3 - Intermediate Queueing Theory

Ing. Daniele Venturi

Dipartimento INFOCOM
Università di Roma “Sapienza”

Outline

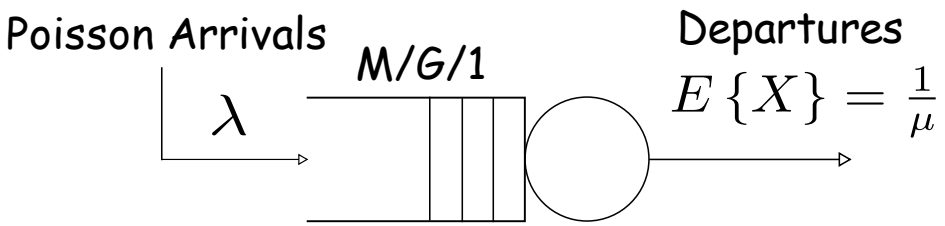
Queues with Non-Exponential Service Time - 1

- All the queues analysed so far are queues where the service time is *exponentially distributed*.
- The exponential distribution is particulary easy to handle in analytical modelling because of its *memory-less* property.
- As we have seen, the key point is that if a job in service is examined at any time while its service is continuing, the distribution of the remaining service time will still be exponentially distributed with the same mean value.
- Hence we can define the system state at any time t by using just a single variable, i. e. the number (of users) in the system at time t.

Queues with Non-Exponential Service Time - 2

- Unfortunately the assumption of exponential distribution cannot be justified in all systems.
- The system state at an arbitrary time instant t would consist of both the number in the system at time t as well as the residual service time for each customer currently in service.
- As a consequence the system is much harder to analyse.
- Exact analytical modelling is however possible for generally distributed service times in the case of single server queues with infinite buffer, i. e. M/G/1/∞ queues (M/G/1 in the following).

Assumptions



- FCFS discipline.
- The arrival process is considered Poisson with average arrival rate λ. We denote with $A(t)$ and $a(t)$, respectively, the *cumulative distribution function* and the *probability density function* of the interarrival times.
- The service time X is generally distributed with mean $E\{X\} = \overline{X}$. For $X = t$ $(t \geq 0)$, the cumulative distribution function is $B(t)$ and the probability density function is $b(t)$.

The Residual Service Time

- Consider a particular arrival of interest that enters the queue; if the queue is non-empty it sees:
  - One or more customers waiting in the queue.
  - A customer currently in service with r seconds of *residual service time*.
- Let $E\{r\} = R$. Since the arrival process is Poisson in nature, PASTA will hold, so that using Little's result:

Mean Time Spent in Queue

$$W_q = \frac{R}{1-\rho} \qquad \text{being } \rho = \lambda E\{X\} \ (< 1 \text{ for stability}).$$

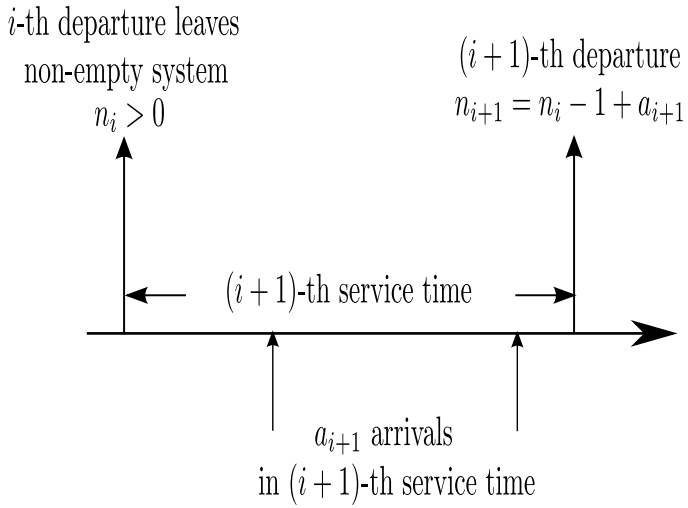
The System State...

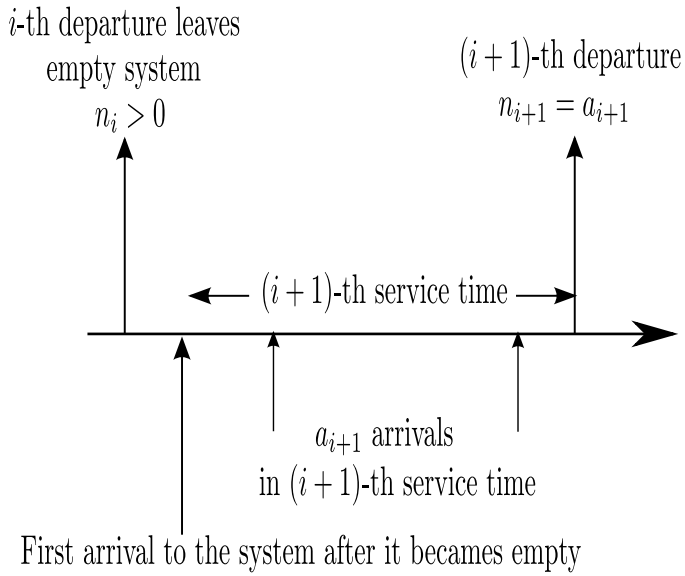
- The main idea is to choose carefully the time points where the system state is observed, so that the Markov property is satisfied and a single state system description will work.
- One such set of time points are the time instants *just after* the departure of a customer following service. We denote with t_i ($i = 1, 2, 3, \cdots, \infty$) the time instants where the i-th customers departs from the system.
- At a time instant t_i we define the system state n_i to be the number of customers left behind when the i-th customer departs.

...and its Evolution

- Let a_{i+1} be the number of arrivals in the $(i+1)$-th sevice time. We have to consider two cases:

$$n_{i+1} = n_i - u_{-1}(n_i) + a_{i+1} = \begin{cases} a_{i+1} & \text{for } n_i = 0 \\ n_i - 1 + a_{i+1} & \text{for } n_i = 1, 2, \cdots \end{cases}$$





State Distribution and Moments at Equilibrium

- The previous equation may be used to obtain the probabilities of the system states as observed by a departing customer.
- At first sight these probabilities seem to be not useful, since they hold only for the departure instant. However:
  - One can show that for systems where the system state can change at most by ± 1, the system distribution as seen by an arriving customer will be the same as that seen by a departing customer (*Kleinrock's result*).
  - Since the arrival process is Poisson, PASTA will hold.
- Thus the results obtained for the departure instants will also be the time averaged (ergodic) results at equilibrium.

The P-K Transform Equation - 1

Taking the expectations of both sides of the system state equation yelds $E\{u_{-1}(n_i)\} = E\{a_{i+1}\}$, with

$$E\{u_{-1}(n_i)\} = 1 - p_0 \qquad E\{a_{i+1}\} = \int_0^\infty (\lambda t) b(t) dt = \lambda \overline{X} = \rho,$$

being p_0 the probability of the system being empty. Thus $p_0 = 1 - \rho$, as expected. Now let $\mathcal{G}_{n_j}(z)$ be the generating function of the state as seen by the j-th departing customer; since future arrivals does not depend on the current state:

$$\mathcal{G}_{n_i}(z) = E\{z^{n_i}\} = \sum_{k=0}^{+\infty} z^k Pr\{n_i = k\}$$

$$\Rightarrow \quad \mathcal{G}_{n_{i+1}}(z) = E\left\{z^{n_i - u_{-1}(n_i) + a_{i+1}}\right\} = E\left\{z^{n_i - u_{-1}(n_i)}\right\} E\{z^{a_{i+1}}\}.$$

The P-K Transform Equation - 2

Since we are at equilibrium, we can drop the dependence on i. Denote with $\mathcal{G}_a(z) = E\{z^a\}$ the generating function of the number arriving in a service time; a simple calculation yelds

$$\mathcal{G}_a(z) = \int_0^{+\infty} E\{z^a \mid \text{service time} = t\}\, b(t)dt =$$
$$= \int_0^{+\infty} e^{-\lambda t(1-z)} b(t)dt = \mathcal{L}_B(\lambda - \lambda_z),$$

being

$$\mathcal{L}_B(s) = \int_0^{+\infty} e^{-st} b(t)dt$$

the Laplace transform of $b(t)$.

The P-K Transform Equation - 3

Manipulating the expression derived for $\mathcal{G}_{n_{i+1}}(z)$ (dropping the dependence on i) we get

$$\mathcal{G}_n(z) = \mathcal{G}_a(z) E\left\{z^{n-u_{-1}(n)}\right\} = \mathcal{G}_a(z)\left[\frac{1}{z}\mathcal{G}_n(z) - \frac{1}{z}p_0(1-z)\right].$$

Hence

Pollaczeck-Khinchin Transform Equation

$$\mathcal{G}_n(z) = \frac{(1-\rho)(1-z)\mathcal{G}_a(z)}{\mathcal{G}_a(z) - z} = \frac{(1-\rho)(1-z)\mathcal{L}_B(\lambda - \lambda z)}{\mathcal{L}_B(\lambda - \lambda z) - z}.$$

Performance Evaluation - 1

At this point the worst is over. First of all, using the properties of both generating functions and Laplace transforms we can write

$$\mathcal{G}_a^{(1)}(z) = -\lambda \mathcal{L}_B^{(1)}(\lambda - \lambda z) \quad \Rightarrow \quad \mathcal{G}_a^{(1)}(1) = -\lambda \mathcal{L}_B^{(1)}(0) = \lambda \overline{X} = \rho$$
$$\mathcal{G}_a^{(2)}(z) = \lambda^2 \mathcal{L}_B^{(2)}(\lambda - \lambda z) \quad \Rightarrow \quad \mathcal{G}_a^{(2)}(1) = \lambda^2 \mathcal{L}_B^{(2)}(0) = \lambda^2 \overline{X^2}.$$

On the other hand using L'Hospital's rule we get

$$\mathcal{G}_n^{(1)}(1) = \lim_{z \to 1} \mathcal{G}_n^{(1)}(z) = \lim_{z \to 1} \frac{(1-\rho)\left[(1-z)\mathcal{G}_a^{(1)}(z) - \mathcal{G}_a(z)\right]}{\mathcal{G}_a^{(1)}(z) - 1} =$$
$$= -\frac{1-\rho}{\rho - 1} = 1.$$

Performance Evaluation - 2

Finally differentiating the P-K Transform Equation (up to the second order), putting $z = 1$ and using earlier results we get:

$$-2(1-\rho)\mathcal{G}_n^{(1)}(1) + \lambda^2\overline{X^2} = -2\rho(1-\rho).$$

Lastly, using the properties of the generating functions and Little's result we are done:

$$N = \mathcal{G}_n^{(1)}(1) = \rho + \frac{\lambda^2\overline{X^2}}{2(1-\rho)} \qquad W = \frac{N}{\lambda} = \overline{X} + \frac{\lambda\overline{X^2}}{2(1-\rho)}$$

$$W_q = W - \overline{X} = \frac{\lambda\overline{X^2}}{2(1-\rho)} = \frac{R}{1-\rho} \qquad N_q = \lambda W_q = \frac{\lambda^2\overline{X^2}}{2(1-\rho)}.$$

Distributions of Time Spent in System - 1

Now we are able to derive the distribution of time spent waiting in queue and total time spent in system by a customer when the queue is FCFS.
Consider the n-th arrival to the queue; it waits in the queue for a time interval Q_n before its service can start ($Q_n = 0$ if the arrival enters an empty queue). Once service starts, the customer engages the server for the duration of a service time X_n. Hence the total time spent in the system is $T_n = Q_n + X_n$.
Denote with $f_Q(t)$ ($f_T(t)$) the probability density function of the queueing (total) delay Q_n (T_n) spent in system by the n-th arrival, with Laplace transform $\mathcal{L}_Q(s)$ ($\mathcal{L}_T(s)$).

Distributions of Time Spent in System - 2

Since the queue is FCFS in nature, the number of customers that the n-th user will see left behind when he departs will be the number of arrivals that occur while it is in the system. The generating function for this random number is $\mathcal{G}_n(z)$, hence we can write:

$$\mathcal{G}_n(z) = \int_0^{+\infty} E\left\{z^n \mid \text{total time spent in the system} = t\right\} f_T(t)dt =$$

$$= \int_0^{+\infty} e^{-\lambda t(1-z)} f_T(t)dt = \mathcal{L}_T(\lambda - \lambda_z),$$

and using the P-K Transform Equation yelds

$$\mathcal{L}_T(s) = \frac{s(1-\rho)\mathcal{L}_B(s)}{s - \lambda + \lambda\mathcal{L}_B(s)} \qquad \mathcal{L}_Q(s) = \frac{\mathcal{L}_T(s)}{\mathcal{L}_B(s)} = \frac{s(1-\rho)}{s - \lambda + \lambda\mathcal{L}_B(s)}.$$

The M/D/1 Queue - 1

- The M/D/1 queue has Poisson arrivals like the M/G/1 queue, but the service times are fixed, i. e. deterministic.
- This is a convenient way to model an *Asynchronous Transfer Mode* (ATM) node with fixed size cells as the jobs requiring service or a packet switching node in a computer network where the packets are of fixed size.
- Since the service time is deterministic, denoted with m the (fixed) duration of service, we have $b(t) = u_0(t - m)$ so that

$$\mathcal{L}_B(s) = e^{-sm} \qquad \rho = \lambda m$$
$$\overline{X} = m \qquad \overline{X^2} = m^2.$$

The M/D/1 Queue - 1

Hence the analysis is straightforward:

M/D/1 Queue Performance

$$N = \rho + \frac{\rho^2 m^2}{2(1-\rho)} \qquad W = \frac{m(2-\rho)}{2(1-\rho)}$$

$$W_q = \frac{m\rho}{2(1-\rho)} \qquad N_q = \frac{\rho^2 m^2}{2(1-\rho)}.$$

$$\mathcal{G}_n(z) = \frac{(1-\rho)(1-z)e^{-\rho(1-z)}}{e^{-\rho(1-z)} - z}.$$

Exercise 1

For a particular M/G/1 queue, the Laplace transform of the service time is given by

$$\mathcal{L}_B(s) = \frac{0.5s(\mu_1 + \mu_2) + \mu_1\mu_2}{(s + \mu_1)(s + \mu_2)}.$$

Analyse the queue using the imbedded Markov chain approach.

Exercise 1 - Solution

The evaluation of the mean results on the mean queueing parameters W, W_q, N and N_q is straightforward, since it suffices to compute $\overline{X}$ and $\overline{X^2}$. Manipulating:

$$\overline{X} = -\left(\frac{d}{ds}\mathcal{L}_B(s)\right)_{s=0} = \frac{0.5}{\mu_1} + \frac{0.5}{\mu_2}$$

$$\begin{aligned}\overline{X^2} &= \left(\frac{d^2}{ds^2}\mathcal{L}_B(s)\right)_{s=0} \\ &= \left(\frac{s^3\mu_2 + 3s\mu_1^2\mu_2 + \mu_1^3\mu_2 + \mu_1(s^3 + 6s^2\mu_2 + 3s\mu_2^2 + \mu_2^3)}{(s+\mu_1)^3(s+\mu_2)^3}\right)_{s=0} \\ &= \frac{\mu_1^3\mu_2 + \mu_1\mu_2^3}{\mu_1^3\mu_2^3}.\end{aligned}$$

Exercise 1 - Solution

Finally using the P-K Transform Equation we can easily evaluate the generating function of the number in the system:

$$\mathcal{G}_n(z) = \frac{(1-\rho)(1-z)\mathcal{L}_B(\lambda - \lambda z)}{\mathcal{L}_B(\lambda - \lambda z) - z}.$$

Hence we can conclude:

$$\mathcal{L}_T(s) = \frac{s(1-\rho)\mathcal{L}_B(s)}{s - \lambda + \lambda\mathcal{L}_B(s)} \qquad \mathcal{L}_Q(s) = \frac{s(1-\rho)}{s - \lambda + \lambda\mathcal{L}_B(s)}.$$

Note that inverting the Laplace Transforms $\mathcal{L}_T(s)$, $\mathcal{L}_Q(s)$ will give the associated probability density functions $f_T(t)$ and $f_Q(t)$.

For Further Reading I

S. K. Bose
An Introduction to Queueing Systems.
Kluwer Academic/Plenum Publishers, 2002.

R. B. Cooper
Introduction to Queueing Theory.
Elsevier North Holland, 1981.

L. Kleinrock
Queueing Systems. Volume 1: Theory.
Wiley-Interscience, 1975.