CS 590R: Algorithms for Communications Networks	March 4th, 2003
Lecture 15	
Lecturer: Gopal Pandurangan	Scribe: Amit Shirsat

Queueing Theory

1 Introduction

Queuing theory is the primary methodological framework for analyzing network delay. It often involves simplyifying assumptions since more realistic assumptions make analysis extremely difficult. Nevertheless, the Queueing models which will be discussed in the course, are often used to provide a basis for estimating delays in networks. At the least, they provide valuable qualitative results and worthwhile insights. In a typical queueing system scenario is characterized by a variable set of customer(s) which contend for resources served by server(s). A customer leaves the system once it receives the service. For a single queueing system where customers are identified based on their arrival order we denote the following;

$$\begin{aligned} N(t) &= \text{number of customers in the queue at time t.} \\ \alpha(t) &= \text{number of customers who arrived in [0,t].} \\ T_i &= \text{time spent in the system by customer i.} \end{aligned}$$

Assume that the following three limits exist:

$$\begin{aligned} N &= \lim_{t\to\infty} \frac{1}{t}\int_0^t N(t) \\ \lambda &= \lim_{t\to\infty} \frac{\alpha(t)}{t} \\ T &= \lim_{t\to\infty} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)} \end{aligned}$$

1.1 Little's Lemma

Lemma 1

$$N = \lambda T$$

Proof: We will prove the Lemma under two simplifying assumptions.

- The system becomes empty infinitely often after a finite interval of time.

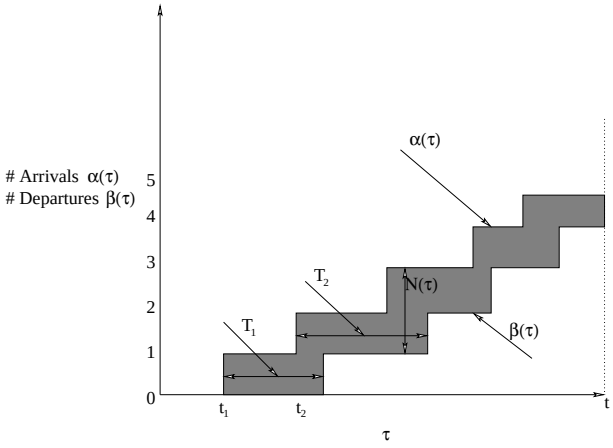

Figure 1: Illustration for the Proof of Little's Lemma

- The customers depart from the system in the same order as they arrive.

The shaded area in Figure 1 between the graphs for arrivals $\alpha(t)$ and departures $\beta(t)$ as functions of t can be expressed as:

$$\int_0^t N(\tau)d\tau$$

If t is the time when the system gets empty i.e. $N(t) = 0$, the shaded area is also equal to the sum of the waiting times for each packet in the system.

$$\begin{aligned}
\int_0^t N(\tau)d\tau &= \sum_{i=1}^{\alpha(t)} T_i \\
\text{Dividing by t;} & \\
\frac{1}{t}\int_0^t N(\tau)d\tau &= \frac{1}{t}\sum_{i=1}^{\alpha(t)} T_i \\
&= \frac{\alpha(t)}{t}\frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)} \\
\text{or equivalently, } N_t &= \lambda_t T_t \\
\text{taking limits as } & \quad t \to \infty \\
\text{proof follows assuming,} & \\
lim_{t\to\infty} N_t &= N \\
lim_{t\to\infty} \lambda_t &= \lambda \\
lim_{t\to\infty} T_t &= T
\end{aligned}$$

■

2 The M/M/1 Queueing system

The name M/M/1 reflects the standard queueing theory nomenclature where by:

1. The first letter indicates the nature of the arrival process. [e.g. M stands for memoryless which means in our context a Poisson arrival process].
2. The second letter indicates the nature of the probability of the server process. [e.g. M stands for memoryless which means in our context an exponential service time model].
3. The last number indicates the number of servers.

2.1 Assumptions

1. Customers arrive according to a Poisson process with rate λ.
2. The probability distribution of the service time is exponential with mean $\frac{1}{\mu}$ sec.
3. The successive interarrival times and service times are statistically independent of each other.

2.2 The Poisson Process

Let $A(t)$ be the number of arrivals in the interval $[0, t]$. Thus $A(t)$ is discrete but a continuous time process. $A(t)$ is a Poisson process if for all $\tau > 0$:

1. $\Pr(A(t+\tau) - A(t) = n) = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}$
2. Number of arrivals in disjoint intervals are independent.

Alternatively if;

1. $A(t)$ is finite $\forall t$.
2. Number of arrivals in disjoint intervals are independent.
3. The number of arrivals between t and $t+\tau$ depends only on τ.

then the resulting continuous time process is Poisson described by the expression $\Pr(A(t+\tau) - A(t) = n) = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}$

2.2.1 Properties of a Poisson Process

1. The expected number of arrivals in an interval of t steps is λt.
2. Let τ_i be the interval between the ith and $i+1$th arrivals. τ_i has an exponential distribution with parameter λ:

$$\Pr(\tau_i \leq s) = 1 - \Pr(A(t+s) - A(t) = 0) = 1 - e^{-\lambda s}$$

3. $\Pr(A(t+\tau) - A(t) = 1) = \lambda\tau + o(\tau)$. [1]

[1] $$f \in o(\tau) \equiv \lim_{\tau \to 0} \frac{f(\tau)}{\tau} = 0$$

4. $\Pr(A(t+\tau) - A(t) \geq 2) = o(\tau)$.
5. Sum of Poisson processes is a Poisson process with sum of rates.
6. If a Poisson process is split randomly, the two processes are Poisson.

2.3 The Exponential Distribution

X is exponentially distributed; denoted as $X \sim Exp(\mu)$ if: $\Pr(X \leq s) = 1 - e^{-\mu s}$

2.3.1 Properties of Exponential Distribution

1. The density function is given by $f_X(s) = \mu e^{-\mu s}$
2. $E[X] = \int_0^\infty s f_X(s) ds = \mu \int_0^\infty s e^{-\mu s} ds = \frac{1}{\mu}$
3. The exponential distribution is **memoryless**:
   $\Pr(X > t + \tau | X > t) = \frac{\Pr(X > t+\tau)}{\Pr(X>t)} = \frac{e^{-\mu(t+\tau)}}{e^{-\mu t}} = e^{-\mu\tau} = \Pr(X > \tau)$

2.4 Interarrival and Waiting time distributions

Let X_n ($n \geq 1$) denote the time from the $(n-1)$st to the nth arrival. Let T_n denote the time for the nth arrival; $T_0 = 0$. Then $X_n = T_n - T_{n-1}$. The sequence $\{X_n, n = 1, 2, \ldots\}$ is the **sequence of interarrival times**. If the arrival process is Poisson with rate λ then each random variable in the above sequence is independently and identically distributed exponential random variable with mean $1/\lambda$.

$$\Pr(X_n \leq s) = 1 - \Pr(X_n > s)$$
$$= 1 - \Pr(A(T_{n-1} + s) - A(T_{n-1}) = 0)$$
$$= 1 - e^{-\lambda s} \frac{(\lambda s)^0}{0!} = 1 - e^{-\lambda s}$$

as required. By identical distribution we mean that all packets have the same arrival distribution.
An intuitive notion for a memoryless distribution is as follows: Given an individual who talks on a phone for t secs then what is the probability that he will talk for the next τ seconds?
Answer: Under the given assumptions, the probability that he will talk for next τ secs is independent of the length of his past connection i.e. t. Hence the distribution is memoryless (exponential). In reality however, we cannot necessarily assume the independence of the length of the past conversation in predicting the future length.
Since the service process is exponentially distributed; so are the waiting times.

3 Discrete Time Markov Chain formulation

The memoryless property and independence of inter arrival and service times imply that once we know the number $N(t)$ of customers in the system at time t, the times at which customers will arrive in future are independent of the arrival times of the customers presently in the system and of how much service the customer in service (if any) has already received. This means that the future number of customers depend on past numbers only through the present

number;i.e. $\{N(t)|t \geq 0\}$ is a **continuous-time Markov chain**. However, for simplicity we will analyze a discrete-time Markov chain.

Let N_k = number of customers in the system at time $k\delta$, $k = 0, 1, \ldots$, and δ is a small positive number.

In the interval $I_k = (k\delta, (k+1)\delta]$, we have the following probabilities associated with the arrival process A and departure process D.

$$\begin{aligned}
\Pr(A((k+1)\delta) - A(k\delta) = 0) &= 1 - \lambda\delta + o(\delta) && (1)\\
\Pr(A((k+1)\delta) - A(k\delta) = 1) &= \lambda\delta + o(\delta) && (2)\\
\Pr(A((k+1)\delta) - A(k\delta) \geq 2) &= o(\delta) && (3)\\
\Pr(D((k+1)\delta) - D(k\delta) = 0) &= 1 - \mu\delta + o(\delta) && (4)\\
\Pr(D((k+1)\delta) - D(k\delta) = 1) &= \mu\delta + o(\delta) && (5)\\
\Pr(D((k+1)\delta) - D(k\delta) \geq 2) &= o(\delta) && (6)
\end{aligned}$$

Let $P_{ij} = \Pr(N_{k+1} = j | N_k = i)$ denote the transition probabilities. If there are no customers then there are no departures; so using (1) we have;

$$P_{00} = 1 - \lambda\delta + o(\delta)$$

If the number of arrivals is equal to the number of departures; $(1) \times (4) + (2) \times (5) + (3) \times (6)$ gives;

$$P_{ii} = 1 - \lambda\delta - \mu\delta + o(\delta) \ldots i \geq 1$$

if the number of arrivals is one more than that of departures; then the transition probability is majored by $(2) \times (4)$

$$P_{i,i+1} = \lambda\delta + o(\delta) \ldots i \geq 0$$

if the number of arrivals is one less than that of departures; then the transition probability is majored by $(5) \times (1)$

$$P_{i,i-1} = \mu\delta + o(\delta) \ldots i \geq 1$$

In all other cases the difference in number of arrivals and departures is atleast two and equations $(3) and (6)$ dominate the transition probabilities.

$$P_{i,j} = o(\delta) \ldots j \neq i, i+1, i-1$$

Notice that we use the independence of arrival and departure processes in calculating the transition probabilities.

3.1 Derivation of steady state distribution

$\{N_k | k = 0, 1, \ldots\}$ is a discrete Markov chain with steady state probabilities:

$$p_n = \lim_{t \to \infty} \Pr(N(t) = n) = \lim_{k \to \infty} \Pr(N_k = n)$$

In steady state, we have the global balance equations:

$$\begin{aligned}
p_n\lambda\delta + o(\delta) &= p_{n+1}\mu\delta + o(\delta)\\
\text{As} \quad \delta \to 0, &\quad \text{we have}\\
p_n\lambda &= p_{n+1}\mu
\end{aligned}$$

$$
\begin{aligned}
p_{n+1} &= \rho^{n+1} p_0 \ldots \text{where } \rho = \lambda/\mu \\
\rho < 1 \Rightarrow \sum_{n=0}^{\infty} p_n &= \sum_{n=0}^{\infty} \rho^n p_0 = 1 \\
\frac{p_0}{1-\rho} &= 1 \\
\text{Thus, } p_n &= \rho^n (1-\rho), n = 0, 1, \ldots .
\end{aligned}
$$

The expected number of customers in the system at steady state is given by:

$$N = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

Using Little's Lemma the average delay per customer is given by:

$$T = N/\lambda = \frac{1}{\mu - \lambda}$$

The average number of customers in the queue is given by:

$$N_Q = \lambda W = \lambda(T - \frac{1}{\mu})$$

$$= \lambda(\frac{1}{\mu - \lambda} - \frac{1}{\mu}) = (\frac{\lambda}{\mu})(\frac{\lambda}{\mu - \lambda}) = \frac{\rho^2}{1-\rho}$$

4 Applications for M/M/1 Queueing system

1. Increasing arrival and transmission rate by the same factor $K > 1$ does not change the utilization factor and hence the average number of packets in the system. However, the average delay per packet reduces to

   $$T = N/(K\lambda)$$

2. Statistical multiplexing versus time-division multiplexing:

   Given m identical Poisson process each with arrival rate λ/m, which scheme gives better delay results?
   Statistical Multiplexing: Mixing in one channel gives a delay corresponding to an arrival rate λ and service rate μ i.e. $T = \frac{1}{\mu - \lambda}$
   Time Division Multiplexing: Transmitting through m separate channels gives a much bigger per packet delay $T' = \frac{m}{\mu - \lambda}$ since the service rate per channel now reduces to $\frac{\mu}{m}$.

4.1 Next Class

We will study more sophisticated Queueing systems, such as $M/M/m$, $M/M/\infty$, $M/G/1$ etc. We will do this in the next class.

References

[1] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, (1992), pp. 149-173.