Mean Delay in M/G/1 Queues with Head-of-Line Priority Service and Embedded Markov Chains

Wade Trappe
Lecture Overview

- Priority Service
  - Setup
  - Mean Waiting Time for Type-1
  - Mean Waiting Time for Type-2
  - Generalization
  - Summary

- Embedded Markov Chain Technique
  - $N_k$ inside of $N(t)$
  - The queue distribution for $M/G/1$
    - *The PGF approach*
      - Example: $M/H/1$ Queue
      - Homework: Alternate Pollaczek-Khinchine Derivation
M/G/1 Priority Services: Setup

- Consider a queue that handles K priority classes of customers
  - Type-k customers arrive as a Poisson Process with rate $\lambda_k$ and service times according to $X_k$.

- Customers are separated as they arrive and assigned to different “prioritized” queues

- Each time the server is free:
  - Selects the next customer from highest priority, non-empty queue
  - This is the “head of line” service
  - We also assume that customers are not pre-empted!

- Server Utilization for type-k customers is $\rho_k = \lambda_k E[X_k]$

- Stability requirement:
  $$\rho = \rho_1 + \rho_2 + \cdots + \rho_K < 1$$
Blue Type-1 Packets always have priority, but cannot preempt Type-2 packets
Our first goal will be to calculate the mean waiting time for Type-1 customers, i.e. $\overline{W}_1$

Assume a type-1 customer arrives and finds $N_{q1}(t)=k_1$ type-1 already in the queue.

Assume we use FCFS within each class.

$W_1$ consists of:
- Residual time $R$ of customer in server
- $k_1$ service times of type-1 in queue

Thus:
$$\overline{W}_1 = \mathbb{E}[R] + \mathbb{E}[N_{q1}] \mathbb{E}[X_1]$$

We get the mean waiting time for type-1 as
$$\overline{W}_1 = \frac{\mathbb{E}[R]}{1-\rho_1}$$
Now, let us look at a type-2 customer. When he arrives he may find

- \( N_{q_1}(t)=k_1 \) type-1 already in the queue-1
- \( N_{q_2}(t)=k_2 \) type-2 already in the queue-2

Thus, \( W_2 \) is the sum of:

- Residual service time \( R \)
- \( k_1 \) service times for type-1 customers
- \( k_2 \) service times for type-2 customers
- AND service times of any new type-1 customers!!

We get the mean waiting time for type-2 as

\[
\bar{W}_2 = E[R] + E[N_{q_1}]E[X_1] + E[N_{q_2}]E[X_2] + E[M_1]E[X_1]
\]
Here, \( M_1 \) is the number of customers of type-1 that arrive during type-2’s waiting period.

By Little’s Theorem applied to the queues:
\[
E[N_{q_1}] = \lambda_1 \bar{W}_1 \\
E[N_{q_2}] = \lambda_2 \bar{W}_2
\]

The mean number of type-1 arrivals during \( \bar{W}_2 \) seconds is
\[
\bar{M}_1 = \lambda_1 \bar{W}_2
\]

Substitute to get
\[
\bar{W}_2 = E[R] + \rho_1 \bar{W}_1 + \rho_2 \bar{W}_2 + \rho_1 \bar{W}_2
\]
Solve for $\bar{W}_2$:

$$\bar{W}_2 = \frac{E[R] + \rho_1 \bar{W}_1}{1 - \rho_1 - \rho_2}$$

use $\bar{W}_1 = \frac{E[R]}{1 - \rho_1}$

$$= \frac{E[R]}{(1 - \rho_1 - \rho_2)(1 - \rho_1)}$$

Generalizing gives

$$\bar{W}_k = \frac{E[R]}{(1 - \rho_1 - \rho_2 - \cdots - \rho_{k-1})(1 - \rho_1 - \rho_2 - \cdots - \rho_k)}$$

Now, we’re left with finding $E[R]$!
When a customer arrives, the one being serviced can be of any type! So R must account for this!

We will use the same formulation as \( E[R] = \frac{\lambda}{2} E[X^2] \)

where \( \lambda = \lambda_1 + \cdots + \lambda_k \) is the aggregate Poisson Process rate.

To get \( E[X^2] \) we use the fact that the fraction of type-\( k \) customers is \( \left( \frac{\lambda_k}{\lambda} \right) \):

\[
E[X^2] = \frac{\lambda_1}{\lambda} E[X_1^2] + \frac{\lambda_2}{\lambda} E[X_2^2] + \cdots + \frac{\lambda_k}{\lambda} E[X_k^2]
\]

Thus, we get

\[
\bar{W}_k = \frac{\sum_{j=1}^{K} \lambda_j E[X_j^2]}{(1 - \rho_1 - \rho_2 - \cdots - \rho_{k-1})(1 - \rho_1 - \rho_2 - \cdots - \rho_k)}
\]
Now, calculating the mean delay for a type-k customer is easy!

\[ \bar{T}_k = \bar{W}_k + \bar{X}_k \]

From this, observe:
- Class-k customers are affected by the behavior of lower priority customers
Embedded Markov Chain Methods for M/G/1
The Embedded Markov Chain

- Let $N_k$ be the number of customers found in the system just after the service completion of the $k$-th customer.
- Let $A_k$ be the number of customers who arrive while the $k$-th customer is served.
- The $A_k$ are i.i.d.
- Under FCFS: $A_k$ is independent of $N_1, N_2, \ldots, N_{k-1}$.
- $N_k$ is not independent of $A_k$.
We may find the following recursion:

\[ N_k = \begin{cases} 
A_k & \text{if } N_{k-1} = 0 \\
N_{k-1} + A_k - 1 & \text{if } N_{k-1} > 0 
\end{cases} \]

Where did first line come from?
- If \( N_{k-1} = 0 \), then the \( k \)-th customer finds the system empty, and on departure he leaves behind those who have arrived during his service.

The second line:
- The queue left by the \( k \)-th customer consists of the previous queue decreased by one plus the new arrivals during his service.

Let \( U(x)=1 \) for \( x>0 \) and \( U(x)=0 \) for \( x<=0 \)

Then we get

\[ N_k = N_{k-1} - U(N_{k-1}) + A_k \]
From
\[ N_k = N_{k-1} - U(N_{k-1}) + A_k \]

we can see that \( N_k \) is a Markov Chain since \( N_k \) depends on the past only through \( N_{k-1} \).

\( N_k \) is called the embedded Markov Chain.

(\textbf{Homework}): The distribution for \( N(t) \) and \( N_k \) are the same:

\begin{itemize}
  \item Step 1: Show, in M/G/1, distributions found by arriving customers is the same as that left by departing customers.
  \item Step 2: Show, in M/G/1, the distribution of customers found by arriving customers is the same as the steady state distribution of \( N(t) \).
\end{itemize}

We will analyze \( N_k \) instead of \( N(t) \).
To find the queue distribution, we use the *probability generating function* method.

The PGF of the distribution \( \{p_n\} \) is

\[
P(z) = \sum_{n=0}^{\infty} p_n z^n = \lim_{k \to \infty} E\left[Z^{N_k}\right]
\]

Substitute \( N_k = N_{k-1} - U(N_{k-1}) + A_k \) and use independence to get

\[
P(z) = \lim_{k \to \infty} E\left[Z^{A_k}\right]E\left[Z^{N_{k-1} - U(N_{k-1})}\right] = A_k(z) \left[p_0 + \sum_{n=1}^{\infty} p_n z^{n-1}\right]
\]

\[
= A_k(z) \left[p_0 + z^{-1}(P(z) - p_0)\right]
\]

Thus

\[
P(z) = \frac{p_0 A_k(z)(z - 1)}{z - A_k(z)}
\]
If the service time of customer k is t, then \( A_k \) has a Poisson distribution with mean \( \lambda t \)

\[
P[A_k = i | X_k = t] = \frac{e^{-\lambda t} (\lambda t)^i}{i!}
\]

Now multiply by the \( f_X(t) \), the pdf of the service time \( X_k \). This gives the joint distribution.

Integrate out \( X \) to obtain the unconditional pdf of \( A_k \)

\[
a_i = P[A_k = i] = \int_0^\infty \frac{e^{-\lambda t} (\lambda t)^i}{i!} f_X(t) dt
\]

What does this remind you of?

Answer: Laplace Transforms!!!
The Probability Generating Function of $A_k$ is given by

$$A_k(z) = E[z^{A_k}] = \sum_{i=0}^{\infty} a_i z^i$$

$$= \sum_{i=0}^{\infty} \int_0^{\infty} e^{-\lambda t} (\lambda t)^i \frac{f_X(t)}{i!} dt$$

$$= \int_0^{\infty} e^{-\lambda t} e^{\lambda z} f_X(t) dt$$

$$= \tilde{f}_X (\lambda - \lambda z)$$

Substitute into earlier result to get:

$$P(z) = \frac{p_0 \tilde{f}_X (\lambda - \lambda z)(z-1)}{z - \tilde{f}_X (\lambda - \lambda z)}$$
We need to now find $p_0$!

To do this, look at limit as $z$ goes to 1, and apply L’Hopital…

First, go back to

$$P(z) = \frac{p_0 A_k(z)(z-1)}{z - A_k(z)}$$

Observe that $P(z=1)=1$, $A_k(z=1)=1$

Thus

$$1 = \frac{p_0 \hat{f}_X(0)}{1 + \lambda \hat{f}_X'(0)} = \frac{p_0}{1 - \lambda E[X]}$$

We used:

$$\hat{f}_X(0) = \int_0^\infty f_X(t)dt = 1$$

$$\frac{d}{ds}\hat{f}_X(s) = \int_0^\infty (-t)f_X(t)e^{-st}dt = -E[X]$$

Show This!
Thus, \( p_0 = 1 - \lambda E[X] = 1 - \rho \)

This is exactly what we got for M/M/1 Queues!!!

Substitute into earlier result to get:

\[
P(z) = \frac{(1 - \rho) \tilde{f}_X(\lambda - \lambda z)(z - 1)}{z - \tilde{f}_X(\lambda - \lambda z)}
\]

What do we do with this? Use it to find the probability distribution \( p_n \)!

How? Equate terms… Best to see an example
Consider the case where $F_X(t)$, the cdf, is a two-term hyper-exponential distribution:

$$F_X(t) = 1 - \pi_1 e^{-\mu_1 t} - \pi_2 e^{-\mu_2 t}$$

Where do hyper-exponentials come from?
- In practice, from situations where service times can be represented as a mixture of exponential distributions…
- That is, suppose there are $k$ types of customers, each occurring with a probability $\pi_i$ (so sum = 1)
- Customer of type $I$ has a service time exponentially distributed with mean $1/\mu_i$

Then the density is

$$f_X(x) = \sum_{i=1}^{k} \pi_i \mu_i e^{-\mu_i x}$$
Back to our 2-phase hyper-exponential problem…

- The mean service time is given by \[ \frac{1}{\mu} = \frac{\pi_1}{\mu_1} + \frac{\pi_2}{\mu_2} \]

- The Laplace Transform of \( f_X(t) \) is:
  \[ \hat{f}_X(s) = \frac{\pi_1 \mu_1}{\mu_1 + s} + \frac{\pi_2 \mu_2}{\mu_2 + s} \]

- We thus get:
  \[ A_k(z) = \hat{f}_X(\lambda - \lambda z) = \frac{\pi_1}{1 + \rho_1 (1 - z)} + \frac{\pi_2}{1 + \rho_2 (1 - z)} \quad \text{for } \rho_i = \lambda / \mu_i \]
The PGF of \( \{p_n\} \) is

\[
P(z) = \frac{(1-\rho)[1+(\rho_1 + \rho_2 - \rho)(1-z)]}{\rho_1\rho_2z^2 - (\rho_1 + \rho_2 + \rho_1\rho_2)z + 1 + \rho_1 + \rho_2 - \rho}
\]

where \( \rho = \pi_1\rho_1 + \pi_2\rho_2 \) = Total Traffic Intensity

The probability distribution is obtained through partial-fraction expansion of \( P(z) \).

To do this, we must find the roots of the denominator… call these \( z_1 \) and \( z_2 \)

Then \( P(z) \) has the form:

\[
P(z) = \frac{C_1z_1}{z_1 - z} + \frac{C_2z_2}{z_2 - z}
\]
M/H₂/1 Queue, pg. 4

- Solving for C1 and C2 involves setting z=0 and z=1 to get

\[ C_1 = \frac{(z_1 - 1)(1 - \rho z_2)}{z_1 - z_2} \]
\[ C_2 = \frac{(z_2 - 1)(1 - \rho z_1)}{z_2 - z_1} \]

- Finally, upon substitution, we get:

\[ p_n = C_1 z_1^{-n} + C_2 z_2^{-n} \]