$G/M/m$ Queueing Systems

John C.S. Lui

Department of Computer Science & Engineering
The Chinese University of Hong Kong
www.cse.cuhk.edu.hk/∼cslui

Outline

Introduction

- Interarrival times are i.i.d according to $A(t)$ with mean time being $1/\lambda$.
- Service times are i.i.d and is exponentially distributed with mean $1/\mu$.
- Instead of keeping track how long since the past arrival occurs, look at the arrival instants, which form an imbedded Markov chain.
- Let $q_n^{'}$ be the number of customers in the system immediately prior to the arrival of customer C_n.
- let $v_{n+1}^{'}$ be the number of customers served during the arrival of C_n and C_{n+1}. We have

$$q_{n+1}^{'} = q_n^{'} + 1 - v_{n+1}^{'}. \tag{1}$$

Now we need to find the transition probabilities of this imbedded Markov chain.

Derivation of transition probabilities

- Define $p_{ij} = P[q'_{n+1} = j | q'_n = i]$.
- p_{ij} is simply the probability that $i + 1 - j$ customers got served during the interarrival time, and it is clear that $p_{ij} = 0$ for $j > i + 1$.
- Transition structure of this imbedded Markov chain:

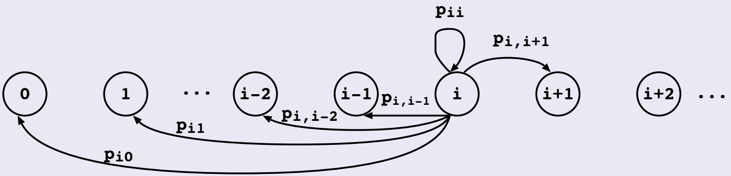

Figure: Transition diagram of $G/M/m$

- Define system utilization as $\rho = \frac{\lambda}{m\mu}$. For system to be stable, we require $\rho < 1$.

Derivation of r_k

- Define $r_k = \lim_{n\to\infty} P[q_n^{'} = k]$ as the steady state probability of finding k customers upon arrival.
- To find all r_k, where $k \geq 0$, we can use

$$\boldsymbol{r} = \boldsymbol{r}\boldsymbol{P} \quad \text{where } \boldsymbol{r} = [r_0, r_1, \ldots]$$

  The only remaining issues is, what are $p_{ij} \in \boldsymbol{P}$?
- **Case 1:** we know that $p_{ij} = 0$ if $j > i + 1$.

Case 2: Derivation of p_{ij} where $j \leq i+1$

- For $j \leq i+1 \leq m$, this is the case which no customers are waiting in the queue.
- Condition that the interarrival time is t, we have

$$
\begin{aligned}
& P[i+1-j \text{ departures within } t \text{ after } C_n \text{ arrives} |\ q_n^{'} = i] \\
&= \binom{i+1}{i+1-j} \left[(1-e^{-\mu t})\right]^{i+1-j} \left[(e^{-\mu t})\right]^{j} \\
&= \binom{i+1}{j} \left[(1-e^{-\mu t})\right]^{i+1-j} \left[(e^{-\mu t})\right]^{j}
\end{aligned}
$$

- Removing the condition of interarrival time t:

$$
p_{ij} = \int_{t=0}^{\infty} \binom{i+1}{j} \left[(1-e^{\mu t})\right]^{i+1-j} \left[(e^{\mu t})\right]^{j} dA(t) \qquad j \leq i{+}1 \leq m. \tag{2}
$$

Case 3: Derivation of p_{ij} where $m \leq j \leq i+1, i \geq m$

- For $m \leq j \leq i+1, i \geq m$, it means that all m servers are busy throughout the interarrival interval.
- Since service time is exponential (memoryless), the number of customers served during this interval will be Poisson distributed with parameter $m\mu$. We have

$$P[k \text{ customers served} |\ t, \text{all } m \text{ busy}] = \frac{(m\mu t)^k}{k!} e^{-m\mu t}.$$

- If we go from state i to j, it means $i+1-j$ customers have been served:

$$p_{ij} = \int_{t=0}^{\infty} P[k \text{ customers served} |\ t, \text{all } m \text{ busy}] dA(t)$$

Case 3: Derivation of p_{ij} where $m \leq j \leq i+1, i \geq m$ (continue)

- Putting everything together, we have:

$$p_{ij} = \int_{t=0}^{\infty} \frac{(m\mu t)^{i+1-j}}{(i+1-j)!} e^{-m\mu t} dA(t) \qquad m \leq j \leq i+1. \tag{3}$$

- Since i and j only appear as the *difference* of $i+1-j$, we define a new quantity with a single index of $\beta_{i+1-j} = p_{ij}$:

$$\beta_n = p_{i,i+1-n} = \int_{t=0}^{\infty} \frac{(m\mu t)^n}{n!} e^{-m\mu t} dA(t) \qquad 0 \leq n \leq i+1-m, m \leq i. \tag{4}$$

Case 4: Derivation of p_{ij} where $j < m < i + 1$

- It means when C_n arrives, there are m customers in service and $i - m$ waiting in the queue, when C_{n+1} joins, there are j customers in service.
- Queue will be empty when $i + 1 - m$ are served. Let $\tilde{y}$ be the time to serve $i + 1 - m$ customers, so $\tilde{y}$ is $(i + 1 - m)^{th}$−Erlangian with

$$f_{\tilde{y}}(y) = \frac{m\mu(m\mu y)^{i-m}}{(i-m)!} e^{-m\mu y}, \qquad y \geq 0.$$

- Also, we have to serve $m - j$ customers so C_{n+1} will find j customers. Let t_{n+1} be the interarrival time between C_{n+1} and C_n, we need to serve $(m - j)$ in $(t_{n+1} - \tilde{y})$ time.

Case 4: Derivation of p_{ij} where $j < m < i+1$ (continue)

- Let $t_{n+1} = t$ and $\tilde{y} = y$, we have

$$P[m-j \text{ served in } t-y | t_{n+1} = t, \tilde{y} = y] = \binom{m}{m-j}\left(1-e^{-\mu(t-y)}\right)^{m-j}\left(e^{-\mu(t-y)}\right)^{j}$$

$$P[m-j \text{ served in } t-y | t_{n+1} = t] = \int_{y=0}^{t}\binom{m}{j}\left(1-e^{-\mu(t-y)}\right)^{m-j}\left(e^{-\mu(t-y)}\right)^{j} f_{\tilde{y}}(y)dy$$

Since we know $f_{\tilde{y}}(y)$, putting it in above and uncondition on t_{n+1}:

$$p_{ij} = \int_{t=0}^{\infty}\binom{m}{j} e^{-j\mu t}\left[\int_{y=0}^{t}\frac{(m\mu y)^{i-m}}{(i-m)!}(e^{-\mu y}-e^{-\mu t})^{m-j} m\mu dy\right] dA(t)dt \tag{5}$$

for $j < m < i+1$.

Putting them together

- Now we can use the standard method to solve for $\boldsymbol{r} = \boldsymbol{r}\boldsymbol{P}$.
- We know all entries of $\boldsymbol{P}$ by the four *marked* equations of p_{ij}.
- Theoretically, we are done, but practically, we have a problem since $\boldsymbol{r}$ is a vector with infinite dimension and $\boldsymbol{P}$ is a two-dimensional infinite matrix.

Importance of β_n

Remember, β_n is the probability that all m servers will finish processing n customers between the interarrival instant. We will use β_n in later section to derive more results.

Introduction

- Define $N_k(t)$ be the number of arrival instants in the interval $(0, t)$ in which the arriving customer finds the system in state E_k, given 0 customer at time $t = 0$.
- Note at the previous figure of transition structure, the system can only move up by at most one state, but may move down by many states in any transition.
- We consider this motion between states and define σ_k (for $m - 1 \leq k$) as the expected number of times E_{k+1} is reached between two successive visits to state E_k.
- The probability of reaching E_{k+1} no times between returns to state E_k is equal to $1 - \beta_0$ (that is, given in state E_k, the only way to reach E_{k+1} before our next visit to E_k is for no customer to be served, which is β_0. So the probability of not reaching E_{k+1} first is $1 - \beta_0$).

Derivation of σ_k (continue)

- Let γ be the probability of leaving E_{k+1} and return to it some time later without passing through E_j, where $j \leq k$.
- Note that γ is independent of k for $k \geq m-1$ (i.e., all m servers are busy). We have:
  P[n visits to E_{k+1} between two successive visits to E_k]=
  $$\beta_0 \gamma^{n-1}(1-\gamma).$$
- We now have:
  $$\sigma_k = \sum_{n=1}^{\infty} n \beta_0 \gamma^{n-1}(1-\gamma) = \frac{\beta_0}{1-\gamma} \quad \text{for } k \geq m-1.$$
  We let $\sigma_k = \sigma$ since it is *independent of* k.
- We also have
  $$\sigma = \lim_{t \to \infty} \frac{N_{k+1}(t)}{N_k(t)} = \frac{\beta_0}{1-\gamma} = \frac{r_{k+1}}{r_k} \quad k \geq m-1.$$

Derivation of σ_k (continue)

- The solution to the last set of equations is clearly

$$r_k = K\sigma^k \quad k \geq m-1. \tag{6}$$

for some constant K.

- Now we have

$$\boldsymbol{r} = [r_0, r_1, r_2, \ldots, r_{m-2}, K\sigma^{m-1}, K\sigma^m, K\sigma^{m+1}, \ldots]$$

- Let us consider the flow balance equation for r_k, $k \geq m$:

$$r_k = K\sigma^k = \sum_{i=0}^{\infty} r_i p_{ik} = \sum_{i=k-1}^{\infty} r_i p_{ik} = \sum_{i=k-1}^{\infty} K\sigma^i \beta_{i+1-k}$$

Derivation of σ_k (continue)

- Cancelling the constant K and common factors of σ:

$$\sigma = \sum_{i=k-1}^{\infty} \sigma^{i+1-k} \beta_{i+1-k} = \sum_{n=0}^{\infty} \sigma^n \beta_n$$

- Since we have derived β_n before, we have

$$\sigma = \sum_{n=0}^{\infty} \sigma^n \int_{t=0}^{\infty} \frac{(m\mu t)^n}{n!} e^{-m\mu t} dA(t) = \int_{t=0}^{\infty} e^{-(m\mu - m\mu\sigma)t} dA(t)$$

- We recognize this as Laplace transform:

$$\sigma = A^*(m\mu - m\mu\sigma). \tag{7}$$

Which is a *functional equation* for σ. So give $A(t)$, we can find σ.

Derivation of conditional distribution of queue size

- Let us find the probability that an arriving customer needs to wait:

$$\text{P[arrival queues]} = \sum_{k=m}^{\infty} r_k = \sum_{k=m}^{\infty} K\sigma^k = \frac{K\sigma^m}{1-\sigma}$$

(note: [TAKA 62] showed that $0 < \sigma < 1$).

- Probability of finding a queue length of n, given that a customer must queue is:

$$\begin{aligned}\text{P[queue size=n|arrival queues]} &= \frac{r_{m+n}}{\text{P[arrival queues]}} \\ &= \frac{K\sigma^{n+m}}{K\sigma^m/(1-\sigma)} = (1-\sigma)\sigma^n \qquad n \geq 0. \quad (8)\end{aligned}$$

The conditional distribution is **geometrically distributed** for $G/M/m$.

Derivation

- Let us define the following conditional Laplace transform:

$$W^*(s|n) = E[e^{-s\tilde{w}}|\text{arrival queues and queue size} = n]$$

- We have

$$W^*(s|n) = \left(\frac{m\mu}{s + m\mu}\right)^{n+1}$$

- The conditional distribution of waiting time is

$$\begin{aligned} W^*(s|\text{arrival queues}) &= \sum_{n=0}^{\infty} \left(\frac{m\mu}{s+m\mu}\right)^{n+1} (1-\sigma)\sigma^n \\ &= (1-\sigma)\frac{m\mu}{s+m\mu-m\mu\sigma} \end{aligned} \quad (9)$$

Derivation (continue)

- Let $w(y|\text{arrival queues})$ be the probability density function for the waiting time, condition that an arriving customer has to wait in the queue. We have

$$w(y|\text{arrival queues}) = (1-\sigma)m\mu e^{-m\mu(1-\sigma)y} \qquad y \geq 0 \quad (10)$$

- The conditional pdf for queueing time is **exponentially distributed** for $G/M/m$.

$G/M/1$

- Let us apply our previous results to $G/M/1$. Since $m = 1$,

$$r_k = K\sigma^k \qquad k = 0, 1, 2, \ldots$$

- Since summing all r_k must be 1, we can find $K = (1 - \sigma)$ and:

$$r_k = (1 - \sigma)\sigma^k \qquad k = 0, 1, 2, \ldots \tag{11}$$

  and $1 - r_0 = \sigma =$ P[arriving customer has to queue]
- σ is the solution to the following functional equation:

$$\sigma = A^*(\mu - \mu\sigma) \tag{12}$$

- Let A be the event "arrival queues":

$$W(y) = 1 - \text{P[queueing time} > y|A]\text{P}[A] = 1 - \sigma e^{-\mu(1-\sigma)y} \quad y \geq 0 \tag{13}$$

Example 1: analyzing $M/M/1$

- For $M/M/1$, $A(t) = 1 - e^{-\lambda t}$ for $t \geq 0$. We have $A^*(s) = \frac{\lambda}{s+\lambda}$.
- To find σ, we have $\sigma = \frac{\lambda}{\mu - \mu\sigma + \lambda}$, or $\mu\sigma^2 - (\mu + \lambda)\sigma + \lambda = 0$.
- This yields $(\sigma - 1)(\mu\sigma - \lambda) = 0$. $\sigma = 1$ is not acceptable due to stability, we then have $\sigma = \frac{\lambda}{\mu} = \rho$.
- Once we have σ, we have:

$$r_k = (1 - \rho)\rho^k \qquad k \geq 0$$

which is our usual solution for $M/M/1$.

Example 2: specific $E_2/M/1$

- Consider an interarrival time distribution such that

$$A^*(s) = \frac{2\mu^2}{(s+\mu)(s+2\mu)}$$

- To find σ, we have

$$\sigma = \frac{2\mu^2}{(\mu - \mu\sigma + \mu)(\mu - \mu\sigma + 2\mu)}$$

  This leads to the cubic equation $\sigma^3 - 5\sigma^2 + 6\sigma - 2 = 0$.

- Since $\sigma = 1$ is always a solution, we have $(\sigma - 1)(\sigma - 2 - \sqrt{2})(\sigma - 2 + \sqrt{2}) = 0$. Solution is $\sigma = 2 - \sqrt{2}$.
- We finally have

$$\begin{aligned} r_k &= (\sqrt{2}-1)(2-\sqrt{2})^k, & k = 0, 1, \ldots, \\ W(y) &= 1 - (2-\sqrt{2})e^{-\mu(\sqrt{2}-1)y} & y \geq 0 \end{aligned}$$

Further analysis of $G/M/m$

- Let's get back to $G/M/m$. We have shown $\boldsymbol{r} = \boldsymbol{r}\boldsymbol{P}$, where $\boldsymbol{r} = [r_0, r_1, \ldots]$. The only remaining unknowns are (a) the constant K, and (b) boundary probabilities r_0,r_1,...,r_{m-2}.
- Since we know $r_k = K\sigma^k$ for $k \geq m-1$, we express

$$\boldsymbol{r} = K\sigma^{m-1}\left[R_0, R_1, \ldots, R_{m-2}, 1, \sigma, \sigma^2, \ldots\right] \quad (14)$$

  where $R_k = \frac{r_k \sigma^{1-m}}{K}$ and $k = 0, 1, \ldots, m-2$.
- For convenience, we define

$$J = K\sigma^{m-1}.$$

- We can apply the flow balance equations on R_i:

$$R_k = \sum_{i=k-1}^{\infty} R_i p_{ik} \qquad k = 0, 1, \ldots, m-2.$$

Continue

- We can express the "tail" of R_k using σ

$$R_k = \sum_{i=k-1}^{m-2} R_k p_{ik} + \sum_{i=m-1}^{\infty} \sigma^{i+1-m} p_{ik}.$$

Solving for R_{k-1}, we have:

$$R_{k-1} = \frac{R_k - \sum_{i=k}^{m-2} R_i p_{ik} - \sum_{i=m-1}^{\infty} \sigma^{i+1-m} p_{ik}}{p_{k-1,k}} \quad k = 1, \ldots, m-1. \tag{15}$$

Note that this is a triangular set, in particular, $R_{m-1} = 1$, so we can solve for $R_{m-2}, \ldots$,1,0.

- The only remaining issue is how to find K (or J).

Continue

- We can use the conservation of probability to evaluate J:

$$J \sum_{k=0}^{m-2} R_k + J \sum_{k=m-1}^{\infty} \sigma^{k-m+1} = 1$$

$$J = \frac{1}{\frac{1}{1-\sigma} + \sum_{k=0}^{m-2} R_k} \tag{16}$$

Derivation of waiting time distribution

- The probability that an arrival customer doesn't need to wait

$$W(0) = \sum_{k=0}^{m-1} r_k = J \sum_{k=0}^{m-1} R_k. \tag{17}$$

- The conditional distribution of waiting is (when $k \geq m$):

$$\text{P}[\tilde{w} < y | \text{ finds } k \text{ in the system}] = \int_{x=0}^{y} \frac{m\mu(m\mu x)^{k-m}}{(k-m)!} e^{-m\mu x} dx.$$

Derivation of waiting time distribution (continue)

- Removing the condition, the waiting time CDF is:

$$\begin{aligned} W(y) &= W(0) + J \sum_{k=m}^{\infty} \int_{x=0}^{y} \frac{(m\mu)(m\mu x)^{k-m} \sigma^{k-m+1}}{(k-m)!} e^{-m\mu x} dx \\ &= W(0) + J\sigma \int_{x=0}^{y} m\mu e^{-m\mu x(1-\sigma)} dx \\ &= 1 - \frac{\sigma}{1 + (1-\sigma) \sum_{k=0}^{m-2} R_k} e^{-m\mu(1-\sigma)y} \quad y \geq 0 \end{aligned} \tag{18}$$

- Let $W = E[\tilde{w}]$, we have:

$$W = \frac{K\sigma^m}{m\mu(1-\sigma)^2} = \frac{J\sigma}{m\mu(1-\sigma)^2} \tag{19}$$

Please refer to Kleinrock's book, Section 6.6, on the example of analyzing $G/M/2$. For example, $\boldsymbol{r}$, $W(y)$, $\ldots$.