,i,k}$ such that $u_{j,i,k} = 1$ indicates that job j is scheduled on Machine i in the position of kth to last, such that

$$\begin{aligned} &\min \textstyle\sum_{j,i,k} k X_{j,i} u_{j,i,k}, \\ &\text{s.t.} \quad \textstyle\sum_{i,k} u_{j,i,k} = 1 \quad j = 1, \ldots, N, \\ &\qquad \textstyle\sum_{j} u_{j,i,k} \le 1 \quad i = 1, \ldots, M, \quad k = 1, \ldots, N, \\ &\qquad u_{j,i,k} \in \{0, 1\} \quad i = 1, \ldots, M, \quad j, k = 1, \ldots, N. \end{aligned}$$

This can be solved in polynomial time, by bipartite weighted matching.

3.3 The SPT-savings parameter of a distribution measures the amount by which knowing the values of the job processing times and using SPT is better than scheduling them in random order. For jobs with processing time distribution G, it is defined as $d^G = \frac{m - m_{1:2}}{\sigma}$, where m is the mean, σ the standard deviation, and $m_{1:2}$ the mean of the smaller in a sample of 2, drawn from G. Calculate d^G for the following distributions:

(i) Exponential,
(ii) Normal,
(iii) Uniform.

Solution:

In general one needs to calculate:

$$m_{1:2} = 2 \int_{-\infty}^{\infty} \int_{x}^{\infty} x f(x) f(y) dy dx = 2 \int_{-\infty}^{\infty} x f(x)(1 - F(x)) dx.$$

The value of d for X is the same as for $(X - a)/b$.

(a) For $X \sim \text{Exp}(\lambda)$, $m = 1/\lambda$, $\sigma = 1/\lambda$ and $m_{1:2} = 1/2\lambda$, hence:

$$d_{\text{Exp}} = \frac{1/\lambda - 1/2\lambda}{1/\lambda} = 1/2.$$

(b) For $X \sim N(\mu, \sigma)$ it is the same as for $Z \sim N(0, 1)$. The value then is:

$$\begin{aligned} d_{\text{Normal}} &= -m_{1:2} = -2 \int_{-\infty}^{\infty} x\phi(x)(1 - \Phi(x)) dx = 2 \int_{-\infty}^{\infty} x\phi(x)\Phi(x)) dx \\ &= \sqrt{\frac{1}{\pi}} = 0.564. \end{aligned}$$

(c) For $X \sim U(a, b)$ we can look at $Y \sim U(0, 1)$, for which: $m = 1/2$, $\sigma^2 = 1/12$, and because $P(Y_{1:2} > y) = (1 - y)^2$, we have: $m_{1:2} = \int_0^1 (1 - y)^2 dy = 1/3$, so:

$$d_{\text{Uniform}} = \frac{1/2 - 1/3}{\sqrt{1/12}} = \sqrt{\frac{1}{3}} = 0.577.$$

3.4 Consider the policy of preemptive LRPT (longest remaining processing time). Describe how this policy works, and show that it maximizes the number of customers in the system at all times.

Solution:
We start on the longest, then when its remainder reaches the size of the second longest, we work on both, splitting the server, till the remainder of both reaches the size of the third longest and we now work on all three splitting the server equally, and so on. By the end we will work on all N jobs splitting the server equally between all N, and all jobs will depart simultaneously at time $X_1 + \cdot + X_N$. Since there are N jobs in the system at all times, it maximizes the number of customers in the system at all times.

3.5 Show that in a $G/G/1$ queue LCFS non-preemptive maximizes the variance of the sojourn and of the waiting time among all non-predictive, non-preemptive, work conserving policies.

Solution:
The proof follows the same ideas as in the proof of Theorem 3.7 that FCFS maximizes the variance.
We know that in each busy period, arrivals will be $A_1 \leq A_2 \leq \cdots \leq A_N$ (not under our control) and departures will be $D_1 \leq D_2 \leq \cdots \leq D_N$, determined by i.i.d. service times also not under our control, and so mean service time will be independent of the policy, and all we control is the permutation in which order we assign arrivals to the server, so as to control $\sum_{j=1}^{N} A_{s_j} D_j$ where $s_1, \ldots, s_N$ is a permutation of $1, \ldots, N$, and we start the arriving jobs in the order $s_1, \ldots, s_N$. By Hardy-Littlewood-Polya inequality this will be maximized, and the variance of sojourn times minimized, if we use $s_j = j$, $j = 1, \ldots, N$. To maximize variance we should minimize $\sum_{j=1}^{N} A_{s_j} D_j$, which would be achieved by $s_j = N - j + 1$, $j = 1, \ldots, N$, but this may be infeasible, since we may then have $D_{j-1} < A_{s_j}$. So we need to minimize $\sum_{j=1}^{N} A_{s_j} D_j$ for given $A_1 \leq A_2 \leq \cdots \leq A_N$, $D_1 \leq D_2 \leq \cdots \leq D_N$, under the added constraint that $A_{s_j} \leq D_{j-1}$ (where we let $D_0 = A_1$). We now show that LCFS assignment achieves that.
Assume that we have chosen an assignment that is not LCFS. Than for some $k < l$, we have A_{s_k}, A_{s_l} are both $\leq D_{k-1}$, but $A_{s_k} < A_{s_l}$, i.e. both were available at time D_{k-1} and we assigned the earlier arrival as the kth job, so that it will depart at D_k. Then we can switch the job assignments, and because $A_{s_k} < A_{s_l}$ and $D_k < D_l$ we would have $A_{s_l} D_k + A_{s_k} D_l < A_{s_k} D_k + A_{s_l} D_l$. We can in this way improve every feasible permutation that is not LCFS and because there are only $N!$ permutations, LCFS is minimizing.

3.6 In a multi-type M/G/1 queue under priority scheduling policy, show that the longterm average amount of relevant work that a customer of type k finds on arrival, is given by (3.4).

Solution:
By work conservation, for non-preemptive policy, the total workload for a

busy period of N_k jobs of type $k = 1, \ldots, K$ consists of

$$\int_0^T \mathcal{W}(t)dt = \sum_{k=1}^{K} \sum_{j=1}^{N_k} \left(V_{j,k} X_{j,k} + X_{j,k}^2/2 \right).$$

Taking long term average, dividing by T, and multiplying and dividing by $\mathcal{A}(T)$, and multiplying and dividing by $\mathcal{A}_k(T)$:

$$\frac{1}{T}\int_0^T \mathcal{W}(t)dt = \frac{\mathcal{A}(T)}{T}\frac{1}{\mathcal{A}(T)}\sum_{k=1}^{K}\sum_{j=1}^{N_k}\frac{X_{j,k}^2}{2} + \sum_{k=1}^{K}\frac{\mathcal{A}_k(T)}{T}\frac{1}{\mathcal{A}_k(T)}\sum_{j=1}^{N_k} V_{j,k}X_{j,k},$$

so for the long time average, recalling that $V_{j,k}$ and $X_{j,k}$ are independent:

$$\bar{\mathcal{W}} = \lambda\frac{\mathbb{E}(X^2)}{2} + \sum_{k=1}^{K}\rho_k \bar{V}_k,$$

which is the total amount of work that an arriving customer of type k sees, by PASTA. Furthermore, the amount of work consisting of waiting customers of type j is $\rho_j \bar{V}_j$. Hence, the amount of relevant work that a customer of priority k will see upon arrival is

$$\sum_{j=1}^{k}\rho_j\bar{V}_j + \lambda\frac{\mathbb{E}(X^2)}{2}.$$

3.7 Prove by induction the formula (3.5) for waiting time of customers of type k under priority scheduling policy.

Solution:

Clearly by Exercise 3.6, for $k = 1$

$$\bar{V}_1 = \rho_1\bar{V}_1 + \lambda\mathbb{E}(X^2)/2,$$

so, as required,

$$\bar{V}_1 = \frac{\lambda\mathbb{E}(X^2)/2}{1-\rho_1}.$$

Consider now $\bar{V}_2$. Assume that the amount of work that he sees on arrival is x. The he will need to wait for a time x, and in addition for all the type 1 customers that arrive in time x, and their entire busy periods of customers of type 1. his expected wait then will be, by (2.13), $x/(1-\rho_1)$. Taking the expected value of x by Exercise 3.6:

$$\bar{V}_2 = \frac{\rho_1\bar{V}_1 + \rho_2\bar{V}_2 + \lambda\mathbb{E}(X^2)/2}{1-\rho_1},$$

or

$$(1-\rho_1-\rho_2)\bar{V}_2 = \rho_1\bar{V}_1 + \lambda\mathbb{E}(X^2)/2 = \frac{\lambda\mathbb{E}(X^2)/2}{1-\rho_1}.$$

and we showed:

$$\bar{V}_2 = \frac{\lambda \mathbb{E}(X^2)/2}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Consider now customer of type k. His wait under the priority policy will be the same as his wait under a policy that serves all customers of types $j < k$ using FCFS. So similar to the result for $\bar{V}_2$,

$$\bar{V}_k = \frac{\lambda \mathbb{E}(X^2)/2}{(1-\sum_{j<k}\rho_j)(1-\sum_{j\le k}\rho_j)}.$$

3.8 Prove that among all priority scheduling policies for M/G/1, SEPT minimizes the average flowtime.

Solution:

Assume priorities $1, \ldots, K$, with average waiting times for customers of all types given by:

$$\bar{V}_1 < \cdots < \bar{V}_K.$$

Recall that by work conservation, $\sum_{j=1}^{K} \rho_j \bar{V}_j$ is constant.

The average waiting time $\mathbf{V}$ satisfies:

$$\lambda \mathbf{V} = \sum_{j=1}^{K} \lambda_j \bar{V}_j = \sum_{j=1}^{K} \mu_j \rho_j \bar{V}_j,$$

where $\lambda = \sum_j \lambda_j$.

Choose types $k, k+1$ and switch their priorities to get new waiting times $\bar{V}'_j$ and objective value $\mathbf{V}'$. The switch will not affect any types except $k, k+1$, and so the difference between the objectives is

$$\lambda(\mathbf{V}' - \mathbf{V}) = \mu_k\left(\rho_k \bar{V}'_k - \rho_k \bar{V}_k\right) + \mu_{k+1}\left(\rho_{k+1}\bar{V}'_{k+1} - \rho_{k+1}\bar{V}_{k+1}\right).$$

But by work conservation,

$$\begin{aligned} &\rho_k \bar{V}'_k + \rho_{k+1}\bar{V}'_{k+1} = \rho_k \bar{V}_k + \rho_{k+1}\bar{V}_{k+1} \\ &\quad \Longrightarrow \left(\rho_k \bar{V}'_k - \rho_k \bar{V}_k\right) = -\left(\rho_{k+1}\bar{V}'_{k+1} - \rho_{k+1}\bar{V}_{k+1}\right) > 0, \end{aligned}$$

so:

$$\lambda(\mathbf{V}' - \mathbf{V}) = (\mu_k - \mu_{k+1})\left(\rho_k \bar{V}'_k - \rho_k \bar{V}_k\right).$$

This is < 0 if $\mu_k < \mu_{k+1}$, and for any static priority policy that is not SEPT priority, switching priorities improves the objective.

3.9 Prove the formulas (3.6) (3.7) for expected waiting times in M/G/1 queue under non-preemptive SPT.

Solution:

Let $\rho(x) = \lambda \int_0^x y dG(y)$. Then this is the offered load of the customers with processing times $\le x$. Let $\bar{V}(x)$ be the average waiting time of a customer that requires x processing time. We refer to such a customer as an x-customer.

We first calculate the expected amount of relevant work in the system that an x-customer finds on arrival, under SPT. We note that it is the same amount that he would find if we had a two type priority policy that gives priority to y-customers with $y \leq x$ over y-customers with $y > x$. This would then be (see Exercise 3.7):

$$\frac{\lambda \mathbb{E}(X^2)/2}{1-\rho(x)}.$$

However, if the relevant work that he finds is y, he would then need to wait for y plus all the work of serving to exhaustion all u-customers with $u < x$, i.e. an EFSBP starting with y, which has expected value $y/(1-\rho(x-))$. Taking expectation over y:

$$\bar{V}(x) = \frac{\lambda \mathbb{E}(X^2)/2}{(1-\rho(x))(1-\rho(x-))}.$$

Finally, averaging over all customer service times we get (3.7).

3.10 Jobs arrive at a single server station in a Poisson stream of rate $\lambda = 0.1$ jobs per Unit time. The processing times of the jobs are i.i.d. uniformly distributed $\sim U(6,12)$.

(i) If no information about the processing time of the jobs is available to the server, calculate the average waiting time per job (in steady state).

(ii) If each arriving job can be classified into one of three types, $\sim U(6,8)$, $\sim U(8,10)$, or $\sim U(10,12)$, and jobs are served according to non-preemptive SEPT, calculate the average waiting time of each type, and the overall average waiting time.

(iii) If the processing time of each job becomes known upon arrival, calculate the average waiting time of each job, and the overall average waiting time for non-preemptive SPT.

Solution

(i)

$$\begin{gathered}\lambda = 0.1,\\ \mathbb{E}(X) = \tfrac{6+12}{2} = 9,\\ \mathbb{E}(X^2)/2 = (\mathbb{V}\mathrm{ar}(X) + (\mathbb{E}(X))^2)/2 = (\tfrac{36}{12} + 81)/2 = 42,\\ \rho = \lambda \mathbb{E}(X) = 0.1 * 9 = 0.9\,.\end{gathered}$$

By Khinchine-Pollaczek formulation, average waiting time is : $\bar{V} = \frac{\lambda E X^2}{2*(1-\rho)} = 0.1\frac{84/2}{1-0.9} = 42$

average sojourn time is 42+9 = 51.

(ii)

$$\begin{aligned}\rho_1 &= \lambda \textstyle\int_6^8 y\, dG(y) = 0.1 * \frac{7}{3} = 0.233\,,\\ \rho_2 &= \lambda \textstyle\int_8^{10} y\, dG(y) = 0.1 * 3 = 0.3\,,\\ \rho_3 &= \lambda \textstyle\int_{10}^{12} y\, dG(y) = 0.1 * \frac{11}{3} = 0.367\,,\end{aligned}$$

$$\bar{V}_3 = \frac{\lambda EX^2/2}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)} = \frac{0.1*42}{0.467*0.1} = 90\,,$$
$$\text{sojourn time}_1 = EX_1 + \bar{V}_1 = 9 + 90 = 99\,,$$
$$\bar{V}_2 = \frac{\lambda EX^2/2}{(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{0.1*42}{0.767*0.467} = 11.74\,,$$
$$\text{sojourn time}_2 = EX_2 + \bar{V}_2 = 9 + 11.74 = 20.74\,,$$
$$\bar{V}_1 = \frac{\lambda EX^2/2}{(1-\rho_1)} = \frac{0.1*42}{0.767} = 5.48\,,$$
$$\text{sojourn time}_3 = EX_1 + \bar{V}_1 = 9 + 5.48 = 14.48\,.$$

$$\bar{V} = \tfrac{1}{3}\bar{V}_1 + \tfrac{1}{3}\bar{V}_2 + \tfrac{1}{3}\bar{V}_3 = 35.74\,.$$

$$\text{mean sojourn time } = 9 + 35.74 = 44.74\,.$$

(iii) $\rho(x) = (x^2 - 36)/120$, so:

$$\bar{V}(x) = \frac{\lambda\mathbb{E}(X^2)/2}{(1-\rho(x))^2} = \frac{0.1 * 42}{\big((156 - x^2)/120\big)^2},$$

$$4.2 = \bar{V}(6) \le \bar{V}(x) \le \bar{V}(12) = 420\,,$$

$$\bar{V} = \lambda\mathbb{E}(X^2)/2 \int_6^{12} \frac{1}{6} \Big/ \big((156 - x^2)/120\big)^2 = 4.2 * 8.19 = 34.4\,.$$

$$\text{mean sojourn time } = 9 + 34.4 = 43.4\,.$$

3.11 (∗) Derive the expected sojourn time of a customer with processing time x in a stationary M/G/1 queue under SRPT (preemptive) policy, and the expected sojourn time over all customers [Schrage and Miller (1966), Wolff (1989) Example 10-5].

Solution:

Assume service time X has distribution G that is continuous with density $g(y)$ and mean m. Under SRPT and this assumption, there is never more than 1 customer with remaining processing time y in the system.

Consider a customer with processing time r, we refer to him as type r. The sojourn time of a customer of type r consists of two parts: V_r is his waiting time before he starts being served. R_r is his residence time from start of service until he leaves, which includes preemptions.

We calculate $\mathbb{E}(R_r)$ first. Once processing starts, our type r will have remaining processing time going down from r to 0, and when his remaining processing time reaches y he can only be preempted by jobs shorter than y, which provide offered load $\rho(y) = \lambda \int_0^y t g(t)dt$. Then:

$$\mathbb{E}(R_r) = \int_0^r \frac{dy}{1-\rho(y)}$$

The waiting time until start of service consists of service of all customers with remaining service time $\leq r$ that are in the system on the arrival of customer type r, which we denote by U_r, and the busy period of all arrivals of types $y < r$ which arrive during U_r (note that this is independent of the service policy, as long as we do not serve any customer with remaining service time $> r$). Relevant to customer r is all the work in the system of customers of processing of length $< r$, as well as processing amount r from all customers of length $\geq r$. Denote $X_r = \min(X, r)$ the processing time of jobs relevant to customer r. Then:

$$\mathbb{E}(X_r) = \int_0^r (1 - G(y))dy, \qquad \mathbb{E}(X_r^2)/2 = \int_0^r y(1 - G(y))dy.$$

Note that because we would preempt any jobs longer than r, we need to consider work in process at the server of only $\mathbb{E}(X_r^2)$ rather than $\mathbb{E}(X^2)$. By PASTA, using the same derivation as of Khinchine-Pollaczek formula for M/G/1:

$$\mathbb{E}(U_r) = \lambda \left(\mathbb{E}(U_r)\mathbb{E}(X_r) + \mathbb{E}(X_r^2)/2 \right).$$

The offered load of jobs with remaining processing times $< r$ is:

$$b_r = \lambda \mathbb{E}(X_r) = \lambda \Big(\int_0^r tg(y)dy + x(1 - G(y)) \Big) = \lambda \int_0^r (1 - G(y))dy,$$

we obtain:

$$\mathbb{E}(U_r) = \frac{\lambda \mathbb{E}(X_r^2)}{2(1 - b_r)} = \frac{\lambda \int_0^r y(1 - G(y))dy}{1 - \lambda \int_0^r (1 - G(y))dy}.$$

to this we need to add processing of arrivals of types $< r$ during U_r, so the total waiting time of customer of type r before he starts processing is:

$$\mathbb{E}(V_r) = \frac{\lambda \mathbb{E}(X_r^2)}{2(1 - b_r)(1 - \rho_r)} = \frac{\lambda \int_0^r y(1 - G(y))dy}{\left(1 - \lambda \int_0^r (1 - G(y))dy\right)\left(1 - \lambda \int_0^r yg(y)dy\right)}.$$

The expected sojourn time is then:

$$\mathbb{E}(W_r) = \mathbb{E}(V_r) + \mathbb{E}(R_r),$$

and the average sojourn time over all the customers is:

$$\mathbb{E}(W) = \int_0^\infty \mathbb{E}(W_r)g(r)dr.$$

3.12 In an M/G/1 queue, under what conditions should one use preemptions or processor splitting to shorten the expected sojourn time.

Solution:

When G has decreasing hazard rate (DHR or DFR decreasing failure rate).

The proof is by using the Gittins index theorem, when we allow any preemptive policy: For IFR we should not use any splitting, since any job we start will immediately be better than all others, while for DFR any job we start should be preempted as soon as possible which proves that PS is optimal for DFR.

3.13 Prove the optimality of Smith's rule for minimizing flow time of a batch of N deterministic jobs, using the analogs of the three proofs for optimality of SPT for flowtime.

Solution:

(a) Pairwise interchange: If we exchange jobs k and $k+1$ in the sequence, the new order minus old will be $\Delta = C_k X_{k+1} - C_{k+1} X_k$ which is < 0 if $\frac{X_{k+1}}{C_{k+1}} < \frac{X_k}{C_k}$, so any non-Smith-rule pair should be interchanged.
(b) The total cost can be written in two ways:

$$V = \sum_{k=1}^{N} C_k \sum_{j=1}^{k} X_j = \sum_{j=1}^{N} X_j \sum_{k=j}^{N} C_k,$$

however, this does not indicate immediately that Smith rule is optimal.
(c) We write total cost as sum of delays, with $\delta_{k,l}$ the indicator that jobe k is scheduled before job l:

$$V = \sum_{k=1}^{N} C_k X_k + \sum_{k \neq l} \delta_{k,l} C_k X_l,$$

from which we see that for every pair it is best to have k before l if $C_k X_l < C_l X_k$ which is Smith rule.

3.14 Show that Smith rule minimizes expected weighted flowtime for stochastic jobs, where priority is given to jobs with the smallest value of $\mathbb{E}(X_j)/\mathbb{E}(C_j)$ where (X_j, C_j), $j = 1, \ldots, N$ are N independent two dimensional random variables.

Solution:

With the same comparison as in Exercise 3.13, we get for the expectation that k should be before l if

$$\mathbb{E}(C_k X_l) = \mathbb{E}(C_k)\mathbb{E}(X_l) < \mathbb{E}(C_l X_k) = \mathbb{E}(C_l)\mathbb{E}(X_k),$$

where we used the fact that the duration and cost of job k are independent of those of job l. So priority should be given to small $\mathbb{E}(C_k)/\mathbb{E}(X_l)$ which is Smith's rule applied to the expected values. Note, X_k, C_k need not be independent.

3.15 Show that the "cμ" rule minimizes expected weighted flowtime among all static priority policies for M/G/1 with K customer types.

Solution:

We proceed as in Exercise 3.8. Consider static priority order k before l if $k < l$, for types $1, \ldots, K$. Denote by $\mathbf{V}$ the overall average cost per customer,

and by $\bar{V}_j$ the average waiting time for type j. Use primes to denote quantities for the policy where we switch the priorities of type k and $k+1$. We have:

$$\mathbf{V} = \sum_{k=1}^{K} \frac{\lambda_j}{\lambda} c_j \bar{V}_j,$$

then the difference in costs will satiisfy:

$$\lambda(\mathbf{V}' - \mathbf{V}) = c_k \mu_k (\rho_k \bar{V}'_k - \rho_k \bar{V}_k) + c_{k+1} \mu_{k+1} (\rho_{k+1} \bar{V}'_{k+1} - \rho_{k+1} \bar{V}_{k+1}).$$

By work conservation, as in Exercise 3.8,

$$\rho_k \bar{V}'_k - \rho_k \bar{V}_k = -(\rho_{k+1} \bar{V}'_{k+1} - \rho_{k+1} \bar{V}_{k+1}) > 0.$$

so:

$$\lambda(\mathbf{V}' - \mathbf{V}) = (c_k \mu_k - c_{k+1} \mu_{k+1})(\rho_k \bar{V}'_k - \rho_k \bar{V}_k)$$

which is negative if $c_k \mu_k < c_{k+1} \mu_{k+1}$, so we should give priority to large $c\mu$ (costly and short over cheap and long).

Part II

Approximations of the Single Queue

4

The G/G/1 Queue

Exercises

4.1 Explain the difference between the sequence $v_n = \max_{0\le j\le n} \sum_{i=j}^{n-1}(X_i - T_i)$ and the sequence: $v_n \circ \theta^{-n} = \max_{-n\le j\le 0} \sum_{i=j}^{-1}(X_i - T_i)$, $n = 0, 1, 2, \ldots$, which is used in the Loynes construction.

Solution: While v_n give the value of the waiting time of customer n when starting from $v_0 = 0$, so for each n it relates to a different customer. In contrast, the sequence $v_n \circ \theta^{-n}$ give an approximation to the waiting time of customer 0, for every n.

Furthermore, the sequence $v_n \circ \theta^{-n}$ is non-decreasing in n, because from n to $n + 1$ all that changes is adding another term to the maximization, but retaining all the previous terms, whereas in the sequence v_n, going from n to $n + 1$ all the terms in the maximization change, with the addition of $X_n - T_n$ which may be positive or negative, and so the sequence is not monotone in n.

4.2 Explain why the sequence V^n is stationary, and verify that it satisfies Lindley's equation.

Solution: $V^n = \sup_{j\le n} \sum_{i=j}^{n-1}(X_i - T_i)$ expresses V^n as a function of the stationary sequence (X_j, T_j), $-\infty < j < \infty$. It is then seen immediately that

$$V^{n+1} = \left(\sup_{j\le n} \sum_{i=j}^{n-1}(X_i - T_i)\right) \circ \theta = V^n \circ \theta.$$

Because the distribution of the sequence (X_j, T_j), $-\infty < j < \infty$ is invariant under θ, the distribution of V^n is also invariant under θ.

$$\begin{aligned} V^{n+1} &= \sup_{j\le n+1} \sum_{i=j}^{n}(X_i - T_i) \\ &= \max\left[\sup_{j\le n}\left(\sum_{i=j}^{n-1}(X_i - T_i) + (X_n - T_n)\right), 0\right] \\ &= \max\left(V^n + (X_n - T_n), 0\right) = \left(V^n + (X_n - T_n)\right)^+, \end{aligned}$$

where we use the convention that summation over empty range is 0.

4.3 Use a Loynes type construction to show that the waiting time V_n process for G/G/s is stable if $\rho = \lambda/s\mu < 1$ and unstable if $\rho > 1$. As before, assume that (T_n, X_n) are a stationary sequence satisfying SLLN. Use the recursion relation analogous to the Lindley's equation for V_n (see Exercise 1.2) [Loynes (1962); Kiefer and Wolfowitz (1955); Brandt et al. (1990) page 165].

Solution:

Assume to start with that the system is empty at time $-r$. We define as in Exercise 1.2, the ordered vector of workloads found by customer n on arrival, as $V_n^r = (V_{n,1}^r, \ldots, V_{n,s}^r)$. Then we have a recursion for V_n^r:

$$V_{n+1}^r = f(V_n^r, X_n, T_n) = \left[R(V_n^r + e_1 X_n) - e_{all} T_n)\right]^+,$$

where $e_1 = (1, 0, \ldots, 0)^\intercal$, $e_{all} = (1, \ldots, 1)^\intercal$, and R is the operator that orders the vector from small to large.

We note first that $f(V, x, t)$ is non-negative and it is monotone non-decreasing in V: increasing any coordinate of V will cause all components of $f(V, x, t)$ to increase (or stay the same). Therefore, $V_{0,j}^r$ is non-decreasing in r for $j = 1, \ldots, s$ for every sample path. It follows that as we let $r \to \infty$, $V_{0,j}^r$ converges to a limit $V_{0,j}^\infty \leq \infty$. So far this is very similar to the single server case.

However, we have not yet excluded the possibility that some components of $V_{0,j}^r$ converge to a finite limit while other components diverge. We show now that this is cannot be. Consider G/G/s starting at time 0 with some initial values of V_0. Let $U_n = X_n - sT_n$, it is a scalar stationary ergodic sequence. Denote $\Sigma x = \sum_{j=1}^s x_j$. We now look at the scalar sequence $b_n = V_{n,s} - \Sigma V_n$, it is the sum of the differences between the largest $V_{n,s}$ and the other components. Then:

$$\begin{aligned}
&\text{if } X_n < V_{n,s} - V_{n,1} \text{ then } b_{n+1} \leq b_n - X_n,\\
&\text{if } X_n \geq V_{n,s} - V_{n,1} \text{ then } b_{n+1} \leq (s-1)X_n,\\
&\text{hence } b_{n+1} \leq \max(b_n - X_n, (s-1)X_n).
\end{aligned}$$

Let now $c_n = b_n - (s-1)X_n$, and let $D_n = (s-1)X_{n-1} - X_n$, then:

$$c_{n+1} \leq (c_n + D_n)^+$$

The recursion $W_{n+1} = (W_n + D_n)^+$ is exactly the recursion for G/G/1 with $\mathbb{E}(D_n) < 0$, so it has a unique stationary solution W_n, $n \in \mathbb{Z}$, and clearly for any initial value $c_0 = W_0$, $c_n \leq W_n$ will not diverge, and hence also b_n cannot diverge.

We have therefore shown that components of V_n either all of them diverge to infinity, or they all of them do not diverge.

We return to the recursion, $V_{n+1} = [R(V_n + X_n e_i - T_n e_{all})]^+$, and the definition $U_n = X_n - sT_n$. We then have: $\Sigma V_{n+1} \geq [\Sigma V_n + U_n]^+$. Using the result for

G/G1, if $\mathbb{E}(U_n) > 0$ then we have that ΣV_n diverges for any initial conditions, i.e. if $\rho > 1$ the G/G/s is unstable.
We now assume that $\rho < 1$, i.e. $\mathbb{E}(X_n - sT_n) < 0$. We will show that in that case $\mathbb{P}\big(\lim_{r\to\infty} V^r_{0,1} < \infty\big) = 1$. The events $\{\lim_{r\to\infty} V^r_{0,1} < \infty\}$ and $\{\lim_{r\to\infty} V^r_{0,1} = \infty\}$ are tail evants, so by ergodicity of $(X_n, T_n$ their probability is either 0 or 1.
We use:

$$\mathbb{E}(V^r_0 - V^r_1) = \mathbb{E}(V^r_0 - V^{r-1}_1) + \mathbb{E}(V^{r-1}_1 - V^r_1) = \mathbb{E}(V^{r-1}_1 - V^r_1) \le 0. \qquad (*)$$

and

$$\sum_{j=1}^{s} V^r_{1,j} = (V^r_{0,1} + X_0 - T_0)^+ + \sum_{j=2}^{s} (V^r_{0,j} - T_0)^+$$
$$= \sum_{j=1}^{s} V^r_{0,j} - \left[V^r_{0,1} \wedge (T_0 - X_0) + \sum_{j=2}^{s} (V^r_{0,j} \wedge T_0) \right]$$

To get that:

$$\mathbb{E}\left[V^r_{0,1} \wedge (T_0 - X_0) + \sum_{j=2}^{s} (V^r_{0,j} \wedge T_0) \right] \le 0.$$

Now, if $\mathbb{P}(V^r_{0,1} \to \infty) = 1$ then of course also $\mathbb{P}(\sum_{j=2}^{s} V^r_{j,1} \to \infty) = 1$, and then the expression in the square parentheses is a.s. equal to $sT_0 - X_0$ and so divergence of $V^r_{0,1}$ implies $E(sT_0 - X_0) \le 0$, which contradicts $\rho < 1$. This proves that $\rho < 1$ implies $V^\infty_{0,1} < \infty$.

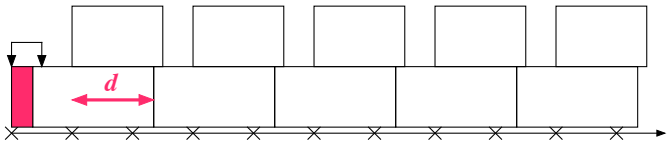
4.4 Show that the sequence of waiting time vectors for G/G/s, obtained by the Loynes construction, is minimal in some sense. What else is needed to show that it is unique ?

Solution:
It can now be shown exactly like the single server case that this sequence is minimal among all stationary sequences that satisfy the generalized Lindley equation.
To show that it is unique one would need to show that starting at 0 at some time $-r$, the vector V^r_n is sure to reach 0 for some n. However this is not the case, as the next example shows.

4.5 Consider the following sequences: $a_n = 1$, $b_{2n} = 2$, $b_{2n+1} = 3/2$, $-\infty < n < \infty$. Let a G/G/2 system have interarrivals $\{T_n = a_n\}_{-\infty<n<\infty}$, and sequence of service times with probability $\frac{1}{2}$ given by $\{X_n = b_n\}_{-\infty<n<\infty}$, and with probability $\frac{1}{2}$ given by $\{X_n = b_{n+1}\}_{-\infty<n<\infty}$. Show that the stationary sequence of waiting times for this system is not unique. Explain why.

Solution:
let $d \in (1, 3/2)$. Then the sequence $V_n = (0, d)$, $V_{n+1} = (1/2, 1-d)$ with probability half and $V_n = (1/2, 1-d)$, $V_{n+1} = (0, d)$ wth probability half is a

stationary sequence that satisfies the recursion, for every $1 < d < 3/2$. This is illustrated by the following figure:



The main point here is that this is different from the single server case. For a single server queue, $\rho < 1$ implies that the queue is repeatedly empty for a whole time interval. With many servers this is not the case, here $\rho < 1$ but the queue is never empty.

Note that in this example, the successive processing times are not i.i.d., X_n and X_{n+1} have the same distribution, but they are not independent: X_{n+1} is completely determined by X_n.

4.6 Let $Y_j = 1$ or $Y_j = -1$ with equal probabilities, and $S_0 = 0$, $S_n = Y_1 + \cdots + Y_n$. S_n is the simple symmetric random walk. The path of the random walk is at (t, k) if it is in position k at time t, i.e. $S_t = k$.

(i) Let $N_{t,k}$ be the number of paths that reach k in t steps, and $p_{t,k} = \mathbb{P}(S_t = k)$. Calculate $N_{t,k}$ and $p_{t,k}$.

(ii) Prove the reflection principle: The number of paths from (t_0, k_0) to (t_1, k_1), both on the same side of the time axis, that cross or touch the time axis equals the total number of paths from $(t_0, -k_0)$ to (t_1, k_1).

(iii) Prove the ballot theorem: For $a > b$, of all the paths with a positive and b negative steps, show that the probability of paths where the number of positives exceeds negatives at all times is $\frac{a-b}{a+b}$.

(iv) Calculate $u_{2n} = \mathbb{P}(S_{2n} = 0)$, the probability to return to 0 at time $2n$.

(v) Show that the probability to return to 0 for the first time at time $2n$,

$$\begin{aligned} f_{2n} &= \mathbb{P}(S_{2n} = 0, S_j \neq 0, j = 1, \ldots, 2n-1) \\ &= u_{2n-2} - u_{2n} = \frac{1}{2n-1} u_{2n} = \frac{1}{2n-1} \binom{2n}{n} \frac{1}{2^{2n}} \end{aligned}$$

(vi) Prove that the simple symmetric random walk returns to 0 infinitely often.

(vii) Show that the expected time to return to 0 is infinite (use Stirling's formula).

Solution:

(i) Let p, q be the number of positive and negative steps respectively, so: $n = p + q$, $x = p - q$. Then: $N_{t,x} = \binom{p+q}{p}$, and $\mathbb{P}(S_t = x) = N_{t,x} 2^{-t} = \binom{p+q}{p} \left(\frac{1}{2}\right)^{p+q}$

(ii) Consider a path from (t_0, k_0) to (t_1, k_1) $(k_0, k_1 > 0)$ that touches or crosses the $(t, 0)$ line. Let T be the first time that it touches or crosses $(t, 0)$, so $(T, 0)$ is on the path. The number of paths from $(t_0, -k_0)$ to $(T, 0)$ equals

those from (t_0, k_0) to $(T, 0)$. So there is a one to one mapping of paths from (t_0, k_0) to (t_1, k_1) $(k_0, k_1 > 0)$ that crosses the line $(t, 0)$ to all paths form $(t_0, -k_0)$ to $(T, 0)$.
(iii) Total number of paths: $N_{a+b,a-b} = \binom{a+b}{a}$. Path where positives exceed negatives go from (0,0) to (1,1) and then continue to $(a + b, a - b)$ without crossing or touching the $(t, 0)$ line. Their number equals all paths from (1,1) to $(a + b, a - b)$, minus all paths from $(-1, 1)$ to $(a + b, a - b)$. We get the number of path where positives exceed negatives:

$$\binom{a+b-1}{a-1} - \binom{a+b-1}{a} = \binom{a+b}{a}\frac{a-b}{a+b}.$$

(iv) $u_{2n} = \binom{2n}{n}\left(\frac{1}{2}\right)^n \sim 1/\sqrt{\pi n}$ by Stirling's formula ($n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$).
(v) We show first:

$$\mathbb{P}(S_1 > 0, S_2 > 0, \ldots, S_{2n} > 0) = \frac{1}{2}\mathbb{P}(S_{2n} = 0) = \frac{1}{2}u_{2n}. \qquad (*)$$

or equaivalently:

$$\mathbb{P}(S_1 \neq 0, S_2 \neq 0, \ldots, S_{2n} \neq 0) = \mathbb{P}(S_{2n} = 0) = u_{2n}. \qquad (**)$$

We write:

$$\mathbb{P}(S_1 > 0, S_2 > 0, \ldots, S_{2n} > 0) = \sum_{r=1}^{\infty} \mathbb{P}(S_1 > 0, \ldots, S_{2n-2} > 0, S_{2n} = 2r).$$

Each term in the sum is the probability of going from $(0, 0)$ to $(2n, 2r)$ with positives exceeding negative at all steps, terms for $r > n$ are 0. By the ballot theorem:

$$\begin{aligned}\mathbb{P}(S_1 > 0, \ldots, S_{2n-2} > 0, S_{2n} = 2r) &= \frac{2r}{2n}\binom{2n}{n+r}2^{-2n} \\ &= \frac{1}{2}\left(p_{2n-1,2r-1} - p_{2n-1,2r+1}\right)\end{aligned}$$

where it is useful to write it in this second way. It is now seen that the sum above collapses, and we are left with: $\frac{1}{2}p_{2n-1,1}$.

$$\mathbb{P}(S_1 > 0, \ldots, S_{2n} > 0) = \sum_{r=1}^{\infty} \mathbb{P}(S_1 > 0, \ldots, S_{2n} = 2r) = \frac{1}{2}p_{2n-1,1}.$$

But u_{2n} is the probability of going from $(0, 0)$ to $(1, \pm 1)$, which is 1, and then from $(1, \pm 1)$ to $(2n, 0)$, and the latter equals the probability of going from $(0, 0)$ to $(2n - 1, \pm 1)$, which is $p_{2n-1,1}$, and we have proved (*).
It is also easy to see also that:

$$\mathbb{P}(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n} \geq 0) = u_{2n}.$$

Since $(S_1 > 0 S_2 > 0, \ldots, S_{2n} > 0)$ happens if we go from $(0, 0)$ to $(1, 1)$

(probability 1/2), and then we have $S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n-1} \geq 0$) which, because $2n-1$ is odd, means $S_{2n-1} > 0$, so also $S_{2n} \geq 0$
We are now ready to show: $f_{2n} = u_{2n-2} - u_{2n}$. f_{2n} is the probability of the event that $S_1 \neq 0, \ldots, S_{2k} \neq 0$, for $k = n-1$ but not for $k = n$, so:

$$f_{2n} = \mathbb{P}(S_1 \neq 0, \ldots, S_{2n-2} \neq 0) - \mathbb{P}(S_1 \neq 0, \ldots, S_{2n} \neq 0) = u_{2n-2} - u_{2n}.$$

We also have:

$$\begin{aligned} u_{2n-2} - u_{2n} &= \binom{2n-2}{n-1} 2^{2n-2} - \binom{2n}{n} 2^{2n} = \frac{(2n-2)!n^2 4 - (2n)!}{2^{2n} n! n!} \\ &= \frac{1}{2n-1} \binom{2n}{n} 2^{2n} = \frac{1}{2n-1} u_2 n. \end{aligned}$$

(vi) The state 0 is recurrent, i.e. state 0 is visited infinitely often since:

$$\sum_{n=1}^{\infty} f_{2n} = \sum_{n=1}^{\infty} (u_{2n-2} - u_{2n}) = u_0 = 1.$$

(vii) The state 0 is null recurrent, i.e. expected time to return to 0 is infinite:

$$\begin{aligned} \text{Expected time ot return} &= \sum_{n=1}^{\infty} 2n f_{2n} = \sum_{n=1}^{\infty} \frac{2n}{2n-1} \binom{2n}{n} 2^{-2n} \\ &\sim \sum_{n=1}^{\infty} \frac{2n}{2n-1} \frac{1}{\sqrt{\pi n}} = \infty. \end{aligned}$$

4.7 Prove the formula (4.7) for the distribution of the stationary workload.
Solution:
When we observe the stationary G/G/1 queue at an arbitrary time t, then the previous arrival was a time T_{Bwd} earlier. At that time, independent of the current interarrival, the last arrival had a waiting time V_∞ and an independent service time X_n. So the workload at t is $W_\infty \stackrel{D}{=} (V_\infty + X_n - T_{Bwd})^+$ where all three components are independent. Recall that $T_{Bwd} \stackrel{D}{=} T_{Fwd} \stackrel{D}{=} T_{eq}$

4.8 Find the average waiting time for the D/M/1 queue and compare it to the Kingman bound.
Solution:
Solution of next Exercise shows that for heavy traffic we have:

$$\bar{W} = \frac{1}{1-\gamma} \approx \frac{1}{1-\rho} \frac{1}{2}$$

4.9 Find the average waiting time for a G/M/1 queue and compare it to the Kingman bound.
Solution:(G/M/1 queue described in most queueing texts)
Take service exponential with rate 1, interarrivals T_n distributed F with rate $\rho < 1$ The stationary distribution of queue length found by an arrival at the regeneration times of arrivals is geometric: $\mathbb{P}(Q^{Arrivals} = k) = (1 -$

$\gamma)\gamma^k$, $k = 0, 1, \ldots$. It follows that the sojourn time for an arrival, given by is the sum of rate 1 exponential service times, is (similar to calculation for M/M/1): $\bar{W} = \frac{1}{1-\gamma}$. Unfortunately, γ cannot be obtained explicitly, it is the < 1 solution of the equation:

$$\gamma = \int_0^\infty e^{-(1-\gamma)y)} dF(y)$$

However, for ρ close to 1, we will have heavy traffic and also γ close to 1. We can then take Taylor expansion of the transform:

$$\begin{aligned}\gamma &= \sum_{k=0}^{\infty} (-1)^k \frac{(1-\gamma^k)}{k!} \mathbb{E}(T^k) \\ &= 1 - (1-\gamma)\mathbb{E}(T) + \frac{1}{2}(1-\gamma)^2 \mathbb{E}(T^2) + o(1-\gamma)^2)\end{aligned}$$

We get:

$$0 \approx 1-\gamma-(1-\gamma)\mathbb{E}(T)+(1-\gamma)^2\frac{1}{2}\mathbb{E}(T^2) = (1-\gamma)(1-\frac{1}{\rho})+(1-\gamma)^2\frac{1}{\rho}\frac{1+c_A^2}{2}$$

muliplying by ρ, cancelling $(1-\gamma)$ we get:

$$1 - \rho = (1-\gamma)\frac{1+c_A^2}{2}$$

and so the expected sojourn time is

$$\bar{W} = \frac{1}{1-\gamma} \approx \frac{1}{1-\rho}\frac{1+c_A^2}{2}$$

which is exactly what we would get from the Kingman bound.

5

The Basic Probability Functional Limit Theorems

Exercises

5.1 A sequence of stochastic processes Z_n converges in probability to a stochastic process Z (written $Z_n \rightarrow_p Z$) if for any $\epsilon > 0$

$$\mathbb{P}(d(Z_n, Z) > \epsilon) \rightarrow 0, \text{ as } n \rightarrow \infty.$$

Here $D(\cdot, \cdot)$ is the u.o.c. distance if $Z_n \in \mathbb{D}$ and $Z \in \mathbb{C}$, or it is the $J1$ topology distance if $Z_n \in \mathbb{D}$ and $Z \in \mathbb{D}$.
Show that $Z_n \rightarrow Z$ a.s. implies $Z_n \rightarrow_p Z$.
Solution
$Z_n \rightarrow Z$ means that except for a set O of measure zero, $d(Z_n(\omega), Z(\omega)) \rightarrow 0$.
Let

$$A_n = \bigcup_{m \geq n} \{\omega : d(Z_m(\omega), Z(\omega) > \epsilon\}.$$

A_n is a decreasing sequence of states, and so $A_\infty = \bigcap_{n \geq 1} A_n$ is well defined. The probabilities of A_n are also decreasing, to $\mathbb{P}(A_\infty)$. We show that this is 0. Assume ω is not in O. Then for some n_ω, $d(Z_n(\omega), Z(\omega) < \epsilon)$ for all $n > n_\omega$, so ω in not in A_∞. Hence $A_\infty \subseteq O$, and so $\mathbb{P}(A_\infty) = 0$.
By the definition of A_n we have $\{\omega : d(Z_n(\omega), Z(\omega)) > \epsilon\} \subseteq A_n$, so:

$$\mathbb{P}(d(Z_n, Z) > \epsilon) \leq \mathbb{P}(A_n) \rightarrow 0.$$

which completes the proof.

5.2 For a seqeunce of random variables X_n, and a constant c, show that if $X_n \rightarrow_w c$, then $X_n \rightarrow_p c$.
Solution
Convergence $X_n \rightarrow_w c$ implies

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n \leq x) = \begin{cases} 0 & x < c \\ 1 & x > c \end{cases}$$

i.e. convergence of F_n to F at all points except the discontinuity of F at c. Hence, for $\epsilon > 0$:

$$\mathbb{P}(X_n - c > \epsilon) \rightarrow 0, \qquad \mathbb{P}(c - X_n > \epsilon) \rightarrow 0$$

i.e. $\mathbb{P}(|X_n - c| > \epsilon) \to 0$ which is the definition of $X_n \to_p c$.

5.3 For a sequence of stochastic processes Z_n with paths in $\mathbb{D}$, and a deterministic continuous function z, show that if $\mathcal{Z}_n \to_w z$, then $\mathcal{Z}_n \to_p z$, i.e. weak convergence of a sequence of stochastic processes to a continuous deterministic function implies convergence in probability.

Solution

$\mathcal{Z}_n \to_w z$ means that we can construct a probability space in which Z'_n is distributed like Z_n, and in which almost surely for ω, $Z'_n(t,\omega) \to z(t)$ u.o.c. as $t \to \infty$. Almost sure convergence implies convergence in probability, so we have:

$$\mathbb{P}(\sup_{0<t\leq T} |Z'_n(t) - z(t)| > \epsilon) \to 0, \text{ for any } T \text{ as } n \to \infty$$

But for every n, Z'_n and Z_n have the same distribution, and because z is constant, the joint distributions of (Z'_n, z) and (Z_n, z) are the same, so

$$\mathbb{P}(\sup_{0<t\leq T} |Z_n(t) - z(t)| > \epsilon) \to 0, \text{ for any } T \text{ as } n \to \infty,$$

which completes the proof. This works because if X, Y have the same distribution, then for every constant c, (X, c) and $Y, c)$ have the same distribution. It would not work for $Z(t)$ non-deterministic, because (Z_n, Z) may have a different joint distribution than (Z'_n, Z').

5.4 Show that the Strong Approximation Theorem 5.11, implies $\bar{S}^n(t) = \frac{1}{n}S(nt) \to_p mt$, $\hat{S}^n(t) = \sqrt{n}\big(\bar{S}^n(t) - mt\big) \to_w \sigma BM(t)$, i.e. the FSLLN (in probability but not a.s.), and the FCLT for random walks.

Solution: We have, by the strong approximation theorem 5.11, assuming existence of $r > 2$ moments, that we can construct copies $S'(t)$ of $S(t)$ and a BM such that:

$$\sup_{0\leq t\leq T} |S'(t) - m\,t - \sigma\, BM(t)| =_{a.s} o(T^{1/r}) \quad \text{as } T \to \infty$$

We then have for any T:

$$\frac{\sup_{0\leq nt\leq nT} |S'(nt) - m\,nt - \sigma\, BM(nt)|}{(nT)^{1/r}} \to_{a.s.} 0, \quad \text{as } n \to \infty \text{ or:}$$

$$\frac{1}{T^{1/r}} \frac{\sup_{0\leq t\leq T} |S'(nt) - m\,nt - \sigma\, BM(nt)|}{n^{1/r}} \to_{a.s.} 0, \quad \text{but } n^{1/r} < \sqrt{n}, \text{ so:}$$

$$\frac{\sup_{0\leq t\leq T} |S'(nt) - m\,nt - \sigma\, \sqrt{n}BM(t)|}{\sqrt{n}} \to_{a.s.} 0, \quad \text{as } n \to \infty \text{ or:}$$

$$\sup_{0\leq t\leq T} \big|\sqrt{n}(\bar{S}'^n(t) - mt) - BM(t)\big| \to 0, \quad \text{as } n \to \infty.$$

which proves the FCLT. Clearly also dividing by n will give

$$\sup_{0\leq t\leq T} \left|(\bar{S}'^n(t) - mt) - \frac{1}{\sqrt{n}}BM(t)\right| \to 0, \quad \text{as } n \to \infty.$$

but clearly, $\sup_{0\le t\le T}\left|\frac{1}{\sqrt{n}}BM(t)\right| \to 0$, so we have:

$$\sup_{0\le t\le T}\left|(\bar{S}'^{n}(t) - mt)\right| \to 0, \quad \text{as } n \to \infty.$$

This implies that $\bar{S}^n(t) \to_w mt$ as $n \to \infty$, but since mt is a deterministic function, by Exercise 5.2, this implies that $\bar{S}^n(t) \to_p mt$ as $n \to \infty$, which is a weaker version of FSLLN.

5.5 Complete the proof of the FSLLN for renewal processes.
Solution:

Proof of the FSLLN for renewal processes For $X_n \ge 0$ i.i.d., $\mathbb{E}(X_n) = m$, $S_n = \sum_{j=1}^{n} X_n$, $S(t) = S_{\lfloor t \rfloor}$, $\mathcal{X}(t) = \max\{n : S_n \le t\}$. We showed that $\frac{1}{n}S(nt) \to mt$ a.s.. This means that for almost all sample paths, the lines mt and $\frac{1}{n}S(nt)$ get arbitrarily close in the sense of u.o.c. as $n \to \infty$. If we reflect those lines around the 45^0 line we get the lines $\frac{1}{n}\mathcal{X}(nt)$ and $\frac{1}{m}t$. This completes the proof. □

5.6 Prove the following result:
If $Z_n(t) \to_w Z(t)$, and $Y_n(t) \to_w y(t)$ where $y(t)$ is deterministic, then the jointly distributed sequence $(Z_n(t), Y_n(t)) \to_w (Z(t), y(t))$.
Solution:
We can construct sequences of stochastic processes (Z'_n, Y'_n) that for every n have the same distribution as the processes (Z_n, Y_n), and a stochastic process Z' distributed as Z, so that $Z_n(t,\omega) \to Z'(t,\omega)$ u.o.c. a.s., and $Y_n(t,\omega) \to y(t)$ u.o.c. a.s. Hence, $\big(Z_n(t,\omega), Y_n(t,\omega)\big) \to (Z'(t,\omega), y(t))$ u.o.c. a.s., and the limit is the same whatever the joint distributions of the (Z'_n, Y'_n), because $(Z'(t,\omega)$ and $y(t))$ are independent. This completes the proof.

6

Scaling of G/G/1 and G/G/∞

Exercises

6.1 Prove by induction that the implicit conditions of the dynamics, the non-negativity, and work conservation, uniquely determine the queue length process.

Solution

We have $Q(0), \mathcal{A}(t), \mathcal{S}(t)$. While $Q(t) > 0$, as time moves on $d\mathcal{I}(t) = 0$ so $d\mathcal{T}(t) = 1$, and we wait for the next arrival or departure at t_1. At the next event, if arrival, Q increases, if departure, Q decreases. If the decrease reaches $Q(t_1) = 0$,

by the requirement that $Q(t_1 + s) \geq 0$ we must have $d\mathcal{T}(t_1 + s) = 0$ until next arrival, so $d\mathcal{I}(t_1 + s) = 1$, and we wait for next arrival. This completes the description.

6.2 Show that $y(t) = -\inf\{0, x(s) : 0 \leq s \leq t\}$ satisfies conditions (i) – (iii) of the Skorohod reflection mapping.

Solution

(i) Let $y(t) = -\inf\{0, x(s) : 0 \leq s \leq t\}$. Then $x(t)+y(t) = x(t)-\inf\{0, x(s) : 0 \leq s \leq t\} \geq 0$.

(ii) Since $x(0) \geq 0$, $y(0) = 0$. Clearly, $\inf\{0, x(s) : 0 \leq s \leq t\}$ is non-increasing in t, so $y(t)$ is non-decreasing

(iii) Assume that $z(t) > 0$ for $a \leq t \leq b$. Then $x(a) > \inf\{0, x(s) : 0 \leq s \leq a\} = -y(a)$. Assume now than $y(b) > y(a)$. Then $\inf\{0, x(s) : 0 \leq s \leq b\} < \inf\{0, x(s) : 0 \leq s \leq a\}$ and so $\inf\{0, x(s) : 0 \leq s \leq b\} = \inf\{0, x(s) : a \leq s \leq b\} < 0$. We now have:

$$\inf_{a\leq s\leq b} z(s) = \inf_{a\leq s\leq b}(x(s)+y(s)) \leq \inf_{a\leq s\leq b} x(s)+\inf_{a\leq s\leq b} y(s) = \inf_{a\leq s\leq b} x(s)-\inf_{a\leq s\leq b} x(s) = 0$$

which contrdicts $z(t) > 0$ for $a \leq t \leq b$.

6.3 Show that $y(t) = -\inf\{0, x(s) : 0 \leq s \leq t\}$ is the minimal function that satisfies conditions (i) and (ii) of the Skorohod reflection mapping

Solution

Let y^*, $z^* = x+y^*$ satisfy (i) and (ii) and assume $y^*(t) < y(t)$ for some t. Let t_0 be the time point at which $\inf\{0, x(s) : 0 \leq s \leq t\} = \min\{x(t_0), \lim_{s\nearrow t_0} x(s)\}$.

Note that

$$y^*(t_0) \le y^*(t) < y(t) = y(t_0) = \inf\{0, x(s), 0 \le s \le t_0\}.$$

But this implies that $\min\{z^*(t_0), \lim_{s \nearrow t_0} z^*(s)\} < 0$, which is a contradiction. Hence, $y(t)$ is minimal.

6.4 Show that $y(t) = -\inf\{0, x(s) : 0 \le s \le t\}$ is the unique function that satisfies conditions (i) – (iii) of the Skorohod reflection mapping.

Solution

For given x assume $y^*, z^* = x + y^*$ are another solution satisfying (i)–(iii). We show that $(z^* - z)^2 = 0$. We have:

$$\begin{aligned}\frac{1}{2}(z^*(t) - z(t))^2 &= \frac{1}{2}(z^*(0) - z(0))^2 + \int_0^t (z^*(s) - z(s))d[(z^*(s) - z(s))] \\ &= 0 + \int_0^t (z^*(s) - z(s))d[(y^*(s) - y(s))] \\ &= \int_0^t (z^*(s) - z(s))dy^*(s) + \int_0^t (z(s) - z^*(s))dy(s)\end{aligned}$$

but $\int_0^t z^*(s)dy^*(s) = 0$ by (iii), and $\int_0^t z(s)dy^*(s) \ge 0$ by (i), (ii), so the first part of the sum is ≤ 0, and by the same argument the second part of the sum is ≤ 0, so $(z^*(t) - z(t))^2 \le 0$ which implies it is 0.

6.5 Show that conditions (i), (ii) and (iii) of the Skorohod reflection mapping are equivalent to conditions (i), (ii) and (iii').

Solution

$$\text{(i) - (iii)} \iff y(t) = -\inf\{0, x(s) : 0 \le s \le t\} \iff \text{(i) - (iii')}.$$

The first implication follows from Exercises 6.2, 6.4. The second follows from Exercises 6.3, 6.4.

6.6 Show that for single server queue with renewal arrivals and i.i.d. service times, under work conserving policy, the busy time $\mathcal{T}(t) \to \infty$ as $t \to \infty$.

Solution

For time t, let $t_0 = \inf\{s : Q(r) > 0,\ s < r \le t\} \ge 0$ be the last time that the system was empty. Then $\mathcal{T}(t) = \mathcal{S}_{\mathcal{A}(t_0)} + t - t_0$. Then as $t \to \infty$, either $t_0 \to \infty$ or $t - t_0 \to \infty$ or both. If $t_0 \to \infty$, $\mathcal{A}(t_0) \to \infty$, and so also does $\mathcal{S}_{\mathcal{A}(t_0)}$, so $\mathcal{T}(t) = \mathcal{S}_{\mathcal{A}(t_0)} + t - t_0 \to \infty$.

6.7 Derive the fluid limit for the workload directly from the fluid limit of the queue length.

Solution

The fluid limit of the queues is $\bar{Q}^n(t) \to [\bar{Q}(0) + (\lambda - \mu)t]^+$. Recall that while the limit is > 0 it incorporates a very large number of customers, hence as $n \to \infty$, $\frac{1}{n}\sum_{j=1}^{Q(nt)} \to \frac{1}{\mu}\bar{Q}(t)$.

6.8 Derive the fluid limit for the workload by scaling and using Skorohod reflection on (6.3)–(6.4).

Solution
We need to look at $\frac{1}{n}S_{\mathcal{A}(nt)}$:

$$\frac{1}{n}S_{\mathcal{A}(nt)} = \frac{1}{n}\sum_{j=1}^{\mathcal{A}(nt)} v_j = \frac{\mathcal{A}(nt)}{n}\frac{1}{\mathcal{A}(nt)}\sum_{j=1}^{\mathcal{A}(nt)} v_j \to \lambda t\frac{1}{\mu} = \rho t.$$

This holds a.s. pointwise for all t. To complete the proof of u.o.c. convergence, see proof of FSLLN, Theorem 5.8.
We have:

$$\bar{\mathcal{W}}^n(t) = \left(\bar{\mathcal{W}}^n(0) + \frac{1}{n}S_{\mathcal{A}(nt)} - t\right) + \bar{\mathcal{I}}^n(t).$$

By (6.4) this satisfies the conditions for Skorohod reflection, with $x^n(t) = \bar{\mathcal{W}}^n(0) + \frac{1}{n}S_{\mathcal{A}(nt)} - t$, where $x^n(t) \to x(t) = \bar{\mathcal{W}}(0) + (\rho - 1)t$ so

$$\bar{\mathcal{W}}^n(t) \to \bar{\mathcal{W}}(t) = \psi(x(t) = (\bar{\mathcal{W}}(0) + (\rho - 1)t)^+.$$

6.9 Show that if $\mathcal{Q}^n(0)/n \to 0$, $\lambda^n \to \lambda$, and $\lambda^n - \mu^n \to 0$, then $\bar{\mathcal{Q}}(t) = 0$ and $\bar{T}(t) = t$.
Solution
Denote by $U(t)$, $V(t)$ the rate 1 unscaled renewal processes of arrivals and services (i.e. $\mathcal{A}^n(t) = U(\lambda_n t)$, $\mathcal{S}^n(t) = V(\mu_n t)$). The netput is

$$\mathcal{X}^n(t) = \mathcal{Q}^n(0) + (\lambda^n - \mu^n)t + (U(\lambda^n t) - \lambda^n t) - (V(\mu^n \mathcal{T}(t)) - \mu^n \mathcal{T}(t)).$$

and for the fluid, scaling time and space by n we have

$$\bar{\mathcal{X}}^n(t) = \bar{\mathcal{Q}}^n(0) + (\lambda^n - \mu^n)t + (\bar{U}(\lambda^n t) - \lambda^n t) - (\bar{V}(\mu^n \bar{\mathcal{T}}^n(t)) - \mu^n \bar{\mathcal{T}}^n(t)).$$

Now, as $n \to \infty$: $\bar{\mathcal{Q}}^n(0) \to 0$, $(\lambda^n - \mu^n) \to 0$, $\lambda^n \to \lambda$, $\mu^n \to \lambda$, and uniformly on compacts, almost surely: $\bar{U}(\lambda^n t) - \lambda^n t \to 0$, since this is true point wise, and it is true uniformly on compacts because for $t \in [0, T]$ we have $\lambda_n t \in [0, \lambda T + \epsilon]$ for n large enough. Similarly, because $\bar{\mathcal{T}}^n(t) \le t$, also $\bar{V}(\mu^n \bar{\mathcal{T}}(t)) - \mu^n \bar{\mathcal{T}}(t) \to 0$. So:

$$\bar{\mathcal{X}}^n(t) \to^{a.s.}_{u.o.c.} 0, \text{ as } n \to \infty$$

But,

$$\bar{\mathcal{Q}}^n(t) = \bar{\mathcal{X}}^n(t) + \bar{\mathcal{Y}}^n(t) = \bar{\mathcal{X}}^n(t) + \mu^n(t - \bar{\mathcal{T}}^n(t))$$

and these converge term by term to

$$\bar{\mathcal{Q}}(t) = \bar{\mathcal{X}}(t) + \bar{\mathcal{Y}}(t) = \bar{\mathcal{X}}(t) + \lambda(t - \bar{\mathcal{T}}(t))$$

but, since $\bar{\mathcal{X}}(t) = 0$, we have $\bar{\mathcal{Q}}(t) = \psi(\bar{\mathcal{X}}(t)) = 0$, and $\bar{\mathcal{Y}}(t) = \phi(\bar{\mathcal{X}}(t)) = 0$, and hence also $\bar{\mathcal{T}}(t) = t$.

6.10 Show that when $Q^n(0)/n \to 0$, $\lambda^n \to \lambda$ and $\lambda^n - \mu^n \to 0$, then $\frac{1}{\sqrt{n}}(\mathcal{A}^n(nt) - \lambda nt) \to_w (\lambda c_a^2)^{\frac{1}{2}} BM(t)$ as well as $\frac{1}{\sqrt{n}}(\mathcal{S}^n(\mathcal{T}(nt)) - \lambda nt) \to_w (\lambda c_s^2)^{\frac{1}{2}} BM(t)$, and the joint distribution of the diffusion scaled arrival and service processes converges weakly jointly to the joint distribution of two independent Brownian motions.

Solution

Denote by $U(t)$, $V(t)$ the rate 1 unscaled renewal processes of arrivals and services (i.e. $\mathcal{A}^n(t) = U(\lambda_n t)$, $\mathcal{S}^n(t) = V(\mu_n t)$), and let c_a, c_s be the coefficient of variation of interarrival and service times.

$$\frac{1}{\sqrt{n}}(\mathcal{A}^n(nt) - \lambda^n nt) = \frac{1}{\sqrt{n}}(U(\lambda^n nt) - \lambda^n nt) = \sqrt{n}(\bar{U}(\lambda^n t) - \lambda^n t)$$
$$\to_w c_a BM(\lambda t) =_D (\lambda c_a^2)^{\frac{1}{2}} BM(t).$$

In fact, we have for the renewal process $U(t)$ that $\sqrt{n}(\bar{U}(t) - t) \to c_a BM(t)$, and by time change lemma, for the deterministic series of functions $\phi^n(t) = \lambda_n t \to \lambda t$ u.o.c., so $\sqrt{n}(\bar{U}(\lambda_n t) - \lambda_n t) \to c_a BM(\lambda t)$.
For $\mathcal{S}^n(\mathcal{T}(nt)$ we use in addition the time change $\mathcal{T}_n(t) \to t$.
Finally, the processes $U(t)$, $V(t)$ are independent, so the convergence holds for their joint distribution, which converges to jointly distributed independent Brownian motions.

6.11 Obtain the fluid and diffusion scaling limits of $\mathcal{S}_{\mathcal{A}(t)}$.

Solution

Fluid scaling limit:

$$\frac{1}{n}\mathcal{S}_{\mathcal{A}(nt)} = \frac{1}{n}\sum_{j=1}^{\mathcal{A}(nt)} v_j = \frac{\mathcal{A}(nt)}{n}\frac{1}{\mathcal{A}(nt)}\sum_{j=1}^{\mathcal{A}(nt)} v_j \to \lambda t \frac{1}{\mu} = \rho t, \text{ u.o.c. a.s.}$$

Diffusion scaling limit:

$$\sqrt{n}\left(\frac{1}{n}\mathcal{S}_{\mathcal{A}(nt)} - \rho t\right) = \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{\mathcal{A}(nt)}\left(v_j - \frac{1}{\mu}\right)\right) + \sqrt{n}\left(\frac{1}{n}\mathcal{A}(nt)\frac{1}{\mu} - \rho t\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{j=1}^{\lfloor n\frac{1}{n}\mathcal{A}(nt)\rfloor}\left(v_j - \frac{1}{\mu}\right) + \frac{1}{\mu}\frac{1}{\sqrt{n}}(\mathcal{A}(nt) - \lambda t)$$
$$\to_w \sigma_s BM_1(\lambda t) + \frac{1}{\mu}(\lambda c_a^2)^{1/2} BM_2(t)$$
$$= \frac{1}{\mu}\left((\lambda c_s^2)^{1/2} BM_1(t) + (\lambda c_a^2)^{1/2} BM_2(t)\right) = \frac{1}{\mu}(\lambda c_a^2 + \lambda c_s^2)^{1/2} BM(t).$$

Here we used time change $\frac{1}{n}\mathcal{A}(nt) \to \lambda t$ in the first limit.

6.12 Show that for all three cases, when $\rho > 1$, when $\rho < 1$ and when ρ^n satisfies

the conditions of heavy traffic (6.12), we have $\mu\hat{\mathcal{W}}^n(t)/\hat{Q}^n(t) \to_w 1$ (where we let $0/0 = 1$).

Solution

We will show that this is true for light traffic and for heavy traffic with $\rho \nearrow 1$. For $\rho > 1$ the actual result needs modification: It is true for the fluid limit but not for the diffusion scaled deviations.

(i) Clearly, when $\rho < 1$, the limits (convergence in probability) are $\hat{\mathcal{W}}(t) = \hat{Q}(t) = 0$.

(ii) Consider next the case $\rho \approx 1$, specfically assume: $\rho^n \to 1$ and $\sqrt{n}(1 - \rho^n) \to \delta$, $0 < \delta < \infty$. Then $\hat{Q}^n(t) = \frac{1}{\sqrt{n}} Q^n(nt)$ and $\hat{\mathcal{W}}^n(t) = \frac{1}{\sqrt{n}} \mathcal{W}^n(nt)$, We note then that $\mathcal{W}^n(nt)$ consists of the workload of exactly the customers in $Q^n(nt)$, (except perhaps some of the work of the job in process. Furthermore, $Q^n(nt)$ is of the order $O(\sqrt{n})$. So, by the SLLN (ignoring the slight effect of the single customer in service),

$$\frac{\hat{\mathcal{W}}^n(t)}{\hat{Q}^n(t)} = \frac{\mathcal{W}^n(nt)}{Q^n(nt)} = \frac{\sum_{j=1}^{Q^n(nt)} v_j}{Q^n(nt)} \to \mu^{-1}, \quad \text{u.o.c., a.s.}$$

(iii) When $\rho > 1$, then by the same argument,

$$\frac{\bar{\mathcal{W}}^n(t)}{\bar{Q}^n(t)} = \frac{\mathcal{W}^n(nt)}{Q^n(nt)} = \frac{\sum_{j=1}^{Q^n(nt)} v_j}{Q^n(nt)} \to \mu^{-1}, \quad \text{u.o.c., a.s.}$$

However, the scaled deviations from the fluid limit do not satisfy $\frac{\hat{\mathcal{W}}^n(t)}{\hat{Q}^n(t)} \to \mu^{-1}$. For example, if arrivals are deterministic and service times have mean m and variance σ^2, then clearly $\hat{Q}^n(t) \to 0$ while:

$$\hat{\mathcal{W}}^n(t) = \frac{\sum_{j=1}^{Q^n(nt)} (v_j - m)}{\sqrt{n}} \sim \frac{\sum_{j=1}^{(\lambda-\mu)nt} (v_j - m)}{\sqrt{n}} \to_w (\lambda - \mu)^{1/2} \sigma BM(t)$$

6.13 (∗) Obtain fluid and diffusion scaling and limits for $G/G/s$, fixed s, $\rho \nearrow 1$ [Iglehart and Whitt (1970a); Borovkov (1965)].

Solution Iglehart and Whitt (1970a)

It is tempting to write:

$$Q(t) = \left[Q(0) + (\lambda - s\mu)t + (\mathcal{A}(t) - \lambda t) - \sum_{j=1}^{s} (\mathcal{S}_j(\mathcal{T}_j(t)) - \mu \mathcal{T}_j(t)) \right] + \left[\mu \left(st - \sum_{j=1}^{s} \mathcal{T}_j(t) \right) \right] = \mathcal{X}(t) + \mathcal{Y}(t)$$

where $\mathcal{X}(t)$ does indeed converge to a Brownian motion under diffusion scaling. However, it is not true that $\mathcal{Y}(t)$ increases only when the system is empty, so we cannot present $Q(t)$ as the Skorohod reflection of $\mathcal{X}(t)$.

Instead we use a trick of Borovkov, we define a modified s server queueing system, $Q'(t)$ as follows: We assume that a server never shuts off, so it has some real and some dummy service completions (potential services), and when a customer arrives to a server that is doing a dummy job, he will actually depart at the end of the dummy job (as if we made him arrive a little earlier, or as if we have made his service shorter), and also any job is assigned to the server that will finish it earliest (join shortest workload). The modified queue has the following two properties: jobs leave in the same order that they arrived, and each completion of a potential service generates a departure as long as there are any jobs in the system.

The modified system now behaves almost like a single server queue with service $s\mu$. To be exact:

$$Q'(t) = \left[Q'(0) + (\lambda - s\mu)t + (\mathcal{A}'(t) - \lambda t) - \sum_{j=1}^{s} (\mathcal{S}_j(\mathcal{T}_j'(t)) - \mu \mathcal{T}_j'(t)) \right] + \left[\mu \left(st - \sum_{j=1}^{s} \mathcal{T}_j'(t) \right) \right] = \mathcal{X}'(t) + \mathcal{Y}'(t)$$

where now we have that the $\mathcal{Y}'(t)$ only increases when $Q(t) = 0$.

Here we wrote $\mathcal{A}'(t)$ for the modified arrival process, but $\mathcal{A}'(t) - \mathcal{A}(t) \leq s$ for all t, so both have the same fluid and diffusion scaling limits. Also, $\sum_{j=1}^{s} \mathcal{S}_j(t)$ as a sum of independent renewal processes has fluid scaling limit $s\mu t$ and diffusion scaling limit $(s\mu c_S^2)^{-1/2} BM(t)$.

It follows that the fluid scaling limit for is $(\bar{Q}(0) - s\mu t)^+$, and the diffusion scaling limit is $RBM_{\hat{Q}(0)}(t, \theta, \lambda c_A^2 + s\mu c_S^2)$, where $\theta = \lim \sqrt{n}(1 - \rho^n)$

6.14 Prove equation (6.18) for the auto-covariance of the diffusion limit of $Q(t)$ for G/D/∞.

Solution

For $t > x$ and $s < x$:

$$\begin{aligned}
&\mathbb{C}\text{ov}(\hat{Q}(t+s), \hat{Q}(t)) \\
&= \lambda c_a^2 \text{Cov}\big(BM(t+s) - BM(t+s-x), BM(t) - BM(t-x)\big) \\
&= \lambda c_a^2 \text{Cov}\big(BM(t+s) - BM(t) + BM(t) - BM(t+s-x), \\
&\qquad BM(t) - BM(t+s-x) + BM(t+s-x) - BM(t-x)\big) \\
&= 0 + \lambda c_a^2 \text{Cov}\Big(BM(t) - BM(t+s-x), BM(t) - BM(t+s-x)\Big) + 0 \\
&= \lambda c_a^2 (x - s).
\end{aligned}$$

For $t > x$, $s > x$ the intervals are disjoint, so the covariance is 0, so

$$\begin{aligned}
&\mathbb{C}\text{ov}(\hat{Q}(t+s), \hat{Q}(t)) \\
&= \lambda c_a^2 (x - s)^+.
\end{aligned}$$

6.15 Calculate the auto-covariance function for the stationary diffusion limit of the queue length process $Q(t)$ for G/Discrete/∞ system.

Solution

We assume arrivals are renewals $\mathcal{A}(t)$ with rate λt and c.o.v. c_a, and processing times are $x_1 < \ldots < x_K$ with probabilities $\alpha_1, \ldots, \alpha_k$. Then we could have an infinite subset of servers serving each of the different processing time types. We note that each of the $\mathcal{A}_i(t)$ is then in itself a renewal process, and for it $Q_i(t) = \mathcal{A}_i(t) - \mathcal{A}_i(t - x_i)$ is a G/D/∞ process. However the $\mathcal{A}_i(t)$ are not independent. We investigate $\mathcal{A}_i(t)$. We note that $(A_1(t), \ldots, \mathcal{A}_k(t) \sim \text{Multinomial}(\mathcal{A}(t), \alpha_1, \ldots, \alpha_K)$.

$$\begin{aligned}
&\mathbb{E}(A_i(t)) = \lambda t \alpha_i, \\
&\mathbb{V}(A_i(t)) = \lambda t[\alpha_i + \alpha_i^2(c_a^2 - 1)], \\
&\mathbb{C}\text{ov}(A_i(t), A_j(t)) = \lambda t(c_a^2 - 1)\alpha_i\alpha_j.
\end{aligned}$$

We then have that the vector $\frac{1}{n}\underline{\mathcal{A}}^n(t)$ scaled with $\lambda^n = n\lambda$, converges to a BM vector with drift $\lambda\underline{\alpha}$, and covariances as above.

To calculate the auto-covariance of $Q(t)$ we need to add up covariance of $\hat{\mathcal{A}}_i(t) - \hat{\mathcal{A}}_i(t - x_i)$ with $\hat{\mathcal{A}}_j(t + s) - \hat{\mathcal{A}}_i(t + s - x_j)$, over all pairs. For $i \neq j$:

$$\mathbb{C}\text{ov}\big(\hat{\mathcal{A}}_i(t) - \hat{\mathcal{A}}_i(t - x_i), \hat{\mathcal{A}}_j(t + s) - \hat{\mathcal{A}}_j(t + s - x_j)\big) = \min(x_j - s, x_i)^+\lambda(c_a^2 - 1)\alpha_i\alpha_j,$$

and also add up the covariances:

$$\mathbb{C}\text{ov}\big(\hat{\mathcal{A}}_i(t) - \hat{\mathcal{A}}_i(t - x_i), \hat{\mathcal{A}}_i(t + s) - \hat{\mathcal{A}}_i(t + s - x_i)\big) = (x_i - s)^+\lambda[\alpha_i^2 c_a^2 + \alpha_i(1 - \alpha_i)],$$

In summary for $s > 0$,

$$\begin{aligned}
\mathbb{C}\text{ov}\big(\hat{Q}(t), \hat{Q}(t + s)\big) = &\sum_{i=1}^{K}(x_i - s)^+\lambda[\alpha_i + \alpha_i^2(c_a^2 - 1)] \\
&+ \sum_{i \neq j} \min(x_j - s, x_i)^+\lambda(c_a^2 - 1)\alpha_i\alpha_j.
\end{aligned}$$

6.16 Give an informal derivation of the expression (6.19) for the auto-covariance function of the stationary diffusion limit of G/G/∞, using the results of Exercise 6.15.

Solution

For the variance:

$$\begin{aligned}
\mathbb{V}(\hat{\boldsymbol{Q}}(t)) &= \lambda \sum_{i=1}^{K} x_i \alpha_i + \lambda(c_a^2 - 1) \left[\sum_{i=1}^{K-1} \alpha_i x_i \sum_{j=i+1}^{K} \alpha_j + \sum_{j=1}^{K} \alpha_j x_j \sum_{i=j}^{K} \alpha_i \right] \\
&= \lambda \int_0^\infty x dG(x) + \lambda(c_a^2 - 1) \left[\int_0^\infty x(1 - G(x)) dG(x) \right. \\
&\qquad \left. + \int_0^\infty x(1 - G(x-)) dG(x) \right] \\
&= \lambda \frac{1}{\mu} + \lambda(c_a^2 - 1) \int_0^\infty (1 - G(x))^2 dx.
\end{aligned}$$

For the covariance:

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\big(\hat{\boldsymbol{Q}}(t), \hat{\boldsymbol{Q}}(t+s)\big) &= \sum_{i=1}^{K} (x_i - s)^+ \lambda [\alpha_i + \alpha_i^2 (c_a^2 - 1)] \\
&\quad + \sum_{i \neq j} \min(x_j - s, x_i)^+ \lambda (c_a^2 - 1) \alpha_i \alpha_j. \\
&= \lambda \int_s^\infty (x - s) dG(x) + 2\lambda(c_a^2 - 1) \sum_{i=1}^{K} x_i \alpha_i (1 - G(x_i + s)) dx \\
&= \lambda \int_0^\infty x dG(x + s) + 2\lambda(c_a^2 - 1) \int_0^\infty x(1 - G(x + s)) dG(x) \\
&= \lambda \int_0^\infty (1 - G(x + s)) dx + \lambda(c_a^2 - 1) \int_0^\infty (1 - G(x + s))^2 dx.
\end{aligned}$$

7

Diffusions and Brownian processes

Exercises

7.1 Let $X(t)$ be a standard Brownian motion. Let $t = t_0 < t_1 < \cdots < t_n = t + \tau$ and let $\epsilon = \min_{k=1,\ldots,n}(t_k - t_{k-1})$. Show that for all t and τ:

$$\lim_{\epsilon \to 0} \mathbb{E}\left(\sum_{k=1}^{n}(X(t_k) - X(t_{k-1}))^2 - \tau\right)^2 = 0.$$

In words, the quadratic variation of $BM(t)$ convergence in mean square to t [Breiman (1992), Section 12.8].

Solution

We write

$$\sum_{k=1}^{n}(X(t_k) - X(t_{k-1}))^2 - \tau = \sum_{k=1}^{n}\Big((X(t_k) - X(t_{k-1}))^2 - (t_k - t_{k-1})\Big)$$
$$= \sum_{k=1}^{n}\Big((X(t_k) - X(t_{k-1}))^2 - \mathbb{E}[(X(t_k) - X(t_{k-1}))^2]\Big)$$

The summands above are independent, and have mean 0. Therefore:

$$\mathbb{E}\left(\sum_{k=1}^{n}\Big((X(t_k) - X(t_{k-1}))^2 - (t_k - t_{k-1})\Big)\right)^2 = \sum_{k=1}^{n}\mathbb{E}\left[\Big((X(t_k) - X(t_{k-1}))^2 - (t_k - t_{k-1})\Big)^2\right]$$

We note that $(X(t_k) - X(t_{k-1}))^2/(t_k - t_{k-1})$ is distributed like Z^2 where $Z \sim N(0,1)$, and Z^2 has mean and variance equal to 1. We obtain:

$$\mathbb{E}\left(\sum_{k=1}^{n}(X(t_k) - X(t_{k-1}))^2 - \tau\right)^2 = \sum_{k=1}^{n}\mathbb{E}(Z^2 - 1)(t_k - t_{k-1})^2$$
$$\leq \sum_{k=1}^{n}\mathbb{E}(Z^2 - 1)(t_k - t_{k-1})\epsilon = \tau\epsilon,$$

Which converges to 0 as $\epsilon \to 0$. (This solution quoted from Breiman (1992), Section 12.8)

7.2 Let $Z \sim N(0, \sigma^2)$ be a mean 0 normal random variable. Calculate the mean and variance of $|Z|$.

Solution

For the expectation:

$$\mathbb{E}(|Z|) = \frac{2}{\sqrt{2\pi}} \int_0^\infty x e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \left[-e^{-x^2/2} \right]_0^\infty = \sqrt{\frac{2}{\pi}}$$

For the second moment:

$$\mathbb{E}(|Z|^2) = \mathbb{E}(Z^2) = \mathbb{V}(Z) = 1$$

so:

$$\mathbb{V}(|Z|) = 1 - \frac{2}{\pi}$$

7.3 Consider a sequence of M/M/1 birth and death queues, with arrival rates λ_n and service rate μ_n, where $\lambda_n \to \lambda$ and $\sqrt{n}(\mu_n - \lambda_n) \to \theta$. let $Q^n(t)$ be its queue length, and let $\hat{Q}^n(t) = \frac{Q^n(nt)}{\sqrt{n}}$. Write the decomposition into netput and regulator, and consider the netput process as a birth and death process. Then use Stone's theorem to show that $\hat{Q}_n(t)$ converges to a reflected Brownian motion.

Solution

We are interested in:

$$\begin{aligned} Q^n(t) = (\lambda_n - \mu_n)t + (\mathcal{A}_n(t) - \lambda_n t) - (\mathcal{S}_n(\mathcal{T}_n(t)) - \mu_n \mathcal{T}_n(t)) \\ +\mu_n(t - \mathcal{T}_n(t)) \quad = \quad \mathcal{X}^n(t) + \mathcal{Y}^n(t). \end{aligned}$$

we will consider instead of $\hat{Q}^n(t) = \sqrt{n}\frac{Q^n(nt)}{n}$, the process $\hat{\mathcal{X}}^n(t) = \sqrt{n}\frac{\mathcal{X}^n(nt)}{n}$, as $n \to \infty$.

$$\hat{\mathcal{X}}^n(t) = \frac{Q^n(0)}{\sqrt{n}} + \frac{\mathcal{A}_n(nt) - \mathcal{S}_n(n\bar{\mathcal{T}}_n(t))}{\sqrt{n}}$$

We have $\bar{\mathcal{T}}_n(t) \to t$, so by time change lemma we can replace $\bar{\mathcal{T}}(t)$ by t to obtain the limit. Let

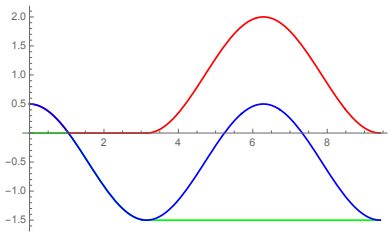
$$\hat{\mathcal{Z}}^n(t) = \frac{\mathcal{A}_n(nt) - \mathcal{S}_n(nt)}{\sqrt{n}}$$

This is a birth and death process, with jumps of $\pm\frac{1}{\sqrt{n}}$ that happen at rate $n\lambda_n$ up and $n\mu_n$ down, so

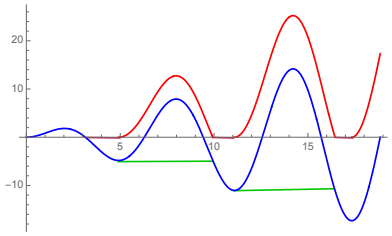
$$\begin{aligned} m_n(x) &= n\lambda_n \frac{1}{\sqrt{n}} - n\mu_n \frac{1}{\sqrt{n}} \\ &= \sqrt{n}(\lambda_n - \mu_n) \to \theta, \\ \sigma_n^2(x) &= n\lambda_n \frac{1}{n} + n\mu_n \frac{1}{n} \\ &= \lambda_n + \mu_n \to 2\lambda. \end{aligned}$$

so, $\hat{\mathcal{Z}}(t) \to \theta t + \sqrt{2\lambda}BM(t)$, and so $\hat{\mathcal{X}}(t) \to \hat{Q}(0) + \theta t + \sqrt{2\lambda}BM(t)$, and $\hat{Q}^n(t)$ goes to the corresponding RBM.

7.4 Obtain the reflection mapping (solution of Skorohod reflection problem) for the following two functions (you can give a formula for $\phi(x), \psi(x)$ or make a drawing).
(a) $x(t) = -0.5 + \cos(t),\ t \geq 0$.
(b) $x(t) = t\ sin(t),\ t \geq 0$.
Solution
(a)



(b)



7.5 Show existence, uniqueness, and minimality of the two sided regulators, and verify the recursive equation (7.17).
Solution
We show existence of solutions to (7.17):

$$L(t) = \sup_{0<s\leq t}(x(s) - U(s))^-, \qquad U(t) = \sup_{0<s\leq t}(b - x(s) - L(s))^-,$$

For given continuous $x(t)$, with $0 \leq x(0) \leq b$, we define a sequence of L^n and U^n upper and lower regulators as follows:

$$\begin{aligned} &L^0(t) = 0, \qquad U^0(t) = 0, \\ &L^{n+1}(t) = \phi(x - U^n)(t) = \sup_{0\leq s\leq t}(x(s) - U^n(s))^-, \\ &U^{n+1}(t) = \phi(b - x - L^n)(t) = \sup_{0\leq s\leq t}(b - x(s) - L^n(s))^-. \end{aligned}$$

Clearly, $L^n(t)$ and $U^n(t)$ increase with n, so they converge to a limit at every t. Furthermore, there exist $T_1 < T_2 < \ldots$ such that for $T_n < t < T_{n+1}$,

$L(t) = L^n(t)$ and $U(t) = U^n(t)$, where in the intervals between T_n $x(t)$ crosses from lower to upper boundary, or from upper to lower boundary. furthermore, $T_n \to \infty$, since otherwise, if $T_n \to T < \infty$, T will be a discontinuity point of x.

It is easy to see that these solutions indeed define the regulators.

7.6 Find the long time average $\lim_{T\to\infty} \frac{1}{T}\int_0^T (t)dt$ for the one sided regulated Brownian motion (reflected Brownian motion).

Solution

For $BM_0(t; m, \sigma^2)$ with $m < 0$, we should expect

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T (t)dt = \begin{cases} -m & m < 0 \\ 0 & m \geq 0 \end{cases}$$

to counter the negative drift. We can obtain this by looking at the two sided regulator, and letting the upper bound go to infinity, i.e. from equation (7.19) by letting $b \to \infty$.

7.7 Calculate the expectation of the stationary two side regulated Brownian motion $\mathcal{Z}(t)$.

Solution

Let $\theta = 2m/\sigma^2$, where m is the drift and σ^2 the diffusion coefficient of the Brownian motion.

For $m = 0$, $\mathcal{Z}(t) \sim \text{Uniform}(0, b)$ with expectation $b/2$.

For $m \neq 0$

$$\mathbb{E}(\mathcal{Z}(t)) = \int_0^b \frac{y\theta e^{\theta y}}{e^{\theta b} - 1} dy = b - \frac{b}{e^{\theta b} - 1} - \frac{1}{\theta}$$

7.8 The equations (7.24), (7.25) give the distributions of the regulator (t) and the reflected Brownian motion $\mathcal{Z}(t)$, with drift $m < 0$ and diffusion coefficient σ^2, starting at $\mathcal{X}(t) = \mathcal{Z}(t) = 0$. Obtain the distributions of (t) and $\mathcal{Z}(t)$ for $\mathcal{Z}(0) = x_0$.

Solution

Let $\mathcal{X}(t)$ be $BM_{x_0}(t; m, \sigma^2)$. Let T be the first time that $\mathcal{X}(t)$ hits 0. Conditioned on T, the process $\mathcal{X}(t)$, $0 \leq t \leq T$ behaves like a Brownian bridge, and $\mathcal{Z}(t) = \mathcal{X}(t)$, $(t) = 0$. Thereafter, for $t > T$, by the strong Markov property it will behave as $RBM_0(\cdot; m, \sigma^2)$ and its regulator. So, conditional on T:

$$\mathcal{Z}(t \mid T) \sim \begin{cases} \sigma\mathcal{B}^0(\frac{t}{T}) + \frac{T-t}{T}x_0, & 0 \leq t \leq T, \\ RBM_0(t - T; m, \sigma^2), & t \geq T. \end{cases}$$

where $\mathcal{B}^0(\cdot)$ is a standard Brownian bridge, and we have (see Section 18.1, Exercise 18.1)

$$\sigma\mathcal{B}^0(\frac{t}{T}) + \frac{T-t}{T}x_0 \sim N\big(\frac{T-t}{T}x_0\,,\ \frac{t}{T}(1 - \frac{t}{T})\sigma^2\big)$$

The probability density of time to hit 0 starting from x_0 is the same as hitting

x_0 starting from 0, with drift $-m$, so by equation (7.23):

$$f_T(t)dt = \mathbb{P}[T \in (t, t+dt)] = \frac{x_0}{\sigma\sqrt{2\pi t^3}} \exp\Big(-\frac{(x_0+mt)^2}{2\sigma^2 t}\Big)dt$$

and we have, by (7.25),

$$\begin{aligned}
\mathbb{P}(\mathcal{Z}(t) \le y) &= \int_0^t \mathbb{P}\Big(RBM_0(t-T; m, \sigma^2) \le y\Big) f_T(T)dT \\
&\quad + \int_t^\infty \mathbb{P}\Big(\sigma\mathcal{B}^0(\frac{t}{T}) + \frac{T-t}{T}x_0 \le y\Big) f_T(T)dT \\
&= \int_0^t \Phi\Big(\frac{y - m(t-T)}{\sigma(t-T)^{1/2}}\Big) - e^{2my/\sigma^2}\Phi\Big(\frac{-y - m(t-T)}{\sigma(t-T)^{1/2}}\Big) f_T(T)dT \\
&\quad + \int_t^\infty \Phi\Big(\frac{y - \frac{T-t}{T}x_0}{\big(\frac{t}{T}(1-\frac{t}{T})\big)^{1/2}\sigma}\Big) f_T(T)dT.
\end{aligned}$$

The distribution of (t) is then, by (7.24),

$$\mathbb{P}((t) \le y) = \int_0^t \Phi\Big(\frac{y + m(t-T)}{\sigma(t-T)^{1/2}}\Big) - e^{-2my/\sigma^2}\Phi\Big(\frac{-y + m(t-T)}{\sigma(t-T)^{1/2}}\Big) f_T(T)dT.$$

Note, when $m > 0$, there is a positive probability that the process will never hit 0, so $f_T(\cdot)$ does not integrate to 1. The above expressions are still correct.

7.9 Provide a mathematical proof that the optimal control of stationary Manufacturing with stationary independent increments demand is to use an upper inventory bound, and produce at maximal rate anywhere below that bound.

Solution

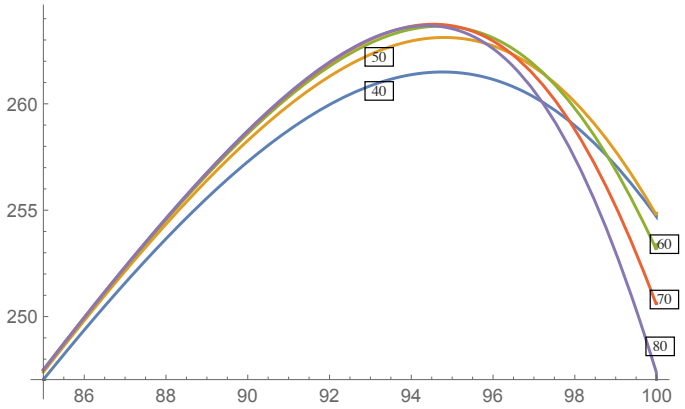
We can formulate a discrete space discrete time Markov decision problem that approximates this problem. If we bound service capacity then it is a problem with finite action space, and has a deterministic optimal solution. it can then be shown that maximal production level and unique upper inventory solve this discrete Markov decision problem.

7.10 Here is some data for a manufacturing system. Average demand rate is $a = 100$, with standard deviation $\sigma = 15$. The rest of the data is: sales price $r = 14$, material cost $c = 5$, workforce cost $w = 6$, inventory holding cost $h = 1$. Determine the optimal workforce k, and the optimal upper inventory level bound b, and calculate the long term average profit V.

How would the solution change if you vary any one of r, c w, h (say in what direction would V, k and b move, and if you can obtain the rates).

Solution

The profit as function of workforce (85-100) and inventory (40-80) is plotted:



Optimal is workforce of 94 or 95, and inventory of 70.

7.11 (∗) The calculations in Section 7.7.5 were for optimal control of the stationary manufacturing system. However, often one wants to take into account the current initial state of the system. In that case it is more reasonable to optimize the discounted profit, with some discount rate λ. For the same policy of upper bound inventory and full production, calculate the discounted infinite horizon profit for initial state z_0, and given k, b, λ [Harrison (1985), Chapter 5, or Harrison (2013), Chapter 6].

Solution(See Harrison (1985) sections 1.5, 3.2, 3.3, and 5.3)
We have $X(t) = BM_x(t; m, \sigma^2)$, and $Z(t) = X(t) + L(t) - U(t)$. We have b, γ, h, r, c. We wish to calculate, for $0 \le x \le b$:

$$k(x) = \mathbb{E}_x \left\{ \int_0^\infty e^{-\gamma t} \left[hZ(t)dt - rdL(t) + cdU(t) \right] \right\}.$$

We start by defining the following Wald martingale:

$$V_\beta(t) = e^{\beta X(t) - q(\beta)t}, \quad \text{where} \quad q(\beta) = m\beta + \tfrac{1}{2}\sigma^2\beta^2.$$

We also define the stopping time $T = T(0) \wedge T(b)$, the first time that $X(t)$ reaches one of the boundaries 0 or b. By the martingale optional stopping theorem,

$$\mathbb{E}(V_\beta(T)) = \mathbb{E}(V_\beta(0)) = e^{\beta x}.$$

We decompose $\mathbb{E}(V_\beta(T)) = \mathbb{E}(V_\beta(T); T{=}T(0)) + \mathbb{E}(V_\beta(T); T{=}T(b))$ where $\mathbb{E}(X; A) = \mathbb{E}(X|A)\mathbb{P}_x(A)$, to obtain:

$$\begin{aligned} e^{\beta x} &= \mathbb{E}(V_\beta(T)) = \mathbb{E}(V_\beta(T); T{=}T(0)) + \mathbb{E}(V_\beta(T); T{=}T(b)) \\ &= \mathbb{E}(e^{-q(\beta)T}; X(T) = 0) + e^{\beta b}\mathbb{E}(e^{-q(\beta)T}; X(T) = b), \end{aligned}$$

This will yield expressions for the expected discounted time up to T, for both

cases, $T = T(0)$, $T = T(b)$. we denote these by:

$$\psi_*(x;\gamma) = \mathbb{E}(e^{-\gamma T}; X(T) = 0), \qquad \psi^*(x;\gamma) = \mathbb{E}(e^{-\gamma T}; X(T) = b),$$
$$\mathbb{E}(e^{-\gamma T}) = \psi_*(x;\gamma) + \psi^*(x;\gamma).$$

To calculate the discounted times we use the equation:

$$e^{\beta x} = \psi_*(x; q(\beta)) + e^{\beta b}\psi^*(x; q(\beta)).$$

There are two values of β (one positive, one negative) that correspond to $\gamma = q(\beta) > 0$,

$$\beta = -\beta_*(\gamma) = \sigma^{-2}[(m^2+2\sigma^2\gamma)^{-1/2}+m] > 0, \quad \beta = \beta^*(\gamma) = \sigma^{-2}[(m^2+2\sigma^2\gamma)^{-1/2}-m] > 0,$$

and we get two equations:

$$e^{-\beta_*(\gamma)x} = \psi_*(x;\gamma)+e^{-b\beta_*(\gamma)}\psi^*(x;\gamma), \quad e^{\beta^*(\gamma)x} = \psi_*(x;\gamma)+e^{b\beta^*(\gamma)}\psi^*(x;\gamma),$$

from which we obtain:

$$\psi_*(x;\gamma) = \frac{e^{\beta_*(\gamma)(b-x)} - e^{-\beta^*(\gamma)(b-x)}}{e^{\beta_*(\gamma)b} - e^{-\beta^*(\gamma)b}}, \quad \psi^*(x;\gamma) = \frac{e^{\beta^*(\gamma)x} - e^{-\beta_*(\gamma)x}}{e^{\beta^*(\gamma)b} - e^{-\beta_*(\gamma)b}},$$

We note that both $\psi_*(x;\gamma)$ and $\psi^*(x;\gamma)$ have boundary values, and satisfy the differential equation:

$$\psi_*(0) = \psi^*(b) = 1, \qquad \psi_*(b) = \psi^*(0) = 0,$$
$$\gamma\psi_*(x) - \Gamma\psi_*(x) = \gamma\psi^*(x) - \Gamma\psi^*(x) = 0,$$

where the initial conditions follow from the definition, and where the operator $\Gamma = m\frac{d}{dx} + \frac{1}{2}\sigma^2\frac{d^2}{dx^2}$.
We have now prepared the background for the calculation of $k(x)$. We first discuss calculation of

$$h(x) = \mathbb{E}_x\Big(\int_0^\infty e^{-\gamma t}hZ(t)dt\Big).$$

It consists of the integral up to time T, where we can replace $\mathcal{Z}(t)$ by $X(t)$, and then of the discounted value of either $h(0)$ if 0 is reached or $h(b)$ if b is reached at T.

$$h(x) = \mathbb{E}_x\Big(\int_0^T e^{-\gamma t}hX(t)dt\Big) + \psi_*(x;\gamma)h(0) + \psi^*(x;\gamma)h(b).$$

We obtain first, by Fubini:

$$\mathbb{E}_x\Big(\int_0^\infty e^{-\gamma t}hX(t)dt\Big) = \int_0^\infty e^{-\gamma t}h\mathbb{E}_x(X(t))dt$$
$$= \int_0^\infty e^{-\gamma t}h(x+mt)dt = h\left(\frac{x}{\gamma} + \frac{m}{\gamma^2}\right),$$

and from this we get:

$$\begin{aligned} h(x) &= \mathbb{E}_x\Big(\int_0^T e^{-\gamma t} hX(t)dt\Big) \\ &= \mathbb{E}_x\Big(\int_0^\infty e^{-\gamma t} hX(t)dt\Big) - \mathbb{E}_x\Big(\int_T^\infty e^{-\gamma t} hX(t)dt\Big) \\ &= h\left[\frac{x}{\gamma} + \frac{m}{\gamma^2} - \psi_*(x)\frac{m}{\gamma^2} - \psi^*(x)(\frac{b}{\gamma} + \frac{m}{\gamma^2})\right] \\ &= h\left[\frac{m}{\gamma^2}(1 - \psi_*(x) - \psi^*(x)) + \frac{1}{\gamma}(x - \psi^*(x)b)\right]. \end{aligned}$$

We again notice that $h(x)$ satisfies boundary conditions and differential equation:

$$\begin{aligned} h(0) = h(b) = 0, \\ \gamma h(x) - \Gamma h(x) = hx, \end{aligned}$$

where the boundary conditions say that at $x = 0$ and at $x = b$ we have $T = 0$. (In fact, this can be generalized to costs $u(x)$ rather then constant cost rate hx, in which case the equation is $\gamma h(x) - \Gamma h(x) = u(x)$ with the same boundary conditions).

We now turn to the calculation of $k(x)$. It is the expectation of the following functional of Z, L, U: $K(Z, L, U) = \int_0^\infty e^{-\gamma t}\,[hZ(t)dt - rdL(t) + cdU(t)]$, We notice that up to time T, $L(T) = U(T) = 0$. By the strong Markov property,

$$k(x) = h(x) + \psi_*(x;\gamma)k(0) + \psi^*(x;\gamma)k(b).$$

We know that $h(x), \psi_*(x;\gamma), \psi^*(x;\gamma)$ all satisfy a differential equation of the form $\gamma f - \Gamma f = g$, and so we have:

$$\gamma k(x) - \Gamma k(x) = hx,$$

(or $= u(x)$ for cost $u(x)$ replacing hx) but now we need different boundary conditions, namely:

$$k'(0) = r, \quad k'(b) = c.$$

Those follow since close to 0 $k(x)$ will gain at rate r of lost sales, and close to b we will gain at rate c of less production costs. Since we already have explicit expressions for $h(x), \psi_*(x;\gamma), \psi^*(x;\gamma)$ all we need now is to calculate $k(0)$, $k(b)$ so as to satisfy the boundary conditions for $k'(0), k'(b)$.

7.12 The following is the Skorohod embedding problem that was discussed in Section 5.2.2: Let X be a random variable with $\mathbb{E}(X) = 0$, $\mathbb{V}\mathrm{ar}(X) < \infty$. Let $BM(t)$ be a standard Brownian motion. Find a stopping time T such that $BM(T) =_D X$ (equal in distribution) and $\mathbb{E}(T) = \mathbb{V}\mathrm{ar}(X)$.

The following exercises lead to an answer to this problem. This answer was

found by Dubins (1968), there are many other answers, including the original one by Skorohod, a survey of results related to this problem is Obłój (2004).

Solution

We construct an increasing sequence of stopping times τ_n such that $\tau = \sup_n \tau_n$ solves the problem.

7.13 Quote a theorem that shows: $\sup_{0\le s\le t} BM(t) \to \infty$ and $\inf_{0\le s\le t} BM(t) \to -\infty$ as $t \to \infty$, almost surely. This means that almost every path of a Brownian motion visits all of the values on the real line.

Solution

The law of the iterated logarithm says that:

$$\limsup_{t\to\infty} \frac{BM(t,\omega)}{\sqrt{t \log\log t}} = \sqrt{2} \text{ almost surely.}$$

7.14 Show that $BM(t)$ and $(BM(t)^2 - t)$ are martingales.

Solution

Let $\mathcal{F}(t)$ be the filtering to which $BM(\cdot)$ is adapted. In other words it is the σ-field generated by $BM(s), 0 \le s \le t$. By the property of independent increments, for $0 \le s < t$,

$$\mathbb{E}(BM(t) - BM(s) \,|\, \mathcal{F}(s)) = 0,$$

so:

$$\mathbb{E}(BM(t) \,|\, \mathcal{F}(s)) = BM(s),$$

and hence $BM(t)$ is a martingale.

Next by independent increments,

$$\mathbb{E}\big((BM(t) - BM(s))^2 \,|\, \mathcal{F}(s)\big) = \mathbb{V}\text{ar}(BM(t-s) = t - s,$$

and

$$\begin{aligned}
&\mathbb{E}\big((BM(t) - BM(s))^2 \,|\, \mathcal{F}(s)\big) \\
&= \mathbb{E}(BM(t)^2 \,|\, \mathcal{F}(s)) - 2\mathbb{E}(BM(t)BM(s) \,|\, \mathcal{F}(s)) + \mathbb{E}(BM(s))^2 \,|\, \mathcal{F}(s)) \\
&= \mathbb{E}(BM(t)^2 \,|\, \mathcal{F}(s)) - 2BM(s)\mathbb{E}(BM(t) \,|\, \mathcal{F}(s)) + BM(s)^2 \\
&= \mathbb{E}(BM(t)^2 \,|\, \mathcal{F}(s)) - BM(s)^2.
\end{aligned}$$

Hence:

$$\mathbb{E}(BM(t)^2 - t \,|\, \mathcal{F}(s)) = BM(s)^2 - s,$$

and hence $(BM(t)^2 - t)$ is a martingale.

7.15 Let Y be a random variable with distribution concentrated on two points, $a < 0 < b$ and mean zero. Find the distribution of Y and its variance.

Solution

$$\mathbb{E}(Y) = a\mathbb{P}(Y = a) + b\mathbb{P}(Y = b) = 0$$

$$\mathbb{P}(Y=a)=\frac{b}{b-a},\quad \mathbb{P}(Y=b)=\frac{-a}{b-a}$$

$$\mathbb{V}\mathrm{ar}(Y)=\mathbb{E}(Y^2)=a^2\frac{b}{b-a}+b^2\frac{-a}{b-a}=|a|b$$

7.16 Let $T_x=\inf\{t: BM(t)=x\}$. Let $T=\min(T_a,T_b)$. Then T solves the Skorohod embedding problem for the two point random variable Y of exercise 7.15. Use the martingale $BM(t)$ to prove that $BM(T)=_D Y$, and the martingale $(BM(t)^2-t)$ to calculate $E(T)$.

Solution

We quote the *martingale optional sopping theorem*

Theorem 7.1 (Doob (1953))**.** *Let X_n be a discrete time martingale, T a stopping time for X_n. Then $\mathbb{E}(X(T))=X(0)$ if one of the following holds:*

(i) T is almost surely bounded, i.e. there exists c such that $T<c$ a.s.

(ii) $\mathbb{E}(T)<\infty$ and $\mathbb{E}(|X_{n+1}-X_n|)<c$ for all $n<T$.

(iii) The truncated stopping times $T\wedge n$ are bounded: there exists c such that $X(T\wedge n)<c$ for all n almost surely.

Furthermore, the same holds for a continuous time martingale if it has continuous paths.

Recall that $BM(t)$ is a martingale. We note first that T is finite almost surely (by a exercise 7.13), and also that $a\le BM(T\wedge t)\le b$. By the *martingale optional stopping theorem* condition (iii), $\mathbb{E}(BM(T))=BM(0)=0$. So $BM(T)$ takes values a or b and has expectation 0. Hence $BM(T)=_D Y$.
Furthermore, since by exercise 7.14 $BM(t)^2-t$ is also a martingale, by the same argument, $\mathbb{E}\left[BM(T)^2-T\right]=BM(0)-0=0$, so $\mathbb{E}(T)=\mathbb{E}(BM(T)^2)=\mathbb{V}\mathrm{ar}(Y)=|a|b$.

7.17 Let X have zero mean and finite variance. Let $m^p=\mathbb{E}(X|0<X<\infty)$, $m^n=\mathbb{E}(X|-\infty<X\le 0)$. Define Y as the two point distribution on m^p and m^n, with mean zero. Show that $\mathbb{P}(Y=m^p)=\mathbb{P}(X>0)$, and $\mathbb{P}(Y=m^n)=\mathbb{P}(X\le 0)$.

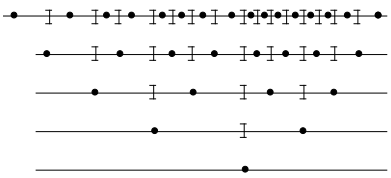
Solution

This follows immediately from:

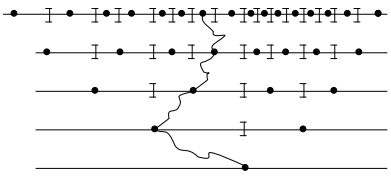
$$\begin{aligned}0=\mathbb{E}(X)&=P(X>0)\mathbb{E}(X|X>0)+\mathbb{P}(X\le 0)\mathbb{E}(X|X\le 0),\\&=\mathbb{P}(X>0)m^p+\mathbb{P}(X\le 0)m^n\\0=\mathbb{E}(Y)&=\mathbb{P}(Y=m^p)m^p+\mathbb{P}(Y=m^n)m^n.\end{aligned}$$

7.18 Define a sequence of stopping times $T^{(k)}$ as follows: Define $y_{0,1}=\mathbb{E}(X)=0$. Start with $y_{1,1}=m^p$, $y_{1,2}=m^n$, and let $T^{(1)}=T_{y_{1,1}}\wedge T_{y_{1,2}}$, the stopping time on $BM(t)$ that stops at $y_{1,1}$ or $y_{1,2}$. Then $\{y_{0,1},y_{1,1},y_{1,2}\}$ divide the real line into $4=2^2$ intervals, $I_{2,j}$, $j=1,\dots,2^2$. Let $y_{2,j}=\mathbb{E}(X\mid X\in I_{2,j})$. Define the stopping time $T^{(2)}=\min\{t: t>T^{(1)}, t=y_{2,j}\text{ for some } j\}$. Next, proceed inductively: given $T^{(k)}$ and the set of values $\{y_{i,j}: i=0,\dots,k,\ j=1,\dots,2^i\}$, these values divide the real line into 2^{k+1} intervals,

$I_{k+1,1}, \ldots, I_{k+1,2^{k+1}}$. Let $y_{k+1,j} = \mathbb{E}(X \mid X \in I_{k+1,j})\,, j = 1, \ldots, 2^{k+1}$, and define $T^{(k+1)} = \min\{t : t > T^{(k)}, t = y_{k+1,j} \text{ for some } j\}$.
Let Y^k take the values $y_{k,j}$ with probability $\mathbb{P}(X \in I_{k,j})$ Prove that:
(i) $BM(T^{(k)}) =_D Y^k$.
(ii) $\mathbb{E}(T^k) = \mathbb{V}\text{ar}(Y^k)$.
Hint: Use the strong Markov property of $BM(\cdot)$, and rules for calculating means and variances from conditional means and variances.
Solution
At this point it is useful to describe the sequences Y^k, $T^{(k)}$, $k = 0, 1, 2, \ldots$ in more detail. We note that Y^k are discretizations of the random variable X, with some very special properties: In the first step we put a probability of 1 at the expected value of X. At stage k we have 2^k intervals, and we take the point which is the conditional expectation of X given that X is in the interval, and put the probability of that interval at that point. Furthermore, we now have a point in each of the 2^k intervals, which divides it into 2 sub intervals, and we use the resulting 2^{k+1} intervals for the next stage. The next figure illustrates these stages:



To understand the stopping times, consider a single sample path of the Brownian motion. Starting at 0, the Brownian motion will at time $T^{(1)}$ reach either $y_{1,1}$ or $y_{1,2}$, say $y_{1,j}$, and be on one or the other side of 0, in the interval $I_{1,j}$. It will then, at time $T^{(2)}$, reach the point $y_{2,2j-1}$, or the point $y_{2,2j}$ which are in the interval $I_{1,j}$, on either one or the other side of $y_{1,j}$. Note that between $T^{(1)}$ and $T^{(2)}$ the Brownian motion does not leave the interval $I_{1,j}$. Once $B(T^{(k)}) = y_{k,j}$ is reached, in the interval $I_{k,j}$, at all future $T^{(\ell)}, \ell > k$ $B(T^{(\ell)})$ will be inside that interval. The next figure illustrates successive stages of the Brownian motion:



We now prove (i) and (ii) by induction. For $k = 1$ this was proved in exercises 7.16 and 7.17. Assume (i), (ii) hold for Y^k, $T^{(k)}$.

We condition on $BM(T^{(k)}) = y_{k,j} \in I_{k,j}$. By definition of $y_{\cdot,\cdot}$:

$$y_{k,j} = \mathbb{E}(X \mid X \in I_{k,j}), \quad y_{k+1,2j-1} = \mathbb{E}(X \mid X \in I_{k,j}, X \leq y_{k,j}),$$
$$y_{k+1,2j} = E(X \mid X \in I_{k,j}, X > y_{k,j}),$$

Let $BM^*(t) = BM(t + T^{(k)}) - BM(T^{(k)})$, by the strong Markov property it is a Brownian motion, starting at 0. Let $T^* = T^{(k+1)} - T^{(k)}$. It is a stopping time for $BM^*(t)$, it is independent of $T^{(k)}$, and it is the first time that $BM^*(t)$ hits one of the values $y_{k+1,2j} - y_{k,j}$ or $y_{k+1,2j-1} - y_{k,j}$. Using again exercises 7.16 and 7.17, we have:

$$\mathbb{P}(BM^*(T^*) = y_{k+1,2j} - y_{k,j}) = P(X > y_{k,j} \mid X \in I_{k,j}) = \mathbb{P}(X \in I_{k+1,2j} \mid X \in I_{k,j}).$$

To un-condition we then have:

$$\begin{aligned}
&\mathbb{P}(BM(T^{(k+1)}) = y_{k+1,2j}) = \mathbb{P}(BM(T^{(k+1)}) = y_{k+1,2j} \cap BM(T^{(k)}) = y_{k,j}) \\
&= \mathbb{P}(BM(T^{(k+1)}) = y_{k+1,2j} \mid BM(T^{(k)}) = y_{k,j}) P(BM(T^{(k)}) = y_{k,j}) \\
&= \mathbb{P}(X > y_{k,j} \mid X \in I_{k,j}) \mathbb{P}(X \in I_{k,j}) = \mathbb{P}(X \in I_{k+1,2j})
\end{aligned}$$

where the first equality follows since $y_{k+1,2j}$ is only reached by $BM(T^{(k+1)}$ if $BM(T^{(k)}) = y_{k,j}$, and the second equality follows by the induction hypothesis. Similar calculation hold for $y_{k+1,2j-1}$. This proves (i).
To prove (ii) note that by the above arguments about the definition of T^*, and exercises 7.16 and 7.17, we have, conditional on $BM(T^{(k)}) = y_{k,j}$:

$$\begin{aligned}
\mathbb{E}(T^* \mid BM(T^{(k)}) = y_{k,j}) &= (y_{k+1,2j} - y_{k,j})(y_{k,j} - y_{k+1,2j-1}) \\
&= \mathbb{V}\mathrm{ar}(Y^{k+1} - Y^k \mid Y^k = y_{k,j}),
\end{aligned}$$

We then have:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(Y^{k+1}) &= \mathbb{E}(\mathbb{V}\mathrm{ar}(Y^{k+1} \mid Y^k)) + \mathbb{V}\mathrm{ar}(\mathbb{E}(Y^{k+1} \mid Y^k)) \\
&= \mathbb{E}(\mathbb{V}\mathrm{ar}(Y^{k+1} - Y^k) \mid Y^k)) + \mathbb{V}\mathrm{ar}(Y^k) \\
&= \mathbb{E}(\mathbb{E}(T^{(k+1)} - T^{(k)} \mid BM(T^{(k)})) + \mathbb{E}(T^{(k)}) \\
&= \mathbb{E}(T^{(k+1)}).
\end{aligned}$$

The first equality is the well known formula for calculating variance, the second inequality uses that $Y^k | Y^k$ has variance 0 and that $\mathbb{E}(Y^{k+1} \mid Y^k) = Y^k$, the third equality was proved in the previous equation, and the last follows from $\mathbb{E}(\mathbb{E}(T^{(k+1)} - T^{(k)} \mid BM(T^{(k)})) = \mathbb{E}(T^{(k+1)} - T^{(k)})$ by the strong Markov propety and independent increments.

7.19 Show that $T = \lim_{k \to \infty} T^{(k)}$ is a stopping time. Show that it solves the Skorohod embedding problem.

Solution

We need to show three things: (i) that $Y^k \to_w X$, (ii) that Y^k are uniformly integrable, and therefore $E(Y^k)$, $\mathbb{V}\mathrm{ar}(Y^k)$ also converge, (iii) that $T^{(k)}$ converges to a stopping time T.

To prove all of these, we couple Y^k, $k = 0, 1, 2 \ldots$ and X, by generating X and taking the value of $Y^k = y_{k,j}$ if $X \in I_{k,j}$.
We now have:

(a) By the construction it is clear that $E(|X - Y^k|) = E(|Y^{k+1} - Y^k|)$.

(b) Let c be such that $\mathbb{P}(X > c) < \epsilon$. Choose (k, j) such that $y_{k,j} < c < y_{k,j+1}$ if that is possible. Then for all $\ell > k$,

$$\mathbb{P}(Y^\ell > y_{k,j+1}) = \mathbb{P}(X > y_{k,j+1}) \leq \mathbb{P}(X > c) < \epsilon,$$

On the other hand, if we cannot find any such (k, j) then $\mathbb{P}(Y^k > c) = 0$ for all k. This proves that Y^k are uniformly integrable.

(c) The Y^k generated coupled to X form a martingale which is uniformly integrable, so by the martingale convergence theorem it converges to some variable Y. But $\mathbb{E}(|X - y^k|) = \mathbb{E}(|Y^{k+1} - Y^k|) \to 0$ so $Y =_D X$.

(d) In particular, because Y^k are uniformly integrable, $\mathbb{V}\text{ar}(Y^k) \to \mathbb{V}\text{ar}(X)$

Next we consider $T^{(k)}$ and generate a coupled version of them, by generating a Brownian motion and the sequence of stopping times $T^{(k)}$ on this Brownian motion. Then:

(a) The T^k are non-decreasing, so they converge to some T

(b) T is a stopping time, since $\{T \leq t\} = \{T^{(k)} \leq t \text{ all } k\}$ is $\mathcal{F}(t)$ measurable.

(c) $T^{(k)} =_D Y^k$ and $Y^k \to_w X$ implies $T =_D X$

(d) $\mathbb{E}(T^k) = \mathbb{V}\text{ar}(Y^k)$ implies $\mathbb{E}(T) = \mathbb{V}\text{ar}(X)$.

Part III

Queueing Networks

8

Product Form Queueing Networks

Exercises

8.1 Show that the conclusions of the Perron Frobenius Theorem **??** continue to hold for A that is non-negative if there exists A such that A^m is positive.

Solution

A non-negative matrix can be approximated by a positive matrix (say $\tilde{a}_{i,j} = a_{i,j} + \epsilon$), for which all the conclusions hold, and the eigenvalues will converge to those of A, so A will have a real eigenvalue with maximal absolute value. Furthermore, A^m will have an isolated maximal real eigenvalue, r_1, and so all maximal eigenvalues of A must be of the (possibly complex) form ${r_1}^{1/m} exp(i\frac{k}{m})$. However, if A^m is positive so is also A^{m+1}, with maximal isolated real eigenvalue r_2, but then all the maximal eigenvalues of A must be of the (possibly complex) form ${r_2}^{1/(m+1)} exp(i\frac{k}{m+1})$. This can only be the case if there is a single real eigenvalue, corresponding to $k = 0$.

8.2 Show that if a transition matrix P satisfies P^m is positive then the chain is irreducible and a-periodic.

Solution

If P^m is positive, all states communicate, with paths of length m, so the chain is irreducible. Furthermore, P^{m+1} is also positive, and therefore $p^m_{i,i} > 0$, $p^{m+1}_{i,i} > 0$ and so the chain is a-periodic.

8.3 Show that a Markov chain is time reversible if and only if it satisfies the detailed balance equations.

Solution

If the stationary chain is time reversible then

$$\mathbb{P}(X(n) = i, X(n+1) = j) = \pi(i)p(i,j) = \mathbb{P}(X(n+1) = i, X(n) = j) = \pi(j)p(j,i).$$

If detailed balance holds than for the stationary process, for any sequence of

states $j_1, \ldots, j_k$,

$$\begin{aligned}
&\mathbb{P}(X(t) = j_1, \ldots, X(t+k) = j_k) = \pi(j_1)p(j_1, j_2)\cdots p(j_{k-1}, j_k)\\
&= p(j_2, j_1)\pi(j_2)p(j_2, j_3)\cdots p(j_{k-1}, j_k)\\
&\quad\vdots\\
&= p(j_2, j_1)p(j_3, j_2)\cdots \pi(j_k)p(j_k, j_{k-1})\\
&= \pi(j_k)p(j_k, j_{k-1})\cdots p(j_2, j_1) = \mathbb{P}(X(t+k) = j_1, \ldots, X(t) = j_k).
\end{aligned}$$

8.4 Prove the Kolmogorov criterion: a stationary Markov chain is time reversible if and only if for any finite sequence of states $j_1, j_2, \ldots, j_k$, the transition probabilities satisfy:

$$p(j_1, j_2)p(j_2, j_3)\cdots p(j_k, j_1) = p(j_1, j_k)p(j_k, j_{k-1}), \cdots p(j_2, j_1).$$

Solution
If $X(t)$ is time reversible then

$$\pi(j_1)p(j_1, j_2)p(j_2, j_3)\cdots p(j_k, j_1) = \pi(j_1)p(j_1, j_k)p(j_k, j_{k-1}), \cdots p(j_2, j_1).$$

and cancelling $\pi(j_1)$ we get Kolmogorov's criterion.
In the opposite direction, we show that detailed balance holds. Start from a reference state j_0, consider arbitrary state i, and a positive probability path $j_0, j_1, \ldots, j_k, i$. Define:

$$\pi(i) = \frac{p(j_0, j_1)\cdots p(j_k, i)}{p(i, j_k)\cdots p(j_1, j_0)}.$$

It may depend on j_0, but no on the path, since by Kolmogorov criterion, fo another path $j_0, j'_1, \ldots, j'_l, i$:

$$\frac{p(j_0, j_1)\cdots p(j_k, i)}{p(i, j_k)\cdots p(j_1, j_0)} = \frac{p(j_0, j'_1)\cdots p(j'_l, i)}{p(i, j'_l)\cdots p(j'_1, j_0)}.$$

Consider now partial balance between i and j: If $p(i, j) = 0$ there is nothing to prove. Else, we define:

$$\pi_j = \frac{p(j_0, j_1)\cdots p(j_k, i)p(i, j)}{p(j, i)p(i, j_k)\cdots p(j_1, j_0)}$$

to obtain $\pi(i)p(i, j) = \pi(j)p(j, i)$. The πs need to be normalized, but are then the unique stationary probabilities.

8.5 Show that every stationary birth and death process is time reversible.
Solution
For a continuous time birth and death process, we set up global balance

equations, for the states $0, 1, 2, \ldots$:

$$\lambda_0\pi_0 = \mu_1\pi_1,$$
$$(\lambda_1 + \mu_1)\pi_1 = \lambda_0\pi_0 + \mu_2\pi_2 \Longrightarrow \lambda_1\pi_1 = \mu_2\pi_2,$$
by using the previous equation and cancelling,
$$\vdots$$
$$(\lambda_n + \mu_n)\pi_n = \lambda_{n-1}\pi_{n-1} + \mu_{n+1}\pi_{n+1} \Longrightarrow \lambda_n\pi_n = \mu_{n+1}\pi_{n+1},$$
using induction and the same argument.

8.6 Write down the formula for the steady state distribution of the queue length vector for Jackson tandem queueing systems and feed forward systems (including solution of the traffic equations). Show that if node i precedes node j then $Q_i(t)$ is independent of $Q_j(s)$ for all $s < t$.

Solution

For the Jackson tandem queue, by Burke's theorem, arrivals to all nodes and departures form all nodes are Poisson at rate α, and if all $\mu_j < \alpha$ the steady state distribution is

$$\text{Tandem:} \qquad \pi(n_1, \ldots, n_I) = \prod_{i=1}^{I} \left(1 - \frac{\alpha}{\mu_i}\right)\left(\frac{\alpha}{\mu_i}\right)^{n_i}.$$

For the feed forward Jackson network, one obtains the flow rates through each node from:

$$\lambda = \left(I + P^{\intercal} + P^{\intercal 2} + \cdots + P^{\intercal I-1}\right)\alpha,$$

put differently, the rate λ_j is the sum over all paths that lead from input to j, say $i_1 \to i_2 \to i_k \to j$ of the product $\alpha_{i_1} p_{i_1,i_2} \cdots p_{i_k,j}$. The steady state distribution, is then

$$\text{Feed Forward:} \qquad \pi(n_1, \ldots, n_I) = \prod_{i=1}^{I} \left(1 - \frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_i}{\mu_i}\right)^{n_i}$$

We now show the independence result, by induction on the number of nodes. There is nothing to show for a single node. Consider then a feedforward network with I nodes, including a *source* node labeled 1, with input Poisson α_1. It has only flow from outside (by feed forward) and its output by Burke's theorem is Poisson, rate α_1. This output is then split into independent Poisson streams of rates $\alpha_1 p_{1,j}$, that flow into nodes $j = 2, \ldots, I$. The system excluding node 1 is a feed forward network of nodes $2, \ldots, I$, with independent Poisson inputs of rates $\alpha_j + \alpha_1 p_{1,j}$. By the induction hypothesis for all the nodes $j = 2, \ldots, I$, the statement of the theorem holds. So we just need to check that $Q_1(t)$ and $Q_j(s)$ are independent for $s < t$. This follows again by Burke's theorem: The state of the network of $2, \ldots, I$ at time s is determined

by the departure process of node 1, and by all the flows of customers that did not enter from node 1. But all of these are independent of $Q_1(t)$ if $s < t$.

8.7 Verify the stationary distribution for the Jackson network with processing rates $\mu_i(n_i)$ as given by Theorem 8.8

Solution

The partial balance equations are

$$\sum_{j\neq i} \pi(x)q(x, T_{i,j}(x)) = \sum_{j\neq i} \pi(T_{i,j}(x))q(T_{i,j}(x), x), \text{ for } x \in S,\ 0 \leq i \leq I,$$

with transition rates:

$$\begin{array}{ll} q(x, T_{i,j}(x)) = \alpha_j, & i = 0, \\ q(x, T_{i,j}(x)) = \mu_i(x_i)p_{i,j}, & i \neq 0,\ x_i > 0,\ j \neq 0, \\ q(x, T_{i,j}(x)) = \mu_i(x_i)\Big(1 - \sum_{k\neq i,0} p_{i,k}\Big), & i \neq 0,\ x_i > 0,\ j = 0, \\ q(x, T_{i,j}(x)) = 0 & i \neq 0,\ x_i = 0. \end{array}$$

We need to verify that

$$\pi(n_1, \ldots, n_I) = B \prod_{i=1}^{I} \frac{{\lambda_i}^{n_i}}{\prod_{m=1}^{n_i} \mu_i(m)},$$

satisfy the balance equations.

For $i \neq 0$ and $n_i > 0$,

$$\pi(x)\mu_i(n_i) = \pi(x)\left[\frac{\mu_i(n_i)}{\lambda_i}\alpha_i + \sum_{j\neq i,0} \frac{\mu_i(n_i)}{\lambda_i} \frac{\lambda_j}{\mu_j(n_j+1)} \mu_j(n_j+1)p_{j,i}\right],$$

which, after canceling and multiplying by λ_i yields the traffic equation for node i:

$$\lambda_i = \alpha_i + \sum_{j\neq i,0} \lambda_j p_{j,i}.$$

For node 0, arrivals to the system out of node 0 need to balance with departures, and on substituting:

$$\pi(x) \sum_{j\neq 0} \alpha_j = \pi(x)\left[\sum_{j\neq 0} \frac{\lambda_j}{\mu_j(n_j+1)} \mu_j(n_j+1)\Big(1 - \sum_{k\neq 0,j} p_{j,k}\Big)\right],$$

which is simply:

$$\mathbf{1}^{\mathsf{T}}(I - P^{\mathsf{T}})\lambda = \mathbf{1}^{\mathsf{T}}\alpha,$$

i.e. summation of the traffic equations.

For the normalizing constant, $1/B$ is the sum of all the product form terms, and the network is stable if and only if the sum is finite, i.e. $B > 0$.

8.8 Prove Kelly's Lemma 8.8.

Solution

We are given a stationary continuous time Markov chain $X(t)$ with transition rates $q(j,k)$, and with $q(j) = \sum_{k\neq j} q(j,k)$. We assume we have values $q'(j,k)$ with $q'(j) = \sum_{k\neq j} q'(j,k) = q(j)$, and a vector of probabilities $\pi(j)$ such that: $\pi(j)q(j,k) = \pi(k)q'(k,j)$. We claim that $q'(j,k)$ are the transition rates of the reversed process, and $\pi(j)$ the stationary probabilities of the process and of the reversed process. For state j we now write:

$$\sum_{k\neq j} \pi_k q(k,j) = \sum_{k\neq j} \pi_j q'(j,k) = \pi_j q(j),$$

so $\pi(j)$ solves the global balance equations, and is the stationary distribution. Furthermore,

$$\begin{aligned}
\mathbb{P}(X(t) = k \,\big|\, X(t+h) = j) &= \frac{\mathbb{P}(X(t) = k, X(t+h) = j)}{\mathbb{P}(X(t+h) = j)} \\
&= \frac{\mathbb{P}(X(t+h) = j \,\big|\, X(t) = k)\mathbb{P}(X(t) = k)}{\mathbb{P}(X(t+h) = j)} \\
&= \frac{\pi(k)}{\pi(j)} q(k,j)h + o(h) = q'(j,k)h + o(h).
\end{aligned}$$

so $q'(j,k)$ are the transition rates of the reversed process.

8.9 Prove that a Jackson network considered in reversed time is again a Jackson network, and calculate its parameters.

Solution

For any stationary Markov chain $\mathcal{X}(t)$ with transitions $q(x,y)$ and stationary distribution $\pi(x)$, the reversed process $\mathcal{X}(-t)$ is a stationary Markov chain with the same stationary distribution and transitions $q^*(x,y)$ given by:

$$q^*(x,y) = \frac{1}{\pi(x)} \pi(y) q(y,x).$$

Clearly, in the reversed Jackson network transitions are again of single customers moving between nodes (including node 0). We then have, for $n = (n_1, \ldots, n_I)$ and $1 \leq i, j \leq I$:

$$\begin{aligned}
q^*(n, T_{i,j}(n)) &= \frac{\pi(T_{i,j}(n))}{\pi(n)} q(T_{i,j}(n), n) \\
&= \frac{\mu_i(n_i)}{\lambda_i} \frac{\lambda_j}{\mu_j(n_j+1)} \mu_j(n_j+1) p_{j,i} \\
&= \mu_i(n_i) p^*_{i,j}
\end{aligned}$$

where the new routing probabilities are $p^*_{i,j} = \frac{\lambda_j}{\lambda_i} p_{j,i}$

For transitions of arrivals we have:

$$\begin{aligned} q^*(n, T_{0,j}(n)) &= \frac{\pi(T_{0,j}(n))}{\pi(n)} q(T_{0,j}(n), n) \\ &= \frac{\lambda_j}{\mu_j(n_j+1)} \mu_j(n_j+1)(1 - \sum_{i=1}^{I} p_{j,i}) \\ &= \alpha_j^* \end{aligned}$$

where the reversed arrival rate to node j is $\alpha_j^* = \lambda_j(1 - \sum_{i=1}^{I} p_{j,i})$
Finally, for transitions of departures we have:

$$\begin{aligned} q^*(n, T_{i,0}(n)) &= \frac{\pi(T_{i,0}(n)}{\pi(n)} q(T_{i,0}(n), n) \\ &= \frac{\mu_i(n_i)}{\lambda_i} \alpha_i \end{aligned}$$

where the reversed departure rate from node i is $\frac{\alpha_i}{\lambda_i}$
We now check first that the $p_{i,j}^*$ add up to 1 for all $i \neq 0$:

$$\sum_{j \neq i,0} p_{i,j}^* + p_{i,0}^* = \sum_{j \neq i,0} \frac{\lambda_j}{\lambda_i} p_{j,i} + \frac{\alpha_i}{\lambda_i} = 1$$

by traffic equation $\lambda_i = \alpha_i + \sum_{j \neq i,0} \lambda_j p_{j,i}$.
We next check that α_j^* and α_j add up to the same sum:

$$\mathbf{1}^\intercal \alpha^* = \mathbf{1}^\intercal (1 - P^\intercal)\lambda = 1^\intercal \alpha$$

by the traffic equations.

8.10 The M/M/1 queue with feedback is a single queue with Poisson rate α arrivals and exponential rate μ service times, where upon completion of service a customer rejoins the queue with probability θ, and leaves the system with probability $1 - \theta$. Calculate the stationary distribution of the queue $\boldsymbol{Q}(t)$ and show it is the same as M/M/1 with $\rho = \frac{\alpha}{\mu(1-\theta)}$, and that customers leave the system as a Poisson process of rate α. However, show that the stream of customers entering service (new arrivals and returns) is not Poisson.

Solution

The simplest way to look at this system is to assume instead of FCFS that each customer that feeds back is served immediately. Under this policy the departure process has the same distribution as under any non-predictive non-preemptive policy, including FCFS, so the distribution of queue length is the same as for FCFS. But under this policy service time of each customer will be Exp($\mu(1 - \theta)$, and the system will behave as an M/M/1 system with service rate $\mu(1 - \theta)$.

Alternatively, consider it as a Jackson network, with one node, arrivals at rate α, routing probability $p_{1,1} = \theta < 1$. Then flow through the node is at

rate $\alpha/(1-\theta)$ and the stationary distribution is

$$\pi(n) = \left(1 - \frac{\alpha}{\mu(1-\theta)}\right)\left(\frac{\alpha}{\mu(1-\theta)}\right)^n.$$

In fact the process is time reversible:

$$\pi(n)q(n,n+1) = (1-\rho)\left(\frac{\alpha}{\mu(1-\theta)}\right)^n \times \alpha,$$

$$\pi(n+1)q(n+1,n) = (1-\rho)\left(\frac{\alpha}{\mu(1-\theta)}\right)^{n+1} \times \mu(1-\theta),$$

and these are equal, so detailed balance holds, which implies reversibility. It follows that the departure process from the system is Poisson.

Consider now a system with small α, large μ and $\theta = 4/5$. Then ρ is small, and most of the time the system is empty. So when there is an outside arrival, that customer is usually served immediately, and that customer will then return on average 5 times, in short intervals of average length $1/\mu$. Clearly, the sequence of arrivals to the server is not Poisson.

8.11 Consider the M/M/1 queue with feedback as a Kelly-type multi-class network, where customers on their kth visit are class k customers. Obtain the stationary distribution of this system.

Solution

In this Kelly-type network there is a single node, we have classes $k = 1, 2, \ldots$, with $\lambda_1 = \alpha$ external arrivals, and $p_{k,k+1} = \theta$ for the routing. To avoid dealing with infinite dimensional states we choose a cutoff N and make class N all the customers that visit N or more times, so we have classes $k = 1, \ldots, N$. Then the flow rates of customers are $\lambda_k = \alpha\theta^{k-1}$, $k = 1, \ldots, N-1$, and $\lambda_N = \alpha\theta^{N-1}/(1-\theta)$. The total flow through the node is $\bar{\lambda} = \alpha/(1-\rho)$ and processing is always at rate μ, so the normalizing constant b is

$$b^{-1} = \sum_{n=0}^{\infty}\left(\frac{\alpha/(1-\theta)}{\mu}\right)^n = \left(1 - \frac{\alpha}{\mu(1-\theta)}\right)^{-1}$$

Let the state of the node be $(k_1, \ldots, k_n)$ when there are n customers at the node, and the customer at position l is type k_l. Then the stationary probability of observing this state is:

$$P(k_1, \ldots, k_n) = b\prod_{l=1}^{n}\frac{\lambda_{k_l}}{\mu}$$

From this we obtain that the joint distribution of the numbers of customers of each type. Let x_j be the number of customers of type j, then

$$P(x_1, \ldots, x_N) = \left(1 - \frac{\alpha}{\mu(1-\theta)}\right)\frac{(x_1 + \cdots + x_N)!}{x_1! \cdots x_N!}\left(\frac{\alpha\theta^{N-1}}{\mu(1-\theta)}\right)^{x_N}\prod_{j=1}^{N-1}\left(\frac{\alpha\theta^{j-1}}{\mu}\right)^{x_j}.$$

8.12 Prove Theorem 8.14, on the stationary distribution of closed Jackson networks, by verifying partial balance.

Solution

The partial balance equations are

$$\sum_{j\neq i}\pi(x)q(x,T_{i,j}(x)) = \sum_{j\neq i}\pi(T_{i,j}(x))q(T_{i,j}(x),x), \text{ for } x\in S,\ 1\leq i\leq I,$$

with transition rates:

$$\begin{aligned} q(x,T_{i,j}(x)) &= \mu_i(x_i)p_{i,j}, \quad && x_i>0,\\ q(x,T_{i,j}(x)) &= 0 && x_i=0. \end{aligned}$$

We need to verify that

$$\pi(n_1,\ldots,n_I) = B(N)\prod_{i=1}^{I}\frac{{\lambda_i}^{n_i}}{\prod_{m=1}^{n_i}\mu_i(m)},$$

satisfy the balance equations.

For $n_i>0$,

$$\pi(x)\mu_i(n_i) = \pi(x)\left[\sum_{j\neq i}\frac{\mu_i(n_i)}{\lambda_i}\frac{\lambda_j}{\mu_j(n_j+1)}\mu_j(n_j+1)p_{j,i}\right],$$

which, after canceling and multiplying by λ_i yields the traffic equation for node i:

$$\lambda_i = \sum_{j\neq i,0}\lambda_j p_{j,i}.$$

which are exactly the traffic equations defining λ.

8.13 Prove the arrival theorem, customers in transit see a stationary state, for both open and closed Jackson networks [Kelly (1979); Sevcik and Mitrani (1981)].

Solution

Consider a customer transiting from queue i to queue j. Insert between these queues a virtual queue 0 with a very high service rate μ_0. In the limit $\mu_0\to\infty$, the added queue does not affect the system at all: the customers transiting from queue i to queue j spend an infinitesimal time in the added virtual queue. The virtual queue, however, enables 'catching the customer in transition. The transition occurs precisely in the short interval when there is a customer in queue 0, i.e. when $n_0(t)=1$. The state distribution seen by the transiting customer is the distribution of the other queues conditioned on $n_0=1$. The new system is itself a Jackson network. The distribution of the rest of the network when the customer is in transition is then given by:

$$\mathbb{P}(n_1,\ldots,n_I|n_0=1) = \frac{\pi(n_0=1,n_1,\ldots,n_I)}{\pi(n_0=1)} = \pi(n_1,\ldots,n_I),$$

as seen immediately by the product form nature of the stationary distribution.

This proof works both for the open and the closed Jackson network, and also for Kelly networks.

8.14 Verify the expressions for the stationary distribution of Kelly networks and Kelly-type multi-class networks as given by (8.14), (8.15) and (8.16), (8.17).

Solution

We give the proof for Kelly-type multi-class networks, Kelly networks with deterministic routes are a special case.

The proof is by Kelly's Lemma 8.11. Let $C(i) = (c(i,1), \ldots, c(i,n_i))$ be the list of customer alsses in positions $r = 1, \ldots, n_i$ at node i, and the state be $C = (C(1) \ldots, C(i))$. Transitions are arrivals of class k to position r at node $i = s(k)$, transitions from class k in position r at node $i = s(k)$ to class l in position h and node $j = s(l)$ and departures from class k in position r at node $i = s(k)$, with rates:

$$\begin{array}{ll} T_{0,k,r}(C) = \alpha_k \delta_{r,n_i+1} & s(k) = i \\ T_{k,r,l,h}(C) = \mu_i(n_i)\gamma_{r,n_i} p_{k,l} \delta_{h,n_j+1} & s(k) = i, s(l) = j \\ T_{k,r,0}(C) = \mu_i(n_i)\gamma_{r,n_i} q_k & q_k = 1 - \sum_{l \neq k} p_{k,l},\ s(k) = i. \end{array}$$

Consider now the time reversed process. We use the same arguments as for Exercise 8.9. The reversed process is again a Kelly-type multi-class network, with the following parameters: The flow rates λ_k are the same. The arrival rates are $\alpha_k^* = \lambda_k q_k$, the routing probabilities are $p*_{k,l} = \frac{\lambda_l}{\lambda_k} p_{l,k}$, and the departure rate is $q_k^* = \frac{\alpha_k}{\lambda_k}$. The policy at node i is given by $\gamma^*_{r,n_i} = \delta_{r,n_i}$, $\delta^*_{r,n_i+1} = \gamma_{r,n_i+1}$.

By exactly the same argument as in 8.9, the $\sum \alpha_k^* = \sum \alpha_k$ and $\sum_l p^*_{k,l} + q^*_k = 1$, so the reversed process is indeed a Kelly-type multi-class network. Hence, we actually know all the reversed transition rates.

It remains to check that for each possible transition, the conjectured π satisfy: $\pi(C)q(C,C') = \pi(C')q^*(C',C)$:

– For arrivals, check that: $\pi(C) \times q(C, T_{0,k,r}(C)) = \pi(T_{0,k,r}(C)) \times q^*(T_{0,k,r}(C), C)$:

$$\pi(C) \times \alpha_k \delta_{r,n_i+1} = \pi(C) \frac{\lambda_k}{\mu_i(n_i+1)} \times \mu_i(n_i+1)\delta_{r,n_i+1} \frac{\alpha_k}{\lambda_k}.$$

– For routing, check that: $\pi(C) \times q(C, T_{k,r,l,h}(C)) = \pi(T_{k,r,l,h}(C)) \times q^*(T_{k,r,l,h}(C), C)$

$$\pi(C) \times \mu_i(n_i)\gamma_{r,n_i} p_{k,l} \delta_{h,n_j+1} = \pi(C) \frac{\mu_i(n_i)}{\lambda_k} \frac{\lambda_l}{\mu_j(n_j+1)} \times \mu_j(n_{j+1}) \delta_{h,n_j+1} \frac{\lambda_k}{\lambda_l} p_{k,l} \gamma_{r,n_i}.$$

– For departures, check that: $\pi(C) \times q(C, T_{k,r,0}(C)) = \pi(T_{k,r,0}(C)) \times q^*(T_{k,r,0}(C), C)$

$$\pi(C) \times \mu_i(n_i)\gamma_{r,n_i} q_k = \pi(C) \frac{\mu_i(n_i)}{\lambda_k} \times \lambda_k q_k \gamma_{r,n_i}.$$

8.15 (∗) Consider an M/M/K/K Erlang loss system, where service rates of the servers differ, say server k has service rate μ_k. Use the policy of assign to

longest idle server (ALIS), so arriving customers go to the server which has been idle for the longest time. Show that this system is insensitive [Adan and Weiss (2012)].

Solution

We do not give the solution here. It is the main result of the paper Adan and Weiss (2012)

8.16 (∗) The following is a model for a Jackson network with positive as well as negative customers. Nodes are $i = 1, \ldots, I$, service rates μ_i. Positive customers arrive at rate α_i^+ and are added to the queue at node i. Negative customers arrive at rate α_i^-, and on arrival they eliminate a customer from node i if not empty. On completion of service, positive customers move as positive customers with routing probabilities $p_{i,j}^+$, and move as negative customers with probabilities $p_{i,j}^-$, where on arrival at node j they eliminate a customer if not empty, and depart at rate $d_i = 1 - \sum_j (p_{i,j}^+ + p_{i,j}^-)$. Arrivals, processing and routings are independent and memoryless. The state of the system is given by the queue lengths of positive customers, $\boldsymbol{Q}(t) = (n_1, \ldots, n_I)$. Prove that $\boldsymbol{Q}(t)$ has product form stationary distribution $\prod_i \rho_i (1 - \rho_i)^{n_i}$, where $\rho_i = \frac{\lambda_i^+}{\mu_i + \lambda_i^-} < 1$, where λ_i^+, λ_i^-, $i = 1, \ldots, I$ solve the non-linear equations:

$$\lambda_i^+ = \alpha_i^+ + \sum_j \mu_j \rho_j p_{j,i}^+,$$

$$\lambda_i^- = \alpha_i^- + \sum_j \mu_j \rho_j p_{j,i}^-,$$

whenever these equations have a unique solution. Such systems are motivated as modeling networks of neurons: a positive signal arriving at a neuron increases its total signal count or potential by one; a negative signal reduces it by one if the potential is positive. When its potential is positive, a neuron"fires", sending positive or negative signals at random intervals to other neurons or to the outside. Positive signals represent excitatory signals and negative signals represent inhibition [Gelenbe (1991)].

Solution

We do not give the solution here. It is the main result in the paper of Gelenbe (1991).

9

Generalized Jackson Networks

Exercises

9.1 Prove by induction that the implicit conditions of the dynamics, the non-negativity, and work conservation (9.1), (9.2), uniquely determine the queue length process.

Solution

We note first that for any time t the number of arrivals and job completions is finite. We therefore can use induction to prove that the equations determine all the queue lengths up to time t. We proceed by induction on the events of arrivals and job completions. Starting at time 0, the queue length is constant $Q(t) = Q(0)$ by (9.1), all the non-empty nodes are processing jobs so $\mathcal{T}(t)$ for those nodes increases at rate 1, and idleness increases at empty nodes, by (9.2). We are waiting for the first event at time t_1, which can be an arrival or a job completion. If an arrival occurs at noe i, $\mathcal{A}_i(t)$ increases by one, and so does $Q_i(t)$ by (9.1). If a job completion occurs at node i, $\mathcal{S}_i(\mathcal{T}_i(t))$ increases by 1, and at the same time, we get a value for $\xi_i(\mathcal{S}_i(\mathcal{T}_i(t))$. As a result, a job leaves so $Q_i(t)$ decreases by 1, and if $\xi_{i,j}(\mathcal{S}_i(\mathcal{T}_i(t))) = 1$, then $\mathcal{R}_{i,j}(\mathcal{S}_i(\mathcal{T}_i(t))$ increases by 1, and $Q_j(t)$ increases by 1, all of this according to (9.1). We now wait for the next event at time t_2, where in (t_1, t_2) the evolution of $\mathcal{T}, \mathcal{I}$ is governed by (9.2), and then at t_2 the changes in $\mathcal{A}, \mathcal{S}, \mathcal{R}, Q$ are governed by (9.1), and we proceed by induction up to time t. This proves that (9.1) and (9.2) determine $Q(t)$, $t > 0$.

9.2 Consider the Jackson network with the following data:

$$\alpha = \begin{bmatrix} 60 \\ 12.5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0.2 & 0 & 0 & 0.8 \\ 0.2 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mu = \begin{bmatrix} 62.5 \\ 30 \\ 50 \\ 30 \\ 75 \end{bmatrix}$$

(a) Draw the network

(b) Classify the states into stable, overloaded and balanced and describe the long term behavior.

(c) Find the steady state limiting distribution of the stable part of the network.

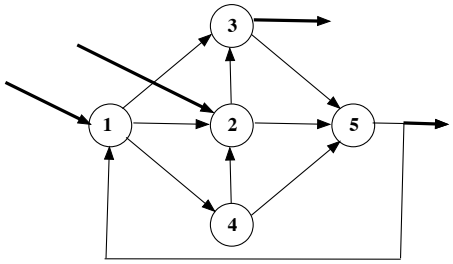
(d) Assume that initial limiting fluid, under fluid scaling

$$\bar{Q}(0) = \begin{bmatrix} 62.5 \\ 30 \\ 50 \\ 30 \\ 75 \end{bmatrix}.$$

Calculate the fluid paths of the network.

Solution

(a)



(b) We use the algorithm developed in Exercise 9.4.
Iteration 1 We calculate the inflow rates to all nodes when the outflows from all nodes are equal to the full processing rates:

$$\nu_i = \mu_i, \lambda = (75., 31., 40., 25., 64.), \mu - \lambda = (-12.5, -1., 10., 5., 11.).$$

We conclude that nodes $\{3, 4, 5\}$ are of type A, they are stable.
Iteration 2 we use equations (9.2) with $U = \{3, 4, 5\}$, and $\overline{U} = \{1, 2\}$. We obtain that if nodes in U are stable, and outflow from nodes in $\overline{U}$ are at rates μ_i, then the inflows are:

$$\nu_1 = \mu_1, \nu_2 = \mu_2, \quad \lambda = (71., 30., 40., 25., 55.), \quad \mu - \lambda = (-8.5, 0., 10., 5., 20.)$$

The set of nodes that must be stable has remained as after iterations 1. Hence

$$\text{Type A: } \{3, 4, 5\}, \qquad \text{Type B: } \{2\}, \qquad \text{Type C: } \{1\}$$

In the long run, node 1 will accumulate inventory at rate 8.5. Node 2 will be transient. Nodes 3,4,5 will behave like a stable Jackson network. The traffic intensities of nodes 3,4,5 are:

$$\rho_3 = 40/50 = 0.8, \qquad \rho_4 = 25/30 = 0.833, \qquad \rho_5 = 55/75 = 0.733.$$

(c) If input of customers is Poisson, and services exponential then

$$\lim_{t\to\infty} \mathbb{P}((Q_3(t), Q_4(t), Q_5(t)) = (n_3, n_4, n_5) = \prod_{i=3}^{5} (1 - \rho_i)\rho_i^{n_1}$$

(d) The evolution is govem on the next table, at time points $t = 5, 6, 6.2$ buffers $3, 4, 5$ empty, buffer 2 reaches constant level 36 at time $t - 6$, and buffer 1 reaches level 140 at $t = 6.2$ and thereafter continues to fill up at rate 8.5. The following table shows levels x and rates of change in level v:

	t_0=0		t_1=5		t_2=6		t_3=6.2	
	x	υ	x	υ	x	υ	x	υ
1	62.5	12.5	125	12.5	137.5	12.5	140	8.5
2	30	1	35	1	36	0	36	0
3	50	-10	0	0	0	0	0	0
4	30	-5	5	-5	0	0	0	0
5	75	-11	20	-16	4	-20	0	0

9.3 Suppose that in the solution of the LCP (9.5), (9.6) the identities of the nodes of type A (underloaded) are known, i.e. $U = \{i : w_i > 0\}$ is given. Derive the solution of the LCP problem and the rate of inflow λ and outflows ν for the network.

Solution

Denote by $P_{U,V}$ the submatrix of P with rows U and columns V. Let the complement of U be $\overline{U}$. then:

$$\lambda_U = (I - P^{\mathsf{T}}_{U,U})^{-1}(\alpha + P^{\mathsf{T}}_{\overline{U},U}\mu_{\overline{U}})$$

and $\nu_U = \lambda_U$. Furthermore, $\nu_{\overline{U}} = \mu_{\overline{U}}$, and

$$\lambda_{\overline{U}} = \alpha_{\overline{U}} + P^{\mathsf{T}}_{U,\overline{U}}\lambda_U + P^{\mathsf{T}}_{\overline{U},\overline{U}}\mu_{\overline{U}}$$

Finally,

$$z_U = 0, \quad z_{\overline{U}} = \lambda_{\overline{U}} - \mu_{\overline{U}}, \quad w_{\overline{U}} = 0, \quad w_U = \mu_U - \lambda_U .$$

It is now possible to identify nodes of type B by $\nu_i = \lambda_i$ and of type C by $\lambda_i > \nu_i$.

9.4 Prove that if you define $\tilde{\lambda} = \alpha + P^{\mathsf{T}}\mu$, then $\tilde{\lambda}_i > \lambda_i$, i.e. $\tilde{\lambda}$ provides an upper bound for the vector of inflows. Generalize this statement for subsets of nodes, and use it to derive an algorithm to solve the LCP (9.5), (9.6) in at most I iterations, each involving one matrix inversion.

Solution

Clearly, by $\nu_i \leq \mu_i$, and non-negativity of P, α, μ, ν:

$$\tilde{\lambda} = \alpha + P^{\mathsf{T}}\mu \geq \alpha + P^{\mathsf{T}}\nu = \lambda \tag{9.1}$$

This implies that all nodes $U(1) = \{i : \tilde{\lambda}_i < \mu_i\}$ must be of type A.

To generalize this, let U be a subset of the nodes. It is immediate to see that

if the spectral radius of P is < 1 so is the spectral radius of $P_{U,U}$. Then, by non-negativity of $(I - P_{U,U}^{\intercal})^{-1}$,

$$\lambda_U = (I - P_{U,U}^{\intercal})^{-1}(\alpha_U + P_{\overline{U},U}\nu_{\overline{U}}) \leq (I - P_{U,U}^{\intercal})^{-1}(\alpha_U + P_{\overline{U},U}\mu_{\overline{U}}) \quad (9.2)$$

From the previous exercise we have, assuming that $i \in \overline{U}$ have outflow rate μ_i, the inflow rates to all nodes are given by:

$$\begin{aligned} \lambda_U &= (I - P_{U,U}^{\intercal})^{-1}(\alpha_U + P_{\overline{U},U}^{\intercal}\mu_{\overline{U}}) \\ \lambda_{\overline{U}} &= \alpha_{\overline{U}} + P_{U,\overline{U}}^{\intercal}\lambda_U + P_{\overline{U},\overline{U}}^{\intercal}\mu_{\overline{U}} \end{aligned} \quad (9.3)$$

The iterative procedure will be:

- iteration 1: Let $U(0) = \emptyset$, solve 9.3, for new values $\lambda^{(1)}$. Set $U(1) = \{i : \lambda_i^{(1)} < \mu_i\}$.
- iteration k: Solve 9.3 with $U(k-1)$, for new values $\lambda^{(k)}$. Set $U(k) = \{i : \lambda_i^{(k)} < \mu_i\}$.
- If $U(k) = U(k-1)$, stop, you found the correct inflow rates $\lambda = \lambda^{(k)}$. Else, go to next iteration.

Proof if in step 1 $U(1) = \{1, \ldots, I\}$, the network is stable. If $U(1) = U(0)$, all nodes are of type B or C. In all other cases, we show by induction that $U(k) \supseteq U(k-1)$. This is so trivially for step 1. Assume $U(k-1) \supseteq U(k-2)$, if they are equal the algorithm will stop. Else assume $U(k-1) \supset U(k-2)$. Then by the above, $\lambda^{(k)} \leq \lambda^{(k-1)}$, and hence $U(k) \supseteq U(k-1)$.
So the sets $U(k)$ increase until they become maximal. At that point, λ_i and $\nu_i = \lambda_i \wedge \mu_i$, $i = 1, \ldots, I$ solve the generalized traffic equations. □

9.5 Consider a memoryless Jackson network $Q(t)$, and let $U = \{i : \rho_i = \lambda_i/\mu_i < 1\}$ be its type A nodes. Define the network $Q^+(t)$ by:

$$\alpha_i^+ = \begin{cases} \alpha_i + \sum_{j \notin U} \mu_j p_{j,i} & i \in U \\ \alpha_i & i \notin U \end{cases}$$
$$p_{i,j}^+ = \begin{cases} 0 & i \notin U \text{ and } j \in U \\ p_{i,j} & \text{otherwise} \end{cases}$$

(i) Show that $Q_i^+(t) \geq_{ST} Q_i(t)$
(ii) Find the stationary distribution of the process $Q_U^+(t)$, that includes only the queues at the nodes in i.
Solution
(i) We couple system $Q(t)$ with system $Q^+(t)$, by letting the arrivals at rate $\mu_j, j \notin U$ in system $^+$ with job completions or dummy events at nodes $j \notin U$. It is then seen that nodes $i \in U$ will receive more input in sytem $^+$, so we will have $Q_i^+(t) \geq Q_i(t)\,, i \in U$, and also $\mathcal{D}_i^+(t) \geq \mathcal{D}_i(t)\,, i \in U$, from which we obtain that nodes $i \notin U$ will also receive more input in system $^+$, so also $Q_i^+(t) \geq Q_i(t)\,, i \notin U$
(ii) In the system $^+$ the nodes $i \in U$ are isolated from the nodes $i \notin U$, as

far as input, while output from the U sub-system goes partly to $i \notin U$ and partly out. Thus it is a Jackson network with λ_i, $i \in U$ solving (9.5), (9.6) with $\nu_i = \mu_i$, $i \notin U$, $\nu_i = \lambda_i$, $i \in U$, i.e. the λ_i of the traffic equations for Q.

9.6 Consider a memoryless Jackson network $Q(t)$, and let $U = \{i : \rho_i = \lambda_i/\mu_i < 1\}$ be its type A nodes. Define the network $Q^{-,\epsilon}(t)$ by:

$$\begin{aligned}
&\alpha_i^{-,\epsilon} = \alpha_i \\
&\mu_i^{-,\epsilon} p_{i,j}^{-,\epsilon} = \mu_i p_{i,j} \\
&\mu_i^{-,\epsilon} q_i^{-,\epsilon} = \begin{cases} \mu_i q_i & i \in U \\ \mu_i q_i + \lambda_i - \mu_i + \epsilon & i \notin U \end{cases}
\end{aligned}$$

where λ_i is the inflow rate obtained from the traffic equations for Q, and $q_i, q_i^{-,\epsilon}$ are the fraction of completed jobs at node i that leave the system.
(i) Show that $Q_i^{-,\epsilon}(t) \leq_{ST} Q_i(t)$
(ii) Show that $Q^{-,\epsilon}(t)$ is stable, and find its stationary distribution.
Solution
(i) We again couple $Q(t)$ and $Q^{-,\epsilon}(t)$ by having the same coupled sequences of arrivals, and job completions running all the time, with dummy completions when nodes are empty. Furthermore, when a job in system $Q^{-,\epsilon}(t)$ is routed into a node $i \notin U$, it leaves the system immediately with probability $(\lambda_i - \mu_i + \epsilon)/\lambda_i$. Then in system $Q^{-,\epsilon}(t)$ there will be less items in nodes $i \notin U$, and so altogether $Q^{-,\epsilon}(t) \leq Q(t)$.
(ii) Let λ_i be the inflow rates in $Q(t)$ (solving $\lambda = \alpha + P^{\mathsf{T}}(\lambda \wedge \mu)$). Consider first the system $Q^{-}(t) = Q^{-,0}(t)$, with $\epsilon = 0$. In this system $\mu_i^{-} = \mu_i$, $i \in U$ and $\mu_i^{-} = \lambda_i$, $i \notin U$, and $p_{i,j}^{-} = p_{i,j}$, $i \in U$ while $p_{i,j}^{-} = \frac{\mu_i}{\lambda_i} p_{i,j}$, $i \notin U$. In the $Q^{-}(t)$ system $\lambda = \alpha + P^{\mathsf{T}}\lambda$, and nodes in U are type A the rest are type B. For the systems $Q^{-,\epsilon}(t)$, the processing rates are higher, and all the routing in and out of node in U as well as routing rates into nodes not in U are same as for Q, the same λ will solve the traffic equations, with $\mu^{-,\epsilon} > \lambda_i$ for all nodes, so the system $Q^{-,\epsilon}(t)$ is stable.

9.7 Use the results of the previous exercises to prove:

Theorem 9.1 (Goodman and Massey (1984)). *In a Jackson network, for the set of nodes $U = \{i : \rho_i = \lambda_i/\mu_i < 1\}$ (nodes of type A, stable nodes),*

$$\lim_{t\to\infty} \mathbb{P}(Q_i(t) = n_i : i \in U) = \prod_{i\in U} (1-\rho_i)\rho_i^{n_i},$$

by showing that $Q^{-,\epsilon}(t)$, $Q(t)$, $Q^{+}(t)$ can be coupled so that: $Q^{-,\epsilon}(t) \leq Q(t) \leq Q^{+}(t)$.
Solution
All the systems $Q(t), Q^{+}(t), Q^{-,\epsilon}(t)$ are Jackson networks. For the nodes in U,

$$\lim_{t\to\infty} Q_U^{-,\epsilon}(t) \leq \lim_{t\to\infty} Q_U(t) \leq \lim_{t\to\infty} Q_U^{+}(t),$$

and with $\rho_i = \lambda_i/\mu_i$, $i \in U$,

$$\lim_{t\to\infty} Q_U^+(t) = \prod_{i\in U}(1-\rho_i)\rho_i^{n_i}, \qquad \lim_{\epsilon\searrow 0}\lim_{t\to\infty} Q_U^+(t) = \prod_{i\in U}(1-\rho_i)\rho_i^{n_i},$$

and the theorem follows.

9.8 We have derived the diffusion limits for a stable GI/GI/1 queueing network, where the queue length converges to $RBM(t;\theta,\lambda(c_a^2+c_s^2))$, where $\theta = \lim \sqrt{n}(\lambda^n - \mu^n)$. For a queue with input consisting of the superposition of several renewal processes, and with s servers, a similar result was derived by Iglehart and Whitt (1970a,b). The proof requires quite a bit of technique, but can you just write down what you think is the result ?

Solution

Assume that there are r renewal arrival streams, $\mathcal{A}_i(t)$, $i = 1,\ldots,r$ with rates λ_i and coefficients of variations $c_{i,a}$ and s processors, with processing completions renewal processes, $\mathcal{S}_j(t)$, $j = 1,\ldots,s$ with rates μ_j and coefficients of variations $c_{j,s}$. All arrivals join single queue and services are started in the order of arrivals. Assume a sequence oof such systems, indexed by superscripts ^{n}, let $\lambda^n = \sum_{i=1}^r \lambda_i^n$ and $\mu^n = \sum_{j=1}^s \mu_j^n$ such that $\lambda_i^n \to \lambda_i$, $\mu_j^n \to \mu_j$. Assume $n^{1/2}(\lambda^n - \mu^n) \to \theta$ with $-\infty < c < \infty$.

Then the centralized queue length satisfies: $Q^n(t)/n^{1/2}$ converges to a reflected Brownian motion with drift θ and diffusion coefficient $\sum_{i=1}^r \lambda_i c_{i,a}^2 + \sum_{j=1}^s \lambda_l c_{j,s}^2$.

9.9 In Section 9.5 we derived the components of the variance covariance matrix Γ when all the stations are heavily loaded. Generalize this for the case when the nodes of the network are of type A, B, and C [Chen and Mandelbaum (1991) or Chen and Mandelbaum (1994)].

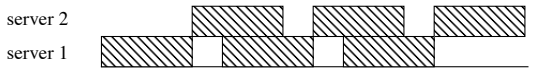
Solution

See calculations in Chen and Mandelbaum (1991) or Chen and Mandelbaum (1994).

9.10 Consider a tandem queueing network with two nodes. Construct an example with deterministic arrivals and deterministic service times, for which $\rho_i < 1$ but the system never empties.

Solution

Arrivals are at times $0, 2, 4, \ldots$, service time at the both servers is $3/2$.



9.11 Consider a parallel service system with two servers, in which arrivals are assigned randomly to one of the servers, and leave at service completion. Construct an example with deterministic arrivals and deterministic service times, for which $\rho_i < 1$ but the system never empties.

Solution

Arrivals are at times $0, 1, 2, \ldots$, service time is $3/2$, arrival assigned to server

1 or server 2, each with probability $1/2$. This system is stable, since arrivals to each server are at rate $1/2$, and service rate is $2/3$, so $\rho < 1$ at each server. Assume the system is not empty during $(n, n+1)$, because a customer arrives at time n, and goes to one of the servers, which may be busy or not busy, but in any case with the new customer that server will be busy at least until $n+3/2$, and at $n+1$ another customer arrives so the system will be non-empty in $(n+1, n+2)$.

Part IV

Fluid Models of Multi-Class Queueing Networks

10

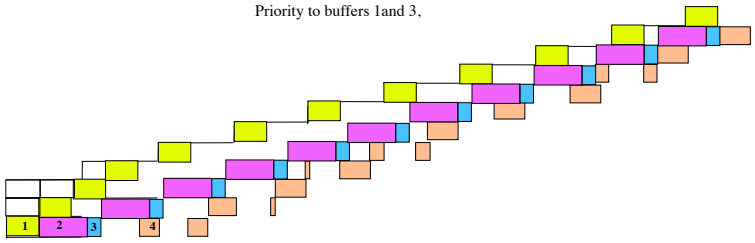
Multi-Class Queueing Networks, Instability and Markov Representations

Exercises

10.1 Prove that the Lu-Kumar system, with $\alpha = 1$ and deterministic arrivals and service times, starting with x parts at time 0^-, does not diverge under any buffer priority policy exept priority to buffers 2 and 4. Analyze the case that priority is given to buffers 1 and 3 in detail.

Solution

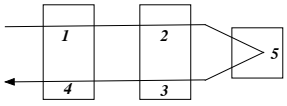
There are 4=2x2 priority policies. We know that priority to 2 and 4 may blow up. The following is a picture of how jobs are processed when priority is given to buffers 1 and 3. The picture remains almost the same with priority to 1 and 2. It is seen that eventually the initial workload of 3 is diminished, and the system is stable.



10.2 The Lu-Kumar network has immediate feedback from class 2 to class 3 at the same node. Show that a similar example can be constructed without immediate feedback.

Solution

Adding a station serving buffer 5 on the route between 2 and 3, will may still be unstable under priority to 2 and 4.



10.3 Explain that the FIFO network of Bramson operates similarly to the Lu-Kumar network under class 2,4 priorities.

Solution

Starting from x customers in 1 and all others empty, they quickly move to 2. all these customers have priority over buffers 3–5, since they arrived earlier. When all are out of 2, they very quickly move through 3–5 snd into 6. Note that at that time 1 is almost empty, since while 2 was processing all input from 1 went straight to 2. Now 6 has a very large quantity of arrivals that arrived almost simultaneously, and further input to 1 has to wait as they arrived later, so while 6 is working input to 2 is again blocked. So 2 and 6 form a virtual machine also for this FIFO network.

10.4 Consider the Lu-Kumar network with Poisson input of rate 1, exponential services with rates μ in buffers 2 and 4, and negligible processing time at buffer 1 and 3. Argue that under priority to buffers 2 and 4, it is stable for $\mu < 2$ and blow up for $\mu > 2$.

Solution

Assume at time 0^- there are altogether x customers in buffer 1. At time 0 they move to buffer 2 which will work until empty. Time to get to this point consists of the sum x independent busy periods of the M/M/1 queue. It can be shown that the average number served in a busy period of an M/M/1 queue is $(1-\rho)^{-1} = \frac{\mu}{\mu-1}$. So when they move to buffer 4, there will be Y customers, with $\mathbb{E}(Y) = \frac{x\mu}{\mu-1}$. The duration of their service T has expectation: $\mathbb{E}(T) = \frac{x}{\mu-1}$, and the number of arrivals to buffer 1 is then X with expectation $\mathbb{E}(X) = \frac{x}{\mu-1}$. So now we compare: x with $\mathbb{E}(X)$:

$$\mathbb{E}(X) = \frac{x}{\mu-1} \lesseqgtr x \quad \Longleftrightarrow \mu \lesseqgtr 2.$$

This shows that for $\mu < 2$ the number in buffer 1 after this cycle is larger than at time 0^-, so the queue blows up. It indicates that it is stable if $\mu > 2$.

The complete proof of stability is: Starting from any state, one of the buffers 2, or 4 will empty first, and if it is buffer 4, then at the end of work in buffer 2 and 4 we will at time T_0 be with x customers in buffer 1 and all others empty. We now look at the stopping times $T_n > T_{n-1}$ of this happening again and use Foster multiplicative criterion to show stability.

10.5 Analyze the KSRS network with deterministic interarrivals and services, under priority to buffers 2 and 4, when it satisfies all 3 conditions, and when it violates the virtual machine condition.

Solution

Assume arrivals to buffer 1 are at $1, 3, 5, \ldots$ and to buffer 3 at $2, 4, 6, \ldots$. Procssing rates at buffers 1,3 are negligible and the are at rate μ in each of the buffers 2,4. Assume the system starts at 0 with x in buffer 1 and all the others empty. We need $\mu > 1/2$ so the two machines are stable. For the virtual machine we need $\mu > 1$.

For the example, buffer 1 transfers to 2 and 2 will work for a time of length $x/(\mu - 1/2)$, at which time buffer 3 will accumulate $x/(2\mu - 1)$ which will

transfer to buffer 4 and buffer 4 will the empty at $x/(2(\mu - 1/2)^2)$, and by the time buffer 4 will be empty, buffer 1 will accumulate $x/(2\mu - 1)^2$. So it will be stable or unstable if:

$$E(X) = x/(2\mu - 1)^2 \lessgtr x \quad \Longleftrightarrow \mu \lessgtr 1$$

10.6 Propose a Markov process to describe MCQN under FCFS, that does not require the ages of all the customers in the system.

Solution

It is enough in addition to the vector (Q, U, V), to keep for each buffer a list of the classes of the items in the buffer in their order of arrival. Then at any event of a customer moving from buffer i to buffer j (including buffer 0 for the outside world), the state of the two buffers i, j is updated, by deleting the leaving item from i (where he was first in the list) and adding his class designation as last in the list of buffer j. The next item to be served at any buffer is then the HOL item from the class that is first in the list. Note that this list of classes for each buffer is finite, the state space of this additional information Y is countable. Hence, for Poisson arrivals and exponential service, the MCQN under FCFS has countable state space. Unfortunately this does not make it any more tractable.

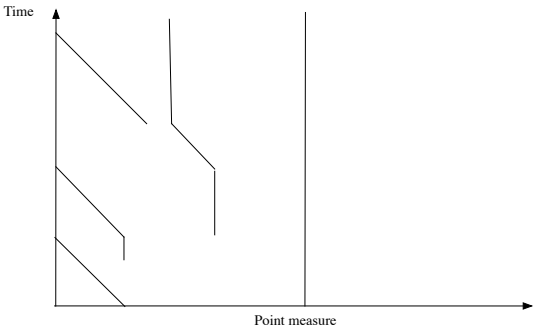
10.7 (∗) To describe the dynamics of a single queue under SPT or SRPT, one requires keeping track of the remaining processing times of all jobs. This requires that the queue be described by a measure valued process. Describe the dynamics of the single queue under SPT or SRPT.

Solution Down et al. (2009)

The state of the system is described by the measure valued process $\mathcal{Z}(t) = \sum_{j \le \mathcal{A}(t)} \delta_{\upsilon_j(t)}$ where $\mathcal{A}(t)$ is the number of initial and arrivals up to time t, $\upsilon_j(t)$ is the remaining processing time of the jth arrival, and $\delta_{\upsilon_j(t)}$ is a point measure of 1 at the remaining processing time of job j if positive, and is 0 if job $\upsilon_j(t) = 0$ so job j has already departed. Denoting by χ the function $\chi(t) = t$, and by 1^+ the function $1(t) = 1$, $t > 0$, $1(0) = 0$, and for function f and measure ξ let $\langle f, \xi \rangle$ be integral of f by measure ξ. Then:

$$\begin{aligned} Q(t) &= \int_{0+}^{t} 1 d\mathcal{Z}(x) = \langle 1^+, \mathcal{Z} \rangle, \\ \mathcal{W}(t) &= \int_{0}^{t} x d(\mathcal{Z}(x) = \langle \chi, \mathcal{Z} \rangle, \\ \mathcal{D}(t) &= \mathcal{Z}(0). \end{aligned}$$

With this measure valued description we need to describe the dynamics: For SRPT, all $\upsilon_j(t)$ except the shortest are unchanged, the shortest decreases at rate 1. A job is preempted if a new shortest job arrives. At draws, lower j has priority. For SPT there are no preemptions. The following figure describes the measure valued process for SRPT:



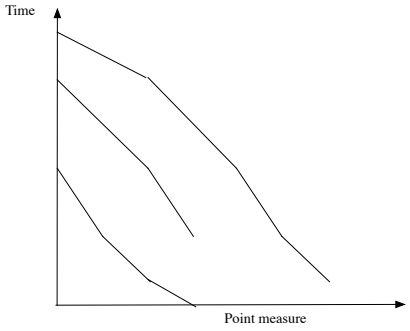
10.8 (∗) Consider a single server queue. Denote by $\phi(x) = 1/x$, $x > 0$, $\phi(0) = 0$. Define the attained service process $\eta(t) = \int_0^t \phi(Q(s))ds$. It is the cumulative amount of service per customer delivered by the server under PS policy. Use it to express the measure valued process and the dynamics of the queue [Gromoll et al. (2002); Gromoll (2004)].

Solution

With the measure valued process $\mathcal{Z}(t)$ as above the dynamics for PS are that the value of

$$\upsilon_j(t) = \big(\upsilon_j - (\eta(t) - \eta(u_j)\big) \vee 0$$

where u_j is the arrival time of customer j, and υ_j is his processing time. The following figure describes the measure valued process for PS:



10.9 Consider the M/G/1 queue, with state described by (n, x) where n is the number of customers in the system, and x the age of the customer in service (0 if empty). Show that $\{(0, x) : 0 \le x < \kappa\}$ is a uniformly small set.

Solution

Let $\mathbb{P}^t_{n,x}(B)$ be the probability to go from state (n, x) at time 0 to set B at time t. For given κ, define the measure $\xi(B) = \mathbb{P}^T_{0,0}(B)e^{-\lambda T}$, where T is any

value $> \kappa$. Then for $x < \kappa$,

$$\begin{aligned}\mathbb{P}^T_{0,x}(B) &= \mathbb{P}^T_{0,x}(B \,|\, \text{arrival in } (0,T)\,)\mathbb{P}(\text{arrival in } (0,T)) \\ &\qquad + \mathbb{P}^T_{0,x}(B \,|\, \text{no arrival in } (0,T))\mathbb{P}(\text{no arrival in } (0,T)) \\ &= \mathbb{P}^T_{0,x}(B \,|\, \text{arrival in } (0,T))\mathbb{P}(\text{arrival in } (0,T)) + \mathbb{P}^T_{0,0}(B)e^{-\lambda T} \\ &\geq \xi(B).\end{aligned}$$

10.10 Consider the GI/GI/1 queue, with state described by (n, x, y) where n is the number of customers in the system, and x the age of the customer in service (0 if empty), and y the time since the last arrival. Assume that interarrival times have infinite support. Show that $\{(0, x, y) : 0 \leq x + y < \kappa\}$ is a uniformly small set.

Solution

The proof is the same, by replacing $e^{-\lambda T}$ with $\mathbb{P}(\text{interarrival} > T)$. The assumption of unbounded support cannot be relaxed if we want the result to hold for any GI/GI/1 queue. Otherwise, if interarrival and service times are deterministic the chain cannot be a Harris chain.

10.11 Show that if there is more than one input to a MCQN (or even a generalized Jackson network), and interarrival times are integer, then the Markov process describing the network cannot be ergodic.

Solution

Consider some stable MCQN, and let u_1, u_2 be the initial residual times to arrival at station 1 and station 2. Assume at time 0 we have on paths of scenario 1, $u_1 = 0, u_2 = 1/\pi$, an on the paths of scenario 2 we have $u_1 = 0, u_2 = 1/e$. Then clearly, the states where the remaining times to arrivals satisfy $U_1(t) - U_2(t) \in A$, will occur in different frequency of times under scenario 1 and under scenario 2. So the process cannot be ergodic.

11

Stability of MCQN via Fluid Limits

Exercises

11.1 Derive the standard fluid model equations under the assumption that $\frac{1}{n}U_k^n(0) \to \bar{U}_k(0) > 0$, and $\frac{1}{n}V_k^n(0) \to \bar{V}_k(0) > 0$, for $k = 1, \ldots, K$, as $n \to \infty$.

Solution

In vector form:

$$\bar{Q}(t) = \bar{Q}(0) + \alpha\big(t - \bar{U}(0)\big)^+ - R\big(\bar{\mathcal{T}}(t) - \bar{V}(0)\big)^+$$

Since exogenous input to k (excluding the first arrival that arrives at time $U_k(0)$) only starts to accumulate after $U_k^n(0)$, and time devoted to processing out of k (excluding time devoted to the first customer, which is equal to $V_k(0)$) only starts after $V_k^n(0)$.

11.2 Explain the stochastic and the fluid model equations (11.5), (11.6) that are added for static priority policies. Show that they determine the stochastic queueing process, and verify that their fluid versions are satisfied by every fluid limit.

Solution

For priority policy: we indeed require that while $Q_k^+(t) > 0$ all processing to buffers that are not with priority equal to or exceeding k should be stopped, and by work conservation there should be no idling, so $t - \mathcal{T}_k^+(t)$ cannot increase. This explains (11.5).

We next show that the standard equations and the added (11.5) determine $Q(t)$. Evolution between events and the change in state at events are determined by the standard equations. So all that remains is the decision at arrivals or job completions which job to start, and by HOL, which class to start. If at node i class k has highest priority among customers present, then we cannot start priority higher than k since empty, and not priority lower than k because $t - \mathcal{T}_k^+(t)$ cannot grow by (11.5). Hence, by non-idling, HOL job of class k will start.

To see (11.6) assume $\bar{Q}_k^+(t) > 0$. By continuity of $\bar{Q}$, $\bar{Q}_k^+(t)$ is positive for some $t_1 \le t \le t_2$, and so $Q_k^{n+}(t) > 0$ in $nt_1 \le t \le nt_2$ for $n > n_0$, and hence $t - \mathcal{T}_k^n(t)$ is not increasing in $nt_1 \le t \le nt_2$ for $n > n_0$, and so $t - \bar{\mathcal{T}}_k^+(t)$ is

not increasing in $t_1 \le t \le t_2$. It follows that (11.6) is satisfied by every fluid limit.

11.3 Write down the stochastic system equations for class and station immediate workload, and the resulting standard fluid model equations for them.

Solution

The amount of work for class k at time t consists of residual service of customer in service plus the work for all other customers in that class. Let $V_k(t)$ be the residual service of job currently in service (or 0 if empty).

$$\mathcal{W}_k(t) = V_k(t) + \sum_{\ell=S_k(\mathcal{T}_k(t))+2}^{S_k(\mathcal{T}_k(t))+Q_k(t)} \upsilon_k(\ell).$$

Assume for simplicity $\frac{1}{n}V_k^n(0) \to 0$ as $n \to \infty$. We have

$$\frac{1}{n}\mathcal{W}_k^n(nt) = \frac{1}{n}V_k^n(nt) + \frac{1}{n}\sum_{\ell=\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+2}^{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+Q_k(nt)} \upsilon_k(\ell)$$

The first summand converges to 0 by Lemma 11.5. The second summand can be written as a difference between two sums starting at 1. Now:

$$\frac{1}{n}\sum_{\ell=1}^{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+1} \upsilon_k(\ell) = \frac{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+1}{n}\frac{1}{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+1}\sum_{\ell=1}^{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+1} \upsilon_k(\ell)$$
$$\to \mu_k\bar{\mathcal{T}}_k(t)m_k = \bar{\mathcal{T}}_k(t), \quad \text{u.o.c. in } t, \text{ a.s.}$$

by the fact that $\bar{\mathcal{T}}_k(t) < t$, time change, and FSLLN. For the second sum:

$$\frac{1}{n}\sum_{\ell=1}^{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+Q_k^n(nt)} \upsilon_k(\ell) = \Big(\frac{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))}{n} + \frac{Q_k^n(nt)}{n}\Big)$$
$$\frac{1}{\mathcal{S}_k^n(\mathcal{T}_k^n(nt)) + Q_k^n(nt)}\sum_{\ell=1}^{\mathcal{S}_k^n(\mathcal{T}_k^n(nt))+Q_k^n(nt)} \upsilon_k(\ell)$$
$$\to \big(\mu_k\bar{\mathcal{T}}_k(t) + \bar{Q}_k(t)\big)m_k = \bar{\mathcal{T}}_k(t) + m_k\bar{Q}_k(t)\big) \quad \text{u.o.c. in } t, \text{ a.s.}$$

So:

$$\frac{1}{n}\mathcal{W}_k^n(nt) \to \bar{\mathcal{W}}_k(t) = m_k\bar{Q}_k(t) \quad \text{u.o.c. in } t, \text{ a.s.}$$

The immediate workload for station i is defined as the sum of $\mathcal{W}_k(t)$. It is the time to finish the current operation on each of the jobs currently in any of the buffers of the station. Arrivals after t from other stations are not included, and also jobs that complete at station i are not readmitted for further visits to station i after t.

Another way to calculate it is to look at what entered exogenously, $\mathcal{A}_k(t)$,

and what enters from other stations, $\sum_{l\neq k}\mathcal{R}_{l,k}(\mathcal{S}_l(\mathcal{T}_l(t)))$. Denote $\mathcal{E}_k(t) = Q_k(0) + \mathcal{A}_k(t) + \sum_{l\neq k}\mathcal{R}_{l,k}(\mathcal{S}_l(\mathcal{T}_l(t)))$, Then

$$\mathcal{W}_i(t) = \sum_{k\in C_i}\sum_{\ell=1}^{\mathcal{E}_k(t)} v_k(\ell) - \sum_{k\in C_i}\mathcal{T}_k(t).$$

Which counts all the work for buffer k of all the jobs that entered before t, added up over the stations, and subtract the total time that the station worked before t.

For the fluid limit we again write:

$$\begin{aligned}
\frac{1}{n}\mathcal{W}_i^n(nt) &= \sum_{k\in C_i}\left[\frac{1}{n}\sum_{\ell=1}^{\mathcal{E}_i^n(nt)} v_k(\ell) - \frac{1}{n}\mathcal{T}_k^n(nt)\right] \\
&= \sum_{k\in C_i}\left[\frac{\mathcal{E}_i^n(nt)}{n}\frac{1}{\mathcal{E}_i^n(nt)}\sum_{\ell=1}^{\mathcal{E}_i^n(nt)} v_k(\ell) - \frac{1}{n}\mathcal{T}_k^n(nt)\right] \\
\to &\sum_{k\in C_i} m_k\left[\bar{Q}_k(0) + \alpha_k t + \sum_{l\neq k} P_{l,k}\mu_l\bar{\mathcal{T}}_l(t) - \mu_k\bar{\mathcal{T}}_k(t)\right] \\
&= \sum_{k\in C_i} m_k\bar{Q}_k(t), \quad \text{u.o.c. in } t, \text{ a.s.}
\end{aligned}$$

11.4 Explain the stochastic and the fluid model equations (11.7) – (11.9) that are added for FCFS policies. Show that they determine the stochastic workload and queueing processes, and verify that their fluid versions are satisfied by every FCFS fluid limit.

Solution

The immediate workload is exactly all that at time t needs to be processed in machine i before any job that arrived, from outside or from any other buffer after time t can be started. Under FCFS, this is exactly what will happen at time t: from time t to time $t + \mathcal{W}_i(t)$, station i will not be empty and wiil be working without idling (work conderving), and it will only be working on jobs that arrived before t by FCFS, until $t + \mathcal{W}_i(t)$, at which time all the jobs that arrived to machine i before t will be exhausted. So all arrivals prior to t will form the departures up to time $t + \mathcal{W}_i(t)$.

This is again an implicit equation, but when considered event by event, it exactly provides the unique sample path of the policy. In particular, when a job joins the queue in buffer k at time t, his start of processing time is determined exactly as $t + \mathcal{W}_i(t)$.

We have already verified equation (11.9), that $\bar{\mathcal{W}}_k(t) = m_k\bar{Q}_k(t)$, the remaining fluid equations follow immediately.

11.5 Prove Lemma 11.5.

Solution

Proof of Lemma 11.5 (Bramson (2008), Lemma 4.13) We have $v_i \geq 0$ i.i.d. with $\mathbb{E}(v_i) = m < \infty$. We wish to show that

$$\frac{1}{n}\max\{v_1, \ldots, v_n\} \to 0 \text{ a.s.}, \quad \frac{1}{n}\mathbb{E}\left(\max\{v_1, \ldots, v_n\}\right) \to 0$$

Consider the random walk $\mathcal{S}(t) = \sum_{i=1}^{\lfloor t \rfloor} v_i$. By FSLLN, $\frac{1}{n}\mathcal{S}(nt) \to mt$ a.s. uniformly for $t \in [0, 1]$. For $\epsilon > 0$, we have for n large enough:

$$\frac{1}{n}\max_{t_1 n \leq i \leq t_2 n} v_i \leq \frac{1}{n}\sum_{i=\lfloor t_1 n \rfloor}^{\lfloor t_2 n \rfloor} v_i \leq (t_2 - t_1 + \epsilon)m,$$

holds uniformly for all $0 \leq t_1 < t_2 \leq 1$. Set $t_2 - t_1 = \epsilon$. We get:

$$\max\{v_1, \ldots, v_n\} = \sup_{t \in [0,1]} \max_{tn \leq i \leq (t+\epsilon)n} v_i \leq 2\epsilon mn,$$

and we have shown that: $\frac{1}{n}\max\{v_1, \ldots, v_n\} \to 0$ a.s. .
Clearly, by $\mathbb{E}(\frac{1}{n}S(nt)) = mt$, $\frac{1}{n}S(nt)$ are uniformly integrable, and so $\frac{1}{n}\max\{v_1, \ldots, v_n\}$ are uniformly integrable, and therefore also:
$\frac{1}{n}\mathbb{E}\left(\max\{v_1, \ldots, v_n\}\right) \to 0$. □

remark The condition $v_i \geq 0$ is not necessary: One can do the proof separately for v_i^+ and for v_i^-.

11.6 Prove Theorem 11.12

Solution

Proof of Theorem 11.12 (Bramson (2008), Proposition 5.21) We assume: For any fluid limit with $\bar{Q}(0) = 0$ exists δ such that $\bar{Q}(\delta) \neq 0$. Need to show: for any fixed $Q(0) = x$ and $\omega \in \mathfrak{G}$, $\lim_{t\to\infty} \frac{Q(t)}{t} > 0$.
Assume to contrary: for $\omega \in \mathfrak{G}$, $\lim_{t\to\infty} \frac{Q(t,\omega)}{t} = 0$. Then for all $a_n \to \infty$, $\frac{Q(a_n,\omega)}{a_n} \to 0$, and also for any fixed δ, $\frac{Q(a_n\delta,\omega)}{a_n} \to 0$. By existence of fluid limits, we have subsequence $a_r \to \infty$ such that $\frac{Q(a_r t,\omega)}{a_r} \to \bar{Q}(t)$. This fluid limit must have $\bar{Q}(0) = 0$, since we had a single path, with fixed initial condition. By assumption, ther exist δ for this fluid limit for which $\bar{Q}(\delta) \neq 0$. But then: $\lim_{a_r\to\infty} \frac{Q(a_r\delta,\omega)}{a_r} \neq 0$, a contradiction. This proves (a)
To prove (b): We assume $\rho_i > 1$ for some station. We argue now for any fluid limit that:

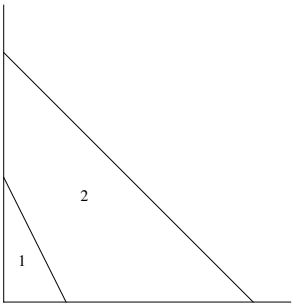
$$CR^{-1}\left(\bar{Q}(t) - \bar{Q}(0)\right) = CR^{-1}\left(\alpha t - (I - P^{\intercal})\bar{\mathcal{D}}(t)\right) = \boldsymbol{\rho} t - C\bar{\mathcal{T}}(t).$$

But $C\bar{\mathcal{T}}(t) \leq t$ componentwise, and therefore there exists $c > 0$ so that for every fluid limit with $\bar{Q}(0) = 0$, $|\bar{Q}(t)| \geq ct$. The same argument as for (a) then leads to $\liminf_{t\to\infty} |Q(t)|/t \geq c$. □

11.7 Prove Lemma 11.13
Solution

Proof of Lemma11.13 Assume at some regular t, $f(t) = 0$ but $\dot{f}(t) \neq 0$. If $\dot{f}(t) = c > 0$, then for some small ϵ and δ, $\frac{f(t-\delta)}{\delta} < -c+\epsilon < 0$, contradicting $f \geq 0$. The case $c < 0$ is similar.
Next, assume if $f(t) > 0$ then $\dot{f}(t) \leq \kappa$. Assume $f(0) > 0$, and assume that $f(s) > 0$ for all $s \in [0, t]$. By absolute continuity we have $f(t) = f(0) + \int_0^t \dot{f}(s)ds \leq f(0) - \kappa t$. So $t \leq f(0)/\kappa$, and so $f(t)$ will reach 0 by time $f(0)/\kappa$.
Assume when $f(t) > 0$, $\dot{f}(t) \leq 0$. We wish to show that if $f(t_0) = 0$ then $f(t) = 0$ for all $t > t_0$. We have: $f(t) = f(t_0) + \int_{t_0}^t \dot{f}(s)ds = 0$ since for $s \in [t_0, t]$, if $f(s) > 0$, $\dot{f}(s) \leq 0$, and if $f(s) = 0$, then as we saw, $\dot{f}(s) = 0$. □

11.8 Show that in a stable Jackson network, $\bar{Q}(t)$ need not be a decreasing function in all coordinates while $t < t_0$.
Solution
Consider a two station tandem queue, with $\alpha = 2$, $\mu_1 = 4$, $\mu_2 = 3$ and initial fluid $\bar{Q}_1(0) = \bar{Q}_2(0) = 5$. The fluid in the two buffers is plotted in the following figure:



11.9 Show that a feed forward MCQN is stable under any work conserving HOL policy.
Solution
Order the stations $1, \ldots, I$ so that customers never go from $k \in C_i$ to $l \in C_j$ if $i > j$. The stability follows immediately, since the fluid model will empty stations 1 in finite time (by Theorem 11.15 for the single station). Following that, station 2, if not empty yet, will empty in finite time, and so on.

11.10 Prove that a re-entant line with $\rho < 1$ is stable under LBFS.
Solution
We take arrival rate 1, and consider $f(t) = |\bar{Q}(t)|$, and its derivative $\dot{f}(t)$ at regular points. First, because $|\bar{Q}(t)| = |\bar{Q}(0)| + t - \mu_K \bar{T}_K(t)$, we have

$\dot{f}(t) = 1 - d_K(t)$. Let k_0 be the last non-empty fluid buffer at time t, so by (b) $d_k(t) = \ldots = d_K(t)$. We have that $\dot{\bar{\mathcal{T}}}_k^+(t) = 1$. On the other hand we have $\dot{\bar{\mathcal{T}}}_k^+(t) = \sum_{l \in H_k} m_k d_k(t) = d_K(t) \sum_{l \in H_k} m_k$, so $d_K(t) = 1/\sum_{l \in H_k} m_k$. Let $\lambda = 1/\max_{1 \le i \le I} \sum_{k \in C_i} m_k = \max_i \rho_i$. Then, by the above, and the assumption that $\rho_i < 1$ we have $d_K(t) \ge \lambda > 1$. Hence the system will be empty at or before the time $|\bar{Q}(0)|/(\lambda - 1)$.

11.11 Consider a fluid re-entrant line with input rate 1, and initial buffer contents $Q_1(0) = 1$ and all the other buffers empty. Show that this fluid re-entrant line under FBFS policy will be empty by time

$$t_K = \sum_{k=1}^{K} m_k \frac{\prod_{l=1}^{k-1} \left(1 - \sum_{j \in H_l \setminus l} m_j\right)}{\prod_{l=1}^{k} \left(1 - \sum_{j \in H_l} m_j\right)}$$

Solution (Dai and Weiss (1996)),

As we saw, under FBFS buffers will empty $1, 2, \ldots$ so that buffer k is empty at time t_k and stays empty thereafter. Since we started with all buffers except 1 empty, fluid will move from 1 to 2, then 2 will empty into 3, etc. We now calculate recursively: from $|Q(t_{k-1})|$ which is the content of buffer k, we obtain $t_k - t_{k-1}$, and then $|Q(t_k)|$. Initially, $|Q(0)| = 1$, and $t_1 = \frac{m_1}{1-m_1}$, so $|Q(t_1)| = (1 + t_1) = \frac{1}{1-m_1}$.

At time t_{k-1}, all the fluid will be in buffer k, the quantity being $|Q(t_{k-1})|$. The outflow rate from buffer k will be $d_k = \frac{1-\sum_{j \in H_k \setminus k} m_j}{m_k} > 1$ so $\dot{Q}_k = -(d_k - 1) < 0$, and:

$$t_k - t_{k-1} = \frac{|Q(t_{k-1})|}{d_k - 1} = |Q(t_{k-1})| \frac{m_k}{1 - \sum_{j \in H_k} m_j}$$

and:

$$|Q(t_k)| = |Q(t_{k-1})| + t_k - t_{k-1} = |Q(t_{k-1})| \left(1 + \frac{m_k}{1 - \sum_{j \in H_k} m_j}\right)$$
$$= |Q(t_{k-1})| \frac{1 - \sum_{j \in H_k \setminus k} m_j}{1 - \sum_{j \in H_k} m_j}$$

We have:

$$t_k - t_{k-1} = |Q(t_{k-1})| \frac{m_k}{1 - \sum_{j \in H_k} m_j},$$
$$|Q(t_k)| = |Q(t_{k-1})| \frac{1 - \sum_{j \in H_k \setminus k} m_j}{1 - \sum_{j \in H_k} m_j}$$

From which:

$$|Q(t_k)| = \prod_{\ell=1}^{k} \frac{1 - \sum_{j \in H_\ell \setminus \ell} m_j}{1 - \sum_{j \in H_\ell} m_j}$$

and

$$t_K = \sum_{k=1}^{K} m_k \frac{\prod_{\ell=1}^{k-1}(1 - \sum_{j \in H_\ell \setminus \ell} m_j)}{\prod_{\ell=1}^{k}(1 - \sum_{j \in H_\ell} m_j)}$$

Note: starting from any $Q(0)$ with $|Q(0)| = 1$, the time to empty under FBFS will be not more than the above value of t_K.

11.12 Find a lower bound to the time needed to empty a fluid re-entrant line, and suggest a policy that will achieve this time.

Solution

Let q_k be the total fluid in buffer k, let α be the inflow rate, m_k processing time for unit fluid at buffer k, and C_i the constituency of station i. Dfine $q_k^+ = \sum_{j=1}^{k} q_j$. Then the emptying time T must satisfy:

$$T > \sum_{k \in C_i} (q_k^+ + T\alpha) m_k, \quad i = 1, \ldots, I.$$

Define

$$t_i = \frac{\sum_{k \in C_i} q_k^+}{1 - \alpha \sum_{k \in C_i} m_k}$$

The the minimum time to empty is

$$T^* = \min(t_1, \ldots, t_i).$$

It can be acieved if we use the processing rates for each of the buffers as

$$d_k^* = \frac{q_k^+ + \alpha T^*}{T^*} = \frac{q_k^+}{T^*} + \alpha.$$

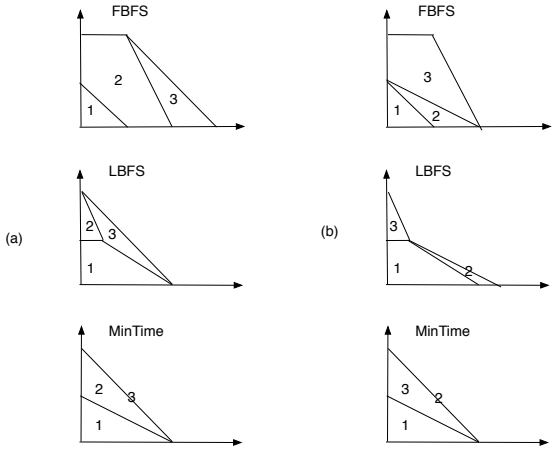
Note that inflow to k is d_{k-1}^*, so the rate of change at the level in each buffer is q_k/T^*. This means that we reduce all the levels at each of the buffers by a constant rate until all the buffers are empty at the same time.

11.13 For the following two examples of fluid re-entrant lines, draw the fluid levels for FBFS, LBFS, and minimum time to empty, and compare time to empty, and inventory:

(i) $C_1 = \{1, 2\}$, $C_2 = \{3\}$, $\alpha = 0$, $m = (1, 0.5, 1)$, $Q(0) = (1, 1, 0)$.

(ii) $C_1 = \{1, 3\}$, $C_2 = \{2\}$, $\alpha = 0$, $m = (1, 2, 0.5)$, $Q(0) = (1, 0, 1)$.

Solution



11.14 Consider a fluid MCQN with $\rho < 1$. Calculate a bound on the minimum time to empty the network, and devise a policy that will achieve that lower bound (use processor splitting, predictive policy). Use this to suggest a policy that will be stable for any MCQN with $\rho < 1$.

Solution

The amount of work needed at stations $i = 1, 2, \ldots, I$ to empty the fluid system by time T is given by $w_1, \ldots, w_i$ as follows:

$$(w_i)_{i=1,\ldots,I} = C\,\mathrm{diag}(m)(I - P^{\mathrm{T}})^{-1}(Q(0) + \alpha T) := M(Q(0) + \alpha T).$$

We can now solve for individual nodes:

$$t_i = \Big(MQ(0)\Big)_i + \Big(M\alpha\Big)_i t_i,$$

to get:

$$t_i = \frac{\Big(MQ(0)\Big)_i}{1 - \Big(M\alpha\Big)_i},$$

and we then have the lower bound on emptying time:

$$T^* = \max(t_1, \ldots, t_I).$$

The amount of work at buffer k if we empty all by time T^* is then:

$$(v_k)_{k=1,\ldots,K} = \mathrm{diag}(m)(I - P^{\mathrm{T}})^{-1}(Q(0) + \alpha T^*) := R^{-1}(Q(0) + \alpha T^*),$$

and so a fluid policy for emptying in minimum time is to allocate a fraction u_k^* of the processing time of node i to buffer k, where

$$u_k^* = \frac{v_k}{T^*} = \frac{\Big(R^{-1}Q(0)\Big)_k}{T^*} + \Big(R^{-1}\alpha\Big)_k.$$

This will again empty the quantities in each buffer gradually at a constant rate, until all buffer are empty at T^*.

These allocated rates, using processor splitting will approximate minimum time for a discrete stochastic system, when the system is large i.e. has many customers in the systems, and high arrival and processing rates.

A non-splitting non-preemptive version will be to choose next to process a job from buffer k for which $\mathcal{T}_k(t) - \frac{\nu_k}{T^*}t$ is maximal.

11.15 Consider a single station fluid re-entrant line with $\rho < 1$. Show that the total amount of fluid in the network is minimized pathwise under LBFS policy, and is maximized pathwise under FBFS policy.

Solution

Note first that the total time to empty is independent of the policy, as long as it is non-idling, say it is T.

With LBFS, we wish to show that the maximal amount leaves the system up to any time $0 < t < T$. Clearly, if we do not work on buffers $1, \ldots, k-1$ we can empty buffers $k, \ldots, K$ faster than if we do also work on buffers $1, \ldots, k-1$. So if we work only on buffers $k, \ldots, K$ until they are empty at time t_k, we cannot do better in terms of total fluid leaving the system by time t_k. This shows that we should empty K first, then $K-1$ etc., so LBFS maximizes outflow and minimizes the quantity in the system pathwise. The rates of outflow are: $\frac{1}{m_K}$ for $0 < t < q_K m_K = t_K$, followed by $\frac{1}{m_{K-1}+m_K}$ for $t_K < t < t_K + q_{K-1}(m_{K-1} + m_K) = t_K$, and so on, so if buffers $K, K-1, \ldots, k+1$ are empty at time t_{k+1} this is followed by outflow at maximal rate $1/(m_k + \cdots + m_K)$ for the time period $t_{k+1} < t < t_{k+1} + q_k(m_k + \cdots + m_K) = t_k$. By the time that buffers $2, \ldots, K$ are empty, we still need to empty buffer 1 which will take an additional time $T - t_2$, to empty the quantity $q_1 + \alpha T$, at rate $1/(m_1 + \cdots, m_K$.

With FBFS no outflow occurs at all until buffers $1, \ldots, K-1$ are empty, so this is least possible outflow for the longest possible duration. After that all the fluid in the system is emptied at rate $1/m_K$. Because of non-idling we cannot retain more in the system during that last period.

11.16 For a MCQN with several types of customers following deterministic routes, devise static priority policies analog to FBFS and to LBFS, and prove their stability. This could be termed a *path priority policy*

Solution

We can concatenate all the buffers to form a single sequence in which we always keep buffers of the same route in their original order. Using FBFS or LBFS will then empty buffers in a given order, and the only difference from a re-entrant line is that by the time we reach a buffer to start working on it, it may have accumulated some input, and while we work on a buffer it may also still have exogenous inflow. if all $\rho_i < 1$, the proofs of stability for this new system remain valid.

11.17 Consider a two station three classes fluid re-entrant line, where $C_1 =$

$\{1,3\}$, $C_2 = \{2\}$, and assume that $\alpha = 1$, $m_1 + m_3 < 1$, $m_2 < 1$. Show it is stable under all work conserving HOL policies [Dai and Weiss (1996)].

Solution

We use piecewise linear Lyapunov functions as in Lemma 11.22. Let $\mathcal{W}_i(t) = \sum_{k \in C_i} Q_k(t)$ be the immediate workload of node i, and $Q_k^+(t) = \sum_{l=1}^{k} Q_l(t)$ be the total current workload at buffer k in the re-entrant line. Let $\theta = \frac{m_1}{m_1+m_3}$. Define:

$$G_1(t) = \theta Q_1^+(t) + (1-\theta) Q_3^+(t),$$
$$G_2(t) = Q_2^+(t)$$

On $\mathcal{W}_1(t) = 0$, $G_1(t) = (1-\theta)Q_2(t) < Q_2(t) = G_2(t)$, and on $\mathcal{W}_1(t) = 0$, $G_2(t) = Q_1(t) \le Q_1(t) + (1-\theta)Q_3(t) = G_1(t)$, which is condition (ii) of Lemma 11.22.

Next we consider $\mathcal{W}_1(t) > 0$. Recall that in a re-entrant line $Q_k^+(t) = Q_k^+(0) + t - \mu_k \mathcal{T}_k(t)$. Hence:

$$G_1(t) = G_1(0) + t - \theta\mu_1 \mathcal{T}_1(t) - (1-\theta)\mu_3 \mathcal{T}_3(t) = G_1(0) + t - (\mathcal{T}_1(t) + \mathcal{T}_3(t))/\rho_1.$$

But when $\mathcal{W}_1(t) > 0$ we have $(\dot{\mathcal{T}}_1(t) + \dot{\mathcal{T}}_3(t)) = 1$, so $\dot{G}_1(t) = 1 - \frac{1}{\rho_1} < 0$.

Similarly, if $\mathcal{W}_2(t) > 0$ then $\dot{G}_2(t) = 1 - \frac{1}{\rho_2} < 0$, so condition (i) of Lemma 11.22 holds.

Take $\epsilon = \min((\frac{1}{\rho_1} - 1), (\frac{1}{\rho_2} - 1))$, then by Lemma 11.22, $G(t) = 0$ for $t > G(0)/\epsilon$, and so the fluid model of the re-entrant line is stable.

11.18 Write down the fluid model equations for the Lu Kumar network, under the static priority policy of priority to buffers 2 and 4. Identify the properties of all the fluid solutions of these equations.

Solution

The fluid model equations are (with $\mu_0 = \alpha$, $\bar{\mathcal{T}}_0(t) = t$):

$$\bar{Q}_k(t) = \bar{Q}_k(t) + \mu_{k-1}\bar{\mathcal{T}}_{k-1}(t) - \mu_k \bar{\mathcal{T}}_k(t),$$
$$\bar{\mathcal{I}}_i(t) = t - \sum_{k \in C_i} \bar{\mathcal{T}}_k(t), \quad \int_0^t \Big(\sum_{k \in C_i} \bar{Q}_k(t) \Big) d\bar{\mathcal{I}}_i(t) = 0,$$
$$\int_0^t \bar{Q}_2(s) d\big(s - \bar{\mathcal{T}}_2(s)\big) = 0,$$
$$\int_0^t \bar{Q}_4(s) d\big(s - \bar{\mathcal{T}}_4(s)\big) = 0.$$

We now consider three situations:

(i) both $\bar{Q}_2(t) > 0$, $\bar{Q}_4(t) > 0$: Then:

$$\dot{\bar{Q}}_1(t) = \alpha,\ \dot{\bar{Q}}_2(t) = -\mu_2,\ \dot{\bar{Q}}_3(t) = \mu_2,\ \dot{\bar{Q}}_4(t) = -\mu_4,$$
$$\dot{\bar{\mathcal{I}}}_1(t) = \dot{\bar{\mathcal{I}}}_2(t) = 0,\ \dot{\bar{D}}(t) = \mu_4.$$

(ii) $\bar{Q}_4(t) = 0$, $\bar{Q}_1(t) > 0$, : Then:

$$\dot{\bar{Q}}_1(t) = \alpha - \mu_1,\ \dot{\bar{Q}}_2(t) = \mu_1 - \mu_2,\ \dot{\bar{Q}}_3(t) = \mu_2,$$
$$\dot{\bar{I}}_1(t) = \dot{\bar{I}}_2(t) = 0, \dot{\bar{D}}(t) = 0.$$

(iii) $\bar{Q}_4(t) = 0$, $\bar{Q}_1(t) = 0$, : Then:

$$\dot{\bar{Q}}_1(t) = 0,\ \dot{\bar{Q}}_2(t) = \alpha - \mu_2,\ \dot{\bar{Q}}_3(t) = \mu_2,$$
$$\dot{\bar{I}}_1(t) = 1 - \alpha/\mu_1,\ \dot{\bar{I}}_2(t) = 0, \dot{\bar{D}}(t) = 0.$$

(iv) $\bar{Q}_2(t) = 0$, $\bar{Q}_3(t) > 0$: Then:

$$\dot{\bar{Q}}_1(t) = \alpha,\ \dot{\bar{Q}}_3(t) = -\mu_3,\ \dot{\bar{Q}}_4(t) = \mu_3 - \mu_4,$$
$$\dot{\bar{I}}_1(t) = \dot{\bar{I}}_2(t) = 0, \dot{\bar{D}}(t) = \mu_4.$$

(v) $\bar{Q}_2(t) = 0$, $\bar{Q}_3(t) = 0$: Then:

$$\dot{\bar{Q}}_1(t) = \alpha,\ \dot{\bar{Q}}_4(t) = -\mu_4,$$
$$\dot{\bar{I}}_1(t) = 0,\ \dot{\bar{I}}_2(t) = 1,\ \dot{\bar{D}}(t) = \mu_4.$$

11.19 For the Lu-Kumar network with input rate α and initial fluid x in buffer 1, show that when $\rho_i < 1$ and $m_1 + m_2 > 1/\alpha$ then $m_2 > m_1$, $m_4 > m_3$, so that Figure 11.3 is correct. With $s_0 = 0$, calculate the values of t_1, t_2, t_3, s_1 and $\bar{Q}(t_1), \bar{Q}(t_2), \bar{Q}(t_3)\bar{Q}(s_1)$.

Solution

$$(1) \qquad m_1 + m_4 < 1/\alpha$$
$$(2) \qquad m_2 + m_3 < 1/\alpha$$
$$(3) \qquad m_2 + m_4 > 1/\alpha$$

Subtract (1) from (3) to get $m_2 - m_1 > 0$, and subtract (2) from (3) to get $m_4 - m_3 > 0$, so while machine 1 is emptying buffer 1 into buffer 2, buffer 2 is filling up at rate $1/m_1 - 1/m_2$, and while buffer 3 is emptying into buffer 4, buffer 4 is filling up at rate $1/m_3 - 1/m_4$.

We use $\mu_i = 1/m_i$. Buffer 1 is filling at rate α, and emptying at rate μ_1 and is empty at t_1. So $x + \alpha t_1 = \mu_1 t_1$ and $t_1 = \frac{x}{\mu_1 - \alpha}$.

At time t_1, buffer 1 is empty and all the fluid quantity $x + \alpha t_1 = \mu_1 t_1$ is in buffers 2 and 3. Buffer 3 has been filling up at rate μ_2, while buffer 2 has been filling at rate $\mu_1 - \mu_2$. So, buffer 2 contains $t_1(\mu_2 - \mu_1)$ and buffer 3 contains $t_1\mu_2$. we have:

$$t_1 = \frac{x}{\mu_1 - \alpha}, \qquad \bar{Q}(t_1) = \Big(0, \frac{(\mu_1 - \mu_2)x}{\mu_1 - \alpha}\frac{\mu_2 x}{\mu_1 - \alpha}, 0\Big).$$

At time t_2, buffers 1 and 2 are empty, and all the initial fluid and the fluid that

came in over $0, t_2$ is in buffer 3. Station 2 is working all that time on buffer 2, and the equation is $x + \alpha t_2 = t_2\mu_2$. All the fluid is now in buffer 3, so:

$$t_2 = \frac{x}{\mu_2 - \alpha}, \qquad \bar{Q}(t_2) = \Big(0, 0, \frac{\mu_2 x}{\mu_2 - \alpha}, 0\Big).$$

During (t_2, t_3) Buffer 3 is emptying at rate μ_3 with no further input, while buffer 4 has inflow at rate μ_3 and outflow at rate μ_4, so that it is filling at rate $\mu_3 - \mu_4$. Also, buffer 1 is filling up during all that time at rate α. So:

$$t_3 - t_2 = \frac{\mu_2 x}{\mu_2 - \alpha}\frac{1}{\mu_3}, \qquad \bar{Q}(t_3) = \Big(\frac{\mu_2 x}{\mu_2 - \alpha}\frac{\alpha}{\mu_3}, 0, 0, \frac{\mu_2 x}{\mu_2 - \alpha}\frac{\mu_3 - \mu_4}{\mu_3}\Big).$$

Next, during the entire interval (t_2, s_1) station 1 is working all the time on buffer 4, which is empty at the end, so $(s_1 - t_2)\mu_4 = \frac{\mu_2 x}{\mu_2 - \alpha}$. Also during this whole period buffer 1 is filling up, and we have:

$$s_1 - t_2 = \frac{\mu_2 x}{\mu_2 - \alpha}\frac{1}{\mu_4}, \qquad \bar{Q}(s_1) = \Big(\frac{\mu_2 x}{\mu_2 - \alpha}\frac{\alpha}{\mu_4}, 0, 0, 0\Big).$$

Reverting to $m_i = 1/\mu_i$,

$$s_1 = \frac{x}{\mu_2 - \alpha} + \frac{\mu_2 x}{\mu_2 - \alpha}\frac{1}{\mu_4} = \frac{(\mu_4 + \mu_2)x}{(\mu_2 - \alpha)\mu_4} = \frac{(m_2 + m_4)x}{1 - \alpha m_2},$$
$$\bar{Q}_1(s_1) = \frac{m_4}{1/\alpha - m_2}x > x.$$

Substituting $x = 1$, $\alpha = 1$ we verify the values in the text:

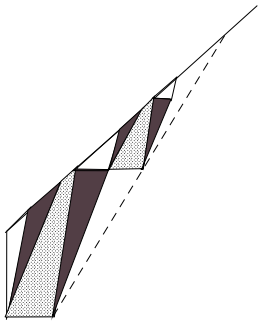
$$s_1 = s_0 + \frac{m_2 + m_4}{1 - m_2}, \quad \bar{Q}(s_1) = \Big(\frac{m_4}{1 - m_2}, 0, 0, 0\Big).$$

We also calculate the time for the backward interpolation to reach 0:

$$t^* = \frac{t^*}{1} = \frac{s_1 - s_0}{\bar{Q}_1(s_1) - \bar{Q}_1(s_0)} = \frac{m_2 + m_4}{1 - m_2} \Big/ \Big(\frac{m_4}{1 - m_2} - 1\Big) = \frac{m_2 + m_4}{m_2 + m_4 - 1}$$

11.20 Plot a fluid solution for the Lu-Kumar network when $\rho_i < 1$ and $m_1 + m_2 < 1/\alpha$, for some initial fluid in buffer 1, and priority to buffers 2 and 4.

Solution



11.21 For the Lu-Kumar network, verify that the function G satisfies properties (i) and (ii) necessary for the piecewise linear Lyapunov function to prove stability, and that the linear program (11.14) is feasible with $\epsilon > 0$, if $m_1 + m_4 < 1$, $m_2 + m_3 < 1$, $m_2 + m_4 < 1$ [Dai and Weiss (1996)].

Solution

We see that $\mathcal{W}_1 = 0$ implies $G_1 \le G_2$ and that $\mathcal{W}_2 = 0$ implies $G_2 \le G_1$ if $\theta_2 \le \theta_1$.

For $\mathcal{W}_1(t) > 0$, $\dot{\mathcal{T}}_1(t) + \dot{\mathcal{T}}_4(t) = 1$, and we get if $m_1 \le \theta_1 \le 1 - m_4$ then:

$$\begin{aligned}\dot{G}_1(t) &= 1 - \theta_1\mu_1\dot{\mathcal{T}}_1(t) - (1-\theta_1)\mu_4\dot{\mathcal{T}}_4(t)\\ &= (1-\theta_1\mu_1)\dot{\mathcal{T}}_1(t) + (1-(1-\theta_1)\mu_4)\dot{\mathcal{T}}_4(t)\\ &\le -\min\Big(\theta_1\mu_1 - 1), ((1-\theta_4)\mu_4 - 1)\Big) < 0.\end{aligned}$$

Similarly, for $\mathcal{W}_2(t) > 0$, $\dot{\mathcal{T}}_2(t) + \dot{\mathcal{T}}_3(t) = 1$, and if we choose $m_2 \le \theta_1 \le 1 - m_3$,

$$\dot{G}_1(t) \le -\min\Big(\theta_2\mu_2 - 1), ((1-\theta_3)\mu_3 - 1)\Big) < 0.$$

Clearly, any solution of the LP (11.14) will satisfy all the requirements for $G = \max(G_1, G_2)$.

Finally, if the conditions $m_1 + m_4 < 1$, $m_2 + m_3 < 1$, $m_2 + m_4 < 1$ hold then

$$\begin{gathered}\delta_1 = (1 - m_1 - m_4)/2, \quad \delta_2 = (1 - m_2 - m_3)/2, \quad \delta_3 = (1 - m_2 - m_4)/3,\\ \theta_1 = 1 - m_4 - \min(\delta_1, \delta_3), \qquad \theta_2 = m_2 + \min(\delta_2, \delta_3),\\ \epsilon = \min(\delta_1, \delta_2, \delta_3).\end{gathered}$$

is a feasible solution as required.

11.22 Consider a KSRS network with input rates α_1, α_2 and mean service times m_1, m_2, m_3, m_4. Show that the conditions:

$$\alpha_1 m_1 + \alpha_2 m_4 < 1, \quad \alpha_1 m_2 + \alpha_2 m_3 < 1, \quad \alpha_1 m_4 + \alpha_2 m_2 < 1,$$

define a global stability region for the network [Botvich and Zamyatin (1992)].

Solution

Since we are dealing with fluids, we can change the units of flow for each of the items, and get $\tilde{\alpha_1} = 1$, $\tilde{\alpha_2} = 1$, $\tilde{\mu_1} = \mu_1/\alpha_1$ etc., so W.L.G. we can assume both arrival rates equal 1. Furthermore, the fluid model of the KSRS with arrival rates 1 is equivalent to the fluid model for the Lu-Kumar network with input rate 1.

Formally, without the rescaling, we can use $G(t) = \max(G_1(t), G_2(t))$:

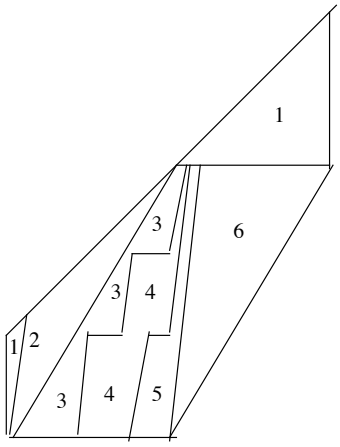
$$\begin{aligned}G_1(t) &= \theta_1 \frac{Q_1^+(t)}{\alpha_1} + (1-\theta_1)\frac{Q_4^+(t)}{\alpha_2},\\ G_2(t) &= \theta_2 \frac{Q_2^+(t)}{\alpha_1} + (1-\theta_2)\frac{Q_3^+(t)}{\alpha_2}.\end{aligned}$$

and the verification is similar to the proof for the Lu-Kumar network.

11.23 Consider the example of the re-entrant line that is unstable under FCFS, of Section 10.2.2. Plot an unstable fluid solution for it.

Solution

The figure below is the fluid picture.



Start with $\bar{Q}_1(t) = 1$ the rest empty, and this is fluid. Let S_1 be the time that buffer 1 empties for the first time. Let S_2 be the time that all fluid present in machine 2 at time S_1 has been served. Let similarly $S_3, S_4, \ldots$ be the times S_l when all the fluid present in machine 2 at time S_{l-1} has been served. By FCFS machine 2 is working on the fluid present at S_{l-1} only, up to S_l, Note that in the interval (S_{l-1}, S_l) no item is served twice in machine 2. What happens is that items in buffer $j = 2, 3, 4$, move to buffer $j + 1$, while buffer 1 continues to feed buffer 2.

Let $c^{-1} \approx 0.9 < 1$ be processing rates at buffers 2,6 and $\delta^{-1} = 0.001$ small be processing rates at buffers 1,3,4,5. The intervals (S_{l-1}, S_l) are of length c^l. At time S_j, fluid in machine 2 is as follows (for convenience we now take $\delta^{-1} = 0$ in the calculations, but these are approximately also the values for $\delta^{-1} = 0.001$):

$$\bar{Q}_2(S_1) = 1,\ \bar{Q}_3(S_1) = \bar{Q}_4(S_1) = \bar{Q}_5(S_1) = 0$$
$$\bar{Q}_2(S_2) = c,\ \bar{Q}_3(S_1) = 1,\ \bar{Q}_4(S_1) = \bar{Q}_5(S_1) = 0$$
$$\bar{Q}_2(S_3) = c^2,\ \bar{Q}_3(S_1) = c,\ \bar{Q}_4(S_1) = 1,\ \bar{Q}_5(S_1) = 0$$
$$\bar{Q}_2(S_4) = c^3,\ \bar{Q}_3(S_1) = c^2,\ \bar{Q}_4(S_1) = c,\ \bar{Q}_5(S_1) = 1$$

In all that time, machine 1 is empty. Next:

$$\bar{Q}_2(S_5) < c^4,\ \bar{Q}_3(S_1) = c^3,\ \bar{Q}_4(S_1) = c^2,\ \bar{Q}_5(S_1) = c$$

Then at time S_5 we have: $\bar{Q}_6(S_5) = 1$, and since still $\bar{Q}_1(S_5) \approx 0$, input to buffer 2 will almost stop, and machine 1 will now work on buffer 6 exclusively. Ashort time later, buffer 6 will contain $\approx 1 + c + c^2 + c^3$. and by the time it empties, buffer 1 will contain more fluid than 1.

11.24 (∗) *HLPPS policy* Under head of the line proportional processor sharing policy, each station is splitting its service capacity between classes in proportion to the number of customers present, and is then serving the head of the line of each class. Show that a fluid MCQN with $\rho < 1$ is stable under HLPPS [Bramson (1996, 2008)].

Solution

The proof for this theorem is quite involved and requires studying one of the papers [Bramson (1996, 2008)].

11.25 (∗) *Early due date policy* Consider the policy in which each customer receives a due date upon arrival, and the policy is to give priority to the customer with the earliest due date (EDD). Show that MCQN under EDD policy is stable [Bramson (2001, 2008)].

Solution

The proof for this theorem is quite involved and requires studying one of the papers [Bramson (2001, 2008)].

11.26 (∗) *Kelly-type networks under FCFS:* Consider a Kelly-type network as in Section 8.9, with general renewal arrivals and general processing times of rate μ_i at node i. Show that with $\rho_i < 1$ it is stable under FCFS [Bramson (1996, 2008)].

Solution

The proof for this theorem is quite involved and requires studying one of the papers [Bramson (1996, 2008)].

12

Processing Networks and Maximum Pressure Policies

Exercises

12.1 Show that there is always a feasible extreme allocation, i.e. if $\mathscr{A} \neq \emptyset$, then $\mathscr{E}(t)$ is not empty at any t.

Solution

We need to satisfy:

$$\sum_j \mathscr{A}_{i,j} = 1, \qquad \text{for every input source } i,$$

$$\sum_j \mathscr{A}_{i,j} \leq 1, \qquad \text{for every internal processor } i,$$

Since $\mathscr{A}$ is not empty, the first set of constraint is feasible, its solutions are a convex set, with feasible extreme points, and because buffer 0 is never empty those allocations to input activities are available. For the other activities idling is always available, and the combination of the extreme allocations to the input activities and idling for the other activities is an extreme allocation.

12.2 Show that $\mathscr{E}(t)$ defined as the extreme points of the available allocations $\mathscr{A}(t)$, are also extreme points of all allocations, $\mathscr{A}$, i.e. $\mathscr{E}(t) \subseteq \mathscr{E}$.

Solution

Assume $a \in \mathscr{E}(t)$. Assume it is not an extreme point of $\mathscr{A}$, then it can be written as a convex combination of extreme points $a^{(k)} \in \mathscr{A}$. But all 0 coordinates of a will be 0 in each of $a^{(k)}$. It follows that $a^{(k)} \in \mathscr{A}(t)$ contradicting the extremality of a in $\mathscr{A}(t)$.

12.3 Formulate and verify the analog of Theorems 11.10, 11.12 for processing networks.

Solution

Theorem. If the fluid limit model of a processing network under some fixed policy is weakly stable, then the processing network is rate stable, in the sense that starting with any fixed $Q(0)$, one has $\lim_{t\to\infty} Q(t)/t = 0$.

The proof is exactly the same as the proof for Theorem 11>10

Theorem. (a) If the fluid limit model of a processing network under some

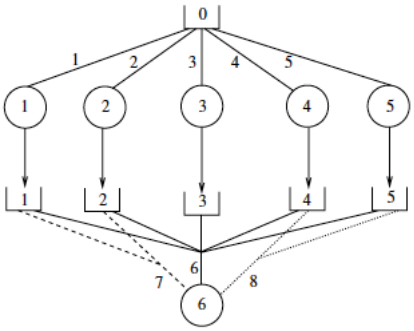
fixed policy is unstable, then for any initial fixed state $Q(0) = x$ and every $\omega \in \mathfrak{G}$, $\liminf_{t\to\infty} Q(t)/t > 0$.
(b) If $\rho > 1$ in the solution of the LP (12.7) then every fluid limit is unstable, and the fluid limit model is unstable, and the processing network is unstable in the stronger sense that there exists positive c such that $\liminf_{t\to\infty} |Q(t)|/t \geq c$.

Proof The proof of (a) is just as for Exercise 11.6.
The proof of (b) needs to be modified from the proof of Exercise 11.6: Recall that $\dot{\mathcal{T}}_j(t)$ needs to satisfy all the constraints on x_j, for every activity j. If $\rho > 1$ then there must be at least one service processor i for which $\sum_{j\in\mathcal{J}} A_{i,j}x_j \geq \rho > 0$ for every feasible solutions.
We assume $\rho_i > 1$ for some station. We argue now for any fluid limit that:

$$CR^{-1}\left(\bar{Q}(t) - \bar{Q}(0)\right) = CR^{-1}\left(\alpha t - (I - P^{\intercal})\bar{\mathcal{D}}(t)\right) = \boldsymbol{\rho} t - C\bar{\mathcal{T}}(t).$$

But $C\bar{\mathcal{T}}(t) \leq t$ componentwise, and therefore there exists $c > 0$ so that for every fluid limit with $\bar{Q}(0) = 0$, $|\bar{Q}(t)| \geq ct$. The same argument as for (a) then leads to $\liminf_{t\to\infty} |Q(t)|/t \geq c$. □

12.4 Show that the following processing network (taken from Dai and Lin (2005)) with the following data has $\rho < 1$, and is stable under some allocation policy, but it is unstable under maximum pressure policy, because it does not satisfy EAA assumptions.



In this network circles are processors, open boxes are buffers, lines represent activities. Activities 1–5 are input activities each with its own input processor. Processor 6 is a service processor with activities 6,7,8. Activity 6 processes an item from each buffer, activity 7 (activity 8) processes an item from buffers 1,2 (buffers 4,5). Each processed item leaves the system. Activity durations are deterministic:

$$\upsilon_1(\ell) = \upsilon_2(\ell) = \upsilon_3(\ell) = 2,\ \ell \geq 1, \quad \upsilon_4(1) = \upsilon_5(1) = 1,\ \upsilon_4(\ell) = \upsilon_5(\ell) = 2,\ \ell \geq 2,$$
$$\upsilon_6(\ell) = \upsilon_7(\ell) = \upsilon_8(\ell) = 1,\ \ell \geq 1, \qquad \mu_i = \begin{cases} 0.5, & i = 1,\ldots,5, \\ 1, & i = 6,7,8. \end{cases}$$

Solution

It is easily checked that an optimal solution to the LP (12.7) has $\rho < 1$, $x_6 = 1/2$, and indeed, using only activity 6, starting processing at $1, 3, 5, \ldots$ will be a stable policy. The only extreme allocations for activities 6,7,8 are: $a^1 = (1,0,0)$, $a^2 = (0,1,0)$, $a^3 = (0,0,1)$, $a^4 = (0,0,0)$. Starting with $Q(0) = 0$, only a^4 is feasible. at time 1 $Q_4(1) = Q_5(1) = 1$ while the other buffers are empty, so only a^3, a^4 are feasible, and a^3 is the max pressure policy, then at time 2 a^2 is max pressure, and so on. Allocation a^1 is never max pressure, so activity 6 is never used, and $Q_3(t) \sim t/2 \to \infty$. This can also be shown for parametrized maximum pressure.

We now show that for this processing network EAA fails: Take $Q(t) = (0,0,1,0,0)$. Then the unique maximum pressure allocation is a^1, i.e. use activity 6. But this activity requires items from all the buffers so it is not available for $Q(t)$.

12.5 Show that in a MCQN, the following conditions are equivalent:

- There exists $x \geq 0$ such that $Rx > 0$ in all components.
- The routing matrix P has spectral radius < 1.

Solution

Recall for MCQN, $R = (I - P^{\intercal})\begin{pmatrix} \ddots & & \\ & \mu & \\ & & \ddots \end{pmatrix}$.

(i) If P has spectral radius < 1 then $(I - P)$ is invertible and $(I - P)^{-1} = I + P + P^2 + \cdots \geq 0$, and furthermore, $x = R^{-1}\mathbf{1}$ has all components > 0, and $Rx = \mathbf{1} > 0$.

(ii) Assume for some $x \geq 0$, $(I - P)x > 0$ in all coordinates. This implies actually that $x > 0$ in all coordinates. Then we have: $(I - P)x > 0$, so $x_i > (Px)_i$, so $(Px)_i/x_i < 1$ all i. By the "Min-max" Collatz-Wielandt formula, define $g(x) = \max_i\{(Px)_i/x_i\}$, then: $\min_{y>0} g(y) = r$ where r is the spectral radius. So we have: $1 > g(x) \geq \min_{y>0} g(y) = r$.

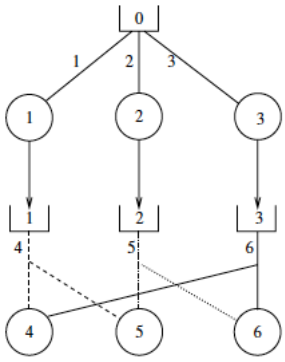
12.6 Show that the set of $\{0,1\}$ allocations is composed only of extreme points of $\mathscr{A}$, i.e. $\mathscr{N} \subseteq \mathscr{E}$.

Solution

Recall the definition of an extreme point of convex set: it cannot be on an interior point of an interval contained in the set, i.e. it is not of the form $\alpha u + (1-\alpha)v$ for u, v, in the set, with $u \neq v$ and $0 < \alpha < 1$.

Clearly, if $1 = \alpha x + (1-\alpha)y$, $x \neq y$, then one of x or y is > 1. Similarly, if $0 = \alpha x + (1-\alpha)y$, $x \neq y$, then one of x or y is < 0. Hence, if $a \in \mathscr{A}$ is a $\{0,1\}$ allocation then for any vectors $x \neq y$ such that $a = \alpha x + (1-\alpha)y$, either x has some > 1 or < 0 components, or y has some > 1 or < 0 components, so either x or y is not in $\mathscr{A}$. Hence a is an extreme point of $\mathscr{A}$.

12.7 Show that the following processing network (taken from Dai and Lin (2005)) with the following data has $\rho < 1$, and is stable under maximum pressure policy which allows processor splitting and preemptions, but is not stable under non-splitting allocations.



In this network circles are processors, open boxes are buffers, lines represent activities. Activities 1–3 are input activities each with its own input processor. Activities 4,5,6 are service activities, each uses two processors, in order to process one item out of buffers 1,2,3. The input activities have processing rates $\mu_1 = \mu_2 = \mu_3 = 0.4$, the service activities have processing rates $\mu_4 = \mu_5 = \mu_6 = 1$.

Solution

For the processing part of the network the input output matrix and the resource consumption matrices are:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$\rho = 0.8$, $x = (0.4, 0.4, 0.4)$ solves LP 12.7, and the system is strictly Leontief, so maximum pressure allowing splitting is rate stable.

The extreme allocations are:

$$(0,0,0), \quad (1,0,0), \quad (0,1,0), \quad (0,0,1), \quad (1/2,1/2,1/2).$$

The system is stable under $(1/2, 1/2, 1/2)$. However, under any non-splitting allocation only one activity can be performed at any time, so the maximum total departure rate is at most 1, while the input rate is 1.2, so the system cannot be stable without processor splitting.

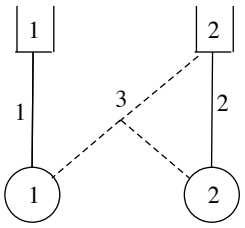
12.8 A processing network is reverse Leontief if every activity is using only a single processor. Show that for reverse Leontief networks, all extreme allocations are integer, i.e. processors are not split, with each allocation $a_j \in \{0, 1\}$, $j = 1, \ldots, J$.

Solution

In a network that is reverse Leontief, all the columns of the resource consumption matrix are unit columns. The columns that correspond to slacks in the allocation constraints (for serivce activities) are also unit columns. Extreme points correspond to basic solutions of the constraints, but the columns

of every basis form a unit matrix, so their inverse is a unit matrix and the allocation then consists of 0's and 1's (including the slacks). So each allocation is integer.

12.9 Consider the following processing network (taken from Dai and Lin (2005)) with the following data:



In this network input processors are not included, circles are service processors, open boxes are buffers, lines represent activities. Arrivals to buffer 1 occur at times $0.5 + 2n$, $n = 1, 2, \ldots$, Arrivals to buffer 2 occur at times $1 + 1.5n$, $n = 1, 2, \ldots$. Activity 1 uses processor 1 and serves buffer 1, with $m_1 = 1$. Activity 2 uses processor 2 and serves buffer 2, with $m_2 = 2$. Activity 3 uses both processors, and serves buffer 2, with $m_3 = 1$. All processing times are deterministic, and the system is empty at time 0.

Solve the static planning problem for this network to calculate $\rho < 1$, explain why EAA holds. Show that using processor splitting non-preemptive maximum pressure policy the system diverges. Describe how the system operates under preemptive processor splitting policy.

Solution

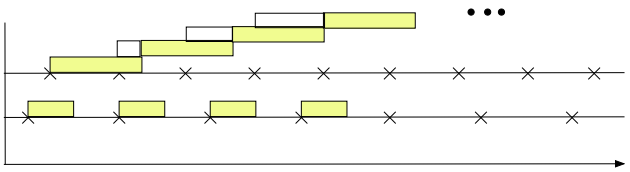
The static planning problem is:

$$\begin{array}{llllll} \min & \rho & & & & \\ s.t. & x_1 & & & & = \frac{1}{2} \\ & & \frac{1}{2}x_2 & +x_3 & & = \frac{2}{3}, \\ & x_1 & & +x_3 & & \leq \rho, \\ & & x_2 & +x_3 & & \leq \rho, \\ & & x_i & & \geq 0. & \end{array}$$

Solution, rate stable: $\rho = \frac{11}{12}$, $x_1 = \frac{1}{2}$, $x_2 = \frac{1}{2}$, $x_3 = \frac{5}{12}$, with both processors tight at this ρ.

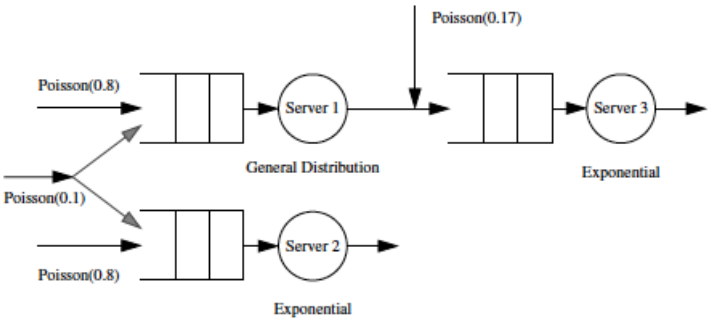
The network is strictly Leontief, since every activity processes a single buffer. Hence, maximum pressure policy with preemption and processor splitting is rate stable.

We now track the performance of maximum pressure policy that allows processor splitting but does not allow preemptions. The following figure shows the processing at the two buffers. We see the at times that jobs arrive to buffer 1, buffer 1 is empty, and activity 2 is serving buffer 2, so the maximum pressure policy is allocation $(1, 1, 0)$ and activity 1 is used to



serivce buffer 1 in the intervals $(1/2+2n, 3/2+2n)$. Buffer 2 is never empty, and when activity 2 completes a service at time $1+2n$, activity 3 is never available since activity 1 is active, and so again the non-preemptive allocation is $(1,1,0)$. Clearly, buffer 2 is accumulating jobs at rate $2/3 - 1/2$, and the system is unstable.

12.10 Consider the following network (taken from Dai and Lin (2005)). It is modeled as a processing network with 3 service buffers, each with its own processor and service activity, and with four input processors, one of which has two input activities, which provide discretionary routing to buffers 1 or 2. All input processors have exponential service times (Poisson inputs) with



the rates indicated in the figure. Processing for the service activities are rate 1, the processing times of servers 2, 3, are exponential, while processing time of buffer 1 is generally distributed. Explain or show the following properties:
(i) Under HOL policy, and join the shortest queue for the discretionary routing, the system is unstable if buffer 1 processing times are exponential, but it is stable if it is hyper-exponential with large enough variability.
(ii) It is stable under round Robin policy that directs a fraction of 90% of the discretionary input to buffer 1, but if the input rates are changed slightly, we will need to change the fraction routed to buffer 1.
(iii) Formulate the static planning problem and show that $\rho < 1$. Find how maximum pressure is implemented here and check that it is stable whenever $\rho < 1$.

Solution

(i) If processing at buffer 1 is exponential, buffers 1 and 2 are homogenous,

and so input from join the shortest queue is divided equally between the two buffers, so the input rate to buffer 3 is $0.8 + 0.05 + 0.17 = 1.02 > 1$ so buffer 3 will diverge. If buffer 1 processing is hyper-exponential, the variability of processing times is greater, so the queue length at buffer 1 tends to be longer than at buffer 2, and a larger fraction of join the shortest queue customers are directed to buffer 2, and the input rate to buffer 3 decreases. This leads to stability.

(ii) With round robin routing as described this is close to a generalized Jackson network (the inputs to buffers 1 and 2 are not independent, but we can calculate their correlations). With the given rates it is stable.

(iii) This network viewed as a processing network has 4 input processors, and the processor with the discretionary input processor has two activities, corresponding to routing to buffer 1 or 2. This is a unitary network: each activity uses a single processor, and serves a single buffer. Hence, maximum pressure policy is rate stable, irrespective of processing time distributions, for any input rates as long as $\rho < 1$ in the LP 12.7. Maximum pressure policy works as follows: Buffers 2 and 3 use HOL non-idling service. Routing input activities choose at any time to route to the shorter of the queues at buffers 1 and 2. Finally, at buffer 1, buffer 1 idles when $Q_3(t) > Q_1(t)$. With this policy, the queue at buffer 3 is controlled and does not diverge, and the queue at buffer 1 grows when buffer 3 gets more congested, which diverts input to the shorter queue at buffer 2.

12.11 Model the input queued switch as a processing network, and show that the maximum weight policy is in fact maximum pressure for this processing network.

Solution

There is one input buffer 0, and one service buffer $k = (i, j)$ for each pair of input output ports. There is one input processor and one input activity for each of the service buffers i.e. for each of the Q_{ij} input queues. An activation of input activity $k = (i, j)$ introduces a number of packets into buffer (i, j) in a single time slot, where inputs are a stationary ergodic sequence, with mean $\lambda_{i,j}$. Service processors are the input ports i and the output ports j. There is one service activity for each service buffer $k = (i, j)$, and it is using two processors, the input processor i and the output processor j. Each activation of service activity $k = (i, j)$ lasts one time slot, and removes one packet from buffer (i, j) if $Q_{i,j} > 0$.

Clearly for all input processors and activities, they work at all times, at rates λ_{ij}. For service activity (i, j) on buffer (i, j), each application takes one time unit, and therefore has rate 1, and item from buffer (i, j) immediately leaves the system, so the elements of the input-output matrix R are $R_{(i,j),(i,j)} = 1$, and R is in fact an $N^2 \times N^2$ identity matrix, and so $Q^{\mathsf{T}}R = Q$, and the pressure for allocation π is $Q^{\mathsf{T}}R\pi = \sum_{i,j} \pi_{i,j} Q_{i,j}(r) = f_\pi(r)$.

12.12 Show that the conditions (12.14) are exactly the conditions that the static planning problem (12.7) has solution $\rho \leq 1$.

Solution

Let $y_{i,j}$ denote the unknown allocations to input activity (i, j), and $x_{i,j}$ denote the unknown allocations to service activity (i, j). The $N^2 \times 2N^2$ input output matrix is of the form $R = [-\Lambda; I]$ with Λ the diagonal matrix of $\lambda_{i,j}$ elements. The resource constrains for the input processors result in $y_{i,j} = 1$ and the constraint $R \begin{bmatrix} y \\ x \end{bmatrix} = 0$ results in $x_{i,j} = \lambda_{i,j}$, so the constraints of the input processors $i = 1, \ldots, N$, and on the output processors $j = 1, \ldots, N$ are then:

$$\sum_i \lambda_{i,j} \leq \rho, \qquad \sum_j \lambda_{i,j} \leq \rho.$$

and so $\rho \leq 1$ if and only if (12.14) hold.

12.13 Show that for the processing network modeling the input queued switch, $\mathscr{E} = \mathscr{N}$.

Solution

The set $\mathscr{A}$ consists of all the doubly sub-stochastic matrices, and all the constraints of $\mathscr{A}$ are tight exactly for doubly stochastic matrices. The following is the basic theorem about doubly stochastic matrices: Every doubly stochastic matrix is a convex combination of permutation matrices, i.e. the extreme points are exactly permutation matrices. The maximization of the pressure is always an assignment problem, and assignment problems have $\{0, 1\}$ solutions, in fact the optimum is alway a permutation.

12.14 $(*)$ Discuss solution of the pressure maximization as an assignment problem, using the Hungarian method, and using the transportation simplex algorithm, and examine their online implementation using efficient use of previous slot solution for next slot solution.

Solution

(i) Hungarian method:

We wish to find at each time slot the assignment π that will maximize $f_\pi(r) = \langle \pi, Q(r) \rangle = \sum_{i,j} \pi_{i,j} Q_{i,j}(r)$. The Hungarian method actually minimizes assignment costs, so one should use $a_{i,r} = q_{\max} - Q_{i,j}$ with $q_{\max}$ the value of the longest queue. One then uses the Hungarian method to obtain the cheapest assignment for the costs $a_{i,j}$.

Use the link `https://en.wikipedia.org/wiki/Hungarian_algorithm` for a tutorial on the Hungarian method.

Once you obtain the best assignment at time slot $t - 1$, what you have is a matrix with a single zero in each row and in each column, which is the best assignment for time slot $t - 1$, and all other entries are residual costs, $\tilde{a}_{i,r}$. To go to time slot t, let $\delta Q_{i,r} = A_{i,r}(t) - \pi_{i,j}(t - 1)$, where $A_{i,r}(t)$ are the new arrivals. Let $b_{i,r} = q_{\max} - \delta Q_{i,r}$ where now $q_{\max} =$

$\max\{\delta Q_{i,r}\} - \min\{\delta Q_{i,r}\}$. Now re-solve the assignment problem using the Hungarian method on $a_{i,r} = \tilde{a}_{i,r} + b_{i,r}$.
(ii) Transportation problem. Use the optimal solution of the transportation problem for time slot $t-1$, and then modify the vector of costs by adding $\delta Q_{i,r}(t)$ to the current costs.

12.15 Show that for a network of switches, the conditions (12.17) are exactly the conditions that the static planning problem (12.7) is feasible with $\rho \leq 1$.

Solution

The conditions $Rx = 0$ of (12.7) are simply that $x_(c, l_1, l_2) = \alpha_c$. The conditions $A_{\text{input}} = 1$ are satisfied automatically since input processor for c have single activity that is active all the time and produces inflow at rate α_c. The constraints $A_{\text{service}} \leq \rho \leq 1$ for every the processing at every switch, as given by (12.17) are exactly the same as for (12.14) summarized over the flows c that go through the switch, which are exactly (12.7).

12.16 Prove proposition12.17.

Solution

We model the system as a processing network. We saw in the previous exercise that if (12.17) holds the than the the static planning problem (12.7) is feasible with $\rho < 1$. The processing network is strictly Leontief, so EAA assumption holds. Further more, the extreme allocations are all integer, and decisions are at slotted time points so preemptions are not an issue. Hence the system is rate stable under any non-splitting, non-preemptive maximum pressure policy.

12.17 Prove proposition 12.18.

Solution

In the description of the system as a processing network there is a single infinite input buffer, an input processor and corresponding input activity to provide input of packets of type c at rate α_c. Type c has route steps $s = 1, \ldots, S_c$ that go through switches $i_1, \ldots, i_s, \ldots, i_{S_c}$, and classes/buffers/queues are labeled (c, i_s, ℓ_1, ℓ_2), where ℓ_1, ℓ_2 are the input and output ports of switch i_s that the route of c uses on step s. Recall that each node is only visited at most once by the route of type c. There is a single service activity labeled also by (c, i_s, ℓ_1, ℓ_2) for each class/buffer/queue, and it is using two service processors: input port $(i_s, \ell_1, \cdot)$ and output port $(i_s, \cdot, \ell_2)$. Each activation of activity $(c, i_s.\ell_1, \ell_2)$ moves 1 item out of (c, i_s, ℓ_1, ℓ_2) and into $(c, i_{s+1}, \ell_1, \ell_2)$. So the input output matrix for the service buffers and activities has 1 in the main diagonal positions, and -1 in the diagonal below the main diagonal for each of the steps of the route of type c. It follows that for activity $j = (c, i_s, \ell_1, \ell_2)$, the element of $Q^\intercal R$ is:

$$(Q^\intercal(t)R)_{(c,i_s,\ell_1,\ell_2)} = Q_{(c,i_s,\ell_1,\ell_2)}(t) - Q_{(c,i_{s+1},\ell_1,\ell_2)}(t),$$

The set of extreme allocations includes for each switch a permutation matching of input and output switches, where for each match of input and output

switches ℓ_1, ℓ_2 one can choose which of the types c that goes through that pair of switches to use. Denote:

$$H_i(c, l_1, l_2) = \begin{cases} 1 & \text{if flow } c \text{ uses pair } (l_1, l_2) \text{ of one of the switches,} \\ 0 & \text{otherwise.} \end{cases}$$

and let

$$Z_{i,\ell_1,\ell_2} = \max \left\{ Q_{(c,i_s,\ell_1,\ell_2)}(t) - Q_{(c,i_{s+1},\ell_1,\ell_2)}(t) \middle| c : H_i(c, l_1, l_2) = 1 \right\}.$$

Then the pressure is maximized by choosing at switch i the permutation that maximizes the sum of these Z_{i,ℓ_1,ℓ_2}.

13

Processing networks with Infinite Virtual Queues

Exercises

13.1 For the two node network with IVQ's of Section 13.2.1, what are the marginal stationary distributions of each of the standard queues.

Solution

Each of the processors is working all the time, and each node provides continuous input, at rate $\mu_i(1-p_i)$ to the other node. Processing of the standard queue at node i is at rate μ_i whenever it is not empty, since the standard queue has preemptive priority, and customers of the standard queue that complete processing leave the queue. So standard queue 1 operates like an M/M/1 queue with arrival rate $\mu_2(1-p_2)$, and processing rate μ_1, and with $\rho_1 = \frac{\mu_2(1-p_2)}{\mu_1}$, it has stationary distribution $\mathbb{P}(Q_1 = k) = (1-\rho_1)\rho_1^k$. Condition for stability is $\rho_1 < 1$. Queue 2 is similar.

13.2 (∗) For the two node network with IVQ's of Section 13.2.1, derive the two dimensional stationary distribution, using the compensation method of section 15.1 [Adan and Weiss (2005)].

Solution

For the complete derivation see Adan and Weiss (2005).

13.3 The following is a generalization of the two node IVQ network of Section 13.2.1. Consider a Jackson network, with nodes $i \in \mathcal{I}$, with exogenous input rates α_i, service rates μ_i, and routing probabilities p_{ij}. Assume that the nodes are partitioned into standard nodes $\mathcal{I}_0$, and IVQ nodes $\mathcal{I}_\infty$. each node $i \in \mathcal{I}_\infty$ has in addition to the queue of items received from outside or from other nodes, also an infinite supply of work. At the IVQ nodes there is preemptive priority to customers that arrive from outside or from other nodes. However when the queue of such customers is empty, the IVQ node serves customers from its infinite buffer [Weiss (2005)].

(a) Find the flow rates, and conditions for stability for this system.

(b) Show that queue lengths at the standard nodes have a product form joint stationary distribution.

(c) Find the stationary marginal distribution of the queue lengths at the IVQ nodes, and show that they have Poisson input and output.

Solution

(a) The traffic equations are:

$$\lambda = \alpha + P\nu$$

where λ is the inflow rates, and ν is the outflow rates, and $\nu_i = \lambda_i$, $i \in I_0$, while $\nu_i = \mu_i$, $i \in I_\infty$. The solution of these equations provides us with the inflow and outflow rates from the standard queues, and the inflow rates for the IVQs. The condition for stability is that $\lambda_i < \mu_i$, $i \in I$.
(b) The main feature here is that nodes $i \in I_\infty$ work all the time and produce an outflow which is Poisson rate μ_i independent of all else. So the set of nodes $j \in I_0$ act like a jackson network, where the exogenous input to node j is Poisson with rate $\alpha_j + \sum_{i \in I_\infty} \mu_i p_{i,j}$. So, once we obtain λ_j, $j \in I_0$ and calculate $\rho_j = \frac{\lambda_j}{mu_j}$ the stationary distribution of this part of the network is:

$$\pi(n_j : j \in I_0) = \prod_{:j \in I_0} (1 - \rho_j)\rho_j^{n_j}.$$

(c) The outflow form each of the nodes $j \in I_\infty$ is Poisson rate μ_j. The inflow to node $j \in I_\infty$ consists of Poisson flows at rates $\mu_i p_{i,j}$ for $i \in I_\infty$, which are independent of the rest of the system, and from the flows that exit the Jackson part of the system. But output from a Jackson network consists of independent Poisson processes, of rates $\lambda_i(1 - \sum_{k \in I_0} p_{i,k})$, for $i \in I_0$, and so the input to node $j \in I_\infty$ from node $i \in I_0$ is a Poisson process, independent of all other inputs to node j, and of rate $\lambda_i p_{i,j}$. Altogether, by the traffic equations, input to each node $i \in I_\infty$ is Poisson rate λ_i and the outflow is Poisson rate μ_i.
Remark The joint distribution of the standard queue at the IVQ nodes is more complex, as the two node example of Section 13.2.1 shows.

13.4 For the 3 buffer re-entrant line with IVQ of Section 13.2.2, write down the balance equations for the random walk describing the queue length of the standard queues, and derive the stationary distribution of $(Q_2(t), Q_3(t))$ [Adan and Weiss (2006)].
Solution
The balance equations are:

$$(\mu_2 + \mu_3)P(n_2, n_3) = \mu_3 P(m_2, n_3 + 1) + \mu_2 P(n_2 + 1, n_3 - 1),\ n_2, n_3 > 0,$$
$$(\mu_1 + \mu_2)P(n_2, 0) = \mu_3 P(n_2, 1) + \mu_1 P(n_2 - 1, 0),\ n_2 > 0,$$
$$\mu_3 P(0, n_3) = \mu_3 P(0, n_3 + 1) + \mu_2 P(1, n_3 - 1)),\ n_3 > 0,$$
$$\mu_1 P(0, 0) = \mu_3 P(0, 1).$$

We use trial solution $P(n_1, n_2) = C\alpha_3^{n_2}\alpha_3^{n_3}$ and obtain from the first two equations:

$$(\mu_2 + \mu_3)\alpha_3 = \mu_3\alpha_3^2 + \mu_2\alpha_3,$$
$$(\mu_1 + \mu_2)\alpha_2 = \mu_3\alpha_3 + \mu_1.$$

We then have from the second equation

$$\alpha_2 = \frac{\mu_1}{\mu_1 + \mu_2 - \mu_3\alpha_3},$$

and substituting into the first we obtain a cubic equation for α_3. One root is $\alpha_3 = \frac{\mu_2}{\mu_3}$, with two other real roots, one smaller and one larger. The smallest root is:

$$\alpha_3 = \frac{\mu_1 + \mu_2 + \mu_3 - \sqrt{(\mu_1 + \mu_2 + \mu_3)^2 - 4\mu_1\mu_3}}{2\mu_3}$$

and substituting back,

$$\alpha_2 = \frac{\mu_1}{\mu_2}\frac{\mu_1 + \mu_2 + \mu_3 - \sqrt{(\mu_1 + \mu_2 + \mu_3)^2 - 4\mu_1\mu_3}}{2\mu_3}$$

It can now be checked that to satisfy also the third and fourth balance equations, the stationary distribution is given by (**??**). For the details see Adan and Weiss (2006).

13.5 Obtain the stationary distribution for the push pull system of Section 13.2.3, with exponential processing times, in the inherently stable case under pull priority [Kopzon et al. (2009)].

Solution

Under pull priority this system is stable and its sample path alternates between M/M/1 busy periods of the top stream, with arrival rate λ_1 and processing rate μ_1, and busy periods of the bottom stream, with arrival rate λ_2 and processing rate μ_2. When it reaches state $(0, 0)$ it will start a top busy period with probability $\lambda_1/(\lambda_1 + \lambda_2)$, and a busy period of the bottom stream with probability $\lambda_2/(\lambda_1 + \lambda_2)$. The balance equations are birth and death type equations. The stationary distribution is:

$$\pi(n_2, 0) = \pi(0, 0)\left(\frac{\lambda_1}{\mu_1}\right)^{n_2},\ n_2 > 0, \quad \pi(0, n_4) = \pi(0, 0)\left(\frac{\lambda_2}{\mu_2}\right)^{n_2},\ n_4 > 0,$$

$$\pi(0, 0) = \frac{(\mu_1 - \lambda_1)(\mu_2 - \lambda_2)}{\mu_1\mu_2 - \lambda_1\lambda_2}.$$

13.6 Show the equivalence of the two definitions of the policy for the inherently unstable push pull system of Section 13.2.3 [Kopzon et al. (2009)].

Solution

Consider the second definition of the policy: Use full utilization, when $n_2 < s_2$ do not serve Q_2, when $n_1 < s_1$ do not serve Q_4, and in all other cases give priority to pull over push. Then: if $n_1 < s_1$, $n_2 < s_2$ we do not serve either standard queue, so for full utilization we push at both IVQs. When $n_2 = 0$, $n_1 \geq s_1$, we need to push at at Q_1 because Q_4 is empty, and we do not serve Q_2 since $0 = n_2 < s_2$, so we also push at Q_3. Similarly, when $n_1 = 0$, $n_2 \geq s_2$, we push at both streams. When $0 < n_1 < s_1$, we do not serve

Q_4, so by full utilization we push at Q_1, but in machine 2 we give priority to Q_2 so we both push and pull at stream 1. Similarly, when $0 < n_2 < s_2$, we push and pull at stream 2. Finally, when $n_1 \geq s_1$, $n_2 \geq s_2$ we are allowed to pull at both streams, and since pull has priority, we pull at both streams. These are exactly the actions under the first definition of the policy.

13.7 Obtain the stationary distribution for the push pull system with exponential processing times in the inherently unstable case, under the policy described in Section 13.2.3 [Kopzon et al. (2009)].

Solution

L et m, n denote the queue lengths of Q_2, Q_4 respectively. Consider $m \geq s_1$, and the states $(m, 0), (m, 1), \ldots, (m, n)$, with $n < s_1$. For this subset of states we get a balance:

$$\lambda_2 P(m,n) + \lambda_1 P(m,0) = \mu_2 P(m,n+1) + \lambda_1 P(m-1,0).$$

and we then have a recursion:

$$P(m,n) = \left(\frac{\lambda_2}{\mu_2}\right)^n P(m,0) - \frac{\lambda_1}{\lambda_2 - \mu_2}\left(\left(\frac{\lambda_2}{\mu_2}\right)^n - 1\right)(P(m-1,0) - P(m,0)).$$

We use this first to get expressions for $P(s_1, n)$. We use $P(s_1 - 1, 0) = 0$ to express $P(s_1, n)$ in terms of $P(s_1, 0)$, and then express $P(s_1, 0)$ in terms of $P(s_1, s_2)$. The result is:

$$P(s_1,n) = P(s_1,s_2)\frac{\left(\frac{\lambda_2}{\mu_2}\right)^n + \frac{\lambda_1}{\lambda_2-\mu_2}\left(\left(\frac{\lambda_2}{\mu_2}\right)^n - 1\right)}{\left(\frac{\lambda_2}{\mu_2}\right)^{s_2} + \frac{\lambda_1}{\lambda_2-\mu_2}\left(\left(\frac{\lambda_2}{\mu_2}\right)^{s_2} - 1\right)},\ 0 \leq n \leq s_2.$$

Next, observing transitions between $Q_2(t) = m - 1$ and $Q_2(t) = m$, we get the balance:

$$\lambda_1 P(m-1,0) = \mu_1 P(m,s_2),\ m > s_1,$$

and following several substitution steps we finally get:

$$P(s_1,n) = P(s_1,s_2)\frac{\left[\frac{\lambda_1}{\mu_1} + \frac{\lambda_1}{\lambda_2-\mu_2}\left(\left(\frac{\lambda_2}{\mu_2}\right)^{s_2} - 1\right)\right]^{m-s_1-1}}{\left[\left(\frac{\lambda_2}{\mu_2}\right)^{s_2} + \frac{\lambda_1}{\lambda_2-\mu_2}\left(\left(\frac{\lambda_2}{\mu_2}\right)^{s_2} - 1\right)\right]^{m-s_1+1}}$$
$$\times\left[\frac{\lambda_1}{\lambda_2-\mu_2}\left(\left(\frac{\lambda_2}{\mu_2}\right)^{s_2} - \frac{\lambda_1}{\mu_1}\right) + \frac{\lambda_1}{\mu_1}\frac{\lambda_1+\lambda_2-\mu_1-\mu_2}{\lambda_2-\mu_2}\left(\frac{\lambda_2}{\mu_2}\right)^n\right],$$
$$m > s_1,\ \ 0 \leq n \leq s_2$$

The same derivation leads to analogous expressions for $P(m, n)$ when $n \geq s_2$, $0 \leq m \leq s_1$. Finally, $P(s_1, s_2)$ is calculated as the normalizing value. Further details can be seen in Kopzon et al. (2009).

13.8 Consider the push pull system in Section 13.2.3, with symmetric rates, i.e. $\lambda_1 = \lambda_2 = \lambda, \mu_1 = \mu_2 = \mu$, with exponential processing times in the inherently

unstable case of $\lambda > \mu$, under the following policy: While $|Q_1(t) - Q_2(t)| > 1$, serve the shorter queue, using both push and pull for this queue. When the queues differ by no more than 1, use pull on both queues. Show that this policy is stable, and obtain its stationary distribution [Kopzon et al. (2009)].

Solution

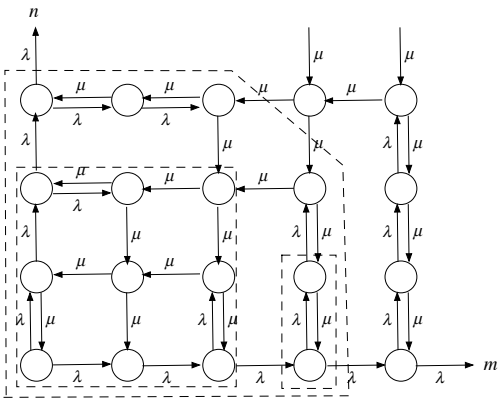
Theorem The stationary distribution for the Markovian symmetric push-pull system is given for $Q_2(t) = m \geq n = Q_4(t)$ by:

$$P(m,0) = P(0,0) \prod_{j=0}^{m-1} \frac{\frac{\lambda}{\mu} + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^j - 1\right)}{\left(\frac{\lambda}{\mu}\right)^j + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^j - 1\right)}, \quad m > 0,$$

$$P(m,n) = P(m,0) \frac{\frac{\lambda}{\lambda-\mu}\left(\frac{\lambda}{\mu}\right)^{m-1} + 2\left(\frac{\lambda}{\mu}\right)^{n+1} - \frac{\lambda}{\mu}\frac{\lambda}{\lambda-\mu}}{\frac{\lambda}{\mu} + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^{m-1} - 1\right)}, \quad m > n > 0,$$

$$P(m,m) = P(m,0)\frac{\lambda}{\mu}.$$

The transition rates for the diagonal policy are given in the following figure, which also shows contours around which we obtain balance equations.



From transitions in and out of the region $Q_2(t), Q_4(t) \leq m-1$ (square region in figure), we get, using symmetry,

$$\lambda P(m-1,0) = \mu P(m,m-1)$$

The recursion reached in Exercise 13.7 is still valid here for $m > 1$ and $0 \leq n < m$, and we use it together with the above balance, for $(m, m-1)$ to

get the recursion

$$P(m,0) = P(m-1,0)\frac{\frac{\lambda}{\mu} + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^{m-1} - 1\right)}{\left(\frac{\lambda}{\mu}\right)^{m-1} + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^{m-1} - 1\right)}, \; m > 0.$$

From which the expression for $P(m,0)$ is obtained.
The expression for $P(m,n)$, $m > n > 0$ is obtained from the full recursion of Exercise 13.7:

$$P(m,n) = \left(\frac{\lambda}{\mu}\right)^n P(m,0) - \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^n - 1\right)(P(m-1,0) - P(m,0)).$$

by substituting the expressions that we got for $P(m,0)$, $P(m-1,0)$.
Finally, the expression for $P(m,m)$ follows from the balance across the large domain in the figure.

13.9 Compare the performance of the two policies, of exercises 13.7, 13.8 [Kopzon et al. (2009)].

Solution

We consider the symmetric system, with $\lambda > \mu$. The fixed threshold policy with $s_1 = s_2 = s$ will be stable whenever $s \geq 2$.
The fixed threshold policy has the following features:
- One of the queues is always longer than s.
- The decay rate of the probabilities $P(m,n)$, $m > s$, $n \leq s$ is seen from exercise 13.7 to be asymptotically at rate:

$$R_1 = \frac{\frac{\lambda}{\mu} + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^s - 1\right)}{\left(\frac{\lambda}{\mu}\right)^s + \frac{\lambda}{\lambda-\mu}\left(\left(\frac{\lambda}{\mu}\right)^s - 1\right)} < 1$$

which improves as s increases.
- The behavior of the sample path is erratic: we alternate between busy periods in which one queue is longer then the other (states in the corridor $0 \leq Q_4 \leq 2$, $Q_2 \geq s$ and the symmetric corridor).
The diagonal policy is less erratic. It can be shown that its decay rate of probabilities is increasing with m, for $m \geq n$ towards:

$$R = \frac{\lambda}{2\lambda - \mu}$$

Which is also the limit of R_1 as $s \to \infty$.
For more details see Kopzon et al. (2009).

13.10 Verify that fluid limits for processing networks with IVQs exist and satisfy the standard fluid equations.

Solution

We have for a processing networks with IVQs the following equations for

work conserving HOL policies:

$$Q_k(t) = Q_k(0) - \sum_{j\in\mathcal{J}} C_{j,k}\mathcal{S}_j(\mathcal{T}_j(t)) + \sum_{j\in\mathcal{J}}\sum_{l\neq k} C_{j,l}\mathcal{R}^j_{l,k}(\mathcal{S}_j(\mathcal{T}_j(t))), \quad k \in \mathcal{K}_0,$$

$$Q_k(t) = Q_k(0) - \sum_{j\in\mathcal{J}} C_{j,k}\mathcal{S}_j(T_j(t)) + \alpha_k t, \quad k \in \mathcal{K}_\infty.$$

and

$$\sum_{j\in\mathcal{J}} A_{i,j}\mathcal{T}_j(t) + \mathcal{I}_i(t) = t, \qquad \int_0^t \Big(\sum_{k\in\mathcal{K},\, j\in\mathcal{J}} A_{i,j}C_{j,k}Q_k(s)\Big)d\mathcal{I}_i(s) = 0.$$

We assume: All service processes $\mathcal{S}_j(t)$ and routing processors satisfy FSLLN: $\mathcal{S}_j(nt)/n \to \mu_j t$ u.o.c. a.s., $\mathcal{R}^j_{k,l}(ns)/n \to p^j_{k,l}s$ u.o.c. a.s.. We denote by $\mathfrak{G}$ the set of ω (of measure 1) for which convergence holds. We consider a sequence of systems, indexed by n, that have the same processing time and routing sequences, but differ in the initial conditions $Q^n(0)$, and assume $Q^n(0)/n \to \bar{Q}(0)$. We state:

Theorem Fluid limits of $Q, \mathcal{T}, \mathcal{I}$ exist and are Lipschitz continuous, and for work conserving HOL policies, every fluid limit for $\omega \in \mathfrak{G}$ satisfies the *fluid model equations* for $k = 1, \ldots, K$:

$$\bar{Q}_k(t) = \bar{Q}_k(0) - \sum_{j\in\mathcal{J}} C_{j,k}\mu_j\bar{\mathcal{T}}_j(t) + \sum_{j\in\mathcal{J}}\sum_{l\neq k} p^j_{l,k}\mu_j\bar{\mathcal{T}}_j(t) \geq 0, \quad k \in \mathcal{K}_0,$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) - \sum_{j\in\mathcal{J}} C_{j,k}\mu_j\bar{\mathcal{T}}_j(t) + \alpha_k, \quad k \in \mathcal{K}_\infty,$$

$$\sum_{j\in\mathcal{J},\, k\in\mathcal{K}} A_{i,j}C_{j,k}\bar{\mathcal{T}}_j(t) + \bar{\mathcal{I}}_i(t) = t, \quad i = 1, \ldots, I,$$

$$\bar{\mathcal{I}}_i(t) \textit{ increases only when } \sum_{k\in\mathcal{K},\, j\in\mathcal{J}} A_{i,j}C_{j,k}Q_k(t) = 0.$$

Proof As for standard MCQN, $\mathcal{T}_k(t,\omega)$ are nondecreasing Lipschitz continuous, and the same holds for their fluid scaling $\bar{\mathcal{T}}^n_k(t) = \mathcal{T}^n_k(nt,\omega)/n$. By the Arzela-Ascoli theorem, for every ω there exists a subsequence of indexes $r \to \infty$ such that $\mathcal{T}^r_j(rt)/r = \bar{\mathcal{T}}_j(t)$ u.o.c.. Therefore there exists also a subsequence for which it holds for all $j \in \mathcal{J}$. So fluid limits for $\mathcal{T}$ exist.

The dynamics equations then follow for each $\omega \in \mathfrak{G}$, since $\bar{\mathcal{T}}_j(t) \leq t$. Lipschitz continuity follows from Lipschitz continuity of $\mathcal{T}$. The work conservation equation for the fluid is proved similar to proof of Theorem 11.2.

13.11 Outline the proof of Theorem 13.1, in analogy with Theorems 11.10, 11.8.

Solution

(a) If the fluid model is weakly stable, then the system is rate stable: The proof is exactly the same as for MCQN. Assume by contradiction that for some $\omega \in \mathfrak{G}$, and $a_n \to \infty$, $|Q_k(a_n,\omega)|/a_n > c > 0$, where we recall that IVQs can be negative, so we use $|\quad|$. By the Theorem on existence of fluid

limits, for a subsequence a_r we have $\lim_{r\to\infty} Q_k(a_r t,\omega)/a_r = \bar{Q}_k(t)$ u.o.c., where $\bar{Q}_k(t)$ is a fluid limit. Since we are looking at a single process $Q(t)$, $Q(0)$ is fixed, and therefore $\bar{Q}_k(0) = 0$ for both standard and IVQs $k \in \mathcal{K}$. On the other hand $|\bar{Q}_k(1)| > c$ by the contrary assumption. This contradicts the assumption that the fluid limit model is weakly stable.

(b) We can describe the processing network with IVQs by a Markov process $\mathcal{X}(t)$, and we have as norm for its state $|x| = \sum_{k\in\mathcal{K}} |Q_k(t)| + \sum_{j\in\mathcal{J}} |V_j(t)|$ The assumptions are: B_κ is petite (uniformly small), and the fluid limit model is stable.

By Theorem 10.7 we need to show that $\lim_{|x|\to\infty} \left|\mathbb{E}_x \frac{1}{|x|}\left(\mathcal{X}(|x|\delta)\right)\right| = 0$ for some δ. The same argument that is used for Theorem 10.8 shows that $\lim \frac{1}{|x_r|} V_j^{x_r}(|x_r|t) = 0$ u.o.c. a.s. and the same convergence holds for the expectation. The argument for this is Lemma 11.5.

All that needs to be shown still is that $\lim \frac{1}{|x_r|} Q_k^{x_r}(|x_r|t) = 0$ u.o.c. a.s. and same holds for the expectation. That $\lim \frac{1}{|x_r|} Q_k^{x_r}(|x_r|t) = 0$ follows from the stability of the fluid model. The convergence for the expected value follows if the sequences $\frac{1}{|x_n|} Q_k(|x_n|t)$ are uniformly integrable. We now check that the statement of Proposition 11.6, that $\frac{1}{a_n} Q_k(a_n t)$ are uniformly integrable, holds for our processing networks with IVQs. Indeed, we again have the bound:

$$\frac{1}{a_n} Q_k(a_n t) \le \kappa + \sum_{j\in\mathcal{J}} \frac{1}{a_n} \mathcal{S}_j(a_n t)$$

which will hold for all buffers, standard of IVQs. Since $\mathcal{S}_j$ are renewal processes, by the elementary renewal theorem they satisfy $\mathbb{E}\left(\frac{1}{a_n}\mathcal{S}_j(a_n t)\right) \to \mu_j t$ so they are uniformly integrable and so are $\frac{1}{a_n} Q_k(a_n t)$.

13.12 Outline the proof of Theorem13.5, in analogy with Theorem 12.6.

Solution

In the proof of Theorem 12.6 the main step was Lemma 12.7, that under maximum pressure policy every fluid limit maximizes the pressure over all of $\mathscr{A}$. The proof of Lemma 12.7 remains unchanged for processing networks with IVQs.

We need a another property of processing networks with IVQs.

Proposition If EAA holds then every activity that is processing a buffer $k \in \mathcal{K}_\infty$ does not process any buffer $\in K_0$.

Proof Assume to the contrary that j serves $k \in \mathcal{K}_\infty$ as well as $l \in \mathcal{K}_0$. Assume first the j is the only activity processing k. Take state $Q_k(t) > 0$, all other buffers empty. Then the maximizing allocation has $a_j > 0$, but $Q_l(t) = 0$, contradiction to EAA. Assume next that some other activities are also serving buffer k. Take $Q_k(t) > 0$, let all the buffers downstream of the other activities that serve k excluding buffer l contain very large number of items, all remaining buffers empty. Then again, the maximizing allocation

has $a_j > 0$, but $Q_l(t) = 0$, contradiction to EAA. □

Corollary If EAA assumption holds, and $Q_k(t) < 0$ for some $k \in \mathcal{K}_\infty$, then $a_j = 0$ for $\{j : C_{j,k} = 1\}$.

Proof Clearly, if $a_j > 0$ this contributes a negative term to the pressure and $a_j = 0$ will increase the pressure. □

Corollary If EAA holds, under maximum pressure policy, for every fluid limit $Q_k(t) \geq 0$ also for the buffers $k \in \mathcal{K}_\infty$.

Proof This follows from $\dot{\bar{Q}}_k(t) = \alpha_k - \sum_{j \in \mathcal{J}} C_{j,k} \mu_j \dot{\bar{T}}(t),\ k \in \mathcal{K}_\infty$, But from the previous corollary it is easy to see that when $\bar{Q}_k(t) < 0$ then $\dot{\bar{T}}(t) = 0$, so by Lemma 11.13, once $\bar{Q}_k(t) \geq 0$ it is never negative again. □

We are now ready to prove Theorem 13.5. Assume we are using some $\alpha \leq \alpha^*$, and maximum pressure policy. Let $\bar{Q}(t), \bar{T}(t)$ be a fluid solution. We again consider the quadratic Lyapunov function:

$$f(t) = \sum_{k=1}^{K} (\bar{Q}_k(t))^2 = \bar{Q}(t) \cdot \bar{Q}(t),$$

We now note the by (13.9),

$$\dot{\bar{Q}}_{\mathcal{K}_\infty}(t) = -R_{\mathcal{K}_\infty} \dot{\bar{T}}(t) + \alpha,$$
$$\dot{\bar{Q}}_{\mathcal{K}_0}(t) = -R_{\mathcal{K}_0} \dot{\bar{T}}(t),$$

So we obtain:

$$\begin{aligned}
\dot{f}(t) &= 2\dot{\bar{Q}}(t) \cdot \bar{Q}(t) \\
&= 2\dot{\bar{Q}}_{\mathcal{K}_\infty}(t) \cdot \bar{Q}_{\mathcal{K}_\infty}(t) + 2\dot{\bar{Q}}_{\mathcal{K}_0}(t) \cdot \bar{Q}_{\mathcal{K}_0}(t) \\
&= -2(R_{\mathcal{K}_\infty} \dot{\bar{T}}(t) + \alpha) \cdot \bar{Q}_{\mathcal{K}_\infty}(t) - 2R_{\mathcal{K}_0} \dot{\bar{T}}(t) \cdot \bar{Q}_{\mathcal{K}_0}(t) \\
&= -2R\dot{\bar{T}}(t) \cdot \bar{Q}(t) + 2\alpha \cdot \bar{Q}_{\mathcal{K}_\infty}(t) \\
&= -2 \max_{a \in \mathcal{A}} Ra \cdot \bar{Q}(t) + 2\alpha \cdot \bar{Q}_{\mathcal{K}_\infty}(t) \qquad \text{by Lemma 12.7}
\end{aligned}$$

Let $x^*,\ \alpha^*$ be an optimal solution of the production planning problem (13.11). Then $R_{\mathcal{K}_\infty} x^* = \alpha^*\ R_{\mathcal{K}_0} x^* = 0$, and from the constraints of (13.11), clearly $x^* \in \mathcal{A}$. We then have:

$$\begin{aligned}
\dot{f}(t) &= -2 \max_{a \in \mathcal{A}} Ra \cdot \bar{Q}(t) + 2\alpha \cdot \bar{Q}_{\mathcal{K}_\infty}(t) \\
&\leq -2Rx^* \cdot \bar{Q}(t) + 2\alpha^* \cdot \bar{Q}_{\mathcal{K}_\infty}(t) \quad \text{since } x^* \in \mathcal{A},\ \alpha \leq \alpha^*, \text{ and } \bar{Q}_{\mathcal{K}_\infty}(t) \geq 0 \\
&= 0 \qquad \text{since } (Rx^*)_{\mathcal{K}_0} = 0, \text{ and } (Rx^*)_{\mathcal{K}_\infty} = \alpha^*.
\end{aligned}$$

The Lyapunov function f then has the properties: $f(t) \geq 0$, it is 0 only when $\bar{Q}_k(t) = 0,\ k = 1, \ldots, K$, and $\dot{f}(t) \leq 0$. It follows by Lemma 11.13 that if $f(0) = 0$ then $f(t) = 0$ for all $t > 0$. Hence, $\bar{Q}_k(0) = 0,\ k = 1, \ldots, K$ implies $\bar{Q}_k(t) = 0$ for all $t > 0$, so the fluid limit model is weakly stable.

13.13 Outline the proof of Corollary 13.6, in analogy with 12.14.

Solution

The proof that no splitting is needed is again because all extreme allocations are $\{0, 1\}$. To see that preemptions are needed, follow all the steps of the proof in Dai and Lin (2005), Section 8 and Appendix B.

13.14 Solve the static production planning problem for the push pull system of Section 13.5, and plot the feasible region for all possible combinations of parameter values [Guo et al. (2014)].

Solution

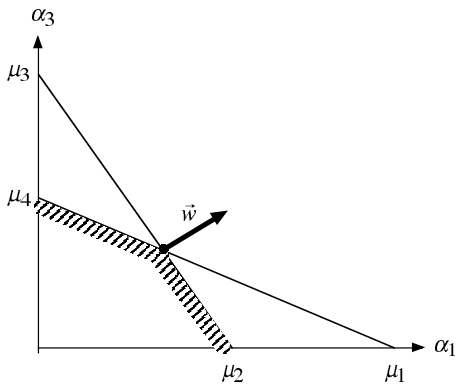
The dynamics of the push pull system are:

$$\begin{aligned} Q_k(t) &= \alpha_k t - S_k(T_k(t)), \qquad k = 1, 3, \\ Q_k(t) &= Q_k(0) + S_{k-1}(T_{k-1}(t)) - S_k(T_k(t)), \qquad k = 2, 4. \end{aligned}$$

with mean processing times $m_k = \mu_k^{-1}$, $k = 1, \ldots, 4$. The static production planning problem for the push-pull network is then:

$$\begin{aligned} \max_{u,\alpha} \quad & w_1\alpha_1 + w_3\alpha_3 \\ s.t. \quad & \begin{bmatrix} \mu_1 & 0 & 0 & 0 \\ -\mu_1 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_3 & 0 \\ 0 & 0 & -\mu_3 & \mu_4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ 0 \\ \alpha_3 \\ 0 \end{bmatrix}, \\ & \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ & u, \alpha \geq 0. \end{aligned}$$

The solution of this linear program is easily read from the following figure,



or from similar figures for any parameter values. According to the values of

the parameters w, μ the optimal nominal inputs can be one of three:

$$\begin{array}{lll} \text{(i)} & \text{either} & \alpha_1 = \min\{\mu_1, \mu_2\}, \qquad \alpha_3 = 0, \\ \text{(ii)} & \text{or} & \alpha_1 = 0, \qquad \alpha_3 = \min\{\mu_3, \mu_4\}, \\ \text{(iii)} & \text{or} & \alpha_1 = \frac{\mu_1\mu_2(\mu_3-\mu_4)}{\mu_1\mu_3-\mu_2\mu_4}, \quad \alpha_3 = \frac{\mu_3\mu_4(\mu_1-\mu_2)}{\mu_1\mu_3-\mu_2\mu_4}. \end{array}$$

If we exclude the singular cases of $\mu_1 = \mu_2$ or $\mu_3 = \mu_4$, we then have the following results: In (i) only queues 1 and 2 are processed, and $\rho_1 = 1, \tilde{\rho}_1 = 0$ while $\rho_2 = \tilde{\rho}_2 = \frac{\mu_1}{\mu_2}$ and this is clearly stable for $\mu_1 < \mu_2$. The case (ii) is similar, with only queues 3, 4 being processed. Case (iii) is the interesting one: We have $\rho_1 = \rho_2 = 1$, but when we define $\tilde{\rho}_i = \sum_{k \in C(i) \cap \mathcal{K}_0} u_k$, as the actual load imposed by the standard queues we have:

$$\tilde{\rho}_1 = \frac{\mu_3(\mu_1 - \mu_2)}{\mu_1\mu_3 - \mu_2\mu_4} < 1, \qquad \tilde{\rho}_2 = \frac{\mu_1(\mu_3 - \mu_4)}{\mu_1\mu_3 - \mu_2\mu_4} < 1.$$

13.15 Obtain the maximum pressure policy for the push pull system Section 13.5, and show that the fluid model is weakly stable but not stable.

Solution

For the push pull network the matrix R is:

$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ -\lambda_1 & \mu_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & -\lambda_2 & \mu_2 \end{pmatrix}$$

and we have that maximum pressure has the following instructions for machines 1 and 2:

$$\begin{array}{cc} \text{for machine 1} & \max\{0;\ \lambda_1(Q_1(t) - Q_2(t);\ \mu_2 Q_4(t);\} \\ \text{choose correspondingly} & a = 0;\ a_1 = 1;\ a_4 = 1; \\ \text{that is:} & \text{idle;} \quad \text{push;} \quad \text{pull;} \end{array}$$

$$\begin{array}{cc} \text{for machine 2} & \max\{0;\ \lambda_2(Q_3(t) - Q_4(t);\ \mu_1 Q_2(t);\} \\ \text{choose correspondingly} & a = 0;\ a_3 = 1;\ a_2 = 1; \\ \text{that is:} & \text{idle;} \quad \text{push;} \quad \text{pull;} \end{array}$$

We now see that if we start at time 0 with

$$\kappa = \lambda_1\big(\bar{Q}_1(0) - \bar{Q}_2(0)\big) = \mu_2\bar{Q}_4(0) = \lambda_2(\bar{Q}_3(0) - \bar{Q}_4(0) = \mu_1\bar{Q}_2(0) > 0$$

then the fluid solution starting from this value will stay with $\bar{Q}(t) = \bar{Q}(0)$ for all $t > 0$. To see this we note: in this state, all four combinations of push or pull are max-pressure policies. It is now seen that if any of them is used,

it needs to be followed by the opposite policy in a sense:

pull - pull will be followed by push - push
push - push will be followed by pull - pull
push - pull will be followed by pull - push for inherently stable
push - pull will be followed by pull - push or by pull - pull for inherently unstable

The fluid limit will then be to use a convex combination:

$$\begin{array}{ll} \text{machine 1: pull } \theta_1 = \frac{\mu_1(\lambda_2-\mu_2)}{\lambda_1\lambda_2-\mu_1\mu_2}; & \text{push } 1-\theta_1 = \frac{\lambda_2(\lambda_1-\mu_1)}{\lambda_1\lambda_2-\mu_1\mu_2}; \\ \text{machine 1: pull } \theta_2 = \frac{\mu_2(\lambda_1-\mu_1)}{\lambda_1\lambda_2-\mu_1\mu_2}; & \text{push } 1-\theta_2 = \frac{\lambda_1(\lambda_2-\mu_2)}{\lambda_1\lambda_2-\mu_1\mu_2}; \end{array}$$

and this will keep the fluid buffer levels constant.

13.16 For the inherently stable push pull system of Section 13.5, under pull priority, show that the fluid model is stable, and show that the policy is a weak pull priority policy [Nazarathy and Weiss (2010)].

Solution

See Nazarathy and Weiss (2010)

13.17 For the inherently unstable push pull system of Section 13.5, show that the linear threshold policy described in Figure 13.8 has a stable fluid limit model [Nazarathy and Weiss (2010)].

Solution

See Nazarathy and Weiss (2010)

13.18 Consider the process $\mathcal{D}(t)$ of departures from the push pull system. Calculate the correlation between $\mathcal{D}_1(t)$ and $\mathcal{D}_2(t)$. For simplicity consider the symmetric case $\mu_1 = \mu_2 = \mu$, $\lambda_1 = \lambda_2 = 1$ [Nazarathy and Weiss (2010)].

Solution

See Nazarathy and Weiss (2010)

14

Optimal Control of Transient Networks

Exercises

14.1 The objective (14.4) gives a reward of $(T-t)h_k$ for each departure from buffer k at time t. Find an objective that will give a reward of $w_{j,k}(T-t)$ for each completion of service of an item in buffer k by activity j, write the equation for this objective in terms of buffer contents, and write its fluid approximation.

Solution

Instead of

$$\max \sum_{k=1}^{K} \int_0^T h_k \mathcal{D}_k(t) dt$$
$$= \max \sum_{k=1}^{K} \int_0^T h_k (T-t) d\mathcal{D}_k(t).$$

We now want to maximize rewards for completed activities, which amounts to:

$$\max \sum_{j=1}^{J} \int_0^T w_j \mathcal{S}_j(\mathcal{T}_j(t)) dt$$
$$= \max \sum_{j=1}^{J} \int_0^T w_j (T-t) d\mathcal{S}_j(\mathcal{T}_j(t)).$$

The fluid approximation for this is:

$$\max \sum_{j=1}^{J} \int_0^T w_j \mu_j \bar{\mathcal{T}}_j(t) dt$$
$$= \max \sum_{j=1}^{J} \int_0^T w_j \mu_j (T-t) \dot{\bar{\mathcal{T}}}_j(t) dt.$$

More generally, if activity j processes simultaneously items from several buffers we have

$$\max \sum_{j=1}^{J} \sum_{k=1}^{K} C_{j,k} \int_0^T w_{j,k}(T-t) d\mathcal{S}_j(\mathcal{T}_j(t)),$$

with fluid approximation:

$$\begin{aligned} &\max \sum_{j=1}^{J} \int_0^T \Big(\sum_{k=1}^{K} C_{j,k} w_{j,k} \Big) \mu_j(T-t) \dot{\bar{\mathcal{T}}}_j(t) dt \\ =: &\max \sum_{j=1}^{J} \int_0^T r_j(T-t) \dot{\bar{\mathcal{T}}}_j(t) dt \end{aligned}$$

14.2 Repeat the proof that fluid limits exist and are Lipschitz continuous, and that almost surely for all ω, the scaled queue lengths and allocations converge to fluid limits that satisfy equations (14.6).

Solution

The proof of Theorem 11.2 needs almost no modifications, for general processing networks with IVQs.

14.3 Show that the problems SCLP and its symmetric dual SCLP* satisfy weak duality, that is if x, u and y, υ are feasible solutions then the objective value of SCLP* is greater or equal to the objective value of SCLP. Show that the objective values are equal if and only if the complementary slackness condition holds, in which case they are optimal solutions.

Solution

Let x, u be feasible solutions of SCLP and let y, υ be feasible solutions of SCLP*. We compare their objective values:

$$\begin{aligned} &\text{Dual objective } = \int_0^T (\alpha^\intercal + (T-t)a^\intercal)\upsilon(t)\ dt + \int_0^T b^\intercal y(t)\ dt \\ &\geq \int_0^T \left(\int_0^{T-t} u(s)^\intercal G^\intercal ds + x(T-t)^\intercal F^\intercal \right) \upsilon(t)\ dt + \int_0^T u(T-t)^\intercal H^\intercal y(t)\ dt \\ &= \int_0^T u(T-t)^\intercal \left(\int_0^t G^\intercal \upsilon(s)\ ds + H^\intercal y(t) \right) dt + \int_0^T x(T-t)^\intercal F^\intercal \upsilon(t)\ dt \\ &\geq \int_0^T u(T-t)^\intercal (\gamma + ct)\ dt + \int_0^T x(T-t)^\intercal d\ dt = \text{Primal objective} \end{aligned}$$

where the first inequality follows from the primal constraints and from υ, y non-negative, the equality follows by exchange in the order of integration, and the second inequality follows from the dual constraints and from u, x nonnegative. □

Clearly, if equality holds then both solutions are optimal. I is not clear that

optimal solutions must achieve equality – if that is the case then one says that strong duality holds.

14.4 Show how to reconstruct the primal and dual solutions of SCLP, SCLP*, from $x^0, y^M, B_1, \ldots, B_M$. and T.

Solution

(step 1) Solve the Boundary-LP and Boundary-LP*. You now have $x^0 = x(0)$ and $y^M = y(0)$.

(step 2) Solve the Rates-LP and Rates-LP* with the basis B_1. Since the basis is given, this requires solving the set of linear equations. Retain the dictionary $\text{Dict}_1 = B_1^{-1}N_1$. You now have $u^1, \dot{x}^1, \upsilon^1, \dot{y}^1$.

(step 3) Form the list of adjacent Rate-LP bases, obtain the variables leaving and entering, ζ_m, ξ_m, $m = 1, \ldots, M-1$

(step 4) Obtain the remaining rates for interval $m = 2, \ldots, M$, where Dict_{m+1} is obtained by a single pivot from Dict_m, in which ζ_m leaves and ξ_m enters. You now have $u^n, \dot{x}^n, \upsilon^n, \dot{y}^n$, $n = 2, \ldots, N$.

(step 5) Solve the interval equations for $x_k(t_m) = 0$ if $\zeta_m = \dot{x}_k$, and $y_j(T - t_m) = 0$ if $\zeta_m = u_j$, where you use the coefficients $\dot{x}, \dot{y}$, and the added equation $\sum \tau_m = T$. You now have $\tau_1, \ldots, \tau_m$.

(step 6) Calculate all other x^m, y^m using x^0, y^M, the $\dot{x}, \dot{y}$, and the τ.

You now have the complete solution.

14.5 Prove that solutions that satisfy all the conditions listed in section 14.4 are optimal solutions of SCLP, SCLP*.

Solution

We will assume that all the bases of the Rates-LP are primal and dual non-degenerate. The proof otherwise is by perturbation.

From the construction it follows that u, x, υ, y are feasible and complementary slack. Therefore they are optimal.

The key point to notice is that in each of the intervals, if $\dot{x}_k = 0$ then also $x_k = 0$, and if $\dot{y}_j = 0$ then also $y_j = 0$. The argument is as follows: Consider $x_k(t)$. If it has $x_k^0 > 0$ then $\dot{x}_k^1$ is in the basis (by compatibility of B_1 with $x(0)$), and by non-degeneracy it is non-zero. Hence, if $\dot{x}_k^1 = 0$, then $x_k(0) = 0$, and $\dot{x}_k(t) = 0$ in the interval $(0, t_1)$, so $x_k(t) = 0$ in the interval $(0, t_1)$. Proceeding we now have that if $x_k(t) > 0$ in the interval (t_{m-1}, t_m), then by non-degeneracy $\dot{x}_k^{m-1}$ in non-zero. Then if $\zeta_m = \dot{x}_k$ then $x_k(t_m) = 0$, and in the next interval both $\dot{x}_k(t) = 0$ and $x_k(t) = 0$, while if $\zeta_m \neq \dot{x}_k$, then $\dot{x}_k$ is basic in the next interval, and so it is non-zero and so is also $x_k(t)$. Next, if $x_k(t) = 0$ in the interval (t_{m-1}, t_m), then of course $\dot{x}_k(t) = 0$ in the interval. Then if $\xi_m \neq \dot{x}_k$, both $\dot{x}_k(t) = 0$ and $x_k(t) = 0$ in the next interval, while if $\xi_m = \dot{x}_k$, then $\dot{x}_k$ is in the basis B_{m+1} and is non-zero by the non-degeneracy. This argument works for all x_k and similarly for all y_j.

14.6 For the fluid solution example described in section 14.5, describe what is happening in intervals 4–9.

Solution

The following table shows in each of the 9 intervals which buffers are processed by the 3 machines.

	machine 1	machine 2	machine 3	empty buffers
Interval 1	1	5	9	none
Interval 2	7	5	9	none
Interval 3	7	5,8	9	9
Interval 4	4	5,8	9	7,9
Interval 5	7	5,8	6,9	7,9
Interval 6	7	8	6,9	5,7,9
Interval 7	7	8	6,9	5,9
Interval 8	7	8	3,6,9	5,9
Interval 9	7	8	3,6,9	5,9

14.7 For the example in section 14.5, write the Boundary-LP/LP*, write the Rates-LP/LP*, list the primal and dual bases of the Rates-LP/LP* for the 9 intervals, and write the equations for the breakpoints.

Solution

We use the notation of Section 14.4, with x_k primal state variables $k = 1, \ldots, 10$, and dual variables $y_j = 1, \ldots, 4$. Here $x_k^n = Q_{k-1}(t_n)$ are the fluid buffer contents. The primal boundary LP is empty since $F = \emptyset$, and its solution is $x_k^0 = \alpha$, initial fluid level. The dual boundary LP* has H and b all non-negative values, and we set $\gamma = 0$, so $y_j^{N+1} = 0$. These are the boundary values for all $0 < t < T$.
The rates LP is:

$$\begin{bmatrix} \mu_1 & & & & & & & & & \\ -\mu_1 & \mu_2 & & & & & & & & \\ & -\mu_2 & \mu_3 & & & & & & & \\ & & -\mu_3 & \mu_4 & & & & & & \\ & & & -\mu_4 & \mu_5 & & & & & \\ & & & & -\mu_5 & \mu_6 & & & & \\ & & & & & -\mu_6 & \mu_7 & & & \\ & & & & & & -\mu_7 & \mu_8 & & \\ & & & & & & & -\mu_8 & \mu_9 & \\ & & & & & & & & -\mu_9 & \mu_{10} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \end{bmatrix} + \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \\ \dot{x}_6 \\ \dot{x}_7 \\ \dot{x}_8 \\ \dot{x}_9 \\ \dot{x}_{10} \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & & & 1 & & & 1 & & \\ & & 1 & & & 1 & & & 1 & \\ & & & 1 & & & 1 & & & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

While the r.h.s. has 0 inflow for buffers $1, \ldots, 9$, we put some small values a_k to avoid degeneracy. Specifically, this will ensure that every non-empty buffer

The bases and pivots for the 9 intervals of the solution are:

1	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	u_2	u_6	u_{10}
pivot		u_2	$\rightarrow$	u_8									
2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	u_6	u_8	u_{10}
pivot		x_{10}	$\rightarrow$	u_9									
3	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	u_6	u_8	u_9	u_{10}
pivot		x_8	$\rightarrow$	u_5									
4	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_9	u_5	u_6	u_8	u_9	u_{10}
pivot		u_5	$\rightarrow$	u_7									
5	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_9	u_6	u_7	u_8	u_9	u_{10}
pivot		x_6	$\rightarrow$	I_2									
6	x_1	x_2	x_3	x_4	x_5	x_7	x_9	u_6	u_7	u_8	u_9	u_{10}	I_2
pivot		x_9	$\rightarrow$	x_8									
7	x_1	x_2	x_3	x_4	x_5	x_7	x_8	u_6	u_7	u_8	u_9	u_{10}	I_2
pivot		x_7	$\rightarrow$	u_4									
8	x_1	x_2	x_3	x_4	x_5	x_8	u_4	u_6	u_7	u_8	u_9	u_{10}	I_2
pivot		x_8	$\rightarrow$	u_5									
9	x_1	x_2	x_3	x_4	x_5	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	I_2

The equations for the breakpoints are:

$$\begin{array}{lrcll}
y_8(t_1): & y_8^0 & - & \sum_{n=2}^{9} \tau_n \dot{y}_8^n & = 0, \\
x_9(t_2): & x_9^0 & - & \sum_{n=1}^{2} \tau_n \dot{x}_9^n & = 0, \\
x_8(t_3): & x_8^0 & - & \sum_{n=1}^{3} \tau_n \dot{x}_8^n & = 0, \\
y_5(t_4): & y_5^0 & - & \sum_{n=5}^{9} \tau_n \dot{y}_8^n & = 0, \\
x_6(t_5): & x_6^0 & - & \sum_{n=1}^{5} \tau_n \dot{x}_8^n & = 0, \\
x_9(t_6): & x_9^0 & - & \sum_{n=1}^{6} \tau_n \dot{x}_9^n & = 0, \\
x_7(t_6): & x_7^0 & - & \sum_{n=1}^{7} \tau_n \dot{x}_7^n & = 0, \\
x_8(t_6): & x_8^0 & - & \sum_{n=1}^{8} \tau_n \dot{x}_8^n & = 0,
\end{array}$$

and the equation $\sum_{n=1}^{8} \tau_n = T$.

14.8 (∗) Reverse engineer the SCLP problem with the solution described in section 14.5.

Solution

I do not know what is the best way of reverse engineering the problem data to obtain a given solution. Here is the actual data for this example:

Re-entrant line with 9 steps and 3 machines, having three successive passages through the 3 machines. Buffer 1 contains the input, served by input processor 1. Buffers $2, \ldots, 10$ are queues for the various steps, going through machines $2, 3, 4$.

Initial fluid levels:

$$x_k^0: \quad 50,\ 24,\ 32,\ 47,\ 7,\ 43,\ 15,\ 20,\ 30,\ 32.$$

The matrix H contains average processing times, which are the values of μ_k^{-1},

for the input and queue buffers:

$$H = \begin{bmatrix} 0.0416 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0579 & 0 & 0 & 0.0432 & 0 & 0 & 0.072 & 0 & 0 \\ 0 & 0 & 0.0323 & 0 & 0 & 0.0371 & 0 & 0 & 0.0173 & 0 \\ 0 & 0 & 0 & 0.0365 & 0 & 0 & 0.0268 & 0 & 0 & 0.0311 \end{bmatrix}.$$

Very small inflow rates to all buffers, to avoid degeneracy, are given by:

$a = 0.00103, 0.00211, 0.00096, 0.00137, 0.00118, 0.00131, 0.00092, 0.00121, 0.00167, 0.00088.$

Costs are 350 + the following:

c : 5.03, 22.19, 19.31, 21.47, 17.04, 54.28, 25.31, 13.53, 10.11, 12.23.

14.9 Use the data for the 3-buffer re-entrant problem of section 14.8 to calculate the minimum time to empty solution, and the last buffer first served solution.

Solution

The data is: $C_1 = 1, 3$, $C_2 = 2$, $q(0) = (8, 1, 15)$, $\mu = (1, 0.25, 1)$.
Accordingly, $q^+ = (8, 9, 24)$, $W_1 = 1 \times 8 + 1 \times 24 = 32$, $W_2 = 4 \times 9 = 36$.
Machine 2 is bottleneck, minimum time to empty is 36. We can use $u = (8/36, 1, 24/36)$ for the inventories to reduce gradually to 0 in that time.
Unde LBFS we have $u(t) = (0, 1, 1)$, $0 \le t < 4$, $u(t) = (0, 0, 1)$, $4 \le t < 15$, $u(t) = (3/4, 1, 1/4)$, $15 \le t < 25.7$, $u(t) = (0, 1, 1/4)$, $25.7 \le t \le 47$.

14.10 Find the optimal fluid solution for the 3-buffer re-entrant problem of section 14.8. Discuss its properties in terms of bases etc.

Solution

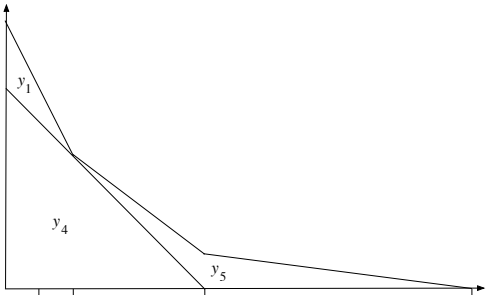
The Boundary-LP solution is $q(0) = (8, 1, 15)$, and the dual variables are $y(0) = (0, 0, 0)$
The reward for the controls u are: $w = h^\intercal R = (0, 0, \mu_3) = (0, 0, 1)$, since only activity 3 reduces the inventory.
The Rates-LP is:

$$\begin{array}{lllll} \max & c_1 u_1 & c_2 u_2 & \mu_3 u_3 & \\ s.t. & \mu_1 u_1 & & & +\dot{q}_1 = a_1, \\ & -\mu_1 u_1 & +\mu_2 u_2 & & +\dot{q}_2 = a_1, \\ & & -\mu_2 u_1 & +\mu_3 u_3 & +\dot{q}_3 = a_3, \\ & u_1 & & +u_3 & \le 1, \\ & & u_2 & & \le 1, \\ & u \ge 0. & & & \end{array}$$

where $a_i > 0$ are exogenous inflows, which we assume are close to 0 (0 would make the problem degenerate). Similarly, $c_1 > 0, c_2 > 0$ are reward rates, close to 0. There are also two slack variables, u_4, u_5. The dual, Rates-LP is:

$$\begin{array}{llllll} \min & a_1 \upsilon_1 & +a_2 \upsilon_2 & +a_3 \upsilon_3 & & \\ s.t. & \mu_1 \upsilon_1 & -\mu_1 \upsilon_2 & & +\dot{y}_4 & -\dot{y}_1 = c_1, \\ & & \mu_2 \upsilon_2 & -\mu_2 \upsilon_3 & +\dot{y}_5 & -\dot{y}_2 = c_2, \\ & & & +\mu_3 \upsilon_3 & +\dot{y}_4 & -\dot{y}_3 = 1, \\ & \upsilon \ge 0. & & & & \end{array}$$

In the following we put $\dot{0}$ for infinitesimal values, replacing c_i, a_i.
– First interval: primal basis is $B_1 = \{\dot{q}_1 = \dot{0}, \dot{q}_2 = -1/4, \dot{q}_3 = -1, u_2 = 1, u_3 = 1\}$. Dual basis is $B_1^* = \{\dot{y}_1 = 1, \dot{y}_4 = 1, \dot{y}_5 = \dot{0}$
At time $t_1 = 4$, $q_2 = 0$. Pivot: $\dot{q}_2$ leaves, u_5 enters. Dual pivot in reversed time: υ_2 leaves, y_5 enters.
–Second interval: primal basis is $B_2 = \{\dot{q}_1 = \dot{0}, \dot{q}_3 = -1, u_2 = \dot{0}, u_3 = 1, u_5 = 1\}$. Dual basis is $B_2^* = \{\upsilon_2 = \dot{0}, \dot{y}_1 = 1, \dot{y}_4 = 1\}$
At time $t = 8$, we start pumping out of 1 into 2 and out of 2. Pivot u_5 leaves, u_1 enters. Dual pivot in reverse time is $\dot{y}_5$ leaves, $\dot{y}_1$ enters.
– Third interval: primal basis is $B_3 = \{\dot{q}_1 = -1/4, \dot{q}_3 = -1/2, u_1 = 1/4, u_2 = 1, u_3 = 3/4\}$. Dual basis is $B_3^* = \{\upsilon_2 = 1, \dot{y}_4 = 1, \dot{y}_5 = -1/4\}$.
At time $t = 24$, buffer 3 is empty. Pivot $\dot{q}_3$ leaves the basis, u_4 enters. Dual pivot in reverse time is υ_3 leaves, $\dot{y}_4$ enters.
– Fourth interval: primal basis is: $B_4 = \{\dot{q}_1 = -1/4, u_1 = 1/4, u_2 = 1, u_3 = 1/4, u_4 = 1/2\}$. Dual basis is $B_4^* = \{\upsilon_2 = \dot{0}, \upsilon_3 = 1, \dot{y}_5 = 1/4\}$.
At time $t = 40$ the system is empty. Primal basis is $B_5 = \{u_1 = u_3 = u_3 = \dot{0}, u_4 = u_5 = 1\}$. Dual basis is $B_5^* = \{\upsilon_1 = \upsilon_2 = \upsilon_1 = 1\}$.
To complete the description of the solution we plot the dual variables:



14.11 Write the equations for the tracking process for the 3-buffer re-entrant problem of section 14.8, and describe the policies for tracking the fluid in each of the intervals of the fluid solution.
Solution

– First period: Buffer 1 stays constant, buffers 2,3 are IVQs, buffer 2 empties as stochastic process rate μ_2, buffer 3 is filled by output of buffer 2, and empties as stochastic process rate μ_3. Interval ends when buffer 2 is empty, at ≈ 4.
– Second period: Buffer 1 stays constant, buffer 2 is empty with no inflow or outflow, buffer 3 is an IVQ, emptying as stochastic process rate μ_3. Period ends at time $t = 8$.
– Third period: Buffers 1 and 3 are IVQs. Buffer 2 is a standard queue. Nominal rate for buffer 2 is μ_2. Tracking using maximum pressure would

suggest that when machine 1 is available it compares pressure at buffers 1 and 3. pressure at 3 is deviation from nominal. Pressure at 1 is deviation from nominal minus level of Q_2. Better alternative in this case: when buffer 2 is close to empty, work on buffer1, otherwise on buffer 3.

– Fourth period: Buffer 1 is IVQ, buffer 2, 3, are standard queues. Policy is: work on 2 always, Chose between 1 and 3 accordigng to pressure. Pressure for 3 is deviation from nominal. pressure for 1 is deviation from nominal minus $Q_2(t)$.

14.12 Simulate the control of the 3-buffer re-entrant problem of section 14.8, for a scaling by 10 and by 100. Use exponential processing times for each activity.

Part V

Diffusion Scaled Balanced Heavy Traffic

15

Join the Shortest Queue in Parallel Servers

Exercises

15.1 (∗) For the case of two symmetric ·/M/1 queues, under Join the Shortest Queue policy, complete the derivation of the stationary distribution as an infinite sum of product forms. Perform the following steps [Adan et al. (1990)]:

(a) Write the balance equations, eliminate those for $n = 0$, and set up the three sets of equations, for states in the interior of the quadrant, for states on the horizontal $n = 1$ boundary, and for states on the vertical $m = 0$ boundary.

(b) Derive the quadratic equation for α, β that give product form solutions for the states in the interior.

(c) Verify that α_0, β_0 satisfy balance equations for states (i),(ii).

(d) Do the first compensation, to satisfy (i),(iii).

(e) Derive the sequence of α_k, β_k and the coefficients c_k, d_k.

(f) Show that the coefficients and the roots converge geometrically.

Solution

(a) The balance equations are:

$$
\begin{aligned}
&2(\rho+1)P(m,n) = 2\rho\mathbb{P}(m-1,n+1) + P(m,n+1) + P(m+1,n-1), && m>0, n>1, && (i)\\
&2(\rho+1)P(m,1) = 2\rho P(m-1,2) + P(m,2) + \tfrac{\rho}{1+\rho}(2\rho P(m-1,1) + P(m,1)) && && (ii)\\
&\quad + \tfrac{1}{1+\rho}(2\rho P(m,1) + P(m+1,1)), && m>0, && \\
&(1+2\rho)P(0,n) = P(0,n+1) + P(1,n-1), && n>1, && (iii)\\
&2(1+\rho)P(m,0) = 2\rho P(m-1,1) + P(m,1), && m>0, && (iv)\\
&2\rho P(0,0) = P(0,1). && && (v)
\end{aligned}
$$

Where in the equation for $P(m,1)$ we substituted the values of $P(m,0)$, $P(m+1,0)$ from equation (iv). Equations (iv), (v) now serve as definitions.

(b) Solutions of equation (i), for $m > 0, n > 1$ are $P(m,n) = \alpha^m \beta^n$ where α, β are any solutions to the bi-quadratic equation:

$$2(\rho+1)\alpha\beta = 2\rho\beta^2 + \alpha\beta^2 + \alpha^2.$$

which we can write as:

$$\beta^2(2\rho+\alpha) - 2\beta\alpha(\rho+1) + \alpha^2 = 0$$

or as:

$$\alpha^2 - \alpha(2\beta(\rho+1) - \beta^2) + 2\beta^2\rho = 0$$

to illustrate its role as quadratic for β with given α, and for α with given β. Any linear combination of such solutions will satisfies (i).
(c) Our compensation method will give an infinite sum of product forms, where each successive term will be smaller, so the first term should describe the asymptotic behavior for $m, n \to \infty$. Since we expect for large m that n will be comparatively small, we should get behavior close to M/M/2, so $\sum_{n=0}^{\infty} P(m,n) \approx \mathbb{P}(\text{M/M/2 queue} = 2m) = K\rho^{2m}$, so $\alpha_0 = \rho^2$.
If we take $\alpha_0 = \rho^2$, there are two values that of β that satisfy the bi-quadratic equation, two quadratic roots, namely ρ and $\beta_0 = \frac{\rho^2}{2+\rho}$. The value ρ would give $P(m,n)$ of two independent M/M/1 queues, which is not what we want. To satisfy equation (ii), α, β must satisfy (by rewriting (ii)):

$$\alpha^{m-1}\left[(2\rho+\alpha)(1+\rho)\beta + \alpha^2 - \alpha\left(2\rho^2+\rho+2\right) + 2\rho^2\right] = \alpha^{m-1}B(\alpha,\beta) = 0.$$

So substituting $\alpha = \rho^2$, and $\beta = \frac{\rho^2}{2+\rho}$ this holds, so $\alpha_0^m\beta_0^n$ will satisfy both (i) and (ii).
(d) To satisfy equations (iii), with a single term we would need to have:

$$\beta^{n-1}\left[\beta^2 - (1+2\rho)\beta + \alpha\right] = \beta^{n-1}A(\alpha,\beta) = 0.$$

However, in the next compensation step, we now have two terms, in the form $\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta^n$, and to satisfy (iii) we need to have that $\beta = \beta_0$. In the bi-quadratic, solving for α with the fixed value β_0, one of the roots was α_0. So this determines α_1 as the other root with β_0. c_1 is now determined from the linear equation:

$$A(\alpha_0,\beta_0) + c_1A(\alpha_1,\beta_0) = 0,$$

so:

$$c_1 = -\frac{A(\alpha_0,\beta_0)}{A(\alpha_1,\beta_0)} = -\frac{\beta_0^2 - (1+2\rho)\beta_0 + \alpha_0}{\beta_0^2 - (1+2\rho)\beta_0 + \alpha_1} = -\frac{\alpha_1 - \beta_0}{\alpha_0 - \beta_0}.$$

where we used that $\alpha_1 + \alpha_0 = 2\beta_0(\rho+1) - \beta_0^2$ as two roots of the bi-quadratic with β_0.
(e) The next compensation step is to add a term so that (ii) again holds. Since the α_0^m, β_0^m already satisfies (i) and (ii), the compensation needs to consider compensating for α_1^m, β_0^m, so the α, β of the new term will need to satisfy:

$$c_1\alpha_1^{m-1}B(\alpha_1,\beta_0) + c_1d_1\alpha^{m-1}B(\alpha,\beta) = 0$$

From which we see that $\alpha = \alpha_1$, and so β_1 needs to be the second root of the bi-quadratic with α_1 (the first root was β_0). Once we have α_1, β_1, the linear equation for d_1 is:

$$B(\alpha_1, \beta_0) + d_1 B(\alpha_1, \beta_1) = 0,$$

and we obtain:

$$d_1 = -\frac{B(\alpha_1, \beta_0)}{B(\alpha_1, \beta_1)} = -\frac{\frac{\beta_0(2\rho+\alpha_1)(\alpha_1+\rho)}{1+\rho} - \alpha_1^2}{\frac{\beta_1(2\rho+\alpha_1)(\alpha_1+\rho)}{1+\rho} - \alpha_1^2} = -\frac{\frac{\alpha_1+\rho}{\beta_1} - (\rho+1)}{\frac{\alpha_1+\rho}{\beta_0} - (\rho+1)},$$

where we used: $\beta_0\beta_1(2\rho+\alpha_1) = \alpha_1^2$ since β_0, β_1 are roots of the bi-quadratic with α_1.
So far we have:
$\alpha_0^m \beta_0^n$ satisfy (i) and (ii) but not (iii)
$\alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n$ satisfy (i) and (iii) but not (ii),
$\alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n + c_1 d_1 \alpha_1^m \beta_1^n$ satisfy (i) and (ii) but not (iii).

It is now clear how the infinite sequence of product form terms is constructed. With α_k we have the two roots $\beta_k > \beta_{k+1}$, and then we obtain α_{k+1} as the second, smaller root (the first being α_k) for β_{k+1}. We note, either by looking at the equation for the bi-quadratic, or simply by looking at Figure 15.2, where each step crosses the line $\alpha = \beta$, that

$$\alpha_0 > \beta_0 > \alpha_1 > \cdots > \alpha_k > \beta_k > \alpha_{k+1} > \cdots$$

The terms for c_k, d_k are then (with $c_0 = d_0 = 1$):

$$c_{k+1} = -\frac{\alpha_{k+1} - \beta_k}{\alpha_k - \beta_k} c_k, \qquad d_{k+1} = -\frac{\frac{\alpha_{k+1}+\rho}{\beta_{k+1}} - (\rho+1)}{\frac{\alpha_{k+1}+\rho}{\beta_k} - (\rho+1)} d_k.$$

and the complete solution of the balance equations for $m \geq 0$, $n \geq 1$ is:

$$\begin{aligned} x(m,n) &= \sum_{k=0}^{\infty} d_k \left(c_k \alpha_k^m + c_{k+1} \alpha_{k+1}^m\right) \beta_k^n \\ &= \beta_0^n \alpha_0^m + \sum_{k=0}^{\infty} c_{k+1} \left(d_k \beta_k^n + d_{k+1} \beta_{k+1}^n\right) \alpha_{k+1}^m \end{aligned}$$

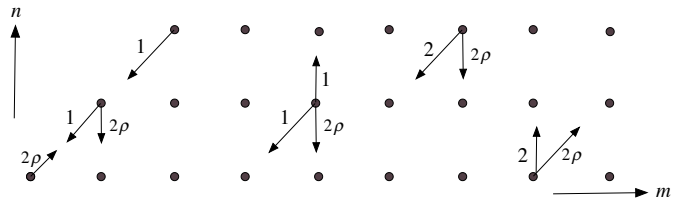
It is seen from these the c_k are all positive, while d_k alternate in sign.
To calculate α_k, β_k explicitly one can use:

$$\alpha_k + \alpha_{k+1} = 2\beta_k(\rho+1) - \beta_k^2, \qquad \alpha_k \alpha_{k+1} = 2\beta_k^2 \rho,$$
$$\beta_k + \beta_{k+1} = \frac{2(\rho+1)\alpha_{k+1}}{2\rho + \alpha_{k+1}}, \qquad \beta_k \beta_{k+1} = \frac{\alpha_{k+1}^2}{2\rho + \alpha_{k+1}},$$

This leads to a 2nd order difference equation for $1/\beta_k$, and to explicit solutions of α_k, β_k.

(f) For the convergence, one can define $u_k = \alpha_k/\beta_k$, $v_k = \alpha_{k+1}/\beta_k$. One can express $\alpha_k, \beta_k, c_k, d_k$ in terms of u_k, v_k and one can easily see that $u_k \uparrow A_1$, $v_k \downarrow A_2$, where $A_1 = \rho + 1 + \sqrt{\rho^2+1}$, $A_2 = \rho + 1 - \sqrt{\rho^2+1}$. One can then show that all the terms in the summation converge geometrically, and so the sum converges absolutely and it converges to a positive term $x(m,n)$ for every m, n.
Finally, looking at the $\sum_{m,n} x(m,n)$, it can be shown that it converges, to a finite B^{-1}, and $p(m,n) = Bx(m,n)$, with $B = \frac{\rho(2+\rho)}{2(1-\rho^2)(2-\rho)}$.
(e) These are obtained as infinite sums of geometric terms. In particular $2m+n$ is the number of customers in the queue, m the length of the shorter queue, and m is also the expected waiting time of a customer that joins the shorter queue, conditional on m.

15.2 (*) The following model is called *join the shortest queue with jockeying* : in the two symmetric ·/M/1 queues customers alway join the shortest queue, however, when the difference in queue length at the two servers exceeds a threshold d a waiting customer is moved from the longer to the shorter queue. For $d = 2$, draw the states and transitions diagram, write the balance equations and suggest how to solve them [Adan et al. (1993, 1994)].
Solution
It is convenient to represent the state as $Q(t) = (m,n)$ where m is the length of longest queue, and $n = 0, 1, 2$ records the difference in length (somwhat counter intuitive, number in system is $2m - n$). The following is the states and transition diagram:

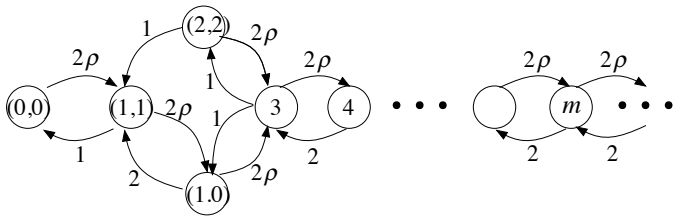


The balance equations are:

$$
\begin{aligned}
&2(1+\rho)P(m,2) = P(m,1), \quad m \geq 2,\\
&2(1+\rho)P(m,1) = 2P(m+1,2) + 2\rho P(m,2) + 2\rho P(m-1,0) + 2P(m,0), \quad m \geq 2,\\
&2(1+\rho)P(m,0) = P(m+1,1) + 2\rho P(m,1), \quad m \geq 1,\\
&(1+2\rho)\,P(1,1) = P(2,2) + 2\rho P(0,0) + 2P(1,0),\\
&\qquad 2\rho\; P(0,0) = P(1,1).
\end{aligned}
$$

A simple approach to solve the balance equations is as follows: We note that the total number of customers fluctuates between even and odd, where odd are the states $O_m = (m,1)$ with $2m-1$ customers, while even are the states $E_m = \{(m,0),(m-1,2)\}$ with $2m$ customers, and for 3 or more customers

in the system, this is just like M/M/2. The transition rates for the new states are



We obtain from this the stationary probabilities:

$$P(0,0) = B,\ P(1,1) = B2\rho,\ P(1,0) = B\frac{2+2\rho}{2+3\rho}2\rho^2,\ P(2,2) = B\frac{2\rho}{2+3\rho}2\rho^2,$$
$$P(3+) = B\frac{2+4\rho}{2+3\rho}2\rho^3,$$
$$P(M) = B\frac{2+4\rho}{2+3\rho}2(1-\rho)\rho^M,\ \text{total number } M \geq 3,$$
$$P(m,1) = B\frac{2+4\rho}{2+3\rho}2(1-\rho)\rho^{2m-1},\ m \geq 2,$$
$$P(m,2) = B\frac{2+4\rho}{(2+3\rho)(2+2\rho)}2(1-\rho)\rho^{2m-1},\ m \geq 2,$$
$$P(m,0) = B\frac{(2+4\rho)(2+\rho)}{(2+3\rho)(2+2\rho)}2(1-\rho)\rho^{2m},\ m \geq 1.$$

This problem can be solved by Matrix-Geometric techniques (a nice introduction to these techniques is Haviv (2013) Chapter 12). In particular it is shown in Adan et al. (1993, 1994) that for any number of parallel exponential servers with different service rates, JSQ with jockeying whenever the difference between the longest and shortest queue reaches a threshold, can be solved explicitly.

15.3 For the case of two symmetric ·/M/1 queues, under join the Shortest Queue policy, obtain the stationary marginal distribution of the difference in length between the two queues, and find an M/M/1 type approximation that is an upper bound for it.

Solution

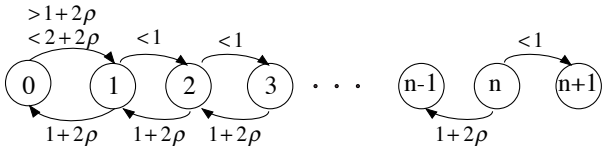
Let $Y(t)$ denote the difference between the longer and shorter queue at time t.

The exact expression for the stationary distribution of $Y(t)$, expressed by an

infinite sum is as follows (for $n \geq 1$):

$$\begin{aligned} p(\cdot, n) &= B \sum_{m=0}^{\infty} \sum_{k=0}^{\infty} d_k \left(c_k \alpha_k^m + c_{k+1} \alpha_{k+1}^m \right) \beta_k^n \\ &= B \sum_{k=0}^{\infty} d_k \left(c_k \frac{1}{1-\alpha_k} + c_{k+1} \frac{1}{1-\alpha_{k+1}} \right) \beta_k^n \\ &= B \sum_{k=0}^{\infty} c_k d_k \frac{(1-\alpha_{k+1}) - \frac{\alpha_{k+1}-\beta_k}{\alpha_k-\beta_k}(1-\alpha_k)}{(1-\alpha_k)(1-\alpha_{k+1})} \beta_k^n \\ &= B \sum_{k=0}^{\infty} c_k d_k \frac{(\alpha_{k+1}-\alpha_k)(1-\beta_k)}{(1-\alpha_k)(1-\alpha_{k+1})(\alpha_k-\beta_k)} \beta_k^n \end{aligned}$$

$Y(t)$ on its own is not a Markov process. However, we can find an upper bound for the stationary distributions of $Y(t)$ by looking at the transition rates of $Y(t)$, as shown in the following figure: Here, the rate at which $Y(t)$ increases



by 1 is bounded: $1 + 2\rho < q(0,1) < 2 + 2\rho$, and $0 < q(n, n+1) < 1$, $n \geq 1$, where the lower bound correspond to the shorter queue being empty, and the upper bound to the shorter queue being not empty. This implies that:

$$\pi(0) = \mathbb{P}(Y(\infty) = 0) \geq \frac{\rho}{1+2\rho},$$

$$\pi(j) = \mathbb{P}(Y(\infty) = j) \leq \frac{2\rho(1+\rho)}{(1+2\rho)^2} \left(\frac{1}{1+2\rho} \right)^{j-1}, \quad j = 1, 2, \ldots$$

The bounds are obtained as follows: We have as upper bound on $\pi(j)$, $\pi(j) \leq \pi(1)\left(\frac{1}{1+2\rho}\right)^{j-1}$, so $\sum_{j=1}^{\infty} \pi(j) < \pi(1)\frac{1+2\rho}{2\rho}$. We also have the upper bound on $\pi(1)$: $\pi(1) < \pi(0)\frac{2+2\rho}{1+2\rho}$. Since probabilities add up to 1, we have for the upper bounds on $\pi(j)$, $j \geq 1$ and lower bound on $\pi(0$ that:

$$1 = \pi(0) + \pi(0)\frac{2+2\rho}{1+2\rho}\frac{1+2\rho}{2\rho} \quad \Longrightarrow \quad \pi(0) = \frac{\rho}{1+2\rho}.$$

15.4 For two identical ·/M/1 servers and any arrival stream, prove that join the shortest queue maximizes the probability that k items will complete service by time s, for all s and k (hint: use induction on k [Winston (1977); Weber (1978)]).

Solution

We wish to show, $\mathbb{P}_x(\mathcal{D}(s) \geq k)$, the probability that starting from $Q(0) = x$ the number of departures by time s will equal or exceed k is maximized by JSQ, for all x, s, k. We show if by induction on k. For $k = 0$ there is nothing to show. Assume the theorem is correct for $k - 1$. We will prove it for k.

After the first departure, by the induction hypothesis we should use JSQ. So we only need to consider the period up to the first departure. Assume at time 0, queue 1 is longer than queue 2. Consider two strategies up to the time of first departure:

– *Strategy S1*: Assigns first customer to queue 1, then continues in some arbitrary fashion, until the first departure.

– *Strategy S2*: Assigns first customer to queue 2. Then imitate strategy S1 at each arrival, until S1 assigns a customer to queue 2 (if that happens before first departure) and assign that arrival not to queue 2 but to queue 1. We also assume that as long as queue 2 is empty under strategy S1 we do not serve it under strategy S2.

At the time of the first departure, there are two possibilities: either the two queues have the same number of customers under both S1,S2, or, if under S1 no arrival has been assigned to queue 2, then there will be one customer more at server 2 and one less at server 1 under S2 than under S1. In the first case we see that S2 is as good as S1. We now discuss the second case, and show that in that case S2 is better than S1 in maximizing $\mathbb{P}_x(\mathcal{D}(s) \geq k)$.

Note that in this second case, at the time of the first departure there is a single customer which is in queue 1 under S1 and in queue 2 under S2. Also, because S1 has only assigned arrivals to queue 1, queue 1 is longer than queue 2 if we do not count the extra customer. Imagine now that instead of the extra customer being in the system before the first departure, it has arrived just after the first departure. Then by the induction hypothesis it should be assigned to queue 2. Hence the action of S2 at the time of the first arrival was better than the action of S1. If under S2 we would have also served queue 2 when it was empty under S1, then the advantage of S2 would be even greater. The same argument then hods also for the 2nd, and for all subsequent arrivals before the first departure. This completes the proof.

Note that the proof is valid for any arrival stream, and for any number of identical servers with exponential service time. We need the exponential service assumption here to make the system Markovian and memoryless.

The same paper of Weber (1978) also shows that JSQ is optimal for identical servers with IHR (increasing hazard rate) distributions.

15.5 Consider Poisson arrivals, and service time that is 0 with probability $1 - \epsilon$, and n with probability ϵ. Consider the following alternative to JSQ: If there is an empty server or if the difference in queue length is 0 or 1, use JSQ. Otherwise, join the longer queue. Show that for fixed λ and n, this policy outperforms JSQ as $\epsilon \to 0$ (Hint: show that under any policy, the probability that a busy cycle contains k long jobs is of order ϵ^k [Whitt (1986)]).

Solution

For any fixed λ, as $\epsilon \searrow 0$ the system goes to light traffic. We refer to the customers according to service time as 0 or n. Consider the story of how things work in a busy period: for most of the time, both queues are empty, and an arriving customer is 0, he waits 0 and leaves immediately. Then we have an arrival n, and for a duration of n he is in one queue and most of the time, all arrivals during n are 0, they go to the empty queue, wait 0, and and leave immediately. Eventually, we may have an arrival of n during n and we now have both servers busy. Arrivals will now join alternating queue 1 and 2, which will remain almost equal, until the earlier of the n leaves, and most of the time that queue will now be empty. We now get to the interesting part: Occasionally, there will be an arrival of a third n before one of the two ns leaves.

With probability $1/2$ it joins the queue that does not empty, and then in most cases the non-empty queue will empty from both n customers before the other queue becomes non-empty.

With probability $1/2$ the new n customer joins the queue of the n customer that leaves first. In that case, the queue that has the other n will have nobody leaving, while the queue of the n that leaves will become shorter, because all the customer before the third n will leave. But now, usually, the second n will complete, and all customers behind it will leave, while the third n is still in process. So if you arrive in this situation, you should join the longest queue.

We have shown that for all cases that the busy period has no more than 3 ns, the policy of join empty or join shortest when difference is 1, but if difference is more than 1 join longest queue is better than the policy of JSQ.

To complete the proof, see Whitt (1986), for proofs of:

Lemma For X a Poisson random variable with parameter λ,

$$\lambda^k \geq \mathbb{P}(X \geq k) \geq \lambda^k e^{-\lambda}/k!.$$

Theorem. The probability of k or more long jobs in a busy period is of order ϵ^k

15.6 Calculate the expected waiting time and the expected sojourn time for an M/M/2 queue, and compare it with the expected waiting time and expected sojourn time of an M/M/1 queue with a server that has twice the speed.

Solution

Denote W the sojourn time, V the waiting time, and Q the number of customers in the system in steady state..

For $M/M/1$ with arrival rate λ and service rate 2μ, with $m = 1/\mu$ and $\rho = \frac{\lambda}{2\mu}$,

$$\mathbb{P}(W = 0) = 1 - \rho, \qquad \mathbb{E}(W) = \frac{m/2}{1-\rho}, \qquad \mathbb{E}(V) = \rho\frac{m/2}{1-\rho}$$

For M/M/2 with arrival rate λ and service rates μ, with $m = 1/\mu$ and $\rho = \frac{\lambda}{2\mu}$,

$$\mathbb{P}(Q = 0) = \frac{1-\rho}{1+\rho}, \qquad \mathbb{P}(Q = k) = \frac{1-\rho}{1+\rho} 2\rho^k, \ k \geq 1$$

$$\mathbb{P}(W = 0) = \mathbb{P}(Q = 0) + \mathbb{P}(Q = 1) = 1 - \rho \frac{2\rho}{1+\rho} > 1 - \rho.$$

and the waiting time once a customer has to wait is the same as for the M/M/1 double speed, so:

$$\mathbb{E}(V) = \rho \frac{2\rho}{1+\rho} \frac{m/2}{1-\rho} = \frac{\rho^2}{1-\rho^2} m, \qquad \mathbb{E}(W) = m + \mathbb{E}(V) = \frac{m}{1-\rho^2}$$

so with two servers, wait is shorter, sojourn longer than with a single double speed server.

When $\rho \nearrow 1$, the differences disappear.

15.7 Calculate the expected waiting time for a system with s ·/M/1 queues, when Poisson customers are routed on a round robin policy to the different queues.

Solution

Assume arrivals are Poisson rate λ, service at each server is exponential rate μ, and let $\rho = \lambda/s\mu$.

Under this policy each server behaves like a G/M/1 queue with interarrival times distributed Erlang(s, λ). As we saw in Section 2.9, the stationary distribution is:

$$\mathbb{P}(Q = k) = \begin{cases} \rho(1-\alpha)\alpha^{k-1}, & k = 1, 2, \dots \\ 1 - \rho, & k = 0. \end{cases} .$$

where α is the unique < 1 solution of

$$\alpha = F^*(\mu(1-\alpha)) = \left(\frac{\lambda}{\lambda + \mu(1-\alpha)} \right)^s$$

and sojourn time is then exponential with parameter $\mu(1-\alpha)$. One can solve for α numerically.

In heavy traffic, i.e when $\rho = \lambda/s\mu \approx 1$, we can use Kingman's bound approximation:

$$\mathbb{E}(V) = \mathbb{E}(W) = \frac{m}{1-\rho} \frac{1/s + 1}{2}$$

For $s = 2$ this is $3/2$ the value for a double speed server, and $2/3$ of the random assignment sojourn. When the number of servers is large, this is half the sojourn time of random assignment.

15.8 Verify the calculated values of the resource pooling effect on expected sojourn time as listed in Table 15.1.

Solution

We already laid out all the necessary theory. The columns for Alternate

routing and for JSQ need to be calculated numerically. The other columns just need substitution in the closed form expressions.

15.9 An alternative to the proof of the bound on the difference between JSQ and G/G/2, is to compare JSQ to a G/G/1 queue, with half the processing times. With $\mathcal{W}(t)$ the unfinished workload of the soulbe speed GI/G/1 queue, show that

$$\mathcal{W}(t) \leq W_1(t) + W_2(t) \leq \tilde{W}(t) + \sup_{0<s<t} \left|\mathcal{W}_1(t) - W_2(t)\right|$$

Solution

We compare JSQ for two identical ·/G/1 systems, with a single G/G_2/1 system with service times distributed as $G(x/2)$ (twice the speed). We couple the streams of arrivals and service requirements for the two systems and denote W_1, W_2, W the remaining workload processes.

Denote $L(t) = \mathcal{W}_1(t) + \mathcal{W}_2(t) - \mathcal{W}(t)$ and $K(t) = |\mathcal{W}_1(t) - \mathcal{W}_2(t)|$, we wish to show $0 \leq L(t) \leq \sup_{0\leq s\leq t} K(s)$. Consider a single busy period of the JSQ system, starting at time 0, so that $L(0) = K(0) = 0$. Let t_0 be the first time that G/G_2/1 is busy, while JSQ has one server idle, if such a time exists, or else we let t_0 be the end of the busy period. Then $L(t) = 0$ for $0 \leq t < t_0$. At the time t_0 w.l.g., assume server 2 is idle and server 1 is busy. Then, $L(t_0) \geq 0$ and $K(t_0) = \mathcal{W}_1(t_0) \geq \mathcal{W}(t_0)$ (they are equal for t_0 but we will need $\geq$ for $t_1, t_2 \ldots$). In the time following t_0, while G/G_2/1 is busy at rate 2, in JSQ only 1 server is busy, so $L(t)$ increases at rate 1, while $K(t)$ decreases at rate 1. However, this can only continue for at most a duration $\mathcal{W}(t_0)/2$, and so in this entire period that $L(t)$ is increasing, $0 \leq L(t) \leq L(t_0) + \mathcal{W}(t_0)/2 \leq K(t_0) - \mathcal{W}(t_0)/2 \leq K(t_0)$.

Before the time $t_0 + \mathcal{W}(t_0)/2$ two things can happen: (i) There may be an arrival, in which case both systems will again have two servers busy, or (ii) we may reach a time t when $\mathcal{W}(t) = 0$ while $L(t) = \mathcal{W}_1(t) \geq 0$. At that time the JSQ system will have one busy server while the G/G_2/1 system will be idle, and $L(t)$ will be decreasing until it reaches 0 and the busy period ends, or there is a new arrival, and both systems will again be processing at rate 2. Throughout, $0 \leq L(t) \leq K(t_0)$ will be maintained, until the busy period is over, or until at some time $t_1 > t_0$, again for the first time, system G/G_2/1 is busy while the JSQ system has one idle server.

Then at t_1, $L(t_1) \geq 0$ and $K(t_1) = \mathcal{W}_i(t_1)$, with $\mathcal{W}_{3-i}(t_1) = 0$ for $i = 1$ or $i = 2$. Note that $K(t_1) = \mathcal{W}_i(t_1) = \mathcal{W}_1(t_1) + \mathcal{W}_2(t_1) = \mathcal{W}(t_1) + L(t_1) \geq \mathcal{W}(t_1)$. In the next period, by the same argument, $0 \leq L(t) \leq K(t_1)$. This can be repeated with $t_0 < t_1 < \cdots t_k < \cdots$ being successive times at which G/G_2/1 is busy and in the JSQ system one of the servers becomes idle. This shows that $L(t) \leq \sup_{0\leq s\leq t} K(s)$ holds throughout the busy period.

15.10 Complete the proof of proposition 15.3.

Solution

Because the longest queue is not receiving JSQ customers, we can assume

that $\bar{\mathcal{W}}_1(0) = \bar{\mathcal{W}}_2(0) = 0$, and thereafter $\bar{\mathcal{W}}_i(t) = 0$, and we have $\hat{\mathcal{W}}_i^n(t) = \mathcal{W}_i^n(t)/\sqrt{n}$.

We will show that $|\hat{\mathcal{W}}_1^n(t) - \hat{\mathcal{W}}_2^n(t)| \to_p 0$ u.o.c., which is equivalent to $|\hat{\mathcal{W}}_1^n(t) - \hat{\mathcal{W}}_2^n(t)| \to_w 0$. We wish to show for $t > 0$ and $\epsilon > 0$:

$$\mathbb{P}(\sup_{0<s\le t} |\hat{\mathcal{W}}_1^n(t) - \hat{\mathcal{W}}_2^n(t)| > \epsilon) \to 0, \quad \text{as } n \to \infty.$$

Assume first that $\hat{\mathcal{Q}}(0) = 0$. Then for n large enough, $\hat{\mathcal{Q}}^n(0) < \epsilon/2$. Consider the event $\sup_{0<s\le t} \hat{\mathcal{Q}}^n(s) > \epsilon$. If $\hat{\mathcal{Q}}^n(0) < \epsilon/2$ and $\sup_{0<s\le t} \hat{\mathcal{Q}}^n(s) > \epsilon$, then we can define $0 \le \tau^{*n} < \tau^n \le t$ such that:

$$\tau^n = \inf\{s : \hat{\mathcal{Q}}^n(s) > \epsilon\}, \qquad \tau^{*n} = \sup\{s : 0 \le s < \tau^n,\ \hat{\mathcal{Q}}^n(s) \le \epsilon/2\}.$$

The rest of the proof follows the same steps as the proof of Theorem 6.3.

15.11 Complete the proof of Theorem 15.1 without assuming $\lambda_1 = \lambda_2 = 0$.

Solution

The proof needs only minor changes. We now have in addition to the JSQ customers also arrivals to either of the queues. As long as both servers work in both systems, $L(t)$ and $K(t)$ do not change. In those periods when G/G/2 has two busy servers and JSQ only one, we now need to consider also arrivals of server 1 and of server 2. Assume t_k is a time when $\mathcal{W}_2(t_k)$ reaches 0, $\mathcal{W}_1(t_k) > 0$, in G/G/2 both servers busy, and we have: $L(t_k) \ge 0$, $\mathcal{W}_1(t_k) \ge \mathcal{W}(t_k)$, and consider the following period, in which $L(t)$ increases and $K(t)$ decreases, but $L(t) \le K(t) \le K(t_k)$ is maintained. Then we need to consider two possibilities: we may have an arrival that joins server 2 in both systems (it may be a server 2 customer or a JSQ customer), and both servers are busy in both systems, and $L(t), K(t)$ stop changing. Or it may be a server 1 customer that joins server 1 in both systems at time s, and requires service S. When he joins he leaves $\mathcal{W}_2(t) = 0$, $K(t)$ increases by S, $L(t)$ remains unchanged, and $\mathcal{W}(t)$ increases by S. Then the following period, when $L(t)$ increases at rate 1 and $K(t)$ decreases by 1, can only last for a time $(\mathcal{W}(s-) + S)/2$ and $L(t) \le \max\{K(t_k), K(s)\}$ is maintained.

Consider then the sequence of times $t_0, \ldots, t_k, \ldots$ at which one JSQ server becomes idle and both G/G/2 servers are busy, and the times $s_1, \ldots, s_l, \ldots$, at which a customer joins the longer queue when the other queue is empty, and both G/G/2 servers are busy. The at all times, $L(t) \le \max\left\{L(s) : s \in \{t_0, t_1, \ldots,\ s_1, s_2, \ldots\}\right\} \le \sup_{0\le s\le t} K(s)$. This completes the proof.

15.12 Prove that similar to Propositions 15.2, 15.3 about workloads, the following theorem holds for the queue lengths:

(i) Under join the shortest queue, $|\hat{Q}_1^n - \hat{Q}_2^n| \to_p 0$ as $n \to \infty$.

(ii) $\hat{Q}_1^n + \hat{Q}_2^n \to_w \text{RBM}(\theta, \sum \lambda_i c_{a,i}^2 + \sum c_{s,i}^2)$ as $n \to \infty$.

Solution

Let $Q(t)$ be the queue length of the coupled G/G/2 system. We then have:

(i) $Q_1(t) + Q_2(t) \geq Q(t)$, (ii) By $|\hat{\mathcal{W}}_1(t) - \hat{\mathcal{W}}_2(t)| \rightarrow_p 0$ we also have $|\hat{Q}_1(t) - \hat{Q}_2(t)| \rightarrow_p 0$, (iii) $\hat{Q}(t) \rightarrow_w \text{RBM}(\theta, \sum \lambda_i c_{a,i}^2 + \sum c_{s,i}^2)$ as $n \rightarrow \infty$. The Theorem follows.

16

Control in Balanced Heavy Traffic

Exercises

16.1 Derive the diffusion limits for the MCQN netput process $\frac{1}{N}\mathcal{X}(N^2t)$, using the property that in balanced heavy traffic $\lim_{N\to\infty}\frac{1}{N}\mathcal{T}_k(Nt) = \nu_k t$ u.o.c. a.s..

Solution

We have

$$\mathcal{X}_k(t) = \big(\mathcal{A}_k(t) - \alpha_k t\big) - \big(\mathcal{S}_k(\mathcal{T}_k(t)) - \mu_k\mathcal{T}_k(t)\big) + \sum_{l=1}^{K}\big(\mathcal{R}_{l,k}(\mathcal{S}_l(\mathcal{T}_l(t))) - p_{l,k}\mu_l\mathcal{T}_l(t)\big)$$

We then have, by the FCLT for renewal processes:

$$\frac{1}{N}\big(\mathcal{A}_k(N^2t) - \alpha_k n^2 t\big) \to_w (\alpha_k c_a^2)^{1/2} BM(t).$$

Also, since $\frac{1}{N}\mathcal{T}(Nt) \to t$ u.o.c. a.s., using time change, we have:

$$\begin{aligned}&\frac{1}{N}\big(\mathcal{S}_k(\mathcal{T}_k(N^2t)) - \mu_k\mathcal{T}_k(N^2t)\big)\\ &\sim \frac{1}{N}\big(\mathcal{S}_k(N^2t)) - \mu_k\mathcal{T}_k(N^2t)\big) \to_w (\mu_k c_s^2)^{1/2} BM(t),\end{aligned}$$

and, since $\frac{1}{N}\mathcal{S}_l(\mathcal{T}_l(Nt)) \to \mu t$ u.o.c. a.s., using time change, we have:

$$\begin{aligned}&\frac{1}{N}\big(\mathcal{R}_{l,k}(\mathcal{S}_l(\mathcal{T}_l(N^2t))) - p_{l,k}\mu_l\mathcal{T}_l(N^2t)\big)\\ &\sim \frac{1}{N}\big(\mathcal{R}_{l,k}(\lfloor\mu_l N^2 t\rfloor) - p_{l,k}\mu_l N^2 t\big) \to_w \big(\mu_l p_{l,k}(1-p_{l,k})\big)^{1/2} BM(t)\end{aligned}$$

The diffusion limits of arrivals and of service times for all k are independent. The diffusion limits of the quantities $\mathcal{R}_{l,k}, k = 1,\dots,K$ are independent of the arrivals and service time limits, but are correlated between themselves. As a result:

$$\frac{1}{N}\mathcal{X}_k(N^2t) \to_w \Big(\alpha_k c_a^2 + \mu_k c_s^2 + \sum_{l\in\mathcal{K}}\mu_l p_{l,k}(1-p_{l,k})\Big)^{1/2} BM(t).$$

16.2 Verify equation (16.2) for the drift and covariance of the netput process,
Solution
The calculations are similar to the derivations of (9.17), Theorem 9.4 in Section 9.5 and to Exercise 9.5. See also Chen and Mandelbaum (1991, 1994).

16.3 Show that the LP (16.8), that solves for $\hat{Q}(t)$ in terms of $\hat{\mathcal{W}}(t)$, is feasible and bounded.
Solution
In the LP (16.8) the coefficients matrix M is non-negative full rank, and the r.h.s. $\hat{\mathcal{W}}(t) \geq 0$. This proves the solution is bounded and the dual is feasible. The dual is:

$$\begin{aligned} \max & \sum_{i=1}^{I} \hat{\mathcal{W}}_i(t)\pi_i \\ & M^{\top}\pi_i \leq h_k, \end{aligned}$$

which also has non-negative coefficient matrix where every row has some positive terms, and non-negative r.h.s., and although the variables π_i are not restricted in sign, its objective coefficients are non-negative, so the maximizing optimal solution is bounded above. This implies that the primal is feasivle.

16.4 Consider the criss-cross network with Poisson arrivals and exponential service times. Assume that customers that arrive at station 1 are served on FCFS basis. Analyze the performance of the network under that policy.
Solution
This is a feed forward Jackson type network. The situation in station 1 is that it acts like an M/M/1 queue with $\lambda = 2\rho$, $\mu = 2$. So in steady state, the number of customers in queue 1 is

$$\mathbb{P}(Q_1 + Q_2 = n) = (1-\rho)\rho^n,$$

with expectation $\frac{\rho}{1-\rho}$.
conditional on $Q_1 + Q_2 = n$ the number of those in Q_1 is $\sim$Bin$(n, 1/2)$, since they arrived randomly and are served in order of arrival, so:

$$\mathbb{P}(Q_1 = k, Q_2 = l) = (1-\rho)\rho^{k+l}\binom{k+l}{k}\left(\frac{1}{2}\right)^{k+l} = \binom{k+l}{k}(1-\rho)\left(\frac{\rho}{2}\right)^{k+l},$$

The marginal distributions of $Q_1$is:

$$\begin{aligned}\mathbb{P}(Q_1 = k) &= \sum_{l=0}^{\infty} \binom{k+l}{k} (1-\rho) \left(\frac{\rho}{2}\right)^{k+l} \\ &= (1-\rho) \left(\frac{\rho}{2}\right)^{k} \sum_{l=0}^{\infty} \binom{k+l}{k} \left(\frac{\rho}{2}\right)^{l} \\ &= (1-\rho) \left(\frac{\rho}{2}\right)^{k} \left(1-\frac{\rho}{2}\right)^{-(k+1)} \\ &= \left(1-\frac{\rho}{2-\rho}\right) \left(\frac{\rho}{2-\rho}\right)^{k},\end{aligned}$$

by the binomial theorem formula $\frac{1}{(1-x)^{k+1}} = \sum_{n=0}^{\infty} \binom{n+k}{k} x^n$. The marginal distribution of Q_2 is the same.

It is interesting to note that each behaves like M/M/1 with arrival rate ρ and service rate $2-\rho$, which is the service rate left over after using rate ρ for the other class. This is a higher rate than the rate 1 it would get if it were on its own. This is exactly the pooling effect.

Because this is a feed forward network, all the streams of customers are Poisson, so input to station 2 is Poisson rate ρ, with service rate 1.

$$\mathbb{P}(Q_3 = m) = (1-\rho)\rho^m.$$

The value of the objective:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T (Q_1(t) + Q_2(t) + Q_3(t))dt = \mathbb{E}(Q_1 + Q_2 + Q_3) = \frac{2\rho}{1-\rho}$$

The sojourn time in station 1, for either type of customer is $\frac{1}{2}\frac{1}{1-\rho}$. The soujourn time for customers of type B composed of the two stations is $\frac{3}{2}\frac{1}{1-\rho}$.

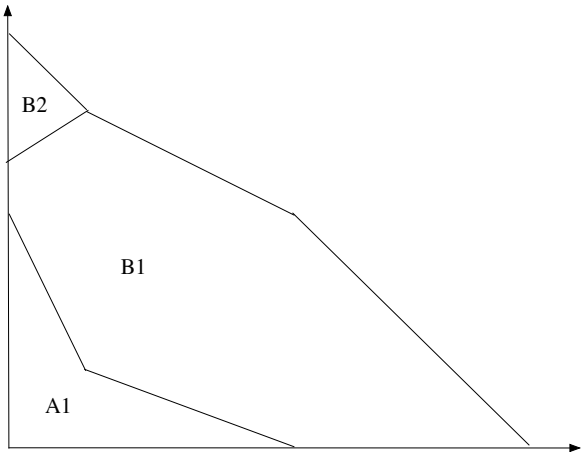
16.5 Consider the criss-cross network with Poisson arrivals and exponential service times. Analyze the performance of the network under the policy of priority to class 1 and under the policy of priority to class 2.

Solution

Under priority to customers of type B at station 1, the flow of customers of type B will be as follows: They experience M/M/1 service at the first station, with sojourn time exponential of rate $2-\rho$, and arrive in a Poisson stream to station 2, and have sojourn time time exponential of rate $1-\rho$ at station 2, and the two quantities are independent. The customers of type A will have a much longer waiting time: The average waiting time will be (see equation (3.5)) $\bar{V} = \frac{2\rho\cdot 1/2\cdot 2/4}{(1-\rho/2)(1-rho)}$, and the sojourn time will be $\frac{2-\rho+\rho^2}{2-3\rho+\rho^2}$.

Under priority to customers of type A at station 1, the sojourn times at station 1 will be reversed. The sojourn time at station 2 will however be longer, since the arrival process will no longer be Poisson but will be much more variable.

16.6 For the criss-cross network, plot a schematic path of the buffer contents, using drift and deviations from the drift, to illustrating all possible states.
Solution
The following is a schematic picture of how fluid will empty under the optimal Brownian policy.



16.7 For the criss-cross network calculate the drift and the covariances for the netput Brownian motion $\hat{\mathcal{X}}(t)$ and for the netput Brownian workload $\hat{\mathcal{Z}}(t)$.
Solution
The scaled drift is $\hat{d}_k t = N(\rho - 1)t$ for each the queue lengths at station 1, which approximates $\hat{dt} = \frac{1}{N} dN^2 t) = N(\rho - 1)$ from which we get the the original drifts are $d(t) = \rho - 1$. The drift for the third queue is 0. The covariance matrix of the $\mathcal{X}(t)$, recalling that for exponential service and interarrivals $c_{a,i}$, $c_{s,k} = 1$, is:

$$\Gamma_{\mathcal{X}} = \begin{bmatrix} 2\rho & 0 & 0 \\ 0 & 2\rho & -\rho \\ 0 & -\rho & 2\rho \end{bmatrix}$$

The drifts of $Z(t)$ is:

$$1/2(\rho - 1) + 1/2(rho - 1) = \rho - 1; \qquad (\rho - 1) + 0 = \rho - 1.$$

The covariance matrix is:

$$\Gamma_{\mathcal{Z}} = \begin{bmatrix} \rho & \frac{1}{2}\rho \\ \frac{1}{2}\rho & 2\rho \end{bmatrix}$$

16.8 Obtain the marginal distributions of the Brownian workloads $\hat{\mathcal{W}}_i(t)$ for the criss-cross network under the optimal policy. Note they are not independent and we cannot obtain their joint distribution. Try and estimate the mean objective.

Solution
We have: $\hat{\mathcal{W}}_i(t) = \hat{\mathcal{Z}}_i(t) + \hat{\mathcal{I}}_i(t)$, where $\hat{\mathcal{Z}}_i(t)$ is a one dimensional Brownian motion with known drift and variance, and $\hat{\mathcal{W}}_i(t)$ is its one sided Skorohod reflection, i.e. $\hat{\mathcal{W}}_i(t) \sim RBM$. The drifts are negative, so the stationary distributions are exponential. for drift $m < 0$ and variance σ^2 we get $\sim\text{Exp}(-\frac{2m}{\sigma^2})$:

$$\hat{\mathcal{W}}_1(t) \sim \text{Exp}\left(\frac{2(1-\rho)}{\rho}\right), \qquad \hat{\mathcal{W}}_2(t) \sim \text{Exp}\left(\frac{(1-\rho)}{\rho}\right)$$

with means and standard deviations: $\frac{\rho}{2(1-\rho)}$ and $\frac{\rho}{1-\rho}$.
Under heavy traffic the lower bound (**??**) should approximate the actual objective:

$$\mathbb{E}(Q_1(t) + Q_2(t) + Q_3(t)) \approx \mathbb{E}\big(Q_1(t) + Q_2(t) \vee Q_2(t) + Q_3(t)\big)$$
$$= \mathbb{E}\big(2\mathcal{W}_1(t) \vee \mathcal{W}_2\big) \leq \frac{3}{2}\frac{\rho}{1-\rho},$$

Because both $2\mathcal{W}_1(t)$ and $\mathcal{W}_2(t)$ are exponential random variables, with the same parameter, and they are positively correlated. For two independent rate λ exponential random variables the maximum has expectation $\frac{3}{2}\frac{1}{\lambda}$. If they are positively correlated we get a value that is between $\frac{1}{\lambda} < \frac{3}{2}\frac{1}{\lambda}$

16.9 Consider a modified criss-cross network, where customers of type B can start their service at station 2 without waiting to complete service at station 1, and customers of type B leave the system when both services are complete. Assume Poisson arrivals and exponential service times, and server 1 is giving non-preemptive priority to customers of type A. Calculate the expected number of customers in the system as $\rho \nearrow 1$.

Solution
This modified system is obviously an improvement on the actual system. The expected number of customers waiting to complete service by server 1 is $\frac{\rho}{1-\rho}$. In addition there will be Y customers of type B that have been served by server 1, and are still waiting for server 2. Since customers of type B have lower priority for service by server 1 and have no competition for server 2, Y will usually be small, and so we can ignore Q_3. For Q_1 and Q_2 the total number in workstations 1 is independent of policy, and is queue length for M/M/1 with traffic intensity ρ. Hence: $\mathbb{E}(Q_1 + Q_2 + Q_3) \approx \frac{\rho}{1-\rho}$. This lower bound is a factor 2 improvement on FCFS.
Together with the previous exercise we get the expected objective is bounded between $\frac{\rho}{1-\rho}, \frac{3}{2}\frac{\rho}{1-\rho}$ as compared to $2\frac{\rho}{1-\rho}$ for FCFS.
We can also calculate the values of the (approximate) expected waiting times for type A and B, using the equations for waiting times in priority M/G/1, Section **??**, Equation (**??**). From it we get (ignoring station 2):

$$\bar{W}_A \approx 2.5, \qquad \bar{W}_B \approx \frac{2}{2-r}.$$

16.10 Consider the criss-cross network under the policy of Maximum pressure, and compare this to the proposed threshold policy. Evaluate both by simulation.

Solution

The input output matrix is:

$$R = (I - P^{\mathsf{T}})\mathrm{diag}(\mu) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & -2 & 1 \end{bmatrix}.$$

For buffer contents Q we have:

$$\begin{aligned} Q^{\mathsf{T}}R &= \begin{bmatrix} Q_A & Q_{B1} & Q_{B2} \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2Q_A, & 2(Q_{B1} - Q_{B2}), & Q_{B2} \end{bmatrix}. \end{aligned}$$

The maximization is over $u_1 + u_2 \leq 1$ and $u_3 \leq 1$, so maximum pressure policy will require:

– At station 2, work fully on buffer B2.

– At station 1, work on A when $Q_A > Q_{B1} - Q_{B2}$,

– At station 1, work on B1 when $Q_{B1} - Q_{B2} > Q_A$.

This works out as follows: Always work on B2, which will change at rate $u_2 - 1$. While $Q_A > Q_{B1} - Q_{B2}$ work on A, so it empties at rate $\rho - 2$. Meanwhile pressure at Q_{B1} increases if buffer $B2$ is emptying, or stays constant if buffer $B2$ is empty, so eventually equality $Q_A = Q_{B1} - Q_{B2}$ is reached. If $Q_{B1} - Q_{B2} > Q_A$ work on B1, so Q_{B1} will empty at rate $\rho - 2$, while Q_{B2} will fill up at rate $2-1$, so also in this case equality $Q_A = Q_{B1} - Q_{B2}$ will be reachd. Once equality is reached, work on both buffers A and B1, with $u_1 = u_2 = 0.5$ so that buffer B2 will receive input at rate 1, and process at rate 1, which means it will stay constant (on the fluid scale), so maximum pressure will equalize the pressure at station 1, and share the processing at station 1 equally. This will continue until buffer A will become empty, and $Q_{B1} = Q_{B2}$.

When fluid is 0, maximum pressure will be stable, but it is harder to say how random fluctuation will affect the processing.

It is significant to note that even when B2 is above a threshold, i.e. it is in no danger of starvation, maximum pressure will not give priority to A, but will share processing of A and B1. This is in contrast to the Brownian inspired threshold policy.

16.11 Consider the criss-cross network with general parameters, arrival rates α_i, service rates μ_k. Assume renewal arrivals and services, with coefficients of variation $c_{a,i}, c_{s,k}$. Follow the next steps:

(a) Write the dynamics of $Q_k(t)$, using nominal rates ν_k.

(b) Define netput processes $\mathcal{X}_k(t)$, and write $Q_k(t)$ in terms of the netput $\mathcal{X}_k(t)$ and the free times $\mathcal{J}_k(t)$.

(c) Write the dynamics of the workload processes $\mathcal{W}_i(t)$, define workload netput $\mathcal{Z}_i(t)$, and write $\mathcal{W}_i(t)$ in terms of the workload netput $\mathcal{Z}_i(t)$ and the idle time processes $\mathcal{I}_i(t)$.
(d) Calculate the offered load for each of the servers, ρ_i.
(e) Define a sequence of systems indexed by N. Use fluid scaling with time and space scaled by N, and obtain the fluid limits of $\mathcal{Q}^N, \mathcal{X}^N, \mathcal{J}^N$ and of $\mathcal{W}^N, Z^N, \mathcal{I}^N$.
(f) Formulate the conditions for balanced heavy traffic and determine the diffusion approximation to $\mathcal{X}(t)$ and $\mathcal{Z}(t)$.
(g) Calculate the drift and and covariance matrix for the diffusion limits $\hat{\mathcal{X}}$ and $\hat{\mathcal{Z}}$.

Solution

(a,b) The nominal rates are: $\nu_1 = \frac{\alpha_1 m_1}{\alpha_1 m_1 + \alpha_2 m_2}$, $\nu_2 = \frac{\alpha_2 m_2}{\alpha_1 m_1 + \alpha_2 m_2}$, $\nu_3 = 1$. If the system is to be rate stabe, server 1 needs to devote a fraction ν_1 of its time to class 1, and ν_2 of its time to class 2.

$$\begin{aligned}
\mathcal{Q}_1(t) &= [(\alpha_1 - \mu_1\nu_1)t + (\mathcal{A}_1(t) - \alpha_1 t) - (\mathcal{S}_1(\mathcal{T}_1(t)) - \mu_1\mathcal{T}_1(t))] \\
&\quad + [\mu_1(\nu_1 t - \mathcal{T}_1(t)] = \mathcal{X}_1(t) + \mu_1\mathcal{J}_1(t) \\
\mathcal{Q}_2(t) &= [(\alpha_2 - \mu_2\nu_2)t + (\mathcal{A}_2(t) - \alpha_2 t) - (\mathcal{S}_2(\mathcal{T}_2(t)) - \mu_2\mathcal{T}_2(t))] \\
&\quad + [\mu_2(\nu_2 t - \mathcal{T}_2(t)] = \mathcal{X}_2(t) + \mu_2\mathcal{J}_2(t) \\
\mathcal{Q}_3(t) &= [(\mu_2\nu_2 - \mu_3)t + (\mathcal{S}_2(\mathcal{T}_2(t)) - \mu_2\mathcal{T}_2(t)) - (\mathcal{S}_3(\mathcal{T}_2(t)) - \mu_3\mathcal{T}_3(t))] \\
&\quad - [\mu_2(\nu_2 t - \mathcal{T}_2(t)] + [\mu_3(t - \mathcal{T}_3(t)] = \mathcal{X}_3(t) + \mu_3\mathcal{J}_3(t) - \mu_2\mathcal{J}_2(t)
\end{aligned}$$

Note that $\mathcal{Q}_3$ includes a combination of its own free time and the free time of $\mathcal{Q}_2$ in its decomposition.

(c) The workloads are $\mathcal{W}(t) = M\mathcal{Q}(t)$, with $M = \begin{bmatrix} m_1 & m_2 & 0 \\ 0 & m_3 & m_3 \end{bmatrix}$ where $m_i = 1/\mu_i$ are mean processing times. We have:

$$\begin{aligned}
\mathcal{W}_1(t) &= m_1\mathcal{X}_1(t) + m_2\mathcal{X}_2(t) + \mathcal{J}_1(t) + \mathcal{J}_2(t) \\
&= (\alpha_1 m_1 + \alpha_2 m_2 - 1)t + [m_1(\mathcal{A}_1(t) - \alpha_1 t) + m_2(\mathcal{A}_2(t) - \alpha_2 t) \\
&\quad - m_1((\mathcal{S}_1(\mathcal{T}_1(t)) - \mu_1\mathcal{T}_1(t)) - m_2(\mathcal{S}_2(\mathcal{T}_2(t)) - \mu_2\mathcal{T}_2(t))] \\
&\quad + [t - \mathcal{T}_1(t) - \mathcal{T}_2(t)] = \mathcal{Z}_1(t) + \mathcal{I}_1(t) \\
\mathcal{W}_2(t) &= m_3\,[\mathcal{X}_2(t) + \mathcal{X}_3(t)] + \mathcal{J}_3(t) \\
&= (\alpha_2 m_3 - 1)t + [m_3(A_2(t) - \alpha_2 t) - m_3(\mathcal{S}_3(\mathcal{T}_3(t)) - \mu_3\mathcal{T}_3(t))] \\
&\quad + [t - T_3(t)] = \mathcal{Z}_2(t) + \mathcal{I}_2(t)
\end{aligned}$$

(d) $\rho_1 = \alpha_1 m_1 + \alpha_2 m_2$, $\rho_2 = \alpha_2 m_3$. We assume that $\rho_i < 1$
(e) We assume that $\mathcal{Q}^N(0) = 0$, and the arrival and service processes are obtained from the same sequence of i.i.d. interarriaval and service times, differentiated only by their scaling (using different rates) of the parameters. We define fluid scaling of $z(t)$ by $\bar{z}(t) = \frac{1}{N}z(Nt)$. Then, by the FSLLN, as

$N \to \infty$, u.o.c. a.s. we get the fluid limits:

$$\bar{\mathcal{T}}_1(t) = \nu_1\rho_1 t, \quad \bar{\mathcal{T}}_2(t) = \nu_2\rho_1 t, \quad \bar{\mathcal{T}}_3(t) = \rho_2 t,$$
$$\bar{\mathcal{J}}_k(t) = \nu_k(1-\rho_{s(k)})t, \ k = 1,2,3,$$
$$\bar{Q}_1(t) = (\alpha_1 - \nu_1\mu_1)^+ t, \quad \bar{Q}_2(t) = (\alpha_2 - \nu_2\mu_2)^+ t, \quad \bar{Q}_3(t) = (\nu_2\mu_2 - \mu_3)^+ t,$$
$$\bar{\mathcal{Z}}_1(t) = (\alpha_1 + \alpha_2 - \nu_1\mu_1 - \nu_2\mu_2)^+ t, \quad \bar{\mathcal{Z}}_2(t) = (\alpha_2 - \mu_3)^+ t, \quad \bar{I}_i(t) = (1-\rho_i)t, \ i = 1,2.$$

Here we assumed that the policy is work conserving, and that the system is made rate stable, so we follow the nominal rates. In particular, by $\rho_i < 1$, since we start from $\bar{Q}(0) = 0$, we have $\bar{Q}_k(t) = 0, \ k = 1,2,3, \quad \bar{\mathcal{Z}}_i(t) = 0, \ i = 1,2$.

(f) The system will be in balanced heavy traffic if $N(1-\rho_i)$ is of moderate size for some large N. In that case the scaled netputs, $\frac{1}{N}\mathcal{X}(N^2 t)$ and $\frac{1}{N}\mathcal{Z}(N^2 t)$, can be approximated by Brownian motions $\hat{\mathcal{X}}_k^N(t), \hat{\mathcal{Z}}_i^N(t)$.

$$\hat{\mathcal{X}}_1(t) \Longrightarrow \theta_1 t + \left(\alpha_1(c_{a,1}^2 + c_{s,1}^2)\right)^{1/2} BM(t)$$
$$\hat{\mathcal{X}}_2(t) \Longrightarrow \theta_2 t + \left(\alpha_2(c_{a,2}^2 + c_{s,2}^2)\right)^{1/2} BM(t)$$
$$\hat{\mathcal{X}}_3(t) \Longrightarrow (\theta_3 - \theta_2)t + \left(\alpha_2(c_{s,2}^2 + c_{s,3}^2)\right)^{1/2} BM(t).$$
$$\hat{\mathcal{Z}}_1(t) \Longrightarrow (m_1\theta_1 + m_2\theta 2)t + \left(m_1^2\alpha_1(c_{a,1}^2 + c_{s,1}^2) + m_2^2\alpha_2(c_{a,2}^2 + c_{s,2}^2)\right)^{1/2} BM(t).$$
$$\hat{\mathcal{Z}}_2(t) \Longrightarrow m_3\theta_3 t + m_3\left(\alpha_2(c_{a,2}^2 + c_{s,3}^2)\right)^{1/2} BM(t).$$

where we define:

$$\theta_1 = N(\alpha_1 - \nu_1\mu_1), \quad \theta_2 = N(\alpha_2 - \nu_2\mu_2), \quad \theta_3 = N(\nu_2\mu_2 - \nu_3\mu_3),$$

The renewal processes $\mathcal{A}_i$ $\mathcal{S}_k$, counting arrivals and service completions at rates α_i μ_k, with interval c.o.v. $c_{a,i}$ and $c_{s,k}$. Then, because $\bar{\mathcal{T}}_k(t) \to \nu_k t, \ k = 1,2$, we have that $\frac{1}{N}(\mathcal{S}_k(\mathcal{T}_k(N^2 t)) - \mu_k\mathcal{T}_k(N^2 t)) \Longrightarrow (\mu_k c_{s,k}^2)^{1/2} BM(\nu_k t) =_D (\nu_k\mu_k c_{s,k}^2)^{1/2} BM(t)$.

(g) For $\hat{Q}(t)$ we obtained the drifts already. The covariances are:

$$\Gamma = \begin{bmatrix} \alpha_1(c_{a,1}^2 + c_{s,1}^2) & 0 & 0 \\ 0 & \alpha_2(c_{a,2}^2 + c_{s,2}^2) & -\alpha_2 c_{s,2}^2 \\ 0 & -\alpha_2 c_{s,2}^2 & \alpha_2(c_{s,2}^2 + c_{s,3}^2) \end{bmatrix}$$

For $\hat{\mathcal{Z}}(t)$ the drifts are: $m_1\theta_1 + m_2\theta_2 = N(1-\rho_1)$ and $m_3\theta_2 + \theta_2 = N(1-\rho_2)$. The covariances are $M\Gamma M^T t$:

$$\begin{bmatrix} m_1^2\alpha_1(c_{a,1}^2 + c_{s,1}^2) + m_2^2\alpha_2(c_{a,2}^2 + c_{s,2}^2) & m_2 m_3\alpha_2 c_{a,2}^2 \\ m_2 m_3\alpha_2 c_{a,2}^2 & m_3^2\alpha_2(c_{a,2}^2 + c_{s,3}^2) \end{bmatrix}$$

16.12 Formulate the BCP of the general criss-cross network in heavy traffic, and solve it for general parameters, $\alpha_i, i = 1, 2, \mu_k, k = 1, 2, 3$ and cost coefficients $h_k, k = 1, 2, 3$. Note that there are several cases, according to the parameters.

Solution

The general formulation is, find Q_k, I_i such that:

$$\begin{aligned}
\min \ & \lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{k=1}^{3} h_k \hat{Q}_k(t) dt \\
\text{s.t.} \ & m_1 \hat{Q}_1(t) + m_2 \hat{Q}_2(t) = \hat{Z}_1(t) + \hat{\mathcal{T}}_1(t), \quad t \geq 0 \\
& m_3 \hat{Q}_2(t) + m_3 \hat{Q}_3(t) = \hat{Z}_2(t) + \hat{\mathcal{T}}_2(t), \quad t \geq 0 \\
& \hat{Q}(t) \geq 0, \qquad \hat{\mathcal{T}}(0) = 0, \quad \hat{\mathcal{T}} \text{ non-decreasing} \quad t \geq 0 \\
& \hat{Q}, \hat{\mathcal{T}} \text{ non-anticipating with respect to } \hat{\mathcal{X}}(t)
\end{aligned}$$

and for given $\hat{I}_1(t), \hat{I}_2(t)$, we can solve it separately for each t, find $\hat{Q}_k$:

$$\begin{aligned}
\min \ & V = \sum_{k=1}^{3} h_k \hat{Q}_k \\
\text{s.t.} \ & m_1 \hat{Q}_1 + m_2 \hat{Q}_2 = \hat{W}_1, \\
& m_3 \hat{Q}_2 + m_3 \hat{Q}_3 = \hat{W}_2, \\
& \hat{Q} \geq 0,
\end{aligned}$$

It is useful to look at the dual problem, find y_1, y_2 :

$$\begin{aligned}
\max \ & V = \sum_{i=1}^{2} \hat{\mathcal{W}}_i y_i, \\
\text{s.t.} \ & m_1 y_1 \leq h_1, \\
& m_2 y_1 + m_3 y_2 \leq h_2, \\
& m_3 y_2 \leq h_3.
\end{aligned}$$

The following figure illustrates the possible extreme point solutions for the dual,

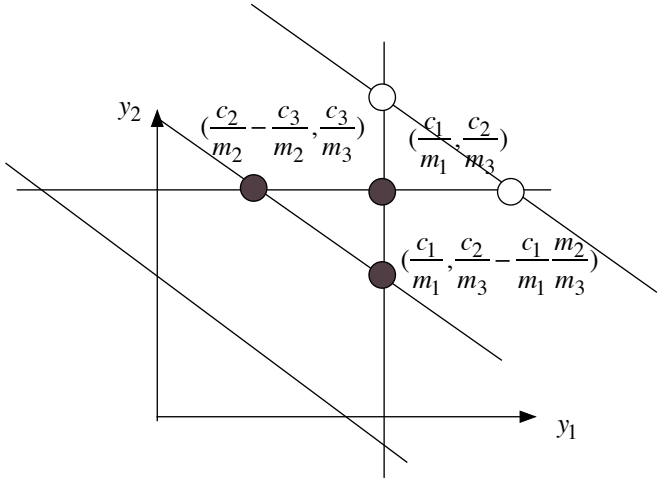
The solutions for the dual are:

Case I: If $h_1 \frac{m_2}{m_1} + h_3 \leq h_2$, then $y_1 = \frac{h_1}{m_1}$, $y_2 = \frac{h_2}{m_3}$, and the primal solution is: $\hat{Q}_1 = \frac{\hat{\mathcal{W}}_1}{m_1}$, $\hat{Q}_2 = 0$, $\hat{Q}_3 = \frac{\hat{\mathcal{W}}_2}{m_3}$. The interpretation of this is that we always give priority to class 2 at station 1, so it is always almost empty, and customers are mainly kept in class 1 and class 3.

Case II: If $h_1 \frac{m_2}{m_1} + h_3 > h_2$, there are two extreme points, corresponding to the solutions:

Case IIa:

$$(y_1, y_2) = (\frac{h_2}{m_2} - \frac{h_3}{m_2}, \frac{h_3}{m_3}), \qquad (\hat{Q}_1, \hat{Q}_2, \hat{Q}_3) = (0, \frac{\hat{\mathcal{W}}_1}{m_2}, \frac{\hat{\mathcal{W}}_2}{m_3} - \frac{\hat{\mathcal{W}}_1}{m_2}),$$



and objective value:

$$V_a = \hat{\mathcal{W}}_1(\frac{h_2}{m_2} - \frac{h_3}{m_2}) + \hat{\mathcal{W}}_2\frac{h_3}{m_3}$$

Case IIb:

$$(y_1, y_2) = (\frac{h_1}{m_1}, \frac{h_2}{m_3} - \frac{h_1}{m_1}\frac{m_2}{m_3}), \qquad (\hat{Q}_1, \hat{Q}_2, \hat{Q}_3) = (\frac{\hat{\mathcal{W}}_1}{m_1} - \frac{\hat{\mathcal{W}}_2}{m_3}\frac{m_2}{m_1}, \frac{\hat{\mathcal{W}}_2}{m_3}, 0),$$

and objective value:

$$V_b = \hat{\mathcal{W}}_1\frac{h_1}{m_1} + \hat{\mathcal{W}}_2(\frac{h_2}{m_3} - \frac{h_1}{m_1}\frac{m_2}{m_3}))$$

We will have:

$$V_b - V_a = \left(h_1\frac{m_2}{m_1} + h_3 - h_2\right)\left(\frac{\hat{\mathcal{W}}_1}{m_2} - \frac{\hat{\mathcal{W}}_2}{m_3}\right)$$

because $h_1\frac{m_2}{m_1} + h_3 - h_2 > 0$, we have that $V_b > V_a \iff \frac{\hat{\mathcal{W}}_1}{m_2} > \frac{\hat{\mathcal{W}}_2}{m_3}$, which is exactly the case that IIb is feasible, and $V_a > V_b \iff \frac{\hat{\mathcal{W}}_2}{m_3} > \frac{\hat{\mathcal{W}}_1}{m_2}$, which is exactly the case that IIa is feasible.

The solution for case II is then:

$$\hat{Q}_2(t) = \frac{\hat{\mathcal{W}}_2}{m_3} \vee \frac{\hat{\mathcal{W}}_1}{m_2}, \qquad \hat{Q}_1 = \frac{m_2}{m_1}\left(\frac{\hat{\mathcal{W}}_1}{m_2} - \frac{\hat{\mathcal{W}}_2}{m_3}\right)^+, \qquad \hat{Q}_3 = \left(\frac{\hat{\mathcal{W}}_2}{m_3} - \frac{\hat{\mathcal{W}}_1}{m_2}\right)^+,$$

$$\begin{aligned} V &= \left(\hat{\mathcal{W}}_1\frac{h_1}{m_1} + \hat{\mathcal{W}}_2\left(\frac{h_2}{m_3} - \frac{h_1}{m_1}\frac{m_2}{m_3}\right)\right) \vee \left(\hat{\mathcal{W}}_1\left(\frac{h_2}{m_2} - \frac{h_3}{m_2}\right) + \hat{\mathcal{W}}_2\frac{h_3}{m_3}\right) \\ &= h_2\left(\frac{\hat{\mathcal{W}}_2}{m_3} \vee \frac{\hat{\mathcal{W}}_1}{m_2}\right) + h_1\frac{m_2}{m_1}\left(\frac{\hat{\mathcal{W}}_1}{m_2} - \frac{\hat{\mathcal{W}}_2}{m_3}\right)^+ + h_3\left(\frac{\hat{\mathcal{W}}_2}{m_3} - \frac{\hat{\mathcal{W}}_1}{m_2}\right)^+. \end{aligned}$$

16.13 For the closed two station multi-class network show that to maximize throughput one should use a work conserving policy.

Solution

Assume at time t we can start job in buffer k, and instead we idle for a period δ. Without loss of generality, assume that we count completion of circulations at buffer k. Then starting the job earlier will in crease the circulation at its completion, without interfering with anything else.

16.14 For the closed two station multi-class network, write down the dynamics, center and scale all components and derive a decomposition with a netput process and control processes. Derive the drift and covariance matrix of the limiting Brownian netput process.

Solution

For buffer k we have:

$$\begin{aligned} Q_k(t) &= Q_k(0) - \mathcal{S}_k(\mathcal{T}_k(t)) + \sum_{l=1}^{K} \mathcal{R}_{l,k}(\mathcal{S}_l(\mathcal{T}_l(t))) \\ &= \Bigg[Q_k(0) - \big(\mu_k - \sum_{l=1}^{K} p_{l,k}\mu_l\big)t - \big(\mathcal{S}_k(\mathcal{T}_k(t)) - \mu_k\mathcal{T}_k(t)\big) \\ &\quad - \sum_{l=1}^{K}\big(\mathcal{R}_{l,k}(\mathcal{S}_l(\mathcal{T}_l(t))) - p_{l,k}\mathcal{S}_l(\mathcal{T}_l(t))\big) + \sum_{l=1}^{K}\big(p_{l,k}\mathcal{S}_l(\mathcal{T}_l(t)) - p_{l,k}\mu_l\mathcal{T}_l(t)\big)\Bigg] \\ &\quad + \left[\mu_k\big(\nu_k t - \mathcal{T}_k(t)\big) - \sum_{l=1}^{K} p_{l,k}\mu_l\big(\nu_k t - \mathcal{T}_l(t)\big)\right] \\ &= \mathcal{X}(t) + R\mathcal{J}(t) \end{aligned}$$

The drift vector for $\mathcal{X}(t)$ is $\mathcal{N}R\nu$ and the variance covariance matrix is given by:

$$\Sigma_{j,l} = \sum_{k=1}^{K}[\nu_k\mu_k p_{k,j}(\delta_{j,l} - p_{k,l}) + \nu_k\mu_k\sigma_k^2 R_{j,k}R_{l,k}]$$

16.15 Prove that if the stochastic transition matrix P is irreducible, then the matrix $P_{\setminus K}$ has spectral radius < 1 so that $I - P_{\setminus K}^{\intercal}$ is invertible.

Solution

Consider P a $K \times K$ irreducible stochastic matrix, $P_{\backslash K}$ the matrix with last row replaced by 0's, $p_{\cdot,K} = (p_{K,1}, \ldots, p_{K,K})^{\mathsf{T}}$ the last row of P, transposed. Let π be the stationary distribution of the transition probability matrix P, so that $\pi P = \pi$. Then:

$$\pi P_{\backslash K} = (\pi_1 - \pi_K p_{1.K}, \ldots, \pi_K - \pi_K p_{K,K},$$

or:

$$(I - P_{\backslash K}^{\mathsf{T}})\pi^{\mathsf{T}} = \pi_K p_{\cdot,K}.$$

In other words, π^{T} solve the traffic equation for the network without node K with input rates proportional to $p_{\cdot,K}$. This shows that $I - P_{\backslash K}^{\mathsf{T}}$ is invertible, and hence $P_{\backslash K}$ has spectral radius < 1.

16.16 Show that (Υ_1, Υ_2) is proportional to (ρ_1, ρ_2).

Solution

Since Υ_i is the amount of work at machine i for a complete circulation of a customer, which we calculated starting at the exit from buffer K, it obviously should be proportional to the traffic intensity. We now go through a calculation to show that this is indeed the case.

Recall that π is the probability vector solving $P^{\mathsf{T}}\pi = \pi$, and β solves $R\beta = 0$, so $(I - P^{\mathsf{T}})\mathrm{diag}(\mu)\beta = 0$ so we can take $\mathrm{diag}(\mu)\beta = \pi$ or $\beta_k = \pi_k m_k$, and we then define $\rho = C\beta$ where we adjust β so that $\max(\rho_1, \rho_2) = 1$. In summary:

$$\rho = C\beta \propto C\mathrm{diag}(m)\pi$$

Recall the definition of $P_{\backslash K}$ as the matrix P with the last row replaced by zeroes, and denote $p_K = \begin{bmatrix} p_{K,1} \\ \vdots \\ p_{K,K} \end{bmatrix}$

$$\Upsilon = Mp_K = CR^{-}p_K = C\mathrm{diag}(m)(I - P_{\backslash K})^{-1}p_K.$$

so we need to show that $(I - P_{\backslash K}^{\mathsf{T}})^{-1}p_K$ is proportional to π

This will follow from showing that p_K is proportional to $(I - P_{\backslash K}^{\mathsf{T}})\pi$.

We note that

$$P_{\backslash K}^{\mathsf{T}}\pi = \begin{bmatrix} \pi_1 - \pi_K\ p_{K,1} \\ \pi_2 - \pi_K\ p_{K,2} \\ \vdots \\ \pi_K - \pi_K p_{K,K} \end{bmatrix} = \pi - \pi_K p_K,$$

and so $(I - P_{\backslash K}^{\mathsf{T}})\pi = \pi_K p_K$ as required. This completes the proof.

16.17 For the closed two station multi-class network, use the results of Section 7.4 to obtain the stationary distribution of the Brownian workload imbalance $\hat{\mathcal{W}}_\Delta(t)$, when using the optimal policy, and find the stationary rate of idling at the two stations.

Solution

We need first to find the drift and variance of the Brownian motion that drives the work imbalance, and we then use (7.19) to obtain the stationary distribution of the work imbalance and the averaged reflection controls.

To calculate the drift and variance of the Brownian motion that drives the work imbalance we start with

$$\mathcal{X}_k(t) = \mathcal{Q}_k(0) - \mathcal{S}_k(\mathcal{T}_k(t)) + \sum_{l=1}^{K} \mathcal{R}_{l,k}(S_l(\mathcal{T}_l(t))),$$

We scale the queue, as $\hat{\mathcal{Q}}(t) = \mathcal{Q}(N^2 t)/N$, where N is the total number in the closed network, and assume that the scaled $\hat{\mathcal{X}}_k(t)$ is close to a Brownian motion. We treat it as a K-dimensional Brownian motion. Recall the definitions of $\beta,\ \rho,\ \nu$. By the scaling, the drift of $\hat{\mathcal{X}}_k(t)$ is $\theta_k = -N\big(\nu_k\mu_k - \sum_{l=1}^{K} p_{l,k}\nu_l\mu_l\big)$, and using $\nu_k = \beta_k\rho_{s(k)}$ and $R\beta = 0$ we have that the drift of $\hat{\mathcal{X}}$ is:

$$\theta = -NR\nu, \qquad \text{of mderate size.}$$

We now calculate the variance covariance matrix Γ of $\hat{\mathcal{X}}$, using the fact that in balanced heavy traffic, $\bar{\mathcal{T}}_k(t) \approx \nu_k t$. The calculation follows, similar to the proof of Theorem 9.4:

$$\mathbb{E}\Big[\mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} \mathcal{R}_{j,k}(\mathcal{S}_j(\nu_j t))\Big|\mathcal{S}(t)\Big] = \mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} p_{j,k}\mathcal{S}_j(\nu_j t),$$

$$\mathbb{V}\text{ar}\Big[\mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} \mathcal{R}_{j,k}(\mathcal{S}_j(\nu_j t))\Big|\mathcal{S}(t)\Big] = \sum_{j=1}^{K} \mathcal{S}_j(\nu_j t)p_{j,k}(1-p_{j,k}),$$

$$\mathbb{C}\text{ov}\Big[\mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} \mathcal{R}_{j,k}(\mathcal{S}_j(\nu_j t)), \mathcal{S}_l(\nu_l t) - \sum_{j=1}^{K} \mathcal{R}_{j,l}(\mathcal{S}_j(\nu_j t))\Big|\mathcal{S}(t)\Big]$$
$$= -\sum_{j=1}^{K} \mathcal{S}_j(\nu_j t)p_{j,k}p_{j,l},$$

and from this we obtain (using $\mathbb{V}\text{ar}(B) = \mathbb{V}\text{ar}(\mathbb{E}(B|A)) + \mathbb{E}(\mathbb{V}\text{ar}(B|A))$):

$$\Gamma_{k,k} = \mathbb{V}\text{ar}\Big[\mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} \mathcal{R}_{j,k}(\mathcal{S}_j(\nu_j t))\Big]$$
$$= \Big[\mu_k\nu_k c_{s,k}^2(1-p_{k,k})^2 + \sum_{j\neq k} \mu_j\nu_j c_{s,j}^2 p_{j,k}^2 + \sum_{j=1}^{k} \mu_j\nu_j p_{j,k}(1-p_{j,k})\Big]t.$$

To get the covariance, we first calculate

$$\mathbb{C}\mathrm{ov}\Big[\mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} p_{j,k}\mathcal{S}_j(\nu_t), \mathcal{S}_l(\nu_l t) - \sum_{j=1}^{K} p_{j,l}\mathcal{S}_j(\nu_j t)\Big]$$
$$= \Big[-\mu_k \nu_k c_{s,k}^2 p_{k,l} - \mu_l \nu_l c_{s,l}^2 p_{l,k} + \sum_{j=1}^{K} \mu_j \nu_j c_{s,j}^2 p_{j,k} p_{j,l}\Big] t,$$

from which we obtain, for $k \neq l$, (using $\mathbb{C}\mathrm{ov}(B, C) = \mathbb{C}\mathrm{ov}(\mathbb{E}(B|A), \mathbb{E}(C|A)) + \mathbb{E}(\mathbb{C}\mathrm{ov}(B, C|A))$):

$$\Gamma_{k,l} = \mathbb{C}\mathrm{ov}\Big[\mathcal{S}_k(\nu_k t) - \sum_{j=1}^{K} \mathcal{R}_{j,k}(\mathcal{S}_j(\nu_j t)), \mathcal{S}_l(\nu_l t) - \sum_{j=1}^{K} \mathcal{R}_{j,l}(\mathcal{S}_j(\nu_j t))\Big]$$
$$= -\Big[\mu_k \nu_k c_{s,k}^2 p_{k,l} + \mu_l \nu_l c_{s,l}^2 p_{l,k} + \sum_{j=1}^{K} \mu_j \nu_j (1 - c_{s,k}^2) p_{j,k} p_{j,l}\Big] t.$$

The workload for the two machines, for eace circulation that starts with exit from buffer K, is then

$$\hat{\mathcal{W}}(t) = M\hat{Q}(t) = CR^{-}\hat{Q}(t), \text{ and } \hat{\mathcal{Z}}(t) = M\hat{\mathcal{X}}(t) \text{ is the workload netput}$$

We now have the Workload process,

$$\begin{aligned}
&\hat{\mathcal{W}}(t) = \hat{\mathcal{Z}}(t) + MR\mathcal{J}(t) \\
&= \hat{\mathcal{Z}}(t) + C\mathrm{diag}(m)(I - P_{\setminus K}^{\intercal})^{-1}(I - P_{\setminus K}^{\intercal}))\mathrm{diag}(\mu)\,\mathcal{J}(t) \\
&- M \begin{bmatrix} p_{K,1} \\ \vdots \\ p_{K,K} \end{bmatrix} \mu_K \mathcal{J}_K(t) \\
&= \hat{\mathcal{Z}}(t) + C\mathcal{J}(t) - \Upsilon\eta(t) \\
&= \hat{\mathcal{Z}}(t) + \hat{\mathcal{I}}(t) - \Upsilon\eta(t),
\end{aligned}$$

were $\hat{\mathcal{I}}(t)$ is the two dimensional idling process, Υ the workload per circulation at the two machines, and η the balancing control.
The 2-dimensional Brownian motion $\hat{\mathcal{Z}}(t)$ has drift $NMR\nu$, and variance matrix $M\Gamma M^{\intercal}$.
We then define a scalar workload imbalance process $\hat{\mathcal{W}}_{\Delta}(t) = \rho_2\hat{\mathcal{W}}_1(t) - \rho_1\hat{\mathcal{W}}_2(t)$, with Brownian netput to the work imbalance, $\hat{\mathcal{Z}}_{\Delta}(t) = \rho_2\hat{\mathcal{Z}}_1(t) - \rho_1\hat{\mathcal{Z}}_2(t)$. The one-dimensional Brownian motion $\hat{\mathcal{Z}}_{\Delta}(t)$ has drift and variance:

$$\begin{aligned}
m = \quad \hat{\mathcal{Z}}_{\Delta}(t) \text{ drift} \quad &= -\big[\rho_2\,; -\rho_1\big] MR\nu N = N(\rho_1 - \rho 2), \\
\sigma^2 = \hat{\mathcal{Z}}_{\Delta}(t) \text{ variance} \; &= \big[\rho_2\,; -\rho_1\big] M\Gamma M^{\intercal} \begin{bmatrix} \rho_2 \\ -\rho_1 \end{bmatrix}.
\end{aligned}$$

To show the final expression for the drift the calculation is:

$$\begin{aligned} MR\nu N &= NCR^{-}\left[\left(I - P^{\intercal}_{\setminus K}\right)\mathrm{diag}(\mu)\nu - p_K\mu_K\nu_K\right] \\ &= NC\mathrm{diag}(m)\left(I - P^{\intercal}_{\setminus K}\right)^{-1}\left(I - P^{\intercal}_{\setminus K}\right)\mathrm{diag}(\mu)\nu - NMp_K\mu_K\nu_K \\ &= NC\nu - N\Upsilon\mu_K\nu_K \end{aligned}$$

but, $C\nu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and $\Upsilon \propto \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$, and the expression for the drift follows.

The process $\hat{\mathcal{W}}(t)$ is controlled between two barriers: $b = \delta_1 > \cdots > \delta_K = a$ where $\delta_k = \rho_2 M_{1,k} - \rho_1 M_{2,k}$. We write:

$$\begin{aligned} \hat{\mathcal{W}}_\Delta(t) &= \hat{\mathcal{Z}}_\Delta(t) + \hat{\mathcal{R}}(t) - \hat{\gamma}(t), \\ a \le \hat{\mathcal{W}}_\Delta(t) &\le b, \quad 0 < t < T, \end{aligned}$$

and we can now obtain the "rate" of control which is the long term average of the cumulative control, and the distribution of $\hat{\mathcal{Z}}_\Delta(t)$ in the range $a < y < b$, from (7.19):

$$\lim_{t\to\infty} \hat{\mathcal{R}}/t = \begin{cases} \sigma^2/2(b-a) & \\ \dfrac{m}{e^{2m(b-a)/\sigma^2} - 1} & \end{cases} \qquad \lim_{t\to\infty} \hat{\gamma}/t = \begin{cases} \sigma^2/2(b-a) & \\ \dfrac{m}{1 - e^{-2m(b-a)/\sigma^2}} & \end{cases}$$

$$\lim_{t\to\infty} \mathbb{P}(\hat{\mathcal{Z}}(t) \le y) = \begin{cases} (y-a)/(b-a) & m = 0 \\ \dfrac{e^{2m(y-a)/\sigma^2} - 1}{e^{2m(b-a)/\sigma^2} - 1} & m \neq 0 \end{cases}.$$

We can now retrieve the idling rates

$$\hat{\mathcal{I}}_1(t) = \hat{\mathcal{R}}(t)/\rho_2, \qquad \hat{\mathcal{I}}_2(t) = \hat{\gamma}(t)/\rho_1.$$

16.18 Analyze the closed two station queueing network as a Kelly network under a symmetric policy, and find the distribution of the stationary queue lengths and the expected circulation time.

Solution

I do not see an easy way to calculate the circulation time. What I can say is: In Exercise 16.23, we obtained the stationary distribution for such a network with Poisson arrivals. Since a Kelly-type network satisfies partial balance, the closed network has the same stationary probabilities as the open one, with a different normalizing constant, and a truncated set of states. So the stationary distribution is:

$$\mathbb{P}(\text{node } i \text{ has } n_i \text{ customers}, i = 1, 2) = B\rho_1^{n_1}\rho_2^{N-n_1}$$

To see n_k custmers of type $k = 1, \ldots, K$ we have, subject to total N:

$$\mathbb{P}(n_1, \ldots, n_K) = B\prod_{k=1}^{K} \beta_k^{n_k}.$$

16.19 For the closed Kumar-Seidman Rybko-Stolyar two station network, described in Figure 16.1, with N customers and feed back to the top or bottom route with probabilities $\alpha, 1-\alpha$, perform all the steps of the analysis, to reach

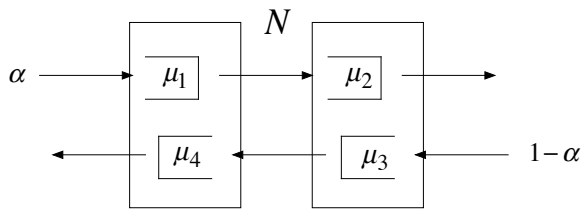


Figure 16.1 A closed KSRS network, with N customers

the BCP and solve it.
Solution

We take $K = 4$ as the exit buffer, so:

$$\pi = \tfrac{1}{2}\big[\alpha;\alpha;1-\alpha;1-\alpha\big], \quad \beta = \big[\alpha m_1;\alpha m_2;(1-\alpha)m_3;(1-\alpha)m_4\big]$$

and the system is in balanced heavy traffic if:

$$\frac{\alpha}{1-\alpha} \approx \frac{m_4 - m_3}{m_2 - m_1}, \qquad \rho_1 \approx \rho_2 \approx 1.$$

If we take $\alpha = \frac{m_4-m_3}{m_2-m_1+m_4-m_3}$ and $\rho_1 = \rho_2 = 1$, there is no drift in the netput of the workload imbalance process.

$$R = (I - P^{\intercal})\mathrm{diag}(\mu) = \begin{bmatrix} \mu_1 & -\alpha\mu_2 & 0 & -\alpha\mu_4 \\ -\mu_1 & \mu_2 & 0 & 0 \\ 0 & -(1-\alpha)\mu_2 & \mu_3 & -(1-\alpha)\mu_4 \\ 0 & 0 & -\mu_3 & \mu_4 \end{bmatrix},$$

$$P_{\setminus K} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \alpha & 0 & 1-\alpha & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad p_4 = \begin{bmatrix} \alpha \\ 0 \\ 1-\alpha \\ 0 \end{bmatrix},$$

$$R^{-} = \mathrm{diag}(m)(I - P_{\setminus K})^{-1} = \begin{bmatrix} \frac{m_1}{1-\alpha} & \frac{m_1\alpha}{1-\alpha} & 0 & 0 \\ \frac{m_2}{1-\alpha} & \frac{m_2}{1-\alpha} & 0 & 0 \\ m_3 & m_3 & m_3 & 0 \\ m_4 & m_4 & m_4 & m_4 \end{bmatrix},$$

$$M = CR^{-} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} R^{-} = \begin{bmatrix} \frac{m_1+m_4(1-\alpha)}{1-\alpha} & \frac{m_1\alpha+m_4(1-\alpha)}{1-\alpha} & m_4 & m_4 \\ \frac{m_2+m_3(1-\alpha)}{1-\alpha} & \frac{m_2+m_3(1-\alpha)}{1-\alpha} & m_3 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} m_1 + \frac{m_2m_4-m_1m_3}{m_2-m_1} & \frac{m_2m_4-m_1m_3}{m_2-m_1} & m_4 & m_4 \\ m_2 + \frac{m_2m_4-m_1m_3}{m_2-m_1} & m_2 + \frac{m_2m_4-m_1m_3}{m_2-m_1} & m_3 & 0 \end{bmatrix}.$$

We can now obtain the optimal strategy for the BCP: Since $\rho_1 = \rho_2 = 1$, the workload imbalance is $\hat{\mathcal{W}}_\Delta = \hat{\mathcal{W}}_1 - \hat{\mathcal{W}}_2$, and for the 4 buffers we have $M_{1,k} - M_{2,k}$:

$$(\delta_k)_{k=1}^4 = [m_1 - m_2 \; ; \; -m_2 \; ; \; m_4 - m_3 \; ; \; m_4].$$

Clearly,

$$b = \max(\delta_k) = \delta_4 = m_4, \qquad a = \min(\delta_k) = \delta_2 = -m_2.$$

The optimal policy is to give lowest priority to buffer 4 at machine 1, and buffer 2 at machine 2 and idle only when a machine has 0 customers..

It is quite clear that by doing so, we make sure that machines will not be starved: we give priority to buffer 1 that feeds machine 2 and to buffer 3 that feeds machine 1.

Note that this system is similar to the push-pull system of Chapter 13, but the optimal policy here is the opposite of that for the push pull, where we give priority to buffers 2 and 4, unless starvation is threatened.

16.20 For the open two station network with admission controls, follow the steps necessary to derive the formulation of the work imbalance BCP (16.29).

Solution

In this problem we wish to schedule service at the two stations, and control the input to have rate of at least $\bar{\alpha}$. We then have flows $\lambda = (I - P^{\intercal})^{-1} p_0 \bar{\alpha}$,

with individual per buffer offered load of $\beta_k = \lambda_k/\mu_k$, and $\rho_i = \sum_{k\in C_i}\beta_k$ (i.e. $\rho = C\beta$). We also have nominal flows $\nu_k = \beta_k/\rho_{s(k)}$.
We have found from (16.25)–(16.27) that

$$Q(t) = \mathcal{X}(t) + R\mathcal{J}(t) - p_0\eta(t),$$

where $\mathcal{J}(t)$ are the free time controls $\nu t - \mathcal{T}(t)$, $\eta(t)$ is the admission control $\bar{\alpha}t - \mathcal{A}(t)$, and the netput $\mathcal{X}(t)$ has drift and variance given by θ, Γ. In order to achive input at rate $\geq \bar{\alpha}$ we need to idle each of the two stations for no more than a fraction $1-\rho_i$ of the time, i.e.

$$\lim_{T\to\infty} \mathcal{A}(T)/T \geq \bar{\alpha} \Longleftrightarrow \lim_{T\to\infty} \mathcal{I}_i(T)/T \leq 1-\rho_i.$$

We now approximate the scaled $\mathcal{X}(n^2t)/N$ by a Brownian motion $\hat{\mathcal{X}}(t)$, with drift $N\theta$ and variance matrix Γ, and our balanced heavy traffic assumption is that $N(1-\rho)$ and hence also $N\theta$ is of moderate size. With the same scaling we now have the approximation

$$\hat{Q}(t) = \hat{\mathcal{X}}(t) + R\hat{\mathcal{J}}(t) - p_0\hat{\eta}(t),$$

and our Brownian control problem is: For Brownian motion $\hat{\mathcal{X}}(t)$ find $\hat{Q}(t), \hat{\mathcal{J}}(t), \hat{\eta}(t)$ such that:

$$\begin{aligned}
\min\ & \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\int_0^T \sum_{k=1}^K h_k\hat{Q}_k(t)dt\Big] \\
\text{s.t. } & \hat{Q}(t) = \hat{\mathcal{X}}(t) + R\hat{\mathcal{J}}(t) - p_0\hat{\eta}(t) \\
& \hat{\mathcal{I}}_i(t) = \sum_{k\in C_i} \hat{\mathcal{J}}_k(t), \\
& \lim_{T\to\infty} \frac{1}{T}\mathbb{E}(\hat{\mathcal{I}}_i(T)) \leq N(1-\rho_i), \\
& \hat{Q}_k(t) \geq 0, \quad \hat{\mathcal{I}}(0), \hat{\mathcal{J}}(0), \hat{\eta}(0) = 0,\ \hat{\mathcal{I}} \text{ non-decreasing}, \\
& \hat{Q}, \hat{\mathcal{I}}, \hat{\mathcal{J}}, \hat{\eta} \text{ are non-anticipating w.r.t. } \hat{\mathcal{X}}(t).
\end{aligned}$$

We now derive the workload formulation:

$$\hat{\mathcal{W}}(t) = M\hat{Q}(t) = CR^{-1}\hat{Q}(t) = C\mathrm{diag}(m)(I - P^T t)^{-1}\hat{Q}(t)$$

which is the expected amount of work at the two stations, for the current queues. We let $\hat{\mathcal{Z}}(t) = M\hat{\mathcal{X}}(t)$, which is then a Brownian motion with drift $N(\rho - 1)$ and variance $M^{\intercal}\Gamma M$.
We then have the workload formulations: For a given Brownian motion $\hat{\mathcal{Z}}(t)$

find $\hat{Q}(t), \hat{I}(t), \hat{\eta}(t)$ such that:

$$\begin{aligned}
\min \ & \lim_{T\to\infty} \sup \frac{1}{T}\mathbb{E}\Big[\int_0^T \sum_{k=1}^K h_k \hat{Q}_k(t)dt\Big] \\
\text{s.t.} \ & \hat{\mathcal{W}}(t) = M\hat{Q}(t), \\
& \hat{\mathcal{W}}(t) = \hat{\mathcal{Z}}(t) + \hat{I}(t) - \Upsilon\hat{\eta}(t) \\
& \lim_{T\to\infty} \frac{1}{T}\mathbb{E}(\hat{I}_i(T)) \le N(1-\rho_i), \\
& \hat{Q}_k(t) \ge 0, \quad \hat{I}(0), \hat{\eta}(0) = 0, \ \hat{I} \text{ non-decreasing}, \\
& \hat{Q}, \hat{I}, \hat{\eta} \text{ are non-anticipating w.r.t. } \hat{X}(t).
\end{aligned}$$

We claim that any policy that there is a one-to-one correspondence between policies feasible for the $\hat{X}(t)$ problem and for the $\hat{\mathcal{Z}}(t) = M\hat{X}(t)$ problem. Clearly given $\hat{Q}, \mathcal{J}, \hat{\eta}$ feasible for the $\hat{X}(t)$ problem, $\hat{Q}, \hat{I}_i = \sum_{k\in C_i} \hat{\mathcal{J}}_k, \hat{\eta}$ is feasible for the $\hat{\mathcal{Z}}(t)$ problem, and vice versa, if $\hat{Q}, \hat{I}, \hat{\eta}$ are feasible for the $\hat{\mathcal{Z}}(t)$ problem we retrieve $\hat{\mathcal{J}} = R^{-1}(\hat{Q}(t) - \hat{X}(t)$. So all we need is to solve the workload formulation.

We now define workload imbalance: $\hat{\mathcal{W}}_\Delta = \rho_2\hat{\mathcal{W}}_1 - \rho_1\hat{\mathcal{W}}_2$, and similarly: $\hat{\mathcal{Z}}_\Delta = \rho_2\hat{\mathcal{Z}}_1 - \rho_1\hat{\mathcal{Z}}_2$. $\hat{\mathcal{Z}}_\Delta$ is a scalar Brownian motion with drift $N(\rho_2 - \rho_1$, and variance $\rho^\intercal M^\intercal \Gamma M\rho$. With this the workload imbalance formulation is (19.29).

The workload and the workload imbalance problems are equivalent, since we can retrieve $\hat{\eta}$ from $\hat{\mathcal{W}} = M\hat{Q}$ and $\hat{\eta} = \frac{1}{\Upsilon_i}(\hat{\mathcal{Z}}_i + \hat{I}_i - \hat{\mathcal{W}}_i)$.

16.21 For the open two station network with admission controls, show that $\delta_1 \ge 0 \ge \delta_K$.

Solution

We have:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = C\text{diag}(m)(I - P^\intercal)^{-1}p_0\bar{\alpha} = Mp_0\bar{\alpha},$$

hence,

$$0 = [\rho_2 \ ; \ \rho_1]\begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = [\rho_2 \ ; \ -\rho_1]Mp_0$$

or:

$$\sum_{k=1}^K (\rho_2 M_{1,k} - \rho_1 M_{2,k})p_{0,k} = 0.$$

p_0 is a probability vector, assume first that $p_{0,k} > 0$, $k = 1, \ldots, K$. Then either for all k $\rho_2 M_{1,k} - \rho_1 M_{2,k} = 0$, or else at least for one k', $\rho_2 M_{1,k'} - \rho_1 M_{2,k'} > 0$, and for another k'', $\rho_2 M_{1,k''} - \rho_1 M_{2,k''} < 0$.

Note that $M_{i,k}$ do not depend on the values of the $p_{0,k}$. Sonce for the given $\bar{\alpha}$, $\rho_1 < 1$ and $\rho_2 < 1$ we can define a new probability vector. $\tilde{p}_0$

with $\tilde{p}_{0,k} = p_{0,k} + \epsilon_k$, such that $\tilde{p}_{0,k} > 0$, $k = 1, \ldots, K$ and still the new system has $\tilde{\rho}_1 < 1$, $\tilde{\rho}_2 < 1$, and for this system we can deduce that either $\rho_2 M_{1,k} - \rho_1 M_{2,k} = 0$ for all k, or else we must have $\max_k \delta_k > 0$, $\min_k \delta_k < 0$.

16.22 For the open two station network with admission controls, analyze the cases when $\delta_1 = 0 > \delta_K$, $\delta_1 > 0 = \delta_K$, and $\delta_1 = \delta_K = 0$. When can that happen?

Solution

We have answered this question in the solution of the previous exercise 16.21.

16.23 Analyze the two station queueing network with admission control as a Kelly network. Assume Poisson arrivals with rate $\bar{\alpha}$, and find the stationary distribution of the queue lengths and the expected waiting times.

Solution

We now assume independent Poisson arrivals of rates $\alpha_k = \bar{\alpha} p_{0,k}$, so arrival rates of customers of class k are $\lambda = (I - P^{\intercal})^{-1}\alpha$, and the offered load of class k is $\beta = \text{diag}(m)\lambda = R^{-1}\alpha$, with $\rho_i = \sum_{k \in C_i} \beta_k$.

If we treat all customers in each station without paying attention to their class, and use a symmetric policy, e.g. PS of LCFS-premeptive, then this is a Kelly-type network, and the stationary distribution is given by (see Section (8.7):

$$\mathbb{P}(\text{node } i \text{ has } n_i \text{ customers}, i = 1, 2) = \prod_{i=1,2} (1 - \rho_i)\rho_i^{n_i}$$

and each customer at a node is type k with probability $\beta_k / \rho_{s(k)}$.

We also have by the arrival theorem (see Section 8.4) that customers that arrive see time average, so they will see the stationary M/M/1 number of customers at the node.

Furthermore, a customer at a node i that will have expected sojourn $x/(1-\rho_i)$ since the policy is a symmetric policy (see Section 3.7). So average sojourn of a class k customer is $\theta_k = m_k/(1 - \rho_{s(k)})$ per visit. Since route and processing are independent, we can add up sojourn expectations over the random stages of the route. We obtain the vector of expected sojourn times for customers of the various types (see Section 8.5) as: $\overline{W} = (I - P)^{-1}\theta$.

The overall average is $\sum_{k=1}^{K} (\alpha_k/\bar{\alpha})\overline{W}_k$

16.24 For the open two station network with admission controls, prove equation (16.33) in Proposition 16.2 [Wein (1990)].

Solution

The details of the proof appear in Wein (1990).

16.25 For the open Kumar-Seidman Rybko-Stolyar two station network, described in Section 10.2.3, with arrival rates α_1, α_2 to the top and bottom routes respectively, perform all the steps of the analysis, to reach the BCP and solve it.

Solution

We start with the queue lengths process $Q(t)$ and its netput $\mathcal{X}(t)$. Parameters

are $\bar{\alpha} = \alpha_1 + \alpha_2$, $p_{0,1} = \alpha_1/\bar{\alpha}$, $p_{0,2} = \alpha_2//\bar{\alpha}$, $m_i = 1/\mu_i$, $c_{s,i}^2 = \sigma_i^2\mu_i^2$, $i = 1, 2, 3, 4$.

$$\beta = \begin{pmatrix} \alpha_1 m_1 \\ \alpha_1 m_2 \\ \alpha_2 m_3 \\ \alpha_3 m_4 \end{pmatrix} \qquad \rho = \begin{pmatrix} \alpha_1 m_1 + \alpha_3 m_4 \\ \alpha_1 m_2 + \alpha_2 m_3 \end{pmatrix}, \qquad \nu = \begin{pmatrix} \frac{\alpha_1 m_1}{\alpha_1 m_1 + \alpha_3 m_4} \\ \frac{\alpha_1 m_2}{\alpha_1 m_2 + \alpha_2 m_3} \\ \frac{\alpha_2 m_3}{\alpha_1 m_2 + \alpha_2 m_3} \\ \frac{\alpha_3 m_4}{\alpha_1 m_1 + \alpha_3 m_4}, \end{pmatrix}$$

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad R^{-1} = \begin{pmatrix} m_1 & 0 & 0 & 0 \\ m_2 & m_2 & 0 & 0 \\ 0 & 0 & m_3 & 0 \\ 0 & 0 & m_4 & m_4 \end{pmatrix}.$$

Then the netput $\mathcal{X}(t)$ has drift:

$$\theta = \begin{pmatrix} \alpha_1 - \frac{\alpha_1}{\alpha_1 m_1 + \alpha_3 m_4} \\ \frac{\alpha_1}{\alpha_1 m_1 + \alpha_3 m_4} - \frac{\alpha_1}{\alpha_1 m_2 + \alpha_2 m_3} \\ \alpha_2 - \frac{\alpha_2}{\alpha_1 m_2 + \alpha_2 m_3} \\ \frac{\alpha_2}{\alpha_1 m_2 + \alpha_2 m_3} - \frac{\alpha_3}{\alpha_1 m_1 + \alpha_3 m_4} \end{pmatrix},$$

and variance $\Gamma_0 + \sum_{j=1}^4 \Gamma_j$ with

$$\Gamma_0 = \begin{pmatrix} \bar{\alpha} p_{0,1}(1 - p_{0,1}) & 0 & -\bar{\alpha} p_{0,1} p_{0,2} & 0 \\ 0 & 0 & 0 & 0 \\ -\bar{\alpha} p_{0,1} p_{0,2} & 0 & \bar{\alpha} p_{0,2}(1 - p_{0,2}) & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \Gamma_j \text{ see (16.27).}$$

The scaled process is $\hat{\mathcal{X}}(t) = \mathcal{X}(N^2 t)/N$ which in balanced heavy traffic is approximately a Brownian process, with drift $N\theta$ and variance Γ. We treat is as a Brownian process and have the Brownian control problem.
The workload formulation has $\hat{\mathcal{W}}(t) = M\hat{\mathcal{Q}}(t)$, $\hat{\mathcal{Z}}(t) = M\hat{\mathcal{X}}(t)$ where

$$M = CR^{-1} = \begin{pmatrix} m_1 & 0 & m_4 & m_4 \\ m_2 & m_2 & m_3 & 0 \end{pmatrix}.$$

and the work imbalance is then $\hat{\mathcal{W}}_\Delta = \rho_2\hat{\mathcal{W}}_1 - \rho_1\hat{\mathcal{W}}_2$ driven by the Brownian motion $\hat{\mathcal{Z}}_\Delta = \rho_2\hat{\mathcal{Z}}_1 - \rho_1\hat{\mathcal{Z}}_2$.
The drift of $\hat{\mathcal{Z}}_\Delta$ is $m = N(\rho_2 - \rho_1$. The variance $sigma^2$ is calculated from ρ, M, Γ.
From the values of these expressions it is now straightforward to calculate the optimal range of imbalance, and the boundaries of the cone of $(\hat{\mathcal{W}}_1, \hat{\mathcal{W}}_2)$ outside of which customers are admitted.

16.26 Provide an argument to show that if $\hat{\mathcal{I}}(t)$ is given, then the LP (10.37) and its dual (10.39) are bounded and feasible.
Solution

Clearly the dual is feasible because h_k are positive, and so $y_k(t) = 0$ is a feasible solution.

To see that the primal is feasible consider the constraints of the workload formulation, rather than the workload imbalances formulation:

$$\sum_{k=1}^{K} M_{i,k}\hat{Q}_k(t) = \hat{\mathcal{W}}_i(t),$$
$$\hat{Q}(t) \geq 0.$$

Here the r.h.s. is positive as are the coefficients $M_{I,k}$. Choose one buffer $k_i \in C_i$ for each machine, and take the solution $\hat{Q}_{k_i}(t) = \hat{\mathcal{W}}_i(t)/M_{i,k_i}$, $i = 1, \ldots, I$, and all other $\hat{Q}_k(t) = 0$. This is a feasible solution, and clearly, these values of $\hat{Q}(t)$ are feasible for the workload imbalance formulation, (19.37).

16.27 For the multistation network with admission control, show that the objective function $H(\hat{\mathcal{W}}(t))$ is convex [Wein (1992)].

Solution

Consider the dual problem (19.39), and its solution as a function of the objective coefficients $\hat{\mathcal{W}}_{\Delta,i}$.

For each vector of $\hat{\mathcal{W}}_{\Delta,i}$ there is an optimal basic solution that maximizes the objective. Consider then the finite collection of all the feasible basic solutions. For each feasible basic solution, the value of the objective is a linear function of $\hat{\mathcal{W}}_\Delta$. The optimal objective is the maximum of these linear functions of $\hat{\mathcal{W}}_\Delta$, but a maximum of linear functions is a convex continuous piecewise linear function.

16.28 For the multistation network with admission control, show that the boundary forms a prism parallel to $(\rho_1, \ldots, \rho_I)$ [Wein (1990)].

Solution

The solution of the singular control problem (16.42) defines a bounded region in $\mathbb{R}^{I-1}$ that includes the origin, in which the $I-1$ dimensional transformed workload imbalance $\hat{\mathcal{W}}^\circ_\Delta$ can move, and the singular controls $\hat{\mathcal{I}}^\circ(t)$ keep the workload imbalance in this region. This region corresponds to a region in which the original (untransformed) workload imbalance $\hat{\mathcal{W}}_\Delta$ moves, and it is kept this region by the singular controls $\hat{\mathcal{I}}(t)$, denote this region in $\mathbb{R}^{I-1}$ by G. G is quite close to the transformed region, because all ρ_i are close to 1, it is bounded, and it contains the origin. Consider now the I dimensional process, $\hat{\mathcal{W}}(t)$. If $\hat{\mathcal{W}}_I(t) = y$, then $\hat{\mathcal{W}}_{\Delta,i}(t) = \rho_I\hat{\mathcal{W}}_i(t) - \rho_i y$, $i \neq I$, so that

$$\hat{\mathcal{W}}_i(t) = \frac{1}{\rho_I}\hat{\mathcal{W}}_{\Delta,i}(t) + \frac{\rho_i}{\rho_I}y$$

So the I dimensional region in $\mathbb{R}^I$ in which $\hat{\mathcal{W}}(t)$ moves is a prism lifting G/ρ_I in the direction parallel to ρ.

16.29 (∗) Analyze the control of a closed MCQN with I workstations, in analogy with Sections 16.4 and 16.6.

Solution

This is as far as I know an open problem. Here are some thoughts on it.
The workload imbalance formulation of the problem is: For an $I-1$ dimensional Brownian motion $\hat{\mathcal{Z}}_\Delta(t)$ with known drift and variance, find $\hat{Q}(t)$ and $\hat{I}_i(t)$ to solve

$$\begin{aligned}
\min \ & \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{I-1} \hat{I}_i(T) \\
s.t. \ & \hat{\mathcal{W}}_\Delta = \sum_{k=1}^{K} (\rho_I M_{i,k} - \rho_i M_{I,k}) \hat{Q}_k(t), \\
& \hat{\mathcal{W}}_{i,\Delta}(t) = \hat{\mathcal{Z}}_{i,\Delta}(t) + \rho_I \hat{I}_i(t) - \rho_i \hat{I}_I(t), \\
& \mathbf{1}^{\mathrm{T}} \hat{Q}(t) = 1, \quad \hat{Q}(t) \geq 0, \quad \hat{I}_i(0) = 0, \ \hat{I}_i \text{ non-decreasing}, \\
& \hat{Q}(t), \hat{I}(t) \text{ non-anticipating w.r.t. } \hat{\mathcal{Z}}_\Delta.
\end{aligned}$$

Clearly, since each control $\hat{I}_i$ affects only one the machines, we should idle only when the workload at a machine is 0.
Next we want to minimize the imbalance between the machines. A heuristic for this is to solve at each moment t the problem:

$$\begin{aligned}
\min \ & \sum_{i=1}^{I-1} \hat{\mathcal{W}}_i(t)^2 \\
s.t. \ & \hat{\mathcal{W}}_i = \sum_{k=1}^{K} (\rho_I M_{i,k} - \rho_i M_{I,k}) \hat{Q}_k(t), \\
& \hat{\mathcal{W}}_{i,\Delta}(t) = \sum_{k=1}^{K} M_{i,k} \hat{Q}_k(t), \\
& \sum_{k=1}^{K} \hat{Q}_k(t) = 1, \qquad \hat{Q}(t) > 0.
\end{aligned}$$

17

MCQN with Discretionary Routing

Exercises

17.1 Derive the factor three saving for the two station model with criss cross customers and routing, over a random routing FCFS scheduling. Assume Poisson arrivals and exponential service.

Solution

For random allocation of type A to the two stations and FCFS at each station, station 1 will experience Poisson arrivals at rate 2λ, station 2 will have arrivals of type A as Poisson of rate λ, and an independent Poisson stream of customers of rate λ that exit from station 1. So each station acts as an M/M/1 queue with arrival rate $(3/2)\lambda$ and service rate μ. Sojourn time at each queue is then $\frac{1}{\mu-(3/2)\lambda}$, and half the customers wait only once, the other half waits at both queues, so in total, average sojourn is:

$$\tfrac{1}{2}\frac{1}{\mu-(3/2)\lambda}+\tfrac{1}{2}\frac{2}{\mu-(3/2)\lambda}=\frac{3}{2\mu-3\lambda}.$$

For the optimized system in heavy traffic, both servers will work all the time, and there will be no queue of type B customers between stations 1 and 2. So it will behave like a single system (state space collapse) with Poisson arrivals of rate 3λ and exponential service of rate 2μ with expected sojourn time $\frac{1}{2\mu-3\lambda}$. To be exact, we control the queue between the stations to be short, and service of 2 exponential servers in heavy traffic is the same as single server with sum of the rates. So sojourn time is

$$\approx \frac{1}{2\mu-3\lambda},$$

for a saving factor of 3.

17.2 For the two station model with criss cross customers and routing, repeat the Brownian problem formulation and solution for general parameters, $\lambda_A, \lambda_B, \mu_1, \mu_2$.

Solution

For stability we need to have: $\rho = \frac{\lambda_A+2\lambda_B}{\mu_1+\mu_2} < 1$, and also $\lambda_B < \mu_1$ and $\lambda_B < \mu_2$. in heavy traffic we assume $N(1-\rho)$ is of mederate size for large

N, so $\rho \approx 1$. We equate the load on the two machines by directing in the long run a fraction θ of type A customers to machine 1, so that $\theta = \frac{\rho\mu_1 - \lambda_B}{\lambda_A}$. We could use any $\rho_1, \rho_2 < 1$ with $N(\rho_i - \rho)$ of moderate size, but the choice of $\rho_1 = \rho_2 = \rho$ will result in a drift of zero in the final analysis.
So we have:

$$\beta_1 = \lambda_B/\mu_1, \quad \beta_2 = \lambda_B/\mu_2, \quad \beta_3 = \rho - \lambda_B/\mu_1, \quad \beta_4 = \rho - \lambda_B/\mu_2,$$

and we have nominal allocations $\nu_k = \beta_k/\rho$.
Our controls for this system are the time allocations, $\mathcal{T}_k(t)$ to processing buffer k, and the admissions $\mathcal{A}_k(t)$ for the discretionary buffers, $k = 3, 4$, where $\mathcal{A}_3(t) + \mathcal{A}_4(t) = \mathcal{A}_A(t)$, and the nominal rates of admissions are $\alpha_3 = \theta\lambda, \alpha_4 = (1-\theta)\lambda$. We then have the deviations from nominal admission rates, $\mathcal{V}_k(t) = \alpha_k t - \mathcal{A}_k(t)$, $k = 3, 4$, and free times $\mathcal{J}_k(t) = \nu_k t - \mathcal{T}_k(t)$, which embody our controls.
We scale the system: $\hat{Q}(t) = Q(N^2 t)/N$, and approximate the netput by a Brownian motion, and obtain for the Brownian approximation the Brownian control problem:

$$\begin{aligned}
\min \ & \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^4 \hat{Q}_k(t)dt\,] \\
\text{s.t.} \ & \hat{Q}(t) = \hat{X}(t) + R\,\hat{\mathcal{J}}(t) - G\hat{\mathcal{V}}(t), \quad t \geq 0. \\
& \hat{I}(t) = C\hat{\mathcal{J}}(t), \qquad H\hat{\mathcal{V}}(t) = 0, \quad t \geq 0. \\
& \hat{Q}(t) \geq 0, \qquad \hat{I}(0) = 0, \quad \hat{I} \text{ non-decreasing}, \quad t \geq 0, \\
& \hat{Q}, \hat{\mathcal{J}}, \hat{\mathcal{V}} \text{ are non-anticipating with respect to } \hat{X}(t),
\end{aligned}$$

where the controls $\hat{\mathcal{J}}_k(t)$ are the buffer free times $\hat{\mathcal{J}}_k(t) = -\nu_k\mu_k$ with

$$R = (I - P^{\mathrm{T}})\mathrm{diag}(\mu) = \begin{bmatrix} \mu_1 & 0 & 0 & 0 \\ -\mu_1 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_1 & 0 \\ 0 & 0 & 0 & \mu_2 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \qquad H = [1 \ -1]$$

$$R^{-1} = \begin{bmatrix} m_1 & 0 & 0 & 0 \\ m_2 & m_2 & 0 & 0 \\ 0 & 0 & m_1 & 0 \\ 0 & 0 & 0 & m_2 \end{bmatrix}, \quad M = CR^{-1} = \begin{bmatrix} m_1 & 0 & m_1 & 0 \\ m_2 & m_2 & 0 & m_2 \end{bmatrix}$$

Workloads are given by

$$\hat{\mathcal{W}}(t) = M\hat{Q}(t) = \begin{bmatrix} m_1 & 0 & m_1 & 0 \\ m_2 & m_2 & 0 & m_2 \end{bmatrix} \hat{Q}(t).$$

Similar to Chapter 15, routing pools the resources of the two stations in the system. We therefore look at the sum of the workloads,

$$\hat{\mathcal{W}}_\Sigma = \big[m_1 + m_2 \,;\, m_2 \,;\, m_1 \,;\, m_2\big]\hat{Q}(t)$$

The pooled workload control problem is to find $\hat{Q}, \hat{I}$ such that:

$$\begin{aligned}
\min \ & \lim_{T\to\infty} \sup \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^4 \hat{Q}_k(t)dt\,] \\
\text{s.t.} \ & \big((m_1+m_2)\hat{Q}_1(t) + m_2\hat{Q}_2(t) + m_1\hat{Q}_3(t) + m_2\hat{Q}_4(t)\big) = \hat{\mathcal{W}}_\Sigma(t), \\
& \hat{\mathcal{W}}_\Sigma(t) = \hat{\mathcal{Z}}(t) + \hat{I}_1(t) + \hat{I}_2(t), \\
& \hat{Q}(t) \ge 0, \quad \hat{I}(0) = 0, \quad \hat{I} \text{ non-decreasing}, \quad t \ge 0, \\
& \hat{Q}(t),\, \hat{I}(t) \text{ are non-anticipating with respect to } \hat{\mathcal{Z}}(t).
\end{aligned}$$

So clearly, inventory should be kept in buffer 1, at station 1 we give priority to buffer 1 over 2, and we keep station 2 by means of the control $\mathcal{V}(t)$. We idle when both buffers are empty.

In summary, the optima Brownian control policy is the same as for the case where $\mu_1 = \mu_2$, and $\lambda_A = \lambda_B$.

17.3 For the two station model with criss cross customers and routing, perform the Brownian problem formulation and solution for general parameters, λ_A, λ_B and individual processing rates, $\mu_k,\ k = 1, 2, 3, 4$.

Solution

There is not much different about this system for Exercise 17.2. Once it is in balanced heavy traffic, the calculations of workload at each station, and of $\hat{\mathcal{W}}_\Sigma$ are as before, and the optimal Brownian policy is the same. The only question is when is it in balanced heavy traffic, and to find this we solve the static planning problem. Assume there are a rewards of $w_A,\ w_B$ for unit items of type $A,\ B$. Then:

$$\begin{aligned}
\max_{\nu,\alpha} \ & w_A\alpha_{A,3} + w_A\alpha_{A,4} + w_B\alpha_B \\
s.t. \ & \begin{bmatrix} \mu_1 & 0 & 0 & 0 \\ -\mu_1 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_1 & 0 \\ 0 & 0 & 0 & \mu_2 \end{bmatrix} \nu = \begin{bmatrix} \alpha_B \\ 0 \\ \alpha_{A,3} \\ \alpha_{A,4} \end{bmatrix}, \\
& \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \nu \le \mathbf{1} \\
& \alpha,\ \nu \ge 0
\end{aligned}$$

For properly chosen w_A, w_B this will have a solution in which both servers are fully utilized, and both α_A, α_B are positive, and if we choose $\rho < 1$ such that for large N we have $N(1-\rho)$ of moderate size and take $(\lambda_A, \lambda_B) = \rho(\alpha_A, \alpha_B)$, then the system will be in balanced heavy traffic.

17.4 For the network of Laws and Louth, derive the formulation of the BCP, equation (17.2).

Solution

There is little to say in this exercise. The matrix R is 8×8 and composed of 4 blocks reflecting transitions from entry to exit buffers:

$$R = \begin{bmatrix} I_{4\times4} & 0_{4\times4} \\ -I_{4\times4} & I_{4\times4} \end{bmatrix}$$

The matrix G reflects input to the four entry buffers,

$$G = \begin{bmatrix} I_{4\times4} \\ 0_{4\times4} \end{bmatrix}$$

The matrix C is the resource consumption matrix,

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

and the matrix H embodies the constraint that the sum of deviations from the nominal inputs for two horizontal buffers and for the two vertical buffers is 0,

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

In the pooled workload processes, $\mathcal{W}_{1+4}$ includes all the work for workstation 1, which includes buffers 1,2, and for station 4 which includes what is currently in 7,8, and also what is currently in buffers 3,4, that will feed into 7,8. Similarly, $\mathcal{W}_{2+3}$ includes work for workstation 2, which is what is currently in buffers 4,5 and also what is currently in buffer 1 and will feed into buffer 5, and work for work station 3, which includes what is currently in buffers 3,6, and also what is currently in buffer 2 and will feed into buffer 6. so:

$$\mathcal{W}_p = \begin{bmatrix} \mathcal{W}_{1+4} \\ \mathcal{W}_{2+3} \end{bmatrix} = m \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \hat{\mathcal{X}}(t)$$

17.5 For the network of Laws and Louth, obtain the drift and variance of $\hat{\mathcal{X}}(t)$ and of $\hat{\mathcal{Z}}(t)$.

Solution

For simplicity we assume that both input streams have the same c.o.v. c_a, and the four service processes have the same c.o.v. c_s.

We rewrite

$$Q(t) = \mathcal{A}(t) - \sum_{j=1}^{K} F^j(\mathcal{T}_j(t)),$$

where $\mathcal{A}_k(t)$ is the controlled (routed) input to buffer k, and $F^j(t)$ is the controlled flow due to completions of service at buffer j, so that $F^j_k(\mathcal{T}_j(t)) = \begin{cases} \mathcal{S}_j(\mathcal{T}_j(t)) & k = j \\ \mathcal{R}_{j,k}(\mathcal{S}_j(\mathcal{T}_j(t))) & k \neq j \end{cases}$.

We then have, for the scaled process, $\hat{Q}(t) = Q(N^2t)/N$, that a fraction $1/2$ of the input goes to each route, and that $\nu_k(t) = p_H = \frac{\lambda_H}{\lambda_H+\lambda_V}$, $k = 1, 3, 5, 7$, $\nu_k(t) = p_V = \frac{\lambda_V}{\lambda_H+\lambda_V}$, $k = 2, 4, 6, 8$. For the total system in balanced heavy traffic, we have that $N(2\lambda_H + 2\lambda_V - 4\mu)$ for large N is of moderate size. Accordingly, the drift for $\hat{X}_k(t)$ i.e. for $\hat{Q}_k(t)$ when positive, is:

$$\text{Drift of } \hat{X}_k(t) \qquad \theta_k = \begin{cases} N(\frac{1}{2}\lambda_H - p_H\mu) & k = 1, 3, 5, 7 \\ N(\frac{1}{2}\lambda_V - p_V\mu) & k = 2, 4, 6, 8 \end{cases}.$$

For the variance matrix of $\hat{X}(t)$, we calculate the unscaled covariances, which are the same as the scaled covariances. We note that $\mathcal{A}$, F^j, $1, \ldots, \mathcal{K}$ are independent, to get $\Gamma = \Gamma_A + \sum_{j=1}^{K} \Gamma_{F^j}$.

Assume that arrivals are allocated randomly with probability $1/2$ to each route.

$$\Gamma_A = \begin{bmatrix} \frac{1}{4}\lambda_H(1+c_a^2) & 0 & -\frac{1}{4}\lambda_H(1+c_a^2) & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4}\lambda_V(1+c_a^2) & 0 & -\frac{1}{4}\lambda_V(1+c_a^2)0 & 0 & 0 & 0 & \\ -\frac{1}{4}\lambda_H(1+c_a^2) & 0 & \frac{1}{4}\lambda_H(1+c_a^2) & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4}\lambda_V(1+c_a^2) & 0 & -\frac{1}{4}\lambda_V(1+c_a^2) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The variances for F^j are calculated similarly, we write just the first:

$$\Gamma_{F^1} = \begin{bmatrix} \nu_1\mu c_s^2 & 0 & 0 & 0 & -\nu_1\mu c_s^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\nu_1\mu c_s^2 & 0 & 0 & 0 & +\nu_1\mu c_s^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

or, more sucsicntly, the non-zero elements are:

$$(\Gamma_{F^j})_{j,j} = p_H\mu c_s^2,\ (\Gamma_{F^j})_{j,j+4} = -p_H\mu c_s^2,\ (\Gamma_{F^j})_{j+4,j} = -p_H\mu c_s^2,\ (\Gamma_{F^j})_{j+4,j+4} = p_H\mu c_s^2, \quad j = 1, 3,$$
$$(\Gamma_{F^j})_{j,j} = p_V\mu c_s^2,\ (\Gamma_{F^j})_{j,j+4} = -p_V\mu c_s^2,\ (\Gamma_{F^j})_{j+4,j} = -p_V\mu c_s^2,\ (\Gamma_{F^j})_{j+4,j+4} = p_V\mu c_s^2, \quad j = 2, 4,$$
$$(\Gamma_{F^j})_{j,j} = p_H\mu c_s^2, \quad j = 5, 7,$$
$$(\Gamma_{F^j})_{j,j} = p_V\mu c_s^2, \quad j = 6, 8.$$

We do not write down explicitly the sum of the 9 matrices in $\Gamma_{\mathcal{X}} = \Gamma_A + \sum_{j=1}^{K} \Gamma_{F^j}$.

In the pooled workload, we have (with $m = \frac{1}{\mu}$):

$$\hat{\mathcal{Z}}(t) = M\hat{\mathcal{X}}(t) = m \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \hat{\mathcal{X}}(t).$$

For $\hat{\mathcal{Z}}(t)$, it would at first glance appear that the drift of $\hat{\mathcal{Z}}(t)$ is $M\theta_{\mathcal{X}}$ and the covariance should be $M\Gamma_{\mathcal{X}}M^T$ but this is misleading. We note that

$$\mathcal{W}_{1+4}(t) = m\big(\mathcal{A}_1(t) + \mathcal{A}_2(t) + \mathcal{A}_3(t) + \mathcal{A}_4(t) - \mathcal{S}_1(\mathcal{T}_1(t)) - \mathcal{S}_4(\mathcal{T}_1(t))\big)$$
$$\mathcal{W}_{2+3}(t) = m\big(\mathcal{A}_1(t) + \mathcal{A}_2(t) + \mathcal{A}_3(t) + \mathcal{A}_4(t) - \mathcal{S}_3(\mathcal{T}_1(t)) - \mathcal{S}_4(\mathcal{T}_1(t))\big)$$

and we obtian immediately that:

$$\text{Drift of } \hat{\mathcal{Z}}(t) \qquad \theta_{1+4} = \theta_{2+3} = Nm(\lambda_H + \lambda_V - 2\mu)$$

and the variance covariance matrix is:

$$\Gamma_{\mathcal{Z}} = \begin{bmatrix} (\lambda_H + \lambda_V)c_a^2 + 2\mu c_s^2 & (\lambda_H + \lambda_V)c_a^2 \\ (\lambda_H + \lambda_V)c_a^2 & (\lambda_H + \lambda_V)c_a^2 + 2\mu c_s^2 \end{bmatrix}$$

17.6 Explain why the Brownian Laws and Louth network under optimal control behaves exactly like the fork-join network, and explain why for the original network the fork-join network provides a lower bound.

Solution

Consider the following fork-join system: There are two servers of service rate 2μ each, these represent the combined servers 1+4 and the combined servers 2+3. Customers arrive at rate $\lambda_H + \lambda_V$, and each arrival sends a task to each server, and the customer leaves when both tasks are complete. Let b_1, b_2 be the queue lengths at the two servers. Then the number of customers in the system that are waiting to be completed is $b_1 \vee b_2$.

But in the Laws and Louth system, under the optimal policy for the BCP, the total number of customers in the system is also $b_1 \vee b_2$, where $b_1 = \mu\hat{\mathcal{W}}_{1+4}$, $b_2 = \mu\hat{\mathcal{W}}_{2+3}$. So under balanced heavy traffic the fork-join system approximates the behavior of the Lasw and Louth network.

It is a lower bound, since in reality, the fork join system will indeed achieve $b_1 \vee b_2$, the 4 station system will not. So in the Brownian solution the servers 1+4 are pooled and servers 2+3 are pooled, and the two services of each customer are performed in parallel rather than in series. The analogy to the fork-join can be though of as if an arriving customer flips a coin and goes to 1+4 or to 2+3, and we clock the departure of the latest of the two.

17.7 In the network of Laws and Louth, assume Poisson arrivals and exponential service times, and use the policy of random routing and FCFS. Analyze this as a Jackson network, and find total expected number of customers in the system, expected total workloads, and expected sojourn times.

Solution

Under this regime, the network behaves like a feed forward network, and all the flows of customers will be poisson. Then each of the stations will have Poisson input at rate $\lambda = \lambda_H + \lambda_V$, and will serve it with exponential service of rate μ, with $\rho = \lambda/\mu$. Hence, the distribution of the number of customers at the stations, given by $n_1, \ldots, n_4$ is

$$\mathbb{P}(n_1, \ldots, n_4) = (1-\rho)^4 \prod_{i=1}^{4} \rho^{n_i}.$$

Each customer will require two services, each distributed $\sim \text{Exp}(\mu - \lambda)$. The expected sojourn time of each customer will be $\frac{2}{\mu-\lambda}$.

17.8 (∗) For the fork-join network that imitates that Laws and Louth network, under Poisson arrivals and exponential service times, find estimates for expected number in system, total workload, and sojourn time. Compare it to the uncontrolled Jackson network.

Solution

We have a fork-join with two ·/M/1 service stations. Arrivals are at rate 2λ service of each task is at rate 2μ.

It has been shown by Nelson and Tantawi (1988) that for this fork-join system, the expected sojourn time is

$$T_2 = \frac{12-\rho}{8} T_1, \quad T_1 = \text{M/M/1 sojourn time.}$$

Hence, in heavy traffic, when $\rho \approx 1$, the sojourn time will be $\approx \frac{11}{8} \frac{1}{2\mu - 2\lambda}$

The savings compared to random routing FCFS are by a factor of $32/11 \approx 3$.

17.9 For the network of Laws and Louth, repeat the derivations when service rates of the four stations are not all equal. Notice that there are some conditions on the processing rates that are needed so that the resulting balanced heavy traffic network will behave like the symmetric network.

Solution

Assume now that processing rates are μ_i at station i, and that c.o.v. of arrivals are $c_{a,H}, c_{a,V}$ and for service are $c_{s,i}$, $i = 1, \ldots, 4$. The new matrix R will be $(I - P^{\intercal})\text{diag}(\mu)$ where the elements $i, i+4$ share the same μ_i. Input to buffer k is α_k, $k = 1 \ldots, 4$ and $\alpha_k = 0$, $k = 5, \ldots, 8$. The static planning problem is:

$$\begin{aligned}
\max_{\nu,\alpha} \quad & w_H \alpha_1 + w_H \alpha_3 + w_V \alpha_2 + w_V \alpha_4 \\
s.t. \quad & R\nu = \alpha, \\
& C\nu \le \mathbf{1}, \\
& \alpha_1 + \alpha_3 = \lambda_H, \\
& \alpha_2 + \alpha_4 = \lambda_V, \\
& \alpha, \nu \ge 0
\end{aligned}$$

A sufficient condition for balanced heavy traffic is:

$$\begin{aligned}
&\lambda_H + \lambda_H \approx \mu_1 + \mu_4 \approx \mu_2 + \mu_3,\\
&\lambda_H < (\mu_1 + \mu_3) \wedge (\mu_2 + \mu_4),\\
&\lambda_V < (\mu_1 + \mu_2) \wedge (\mu_3 + \mu_4),
\end{aligned}$$

where the approximation is such that the difference time N is of moderate size for large N.

Once this the system in in balanced heavy traffic, the Brownian control problem is exactly the same as for the symmetric case (17.2), and for the "workload formulation" it is convenient to take the number of customers in the buffers of 1+4 and of 2+3, rather then workload, and thus define $\hat{\mathcal{W}}_{1+4}, \hat{\mathcal{W}}_{2+3}$, and the Brownian motion $\hat{\mathcal{Z}}_{1+4}, \hat{\mathcal{Z}}_{2+3}$.

The drift and variance of $\hat{\mathcal{Z}}$ are:

$$\text{Drift of } \hat{\mathcal{Z}}(t) \quad \theta_{1+4} = N(\lambda_H + \lambda_V - \mu_1 - mu_4), \quad \theta_{2+3} = N(\lambda_H + \lambda_V - \mu_2 - mu_3)$$

and the variance covariance matrix is:

$$\Gamma_{\mathcal{Z}} = \begin{bmatrix} \lambda_H c_{a,H}^2 + \lambda_V c_{a,V}^2 + \mu_1 c_{s,1}^2 + + \mu_4 c_{s,4}^2 & \lambda_H c_{a,H}^2 + \lambda_V c_{a,V}^2 \\ \lambda_H c_{a,H}^2 + \lambda_V c_{a,V}^2 & \lambda_H c_{a,H}^2 + \lambda_V c_{a,V}^2 + \mu_2 c_{s,2}^2 + + \mu_3 c_{s,3}^2 \end{bmatrix}.$$

This is independent of the controls and the policy. The solution of the Brownian control problem is therefore the same as for the symmetric problem.

17.10 (∗) Formulate the control problem for the Laws and Louth network with additional admission control, and required total input rate of $\bar{\lambda}$ and describe its optimal solution.

Solution This problem is open, to the best of my knowledge it has not been discussed in the literature.

We now assume that $\mathcal{A}(t)$ is controlled, and each admitted customer will be horizontal with probability p_H and vertical with probability p_V. We write the queue dynamics for some of the buffers. For buffer 1,

$$\begin{aligned}
\mathcal{Q}_1(t) &= \mathcal{Q}_1(0) + \mathcal{N}_1(t) - \mathcal{S}_1(\mathcal{T}_1(t))\\
&= \left[\mathcal{Q}_1(0) + (\tfrac{1}{2} p_H \bar{\alpha} t - p_H \mu_1 t) - (\mathcal{S}_1(\mathcal{T}_1(t) - \mu_1 \mathcal{T}_1(t))\right]\\
&\quad + \mu_1 (p_H t - \mathcal{T}_1(t)) - \tfrac{1}{2} p_H (\bar{\alpha} t - \mathcal{A}(t)) + (\mathcal{N}_1(t) - \tfrac{1}{2} p_H \mathcal{A}(t))\\
&= \mathcal{X}_1(t) + \mu_1 \mathcal{J}_1(t) - \tfrac{1}{2} p_H \eta(t) + \mathcal{V}_1(t).
\end{aligned}$$

Here the input to buffer 1 is totally controlled, we decide when to admit a customer, and we then send it horizontally or vertically in deterministic splitting according to p_H, p_V. Once a customer is admitted to the horizontal stream, we decide to route it to buffer 1 or to buffer 3. Hence in the netput we include just the deterministic input to buffer 1, $\frac{1}{2} p_H \bar{\alpha} t$.

We then have 3 controls: $\mathcal{J}_1$ is the free time for processing at buffer 1, $\mathcal{V}_1$ is

the control for routing admitted customers to buffer 1, beyond the nominal $\frac{1}{2}$, and $\eta(t)$ is the admission control.
The dynamics of buffers 2,3,4 are analogous. Similarly for buffer 5 (analogously for buffers 6,7,8),

$$\begin{aligned} Q_5(t) &= Q_5(0) + \mathcal{S}_1(\mathcal{T}_1(t)) - S_5(T_2(t)) \\ &= \Big[Q_5(0) + p_H\mu_1 t - p_H\mu_2 t \\ &\quad +(\mathcal{S}_1(\mathcal{T}_1(t) - \mu_1\mathcal{T}_1(t)) - (\mathcal{S}_5(\mathcal{T}_1(t) - \mu_2\mathcal{T}_5(t))\Big] \\ &\quad -\mu_1(p_H t - \mathcal{T}_1(t)) + \mu_2(p_H t - \mathcal{T}_5(t)), \end{aligned}$$

with free time controls $\mathcal{J}_1, \mathcal{J}_5$.
We can now scale time by N^2 and space by N and write the approximately Brownian dynamics as

$$\hat{Q}(t) = \hat{X}(t) + R\hat{\mathcal{J}}(t) - G\hat{\mathcal{V}}(t) + D\hat{\eta}(t),$$

where $D^{\intercal} = \Big[\frac{1}{2}p_H\ ;\frac{1}{2}p_V\ ;\frac{1}{2}p_H\ ;\frac{1}{2}p_V\ ;0\ ;0\ ;0\ ;0\Big]$.
The Brownian control problem is

$$\begin{aligned} \min\ & \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^{8} \hat{Q}_k(t)dt\,], \\ \text{s.t. } & \hat{Q}(t) = \hat{X}(t) + R\,\hat{\mathcal{J}}(t) + G\hat{\mathcal{V}}(t) + D\hat{\eta}(t), \\ & \hat{I}(t) = C\hat{\mathcal{J}}(t), \qquad H\hat{\mathcal{V}}(t) = 0, \\ & \lim_{T\to\infty} \frac{1}{T}\mathbb{E}(\mathcal{A}(T)) \geq \bar{\alpha}, \\ & \hat{Q}(t) \geq 0, \quad \hat{I}(0) = 0,\ \hat{I} \text{ non-decreasing}, \\ & \hat{Q}, \hat{\mathcal{J}}, \hat{\mathcal{V}}, \hat{\eta} \text{ are non-anticipating with respect to } \hat{X}(t). \end{aligned}$$

This leads to the pooled workload formulation at the cut defined by stations 1+4 and 2+3, with (note, we call this workload, but in fact it is the pooled number of customers)

$$\hat{\mathcal{W}}_P = \begin{bmatrix} \hat{\mathcal{W}}_{1+4} \\ \hat{\mathcal{W}}_{2+3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} Q(t) := MQ(t).$$

and with the two dimensional Brownian motion $\hat{\mathcal{Z}}(t) = M\mathcal{X}(t)$, where:

$$\text{Drift of } \hat{\mathcal{Z}}(t) \quad \theta_{1+4} = \theta_{2+3} = N(\bar{\alpha} - 2\mu),$$

$$\text{Variance of } \hat{\mathcal{Z}}(t) \quad \Gamma_{\mathcal{Z}} = \begin{bmatrix} 2\mu c_s^2 & 0 \\ 0 & 2\mu c_s^2 \end{bmatrix}$$

The pooled workload Brownian control problen is: For $\hat{\mathcal{Z}}(t)$ find $\hat{Q}, \hat{I}, \hat{\mathcal{V}}, \hat{\eta}$

such that

$$
\begin{aligned}
V = \min \ & \lim_{T\to\infty} \sup \frac{1}{T} E[\int_0^T \sum_{k=1}^{8} \hat{Q}_k(t)dt\,], \\
\text{s.t. } & \hat{\mathcal{W}}_P = M\hat{Q}(t), \\
& \hat{\mathcal{W}}_{1,4}(t) = \hat{\mathcal{Z}}_1(t) + \hat{\mathcal{I}}_1(t) + \hat{\mathcal{I}}_4(t) + \hat{\eta}(t), \\
& \hat{\mathcal{W}}_{2,3}(t) = \hat{\mathcal{Z}}_2(t) + \hat{\mathcal{I}}_2(t) + \hat{\mathcal{I}}_3(t) + \hat{\eta}(t), \\
& \lim_{T\to\infty} \frac{1}{T}\mathbb{E}(\mathcal{I}_1(T) + \mathcal{I}_4(T)) \le N(1-\rho), \\
& \lim_{T\to\infty} \frac{1}{T}\mathbb{E}(\mathcal{I}_2(T) + \mathcal{I}_3(T)) \le N(1-\rho), \\
& \hat{Q}(t) \ge 0, \quad \hat{\mathcal{I}}(0) = 0, \ \hat{\mathcal{I}} \text{ non-decreasing}, \quad t \ge 0, \\
& \hat{Q}, \hat{\mathcal{I}}, \hat{\eta}, \text{ are non-anticipating with respect to } \hat{\mathcal{Z}}(t),
\end{aligned}
$$

with $\rho = \bar{\alpha}/\mu$

In the original problem we were given $\hat{X}$ and wished to solve for $\hat{Q}, \hat{\mathcal{J}}, \hat{\mathcal{V}}, \hat{\eta}$. We now defined $\hat{\mathcal{Z}}$ to solve for $\hat{Q}, \hat{\mathcal{I}}, \hat{\eta}$. We claim that any solution of the workload formulation, i.e. given $\hat{X}$, and a solution $\hat{Q}, \hat{\mathcal{I}}, \hat{\eta}$ for the workload formulation, we can retrieve $\hat{\mathcal{J}}, \hat{\mathcal{V}}$. They should come from the equations:

$$
\begin{aligned}
& R\hat{\mathcal{J}} + G\hat{\mathcal{V}} = \hat{Q} - \hat{X} - D\hat{\eta}, \\
& C\hat{\mathcal{J}} = \hat{\mathcal{I}},
\end{aligned}
$$

which have unique solution, since the ranks of the coefficients matrix without or with the added column of the r.h.s. agree and are 10.

We now note that given the pooled workloads the optimal solution will be:

$$
\begin{aligned}
& V = \hat{\mathcal{W}}_{1+4} \vee \hat{\mathcal{W}}_{2+3}, \\
& \hat{Q}_1 + \hat{Q}_2 + \hat{Q}_3 + \hat{Q}_4 = \hat{\mathcal{W}}_{1+4} \wedge \hat{\mathcal{W}}_{2+3} \\
& \hat{Q}_5 + \hat{Q}_6 = (\hat{\mathcal{W}}_{1+4} - \hat{\mathcal{W}}_{2+3})^-, \quad \hat{Q}_7 + \hat{Q}_8 = (\hat{\mathcal{W}}_{1+4} - \hat{\mathcal{W}}_{2+3})^+.
\end{aligned}
$$

Similar to the control of the Laws and Louth network with uncontrolled input, we would keep 0 inventory in one of the sets of buffers at all times, i.e. $\hat{Q}_5 + \hat{Q}_6 \wedge \hat{Q}_7 + \hat{Q}_8 = 0$. But we now have an additional control, $\hat{\eta}(t)$ for admissions. Using the admissions control we can keep all the input buffers, $\hat{Q}_k$, $k = 1, 2, 3, 4$ empty while they still process customers. So all the variability is concentrated in the otput buffers $\hat{Q}_k$, $k = 5, 6, 7, 8$.

We now look at the pooled workload imbalance: $\hat{\mathcal{W}}_\Delta = \hat{\mathcal{W}}_{1+4} - \hat{\mathcal{W}}_{2+3}$. It satisfies the following dynamics:

$$
h\mathcal{W}_\Delta = \hat{\mathcal{Z}}_\Delta + (\hat{\mathcal{I}}_1(t) + \hat{\mathcal{I}}_4(t)) - (\hat{\mathcal{I}}_2(t) + \hat{\mathcal{I}}_3(t))
$$

where the Brownian motion $\hat{\mathcal{Z}}_\Delta = \hat{\mathcal{Z}}_{1+4} - \hat{\mathcal{Z}}_{2+3}$ has drift $m = 0$ and variance $\sigma^2 = 4\mu c_s^2$.

Our optimal policy then is: use $\hat{\eta}$ to keep the input buffers empty. Keep one of the output buffer pairs $\hat{Q}_5 + \hat{Q}_6$ or $\hat{Q}_7 + \hat{Q}_8$ empty, using the routing controls, and the sequencing at server 1. Use idling so as to keep $-a < \hat{\mathcal{W}}_\Delta < a$ where a is determined so that each server is idle a fraction $N(1-\rho)$ of the time, calculated from (7.19). Sequencing at servers 2 and 3 gives priority to output buffers, sequencing at server 4 is FCFS.

17.11 For the cube network with three types of customers, formulate the BCP and derive the optimal Brownian solution. Show the analogy to a fork-join network.

Solution

In the cube network there are 8 servers, and there are 24 buffers, 3 for each server. 12 of the buffers are input buffers, the other 12 are exit buffers.

Server 1 serves 3 entry buffers, servers 2,3,5 serve 2 entry buffers and 1 exit buffer, servers 4,6,7 server 1 entry and 2 exit buffers, server 8 serves 3 exit buffers.

We can number the buffers according to the route, (a, i_1, i_2), (b, i_1, i_2) for entry and exit buffers of the route of servers i_1, i_2, with dynamics:

$$\begin{aligned}
Q_{a,i_1,i_2}(t) &= \Big[Q_{a,i_1,i_2}(0) + (\frac{1}{4}\lambda - \frac{1}{3}\mu)t - (\mathcal{S}_{i_1}(\mathcal{T}_{i_1}(t)) - \mu\mathcal{T}_{i_1}(t))\Big] \\
&\quad +\mu(\frac{1}{3}t - \mathcal{T}_{i_1}(t)) + (\mathcal{V}_{i_1}(t) - \frac{1}{4}\lambda t), \\
Q_{b,i_1,i_2}(t) &= \Big[Q_{b,i_1,i_2}(0) + (\mathcal{S}_{i_1}(\mathcal{T}_{i_1}(t)) - \mu\mathcal{T}_{i_1}(t)) - (\mathcal{S}_{i_2}(\mathcal{T}_{i_2}(t)) - \mu\mathcal{T}_{i_2}(t))\Big] \\
&\quad -\mu(\frac{1}{3}t - \mathcal{T}_{i_1}(t)) + \mu(\frac{1}{3}t - \mathcal{T}_{i_2}(t)).
\end{aligned}$$

We then scale and use a Gaussian approximation to get the Brownian control problem:

$$\begin{aligned}
\min\ & \lim\sup_{T\to\infty} \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^{2} 4\hat{Q}_k(t)dt\,], \\
\text{s.t. } & \hat{Q}(t) = \hat{X}(t) + R\,\hat{\mathcal{J}}(t) + G\hat{\mathcal{V}}(t), \\
& \hat{\mathcal{I}}(t) = C\hat{\mathcal{J}}(t), \qquad H\hat{\mathcal{V}}(t) = 0, \\
& \hat{Q}(t) \geq 0, \quad \hat{\mathcal{I}}(0) = 0,\ \hat{\mathcal{I}} \text{ non-decreasing}, \\
& \hat{Q}, \hat{\mathcal{J}}, \hat{\mathcal{V}} \text{ are non-anticipating with respect to } \hat{X}(t).
\end{aligned}$$

Where $R = \mu(I - P^{\mathsf{T}})$, G is identity matrix for the 12 entry buffers and 0 for the exit buffers, C is the constituency matrix, and H has 3 rows, adding up the 4 routing controls of each face.

We now note that each customer has to go through one of the servers $\mathcal{S}_1 = (1,4,6,7)$, and also through one of the servers $\mathcal{S}_2 = (2,3,5,8)$. We define

the pooled workloads (actually numbers of customers):

$$\begin{aligned}
\hat{\mathcal{W}}_{\mathcal{S}_1} &= \hat{Q}_{a,1,2} + \hat{Q}_{a,1,3} + \hat{Q}_{a,1,5} + \hat{Q}_{a,2,4} + \hat{Q}_{a,2,6} + \hat{Q}_{a,3,4} \\
&\quad + \hat{Q}_{a,3,7} + \hat{Q}_{a,4,8} + \hat{Q}_{a,5,6} + \hat{Q}_{a,5,7} + \hat{Q}_{a,6,8} + \hat{Q}_{a,7,8} \\
&\quad + \hat{Q}_{b,2,4} + \hat{Q}_{b,2,6} + \hat{Q}_{b,3,4} + \hat{Q}_{b,3,7} + \hat{Q}_{b,5,6} + \hat{Q}_{b,5,7}, \\
\hat{\mathcal{W}}_{\mathcal{S}_2} &= \hat{Q}_{a,1,2} + \hat{Q}_{a,1,3} + \hat{Q}_{a,1,5} + \hat{Q}_{a,2,4} + \hat{Q}_{a,2,6} + \hat{Q}_{a,3,4} \\
&\quad + \hat{Q}_{a,3,7} + \hat{Q}_{a,4,8} + \hat{Q}_{a,5,6} + \hat{Q}_{a,5,7} + \hat{Q}_{a,6,8} + \hat{Q}_{a,7,8} \\
&\quad \hat{Q}_{b,1,2} + \hat{Q}_{b,1,3} + \hat{Q}_{b,1,5} + \hat{Q}_{b,4,8} + \hat{Q}_{b,6,8} + \hat{Q}_{b,7,8}.
\end{aligned}$$

We note that we can write

$$\hat{\mathcal{W}}_{\mathcal{S}_1} = \hat{Q}_{in} + \hat{Q}_{out-\mathcal{S}_1}, \quad \hat{\mathcal{W}}_{\mathcal{S}_2} = \hat{Q}_{in} + \hat{Q}_{out-\mathcal{S}_2},$$

where $\hat{Q}_{in}$ are all the input buffers, and $\hat{Q}_{out-\mathcal{S}_i}$, $i = 1, 2$ are the exit buffers of $\mathcal{S}_i$, $i = 1, 2$.
The workload formulation of the workload Brownian control problem is

$$\begin{aligned}
\min \ & \lim_{T\to\infty} \sup \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^{2} 4\hat{Q}_k(t)dt\,], \\
\text{s.t.} \ & \hat{\mathcal{W}}_{pooled}(t) = M\hat{Q}(t) \\
& \hat{\mathcal{W}}_{\mathcal{S}_1} = \hat{\mathcal{Z}}_{\mathcal{S}_1} + \hat{I}_1(t) + \hat{I}_4(t) + \hat{I}_6(t) + \hat{I}_7(t), \\
& \hat{\mathcal{W}}_{\mathcal{S}_2} = \hat{\mathcal{Z}}_{\mathcal{S}_1} + \hat{I}_2(t) + \hat{I}_3(t) + \hat{I}_5(t) + \hat{I}_8(t), \\
& \hat{Q}(t) \geq 0, \quad \hat{I}(0) = 0, \ \hat{I} \text{ non-decreasing}, \\
& \hat{Q}, \hat{I} \text{ are non-anticipating with respect to } \hat{\mathcal{X}}(t).
\end{aligned}$$

It needs to be checked that for given $\hat{\mathcal{X}}, \hat{\mathcal{Z}}, \hat{Q}, \hat{I}$ we can reconstruct also $\hat{\mathcal{J}}_k$, $\hat{\mathcal{V}}_k$. This is done by solving:

$$\begin{aligned}
&R\hat{\mathcal{J}} + G\hat{\mathcal{V}} = \hat{Q} - \hat{\mathcal{X}}, \\
&C\hat{\mathcal{J}} = \hat{I}, \\
&H\hat{\mathcal{V}} = 0.
\end{aligned}$$

There are 44 equations, for 36 unknowns. However, a solution exists, because one can check that the ranks of the coefficient matrix with and without the r.h.s. is 36.
The optimal policy is to idle only if $\hat{\mathcal{W}}_{\mathcal{S}_1}$ or $\hat{\mathcal{W}}_{\mathcal{S}_2}$ or both are empty, and

$$\begin{aligned}
&V = \hat{\mathcal{W}}_{\mathcal{S}_1} \vee \hat{\mathcal{W}}_{\mathcal{S}_1}, \quad \hat{Q}_{in} = \hat{\mathcal{W}}_{\mathcal{S}_1} \wedge \hat{\mathcal{W}}_{\mathcal{S}_1} \\
&\hat{Q}_{out-\mathcal{S}_1} = (\hat{\mathcal{W}}_{\mathcal{S}_1} - \hat{\mathcal{W}}_{\mathcal{S}_1})^+ \quad \hat{Q}_{out-\mathcal{S}_2} = (\hat{\mathcal{W}}_{\mathcal{S}_1} - \hat{\mathcal{W}}_{\mathcal{S}_1})^-
\end{aligned}$$

This behaves exactly like the Laws and Louth network, and in balanced heavy traffic limit it behaves like a two station fork join network.

17.12 For the ring network with six stations and three types of customers, formulate the BCP and derive the optimal Brownian solution. Show the analogy to a fork-join network.

Solution

The system has 18 buffers, that corresponds to entry, passage and exit of each of the 6 routes (2 for each type). We index the buffers by (x, i_1, i_2, i_3, x) where $x = 1, 2, 3$ is the type of customer on the route, , i_1, i_2, i_3 is the route, and $y = I, P, E$ for entry, passage and exit, for processor is the stage on the route. The dynamics are:

$$
\begin{aligned}
Q_{x,i_1,i_2,i_3,I}(t) &= \Big[Q_{x,i_1,i_2,i_3,I}(0) + (\tfrac{1}{2}\lambda - \tfrac{1}{3}\mu)t \\
&\quad + \tfrac{1}{2}(\mathcal{A}_x(t) - \lambda t) - (S_{x,i_1,i_2,i_3,I}(\mathcal{T}_{x,i_1,i_2,i_3,I}(t) - \mu\mathcal{T}_{x,i_1,i_2,i_3,I}(t))\Big] \\
&\quad + \mu(\tfrac{1}{3}t - \mathcal{T}_{x,i_1,i_2,i_3,I}(t)) + (\mathcal{A}_{i_1,i_2,i_3} - \tfrac{1}{2}\mathcal{A}_x(t)), \\
Q_{x,i_1,i_2,i_3,P}(t) &= \Big[Q_{x,i_1,i_2,i_3,P}(0) + (S_{x,i_1,i_2,i_3,I}(\mathcal{T}_{x,i_1,i_2,i_3,I}(t) - \mu\mathcal{T}_{x,i_1,i_2,i_3,I}(t)) \\
&\quad - (S_{x,i_1,i_2,i_3,P}(\mathcal{T}_{x,i_1,i_2,i_3,P}(t) - \mu\mathcal{T}_{x,i_1,i_2,i_3,P}(t))\Big] \\
&\quad - \mu(\tfrac{1}{3}t - \mathcal{T}_{x,i_1,i_2,i_3,I}(t)) + \mu(\tfrac{1}{3}t - \mathcal{T}_{x,i_1,i_2,i_3,P}(t)), \\
Q_{x,i_1,i_2,i_3,E}(t) &= \Big[Q_{x,i_1,i_2,i_3,E}(0) + (S_{x,i_1,i_2,i_3,P}(\mathcal{T}_{x,i_1,i_2,i_3,P}(t) - \mu\mathcal{T}_{x,i_1,i_2,i_3,P}(t)) \\
&\quad - (S_{x,i_1,i_2,i_3,E}(\mathcal{T}_{x,i_1,i_2,i_3,E}(t) - \mu\mathcal{T}_{x,i_1,i_2,i_3,E}(t))\Big] \\
&\quad - \mu(\tfrac{1}{3}t - \mathcal{T}_{x,i_1,i_2,i_3,P}(t)) + \mu(\tfrac{1}{3}t - \mathcal{T}_{x,i_1,i_2,i_3,E}(t)).
\end{aligned}
$$

We scale time by N^2 and space by N and we write:

$$\hat{Q}(t) = \hat{\mathcal{X}}(t) + R\hat{\mathcal{J}}(t) + G\hat{\mathcal{V}}(t),$$

where $\mathcal{X}$ is the Brownian motion netput, $\hat{\mathcal{J}}(t)$ the sequencing control, and $\hat{\mathcal{V}}$ the routing control. The matrices R and G are obtained from examining the dynamics equations above.

We note that every customer needs to go through one of the processors $(1, 4)$ one of $(2, 5)$ and one o $(3, 6)$. We look now at the workloads (number of customers) in the system that require processing at the pooled server pairs $(1, 4)$, $(2, 5)$, $(3, 6)$. This is non-overlapping, and includes all the work in the system. The scaled Pooled workload is $\hat{\mathcal{W}}_{Pooled}(t) = \begin{bmatrix} \hat{\mathcal{W}}_{1+4} \\ \hat{\mathcal{W}}_{2+5} \\ \hat{\mathcal{W}}_{3+6} \end{bmatrix} = MQ(t)$.

The matrix M is described in the following table: on top are listed the buffer, given by the route (vertical) and the step, I,P,E, and the three rows correspond to the pairs of stations, $(1, 4)$, $(2, 5)$, $(3, 6)$.

	1			2			3			2			3			4	
	6			1			2			3			4			5	
	5			6			1			4			5			6	
I	P	E	I	P	E	I	P	E	I	P	E	I	P	E	I	P	E
1			1	1		1	1	1	1	1	1	1	1		1		
1	1	1	1			1	1		1			1	1	1	1	1	
1	1		1	1	1	1			1	1		1			1	1	1

Having gone through the laborious process of identifying
The pooled Brownian control problem is now:

$$
\begin{aligned}
\min\ & \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^{1} 8\hat{Q}_k(t)dt\,], \\
\text{s.t. } & \hat{\mathcal{W}}_{pooled}(t) = M\hat{Q}(t) \\
& \hat{\mathcal{W}}_{1+4}(t) = \hat{\mathcal{Z}}_{1+4}(t) + \hat{I}_1(t) + \hat{I}_4(t), \\
& \hat{\mathcal{W}}_{2+5}(t) = \hat{\mathcal{Z}}_{2+5}(t) + \hat{I}_2(t) + \hat{I}_5(t), \\
& \hat{\mathcal{W}}_{3+6}(t) = \hat{\mathcal{Z}}_{3+6}(t) + \hat{I}_3(t) + \hat{I}_6(t), \\
& \hat{Q}(t) \geq 0, \quad \hat{I}(0) = 0,\ \hat{I} \text{ non-decreasing}, \\
& \hat{Q}, \hat{I} \text{ are non-anticipating with respect to } \hat{\mathcal{X}}(t).
\end{aligned}
$$

We check that for given $\hat{\mathcal{X}}, \hat{\mathcal{Z}}$, the values of $\hat{Q}, \hat{I}$ can be used to obtain the controls of the original scaled network $\hat{\mathcal{J}}, \hat{\mathcal{V}}$. There are 27 equations for 24 unknowns, but 3 equations are redundant because they just say that the sum of the relevant buffers in $\hat{\mathcal{X}}$ equal the component of $\hat{\mathcal{Z}}$, on both sides of the equations, and a solution exists (in fact is unique)

We can now solve the pooled Brownian control problem: We keep all the inventory in the 6 input buffers, balancing the contents of the two routes, we balances the workloads in the two servers of each pooled pair, and we idle only when the pooled workload is 0.

We return to determination of the queue lengths. Denote $b_1(t) = \hat{\mathcal{W}}_{1+4}(t),\ b_2(t) = \hat{\mathcal{W}}_{2+5}(t),\ b_3 = \hat{\mathcal{W}}_{3+6}(t)$. Then the minimization pathwise is to solve for each

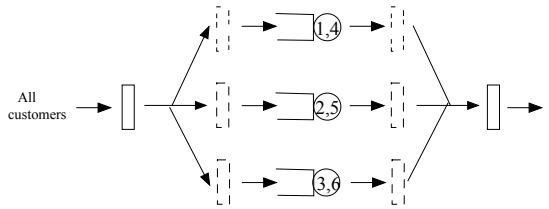
t the problem:

$$\begin{aligned}
\min \ & \sum_{k=1}^{18} \hat{Q}_k(t), \\
\text{s.t.} \ & \sum_{k=1}^{6} \hat{Q}_k + \hat{Q}_8(t) + \hat{Q}_9(t)\hat{Q}_{10}(t) + \hat{Q}_{11}(t) + \hat{Q}_{14}(t)\hat{Q}_{17}(t) = b_1(t), \\
& \sum_{k=1}^{6} \hat{Q}_k + \hat{Q}_7(t) + \hat{Q}_{10}(t)\hat{Q}_{11}(t) + \hat{Q}_{12}(t) + \hat{Q}_{13}(t)\hat{Q}_{16}(t) = b_2(t), \\
& \sum_{k=1}^{6} \hat{Q}_k + \hat{Q}_7(t) + \hat{Q}_8(t)\hat{Q}_9(t) + \hat{Q}_{12}(t) + \hat{Q}_{15}(t)\hat{Q}_{18}(t) = b_3(t),
\end{aligned}$$

and the solution to this is:

$$\begin{aligned}
& V^*(t) = b_1(t) \vee b_2(t) \vee b_3(t), \\
& \sum_{k=1}^{6} \hat{Q}_k = b_1(t) \wedge b_2(t) \wedge b_3(t), \\
& \hat{Q}_7(t) + \hat{Q}_{12}(t) = \big(b_2(t) \wedge b_3(t) - b_1(t)\big)^+ \\
& \hat{Q}_8(t) + \hat{Q}_9(t) = \big(b_1(t) \wedge b_3(t) - b_2(t)\big)^+ \\
& \hat{Q}_{10}(t) + \hat{Q}_{11}(t) = \big(b_1(t) \wedge b_2(t) - b_3(t)\big)^+ \\
& \hat{Q}_{13}(t) + \hat{Q}_{16}(t) = \big(b_2(t) - b_1(t) \vee b_3(t)\big)^+ \\
& \hat{Q}_{14}(t) + \hat{Q}_{17}(t) = \big(b_1(t) - b_2(t) \vee b_3(t)\big)^+ \\
& \hat{Q}_{15}(t) + \hat{Q}_{18}(t) = \big(b_3(t) - b_1(t) \vee b_2(t)\big)^+
\end{aligned}$$

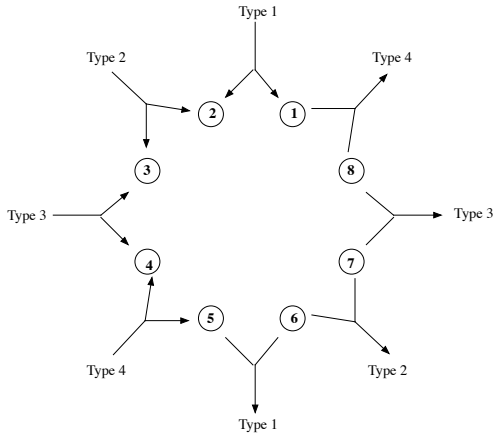
This is analogous to a fork join network with 3 parallel servers, where each arriving customer splits into 3 tasks that are processed in parallel at the 3 servers, and a customer departs when his three tasks are complete.



17.13 (∗) For a ring network with 8 stations, obtain a pooled workload Brownian problem, and show how to minimize the pooled workloads a.s. pathwise, but then show that this solution does not minimize the sum of queue lengths, and there is no a.s. pathwise optimal solution to the original BCP.

Solution

This is based on Laws (1990). In this network there are eight workstations and 4 types of customers, each type can choose to go clockwise or anticlockwise to receive service at four stations, as illustrated here:



Following the same steps as for the six station network, we pool the queues for stations $(1,5),(2,6),(3,7),(4,8)$, and obtain $\hat{\mathcal{W}}_{pooled}(t) = M\hat{Q}(t)$, $\hat{\mathcal{Z}}_{pooled}(t) = M\hat{X}(t)$, where

$$M_{r,k} = \begin{cases} 1 & k \in C_r \cup C_{r+4} \\ 0 & \text{otherwise} \end{cases}, \quad r = 1,\ldots,4,\ k = 1,\ldots,32.$$

and we get the workload BCP:

$$\begin{aligned}
\min\ & \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[\int_0^T \sum_{k=1}^{32} \hat{Q}_k(t)dt\,], \\
\text{s.t. } & \hat{\mathcal{W}}_{pooled}(t) = M\hat{Q}(t) \\
& \hat{\mathcal{W}}_{1+5}(t) = \hat{\mathcal{Z}}_{1+5}(t) + \hat{I}_1(t) + \hat{I}_5(t), \\
& \hat{\mathcal{W}}_{2+6}(t) = \hat{\mathcal{Z}}_{2+6}(t) + \hat{I}_2(t) + \hat{I}_6(t), \\
& \hat{\mathcal{W}}_{3+7}(t) = \hat{\mathcal{Z}}_{3+7}(t) + \hat{I}_3(t) + \hat{I}_7(t), \\
& \hat{\mathcal{W}}_{4+8}(t) = \hat{\mathcal{Z}}_{4+8}(t) + \hat{I}_4(t) + \hat{I}_8(t), \\
& \hat{Q}(t) \geq 0, \quad \hat{I}(0) = 0,\ \hat{I} \text{ non-decreasing}, \\
& \hat{Q}, \hat{I} \text{ are non-anticipating with respect to } \hat{X}(t).
\end{aligned}$$

The solution to this workload BCP is similar to the one for the six stations network: Since the pooled stations do not overlap, we can choose the idling for each pair of stations independent from the others, and so, for given Brownian

motion $\hat{\mathcal{Z}}$, we let

$$\hat{\mathcal{I}}_r(t) + \hat{\mathcal{I}}_{r+4}(t) = \sup_{0<s<t} \left(\hat{\mathcal{Z}}_r(s) + \hat{\mathcal{Z}}_{r+4}(s)\right)^-, \quad r = 1, \ldots, 4.$$

We now try to minimize the queue lengths, by solving the LP at each point t, and see that there is a problem. Denote $b_r(t) = \hat{\mathcal{W}}_{r,r+4}(t)$, $r = 1, \ldots, 4$ we get:

$$\begin{aligned} \min V(t) \ & \sum_{k=1}^{32} \hat{Q}_k(t), \\ \text{s.t.} \ & \sum_{k \in C_r \cup C_{r+4}} \hat{Q}_k(t) = b_r(t), \quad r = 1, \ldots, 4, \\ & \hat{Q}(t) \geq 0. \end{aligned}$$

It is immediate to see that (as in the six station case), $V(t) \geq \max(b_r(t),\ r = 1, \ldots, 4)$. In the six station case for any values of b_r this lower bound could be achieved. However, this is no longer the case for the eight station case. One can see that:

$$b_4 < b_2 \wedge (b_1 \wedge b_3 - b_2) \Longrightarrow \sum_{k=1}^{32} \hat{Q}_k(t) > \max(b_r(t),\ r = 1, \ldots, 4.$$

It then turns out that to get close to this lower bound value it is better some times to idle some stations even when there ar customers in the station. For details see Laws (1990)

17.14 Obtain the exact expressions for R, G, C, H in equation (17.3).

Solution

The matrix R is the input output matrix, $(I - P^{\intercal})\text{diag}(\mu)$. The matrix G has dimension $K \times L$ where L is the number of routes, K the number of buffers (classes), and $G_{k,l} = 1$ if k is the first buffer on route l and 0 otherwise. C is the constituency matrix, with $C_{i,k} = 1$ when buffer k is processed at station i, 0 otherwise. H is the flow allocation matrix, of dimension $M \times L$, where M is the number of customer types, and $H_{m,l} = 1$ if route l is of customers of type m, 0 otherwise, so multiplying by M sums the total flow rate of customers of type m.

Thus, the first constraint writes the scaled buffer contents $\hat{Q}(t)$, as netput under the nominal control, approximated by a Brownian motion $\hat{\mathcal{X}}(t)$, plus sequencing control given by the free times $\hat{\mathcal{J}}(t)$ times R, plus discretionary routing controls $\hat{\mathcal{V}}(t)$ times G. The next two constraints say that the free times have to add up to the total idling $\hat{\mathcal{I}}(t)$, and the routing controls which are deviations form the nominal controls that determine the netput, have to add up to zero.

17.15 Justify the formulation and the solution of the BCP (17.4).

Solution

To wish to show that the solution of the BCP (17.4) is indeed the solution of the Brownian control problem (17.3).
Consider first the workload Brownian control problem (17.4). We can solve it optimally for every sample path and the solution is path-wise optimal:

$$\begin{aligned}
\min \ & \int_0^T \sum_{k=1}^K \hat{Q}_k(t,\omega)dt \,] \\
\text{s.t. } & \hat{\mathcal{W}}(t,\omega) = \sum_{k=1}^K \bar{m}_k \hat{Q}_k(t,\omega) \\
& \hat{\mathcal{W}}(t,\omega) = \hat{\mathcal{Z}}(t,\omega) + \sum_{i=1}^I \pi_i^* \mu_i \hat{I}_i(t,\omega), \\
& \hat{Q}(t,\omega) \geq 0, \qquad \hat{I}(0) = 0, \quad \hat{I} \text{ non-decreasing}, \\
& \hat{Q}, \hat{I} \text{ are non-anticipating with respect to } \hat{\mathcal{Z}}(t,,\omega),
\end{aligned}$$

Then for given $\hat{\mathcal{W}}(t,\omega)$, we can solve for $\hat{Q}_k(t,\omega)$ pathwise at every t:

$$\begin{aligned}
\min \ & \sum_{k=1}^K \hat{Q}_k(t,\omega)dt \,] \\
\text{s.t. } & \hat{\mathcal{W}}(t,\omega) = \sum_{k=1}^K \bar{m}_k \hat{Q}_k(t,\omega) \\
& \hat{Q}(t,\omega) \geq 0,
\end{aligned}$$

and the solution is any choice of $\hat{Q}_k(t,\omega)$ such that only buffers with $\bar{m}_k = \bar{\bar{m}}$ have any fluid.

$$\sum_{k \in K^*} \hat{Q}_k(t) = \frac{1}{\bar{\bar{m}}} \hat{\mathcal{W}}(t), \qquad \hat{Q}_k(t) = 0, \quad k \notin K^*.$$

These will be first buffers on their route, and for those we have $\bar{\bar{m}} = \max_c h_c^*$. Clearly we should try and minimize $\hat{\mathcal{W}}(t,\omega)$, and this is done by the Skorohod reflection. we can choose any combination of idle times that satisfy:

$$\sum_{i=1}^I \pi_i^* \mu_i \hat{I}_i(t,\omega) = -\inf_{0 \leq s \leq t} \hat{\mathcal{Z}}(s,\omega),$$

We now need to show that the solution of (17.4) solves (17.3). We first claim that solution of (17.4) provides a lower bound for (17.3). We can see that the single constraint expressing $\hat{\mathcal{W}}$ in terms of $\hat{\mathcal{Z}}$ in (17.4), is obtained by multiplying each buffer constraint in (17.3) by $\bar{m}_k$ and summing up. Hence (17.4) is a relaxation of (17.3) and provides a lower bound.
Next needs to show that from a solution of (17.4) we can construct a solution for (17.3). Solution of (17.4) determines $\hat{Q}(t)$, $\hat{I}(t)$, and we still also have

the original $X_k(t)$. Form these we wish to determine $\mathcal{J}_k(t)$ and $\hat{V}_r(t)$. They are connected by a set of linear equations. One can show that these equations always have a solution, which is a feasible set of controls for (17.3).
For details of the proofs see Laws (1990, 1992).

17.16 For the six node network of Figure 17.6, write the compatibility linear program and its dual, and identify all the 29 cut constraints. Show that there are no more than 29.

Solution

The network has machines $1, \ldots, 6$ and flows $1, \ldots, 5$. The compatibility LP is:

$$\begin{aligned}
&\min \quad \eta \\
&\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} - \mathbf{1}\eta \leq \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} \\
&\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \\
&f_r \geq 0
\end{aligned}$$

The dual compatibility LP is:

$$\max \lambda_1 h_1 + \lambda_2 h_2 - \mu_1\pi_1 - \mu_2\pi_2 - \mu_3\pi_3 - \mu_4\pi_4 - \mu_5\pi_5 - \mu_6\pi_6$$

$$\begin{aligned}
&\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \leq \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \pi_6 \end{bmatrix} \\
&\mathbf{1}\pi = 1, \qquad \pi_i \geq 0
\end{aligned}$$

The matrix of coefficients, including slacks y_r, looks like this:

$$\left[\begin{array}{cccccccc|cccccc}
\lambda_1 & \lambda_2 & \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 & \mu_6 & & & & & & \\
h_1 & h_2 & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 & \pi_6 & y_1 & y_2 & y_3 & y_4 & y_5 & r.h.s. \\
-1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & C
\end{array}\right]$$

In principle we may have as many as $\binom{13}{6=1716}$ basic solutions. We are interested in feasible bases, i.e. basic solutions which are non-negative. By putting C on the r.h.s. of the last constraint, we can make all solution be integer. We are interested in cut of flow, so each basic solution must involve at least on λ_c i.e. have at least one of the $h_c > 0$. Also we need at least one π_i positive by the last constraint. in fact, to be a cut for λ_1 we need at east one $\pi_i > 0, i = 1, 2, 3$ and one of $\pi_i > 0.i = 4, 5, 6$. For λ_2 we need at east one $\pi_i > 0, i = 1, 4$ and one of $\pi_i > 0.i = 2, 5$ and one of $\pi_i > 0.i = 3, 6$.

The 29 generalized cuts are.

(a) Cuts for λ_1, no cost for λ_2: $\lambda_1 < \mu_1 + \mu_4$ or any $\lambda_2 < \mu_k + \mu_l$, $k \in 1, 2, 3$, $l \in 4, 5, 6$. In these basic solutions the slacks $y_3 = y_4 = y_5 = 1$. If this cut is the only tight cut, customers of type $c = 2$ have routes that are not moving through bottlenecks. Value of $C = 2$. Total of 9.

(b) Cuts for λ_2 no cost for λ_1: $\lambda_2 < \mu_1 + \mu_2 + \mu_3$ or $\lambda_2 < \mu_4 + \mu_5 + \mu_6$. In these basic solutions the slacks $y_1 = y_2 = 1$. If this cut is the only tight cut, customers of type $c = 1$ have routes that are not moving through bottlenecks. Value of $C = 3$. Total of 2.

(c) Cuts for λ_1, λ_2, involving just three positive π_i, like: $\lambda_1+\lambda_2 < \mu_1+\mu_2+\mu_6$, or any $\lambda_1+\lambda_2 < \mu_k+\mu_l+\mu_m$, where $m \in (1, 2, 3, 4, 5, 6)$, and conditional on m, $k < l$ are in another row, in the other two columns. In these basic solutions one of the slacks y_1 or y_2 equals 1, in fact the one of the row that has k, l. If this cut is the only tight cut, customers of type $c = 1$ will not use the route through k, l, because it is too expensive. Value of $C = 3$. Total of 6.

(d) Cuts for λ_1, λ_2, involving just three positive π_i, like: $2\lambda_1 + \lambda_2 < \mu_1 + \mu_2 + 2\mu_6$, or any $\lambda_1 + \lambda_2 < \mu_k + \mu_l + \mu_m$, where $m \in (1, 2, 3, 4, 5, 6)$, and conditional on m, $k < l$ are in another row, in the other two columns. In these basic solutions the slack corresponding to node m i.e. y_3 for $m = 1, 4$, y_4 for $m = 2, 5$, y_5 for $m = 3, 6$ equals 1. If this cut is the only tight cut, customers of type $c = 2$ will not use the route through m, because it is too expensive. Value of $C = 4$. Total of 6.

(e) Cuts for λ_1, λ_2, involving foru positive π_i, and no slacks, like: $3\lambda_1+2\lambda_2 < 2\mu_1 + 2\mu_5 + \mu_3 + \mu_6$, or any $3\lambda_1 + 2\lambda_2 < 2\mu_k + 2\mu_l + \mu_m + \mu_n$, where

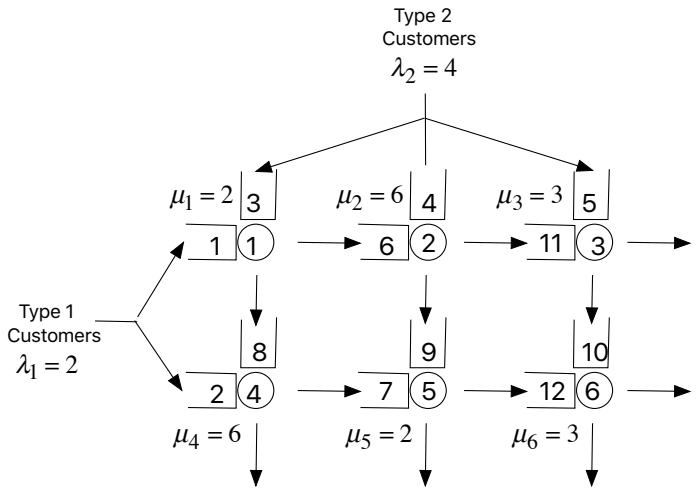
$m < n$ are one of $(1,4)$, $(2,5)$, $(3,6)$ and m, n, $k < l$ are located k in top row, l in bottom row, occupying the two columns left by m, n. There are no slacks. If one of these cuts is the only tight cut, all the nodes are heavily loaded, all the customer types need to use bottlenecks, and all the routes of each type have the same cost and will be used. Value of $C = 6$ Total of 6.

It is clear that these 29 cuts comprise all the possible cuts (i.e. some positive h_c, cuts for some λ_c, with 3, 2, 1, and 0 slacks.

17.17 Consider the six node network of Figure 17.6, with the following arrival and processing rates: $\lambda = (2,4)$, $\mu = (2,6,3,6,2,3)$. Locate the critical generalized cut constraint and formulate and solve the BCP. Obtain the distribution of the pooled workload of the solution.

Solution

The network is given by the following figure:



We can do one of three things: solve the LP(λ, μ) to obtain the optimal flows and η, or solve the dual LP$^*(\lambda, \mu)$, to obtain the tightest cut, or calculate the 29 cut constraints, get those that go intto Heavy traffic first.

We take the third option:

#	typ	λ	$l.h.s.$	μ	$r.h.s$	δ
1	a	1	2	1, 4	8	4
2	a	1	2	1, 5	4	2
3	a	1	2	1, 6	5	2.5
4	a	1	2	2, 4	12	6
5	a	1	2	2, 5	8	4
6	a	1	2	2, 6	9	4.5
7	a	1	2	3, 4	9	4.5
8	a	1	2	3, 5	5	2.5
9	a	1	2	3, 6	6	3
10	b	2	4	1, 2, 3	11	2.75
11	b	2	4	4, 5, 6	11	2.75
12	c	1, 2	6	5, 6, 1	7	1.167
13	c	1, 2	6	4, 6, 2	15	2.5
14	c	1, 2	6	4, 5, 3	11	1.833
15	c	1, 2	6	2, 3, 4	15	2.5
16	c	1, 2	6	1, 3, 5	7	1.167
17	c	1, 2	6	1, 2, 6	11	1.833
18	d	1, 2	8	5, 6, 1	9	1.125
19	d	1, 2	8	4, 6, 2	21	2.625
20	d	1, 2	8	4, 5, 3	14	1.75
21	d	1, 2	8	2, 3, 4	21	2.625
22	d	1, 2	8	1, 3, 5	9	1.125
23	d	1, 2	8	1, 2, 6	14	1.75
24	e	1, 2	14	1, 5, 3, 6	14	1
25	e	1, 2	14	4, 2, 3, 6	30	2.143
26	e	1, 2	14	1, 6, 2, 5	18	1.286
27	e	1, 2	14	4, 3, 2, 5	26	1.857
28	e	1, 2	14	2, 6, 1, 4	26	1.857
29	e	1, 2	14	5, 3, 1, 4	18	1.286

We obtain that the tightest generalized cut is

$$3\lambda_1 + 2\lambda_2 \leq 2\mu_1 + 2\mu_5 + \mu_3 + \mu_6$$

and in fact with the values $\lambda = (2, 4), \mu = (2, 6, 3, 6, 2, 3)$ this cut becomes critical, and the system becomes unstable. We assume then that the input rates are $\lambda = (2\rho, 4\rho)$ with $\rho < 1, \approx 1$.
For this cut we have: $h_1^* = 3, h_2^* = 2, \pi^* = (2, 0, 1, 0, 2, 1)$ (not normalized)
Looking at the primal problem, we note first that stations 2,4 are not bottleneck, and we can discard those stations, and also discard buffers 2,4,6,8 which will be empty.
We now solve LP(λ, μ for the various flows. Given the dual solution we can

obtain these flows for the equations (note that $\eta = 0$):

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 2\rho \\ 2\rho \\ 3 \\ 3 \\ 2 \\ 4 \end{bmatrix} \quad \begin{matrix} \mu_1 \\ \mu_5 \\ \mu_3 \\ \mu_6 \\ \lambda_1 \\ \lambda_2 \end{matrix}$$

(note that $\eta = 0$) This is solved uniquely by the flows: $f_1 = f_2 = \rho,\ f_3 = f_4 = 1,\ f_5 = 2$. Recall, these are the nominal flows.

These flows are the nominal rates. At the various buffers, $k = 1, \ldots, 12$ we then have the free time controls $\mathcal{J}_k(t) = \nu_k t - \mathcal{T}_k(t)$, and for the routing controls we have $\mathcal{V}_k(t) = \mathcal{A}_k(t) - f_k t$. The Brownian control problem is formulated for the scaled queues, $\hat{Q}_k(t) = \frac{1}{N} Q_k(N^2 t)$, as (17.3), where:

$$R = \begin{bmatrix} & 1 & 3 & 5 & 7 & 9 & 10 & 11 & 12 \\ 1 & \mu_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & \mu_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & \mu_3 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & \mu_5 & 0 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & \mu_5 & 0 & 0 & 0 \\ 10 & 0 & 0 & -\mu_3 & 0 & 0 & \mu_6 & 0 & 0 \\ 11 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_3 & 0 \\ 12 & 0 & 0 & 0 & -\mu_5 & 0 & 0 & 0 & \mu_6 \end{bmatrix}$$

For the various buffers we have $m_{i,k}$ as follows:

$$R = \begin{bmatrix} i\backslash k & 1 & 3 & 5 & 7 & 9 & 10 & 11 & 12 \\ 1 & 1 & 1 & & & & & & \\ 3 & 1 & & 1 & & & & 1 & \\ 5 & & & & 1 & 1 & & & \\ 6 & & & 1 & 1 & & 1 & & 1 \end{bmatrix}$$

and we obtain $\bar{m}_k$ as $\pi^{\mathsf{T}}[m_{i,k}]_{4\times 8}$ (recall, $\pi_1 = \pi_5 = 2$, $\pi_3 = \pi_6 = 1$, so $\pi = (2, 1, 2, 1)$which needs to be normalized):

$$\bar{m} = (3,\ 2,\ 2,\ 3,\ 2,\ 1,\ 2,\ 1)/6$$

We already see that $K^* = (1, 7)$ so we should hold all the inventory in those two buffers.

The pooled workload is:

$$\hat{\mathcal{W}}_P(t) = \frac{1}{2}\hat{Q}_1(t) + \frac{1}{3}\hat{Q}_3(t) + \frac{1}{3}\hat{Q}_5(t) + \frac{1}{2}\hat{Q}_7(t) + \frac{1}{3}\hat{Q}_9(t) + \frac{1}{6}\hat{Q}_{10}(t) + \frac{1}{3}\hat{Q}_{11}(t) + \frac{1}{6}\hat{Q}_{12}(t).$$

and we have:

$$\hat{\mathcal{W}}_P(t) = \hat{\mathcal{Z}}_P(t) + \hat{\mathcal{I}}_P(t),$$

where $\hat{\mathcal{Z}}_P(t)$ is a Brownian motion with drift:

$$\begin{aligned}\theta &= N\big((3\lambda_1 + 2\lambda_2) - (2\mu_1 + 2\mu_5 + \mu_3 + \mu_6)\big)\\ &= N\big((3 \cdot 2 \cdot \rho) + 2 \cdot 4 \cdot \rho) - (2 \cdot 2 + 1 \cdot 3 + 2 \cdot 2 + 1 \cdot 3)\big)\\ &= N(\rho - 1) \cdot 14\end{aligned}$$

If we take $N = 1000$ and $\rho = 0.999$ we get drift $\theta = -14$, so it will take about $1000/14 \approx 70$ time units to drift from 1000 to 0.
The variance of $\hat{\mathcal{Z}}_P(t)$ is

$$\begin{aligned}\sigma^2 &= 9\lambda_1 c_{a,1}^2 + 4\lambda_2 c_{a,2}^2 + 4\mu_1 c_{s,1}^2 + \mu_3 c_{s,3}^2 + 4\mu_5 c_{s,5}^2 + \mu_6 c_{s,6}^2\\ &= 22 \quad \text{for Poisson exponential with } c^2 = 1.\end{aligned}$$

We will not calculate the exact values of $\hat{\mathcal{J}}_k(t)$, $\hat{\mathcal{V}}(t)$ for each realization of $\hat{\boldsymbol{Q}}(t)$, $hX(t)$ that need to be solved at each t. However, for the Brownian system our policy is: Work fully at all stations if there is fluid in the system, keep all fluid in buffers 1 and 7, make sure that they empty together, and idle all the machines whenever the system is empty.

Part VI

Many-Server Systems

18

Infinite Servers Revisited

Exercises

18.1 Calculate the auto-covariance function of the Brownian bridge $BM(x) - xBM(1)$, $0 \le x \le 1$, and conversely, show that the unique Gaussian process with continuous paths and auto-covariance $x \wedge y - xy$ is $BM(x) - xBM(1)$, $0 \le x \le 1$.

Solution

The auto covariance is:

$$\begin{aligned}
&\mathbb{E}\Big(\big(BM(t) - tBM(1)\big)\big(BM(s) - sBM(1)\big)\Big) \\
&= \mathbb{E}\big(BM(t)BM(s)\big) s\mathbb{E}\big(BM(t)BM(1)\big) - t\mathbb{E}\big(BM(s)BM(1)\big) \\
&\quad + st\mathbb{E}\big(BM(1)BM(1)\big) \\
&= s \wedge t - st - st + st = s \wedge t - st.
\end{aligned}$$

Clearly, $BM(t)$ is a Gaussian process, i.e. all its joint distributions are multivariate Gaussian, and so then is $BM(t)tBM(1)$, and a Brownian process is determined by its mean and auto-covariance function. Hence the only process which has mean 0 and auto-covariance function $s \wedge t - st$ is indeed $BM(t) - tBM(1)$.

The true question is to show that $BM(t) - tBM(1)$ is the process $Z(t) = \big(BM(t)\big|BM(1) = 0\big)$. We show that now.

For $t_1 < t_2$, we look at $BM(t_1), BM(t_2)|BM(1) = 0$:

$$\begin{aligned}
&\mathbb{P}\Big(BM(t_1) = x_1, BM(t_2) = x_2\Big|BM(1) = 0\Big) \\
&= \varphi(\frac{x_1}{\sqrt{t_1}})\varphi(\frac{x_2 - x_1}{\sqrt{t_2 - t_1}})\varphi(\frac{0 - x_2}{\sqrt{1 - t_2}})\Big/\varphi(0) \\
&= \frac{1}{2\pi\sqrt{t_1(t_2 - t_1)(1 - t_2)}} \exp\Big\{-\frac{t_2(1 - t_2)x_1^2 + t_1(1 - t_1)x_2^2 - 2t_1(1 - t_2)x_1x_2}{2t_1(t_2 - t_1)(1 - t_2)}\Big\} \\
&= \frac{1}{2\pi|\Sigma|} \exp\Big\{-(x_1, x_2)\,\Sigma^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\Big\}
\end{aligned}$$

which is the joint distribution of $(X, Y) \sim N(0, \Sigma)$ where $\Sigma = \begin{bmatrix} t_1(1-t_1) & t_1(1-t_2) \\ t_1(1-t_2) & t_2(1-t_2) \end{bmatrix}$

So the two Gaussian processes $BM(t) - tBM(1)$ and $Z(t) = \big(BM(t) \big| BM(1) = 0\big)$ have the same auto-covariance function, so they are the same.

18.2 For a Brownian bridge $\mathcal{B}(y)$ and a distribution function $F(x)$ of a non-negative random variable, derive the auto-covariance function of the process $\mathcal{B} \circ F(x)$ (calculated as $\mathcal{B}(F(x),\ x \geq 0)$.

Solution

Since for a random variable X with distribution F, that is strictly increasing from 0 to 1 in an interval $[a, b]$ $(a \geq -\infty, b \leq \infty)$, $F(X) \sim U(0, 1)$, we have that $\mathcal{B}(F(x)$ has the auto-covariance function for $x \leq y$ given by $F(x)(1 - F(y))$, for $a \leq x \leq y \leq b$

18.3 Specialize the results of Theorems 18.3 and 18.4 to the GI/GI/∞, with inter-arrival time distribution F service time distribution G, starting empty.

Solution

We consider a sequence of systems, where the arrival stream for system n is a renewal process $\mathcal{A}^n(t)$ with interarrival times u_j/n, where u_i are i.i.d. with distribution F and mean $1/\lambda$, and coefficient of variation c_a. Hence, by FSLLN $\frac{1}{n}A^n(t) \to \lambda t$ u.o.c. a.s., and by FCLT for renewal processes, $\sqrt{n}\Big(\frac{1}{n}A^n(t) - \lambda t\Big) \to_w \sqrt{\lambda c_a^2}BM(t)$.

Substituting into the expressions in Theorem 18.3, the fluid limit is:

$$\frac{1}{n}Q^n(t) \to_p \lambda \int_0^t \bar{G}(s)ds, \quad \text{u.o.c. a.s. as } n \to \infty.$$

Substituting into the expressions in Theorem 18.4, the diffusion approximation is, that

$$\hat{Q}^n(t) = \sqrt{n}\Big(\frac{Q^n(t)}{n} - \lambda \int_0^t \bar{G}(t-s)ds\Big).$$

converges weakly to $\hat{Q}(t)$,

$$\hat{Q}(t) = \sqrt{\lambda c_a^2} \int_0^t \bar{G}(t-s)dBM(s) - \int_0^t \int_0^t \mathbb{1}(s + x \leq t)d\mathcal{B}(\lambda s, G(x)),$$

Note that for this case, $\hat{Q}(t)$ is obtained as a linear transformation on gaussian processes, and so this is a Gaussian process. We have already derived the form of this Gaussian process for the G/G/∞ model with renewal arrivals and services in Section 6.9.

18.4 Specialize the results of Theorems18.3 and 18.4 to the G/G/∞, with interarrival time distribution F and i.i.d. service time distribution G, starting with initial number of customers $Q^n(0)/n \to q_0$, $\sqrt{n}(Q^n(0)/n - q_0) \to Q_0$ that have i.i.d. residual service time distribution G_{eq}.

Solution

For the fluid limit the only added term is: $q_0\bar{G}_{eq}(t)$.

For the diffusion limit we need to add two random terms: $\bar{G}_{eq}(t)Q_0$, and $\sqrt{q_0}\mathcal{B}^0(G_{eq}(t))$.

Remark: It is tempting to think that if Q_0 has the stationary distribution then this will be the stationary process for this G/G/∞ process. However this is wrong: the diffusion limit is not a diffusion, it is not Markovian, and so its stationary distribution is not determined by its state at a single point in time.

18.5 Specialize the results of Theorems 18.3 and 18.4 to the G/M/∞, with inter-arrival time distribution F and exponential service time, starting with initial number of customers with the same exponential service time.

Solution

The fluid limit is:

$$q_0 e^{-\mu t} + \frac{\lambda}{\mu}(1 - e^{-\mu t})$$

In particular, if $q_0 = \frac{\lambda}{\mu}$, then $\frac{1}{b}Q^n(t) \to \frac{\lambda}{\mu}$.

For the diffusion limit we have:

$$\hat{Q}(t) = e^{-\mu t}Q_0 + \sqrt{q_0}\mathcal{B}^0(1 - e^{-\mu t}) + \sqrt{\lambda c_a^2}\int_0^t e^{\mu(t-s)}dBM(s)$$
$$- \int_0^t \int_0^t \mathbb{1}(s + x \le t)d\mathcal{B}(\lambda s, 1 - e^{-\mu x})$$

One can then show that if $q_0 = \frac{\lambda}{\mu}$, then $\hat{Q}(t)$ satisfies the linear Ito equation:

$$\hat{Q}(t) = \hat{Q}(0) - \mu\int_0^t \hat{Q}(s)ds + \sqrt{\lambda(c_a^2 + 1)}BM(t),$$

that is, $\hat{Q}(t)$ is an Ornstein-Uhlenbeck process. This is given as Theorem 3, part II, of Krichagina and Puhalskii (1997), where the complete proof appears.

18.6 Specialize the results of Theorems 18.3 and 18.4 to the M/G/∞, with service time distribution G, starting with initial number of customers $Q(0)$ where $Q(0)$ is a Poisson random variable, and the initial customers have i.i.d. service time distribution G_{eq}.

Solution

The fluid limit is

$$\bar{Q}(t) = q_0\bar{G}_{eq}(t) + \lambda\int_0^t \bar{G}(s)ds$$

and the diffusion limit is

$$\hat{Q}(t) = \bar{G}_{eq}(t)Q_0 + \sqrt{q_0}\mathcal{B}^0(G_{eq}(t)) + \sqrt{\lambda}\int_0^t \bar{G}(t-s)dBM(s)$$
$$- \int_0^t \int_0^t \mathbb{1}(s + x \le t)d\mathcal{B}(\lambda s, G(x)),$$

There is no simplification beyond that obtained for the general G/G/∞ case, except in the third term we have $c_a^2 = 1$.

18.7 Specialize the results of Theorems 18.3 and 18.4 to the M/M/∞ stationary process,

Solution

The answer is exactly as for Exercise 18.5 with the only simplification that now $c_a^2 = 1$.

19

Asymptotics Under Halfin-Whitt Regime

Exercises

19.1 Assume $(1-\rho_n)\sqrt{n} \to \beta$, and let X_n be a Poisson random variable with parameter $n\rho_n$. Use CLT to show that $\mathbb{P}(X_n \le n-1) \to \Phi(\beta)$.
Solution
$\mathbb{P}(X_n \le n-1) = \mathbb{P}((X_n - n\rho_n)/\sqrt{n\rho_n} \le \nu_n)$ where $\nu_n = (1-\rho_n)\sqrt{n}\,\rho_n^{-1/2} - (n\rho_n)^{-1/2}$, but $\nu_n \to \beta$, so by CLT, $\mathbb{P}(X_n \le n-1) \to \Phi(\beta)$.

19.2 Assume $(1-\rho_n)\sqrt{n} \to \beta$, and let X_n be a Poisson random variable with parameter ρ_n. Use Stirling's approximation to $n!$ to show that $\mathbb{P}(X_n = n)/(1-\rho_n) \to \phi(\beta)/\beta$.
Solution

$$\begin{aligned}
\mathbb{P}(X_n = n)/(1-\rho_n) &= \frac{(n\rho_n)^n}{n!(1-\rho_n)} e^{-n\rho_n} \\
&\sim e^{n(1-\rho_n+\log\rho_n)}/(\sqrt{2\pi n}\,(1-\rho_n)) \\
&\sim e^{-n(1-\rho_n)^2)/2}/(\sqrt{2\pi n}\,(1-\rho_n)) \\
&\to \frac{1}{\beta\sqrt{2\pi}} e^{-\beta^2/2}
\end{aligned}$$

19.3 In an M/M/n system show that if $(1-\rho_n)\sqrt{n} \to 0$ then the probability of delay $\alpha_n \to 1$, and if $(1-\rho_n)\sqrt{n} \to \infty$ then $\alpha_n \to 0$.
Solution
The limits for γ_n, ξ_n are valid also for $\beta = 0$, and $\beta \to \infty$. Recall that the limit is $\alpha = \lim_{n\to\infty}[1 + \frac{\gamma_n}{\xi_n}]^{-1} = [1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}]^{-1}$. Clearly, $\frac{\beta\Phi(\beta)}{\phi(\beta)}$ converges to ∞ for $\beta \to \infty$ and is 0 for $\beta = 0$.

19.4 Show that α is monotone decreasing with β, for $0 < \beta < \infty$.
Solution
Note that $\Phi(\beta)$ is increasing, so $\beta\Phi(\beta)$ is increasing, and $\phi(\beta)$ is decreasing for $0 < \beta < \infty$, so $\frac{\beta\Phi(\beta)}{\phi(\beta)}$ is increasing, and hence $\left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}$ is decreasing for $0 < \beta < \infty$.

19.5 Verify equations (19.7), (19.8).
Solution

To prove (19.7), by CLT for Poisson random variables:

$$\begin{aligned}
\mathbb{P}(Q_n \le \delta_n \,|\, Q_n \le n) &= \sum_{k=0}^{\delta_n} \frac{(n\rho_n)^k}{k!} p_0 \Big/ \sum_{k=0}^{n} \frac{(n\rho_n)^k}{k!} p_0 \\
&= \sum_{k=0}^{\delta_n} \frac{(n\rho_n)^k}{k!} e^{-n\rho_n} \Big/ \sum_{k=0}^{n} \frac{(n\rho_n)^k}{k!} e^{-n\rho_n} \\
&= \mathbb{P}(X_n \le \delta_n) \Big/ \mathbb{P}(X_n \le n) \qquad \text{where } X_n \sim \text{Poisson}(n\rho_n) \\
&= \mathbb{P}\Big(\frac{X_n - n\rho_n}{\sqrt{n\rho_n}} \le \frac{\delta_n - n\rho_n}{\sqrt{n\rho_n}}\Big) \Big/ \mathbb{P}\Big(\frac{X_n - n}{\sqrt{n\rho_n}} \le \frac{n - n\rho_n}{\sqrt{n\rho_n}}\Big) \\
&= \mathbb{P}\Big(\frac{X_n - n\rho_n}{\sqrt{n\rho_n}} \le \frac{(n - n\rho_n) - (n - \delta_n)}{\sqrt{n\rho_n}}\Big) \Big/ \mathbb{P}\Big(\frac{X_n - n}{\sqrt{n\rho_n}} \le \frac{n - n\rho_n}{\sqrt{n\rho_n}}\Big) \\
&\to \Phi(\beta - \delta)/\Phi(\beta)
\end{aligned}$$

To prove (19.8):

$$\begin{aligned}
\mathbb{P}(Q_n(\infty) \ge \delta_n \,|\, Q_n(\infty) \ge n) &= \sum_{k=\delta_n}^{\infty} \frac{n^n \rho_n^k}{n!} p_0 \Big/ \sum_{k=n}^{\infty} \frac{n^n \rho_n^k}{n!} p_0 \\
&= \frac{(n\rho_n)^n}{n!(1-\rho_n)} \rho_n^{\delta_n - n} p_0 \Big/ \frac{(n\rho_n)^n}{n!(1-\rho_n)} p_0 \\
&= \rho_n^{\delta_n - n} \\
&= \exp\big((\delta_n - n)\log(1 - (1-\rho_n))\big) \\
&= \exp\big((\delta_n - n)(-(1-\rho_n))\big) \\
&\sim \exp\big(\frac{\delta_n - n}{\sqrt{n}}[-(1-\rho_n)\sqrt{n}]\big) \\
&\to e^{-\delta\beta}
\end{aligned}$$

19.6 Verify the infinitesimal mean and variance (drift and diffusion coefficients) for M/M/n.

Solution

We have

$$\begin{aligned}
m_{Q_n}(k) &= \lim_{h\to 0} \frac{1}{h} \mathbb{E}\big(Q_n(t+h) - Q_n(t) \,\big|\, Q_n(t) = k\big) \\
&= \lambda - \min(k, n)\mu = \lambda - \mu n + \mu(n-k)^+, \\
\sigma^2_{Q_n}(k) &= \lim_{h\to 0} \frac{1}{h} \mathbb{V}\text{ar}\big(Q_n(t+h) - Q_n(t) \,\big|\, Q_n(t) = k\big) \\
&= \lambda + \min(k, n)\mu = \lambda + \mu n - \mu(n-k)^+.
\end{aligned}$$

We now center and scale the queue length, so that $\hat{Q}_n(t) = \frac{Q_n(t)-n}{\sqrt{n}}$, so that if $\hat{Q}_n(t) = x$ then $Q_n(t) = \sqrt{n}x + n$. When $n \to \infty$ and $\sqrt{n}(1-\rho_n) \to \beta$ we

have:

$$\begin{aligned} m_{\hat{Q}_n}(x) &= \frac{1}{\sqrt{n}} m_{Q_n}(\sqrt{n}x + n) \\ &= \frac{1}{\sqrt{n}} \left(\lambda - n\mu + \mu(n - n - \sqrt{n}x)^+\right) \\ &= \frac{\lambda - n\mu}{\sqrt{n}} + \mu(-x)^+ \\ &\to \begin{cases} -\mu\beta, & x > 0, \\ -\mu(\beta + x), & x < 0. \end{cases} \end{aligned}$$

$$\begin{aligned} \sigma^2_{\hat{Q}_n}(x) &= \frac{1}{n} \sigma^2_{Q_n}(\sqrt{n}x + n) \\ &= \frac{1}{n} \left(\lambda + n\mu - \mu(n - n - \sqrt{n}x)^+\right) \\ &= \frac{\lambda - n\mu}{n} - \frac{\mu(-x)^+}{\sqrt{n}} \\ &\to 2\mu. \end{aligned}$$

19.7 Consider the embedded Markov chain at arrival times of the GI/M/n system. Derive the limit infinitesimal drift and diffusion under Halfin Whitt regime for the sequence of centered and scaled queue length at the embedded times.

Solution

Let $Q_n(M)$, $M = 0, 1, 2, \ldots$ be the embedded Markov chain just before arrival times.

$$\begin{aligned} \mathbb{E}(Q_n(M+1) - Q_n(M) \big| Q_n(M) = k) &= 1 - \frac{1}{\lambda}(\mu n - \mu(n-k)^+) \\ &= \frac{1}{\lambda}\left(\lambda - \mu n + \mu(n-k)^+\right) \end{aligned}$$

To calculate the variance we condition on u, the length of the period between the arrivals:

$$\begin{aligned} &\mathbb{V}\text{ar}(Q_n(M+1) - Q_n(M) \big| Q_n(M) = k) \\ &= \mathbb{E}\big(\mathbb{V}\text{ar}(Q_n(M+1) - Q_n(M) \big| Q_n(M) = k, u)\big) \\ &\quad + \mathbb{V}\text{ar}\big(\mathbb{E}(Q_n(M+1) - Q_n(M) \big| Q_n(M) = k, u)\big) \\ &= \mathbb{E}\big((u(\mu n - \mu(n-k)^+)\big) + \mathbb{V}\text{ar}\big(1 - u(\mu n - \mu(n-k)^+)\big) \\ &= \frac{1}{\lambda}(\mu n - \mu(n-k)^+) + \sigma^2\mu^2(n - (n-k)^+)^2 \\ &= \frac{1}{\lambda}(\mu n - \mu(n-k)^+) + c_a^2\left(\frac{\mu}{\lambda}\right)^2(n - (n-k)^+)^2 \end{aligned}$$

We are centering and scaling the queue, $\hat{Q}_n t = \frac{Q_n(t) - n}{\sqrt{n}}$, so that when $\hat{Q}_n t = x$

then $Q_n(t) = n + \sqrt{n}x$. We then get the limits:

$$m_{\hat{Q}}(x) = \frac{1}{\sqrt{n}}(\lambda - \mu n + \mu\sqrt{n}(-x)^+) \to \begin{cases} -\mu\beta, & x > 0, \\ -\mu(\beta + x), & x < 0. \end{cases}$$

$$\begin{aligned}\sigma^2_{\hat{Q}}(x) &= \lim_{n\to\infty}\Big(\frac{1}{\lambda}(\mu n - \mu\sqrt{n}(-x)^+) + c_a^2\big(\frac{\mu}{\lambda}\big)^2(n - \sqrt{n}(-x)^+)^2\Big) \\ &= \lim_{n\to\infty}\Big\{\frac{\mu n}{\lambda}\Big[1 - \Big(-\frac{x}{\sqrt{n}}\Big)^+\Big] + c_a^2\Big(\frac{\mu n}{\lambda}\Big)^2\Big[1 - \Big(-\frac{x}{\sqrt{n}}\Big)^+\Big]^2\Big\} \\ &= (1 + c_a^2)\mu\end{aligned}$$

We note that events in the embedded Markov chain occur at rate $\approx 2\lambda_n$ as n is large, independent of the state. So these are also the infinitesimal rates for the continuous time process. What we do is to define $\tilde{\hat{Q}}^n(t) = \hat{Q}^n([nt])$, then this continuous time process is a Markov chain, with the same infinite mean and variance, and will converge to the Halfin-Whitt limit. We then also get be time change that $\hat{Q}^n(t)$ converges to the same Halfin-Whitt limit.

19.8 (∗) For the Halfin-Whitt diffusion limit of the centered and scaled M/M/n process, in state $\hat{Q}(t) = x$, calculate the distribution of the times to return to 0 (no queue and all busy), from position $x > 0$ and $x < 0$ (hitting times).

Solution

We will only calculate the hitting time from $x > 0$. The calculation of hitting times for Ornstein-Uhlenbeck process is a hard problem. A recent paper on that question is Lipton and Kaushansky (2018), which gives theoretical derivations and numerical metods.

For $x > 0$, we use the results of Section 7.7.2. For a Brownian motion starting at 0, with drift m and diffusion σ^2, the first passage time distribution to reach level y by time t has the distribution $\mathbb{P}(T(y) > t) = \Phi\Big(\frac{y-mt}{\sigma t^{1/2}}\Big) - e^{2my/\sigma^2}\Phi\Big(\frac{-y-mt}{\sigma t^{1/2}}\Big)$.

For our case, we start at level x and want the hitting time of 0, where the drift is $-\mu\beta$ and the diffusion is $\sigma^2 = 2\mu$. We need to go from $x > 0$ to 0, which is the same as from 0 to x with drift $m = \mu\beta$. Hence:

$$\mathbb{P}(\text{return to 0 from } x > 0 > t) = \Phi\Big(\frac{x - \mu\beta t}{(2\mu t)^{1/2}}\Big) - e^{\beta x}\Phi\Big(\frac{-x - \mu\beta t}{(2\mu t)^{1/2}}\Big)$$

19.9 Solve the renewal type equation for the Example19.9.

Solution

The equation for the fluid limit is:

$$\bar{Q}(t) = \min(q_0, 1)\bar{G}_0(t) + (q_0 - 1)^+\bar{G}(t) + \int_0^t \bar{G}(t-s)d\bar{a}(s)$$
$$+ \int_0^t (\bar{Q}(t-s) - 1)^+ dG(s).$$

We assume: G, G_0 are deterministic 1, $q_0 = 1$, and $a(t) = t$, and we then have:
For $0 < t < 1$:

$$\bar{Q}(t) = 1 + 0 + t + 0 = 1 + t$$

where the last term is 0, since $G(s) = 0$ for $0 < s < 1$.
For $1 < t < 2$:

$$\bar{Q}(t) = 1 + 0 + 0 + (\bar{Q}(t-1) - 1) = \bar{Q}(t-1)$$

where in the last integral, $dG(s)$ has a single jump o 1 at $s = 1$, where the integrand is $(\bar{Q}(t-1) - 1)^+$.
For $m < t < m + 1$:

$$\bar{Q}(t) = 1 + 0 + 0 + (\bar{Q}(t-1) - 1) = \bar{Q}(t-1).$$

This is exactly the saw-tooth discontinuous periodic function.

20

Many Servers with Abandonment

Exercises

20.1 Use Figure 20.1 to obtain the limiting expected sojourn time for an arriving customer if he abandons, or if he waits and and is patient enough to receive service.

Solution

For customers that abandon, the mean sojourn time is:

$$\text{mean sojourn for abandon } = \int_0^w (1 - H(y))dy$$

This is the integral of Figure 20.1, from 0 to w, divided by ρ.

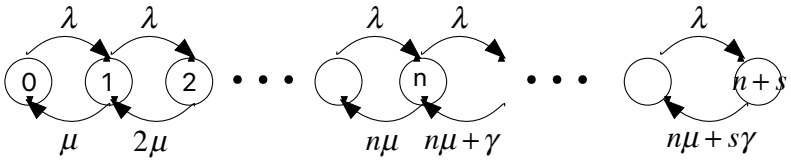
For customers that wait and get served, the sojourn time consists of waiting for w and service of 1,

$$\text{mean sojourn for served } = w + \int_0^\infty (1 - G(y))dy = w + 1.$$

It equals the area in Figure 20.1, under a level of 1.

20.2 For the M/M/n+M system, represent the queue length process as a birth and death process and derive its stationary distribution.

Solution



The stationary distribution is:

$$\pi_k = \begin{cases} \pi_0 \dfrac{1}{k!}\left(\dfrac{\lambda}{\mu}\right)^k, & 0 \le k \le n, \\ \pi_n \displaystyle\prod_{j=1}^{k-n} \frac{\lambda}{n\mu + j\gamma}, & k > n. \end{cases}$$

20.3 For the M/M/n+M system, assume that n is fixed, and that the value of μ increases according to the arrival rate λ, so that $\mu^{(\lambda)} = (\lambda+\beta\sqrt{\lambda})/n$. Calculate the infinitesimal mean and variance for the birth and death queue length process and obtain its limits as $\lambda \to \infty$. Obtain the diffusion approximation for the scaled queue length [Ward (2012)].

Solution

We have

$$\begin{aligned} m_{Q^n}(k) &= \lim_{h\to 0}\frac{1}{h}\mathbb{E}\big(Q^n(t+h) - Q_n(t) \,\big|\, Q_n(t) = k\big) \\ &= \lambda - \min(k,n)\mu - \gamma(k-n)^+ \\ &= \lambda - \mu n + \mu(n-k)^+ - \gamma(k-n)^+, \\ \sigma^2_{Q^n}(k) &= \lim_{h\to 0}\frac{1}{h}\mathbb{V}\mathrm{ar}\big(Q^n(t+h) - Q_n(t) \,\big|\, Q_n(t) = k\big) \\ &= \lambda + \min(k,n)\mu + \gamma(k-n)^+ \\ &= \lambda + \mu n - \mu(n-k)^+ + \gamma(k-n)^+. \end{aligned}$$

We now scale the queue length. Since n is fixed, and λ as well as μ tend to infinity, the proper scaling is $\hat{Q}^n(t) = \frac{Q^n(t)}{\sqrt{\lambda}}$, for which the infinitesimal means and variances are:

$$\begin{aligned} m_{\hat{Q}^n}(x) &= \frac{1}{\sqrt{\lambda}} m_{Q^n}(\sqrt{\lambda}x) \\ &= \frac{1}{\sqrt{\lambda}}\left(\lambda - n\mu + \mu(n - \sqrt{\lambda}x)^+ - \gamma(\sqrt{\lambda}x - n)^+\right) \\ &= \frac{1}{\sqrt{\lambda}}\left(\lambda - (\lambda+\beta\sqrt{\lambda}) + \frac{\lambda+\beta\sqrt{\lambda}}{n}(n - \sqrt{\lambda}x)^+ - \gamma(\sqrt{\lambda}x - n)^+\right) \\ &\to \begin{cases} -\beta - \gamma x, & x > 0, \\ \infty, & x < 0. \end{cases} \qquad \text{as } \lambda \to \infty, \end{aligned}$$

$$\begin{aligned} \sigma^2_{\hat{Q}^n}(x) &= \frac{1}{\lambda}\sigma^2_{Q^n}(\sqrt{\lambda}x) \\ &= \frac{1}{\lambda}\left(\lambda + n\mu - \mu(n - \sqrt{\lambda}x)^+ + \gamma(\sqrt{\lambda}x - n)^+\right) \\ &= \frac{1}{\lambda}\left(2\lambda + \beta\sqrt{\lambda} - \mu(n - \sqrt{\lambda}x)^+ + \gamma(\sqrt{\lambda}x - n)^+\right) \\ &= \to 2, \qquad \text{as } \lambda \to \infty. \end{aligned}$$

What we see here is that the limiting process behaves as an Ornstein-Uhlenbeck process when positive, and is reflected at zero.

20.4 For the M/M/n+M system, assume that both n and μ increase according to the arrival rate λ, so that $\mu^{(\lambda)} = \mu(\lambda^{1-\alpha} + \beta\lambda^{\frac{1}{2}-\alpha})$, and $n^{(\lambda)} = \lambda^\alpha/\mu$ for $0 \le \alpha \le 1$. Show that the system goes into heavy traffic with $\rho_\lambda \to 1$ as $\lambda \to \infty$. Calculate the infinitesimal mean and variance for the birth and death

queue length process and obtain its limits as $\lambda \to \infty$, and obtain the diffusion approximation for the scaled queue length [Ward (2012)].

Solution

The arrival rate is λ. The service rate is

$$\mu^{(\lambda)} n^{(\lambda)} = \mu(\lambda^{1-\alpha} + \beta\lambda^{\frac{1}{2}-\alpha})\lambda^{\alpha}/\mu = \lambda + \beta\sqrt{\lambda},$$

so indeed, $\rho_\lambda = \frac{\lambda}{\mu^{(\lambda)} n^{(\lambda)}} \to 1$ as $\lambda \to \infty$.

As before

$$\begin{aligned}
m_{Q^n}(k) &= \lim_{h\to 0} \frac{1}{h}\mathbb{E}\big(Q^n(t+h) - Q_n(t) \,\big|\, Q_n(t) = k\big) \\
&= \lambda - \min(k,n)\mu - \gamma(k-n)^+ \\
&= \lambda - \mu n + \mu(n-k)^+ - \gamma(k-n)^+, \\
\sigma^2_{Q^n}(k) &= \lim_{h\to 0} \frac{1}{h}\mathbb{V}\mathrm{ar}\big(Q^n(t+h) - Q_n(t) \,\big|\, Q_n(t) = k\big) \\
&= \lambda + \min(k,n)\mu + \gamma(k-n)^+ \\
&= \lambda + \mu n - \mu(n-k)^+ + \gamma(k-n)^+.
\end{aligned}$$

We scale the queue length process, defining $\hat{Q}^n(t) = \frac{Q^n(t) - n^{(\lambda)}}{\sqrt{\lambda}}$, so that $(\hat{Q}^n(t))^+$ is the scaled number of customers that are waiting, and $(-\hat{Q}^n(t))^+$ is the scaled number of idle servers. With $\hat{Q}^n(t) = x$, we have $Q(t) = \sqrt{\lambda}x + n^{(\lambda)}$. We get:

$$\begin{aligned}
m_{\hat{Q}^n}(x) &= \frac{1}{\sqrt{\lambda}} m_{Q^n}(\sqrt{\lambda}x + n^{(\lambda)}) \\
&= \frac{1}{\sqrt{\lambda}}\left(\lambda - n^{(\lambda)}\mu^{(\lambda)} + \mu^{(\lambda)}(-\sqrt{\lambda}x)^+ - \gamma(\sqrt{\lambda}x)^+\right) \\
&= \frac{1}{\sqrt{\lambda}}\left(\lambda - (\lambda + \beta\sqrt{\lambda}) + \mu(\lambda^{1-\alpha} + \beta\lambda^{\frac{1}{2}-\alpha})(-\sqrt{\lambda}x)^+ - \gamma(\sqrt{\lambda}x)^+\right) \\
&\to \begin{cases} -\beta - \gamma x, & x > 0, \\ \infty, & x < 0. \end{cases} \qquad \text{as } \lambda \to \infty,
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{\hat{Q}^n}(x) &= \frac{1}{\lambda}\sigma^2_{Q^n}(\sqrt{\lambda}x + n^{(\lambda)}) \\
&= \frac{1}{\lambda}\left(\lambda + n^{(\lambda)}\mu^{(\lambda)} - \mu^{(\lambda)}(-\sqrt{\lambda}x)^+ + \gamma(\sqrt{\lambda}x)^+\right) \\
&= \frac{1}{\lambda}\left(\lambda + (\lambda + \beta\sqrt{\lambda}) - \mu(\lambda^{1-\alpha} + \beta\lambda^{\frac{1}{2}-\alpha})(-\sqrt{\lambda}x)^+ + \gamma(\sqrt{\lambda}x)^+\right) \\
&\to 2 \text{ for } x > 0.
\end{aligned}$$

which again is a reflected Ornstein-Uhlenbeck process.

20.5 For the M/M/n+M system in Halfin-Whitt regime, derive the asymptotic distribution of the number waiting in queue when positive, the number of

idle servers when positive, and the probability of waiting, as given in equation (20.1) [Browne et al. (1995)].

Solution

We found that $\hat{Q}(t)$ is the diffusion process defined by,

$$d\hat{Q}(t) = m(\hat{Q}(t))dt + \sigma(\hat{Q}(t))dBM(t)$$

where the drift and diffusion coefficient are:

$$m(x) = \begin{cases} \beta - \gamma x, & x > 0, \\ \beta - \mu x, & x \le 0. \end{cases} \qquad \sigma^2(x) = 2.$$

This is a piecewise Ornstein Uhlenbeck (OU) process, which for $x > 0$ is centered at β/γ with drift down at rate γx, and for $x < 0$ it also is centered at β/μ and drifts towards it at rate μx. The OU process is time reversible and the restriction of its state space to an interval has the same stationary distribution as the OU process over the whole real line, renormalized.

The OU process has drift $m(x) = -a(x - m)$ and diffusion parameter $\sigma^2(x) = \sigma^2$, and its stationary distribution is $\sim N(m, \sigma^2/2a)$, with density $\sqrt{\frac{2a}{\sigma^2}}\phi\Big(\sqrt{\frac{2a}{\sigma^2}}(x-m)\Big)$.

So, for $m(x) = \beta - \gamma x$, $\sigma^2 = 2$ we get that the OU process has density $\sqrt{\gamma}\phi\Big(\sqrt{\gamma}(x - \frac{\beta}{\gamma})\Big)$

$$\mathbb{P}\Big(\frac{(Q(t)-n)^+}{\sqrt{\lambda}} \in (x, x+dx) \,\Big|\, \hat{Q}(t) > 0\Big) = \sqrt{\gamma}\frac{\phi\Big(\sqrt{\gamma}\Big(x - \frac{\beta}{\gamma}\Big)\Big)}{\Phi\Big(\frac{\beta}{\sqrt{\gamma}}\Big)}dx, \quad x > 0,$$

and for $m(x) = \beta - \mu x$, $\sigma^2 = 2$ we get that the OU process has density $\sqrt{\mu}\phi\Big(\sqrt{\mu}(x - \frac{\beta}{\mu})\Big)$, and we are interested in $-\hat{Q}(x)$, $x < 0$, so

$$\mathbb{P}\Big(\frac{(n-Q(t))^+}{\sqrt{\lambda}} \in (x, x+dx) \,\Big|\, \hat{Q}(t) < 0\Big) = \sqrt{\mu}\frac{\phi\Big(\sqrt{\mu}\Big(x + \frac{\beta}{\mu}\Big)\Big)}{\Phi\Big(-\frac{\beta}{\sqrt{\mu}}\Big)}dx, \quad x > 0.$$

We now want to obtain $\alpha > 0$, the probability of waiting, i.e. $\hat{Q}(t) > 0$). We wish to have that $\mathbb{P}(\hat{Q}(t) \le x)$ be continuous at 0, so we need to have:

$$\alpha\sqrt{\gamma}\frac{\phi\Big(\sqrt{\gamma}\Big(-\frac{\beta}{\gamma}\Big)\Big)}{\Phi\Big(\frac{\beta}{\sqrt{\gamma}}\Big)} = (1-\alpha)\sqrt{\mu}\frac{\phi\Big(\sqrt{\mu}\Big(\frac{\beta}{\mu}\Big)\Big)}{\Phi\Big(-\frac{\beta}{\sqrt{\mu}}\Big)}$$

$$\alpha = \mathbb{P}(\text{waiting}) = \mathbb{P}(\hat{Q}(t) > 0) = \left(1 + \sqrt{\frac{\gamma}{\mu}}\frac{\phi\Big(-\frac{\beta}{\sqrt{\gamma}}\Big)}{\Phi\Big(\frac{\beta}{\sqrt{\gamma}}\Big)}\frac{\Phi\Big(-\frac{\beta}{\sqrt{\mu}}\Big)}{\phi\Big(\frac{\beta}{\sqrt{\mu}}\Big)}\right)^{-1}.$$

20.6 For the M/M/n+G system, explain why $(N(t), V(t))$ is a Markov process [Baccelli and Hebuterne (1981)].

Solution

Clearly, while not all the servers are busy, the system behaves like an M/M/n system, with arrival rate λ and departure rate $N(t)\mu$. When all the servers are busy and $V(t) = x > 0$ then $V(t)$ decreases at rate 1 until either an arrival occurs, or $V(t)$ reaches 0. If an arrival occurs then all servers remain busy, and the virtual waiting time $V(t)$ is unchanged if the new arrival does not join, which has probability $G(Vt))$, or it is increased with probability $1-G(V(t))$, and the increase is by a quantity which is exponentially distributed with rate $n\mu$.

20.7 For the M/M/n+G system, explain the Kolmogorov transition equations (20.2) for the process $(N(t), V(t))$ when $N(t) = n$, $V(t) = x > 0$ [Baccelli and Hebuterne (1981)].

Solution

The states $N(t+\delta) = n$, and $V(t+\delta) > x > 0$ will be reached from $N(t) = n$ and $V(t) > x + \delta$, if no arrivals occur in $(t, t+\delta)$. If an arrival occurs, which has probability $\lambda\delta$, then this state can be reached from one of the following states: from $(N(t) = n, V(t) > x)$ with certainty, or from $N(t) = n-1$ if the arrival has processing time $> \frac{x}{n\mu}$, which has probability $e^{-n\mu x}$, or from $N(t) = n$ and $V(t) = u$ where $0 \leq u \leq x$ and the arrival has patience exceeding u which has probability $1 - H(u)$, and has processing time $> \frac{x-u}{n\mu}$ [Baccelli and Hebuterne (1981)].

20.8 For the M/M/n+G system, show that for the stationary system, p_j and $v(x)$ satisfy the equations (20.5) [Baccelli and Hebuterne (1981)].

Solution

The expressions for $p_0, \ldots, p_{n-1}$ are those for the standard M/M/n system. In the equation $(\lambda + (n-1)\mu)p_{n-1} = \lambda p_{n-2} + v(0)$, the terms $(n-1)\mu p_{n-1}$ and λp_{n-2} cancel to give: $v(0) = \lambda p_{n-1}$.

Equation (20.2) leads to the differential-integral equation:

$$
\begin{aligned}
&\mathbb{P}(N(t+\delta) = n, V(t+\delta) > x) - \mathbb{P}(N(t) = n, V(t) > x + \delta)) \Big/ \delta \\
&= \lambda \Big[- \mathbb{P}(N(t) = n, V(t) > x + \delta)) + \mathbb{P}(N(t) = n, V(t) > x)) \\
&+ \mathbb{P}(N(t) = n-1)\ e^{-n\mu x} \\
&+ \int_0^x \mathbb{P}(N(t) = n, V(t) = u)\,(1 - H(u))\, e^{-n\mu(x-u)}\, du \Big] + o(\delta)/\delta,
\end{aligned}
$$

Which is in fact

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)\mathbb{P}(N(t)=n,V(t)>x)$$
$$=\lambda\Big[\mathbb{P}(N(t)=n-1)\ e^{-n\mu x}$$
$$+\int_0^x \mathbb{P}(N(t)=n,V(t)=u)\,(1-H(u))\,e^{-n\mu(x-u)}du\Big].$$

When $t\to\infty$, i.e. for steady state, $\frac{\partial}{\partial t}=0$, and we have:

$$v(x)=\lambda\big[p_{n-1}e^{-n\mu x}+\int_0^x v(u)(1-H(u))\,e^{-n\mu(x-u)}du\big].$$

and multiplying by $e^{n\mu x}$ we have:

$$v(x)e^{n\mu x}=\lambda p_{n-1}+\lambda\int_0^x v(u)e^{n\mu u}(1-H(u))du$$

20.9 For the M/M/n+G system, verify the solution of the integral equation (20.6) for $v(x)$ [Baccelli and Hebuterne (1981)].

Solution

Let $f(x)=v(x)e^{n\mu x}$. the equation to be solved is:

$$f(x)=\lambda p_{n-1}+\lambda\int_0^x f(u)(1-H(u))du$$

taking derivatives we get:

$$f'(x)=\lambda f(x)(1-H(x))$$

so

$$\frac{f'(x)}{f(x)}=\lambda(1-H(x))$$

solved by:

$$\log(f(x))=c_1+\int_0^x \lambda(1-H(u))du$$

so

$$f(x)=c_2e^{\lambda\int_0^x(1-H(u))du}$$

and substituting $x=0$ we find: $c_2=\lambda p_{n-1}$, and so

$$f(x)=\lambda p_{n-1}e^{\lambda\int_0^x(1-H(u))du}$$

and we have:

$$v(x)=\lambda p_{n-1}e^{\lambda\int_0^x(1-H(u))du-n\mu x}$$

20.10 For the M/M/n+G system, verify Theorem 20.3 [Zeltyn and Mandelbaum (2005)].

Solution

(i) The probability of needing to wait is p_n. We saw that $p_n = \lambda p_{n-1} J$. By definition, $p_{n-1}(\mathcal{E}\lambda J) = 1$, and (i) follows.

(ii) Start from equation (20.5) and integrate both sides from 0 to ∞ to get:

$$\begin{aligned}
p_n &= \int_0^\infty v(x)dx = \int_0^\infty \lambda p_{n-1} e^{-n\mu x} dx + \int_0^\infty \lambda \int_0^x v(u)(1-H(u))e^{-n\mu(x-u)} du\, dx \\
&= \frac{\lambda}{n\mu} p_{n-1} + \int_0^\infty \int_u^\infty \lambda v(u)(1-H(u))e^{-n\mu(x-u)} dx\, du \\
&= \frac{\lambda}{n\mu} p_{n-1} + \int_0^\infty \lambda v(u)(1-H(u)) \int_u^\infty e^{-n\mu(x-u)} dx\, du \\
&= \frac{\lambda}{n\mu} p_{n-1} + \frac{\lambda}{n\mu} \int_0^\infty v(u)(1-H(u))du \\
&= \frac{\lambda}{n\mu} p_{n-1} + \frac{\lambda}{n\mu} p_n - \frac{\lambda}{n\mu} \int_0^\infty v(u)H(u)du.
\end{aligned}$$

and we obtain:

$$\mathbb{P}(ab) = \int_0^\infty v(u)H(u)du = (1 - \frac{n\mu}{\lambda})p_n + p_{n-1}$$

and (ii) follows.

(iii) follows immediately from

$$\mathbb{E}(V(t)|V(t) > 0) = \frac{\int_0^\infty xv(x)dx}{\mathbb{P}(V(t) > 0} = \frac{\int_0^\infty xv(x)dx}{p_n} = \frac{\int_0^\infty xv(x)dx}{\int_0^\infty v(x)dx} = \frac{\lambda p_{n-1} J_1}{\lambda p_{n-1} J}.$$

20.11 For the M/M/n+G system, verify the asymptotics for QD, ED and QED [Zeltyn and Mandelbaum (2005)].

Solution

The calculations of the explicit expressions for the probability of waiting, the average waiting time, and the probability of abandonment, are quite complex. The interested reader is directed to the internet supplement to Zeltyn and Mandelbaum (2005) and to its internet supplement.

21

Load Balancing in the Supermarket Model

Exercises

21.1 Consider the n server system with Poisson arrivals and exponential service time, and assume that customers are dispatched to the servers in round robin order. Show that in heavy traffic, the expected waiting time of each customer approaches $\frac{1}{2(1-\rho)}$ as $n \to \infty$, i.e. half of the time under random dispatching.

Solution

Under round robin dispatching the interarrival times of each server will by $\sim$ Erlang$(n, \lambda n)$, with mean λ. But then $c_a^2 = \frac{1}{n} \to 0$ as $n \to \infty$ and by Kingman's bound the expected waiting time approaches $\frac{1+c_a^2}{2(1-\rho)} \to \frac{1}{2(1-\rho)}$.

21.2 $(*)$ Show that the stationary queue length for choose shortest of d, ${}^dQ^n$ is smaller in the sense of convex majorization than the stationary queue length under random dispatching ${}^1Q^n$, and prove that the choose shortest of d is a stable ergodic system [Vvedenskaya et al. (1996)].

Solution

We will show convex majorization, i.e.:

$$\mathbb{E}\big(({}^1Q^n(t) - k)^+\big) \geq \mathbb{E}\big(({}^dQ^n(t) - k)^+\big).$$

Clearly, if this holds then the ergodicity of choose shortest of d follows from the ergodicity of random assignment, when $\lambda < 1$.

We note that:

$$\begin{aligned}
&\mathbb{P}({}^dQ^n(t) = k) = \mathbb{E}({}^d\bar{S}_k^n(t) - {}^d\bar{S}_{k+1}^n(t)),\\
&\mathbb{P}({}^dQ^n(t) \geq k) = \mathbb{E}({}^d\bar{S}_k^n(t)),\\
&\mathbb{E}\big(({}^dQ^n(t) - k)^+\big) = \sum_{j=k}^{\infty} \mathbb{P}({}^dQ^n(t) \geq j) = \sum_{j=k}^{\infty} \mathbb{E}({}^d\bar{S}_j^n(t)).
\end{aligned}$$

Similarly,

$$\mathbb{E}\big(({}^1Q^n(t) - k)^+\big) = \sum_{j=k}^{\infty} \mathbb{E}({}^1\bar{S}_j^n(t)).$$

So, what we wish to show is that:

$$M_k^{(1)}(t) := \sum_{j=k}^{\infty} \mathbb{E}({}^1\bar{S}_j^n(t)) \geq \sum_{j=k}^{\infty} \mathbb{E}({}^d\bar{S}_j^n(t)) := M_k^{(d)}(t).$$

The proof of this is quite technical. We recall the equations:

$$\begin{aligned}
&\mathbb{E}({}^d\bar{S}_k^n(t+h) - {}^d\bar{S}_k^n(t) \big| {}^d\bar{S}(t)) \\
&= \lambda h({}^d\bar{S}_{k-1}^n(t)^d - {}^d\bar{S}_k^n(t)^d) - h({}^d\bar{S}_k^n(t) - {}^d\bar{S}_{k+1}^n(t)) + o(h), \quad {}^d\bar{S}_0^n(t) = 1,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}({}^1\bar{S}_k^n(t+h) - {}^1\bar{S}_k^n(t) \big| {}^1\bar{S}(t)) \\
&= \lambda h({}^1\bar{S}_{k-1}^n(t) - {}^1\bar{S}_k^n(t)) - h({}^1\bar{S}_k^n(t) - {}^d\bar{S}_{k+1}^n(t)) + o(h), \quad {}^1\bar{S}_0^n(t) = 1.
\end{aligned}$$

(These are the equations for the generator of the Markov chains $({}^d\bar{S}^n(t), ({}^1\bar{S}^n(t).$)
Adding up the first one from k:

$$\begin{aligned}
&\sum_{j=k}^{\infty} \mathbb{E}({}^d\bar{S}_j^n(t+h) - {}^d\bar{S}_j^n(t) \big| {}^d\bar{S}(t)) \\
&= \lambda h {}^d\bar{S}_{k-1}^n(t)^d - h {}^d\bar{S}_k^n(t) + o(h) \\
&\leq \lambda h {}^d\bar{S}_{k-1}^n(t)^d - h {}^d\bar{S}_k^n(t) + o(h),
\end{aligned}$$

where the inequality follows from $S_{k-1}^n(t) \leq 1$ $(< 1,\ k > 1)$.
Taking expectations, and going to the limit as $h \to \infty$, we get the following difference-differential set of inequalities for $M_k^{(d)}(t)$.

$$\begin{aligned}
\frac{d}{dt} M_k^{(d)}(t) &\leq \lambda \mathbb{E}\big({}^d\bar{S}_{k-1}^n(t)\big) - \mathbb{E}\big({}^d\bar{S}_k^n(t)\big) \\
&= \lambda\big(M_{k-1}^{(d)}(t) - M_k^{(d)}(t)\big) - \big(M_k^{(d)}(t) - M_{k+1}^{(d)}(t)\big), \\
&\quad M_0^{(d)}(t) - M_1^{(d)}(t) = 1.
\end{aligned}$$

Similarly, without inequality, for the random choice policy:

$$\begin{aligned}
\frac{d}{dt} M_k^{(1)}(t) &= \lambda\big(M_{k-1}^{(1)}(t) - M_k^{(1)}(t)\big) - \big(M_k^{(1)}(t) - M_{k+1}^{(1)}(t)\big), \\
&\quad M_0^{(1)}(t) - M_1^{(1)}(t) = 1.
\end{aligned}$$

The result follows by showing the for any initial values m_k with $1 = m_0 - m_1 > m_1 - m_2 > \cdots$, $m_k \to 0$, $k \to \infty$, this implies $M_k^{(1)}(t) \geq M_k^{(d)}(t)$.

21.3 Justify the inequalities:

$$\begin{aligned}
&\tilde{D}_l\left(n \int_0^t \beta_l(\bar{X}^n(s))ds\right) \\
&\leq \sup_{u \leq t} \tilde{D}_l\left(n\bar{\beta}_l u\right) \\
&\leq \left[\tilde{D}_l\left(n\bar{\beta}_l t\right) + n\bar{\beta}_l t\right]
\end{aligned}$$

.

Solution

(i) For the first inequality, we note that $\tilde{D}_l\left(n\int_0^t \beta_l(\bar{X}^n(s))ds\right)$ is the value of the process $D_l(w) - w$ at the single point $w = n\int_0^t \beta_l(\bar{X}^n(s))ds$, and so it is less or equal to $\sup_{u\le w}|D_l(u)-u|$. Next we see that $w = n\int_0^t \beta_l(\bar{X}^n(s))ds \le nt\bar{\beta}_l$, and so it follows that the $\tilde{D}_l\left(n\int_0^t \beta_l(\bar{X}^n(s))ds\right) \le \sup_{u\le nt\bar{\beta}_l}|D_l(u) - u| = \sup_{u\le t}\left|D_l(nt\bar{\beta}_l) - nt\bar{\beta}_l\right| = \sup_{u\le t}|\tilde{D}_l(nt\bar{\beta}_l)|$.

(ii) For the second inequality, we note that both $D(u)$ and u are monotone non-decreasing functions, and hence $\sup_{u\le t}|D(u) - u| \le D(t) + t$.

21.4 Show that with

$$F(s(t)) = \left[\lambda(s_{k-1}(t)^d - s_k(t)^d) - (s_k(t) - s_{k+1}(t))\right]_{k=1,2,\dots},$$

F is Lipschitz continuous, i.e. there exists M such that $|F(x) - F(y)| \le M|x - y|$.

Solution

We use the norm:

$$|u - v| = \sup_{k=1,2,\dots}\frac{|u_k - v_k|}{k}$$

under which the space of decreasing sequences is compact, and convergence is equivalent to component-wise convergence. We have:

$$\begin{aligned}
|F(u) - F(v)| &= \sup_{k=1,2,\dots}\frac{|\lambda(u_{k-1} - u_k) - (u_k - u_{k+1} - \lambda(v_{k-1} - v_k) - (v_k - v_{k+1}}{k}\\
&\le \sup_{k=1,2,\dots}\frac{\lambda|u_{k-1} - v_{k-1}|}{k} + \sup_{k=1,2,\dots}\frac{\lambda|u_k - v_k|}{k} + \sup_{k=1,2,\dots}\frac{|u_k - v_k|}{k} + \sup_{k=1,2,\dots}\frac{|u_{k+1} - v_{k+1}|}{k}\\
&\le 5|u - v|.
\end{aligned}$$

21.5 Complete the steps of the proof of Proposition 21.6.

Solution

We first check that substituting s_k^* gives $\frac{d}{dt}s_k(t) = 0$:

$$\begin{aligned}
&\lambda\left[\left(\lambda^{\frac{d^{k-1}-1}{d-1}}\right)^d - \left(\lambda^{\frac{d^k-1}{d-1}}\right)^d\right] - \left[\lambda^{\frac{d^k-1}{d-1}} - \lambda^{\frac{d^{k+1}-1}{d-1}}\right]\\
&= \lambda\left(\lambda^{\frac{d^k-d}{d-1}} - \lambda^{\frac{d^{k+1}-d}{d-1}}\right) - \left(\lambda^{\frac{d^k-1}{d-1}} - \lambda^{\frac{d^{k+1}-1}{d-1}}\right)\\
&= \left(\lambda^{\frac{d^k-d}{d-1}+1} - \lambda^{\frac{d^{k+1}-d}{d-1}+1}\right) - \left(\lambda^{\frac{d^k-1}{d-1}} - \lambda^{\frac{d^{k+1}-1}{d-1}}\right)\\
&= \left(\lambda^{\frac{d^k-1}{d-1}} - \lambda^{\frac{d^{k+1}-1}{d-1}}\right) - \left(\lambda^{\frac{d^k-1}{d-1}} - \lambda^{\frac{d^{k+1}-1}{d-1}}\right)\\
&= 0
\end{aligned}$$

Next, we show that if $\frac{d}{dt}s_k(t) = 0,\ k \ge 1$ then $s_k(t) = s_k^*,\ k \ge 1$.

The added conditions apply $\sum_{k=1}^{\infty} s_k(t) < \infty$, and $s_k(t) \geq 0$,so we can sum all the equations:

$$\begin{aligned} 0 &= \sum_{k=1}^{\infty} \Big(\lambda(s_{k-1}(t)^d - s_k(t)^d) - (s_k(t) - s_{k+1}(t)) \Big) \\ &= \lambda(s_0(t)^d - s_j(t)^d)) - (s_1(t) - s_{j+1}(t)) \\ &= \lambda - s_1(t), \end{aligned}$$

hence for the fixed point $s_1(t) = \lambda$. We now proceed by induction:
Assuming $s_i(t) = \lambda^{\frac{d^i-1}{d-1}}$, $i \leq k$ we need to evaluate $s_{k+1}(t)$, where we use the kth equation and $\frac{d}{dt} s_k(t) = 0$:

$$0 = \frac{d}{dt} s_k(t) = \lambda \left[\left(\lambda^{\frac{d^{k-1}-1}{d-1}} \right)^d - \left(\lambda^{\frac{d^k-1}{d-1}} \right)^d \right] - \left[\lambda^{\frac{d^k-1}{d-1}} - s_{k+1}(t) \right]$$

and looking at the first derivation above, we get: $s_{k+1}(t) = \lambda^{\frac{d^{k+1}-1}{d-1}}$.
Note that the condition $s_j(t) = 0$ for all $t \geq 0$ is necessary. Without it we get that $S(t) = \{1, 1, \ldots\}$ is also a fixed point.

21.6 For the supermarket model, under choose shortest of d, show that in the limiting infinite server system (21.5) if we increase $s_j(0)$ for some j, this will increase or leave unchanged $s_k(t)$ for all $t > 0$ and all k.

Solution

Consider the n server system with states $S_k(t)$ at time 0. Assume we add a single customer to a queue of length $i - 1$. We now couple the original system to the system with the added customer, with state $\tilde{S}_k(t)$. We do the coupling by uniformizing all events to a single Poisson process of rate $(\lambda + 1)n$, and then choosing randomly if arrival or completion, and in the case of completion, decrease queue if > 0 or have dummy event, but, exclude the event of completion of the added customer. taking out one point from a Poisson process, leaves it as a Poisson process. Then, for as long as that new customer is in the system, the new system has one more customer in that particular queue, but that implies that so $\tilde{S}_k(t) \geq S_k(t)$ for all k and t up ot that time. Once it leaves the system we have equality $\tilde{S}_k(t) = S_k(t)$. This holds for any n, and should also hold for the continuous $s_k(t)$.
A formal proof is as follows: we see from $\frac{d}{dt} s_k(t) = \lambda(s_{k-1}^d - s_k^d) - (s_k - s_{k+1}$ that $\frac{d}{dt} s_k(t)$ is increasing or unchanged by increasing any $s_j(t)$, $j \neq k$. This means that the sequence $s_k(t)$ is quasimonotone. This implies that increasing any component increases or leaves unchanged all others as proved in Deimling (2006), pages 70-74.

21.7 In the proof of Proposition 21.7 show that $M(0) < 1/\lambda^{1/(d-1)}$.

Solution

(i) Let $\pi_k = \lambda^{\frac{d^k-1}{d-1}}$ be the fixed point. Define $M(t) = \sup_k (s_k(t)/\pi_k)^{1/d^k}$
Then: $M(t) \leq M(0)$.

Proof Recall from above, that increasing $s_j(0)$ increases all $s_k(t)$. For the sequence $s_k(t)$, consider the sequence that will be obtained by increasing all components (including s_0, to start with $\check{s}_k(0) = M(0)^{d^k}\pi_k$.
We show that $\check{s}_k(0)$ is a fixed point of the differential equations (with perhaps $\check{s}_0(0) > 1$:

$$\begin{aligned}
&\lambda(\check{s}_{k-1}(0)^d - \check{s}_k(0)^d) - (\check{s}_k(0) - \check{s}_{k+1}(0)) \\
&= \lambda\left[\left(M(0)^{d^{k-1}}\pi_{k-1}\right)^d - \left(M(0)^{d^k}\pi_k\right)^d\right] - \left[M(0)^{d^k}\pi_k - M(0)^{d^{k+1}}\pi_{k+1}\right] \\
&= \lambda\left(M(0)^{d^k}\pi_{k-1}^d - M(0)^{d^{k+1}}\pi_k^d\right) - \left(M(0)^{d^k}\pi_k - M(0)^{d^{k+1}}\pi_{k+1}\right) \\
&= M(0)^{d^k}(\lambda\pi_{k-1} - \pi_k) - M(0)^{d^{k+1}}(\lambda\pi_k - \pi_{k+1}) \\
&= 0
\end{aligned}$$

where in the last expression equality to 0 holds for each of the two terms.
Therefore $\check{s}_k(t) = M(0)^{d^k}\pi_k$ for all t, so $\check{M}(t) = M(0)$ for all t. However, since $\check{s}_k(0) \geq s_k(0)$ for all k, we also have $\check{s}_k(t) \geq s_k(t)$ for all k, t, and so $M(t) \leq \check{M}(t) = M(0)$ for all t. □

(ii) So we have seen that $M(t) \leq M(0)$. Let j be the smallest for which $s_j(0) = 0$ (we assumed there is such a j). We then start from the smallest non-zero, which is $s_{j-1}(0)$. Recall also that $s_k \leq 1$ (by definition), and that $\lambda < 1$ (by stability). Then:

$$M(0) \leq (1/\pi_{j-1})^{1/d^{j-1}} < 1/\lambda^{\frac{1}{d-1}}$$

To see this:

$$M(0) = sup_k(s_k(t)/\pi_k)^{1/d^k} \leq sup_{0\leq k<j}(1/\pi_k)^{1/d^k}$$

and

$$(\pi_k)^{1/d^k} = \left(\lambda^{\frac{d^k-1}{d-1}}\right)^{1/d^k} = \lambda^{\frac{1}{d-1}}\Big/\lambda^{\frac{1}{d^k(d-1)}}.$$

but $\lambda^{\frac{1}{d^k(d-1)}} < 1$ and it is increasing in k. Hence, for all $0 \leq k \leq j-1$:

$$(\pi_k)^{1/d^k} = \lambda^{\frac{1}{d-1}}\Big/\lambda^{\frac{1}{d^k(d-1)}} \geq (\pi_{j-1})^{1/d^{j-1}} = \lambda^{\frac{1}{d-1}}\Big/\lambda^{\frac{1}{d^{j-1}(d-1)}} > \lambda^{\frac{1}{d-1}}.$$

We then have:

$$s_k(t) \leq M(t)^{d^k}\pi_k \leq M(0)^{d_k}\pi_k = \lambda^{-\frac{1}{d-1}}\left(\lambda^{\frac{1}{d-1}}M(0)\right)^{d^k}$$

and we saw that $\lambda^{\frac{1}{d-1}}M(0) < 1$, so:

$$s_k(t) \leq \gamma\alpha^{\beta^k}, \qquad \alpha = \lambda^{\frac{1}{d-1}}M(0) < 1, \quad \beta = d \geq 2, \quad \gamma = \lambda^{-\frac{1}{d-1}}.$$

(iii) If the system starts empty, then $M(0) = 1$ so $s_k(t) \leq \pi_k$.

21.8 Show that one can find an increasing sequence $w_k \geq 1$ and $\delta > 0$ that satisfy

$$w_{k+1} \leq w_k + \frac{(1-\delta)w_k - w_{k-1}}{\lambda(2\pi_k + 1)}$$

so that this sequence is bounded by a geometric sequence.

Solution

We now construct an increasing sequence of w_k starting with $w_0 = 0$, $w_1 = 1$, that satisfies this inequality. For the finite number of k such that $\lambda(2\pi_k + 1) \geq \frac{1+\lambda}{2}$ we take: $w_{k+1} = w_k + \frac{(1-\delta)w_k - w_{k-1}}{3}$, for the rest of k, with $\lambda(2\pi_k + 1) < \frac{1+\lambda}{2}$ we take: $w_{k+1} = w_k + \frac{2(1-\delta)w_k - 2w_{k-1}}{1+\lambda}$. We can choose δ small enough that this sequence is increasing, and it is dominated by a geometric increasing sequence.

To see the choice of δ, note that for $\lambda(2\pi_k + 1) < \frac{1+\lambda}{2}$,

$$(w_k - w_{k-1})(1-\delta)\frac{2}{1+\lambda} \leq w_{k+1} - w_k \leq (w_k - w_{k-1})\frac{2}{1+\lambda}$$

21.9 Prove that

$$\lim_{\lambda \to 1} \frac{\sum_{k=0}^{\infty} \lambda^{d^k}}{\log \frac{1}{1-\lambda}} = \frac{1}{\log d}$$

and use this to prove Proposition 21.10 [Mitzenmacher (1996)].

Solution

The proof of the limiting result is quite lengthy, it is given in Mitzenmacher (2001), as Lemma 3.

We now use this limiting result to prove Proposition 21.10 . Let $\tilde{\lambda} = \lambda^{1/(d-1)}$. Then:

$$\overline{W}_d(\lambda) = \sum_{k=1}^{\infty} \lambda^{\frac{d^k - d}{d-1}} = \frac{\sum_{i=1}^{\infty} \tilde{\lambda}^{d^i}}{\lambda^{d/(d-1)}},$$

Hence:

$$\begin{aligned}
\lim_{\lambda \nearrow 1} \frac{\overline{W}_d(\lambda)}{\overline{W}_1(\lambda)} &= \lim_{\lambda \nearrow 1} \frac{\overline{W}_d(\lambda)}{\log 1/(1-\lambda)} \\
&= \lim_{\lambda \nearrow 1} \frac{\sum_{i=1}^{\infty} \tilde{\lambda}^{d^i}}{\lambda^{d/(d-1)} \log 1/(1-\lambda)} \\
&= \lim_{\lambda \nearrow 1} \frac{\sum_{i=1}^{\infty} \tilde{\lambda}^{d^i}}{\log 1/(1-\tilde{\lambda})} \frac{\log 1/(1-\tilde{\lambda})}{\log 1/(1-\lambda)} \frac{1}{\lambda^{d/(d-1)}},
\end{aligned}$$

The last two terms tend to 1, and the first converges to $1/\log d$, by the above result.

21.10 Show that the M/M/n-JSQ model can be described by a density dependent Markov chain.

Solution

Strictly speaking, this will only be the case if the arrival rate is λn, which is not the case for the Halfin-Whitt regime. If we make this assumption,

$$\begin{aligned}
&S_k^n(t) = \text{number of queues of length } k, \quad \text{Transitions are } \pm e_k\\
&q_{x,x-e_k} = n\Big(\frac{S_k^n}{n} - \frac{S_{k+1}^n}{n}\Big),\ k = 1, 2, \ldots,\\
&q_{x,x+e_1} = n\lambda \mathbb{1}\Big(\frac{S_1^n}{n} < 1\Big),\\
&q_{x,x+e_k} = n\lambda \mathbb{1}\Big(\frac{S_{k-1}^n}{n} = 1,\ \frac{S_k^n}{n} < 1\Big),\ k = 2, 3, \ldots.
\end{aligned}$$

21.11 For the M/M/n-JSQ model, prove that for the limiting system, the scaled counts of the queues $\hat{S}_k(t)$, $k \geq 3$ are given by

$$\begin{aligned}
\hat{S}_k(t) &= \left(\hat{S}_k(0) + \sum_{j=1}^{i-k} \frac{t^j}{j!}\hat{S}_{k+j}(0)\right)e^{-t}, \qquad 3 \leq k < i,\\
\hat{S}_i(t) &= \hat{S}_i(0)e^{-t}
\end{aligned}$$

Solution

The limiting equations for $\hat{S}_k(t)$, $3 \leq k \leq i$ are

$$\begin{aligned}
\hat{S}_k(t) &= \hat{S}_k(0) - \int_0^t \Big(\hat{S}_k(s) - \hat{S}_{k+1}(s)\Big)ds, \quad 2 < k < i,\\
\hat{S}_i(t) &= \hat{S}_i(0) - \int_0^t \hat{S}_i(s)ds,
\end{aligned}$$

For i we have:

$$\frac{d}{dt}\hat{S}_i(t) = -\hat{S}_i(t), \text{ with boundary value } \hat{S}_i(0),$$

which is solved by:

$$\hat{S}_i(t) = \hat{S}_i(0)e^{-t}$$

Next we have:

$$\hat{S}_{i-1}(t) = \hat{S}_{i-1}(0)e^{-t} + \hat{S}_i(0)te^{-t}$$

and it is then easily checked by induction that

$$\frac{d}{dt}\hat{S}_k(t) = -\hat{S}_k(t) + \hat{S}_{k+1}(t), \text{ with boundary value } \hat{S}_k(0)$$

is solved by

$$\hat{S}_k(t) = \left(\hat{S}_k(0) + \sum_{j=1}^{i-k} \frac{t^j}{j!}\hat{S}_{k+j}(0)\right)e^{-t}$$

21.12 Show that if $\tilde{w}_1(t), \tilde{w}_2(t)$ solve equations (21.28), then $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \phi_0(\tilde{w}_1)$ solve equations (21.27), and show that if (w_1, w_2) solve equations (21.27), then $x_1 = \psi_0(w_1)$, $u_1 = \phi_0(w_1)$, $x_2 = \psi_B(w_2)$, $u_2 = \phi_B(w_2)$ solve equations (21.25).

Solution

Define the Skorohod reflection transformation: for $y \in \mathbb{D}$ and a constant B, let z be non-decreasing $z(0) = 0$, $x(t) = y(t) - z(t) \leq B$, and $\mathbb{1}(x < B)dz = 0$, then the transformation is $x = \psi_B(y)$, $z = \phi_B(y)$, which is unique and continuous.

Consider first the equations:

$$\tilde{w}_1(t) = b_1 + y_1(t) - \int_0^t \Big(\psi_0(\tilde{w}_1(s)) - \psi_B\big(\tilde{w}_2(s) + \phi_0(\tilde{w}_1(s))\big)\Big)ds,$$
$$\tilde{w}_2(t) = b_2 + \tilde{y}_2(t) - \int_0^t \psi_B\big(\tilde{w}_2(s) + \phi_0(\tilde{w}_1(s))\big)ds,$$

and let $(\tilde{w}_1, \tilde{w}_2)$ be their solution.

Let $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \phi_0(\tilde{w}_1)$. Then we first show that (w_1, w_2) solve the equations:

$$w_1(t) = b_1 + y_1(t) - \int_0^t (\psi_0(w_1(s)) - \psi_B(w_2(s)))\, ds,$$
$$w_2(t) = b_2 + \tilde{y}_2(t) + \phi_0(w_1(t)) - \int_0^t \psi_B(w_2(s))ds.$$

Proof

Into $w_2(t) = b_2 + \tilde{y}_2(t) + \phi_0(w_1(t)) - \int_0^t \psi_B(w_2(s))ds$,

Substitute $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \phi_0(\tilde{w}_1)$,

$$\tilde{w}_2(t) + \phi_0(\tilde{w}_1(t)) = b_2 + \tilde{y}_2(t) + \phi_0(\tilde{w}_1(t)) - \int_0^t \psi_B\big(\tilde{w}_2(s) + \phi_0(\tilde{w}_1(s))\big)ds,$$

or $\tilde{w}_2(t) = b_2 + \tilde{y}_2(t) - \int_0^t \psi_B\big(\tilde{w}_2(s) + \phi_0(\tilde{w}_1(s))\big)ds$, as required.

Into $w_1(t) = b_1 + y_1(t) - \int_0^t (\psi_0(w_1(s)) - \psi_B(w_2(s)))\, ds$,

Substitute $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \phi_0(\tilde{w}_1)$,

$$\tilde{w}_1(t) = b_1 + y_1(y) - \int_0^t \Big(\psi_0(\tilde{w}_1(s)) - \psi_B\big(\tilde{w}_2(s) + \phi_0(\tilde{w}_1(s))\big)\Big)ds, \text{ as required.}$$

□

Let (w_1, w_2) be the solutions of the second set of equations. Let $x_1 = \psi_0(w_1)$, $u_1 = \phi_0(w_1)$, $x_2 = \psi_B(w_2)$, $u_2 = \phi_B(w_2)$. We next show that

(x_1, x_2) solve the equations:

$$\begin{aligned} x_1(t) &= b_1 + y_1(t) - \int_0^t (x_1(s) - x_2(s))\, ds - u_1(t), \\ x_2(t) &= b_2 + \tilde{y}_2(t) - \int_0^t x_2(s) ds + u_1(t) - u_2(t), \\ 0 &= \int_0^t \mathbb{1}(x_1(s) < 0) du_1(s), \\ 0 &= \int_0^t \mathbb{1}(x_2(s) < B) du_2(s), \\ x_1(t) \le 0, &\quad 0 \le x_2(t) \le B, \quad t \ge 0, \end{aligned}$$

Proof

Into $x_1(t) = b_1 + y_1(t) - \int_0^t (x_1(s) - x_2(s)) ds - u_1(t)$,
Substitute $x_1 = \psi_0(w_1)$, $u_1 = \phi_0(w_1)$, $x_2 = \psi_B(w_2)$,

$$\psi_0(w_1(t)) = b_1 + y_1(t) - \int_0^t \Big(\psi_0(w_1(s)) - \psi_B(w_2(s))\Big) ds - \phi_0(w_1(t))$$

and using $w_1(t) = \psi_0(w_1(t)) + \phi_0(w_1(t))$
we get $w_1(t) = b_1 + y_1(t) - \int_0^t \Big(\psi_0(w_1(s)) - \psi_B(w_2(s))\Big)\, ds$ as required.

Into $x_2(t) = b_2 + \tilde{y}_2(t) - \int_0^t x_2(s) ds + u_1(t) - u_2(t)$,
Substitute $x_2 = \psi_B(w_2)$, $u_1 = \phi_0(w_1)$, $u_2 = \phi_B(w_2)$,

$$\psi_B(w_2(t)) = b_2 + \tilde{y}_2(t) - \int_0^t \psi_B(w_2(s)) ds + \phi_0(w_1(t)) - \phi_B(w_2(t))$$

and using $w_2(t) = \psi_B(w_2(t)) + \phi_B(w_1(t))$
we get $w_2(t) = b_2 + \tilde{y}_2(t) + \phi_0(w_1(t)) - \int_0^t \psi_B(w_2(s)) ds$, as required.

□

21.13 For the M/M/n-JSQ model, show that the transformation:

$$\begin{aligned} T_1(\tilde{w}_1, \tilde{w}_2)) &= b_1 + y_1(t) - \int_0^t (\psi_0(\tilde{w}_1(s)) - \psi_B(\tilde{w}_2(s)) - \phi_0(\tilde{w}_1(s)))\, ds, \\ T_2(\tilde{w}_1, \tilde{w}_2)) &= b_2 + \tilde{y}_2(t) - \int_0^t (\psi_B(\tilde{w}_2(s)) - \phi_0(\tilde{w}_1(s)))\, ds, \end{aligned}$$

is a contraction mapping from $\mathbb{D}^2$ to $\mathbb{D}^2$.
Solution

By standard estimates,

$$||T_1(\tilde{w}) - T_1(\tilde{v})||_t \leq 5t||\tilde{w} - \tilde{v}||_t,$$
$$||T_2(\tilde{w}) - T_2(\tilde{v})||_t \leq 3t||\tilde{w} - \tilde{v}||_t,$$

so for $t_0 < 1/5$ this is a contraction mapping, so the solution exists and is unique for the interval $[0, t_0]$. By the same argument it is unique for $t \in [t_0, 2t_0], [2t_0, 3t_0], \ldots$, so it exists and is unique for all $t > 0$.

21.14 Show that under Halfin-Whitt heavy traffic staffing, with choose shortest of d policy, the stationary average sojourn time grows like $\dfrac{\log n}{2 \log d}$.

Solution

We use Proposition 21.10 by which as $\lambda \nearrow 1$:

$$\bar{W}_d(\lambda_n) \sim \log \bar{W}_1(\lambda_n) \big/ \log d$$

and substitute

$$\bar{W}_1(\lambda_n) = 1/(1 - \lambda_n) = 1/(1 - (1 - \beta/\sqrt{n})) = \sqrt{n}/\beta$$

so

$$\bar{W}_d(\lambda_n) \sim \log(\sqrt{n}/\beta) \big/ \log d \sim \frac{\log n}{2 \log d},$$

and by $\mathbb{E}({}^dW^n) \to \bar{W}_d$ the proposition follows.

22

Parallel Servers with Skill Based Routing

Exercises

22.1 For state $\mathfrak{s} = (S_1, n_1, \ldots, S_i, n_i, S_{i+1}, \ldots, S_J)$ of the PSS system under FCFS-ALIS, write all the transitions out of state $\mathfrak{s}$ and their transition rates [Adan and Weiss (2014); Visschers et al. (2012)].

Solution

Denote:

$$\delta_j(S) = \begin{cases} \dfrac{\lambda_{\mathcal{U}(S_1,\ldots,S_j)}}{\lambda_{\mathcal{U}(S_1,\ldots,S_j,S)}} & \mathcal{U}(S_1,\ldots,S_j,S) \neq \emptyset, \\ 0 & \mathcal{U}(S_1,\ldots,S_j,S) = \emptyset. \end{cases}$$

(i) Arrival of customer that joins the queue:

$$q\big(\mathfrak{s} \to (S_1, \ldots, S_i, n_i + 1, S_{i+1}, \ldots, S_J)\big) = \lambda_{\mathcal{U}(S_1,\ldots,S_i)}$$

(ii) Arrival of customer of type that activates idle server S_j where $i < j \leq J$:

$$q\big(\mathfrak{s} \to (S_1, \ldots, S_i, n_i, S_k, 0, S_{i+1}, \ldots, S_{k-1}, S_{k+1}, \ldots, S_J)\big) = \lambda_{C(S_k)\setminus C(S_{k+1},\ldots,S_J)}$$

(iii) Completion of service by server S_j where $1 \leq j \leq i$ that becomes idle:

$$\begin{aligned} &q\big(\mathfrak{s} \to (S_1, \ldots, S_{j-1}, n_{j-1} + n_j, S_{j+1} \ldots, S_i, n_i, S_j, S_{i+1}, \ldots, S_J)\big) \\ &= \mu_{S_j} \delta_j(S_j)^{n_j} \cdots \delta_i(S_j)^{n_i}. \end{aligned}$$

(iv) Completion of service by server S_j, which immediately starts service of customer l among the n_k customers between S_k and S_{k+1}, where $j \leq k \leq i$ and $1 \leq l \leq n_k$:

$$\begin{aligned} &q\big(\mathfrak{s} \to (S_1, \ldots, S_{j-1}, n_{j-1} + n_j, S_{j+1} \ldots, S_k, l-1, S_j, \\ &\qquad\qquad n_k - l, S_{k+1} \ldots, S_i, n_i, S_j, S_{i+1}, \ldots, S_J)\big) \\ &= \mu_{S_j} \delta_j(S_j)^{n_j} \cdots \delta_k(S_j)^{l-1} (1 - \delta_k(S_j)). \end{aligned}$$

22.2 (continued) For state $\mathfrak{s} = (S_1, n_1, \ldots, S_i, n_i, S_{i+1}, \ldots, S_J)$, write all the possible transition into this state, and find their transition rates.

Solution
Denote:

$$\delta_j(S) = \begin{cases} \dfrac{\lambda_{\mathcal{U}(S_1,\ldots,S_j)}}{\lambda_{\mathcal{U}(S_1,\ldots,S_j,S)}} & \mathcal{U}(S_1,\ldots,S_j,S) \neq \emptyset, \\ 0 & \mathcal{U}(S_1,\ldots,S_j,S) = \emptyset. \end{cases}$$

(i) Transition due to a departure, where a server becomes idle, as illustrated in Fig. 22.1

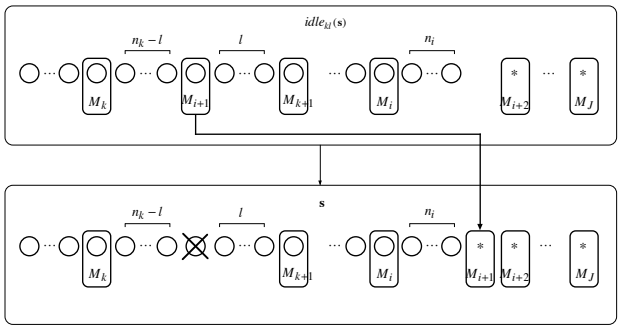


Figure 22.1 Transition from state $idle_{k,l}(\mathfrak{s})$ to state $\mathfrak{s}$

The state from which the transition is made:

$$idle_{kl}(\mathfrak{s}) = (S_1, n_1, \ldots, S_k, n_k - l, S_{i+1}, l, S_{k+1}, \ldots, S_i, n_i, S_{i+2}, \ldots, S_J),$$

and the transition probability, conditional on service completion by server S_{i+1}, is

$$p_{k,l}(\mathfrak{s}) = \delta_k(S_{i+1})^l \delta_{k+1}(S_{i+1})^{n_{k+1}} \ldots \delta_i(S_{i+1})^{n_i}, \quad k \geq 1,\ l = 0, \ldots, n_k,$$
$$p_{0,0}(\mathfrak{s}) = p_{1,n_1}(\mathfrak{s}).$$

(ii) Transition in which a customer departs, and the server starts a new service, as illustrated in figure. 22.2
The originating state is

$$swap_{k,l,j}(\mathfrak{s}) = (S_1, n_1, \ldots, S_k, n_k - l, S_j, l, \ldots, S_{j-1}, n_{j-1} + 1 + n_j, S_{j+1}, \ldots, S_J),$$

and the transition probability, conditional on service completion by server M_j, is

$$q_{k,l,j}(\mathfrak{s}) = \delta_k(S_j)^l \delta_{k+1}(S_j)^{n_{k+1}} \ldots \delta_{j-1}(S_j)^{n_{j-1}} \left(1 - \delta_{j-1}(S_j)\right),$$
$$j = 2, \ldots, J, 1 \leq k < j, l = 0, \ldots, n_k,$$
$$q_{0,0,j}(\mathfrak{s}) = q_{1,n_1,j}(\mathfrak{s}),$$
$$q_{0,0,1}(\mathfrak{s}) = 1.$$

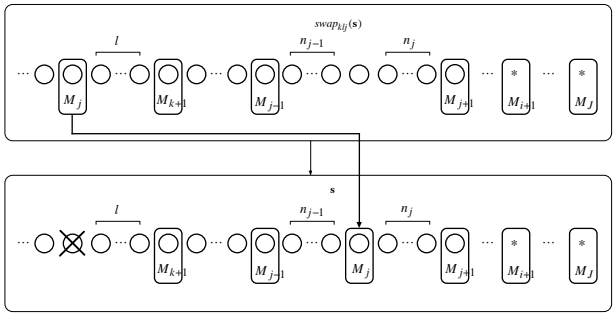


Figure 22.2 Transition from state $swap_{k,l,j}(\mathfrak{s})$ to state $\mathfrak{s}$.

(iii) Transition in which an arrival joins the queue. The originating state is

$$wait(\mathfrak{s}) = (M_1, n_1, \ldots, \ldots, M_i, n_i - 1, M_{i+1}, \ldots, M_J), \quad n_i > 0,$$

and the transition rate is $\lambda_{\mathcal{U}(\{M_1,\ldots,M_i\})}$.

(iv) Transition in which an arrival activates an idle server, as illustrated in figure 22.3

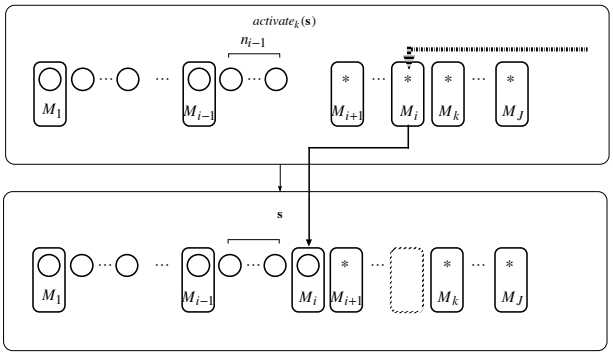


Figure 22.3 Transition from state $activate_k(\mathfrak{s})$ to state $\mathfrak{s}$.

The originating state is

$$activate_k(\mathfrak{s}) = (S_1, n_1, \ldots, S_{i-1}, n_{i-1}, S_{i+1}, \ldots, S_{k-1}, S_i, S_k, \ldots, S_J),$$

and the transition rate is $\lambda_{C(S_i)\setminus C(\{S_k,\ldots,S_J\})}$ for $k = i+1, \ldots, J$, and in the case that S_J is activated, we use the convention that $k = J+1$, and the rate is $\lambda_{C(S_J)}$.

22.3 (continued) Write down the partial balance equations for the four type of balanced transitions.

Solution

Define the sum of all transition rates in which a server completes service and becomes idle,

$$\mathcal{P}_{S_{i+1}}(\mathfrak{s}) = \sum_{k=1}^{i}\sum_{l=0}^{n_k} p_{k,l}(\mathfrak{s})\pi(idle_{k,l}(\mathfrak{s})) + p_{1,n_1}(\mathfrak{s})\pi(idle_{0,0}(\mathfrak{s})),$$

and the sum of transition rates in which a server completes service and moves on to become the jth active server.

$$Q_{S_j}(\mathfrak{s}) = \begin{cases} \displaystyle\sum_{k=1}^{j-1}\sum_{l=0}^{n_k} q_{k,l,j}(\mathfrak{s})\pi(swap_{k,l,j}(\mathfrak{s}))+ \\ \qquad +q_{0,0,j}\pi(swap_{0,0,j}(\mathfrak{s})), & \text{if } \mathcal{U}(\{S_1,\ldots,S_j\}) \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Four types of partial balance need to be verified. In these four categories of transitions, we have transitions out of state $\mathfrak{s}$ on the left, balanced against transitions into state $\mathfrak{s}$ on the right.

(i) The total probability flux out of state $\mathfrak{s}$ due to an arrival that activates a server equals the total probability flux into state $\mathfrak{s}$ due to a departure which idles a server:

$$\lambda_{C(\{S_{i+1},\ldots,S_J\})}\pi(\mathfrak{s}) = \mu_{S_{i+1}}\mathcal{P}_{S_{i+1}}(\mathfrak{s}). \tag{22.1}$$

(ii) The total probability flux out of state $\mathfrak{s}$, due to an arrival that joins the queue, equals the total probability flux into state $\mathfrak{s}$, due to a departure which is followed by another start of service (so that the set of idle servers is unchanged):

$$\lambda_{\mathcal{U}(\{S_1,\ldots,S_i\})}\pi(\mathfrak{s}) = \sum_{j=1}^{i}\mu_{S_j}Q_{S_j}(\mathfrak{s}). \tag{22.2}$$

(iii) The total probability flux out of state $\mathfrak{s}$ in which $n_i = 0$ due to a departure, equals the total probability flux into state $\mathfrak{s}$, due to an arrival of a customer which activates server S_i:

$$\mu_{\{S_1,\ldots,S_i\}}\pi(\mathfrak{s}) = \sum_{k=i+1}^{J}\lambda_{C(S_i)\setminus C(\{S_k,\ldots,S_J\})}\pi(activate_k(\mathfrak{s})) + \lambda_{C(S_i)}\pi(activate_{J+1}(\mathfrak{s})), \qquad n_i = 0. \tag{22.3}$$

(iv) The total probability flux out of state $\mathfrak{s}$ in which $n_i > 0$ due to a departure, equals the total probability flux into state $\mathfrak{s}$, due to an arrival of a customer which joins the queue:

$$\mu_{\{S_1,\ldots,S_i\}}\pi(\mathfrak{s}) = \lambda_{\mathcal{U}(\{S_1,\ldots,S_i\})}\pi(wait(\mathfrak{s})), \qquad n_i > 0. \tag{22.4}$$

22.4 (continued) Verify that the proposed stationary distribution (22.2) satisfies the four partial balance equations.

Solution

We consider the state $\mathfrak{s} = (S_1, n_1, \ldots, S_i, n_i, S_{i+1}, \ldots, S_J)$, so the permutation is $S_1, \ldots, S_J$, the number busy is i, and we denote $a_j = \lambda_{\mathcal{U}(S_1,\ldots,S_j)}$, $b_j = \mu_{S_1,\ldots,S_j}$, $j = 1, \ldots, i$, and $c_j = \lambda_{C(S_j,\ldots,S_J)}$, $j = i+1, \ldots, J$. Denote also $a_{j,S} = \lambda_{\mathcal{U}(S_1,\ldots,S_j,S)}$, $b_{j,S} = b_j + \mu_S$,

Accordingly, the presumed expressions for the various states are:

$$\pi(\mathfrak{s}) = B \prod_{j=1}^{i} \frac{a_j^{n_j}}{b_j^{n_j+1}} \prod_{j=i+1}^{J} \frac{1}{c_j},$$

$$\pi(idle_{kl}(\mathfrak{s})) = \pi(\mathfrak{s}) \frac{1}{b_k + \mu_{S_{i+1}}} \left(\frac{b_k}{b_k + \mu_{S_{i+1}}}\right)^l \left(\frac{a_{k,S_{i+1}}}{a_k}\right)^l \prod_{j=k+1}^{i} \left(\frac{b_j}{b_j + \mu_{S_{i+1}}}\right)^{n_j+1} \left(\frac{a_{j,S_{i+1}}}{a_j}\right)^{n_j} c_{i+1},$$

$$\pi(idle_{0,0}(\mathfrak{s})) = \pi(idle_{1,n_1}(\mathfrak{s})) \frac{b_1}{\mu_{S_{i+1}}}$$

$$\pi(swap_{klr}(\mathfrak{s})) = \pi(\mathfrak{s}) \frac{1}{b_k + \mu_{S_r}} \left(\frac{b_k}{b_k + \mu_{S_r}}\right)^l \left(\frac{a_{k,S_r}}{a_k}\right)^l \prod_{j=k+1}^{r-1} \left(\frac{b_j}{b_j + \mu_{S_r}}\right)^{n_j+1} \left(\frac{a_{j,S_r}}{a_j}\right)^{n_j} a_{r-1,S_r}$$

recall that $\delta_j(S) = \frac{a_j}{a_{j,S}}$, to get that:

$$\begin{aligned}
p_{k,l}(\mathfrak{s}) &= \delta_k(S_{i+1})^l \delta_{k+1}(S_{i+1})^{n_{k+1}} \ldots \delta_i(S_{i+1})^{n_i} \\
&= \left(\frac{a_k}{a_{k,S_{i+1}}}\right)^l \prod_{j=k+1}^{i} \left(\frac{a_j}{a_{j,S_{i+1}}}\right)^{n_j}, \\
p_{0,0}(\mathfrak{s}) &= p_{1,n_1}(\mathfrak{s}), \\
q_{k,l,r}(\mathfrak{s}) &= \delta_k(S_r)^l \delta_{k+1}(S_r)^{n_{k+1}} \ldots \delta_{r-1}(S_r)^{n_{r-1}} \left(1 - \delta_{r-1}(S_r)\right) \\
&= \left(\frac{a_k}{a_{k,S_r}}\right)^l \prod_{j=k+1}^{r-1} \left(\frac{a_j}{a_{j,S_r}}\right)^{n_j} \left(1 - \frac{a_{r-1}}{a_{r-1,S_r}}\right), \\
q_{0,0,j}(\mathfrak{s}) &= q_{1,n_1,j}(\mathfrak{s}), \quad j > 1, \qquad q_{0,0,1}(\mathfrak{s}) = 1
\end{aligned}$$

so that:

$$p_{k,l}(\mathfrak{s})\pi(idle_{kl}(\mathfrak{s})) = \pi(\mathfrak{s})$$
$$\frac{1}{b_k+\mu_{S_{i+1}}}\left(\frac{b_k}{b_k+\mu_{S_{i+1}}}\right)^l \prod_{j=k+1}^{i}\left(\frac{b_j}{b_j+\mu_{S_{i+1}}}\right)^{n_j+1} c_{i+1},$$
$$q_{k,l,r}(\mathfrak{s})\pi(swap_{klr}(\mathfrak{s})) = \pi(\mathfrak{s})$$
$$\frac{1}{b_k+\mu_{S_r}}\left(\frac{b_k}{b_k+\mu_{S_r}}\right)^l \prod_{j=k+1}^{r-1}\left(\frac{b_j}{b_j+\mu_{S_r}}\right)^{n_j+1} (a_{r-1,S_r} - a_{r-1})$$

We are now ready to verify (i) and (ii):
(i) We need to verify:

$$\lambda_{C(\{S_{i+1},\ldots,S_J\})}\pi(\mathfrak{s}) = \mu_{S_{i+1}}\mathcal{P}_{S_{i+1}}(\mathfrak{s}).$$

We need to show that:

$$\mu_{S_{i+1}}\mathcal{P}_{S_{i+1}}(\mathfrak{s}) \Big/ \lambda_{C(\{S_{i+1},\ldots,S_J\})}\pi(\mathfrak{s}) = 1.$$

Recall that $c_{i+1} = \lambda_{C(\{S_{i+1},\ldots,S_J\})}$, so we ned to show, using the above that

$$1 = \sum_{k=1}^{i}\sum_{l=0}^{n_k} \frac{\mu_{S_{i+1}}}{b_k+\mu_{S_{i+1}}}\left(\frac{b_k}{b_k+\mu_{S_{i+1}}}\right)^l \prod_{j=k+1}^{i}\left(\frac{b_j}{b_j+\mu_{S_{i+1}}}\right)^{n_j+1}$$
$$+\prod_{j=1}^{i}\left(\frac{b_j}{b_j+\mu_{S_{i+1}}}\right)^{n_j+1}.$$

This is in fact correct, since we have here the sum of probabilities of a sequence of Bernoulli trials, with success probabilities $\beta_k = \frac{\mu_{S_{i+1}}}{b_k+\mu_{S_{i+1}}}$, for the n_{k+1} attempts, $k = i, i-1, \ldots, 1$, and in the last attempt we count both the probability of success and of failure.
(ii) We need to show:

$$\lambda_{\mathcal{U}(\{S_1,\ldots,S_i\})}\pi(\mathfrak{s}) = \sum_{j=1}^{i}\mu_{S_j}\mathcal{Q}_{S_j}(\mathfrak{s}).$$

To do so we need to show that

$$\sum_{j=1}^{i}\mu_{S_j}\mathcal{Q}_{S_j}(\mathfrak{s}) \Big/ \lambda_{\mathcal{U}(\{S_1,\ldots,S_i\})}\pi(\mathfrak{s}) = 1$$

That is:

$$
\begin{aligned}
1 = \sum_{r=1}^{i} \Big[& \mu_{S_r} \mathbf{1}(\mathcal{U}(s_1, \ldots, S_r) \neq 0) \\
& \sum_{k=1}^{r-1} \sum_{l=0}^{n_k} q_{k,l,r}(\mathfrak{s}) \pi(swap_{k,l,r}(\mathfrak{s})) \\
& \qquad + q_{0,0,r} \pi(swap_{0,0,r}(\mathfrak{s})) \Big] \Big/ \lambda_{C(\{S_{i+1}, \ldots, S_J\})} \\
= \sum_{r=1}^{i} & \mathbf{1}(\mathcal{U}(s_1, \ldots, S_r) \neq 0) \\
& \left[\sum_{k=1}^{r-1} \sum_{l=0}^{n_k} \frac{\mu_{S_r}}{b_k + \mu_{S_r}} \left(\frac{b_k}{b_k + \mu_{S_r}} \right)^l \prod_{j=k+1}^{r-1} \left(\frac{b_j}{b_j + \mu_{S_r}} \right)^{n_j + 1} \right. \\
& \left. + \sum_{l=0}^{n_1} \frac{b_1}{b_1 + \mu_{S_r}} \left(\frac{b_1}{b_1 + \mu_{S_r}} \right)^l \prod_{j=2}^{r-1} \left(\frac{b_j}{b_j + \mu_{S_r}} \right)^{n_j + 1} \right] (a_{r-1, S_r} - a_{r-1}) \Big/ \lambda_{C(\{S_{i+1}, \ldots, S_J\})}
\end{aligned}
$$

The expressions inside the square brackets in the last line add up to 1, as for (i). What is left to show is:

$$
\begin{aligned}
& \sum_{r=1}^{i} \mathbf{1}(\mathcal{U}(s_1, \ldots, S_r) \neq 0)(a_{r-1, S_r} - a_{r-1}) \Big/ \lambda_{C(\{S_{i+1}, \ldots, S_J\})} \\
= & \sum_{r=1}^{i} \mathbf{1}(\mathcal{U}(s_1, \ldots, S_r) \neq 0)(\lambda_{\mathcal{U}(S_1, \ldots, S_r)} - \lambda_{\mathcal{U}(S_1, \ldots, S_{r-1})}) \Big/ \lambda_{C(\{S_{i+1}, \ldots, S_J\})} \\
= & \frac{\sum_{r=1}^{i} \lambda_{\mathcal{U}(s_r)}}{\lambda_{C(\{S_{i+1}, \ldots, S_J\})}} = 1.
\end{aligned}
$$

(iii) When $n_i = 0$ we have that

$$
\begin{aligned}
\pi(activate_k(\mathfrak{s})) &= \pi(S_1, n_1, \ldots, S_{i-1}, n_{i-1}, S_{i+1}, \ldots, S_{k-1}, S_i, S_k, \ldots, S_J) \\
&= \pi(\mathfrak{s}) \frac{\mu_{S_1, \ldots, S_i}}{\lambda_{C(S_k, \ldots, S_J)} + \lambda_{C(S_i) \setminus C(S_k, \ldots, S_J)}} \prod_{j=i+1}^{k-1} \frac{\lambda_{C(S_j, \ldots, S_J)}}{\lambda_{C(S_j, \ldots, S_J)} + \lambda_{C(S_i) \setminus C(S_j, \ldots, S_J)}}
\end{aligned}
$$

We wish to verify:

$$
\begin{aligned}
\mu_{\{S_1, \ldots, S_i\}} \pi(\mathfrak{s}) = & \sum_{k=i+1}^{J} \lambda_{C(S_i) \setminus C(\{S_k, \ldots, S_J\})} \pi(activate_k(\mathfrak{s})) \\
& + \lambda_{C(S_i)} \pi(activate_{J+1}(\mathfrak{s})),
\end{aligned}
$$

So we need to show:

$$1 = \sum_{k=i+1}^{J} \frac{\lambda_{C(S_i)\setminus C(\{S_k,\ldots,S_J\})}}{\lambda_{C(S_k,\ldots,S_J)} + \lambda_{C(S_i)\setminus C(S_k,\ldots,S_J)}} \prod_{j=i+1}^{k-1} \frac{\lambda_{C(S_j,\ldots,S_J)}}{\lambda_{C(S_j,\ldots,S_J)} + \lambda_{C(S_i)\setminus C(S_j,\ldots,S_J)}}$$
$$+ \prod_{j=i+1}^{J} \frac{\lambda_{C(S_j,\ldots,S_J)}}{\lambda_{C(S_j,\ldots,S_J)} + \lambda_{C(S_i)\setminus C(S_j,\ldots,S_J)}}$$

Denote: $l_j = \lambda_{C(S_j,\ldots,S_J)}$ and $l_{\neg j} = \lambda_{C(S_i)\setminus C(S_j,\ldots,S_J)}$. We rewrite the sum of products by rearranging the terms and get:

$$\begin{array}{l}
\frac{l_{\neg i+1}}{l_{i+1}+l_{\neg i+1}} \\
+\frac{l_{i+1}}{l_{i+1}+l_{\neg i+1}} \quad \left(\frac{l_{\neg i+2}}{l_{i+2}+l_{\neg i+2}}\right. \\
\qquad\qquad +\frac{l_{i+2}}{l_{i+2}+l_{\neg i+2}} \quad \left(\frac{l_{\neg i+3}}{l_{i+3}+l_{\neg i+3}}\right. \\
\qquad\qquad\qquad +\frac{l_{i+3}}{l_{i+3}+l_{\neg i+3}} \quad \left(\right. \\
\qquad\qquad\qquad\qquad \vdots \\
\qquad\qquad\qquad\qquad \left(\frac{l_{\neg k}}{l_k+l_{\neg k}}\right. \qquad\qquad\qquad , \\
\qquad\qquad\qquad\qquad +\frac{l_k}{l_k+l_{\neg k}} \quad \left(\right. \\
\qquad\qquad\qquad\qquad\qquad \vdots \\
\qquad\qquad\qquad\qquad\qquad \left(\frac{l_{\neg J}}{l_J+l_{\neg J}}\right. \\
\qquad\qquad\qquad\qquad\qquad +\frac{l_J}{l_J+l_{\neg J}} \quad \left.\right)\cdots\left.\right)
\end{array}$$

It is now seen immediately that the sum within each of the nested parentheses is 1, verifying the balance equation. (iv) when $n_i > 0$, we have:

$$\pi(wait(\mathfrak{s}) = \pi(\mathfrak{s}) \frac{\mu_{S_1,\ldots,S_i}}{\lambda_{\mathcal{U}(S_1,\ldots,S_i)}}.$$

which verifies:

$$\mu_{\{S_1,\ldots,S_i\}}\pi(\mathfrak{s}) = \lambda_{\mathcal{U}(\{S_1,\ldots,S_i\})}\pi(wait(\mathfrak{s})), \qquad n_i > 0.$$

22.5 (continued) Calculate the normalizing constant B given by (22.3).

Solution

We start from

$$\pi(S_1, n_1, \ldots, S_i, n_i, S_{i+1}, \ldots, S_J) = B \prod_{j=1}^{i} \frac{\lambda_{\mathcal{U}(\{S_1,\ldots,S_j\})}{}^{n_j}}{\mu_{\{S_1,\ldots,S_j\}}{}^{n_j+1}} \prod_{j=i+1}^{J} \lambda_{C(\{S_j\ldots,S_J\})}{}^{-1}$$

For fixed $S_1, \ldots, S_J$ and i, we obtain adding over $n_j = 0, 1, \ldots$ for $j = 1, \ldots, i$:

$$\pi(S_1, \cdot, \ldots, S_i, \cdot, S_{i+1}, \ldots, S_J) = B \prod_{j=1}^{i} (\mu_{S_1,\ldots,S_j} - \lambda_{\mathcal{U}(S_1,\ldots,S_j)})^{-1} \prod_{j=i+1}^{J} \lambda_{C(\{S_j\ldots,S_J\})}{}^{-1}$$

and the expression for B follows.

22.6 Use the distributional form of Little's law to prove Theorem 22.3 [Visschers et al. (2012)].

Solution

We note that customers of type c arrive in a Poisson stream, their arrival has no influence on future arrivals and on service times of previous arrivals, and they leave in the same order as they arrived. These are exactly the conditions for the distributional form of Littles's law. Hence we have that the LST of the waiting time of type c customers until entry to service can be obtained from the generating function of the number of waiting customers of type c. That is $\mathbb{E}(e^{-\lambda_c W_c(1-z)} = \mathbb{E}(Z^{N_c})$, where λ is the arrival rate, W_c the waiting time, and N_c the number of type c customers in the system. Equivalently, and useful here: $\mathbb{E}(e^{-sW_c}) = \mathbb{E}(\left(\frac{\lambda_c - s}{\lambda_c}\right)^{N_c})$. Note that W_c does not include the actual service time, since N_c only counts waiting customers.

The following calculations are done conditionally on the permutation $S_1, \ldots, S_J$ with $S_{i+1}, \ldots, S_J$ idle. Form the conditional stationary distribution we see the variables N_j, the number of customers between the j and $j+1$ server, are independent, and $N_{j,c}$, those of type c, where $c \in \mathcal{U}(S_1, \ldots, S_j)$ are binomial:

$$N_j \sim \text{Geometric}_0\left(1 - \frac{\lambda_{\mathcal{U}(S_1,\ldots,S_j)}}{\mu_{S_1,\ldots,S_j}}\right), \qquad N_{c,j} \sim \text{Binomial}\left(N_j, \frac{\lambda_c}{\lambda_{\mathcal{U}(S_1,\ldots,S_j)}}\right).$$

If N is geometric with parameter α, and M conditional on N is binomial with parameters (N, θ) then:

$$\mathbb{E}(z^N) = \frac{1 - \frac{\alpha\theta}{1-\alpha(1-\theta)}}{1 - \frac{\alpha\theta}{1-\alpha(1-\theta)}z}$$

i.e. M is itself a geometric random variable, with parameter $\frac{\alpha\theta}{1-\alpha(1-\theta)}$ (probability of failure). Substituting we get the parameter for $N_{c,j}$:

$$\eta_{c,j} = \frac{\lambda_c}{\mu_{S_1,\ldots,S_j} - \lambda_{\mathcal{U}(S_1,\ldots,S_j)} + \lambda_c}$$

and therefore,

$$\mathbb{E}(z^{N_c}|S_1, \ldots, S_i) = \prod_{\substack{j=1 \\ c \in \mathcal{U}(\{S_1,\ldots,S_j\})}}^{i} \frac{1 - \eta_{c,j}}{1 - \eta_{c,j}z}.$$

We now have for each of the internal waiting times

$$\mathbb{E}(e^{-sW_{c,j}}) = \mathbb{E}[\left(\frac{\lambda_c - s}{\lambda_c}\right)^{N_{c,j}}] = \frac{1 - \eta_{c,j}}{1 - \eta_{c,j}\frac{\lambda_c - s}{\lambda_c}} = \frac{\frac{1-\eta_{c,j}}{\eta_{c,j}}\lambda_c}{\frac{1-\eta_{c,j}}{\eta_{c,j}}\lambda_c + s}$$

which is an exponential random variable, with parameter:

$$\frac{1-\eta_{c,j}}{\eta_{c,j}}\lambda_c = \mu_{S_1,\ldots,S_j} - \lambda\mathcal{U}(S_1,\ldots,S_j),$$

The theorem follows.

22.7 Use equation (22.4) to obtain expressions for the first and second moment of the waiting time.

Solution

Recall equation (22.4)

$$\varphi_{W_c}(s) = \mathbb{E}(e^{-sW_c}) = \sum_{\mathcal{P}(J)} \sum_{i=0}^{J} \pi(S_1,\cdot,\ldots,S_i,\cdot,S_{i+1},\ldots,S_J) \prod_{\substack{j=1 \\ c\in\mathcal{U}(\{S_1,\ldots,S_j\})}}^{i} \frac{\mu_{\{S_1,\ldots,S_j\}} - \lambda\mathcal{U}(\{S_1,\ldots,S_j\})}{\mu_{\{S_1,\ldots,S_j\}} - \lambda\mathcal{U}(\{S_1,\ldots,S_j\}) + s},$$

We need to obtain $\mathbb{E}(W_c) = -\frac{d}{ds}\varphi_{W_c}(s)\Big|_{s=0}$ and $\mathbb{E}(W_c) = \frac{d^2}{ds^2}\varphi_{W_c}(s)\Big|_{s=0}$. However, recall that each of the expressions inside the summation is the LST of the sum of independent exponential random variables. Hence:

$$\mathbb{E}(W_c) = \sum_{\mathcal{P}(J)} \sum_{i=0}^{J} \pi(S_1,\cdot,\ldots,S_i,\cdot,S_{i+1},\ldots,S_J) \sum_{\substack{j=1 \\ c\in\mathcal{U}(\{S_1,\ldots,S_j\})}}^{i} \left(\mu_{\{S_1,\ldots,S_j\}} - \lambda\mathcal{U}(\{S_1,\ldots,S_j\})\right)^{-1},$$

For the secon moment it is still easier to calculate the second derivative:

$$\begin{aligned}\mathbb{E}(W_c^2) = \sum_{\mathcal{P}(J)} \sum_{i=0}^{J} \pi(S_1,\cdot,\ldots,S_i,\cdot,S_{i+1},\ldots,S_J) \\ \sum_{\substack{j=1 \\ c\in\mathcal{U}(\{S_1,\ldots,S_j\})}}^{i} \left(\mathbf{2}\left(\mu_{\{S_1,\ldots,S_j\}} - \lambda\mathcal{U}(\{S_1,\ldots,S_j\})\right)^{-2}\right. \\ + \left.\sum_{\substack{k\neq j \\ c\in\mathcal{U}(\{S_1,\ldots,S_k\})}} \left(\mu_{\{S_1,\ldots,S_j\}} - \lambda\mathcal{U}(\{S_1,\ldots,S_j\})\right)^{-1}\left(\mu_{\{S_1,\ldots,S_k\}} - \lambda\mathcal{U}(\{S_1,\ldots,S_k\})\right)^{-1}\right)\end{aligned}$$

22.8 For the infinite bipartite matching model, show that matching of $s^1, s^2, \ldots$, and $c^1, c^2, \ldots$ is unique, and if each type occurs infinitely often it matches all customers and servers [Adan and Weiss (2012)].

Solution

Proposition 22.1. *For every M, N there exists a full FCFS matching, and it is unique*

Proof We prove this by induction on M, N. For $(M, N) = (1, 1)$, if $(c^1, s^1) \in G$, then $A = \{(1, 1)\}$, else $A = \emptyset$. Clearly this is a full FCFS $(1, 1)$ matching, and it is unique.
To prove existence, assume that a unique full FCFS matching exists for (M, N), denoted by A. We will show how to extend it to $(M, N + 1)$. The extension to $(M + 1, N)$ is analogous. We consider s^{N+1} and define

$$i_0 = \arg\min\{i \,:\, 1 \le i \le M,\, i \notin A_c,\, (c^i, s^{N+1}) \in G\},$$

if the set on the right hand side is not empty, and let $\tilde{A} = A \cup \{(i_0, N + 1)\}$. Else, if the set is empty, let $\tilde{A} = A$. It is immediate to see that $\tilde{A}$ is full and FCFS: it is full, since the added s^{N+1} is either matched or has no match, and to check that it is FCFS, we need to check the condition only for s^{N+1}, but all c^i compatible with s^{N+1} with $i < i_0$ are matched to one of $s^1, \ldots, s^N$, so the condition holds.
To prove uniqueness, assume that for all (M', N') with $M' < M$, $N' \le N$ or $M' \le M$, $N' < N$ there is a unique full FCFS matching, and consider M, N. Assume that there are two full FCFS (M, N) matchings and denote them by A, B. Define $\tilde{A}$ by removing s^N, and if $(i, N) \in A$ for some i, then $\tilde{A} = A\backslash\{(i, N)\}$. It is immediate to see that $\tilde{A}$ is a full FCFS matching on $(M, N - 1)$: If s^N was not matched in A, then A and $\tilde{A}$ consist of the same pairs, so there is nothing to show, and if $(i, N) \in A$, then after removal of s^N, customer c^i cannot have any match, since it was previously matched to s^N, and so there is no earlier unmatched server that is compatible with it. Define $\tilde{B}$ analogously. By the induction hypothesis, since both $\tilde{A}$ and $\tilde{B}$ are $(M, N - 1)$ full FCFS matchings, they must coincide, so $\tilde{A} = \tilde{B}$. It remains to consider pairs $(i, N) \in G$, and see that the same ones appear in A and B, to show that $A = B$, and prove the uniqueness. If s^N has no match in either A or B, there is nothing more to show, $A = B$. Assume that $(i_1, N) \in A$. If s^N has no match in B, then c^{i_1} is unmatched, which contradicts the fact that B is full. If $(i_2, N) \in B$, and $i_1 \neq i_2$, then we have a contradiction to the FCFS property. Hence $A = B$ is proved.
That every item is matched if there are infinitely many of each type follows since if s^n is the earliest not matched up to (M, N), let c^m be earliest with $m > M$ that can match it, which exists by assumprion. Then the full match (m, N) will include the match for s_n. □

22.9 Show that the three conditions of the CRP definition (22.5) are equivalent.
Solution
We wish to show that the following conditions for complete resource pooling

(CRP)

$$\text{(i)}\ \beta_{\mathcal{S}(C)} > \alpha_C, \qquad \text{(ii)}\ \alpha_{\mathcal{C}(S)} > \beta_S, \qquad \text{(iii)}\ \beta_S > \alpha_{\mathcal{U}(S)},$$

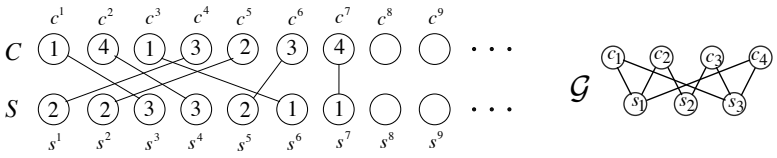
are equivalent.

These conditions essentially say that there is enough service capacity to serve every subset of customer types, and enough service demand to keep each subset of the servers fully occupied. In particular, for every subset of servers there is enough service capacity to serve the unique customers of this subset. We now prove the equivalence: Consider any $C \neq \emptyset, \mathcal{C}$, $S \neq \emptyset, \mathcal{S}$. Assume (i). We have $\mathcal{S}(\mathcal{U}(S)) = S$, so $\beta_S = \beta_{\mathcal{S}(\mathcal{U}(S))} > \alpha_{\mathcal{U}(S)}$ which proves that (i)$\Longrightarrow$(iii). Next, assume (iii). $1-\beta_S = \beta_{\overline{S}} > \alpha_{\mathcal{U}(\overline{S})} = \alpha_{\overline{\mathcal{C}(S)}} = 1-\alpha_{\mathcal{C}(S)}$, which proves (iii)$\Longrightarrow$(ii). So we have shown that (i)$\Longrightarrow$(ii). But the roles of s_j, β_{s_j} and of c_i, α_{c_i} are completely interchangeable. Hence (ii)$\Longrightarrow$(i) follows from (i)$\Longrightarrow$(ii).

22.10 Write down the first 7 states of the processes $\mathring{Y}(n)$ and $\mathring{Z}(n)$, for the matching in Figure 22.2.

Solution

We recall the figure:



For $\mathring{Y}(n)$:

$$\mathring{Y}(1) = (s_2, s_2, \tilde{c}_1),\ \mathring{Y}(2) = (s_2, s_2, \tilde{c}_1, \tilde{c}_4),\ \mathring{Y}(3) = (s_2, s_2, \tilde{c}_1, \tilde{c}_4, s_2, \tilde{c}_1),$$
$$\mathring{Y}(4) = (s_2, \tilde{c}_1, \tilde{c}_4, s_2, \tilde{c}_1),\ \mathring{Y}(5) = (s_2, \tilde{c}_1),\ \mathring{Y}(6) = \emptyset,\ \mathring{Y}(7) = \emptyset.$$

For $\mathring{Z}(n)$:

$$\mathring{Z}(1) = (c_1; z_2),\ \mathring{Z}(2) = (c_1, c_4; s_2, s_2),\ \mathring{Z}(3) = (c_4, c_1; s_2, s_2, \tilde{c}_1),$$
$$\mathring{Z}(4) = (c_1, \tilde{s}_2; s_2, \tilde{c}_1, \tilde{c}_4),\ \mathring{Z}(5) = (c_1, \tilde{s}_2, \tilde{s}_2; s_2),\ \mathring{Z}(6) = \emptyset,\ \mathring{Z}(7) = \emptyset.$$

22.11 For the infinite bipartite matching model, show that $\mathring{X}(n) = (x_1, \ldots, x_L)$ is a possible state of the process $\mathring{X}$ if and only if: for any $1 \leq k < l \leq L$, if $x_k = c_i$ and $x_l = \tilde{s}_j$ then $(c_i, s_j) \notin \mathcal{G}$ with a similar characterization for possible states of $\mathring{Y}$ and $\mathring{Z}$ [Adan et al. (2018b)].

Solution

Proof of only if: Assume that $1 \leq k < l \leq L$, if $x_k = c_i$ and $x_l = \tilde{s}_j$ and that $(c_i, \tilde{s}_j) \in \mathcal{G}$. Then we could have had the earlier match between these two, which contradicts FCFS.

Proof of if: Assume we have a state $\mathring{X}(N) = x_1, \ldots, x_L$. Assume in this

sequence all the matched and exchanged servers are in positions $m_1 < m_2 < \cdots < m_K$, and they are $\tilde{s}_1, \ldots, \tilde{s}_K$. Consider the sequences: $c^1, \ldots, c^L$, $s^1, \ldots, s^K$, where $s^j = \tilde{s}_j$, $j = 1 \ldots, K$, and $c^i = x_i$, $i \notin (m_1, \ldots, m_K)$, and $c^{m_j} = c_j$, $j = 1, \ldots, K$, where $(\tilde{s}_j, c_j) \in \mathcal{G}$. For this initial state of the sequences, under server by server scheduling, by the condition of *if*, $s^1 = \tilde{s}_1$ is incompatible with $x_1, \ldots, x_{m_1-1}$ and will match FCFS with $c^{m_1} = c_1$, etc. so for this initial sequence, $\mathring{X}(K) = x_1, \ldots, x_L$.

22.12 Show how to construct $\mathring{Z}(n)$ from $\mathring{X}(n)$ and $\mathring{Y}(n)$, and show that if $\mathring{Z}(n) = ((x_1, \ldots, x_L), (y_1, \ldots, y_K))$ then $(x_1, \ldots, x_L, y_K, \ldots, y_1)$ is a possible state of $\mathring{X}(n)$.

Solution

Let $\mathring{X}(N) = (c^{\underline{N}}, \ldots, z^N, \ldots, \tilde{s}^{\overline{N}})$, $\mathring{Y}(N) = (s^{\underline{M}}, \ldots, w^N, \ldots, \tilde{c}^{\overline{M}})$, Then: $\mathcal{Z}(N) = ((c^{\underline{N}}, \ldots, z^N), (s^{\underline{M}}, \ldots, w^N))$. To see this, note that in server by server matching up to N, what we do to the original $c^1, \ldots, c^N$ is that we match all c^n, $n < N$ that could be matched by servers up to N, but no others, which gives us $(c^{\underline{N}}, \ldots, z^N)$, and in customer by customer matching up to N, what we do to the original $s^1, \ldots, s^N$ is that we match all s^n, $n < N$ that could be matched by customers up to N, but no others, which gives us $(s^{\underline{M}}, \ldots, w^N)$.

Consider $\mathring{Z}(n) = ((x_1, \ldots, x_L), (y_1, \ldots, y_K))$, then $(x_1, \ldots, x_L)$ is a beginning of $\mathring{X}(N)$, and $(y_1, \ldots, y_K)$ is a beginning of $\mathring{Y}(N)$. So for any $u < v$:

(i) for $x_u = c_i$, $x_v = \tilde{s}_j$ then $(c_i, \tilde{s}_j) \notin \mathcal{G}$, as part of $\mathring{X}(N)$.

(ii) for $y_u = s_j$, $y_v = \tilde{c}_i$ then $(s_j, \tilde{c}_j) \notin \mathcal{G}$, as part of $\mathring{Y}(N)$.

(iii) for $x_k = c_i$, $y_l = s_j$, then $(c_i, s_j) \notin \mathcal{G}$, by definition of $\mathring{Z}(N)$.

This implies that in all cases, for any $u < v$ if the u item and the v item in $(x_1, \ldots, x_L, y_K, \ldots, y_1)$ are a customer and server respectively, then they must be incompatible, so it is a possible state of $\mathring{X}(N)$.

22.13 Prove the subadditivity property of FCFS matching, Proposition 22.11.

Solution

We prove first that if in the FCFS matching of $A = (c^1, \ldots, c^M)$ with $B = (s^1, \ldots, s^N)$ there are K unmatched customers and L unmatched servers, then in the FCFS matching of $c^0, c^1, \ldots, c^M$ with $s^1, \ldots, s^N$ there are no more than $K+1$ unmatched customers and no more than L unmatched servers. In the matching of (c^0, A) and B, if c^0 has no match, then all the other links in the matching are the same as in the matching of A and B, so the total number of unmatched customers is $K + 1$ and unmatched servers is L. If c^0 is matched to s^n and s^n is unmatched in the matching of A and B, then (c^0, s^n) is a new link and all the other links in the matching of (c^0, A) and B are the same as in the matching of A, B, so the total number of unmatched customers is K and unmatched servers is $L - 1$.

If c^0 is matched to s^{n_1} and s^{n_1} was matched to c^{m_1} in the A, B matching, then (c^0, s^{n_1}) is a new link, and the link (s^{n_1}, c^{m_1}) in the A, B matching is disrupted. We now look for a match for c^{m_1} in the matching of (c^0, A) and

B. Clearly, c^{m_1} is not matched to any of s^j, $j < n_1$, since in the construction of the A, B matching c^{m_1} was not matched to any of those. So c^{m_1} will either remain unmatched, or it will be matched to some s^{n_2}, where $n_2 > n_1$. In the former case, all the links of the A, B matching except (s^{n_1}, c^{m_1}) remain unchanged in the (c^0, A) and B matching, and so the numbers of unmatched items remain $K + 1$ and L. In the latter case, there are again two possibilities: If s^{n_2} is unmatched in the A, B matching, then the $(c^0, A), B$ matching will have disrupted one link and added 2 links retaining all other links of the A, B matching, so the numbers of unmatched items are K and $L - 1$. If s^{n_2} is matched to c^{m_2} in the A, B matching, then the link s^{n_2}, c^{m_2} is disrupted, and we now look for a match for c^{m_2} in the $(c^0, A), B$ matching. Similar to c^{m_1}, either c^{m_2} remains unmatched, resulting in $K + 1$ and L unmatched items in the $(c^0, A), B$ matching, or, by the same argument as before, c^{m_2} will be matched to s^{n_3}, where $n_3 > n_2$. Repeating these arguments for any additional disrupted links, we conclude that we either end up with one more link, so the number of unmatched items are K and $L - 1$, or we have the same number of links and the number of unmatched items are $K + 1$ and L.

We now consider matching of A' with B', of A'' with B'', and of $A'A''$ with $B'B''$. Assume in the matching of A' with B' that $c^1, \ldots, c^K$ and $s^1, \ldots, s^L$ are unmatched. The number of unmatched in $A'A''$ with $B'B''$ is the same as in the match of $(c^1, \ldots, c^K), A''$ with $(s^1, \ldots, s^L), B''$. We now add them one by one, from last to first, and by the proof above, at each step the number of unmatched either remains the same, or it decreases by 1 for both customers and servers.

22.14 Prove the monotonicity result of Proposition 22.12.

Solution

proof: This follows directly from the subadditivity. Consider the blocks of customers $0 \le n \le M_0$ and the block $M_0 + 1 \le n \le M_1$. The second block is perfectly matched so it has 0 unmatched. So the union of the two blocks has no more unmatched than the first block, and thatmeans $N_0 \ge N_1$, and by the same argument $N_1 \ge N_2$ etc.

22.15 For an incompatible pair (c^0, s^0), construct $(c^0, c^1, \ldots, c^h, s^0, s^1, \ldots, s^h)$, that are perfectly matched by FCFS, and find a lower bound for the probability of such a sequence, to prove Proposition 22.13.

Solution

proof: Because the bipartite graph is connected, and there is no direct edge between c^0, s^0, there exists a simple path (i.e. with no repeated nodes), $c^0 \to s_{j_1} \to c_{i_1} \cdots \to s_{j_h} \to c_{i_h} \to s^0$ which connects them, with $1 \le h \le \min\{I, J\} - 1$. Clearly, the FCFS matching of $c^0, c^1, \ldots, c^h, s^0, s^1, \ldots, s^h$, where $c^l = c_{i_l}, s^l = s_{j_l}$, $l = 1, \ldots, h$ is perfect, with exactly the links of the path, where c^0 is matched to s^1, and s^0 is matched to c^h. Note that FCFS matching of $c^1, \ldots, c^h, s^1, \ldots, s^h$ consists of h perfectly matched blocks of length one.

22.16 Show that the exchange transformation on a perfectly matched block will retain the same links if we now do FCFS in reversed time, as stated in Proposition (22.16).

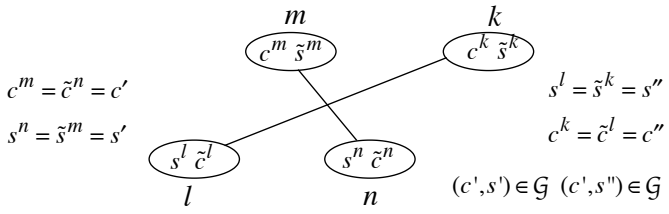
Solution

A necessary and sufficient condition for FCFS perfect matching of a block is that:

(i) If $c^m = c'$ and $s^n = s'$ are matched then for all $l < n$ If $s^l = s''$ is compatible with c' then we cannot have s^l matched with c^k where $k > m$.

(ii) If $c^n = c'$ and $s^m = s'$ are matched then for all $l < n$ If $c^l = c''$ is compatible with s' then we cannot have c^l matched with s^k where $k > m$.

But then when we do the exchange transformation we get that these conditions now hold for the match $\tilde{c}^n = c'$ with $\tilde{s}^m = s'$. The proof for condition (i) is illustrated in the following figure:



22.17 Prove the uniqueness theorem, 22.18 for the ridesharing model [Adan et al. (2018a)].

Solution

The proof is very similar to the proof of Theorem 22.5 for the symmetric bipartite matching system. There is subadditivity, monotonicity, forward coupling and backward coupling. Details are in Adan et al. (2018a)

22.18 Prove the time reversal theorem, 22.19 for the ridesharing model.

Solution

The proof is very similar to the proof of Theorem 22.7, for the symmetric bipartite matching system. The somewhat surprising part of the theorem is that time reversal leaves unmatched servers in their place, unmatched. For the proof one shows that the exchange transformation of a perfectly matched block gives the same links for directed FCFS in the reversed time direction, and that the probability of a perfectly matched block equals that of the reversed block. Next one regards the conditional process, conditioned on being empty at time 0. One then has that this process of perfectly matched blocks has the reversal property. Therefore, by uniqueness of Palm measure the reversal theorem also holds for the unconditional system. Details are in Adan et al. (2018a).

22.19 Verify the Bernoulli type stationary distributions, Theorem 22.20 for the ridesharing model [Weiss (2020)].

Solution

The proof is similar to the proof of Theorem 22.9, for the symmetric bipartite matching system. One defines the detailed matching Markov chain that includes the list of unmatched and of matched and exchanged items in the sequence. By the reversal theorem we then have both the forward and the backward transition rates, and on can then use Kelly's Lemma to verify that the stationary distribution is multi-Bernoulli. An important part in the proof is the characterization of the possible states. Details are in Weiss (2020).

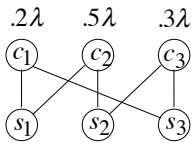
22.20 Prove the ergodicity condition of (22.21) and verify equations (22.14)–(22.17) [Weiss (2020)].

Solution

For details for these calculations see Weiss (2020).

22.21 The following table has data for a system with 3 types of customers and 3 types of servers. Calculate matching rates, and provide designs for ED with $W = 1$, for QD with $T = 0.5$, and for QED, for $\lambda = 20, 50, 100, 200$ and simulate the systems to evaluate the performance [Adan et al. (2019)].

Example – System and Data



Patience time distributions	
	H_{c_i}
c_1	Exp(0.1)
c_2	U(0,10)
c_3	Exp(0.2)

Service time distributions			
G_{c_i,s_j}	c_1	c_2	c_3
s_1	Pareto(2, 3)	Exp(0.125)	
s_2		Exp(0.2)	U(2, 6)
s_3	Pareto(3, 3)		U(1, 5)

Only the *mean* service times are used by the design algorithms. The full distributions are used in the simulations.

Resource allocation design parameters are: $\beta_{s_1} = 0.3$, $\beta_{s_2} = 0.3$, $\beta_{s_3} = 0.4$.

References

Abate, J.C., Gagan, L., and Whitt, W. 1995. Calculating the M/G/1 busy-period density and LIFO waiting-time distribution by direct numerical transform inversion. *Operations Research Letters*, **18**(3), 113–119.

Adan, I., and Weiss, G. 2005. A two-node Jackson network with infinite supply of work. *Probability in the Engineering and Informational Sciences*, **19**(02), 191–212.

Adan, I., and Weiss, G. 2006. Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy. *Queueing Systems*, **54**(3), 169–183.

Adan, I., and Weiss, G. 2012. Exact FCFS matching rates for two infinite multitype sequences. *Operations Research*, **60**(2), 475–489.

Adan, I., and Weiss, G. 2014. A skill based parallel service system under FCFS-ALIS – steady state, overloads, and abandonments. *Stochastic Systems*, **4**(1), 250–299.

Adan, I., Wessels, J., and Zijm, H. 1990. Analysis of the symmetric shortest queue problem. *Stochastic Models*, **6**, 691–713.

Adan, I., Wessels, J., and Zijm, H. 1993. Matrix-geometric analysis of the shortest queue problem with threshold jockeying. *Operations Research Letters*, **13**(2), 107–112.

Adan, I., van Houtum, G. J., and van der Wal, J. 1994. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research*, **48**(2), 197–217.

Adan, I., Kleiner, I., Righter, R., and Weiss, G. 2018a. FCFS parallel service systems and matching models. *Performance Evaluation*, **127**, 253–272.

Adan, I., Busic, A., Mairesse, J., and Weiss, G. 2018b. Reversibility and further properties of FCFS infinite bipartite matching. *Mathematics of Operations Research*, **43**(2), 598–621.

Adan, I., Boon, M., and Weiss, G. 2019. Design heuristic for parallel many server systems. *European Journal of Operational Research*, **273**(1), 259–277.

Baccelli, F., and Hebuterne, G. 1981. *On queues with impatient customers*. Tech. rept. RR-0094 inria-00076467. INRIA Rapports de Recherthe.

Borovkov, A.A. 1965. Some limit theorems in the theory of mass service, II multiple channels systems. *Theory of Probability & Its Applications*, **10**(3), 375–400.

Botvich, D. D., and Zamyatin, A. A. 1992. Ergodicity of conservative communication networks. *Rapport de Recherche, INRIA*, **1772**.

Bramson, M. 1996. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems*, **23**(1-4), 1–26.

Bramson, M. 2001. Stability of earliest-due-date, first-served queueing networks. *Queueing Systems*, **39**(1), 79–102.

Bramson, M. 2008. *Stability of Queueing Networks*. Springer.

Brandt, A., Franken, P., and Lisek, B. 1990. *Stationary stochastic models*. Vol. 227. Wiley.

Breiman, L. 1992. *Probability*. SIAM.

Browne, S., Whitt, W., and Dshalalow, J.H. 1995. Piecewise-linear diffusion processes. *Advances in queueing: Theory, methods, and open problems*, **4**, 463–480.

Chen, H., and Mandelbaum, A. 1991. Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Annals of Probability*, 1463–1519.

Chen, H., and Mandelbaum, A. 1994. Hierarchical modeling of stochastic networks, Part II: Strong approximations. Pages 107–131 of: *Stochastic Modeling and Analysis of Manufacturing Systems*. Springer.

Cohen, J.W. 1982. *The Single Server Queue*. North-Holland.

Dai, J.G., and Lin, W. 2005. Maximum pressure policies in stochastic processing networks. *Operations Research*, **53**(2), 197–218.

Dai, J.G., and Weiss, G. 1996. Stability and instability of fluid models for reentrant lines. *Mathematics of Operations Research*, **21**(1), 115–134.

Deimling, K. 2006. *Ordinary differential equations in Banach spaces*. Vol. 596. Springer.

Doob, J.L. 1953. *Stochastic processes*. Wiley.

Doshi, B.T. 1986. Queueing systems with vacations?a survey. *Queueing systems*, **1**(1), 29–66.

Down, D.G, Gromoll, H.C., and Puha, A.L. 2009. Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research*, **34**(4), 880–911.

Dubins, L.E. 1968. On a theorem of Skorohod. *Annals of Mathematical Statistics*, **39**(6), 2094–2097.

Gelenbe, E. 1991. Product-form queueing networks with negative and positive customers. *Journal of applied probability*, 656–663.

Goodman, J.B., and Massey, W.A. 1984. The non-ergodic Jackson network. *Journal of Applied Probability*, **21**(4), 860–869.

Gromoll, H.C. 2004. Diffusion approximation for a processor sharing queue in heavy traffic. *Annals of Applied Probability*, **14**(2), 555–611.

Gromoll, H.C., Puha, A.L., and Williams, R.J. 2002. The fluid limit of a heavily loaded processor sharing queue. *Annals of Applied Probability*, **12**(3), 797–859.

Guo, Y., Lefeber, E., Nazarathy, Y., Weiss, G., and Zhang, H. 2014. Stability of multiclass queueing networks with infinite virtual queues. *Queueing Systems*, **76**(3), 309–342.

Harrison, J.M. 1985. *Brownian Motion and Stochastic Flow Systems*. Wiley.

Harrison, J.M. 2013. *Brownian models of performance and control*. Cambridge University Press.

Haviv, M. 2013. *Queues: A Course in Queueing Theory*. Springer.

Iglehart, D.L., and Whitt, W. 1970a. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, **2**(1), 150–177.

Iglehart, D.L., and Whitt, W. 1970b. Multiple channel queues in heavy traffic. II. *Advances in Applied Probability*, **2**(2), 355–369.

Kelly, F.P. 1979. *Reversibility and Stochastic Networks*. Wiley, Reprinted Cambridge University Press, 2011.

Kendall, D.G. 1964. Functional equations in information theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **2**(3), 225–229.

Kiefer, J., and Wolfowitz, J. 1955. On the theory of queues with many servers. *Transactions of the American Mathematical Society*, **78**(1), 1–18.

Kopzon, A., Nazarathy, Y., and Weiss, G. 2009. A push–pull network with infinite supply of work. *Queueing Systems*, **62**(1-2), 75–111.

Krichagina, E.V., and Puhalskii, A.A. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems*, **25**(1-4), 235–280.

Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., and Shmoys, D.B. 1993. Sequencing and scheduling: Algorithms and complexity. *Handbooks in Operations Research and Management Science*, **4**, 445–522.

Laws, C.N. 1990. *Dynamic routing in queueing networks*. Ph.D. thesis, Cambridge University.

Laws, C.N. 1992. Resource pooling in queueing networks with dynamic routing. *Advances in Applied Probability*, **24**(3), 699–726.

Lipton, Alexander, and Kaushansky, Vadim. 2018. On the first hitting time density of an Ornstein-Uhlenbeck process. *arXiv preprint arXiv:1810.02390*.

Loynes, R.M. 1962. The stability of a queue with non-independent inter-arrival and service times. Pages 497–520 of: *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58. Cambridge Univ Press.

Mitzenmacher, M. 1996. The power of two choices in randomized load balancing. *PhD thesis, University of California at Berkeley*.

Mitzenmacher, M. 2001. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, **12**(10), 1094–1104.

Nazarathy, Y., and Weiss, G. 2010. Positive Harris recurrence and diffusion scale analysis of a push pull queueing network. *Performance Evaluation*, **67**(4), 201–217.

Nelson, Randolph, and Tantawi, Asser N. 1988. Approximate analysis of fork/join synchronization in parallel queues. *IEEE transactions on computers*, **37**(6), 739–743.

Obłój, J. 2004. The Skorokhod embedding problem and its offspring. *Probability Surveys*, **1**, 321–392.

Schrage, L.E., and Miller, L.W. 1966. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, **14**(4), 670–684.

Servi, L.D., and Finn, S.G. 2002. M/M/1 queues with working vacations (m/m/1/wv). *Performance Evaluation*, **50**(1), 41–52.

Sevcik, K.C., and Mitrani, I. 1981. The distribution of queuing network states at input and output instants. *Journal of the ACM (JACM)*, **28**(2), 358–371.

Visschers, J., Adan, I., and Weiss, G. 2012. A product form solution to a system with multi-type jobs and multi-type servers. *Queueing Systems*, **70**(3), 269–298.

Vvedenskaya, N., Dobrushin, R., and Karpelevich, F. 1996. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, **32**(1), 20–34.

Ward, A.R. 2012. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science*, **17**(1), 1–14.

Weber, R.R. 1978. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, **15**(2), 406–413.

Wein, L.M. 1990. Optimal control of a two-station Brownian network. *Mathematics of Operations Research*, **15**(2), 215–242.

Wein, L.M. 1992. Scheduling networks of queues: heavy traffic analysis of a multistation network with controllable inputs. *Operations Research*, **40**(3-supplement-2), S312–S334.

Weiss, G. 2005. Jackson networks with unlimited supply of work. *Journal of Applied Probability*, **42**(3), 879–882.

Weiss, G. 2020. Directed FCFS infinite bipartite matching. *Queueing Systems*, 1–32.

Whitt, W. 1986. Deciding which queue to join: Some counterexamples. *Operations Research*, **34**(1), 55–62.

Winston, W. 1977. Optimality of the shortest line discipline. *Journal of Applied Probability*, **14**(1), 181–189.

Wolff, R.W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice Hall.

Zeltyn, S., and Mandelbaum, A. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n+ G queue. *Queueing Systems*, **51**(3-4), 361–402.