TCOM 501: Networking Theory & Fundamentals

Lectures 9 & 10
M/G/1 Queue
Prof. Yannis A. Korilis
Topics

- M/G/1 Queue
- Pollaczek-Khinchin (P-K) Formula
- Embedded Markov Chain Observed at Departure Epochs
- Pollaczek-Khinchin Transform Equation
- Queues with Vacations
- Priority Queueing
M/G/1 Queue

- Arrival Process: Poisson with rate $\lambda$
- Single server, infinite waiting room
- Service times:
  - Independent identically distributed following a general distribution
  - Independent of the arrival process

Main Results:
- Determine the average time a customer spends in the queue waiting service (Pollaczek-Khinchin Formula)
- Calculation of stationary distribution for special cases only
M/G/1 Queue – Notation

- $W_i$: waiting time of customer $i$
- $X_i$: service time of customer $i$
- $Q_i$: number of customers waiting in queue (excluding the one in service) upon arrival of customer $i$
- $R_i$: residual service time of customer $i$ = time until the customer found in service by customer $i$ completes service
- $A_i$: number of arrivals during the service time $X_i$ of customer $i$

Service Times

- $X_1, X_2, \ldots$, independent identically distributed RVs
- Independent of the inter-arrival times
- Follow a general distribution characterized by its pdf $f_X(x)$, or cdf $F_X(x)$
- Common mean $E[X] = 1/\mu$
- Common second moment $E[X^2]$
M/G/1 Queue

State Representation:
- \( \{N(t) : t \geq 0\} \) is not a Markov process – time spent at each state is not exponential
- \( R(t) \) = the time until the customer that is in service at time \( t \) completes service
- \( \{(N(t), R(t)) : t \geq 0\} \) is a continuous time Markov process, but the state space is not a countable set
- Finding the stationary distribution can be a rather challenging task

Goals:
- Calculate average number of customers and average time-delay without first calculating the stationary distribution
- Pollaczek-Khinchin (P-K) Formula:
  \[
  E[W] = \frac{\lambda E[X^2]}{2(1 - \lambda E[X])}
  \]
- To find the stationary distribution, we use the embedded Markov chain, defined by observing \( N(t) \) at departure epochs only – transformation methods
Proposition: Sum of a Random Number of Random Variables

- $N$: random variable taking values 0,1,2,…, with mean $E[N]$
- $X_1, X_2, \ldots, X_N$: iid random variables with common mean $E[X]$

Then: $E[X_1 + \cdots + X_N] = E[X] \cdot E[N]$

Proof: Given that $N=n$ the expected value of the sum is

$$E\left[\sum_{j=1}^{N} X_j \mid N = n\right] = E\left[\sum_{j=1}^{n} X_j\right] = \sum_{j=1}^{n} E[X_j] = nE[X]$$

Then:

$$E\left[\sum_{j=1}^{N} X_j\right] = \sum_{n=1}^{\infty} E\left[\sum_{j=1}^{N} X_j \mid N = n\right] \times P\{N = n\} = \sum_{n=1}^{\infty} nE[X] \cdot P\{N = n\}$$

$$= E[X] \sum_{n=1}^{\infty} nP\{N = n\} = E[X]E[N]$$
Pollaczek-Khinchin Formula

- Assume FCFS discipline
- Waiting time for customer $i$ is:
  $$W_i = R_i + X_{i-1} + X_{i-2} + \cdots + X_{i-Q_i} = R_i + \sum_{j=i-Q_i}^{i-1} X_j$$
- Take the expectation on both sides and let $i \to \infty$, assuming the limits exist:
  $$E[W_i] = E[R_i] + E\left[\sum_{j=i-Q_i}^{i-1} X_j\right] = E[R_i] + E[X]E[Q_i] \Rightarrow$$
- Averages $E[Q], E[R]$ in the above equation are those seen by an arriving customer.
- Poisson arrivals and Lack of Anticipation: averages seen by an arriving customer are equal averages seen by an outside observer – PASTA property
- Little’s theorem for the waiting area only:
  $$E[Q] = \lambda E[W]$$
  $$E[W] = E[R] + \lambda E[X] \cdot E[W] = R + \rho E[W] \Rightarrow E[W] = \frac{E[R]}{1 - \rho}$$
- $\rho = \lambda E[X] = \lambda / \mu$ : utilization factor = proportion of time server is busy
  $$\rho = \lambda E[X] = 1 - p_o$$
- Calculate the average residual time: $E[R] = \lim_{i \to \infty} E[R_i]$
Average Residual Time

Graphical calculation of the long-term average of the residual time

Time-average residual time over \([0,t]\): \( t^{-1} \int_0^t R(s) ds \)

Consider time \(t\), such that \(R(t)=0\). Let \(D(t)\) be the number of departures in \([0,t]\) and assume that \(R(0)=0\). From the figure we see that:

\[
\frac{1}{t} \int_0^t R(s) ds = \frac{1}{t} \sum_{i=1}^{D(t)} \frac{X_i^2}{2} = \frac{1}{2} \cdot \frac{D(t)}{t} \cdot \sum_{i=1}^{D(t)} \frac{X_i^2}{D(t)} \Rightarrow
\]

\[
\lim_{t \to \infty} \frac{1}{t} \int_0^t R(s) ds = \frac{1}{2} \cdot \lim_{t \to \infty} \frac{D(t)}{t} \cdot \lim_{t \to \infty} \sum_{i=1}^{D(t)} \frac{X_i^2}{D(t)}
\]

Ergodicity: long-term time averages = steady-state averages (with probability 1)

\[
E[R] = \lim_{t \to \infty} E[R_t] = \lim_{t \to \infty} \frac{1}{t} \int_0^t R(s) ds
\]
Average Residual Time (cont.)

\[
\lim_{t \to \infty} \frac{1}{t} \int_0^t R(s)ds = \frac{1}{2} \lim_{t \to \infty} \frac{D(t)}{t} \cdot \lim_{t \to \infty} \sum_{i=1}^{D(t)} X_i^2 / D(t)
\]

\[\lim_{t \to \infty} \frac{D(t)}{t} : \text{long-term average departure rate. Should be equal to the long-term average arrival rate. Long-term averages = steady-state averages (with probability 1):}\]

\[\lim_{t \to \infty} \frac{D(t)}{t} = \lambda\]

\[\text{Law of large numbers:}\]

\[\lim_{t \to \infty} \frac{\sum_{i=1}^{D(t)} X_i^2}{D(t)} = \lim_{n \to \infty} \frac{\sum_{i=1}^{n} X_i^2}{n} = E[X^2]\]

\[\text{Average residual time:}\]

\[E[R] = \frac{1}{2} \lambda E[X^2]\]

\[\text{P-K Formula:}\]

\[E[W] = \frac{E[R]}{1 - \rho} = \frac{\lambda E[X^2]}{2(1 - \rho)}\]
P-K Formula:

- Average time a customer spends in the system
  \[
  E[T] = E[X] + E[W] = \frac{1}{\mu} + \frac{\lambda E[X^2]}{2(1 - \rho)}
  \]

- Average number of customers waiting for service:
  \[
  E[Q] = \lambda E[W] = \frac{\lambda^2 E[X^2]}{2(1 - \rho)}
  \]

- Average number of customers in the system (waiting or in service):
  \[
  E[N] = \lambda E[T] = \rho + \frac{\lambda^2 E[X^2]}{2(1 - \rho)}
  \]

- Averages \(E[W], E[T], E[Q], E[N]\) depend on the first two moments of the service time
**P-K Formula: Examples**

**M/D/1 Queue:** Deterministic service times all equal to $1/\mu$

- $E[X] = \frac{1}{\mu}$, \quad $E[X^2] = \frac{1}{\mu^2}$
- $E[W] = \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{\rho}{2\mu(1-\rho)}$, \quad $E[Q] = \frac{\lambda^2 E[X^2]}{2(1-\rho)} = \frac{\rho^2}{2(1-\rho)}$
- $E[T] = \frac{1}{\mu} + \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{1}{\mu} \left( 1 + \frac{\rho}{2\mu(1-\rho)} \right) = \frac{2-\rho}{2\mu(1-\rho)}$, \quad $E[N] = \lambda E[T] = \frac{\rho(2-\rho)}{2(1-\rho)}$

**M/M/1 Queue:** Exponential service times with mean $1/\mu$

- $E[X] = \frac{1}{\mu}$, \quad $E[X^2] = \frac{2}{\mu^2}$
- $E[W] = \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{\rho}{\mu(1-\rho)}$, \quad $E[Q] = \frac{\lambda^2 E[X^2]}{2(1-\rho)} = \frac{\rho^2}{(1-\rho)}$
- $E[T] = \frac{1}{\mu} + \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{1}{\mu} \left( 1 + \frac{\rho}{\mu(1-\rho)} \right) = \frac{1}{\mu - \lambda}$, \quad $E[N] = \lambda E[T] = \frac{\lambda}{\mu - \lambda}$
Distribution Upon Arrival or Departure

**Theorem 1:** For an M/G/1 queue at steady-state, the distribution of customers seen by an arriving customer is the same as that left behind by a departing customer.

**Proof:** Customers arrive one at a time and depart one at a time.

- $A(t), D(t)$: number of arrivals and departures (respectively) in $(0,t)$
- $U_n(t)$: number of $(n,n+1)$ transitions in $(0,t) = \text{number of arrivals that find system at state } n$
- $V_n(t)$: number of $(n+1,n)$ transitions in $(0,t) = \text{number of departures that leave system at state } n$
- $U_n(t)$ and $V_n(t)$ differ by at most 1 [when a $(n,n+1)$ transition occurs, another $(n,n+1)$ transition can occur only if the state has moved back to $n$, i.e., after a $(n+1,n)$ transition has occurred]
- Stationary probability that an arriving customer finds the system at state $n$:
  \[
  \alpha_n = \lim_{t \to \infty} P\{N(t) = n \mid \text{arrival at } t^+\}
  \]
  - $\alpha_n$ is the proportion of arrivals that find the system at state $n$:
  \[
  \alpha_n = \lim_{t \to \infty} \frac{U_n(t)}{A(t)}
  \]
- Similarly, stationary probability that a departing customer leaves system at state $n$:
  \[
  \beta_n = \lim_{t \to \infty} \frac{V_n(t)}{D(t)}
  \]
- Noting that $\lim_{t \to \infty} \frac{A(t)}{t} = \lim_{t \to \infty} \frac{D(t)}{t} = \lambda$, we have:
  \[
  \lim_{t \to \infty} \frac{U_n(t)}{t} = \lim_{t \to \infty} \frac{V_n(t)}{t} \Rightarrow \lim_{t \to \infty} \frac{U_n(t)}{A(t)} \frac{A(t)}{t} = \lim_{t \to \infty} \frac{V_n(t)}{D(t)} \frac{D(t)}{t} \Rightarrow \alpha_n = \beta_n
  \]
Theorem 2: For an M/G/1 queue at steady-state, the probability that an arriving customer finds \( n \) customers in the system is equal to the proportion of time that there are \( n \) customers in the system. Therefore, the distribution seen by an arriving customer is identical to the stationary distribution.

Proof: Identical to the PASTA theorem due to:
- Poisson arrivals
- Lack of anticipation: future arrivals independent of current state \( N(t) \)

Theorem 3: For an M/G/1 queue at steady-state, the system appears statistically identical to an arriving and a departing customer. Both an arriving and a departing customer, at steady-state, see a system that is statistically identical to the one seen by an observer looking at the system at an arbitrary time.

Analysis of the M/G/1 Queue:
- Consider the embedded Markov chain resulting by observing the system at departure epochs
- At steady-state, the embedded Markov chain and \( \{N(t)\} \) are statistically identical
- Stationary distribution \( p_n \) is equal to the stationary distribution of the embedded Markov chain
Embedded Markov Chain

- $s_j$: time of $j^{th}$ departure
- $L_j = N(s_j)$: number of customers left behind by the $j^{th}$ departing customer

Show that $\{L_j : j \geq 1\}$ is a Markov chain

- If $L_{j-1} \geq 1$: customer $j$ enters service immediately at time $s_{j-1}$. Then:
  $$L_j = L_{j-1} - 1 + A_j, \quad \text{if } L_{j-1} \geq 1$$

- If $L_{j-1} = 0$: customer $j$ arrives after time $s_{j-1}$ and departs at time $s_j$. Then:
  $$L_j = A_j, \quad \text{if } L_{j-1} = 0$$

Combining the above:
$$L_j = L_{j-1} + A_j - 1\{L_{j-1} > 0\}$$

- $A_j$: number of arrivals during service time $X_j$:
  $$P\{A_j = k\} = \int_0^\infty P\{A_j = k \mid X_j = t\} f_X(t)dt = (1/k!) \int_0^\infty e^{-\lambda t} (\lambda t)^k f_X(t)dt$$

- $A_1, A_2, \ldots$: independent – arrivals in disjoint intervals

- $L_j$ depends on the past only through $L_{j-1}$. Thus: $\{L_j : j \geq 1\}$ is a Markov chain
Number of Arrivals During a Service Time

- $A_1, A_2, \ldots$: iid. Drop the index $j$ – equivalent to considering the system at steady state

$$a_k = P\{A = k\} = \int_0^\infty P\{A = k \mid X = t\} f_X(t) dt = \frac{1}{k!} \int_0^\infty e^{-\lambda t} (\lambda t)^k f_X(t) dt, \quad k = 0, 1, \ldots$$

- Find the first two moments of $A$

**Proposition:** For the number of arrivals $A$ during service time $X$, we have:

$$E[A] = \lambda E[X] = \rho$$
$$E[A^2] = \lambda^2 E[X^2] + \lambda E[X]$$

**Proof:** Given $X=t$, the number of arrivals $A$ follows the Poisson distribution with parameter $\lambda t$.

$$E[A] = \int_0^\infty E[A \mid X = t] f_X(t) dt = \int_0^\infty (\lambda t) f_X(t) dt = \lambda \int_0^\infty t f_X(t) dt = \lambda E[X]$$

$$E[A^2] = \int_0^\infty E[A^2 \mid X = t] f_X(t) dt = \int_0^\infty (\lambda^2 t^2 + \lambda t) f_X(t) dt$$
$$= \lambda^2 \int_0^\infty t^2 f_X(t) dt + \lambda \int_0^\infty t f_X(t) dt = \lambda^2 E[X^2] + \lambda E[X]$$

**Lemma:** Let $Y$ be a RV following the Poisson distribution with parameter $\alpha > 0$. Then:

$$E[Y] = \sum_{k=1}^\infty k \cdot e^{-\alpha} \frac{\alpha^k}{k!} = \alpha, \quad E[Y^2] = \sum_{k=1}^\infty k^2 \cdot e^{-\alpha} \frac{\alpha^k}{k!} = \alpha^2 + \alpha$$
Embedded Markov Chain

- Transition Probabilities: $P_{ij} = P\{L_{n+1} = j | L_n = i\}$
  
  $P_{0j} = \alpha_j, \quad j = 0, 1, ...$
  
  $P_{ij} = \begin{cases}
  \alpha_{j-i+1}, & j \geq i - 1 \\
  0, & j < i - 1
  \end{cases}, \quad i \geq 1$

- Stationary distribution: $\pi_j = \lim_{n \to \infty} P\{L_n = j\}$
  
  $\pi_j = \pi_0 \alpha_j + \pi_1 \alpha_j + \cdots + \pi_j \alpha_1 + \pi_{j+1} \alpha_0, \quad j \geq 1$
  
  $\pi_0 = \pi_0 \alpha_0 + \pi_1 \alpha_0$

  or $\pi = \pi P, \quad P = \begin{bmatrix}
  \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\
  \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\
  0 & \alpha_0 & \alpha_1 & \cdots \\
  \vdots & \vdots & \vdots & \ddots
  \end{bmatrix}$

  Unique solution: $\pi_j$ is the fraction of departing customers that leave $j$ customers behind

  From Theorem 3: $\pi_j$ is also the proportion of time that there are $j$ customers in the system
Calculating the Stationary Distribution

- Applying Little’s Theorem for the server, the proportion of time that the server is busy is:
  \[ 1 - \pi_0 = \lambda E[T] = \rho \Rightarrow \pi_0 = 1 - \rho \]

- Stationary distribution can be calculated iteratively:
  \[ \pi_0 = \pi_0 \alpha_0 + \pi_1 \alpha_0 \]
  \[ \pi_1 = \pi_0 \alpha_1 + \pi_1 \alpha_1 + \pi_2 \alpha_0 \]
  ...

- Iterative calculation might be prohibitively involved
- Often, we want to find only the first few moments of the distribution, e.g., \( E[N] \) and \( E[N^2] \)

- We will present a general methodology based on \( z \)-transforms that can be used to
  1. Find the moments of the stationary distribution without calculating the distribution itself
  2. Find the stationary distribution, in special cases
  3. Derive approximations of the stationary distribution
Moment Generating Functions

**Definition:** Moment generating function of random variable $X$; for any $t \in \mathbb{R}$

$$M_X(t) = E[e^{tX}] = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx, & X \text{ continuous} \\ \sum_j e^{tx_j} P\{X = x_j\}, & X \text{ discrete} \end{cases}$$

**Theorem 1:** If the moment generating function $M_X(t)$ exists and is finite in some neighborhood of $t=0$, it determines the distribution (pdf or pmf) of $X$ uniquely.

**Theorem 2:** For any positive integer $n$:

1. $\frac{d^n}{dt^n} M_X(t) = E[X^n e^{tX}]$

2. $\frac{d^n}{dt^n} M_X(0) = E[X^n]$

**Theorem 3:** If $X$ and $Y$ are independent random variables:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$
Z-Transforms of Discrete Random Variables

- For a discrete random variable, the moment generating function is a polynomial of $e^t$.
- It is more convenient to set $z = e^t$ and define the $z$-transform (or characteristic function):

$$G_X(z) = E[z^X] = \sum_{j} z^{x_j} P\{X = x_j\}$$

- Let $X$ be a discrete random variable taking values 0, 1, 2,..., and let $p_n = P\{X = n\}$. The $z$-transform is well-defined for $|z| < 1$:

$$G_X(z) = p_0 + zp_1 + z^2 p_2 + z^3 p_3 + \cdots = \sum_{n=0}^{\infty} p_n z^n$$

- $Z$-transform uniquely determines the distribution of $X$
- If $X$ and $Y$ are independent random variables: $G_{X+Y}(z) = G_X(z)G_Y(z)$

- Calculating factorial moments:

$$\lim_{z \to 1} G'_X(z) = \lim_{z \to 1} \sum_{n=1}^{\infty} n p_n z^{n-1} = \sum_{n=1}^{\infty} n p_n = E[X]$$

$$\lim_{z \to 1} G''_X(z) = \lim_{z \to 1} \sum_{n=2}^{\infty} n(n-1) p_n z^{n-2} = \sum_{n=2}^{\infty} n(n-1) p_n = E[X(X-1)]$$

- Higher factorial moments can be calculated similarly.
## Continuous Random Variables

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform over $(a, b)$</td>
<td>$\frac{1}{b-a}$ $a &lt; x &lt; b$</td>
<td>$\frac{e^{tb} - e^{ta}}{t(b-a)}$</td>
<td>$\frac{a+b}{2}$</td>
<td>$\frac{(b-a)^2}{12}$</td>
</tr>
<tr>
<td>Exponential $\lambda$</td>
<td>$\lambda e^{-\lambda x}$ $x \geq 0$</td>
<td>$\frac{\lambda}{\lambda - t}$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda}$</td>
</tr>
<tr>
<td>Normal $(\mu, \sigma^2)$</td>
<td>$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty &lt; x &lt; \infty$</td>
<td>$e^{\mu t + (\sigma t)^2/2}$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>
## Discrete Random Variables

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binomial</strong> $(n, p)$</td>
<td>$(\binom{n}{k})p^k(1-p)^{n-k}$</td>
<td>$(pe^t + 1 - p)^n$</td>
<td>$np$</td>
<td>$np(1 - p)$</td>
</tr>
<tr>
<td></td>
<td>$k = 0, 1, \ldots, n$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Geometric</strong> $p$</td>
<td>$(1-p)^k p$</td>
<td>$\frac{pe^t}{1-(1-p)e^t}$</td>
<td>$\frac{1}{p}$</td>
<td>$\frac{1-p}{p^2}$</td>
</tr>
<tr>
<td></td>
<td>$k = 1, 2, \ldots$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Negative Bin.</strong> $(r, p)$</td>
<td>$(\binom{k-1}{r-1})p^r(1-p)^{k-r}$</td>
<td>$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$</td>
<td>$\frac{r}{p}$</td>
<td>$\frac{r(1-p)}{p^2}$</td>
</tr>
<tr>
<td></td>
<td>$k = r, r + 1, \ldots$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Poisson</strong> $\lambda$</td>
<td>$e^{-\lambda}\frac{\lambda^k}{k!}$</td>
<td>$e^\lambda(e^t - 1)$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td></td>
<td>$k = 0, 1, \ldots$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
P-K Transform Equation

We have established:

\[ L_j = L_{j-1} - 1 \{L_{j-1} > 0\} + A_j = (L_{j-1} - 1)^+ + A_j \]

Let \( \pi_n = \lim_{j \to \infty} P\{L_j = n\} \) be the stationary distribution and \( G_L(z) = \sum_{n=0}^{\infty} \pi_n z^n \) its \( z \)-transform.

Noting that \((L_{j-1} - 1)^+\) and \(A_j\) are independent, we have:

\[ E[z^{L_j}] = E[z^{(L_{j-1})^+}]E[z^{A_j}] \]

At steady-state, \(L_j \) and \(L_{j-1}\) are statistically identical, with pmf \( \pi_n \). Therefore:

\[ E[z^{L_j}] = E[z^{L_{j-1}}] = G_L(z) \]

Moreover:

\[ E[z^{A_j}] = G_A(z) = \sum_{n=0}^{\infty} \alpha_n z^n \]

Let \( X \) be a discrete random variable taking values 0, 1, 2,\ldots, and let \( p_n = P\{X = n\} \). Then:

\[ E[z^{(X-1)^+}] = p_0 + p_1 + z p_2 + z^2 p_3 + \cdots = p_0 + z^{-1}(E[z^X] - p_0) \]

Therefore:

\[ E[z^{(L_{j-1})^+}] = \pi_0 + z^{-1}(E[z^{L_{j-1}}] - \pi_0) = \pi_0 + z^{-1}(G_L(z) - \pi_0) \]

Then:

\[ G_L(z) = [\pi_0 + z^{-1}(G_L(z) - \pi_0)]G_A(z) \Rightarrow G_L(z) = \frac{\pi_0(z-1)G_A(z)}{z-G_A(z)} \]
P-K Transform Equation

Probability \( \pi_0 \) can be calculated by requiring \( \lim_{z \to 1} G_L(z) = \sum_{n=0}^{\infty} \pi_n = 1 \)

Using L’Hospital’s rule:

\[
1 = \lim_{z \to 1} \frac{\pi_0 [G_A(z) + (z-1)G'_A(z)]}{1-G'_A(z)} = \frac{\pi_0}{1-E[A]} \Rightarrow \pi_0 = 1 - E[A]
\]

Recall that \( E[A] = \lambda E[X] = \rho \). For \( \pi_0 > 0 \), we must have: \( \rho = E[A] < 1 \). Finally:

\[
G_A(z) = \sum_{n=0}^{\infty} z^k \alpha_k = \sum_{n=0}^{\infty} z^k \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^k}{k!} f_X(x)dx
\]

\[
= \int_0^\infty e^{-\lambda x} \sum_{n=0}^{\infty} \frac{(\lambda x)^k}{k!} f_X(x)dx = \int_0^\infty e^{\lambda(z-1)x} f_X(x)dx = M_X(\lambda(z-1))
\]

where \( M_X(t) = E[e^{tx}] = \int_0^\infty e^{tx} f_X(x)dx \) is the moment generating function of the service time \( X \)

- At steady-state the number of customers left behind by a departing customer and the number of customers in the system are statistically identical, i.e., \( \{L_j\} \) and \( \{N(t)\} \) have the same pmf
- Concluding:

\[
G_N(z) = G_L(z) = \frac{(1-\rho)(z-1)G_A(z)}{z-G_A(z)} = \frac{(1-\rho)(z-1)M_X(\lambda(z-1))}{z-M_X(\lambda(z-1))}
\]
**Example 1: M/M/1 Queue.** $X$ is exponentially distributed with mean $1/\mu$. Then:

$$M_X(t) = E[e^{tx}] = \frac{\mu}{\mu-t}$$

Then, the $z$-transform of the number of arrivals during a service time is:

$$G_\lambda(z) = M_X(\lambda(z-1)) = \frac{\mu}{\mu-\lambda(z-1)} = \frac{\mu}{\mu+\lambda-\lambda z}$$

The $P-K$ Transform equation, then, gives:

$$G_N(z) = \frac{(1-\rho)(z-1)G_\lambda(z)}{z-G_\lambda(z)} = \frac{(1-\rho)(z-1)\mu}{\mu+\lambda-\lambda z} = \frac{1-\rho}{1-\rho z}$$

For $|z|<\rho$:

$$\frac{1}{1-\rho z} = 1 + \rho z + \rho^2 z^2 + \cdots$$

Then:

$$G_N(z) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \cdot z^n$$

Therefore:

$$p_n = (1-\rho)\rho^n, \quad n \geq 0$$
Expansion in Partial Fractions

Assume that the $z$-transform is of the form:

$$G(z) = \frac{U(z)}{V(z)},$$

$U(z)$ and $V(z)$ polynomials without common roots. Let $z_1, \ldots, z_m$ be the roots of $V(z)$. Then:

$$V(z) = (z - z_1)(z - z_2)\cdots(z - z_m)$$

Expansion of $G(z)$ in partial fractions:

$$G(z) = \frac{\Phi_1}{z_1 - z} + \frac{\Phi_2}{z_2 - z} + \cdots + \frac{\Phi_m}{z_m - z}$$

Given such an expansion, for $|z| < |z_k|:

$$\frac{1}{1 - z/z_k} = 1 + \frac{z}{z_k} + \left(\frac{z}{z_k}\right)^2 + \cdots$$

Then:

$$G(z) = \sum_{k=1}^{m} \frac{\Phi_k}{z_k} \frac{1}{1 - z/z_k} = \sum_{k=1}^{m} \frac{\Phi_k}{z_k} \sum_{n=0}^{\infty} \left(\frac{z}{z_k}\right)^n = \sum_{n=0}^{\infty} \sum_{k=1}^{m} \frac{\Phi_k}{z_k^{n+1}} z^n$$

Therefore:

$$p_n = \sum_{k=1}^{m} \frac{\Phi_k}{z_k^{n+1}}$$
Expansion in Partial Fractions (cont.)

\[ G(z) = \frac{U(z)}{V(z)}, \quad V(z) = (z - z_1)(z - z_2) \cdots (z - z_m) \]

Expansion of \( G(z) \) in partial fractions:

\[ G(z) = \frac{\varphi_1}{(z_1 - z)} + \frac{\varphi_2}{(z_2 - z)} + \cdots + \frac{\varphi_m}{(z_m - z)} \]

Determining the coefficients of the partial fractions:

\[ \varphi_1 = \lim_{z \to z_1} (z_1 - z)G(z) \]

Note that:

\[ \lim_{z \to z_1} (z_1 - z)G(z) = - \frac{U(z_1)}{(z_1 - z_2) \cdots (z_1 - z_m)} = - \frac{U(z_1)}{V'(z_1)} \]

Therefore, the coefficients can be determined as:

\[ \varphi_k = \lim_{z \to z_k} (z_k - z)G(z) = - \frac{U(z_k)}{V'(z_k)} \]
M/G/1 Queue with Priority Classes

- M/G/1 system with arriving customers divided in $c$ priority classes
- Class 1 highest priority, class 2 second highest, up to class $c$ which is the lowest priority class
- Class $k$ customers arrive according to a Poisson process with rate $\lambda_k$
- Service times of class $k$ customers are iid, following a general distribution with mean $\bar{X}_k = 1/\mu_k$ and second moment $\bar{X}_k^2$
- Arrival processes are independent of each other and independent of the service times
- $\rho_k = \lambda_k \bar{X}_k = \frac{\lambda_k}{\mu_k}$ : utilization factor for class $k$
- $W_k$ : average queueing time (at steady state) of class $k$ customers
- Preemptive or non-preemptive priority discipline
- Develop a formula that gives the average queueing time for each priority class
Non-Preemptive Priority

- Non-preemptive: service of a customer completes uninterrupted, even if customers of higher priority arrive in the meantime

- A separate queue is maintained for each class; each time the server becomes free, the first customer in the highest priority queue (that is not empty) enters service

- Non-preemptive policy: the mean residual service \( R \) time seen by an arriving customer is the same for all priority classes

**Priority Class 1**

- Queueing time of class 1 customer = residual service time + time required to service all class 1 customers found in the queue upon arrival

- Similarly to the derivation of \( P-K \) formula, this implies:

\[
W_1 = R + \frac{1}{\mu_1} Q_1
\]

- Little’s formula: \( Q_1 = \lambda_1 W_1 \)

- Combining the two:

\[
W_1 = \frac{R}{1 - \rho_1}
\]
Non-Preemptive Priority

Priority Class 2:

- Queueing time for class 2 customer is the sum of the following:
  1. Residual service time
  2. Time to service all class 1 customers found in queue upon arrival
  3. Time to service all class 2 customers found in queue upon arrival
  4. Time to service all class 1 customers that arrive while the customer waits in the queue

- Focusing on averages of these times:

\[
W_2 = R + \frac{1}{\mu_1}Q_1 + \frac{1}{\mu_2}Q_2 + \frac{1}{\mu_1}\lambda_1W_2
\]

\[
= R + \frac{1}{\mu_1}\lambda_1W_1 + \frac{1}{\mu_2}\lambda_2W_2 + \frac{1}{\mu_1}\lambda_1W_2
\]

- Solving for \( W_2 \) and using the expression for \( W_1 \):

\[
W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}
\]
Non-Preemptive Priority

- Priority Class $k$: Using induction
  \[ W_k = \frac{R}{(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_{k-1} - \rho_k)} \]

- Mean Residual Service Time: Using the graphical method developed in the proof of the $P-K$ formula, one can show:
  \[ R = \frac{1}{2} \sum_{k=1}^{c} \lambda_k \bar{X}_k^2 \]

- Average time a class $k$ customer spends in the system:
  \[ T_k = \frac{1}{\mu_k} + W_k \]

- Using Little’s formula the number of customers in the system for each class, and for all customers, the average time delay per customer:
  \[ T = \lambda_1 T_1 + \cdots + \lambda_c T_c \]
  \[ \frac{\lambda_1 + \cdots + \lambda_c}{\lambda_1 + \cdots + \lambda_c} \]
Preemptive Priority

- Non-preemptive policy: average queueing time of a customer depends on the arrival rate of lower priority customers
- Preemptive/resume policy: service of a customer is interrupted when a higher priority customer arrives; it resumes from the point of interruption when all higher priority customers have been served
- Priority $k$ customers are not affected by the presence of lower priority customers
- Calculate $T_k = \text{average time a class } k \text{ customer spends in the system. This consists of:}$

1. Average service time of the customer $1/\mu_k$

2. Average time required to service customers of priority 1 through $k$ found in the system upon arrival. This is equal to the average waiting time in an M/G/1 system where customers of priority lower than $k$ are neglected, that is:

$$R_k = {R_k \over 1 - \rho_1 - \cdots - \rho_k}, \quad R_k = \frac{1}{2} \sum_{i=1}^{k} \lambda_i \bar{X}_i^2$$

3. Average time required to service customers of priority higher than $k$ that arrive while the customer is in the system:

$$\sum_{i=1}^{k-1} \frac{1}{\mu_i} \lambda_i T_k = \sum_{i=1}^{k-1} \rho_i T_k, \quad k > 1$$
Preemptive Priority

Combining terms:

\[ T_k = \frac{1}{\mu_k} + \frac{R_k}{1 - \rho_1 - \cdots - \rho_k} + T_k (\rho_1 + \cdots + \rho_{k-1}) \]

Final solution:

\[ T_k = \frac{(1/\mu_k)(1 - \rho_1 - \cdots - \rho_k) + R_k}{(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)} \]

where:

\[ R_k = \frac{1}{2} \sum_{i=1}^{k} \lambda_i \bar{X}_i^2 \]