Computer Networks
A simple queueing system

Saad Mneimneh
Computer Science
Hunter College of CUNY
New York

Thanks to Andrey Markov.

1 Introduction

We now consider a simple queueing system called M/M/1. This terminology arises from a standard notational system in queueing theory, first suggested by David Kendel in 1953. The three parts of the notation describe:

- the inter-arrival time distribution,
- the service time distribution, and
- the number of servers

The letter M stands for memoryless, i.e. the Poisson process with exponential distribution. Other possibilities include D for deterministic inter-arrival times, and G denoting a general distribution of inter-arrival times. Therefore, the M/M/1 queueing system consists of customers arriving according to a Poisson process with rate λ, and one server with a service time per customer that is exponentially distributed with rate μ. This of course does not completely describe the system; for instance, what is the size of the queue where customers wait for service? We will assume an infinite queue for now as depicted below:

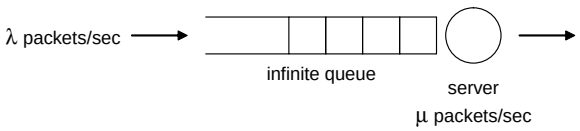

Figure 1: The M/M/1 queueing system for packets (server is transmitter)

We already know a lot about this system from simple application of Little's theorem. For instance, $N = \lambda T$, where N is the average number of customers in the system and T is the average delay per customer. Similarly, $N_Q = \lambda W$, where N_Q is the average number of customers in the queue (i.e. not being currently served), and W is the average waiting time in the queue per customer. Moreover, $\rho = \lambda/\mu$, where ρ is the average number of customers being served, which can be also interpreted as the server utilization, throughput, or efficiency. [1]

Note also that $N = N_Q + \rho$ and $T = W + 1/\mu$. Therefore, it is enough to determining one of N, T, N_Q, and W, to determines all. One approach would be to first determine N.

2 Modeling the system as a Markov chain

An important consequence of the memoryless property is that it allows us to use the theory of Markov chains, named after the mathematician Andrey Markov. A Markov chain consists of a set of states (possibly infinite), with given probabilities of transitions P_{ij} from state i to state j in one time step. For instance, if we let X_k be the state at time $k\delta$ for a fixed δ and $k = 0, 1, 2, \ldots$, then

$$\begin{aligned} P_{ij} &= P(X_{k+1} = j | X_k = i, X_{k-1} = i_{k-1}, \ldots X_0 = i_0) \\ &= P(X_{k+1} = j | X_k = i) \quad \forall k \end{aligned}$$

In other words, conditioned on X_k, X_{k+1} is independent of the past. For the M/M/1 queing system, we can assume that $X_k = N_k$ is the number of customers at time $k\delta$. Therefore, our (infinite) set of states consists of the states $0, 1, 2 \ldots$. The fact that arrivals and service are modeled by Poisson processes (memoryless) and are independent means that a Markov chain is an appropriate model for M/M/1. For a small δ (ignoring the $o(\delta)$ terms):

$$P(N_{k+1} = n | N_k = n) \approx P(0 \text{ arrivals}, 0 \text{ departures}) \approx (1-\lambda\delta)(1-\mu\delta) \approx 1-\lambda\delta-\mu\delta$$

$$P(N_{k+1} = n+1 | N_k = n) \approx P(1 \text{ arrival}, 0 \text{ departures}) \approx \lambda\delta(1-\mu\delta) \approx \lambda\delta$$

$$P(N_{k+1} = n-1 | N_k = n) \approx P(0 \text{ arrival}, 1 \text{ departures}) \approx (1-\lambda\delta)\mu\delta \approx \mu\delta$$

$$P(|N_{k+1} - n| \geq 2 | N_k = n) \approx 0$$

[1] Note that $\rho \leq 1$ because the server can serve at most one customer at a time. However, λ and μ are unconstrained! How can we explain this?

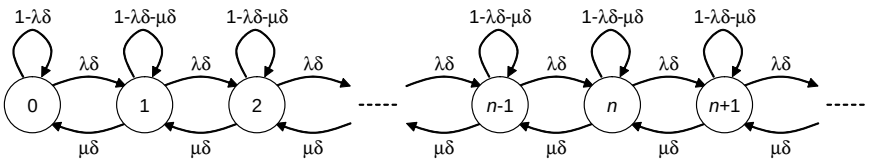

Figure 2: Markov chain of M/M/1

We would like to find N, the expected number of customers in the system at steady state

$$N = \sum_{n=0}^{\infty} p_n n$$

where p_n is the steady state probability of being in state n

$$p_n = \lim_{k \to \infty} P(N_k = n | N_0 = i) \quad i = 0, 1, 2 \ldots$$

But, is there a steady state? To answer this question, we need a little bit of theory.

3 A little bit theory

A Markov chain is *irreducible* iff the directed graph formed by the chain is connected, i.e. given any two states i and j, there is a path from i to j. Moreover, a state i is periodic iff there is a path from i to i and the length of every such path is a multiple of some integer $d > 1$ (d is said to be the period). A Markov chain is *aperiodic* iff none of its states is periodic.

Given an irreducible and aperiodic Markov chain, there are two possibilities for $p_j = \lim_{k \to \infty} P(X_k = j | X_0 = i)$:

- $p_j = 0$ for all j, in which case the chain has no steady state distribution
- $p_j > 0$ for all j, in which case this is the unique steady state distribution of the chain, and it satisfies the equations

  $$\sum_j p_j = 1, \quad p_j = \sum_i p_i P_{ij} \quad \text{(why?)}$$

  where P_{ij} is the transition probability from state i to state j

4 Analysis of M/M/1

According to the above, the markov chain for M/M/1 is irreducible and aperiodic. If a steady state exists, the steady state equation for p_0 is given by:

$$p_0 = p_0(1 - \lambda\delta) + p_1 \mu\delta + o(\delta)$$

Similarly, the steady state equation for p_n for $n > 0$ is given by:

$$p_n = p_n(1 - \lambda\delta - \mu\delta) + p_{n-1}\lambda\delta + p_{n+1}\mu\delta + o(\delta)$$

Since $\lim_{\delta \to 0} o(\delta)/\delta = 0$, dividing by δ and taking the limit as δ goes to 0, we have:

$$\frac{p_n}{p_{n-1}} = \frac{\lambda}{\mu} = \rho \quad \forall n > 0$$

Since $\sum_n p_n = 1$ and $p_n = \rho^n p_0$, we have:

$$p_0 \sum_{n=0}^{\infty} \rho^n = 1$$

Note that the above sum converges only for $\rho < 1$, i.e. for $\lambda < \mu$. If $\rho = 1$ i.e. $p_0 = p_1 = p_2 = \ldots$, all states are equally likely and hence $p_n = 0 \quad \forall n$ (no steady state distribution). If $\rho > 1$ i.e. $p_0 < p_1 < p_2 < \ldots$, further states are more likely (no steady state distribution since the queue is infinite). When $\rho < 1$,

$$\sum_{n=0}^{\infty} \rho^n = \frac{1}{1 - \rho}$$

$$p_0 = 1 - \rho$$

$$p_n = \rho^n(1 - \rho)$$

This result is consistent with the meaning of ρ from Little's theorem: the expected number of customers being served. Since at most one customer can be served, we have:

$$\rho = P(\text{system is empty}) \cdot 0 + P(\text{system is not empty}) \cdot 1$$
$$= p_0 \cdot 0 + (1 - p_0) \cdot 1 = 1 - p_0$$

Now the expected number of customers in the system can be computed:

$$\begin{aligned} N &= \textstyle\sum_{n=0}^{\infty} p_n n \\ &= \rho(1 - \rho) \textstyle\sum_{n=0}^{\infty} n\rho^{n-1} \\ &= \rho(1 - \rho) \frac{\partial}{\partial \rho} \textstyle\sum_{n=0}^{\infty} \rho^n \\ &= \rho(1 - \rho) \frac{\partial}{\partial \rho} \frac{1}{1-\rho} \\ &= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \end{aligned}$$

From Little's theorem, the expected delay per customer is:

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

Therefore, the expected waiting time (in queue) is:

$$W = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

and the expected number of customers in the queue is

$$N_Q = \lambda W = \frac{\rho^2}{1 - \rho}$$

We can also verify that

$$N = N_Q + \rho$$

The parameter ρ can be interpreted as the throughput of the system at steady state. If $\rho \geq 1$, the steady state solution does not exists, but the server is expected to be alwasy busy ($\lambda > \mu$). Therefore, we can define $\min(1, \rho)$ as the throughput of the system as illustrated in the figure below:

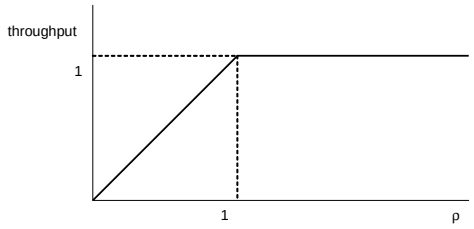

Figure 3: Throughput curve of M/M/1

Similarly, the delay of M/M/1 is illustrated in the figure below:

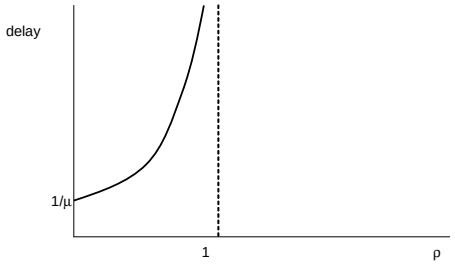

Figure 4: Delay curve of M/M/1

The best operating point of the system is when throughput is high and delay is low. If we define the power as the throughput divided by the delay, we have the following:

$$\text{power} = \frac{\text{throughput}}{\text{delay}} = \lambda(1 - \lambda/\mu)$$

It is easy to see that power is maximized when $\lambda = \mu/2$ ($\rho = 1/2$):

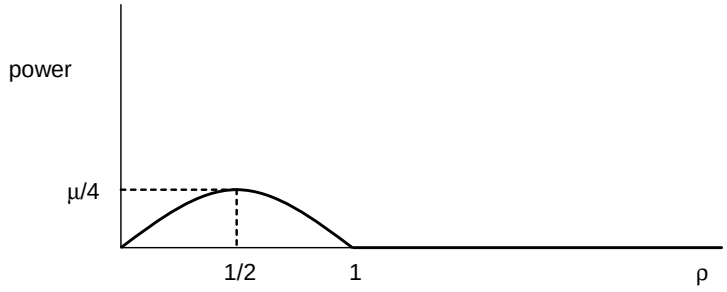

Figure 5: Power curve of M/M/1

5 Example 1 of M/M/1: a fast food restaurant

Conisder a fast food restaurant with Poisson arrivals at a rate of 100 customers per hour. The service time is exponentially distributed with an average of 30 seconds.

- $\lambda = 100$
- $\mu = 1/0.5 = 2$ customers/min $= 120$ customers/hour
- A customers spends on average $T = \frac{1}{\mu - \lambda} = \frac{1}{20}$ hours $= 3$ min until completely served
- A customer waits in line on average $W = T - 1/\mu = 3 - 0.5 = 2.5$ min
- The average number of customers in the restaurant at any time is $\lambda T = 100\frac{1}{20} = 5$
- The throughput (servant utilization) is $\rho = \lambda/\mu = 5/6$

6 Example 2 of M/M/1: Packet switching vs. circuit switching

Consider m sessions each with a Poisson arrival at a rate of λ/m. Packet sizes are exponentially distributed with an average of L bits. The line has a bandwidth of μ bps.

- The transit time is $\frac{\text{packet size}}{\mu}$, thus transit times are exponentially distributed with an average of L/μ
- Packet switching (Poisson processes of m sessions are merged)

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

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

- Circuit switching (each session is given $1/m$ of link bandwidth)

$$T = \frac{1}{\mu/mL - \lambda/m} = \frac{m}{\mu/L - \lambda}$$

$$N = (\lambda/m)T = \frac{\lambda}{\mu/L - \lambda} \quad \text{(per session)}$$

- Delay and number of packets are both multiplied by m in circuit switching

References

Dimitri Bertsekas and Robert Gallager, Data Networks