Advanced Queueing Theory

These slides are created by Dr. Yih Huang of George Mason University. Students registered in Dr. Huang's courses at GMU can make a single machine-readable copy and print a single copy of each slide for their own reference, so long as each slide contains the copyright statement, and GMU facilities are not used to produce paper copies. Permission for any other use, either in machine-readable or printed form, must be obtained from the author in writing.

M/G/1 Queueing Systems

- Service times have a general distribution.
- Other assumptions of M/M/1 are retained.

Implications:
- We can no longer rely on the memoryless property of service times.
- If we were to use the state transition diagram approach, then each state must contain \((N(t), A(t))\), where \(N(t)\) is the number of customers at time \(t\) and \(A(t)\) represents how long the customer in the server has been served up to time \(t\).
- Can you explain why we don't need \(A(t)\) in M/M/*?
Notations

- $W_i$ --- waiting time in queue of the $i$th customer
- $R_i$ --- residual service time of the currently served customer upon the arrival of the $i$th customer
- $X_i$ --- service time of the $i$th customer
- $N_i$ --- number of customers found waiting in queue by the $i$th customer upon his arrival

Analysis

By definition,

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j$$

$$\Rightarrow E[W_i] = E[R_i] + E\left\{ \sum_{j=i-N_i}^{i-1} X_j \right\} = E[R_i] + E[X]E[N_i]$$

$$\Rightarrow W = R + \frac{1}{\mu} N_Q , \text{ (by taking the limit } i \to \infty)$$

$$\Rightarrow W = R + \frac{1}{\mu} (\lambda W) , \text{ (by Little's Theorem)}$$

$$\Rightarrow W = \frac{R}{1 - \rho}$$

where $R$ is the average residual service time.
Residual Times

- $r(t)$ --- the residual service time at time $t$
- $m(t)$ --- the number of service completions up to time $t$.

Let us compute the time average of $r(t)$ in the interval $(0,t)$:

\[ \frac{1}{t} \int_0^t r(s)ds = \frac{1}{t} \sum_{i=1}^{m(t)} \frac{1}{2} X_i^2 = \frac{1}{2} \left( \frac{m(t)}{t} \right) \left( \frac{\sum_{i=1}^{m(t)} X_i^2}{m(t)} \right) \]

\[ \Rightarrow \lim_{t \to \infty} \frac{1}{t} \int_0^t r(s)ds = \frac{1}{2} \left( \lim_{t \to \infty} \frac{m(t)}{t} \right) \left( \lim_{t \to \infty} \frac{\sum_{i=1}^{m(t)} X_i^2}{m(t)} \right) \]

\[ \Rightarrow R = \frac{1}{2} \lambda \bar{X}^2, \quad \text{where } \bar{X}^2 = E[X^2] \]

Recalling that $W = R/(1 - \rho)$, we now have the Pollaczek-Khinchin (P-K) formula:

\[ W = \frac{\lambda \bar{X}^2}{2(1 - \rho)} \]
Average time in system ($X = 1/\mu$ is the average service time):

$$T = \bar{X} + W = \bar{X} + \frac{\lambda \bar{X}^2}{2(1 - \rho)}$$

Average number of customers in queue:

$$N_0 = \lambda W = \bar{X} + \frac{\lambda^2 \bar{X}^2}{2(1 - \rho)}$$

Average number of customers in system:

$$N = \lambda T = \rho + \frac{\lambda^2 \bar{X}^2}{2(1 - \rho)}$$

---

**G/M/1 Queueing Systems**

- Interarrival times form a general distribution with pdf $g(t)$.
- All other M/M/1 assumptions are retained.
- As in the case of M/G/1 queues, we cannot summarize the state of the entire system in a single number, the number of customers in system.
- Instead, we will focus on the behavior of the system at some “special moments” when the state of the system can be summarized in one number.
State Probability Revisited

- For M/M/1 queues, we solved $P_i$, the probability for the system to be in state $i$.
- If you think carefully, the state probability changes over time.
  - consider a barbershop whose $\lambda = 4$ customers per hour and whose $\mu = 5$ customers per hour
  - according to our formula, $P_3 = 0.8^3 \times 0.2 = 0.1024$
  - however, what is the chance of seeing 3 customers in the shop in the first 1 second?
  - must be very very small! (certainly less than 10%)
- What does $P_i$ really mean?

Let $P_i(t)$ be the probability of having $i$ customers during the interval [0,t], that is

$$P_i(t) = \frac{\text{the portion in [0,t] when the system has } i \text{ customers}}{\text{the length of the interval, } t}$$

- $P_i$ is defined as $P_i = \lim_{t \to \infty} P_i(t)$
- That is, $P_i$ is the average probability of state $i$ over an indefinitely long period of time, taking all time points into account.
- It turns out that $P_i$ is difficult to obtain with G/M/1 queues.
Rather than finding the probability over all time points, we shall content ourselves with the system's behavior only at the moments of customer arrivals.

Precisely, let $\pi_i$ be the probability that an arriving customer sees $i$ customers in the system.

Can you see that the knowledge of $\pi_i$ is rather limited?
- with $P_i$, we know the probability of state $i$ at all times, as long as the system has been running long enough
- with $\pi_i$, we know the probability of state $i$ only at the moments of customer arrivals

On the positive side, the system CAN be summarized in a single number at such moments. Why?

---

State Transition Diagram

- A transition represents a customer arrival.
- $P_{i,j}$ represents the probability of moving from $i$ to $j$ upon a new arrival.
Reading the Diagram

- How do we enter state $i$ from $i$?
  - the system had $i$ customers when the previous customer arrived
  - it has $i$ customers when the next customer arrives
  - this means that exactly one customer has been served and left the system between the two arrivals

- In general, a transition from $i$ to $i+1-j$ means $j$ customers have been served between two consecutive arrivals.

It can be shown that (see Appendix):

$$
\pi_k = (1 - \beta) \beta^k
$$

where

$$
\beta = \int_0^\infty e^{-\mu(1-\beta)} dG(t)
$$

- We can obtain the value of $\beta$ through numeric methods.
Notice that, when an arrival sees $k$ customers in system, then he spends $k+1$ service periods in the system, implying

$$T = \sum_{k=0}^{\infty} \frac{k+1}{\mu} \pi_k = \sum_{k=0}^{\infty} \frac{k+1}{\mu} (1-\beta)^k \beta^k = \frac{1}{\mu(1-\beta)}$$

Finally, the average number of customers in the system is

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

Amazingly, the above $T$ and $N$ formulas are unconditional, that is, they are valid at all times, not just the moments of arrivals.

---

**Appendix**

By definition,

$$P_{i,i+1-j} = \int_0^{\infty} e^{-ut} \frac{(ut)^j}{j!} g(t) dt, \quad j = 0,1,\ldots,i$$

$$P_{i,0} = 1 - \sum_{j=0}^{i} P_{i,i+1-j}$$

By the nature of $\pi$, we have

$$\pi_k = \sum_{i} \pi_i P_{ik}, \quad k \geq 0$$

and

$$\sum_k \pi_k = 1$$
The above can be reduced to

\[ \pi_k = \sum_{i=k}^{\infty} \pi_i \int_0^\infty e^{-ut} \frac{(ut)^{i+k}}{(i+k)!} dG(t), \quad k \geq 1 \]

and

\[ \sum_{0}^{\infty} \pi_k = 1 \]

Let us try a solution of the form \( \pi_k = c \beta^k \).
That is,

\[ c \beta^k = \sum_{i=k}^{\infty} c \beta^i \int_0^\infty e^{-ut} \frac{(ut)^{i+k}}{(i+k)!} dG(t) \]

\[ = c \int_0^\infty e^{-ut} \beta^{k-1} \sum_{i=k}^{\infty} \frac{(\beta ut)^{i+k}}{(i+k)!} dG(t) \]

However,

\[ \sum_{i=k}^{\infty} \frac{(\beta ut)^{i+k}}{(i+k)!} = \sum_{j=0}^{\infty} \frac{(\beta ut)^j}{j!} = e^{\beta ut} \]

\[ \Rightarrow \beta^k = \beta^{k-1} \int_0^\infty e^{-ut(1-\beta)} dG(t) \]

\[ \Rightarrow \beta = \int_0^\infty e^{-ut(1-\beta)} dG(t) \]

We can obtain the value of \( \beta \) through numeric methods.

Since \( \sum_{0}^{\infty} \pi_k = \sum_{0}^{\infty} c \beta^k = 1 \), we have \( c = 1 - \beta \).

It follows that the conditional probability

\[ \pi_k = (1 - \beta) \beta^k \]