Single Queuing Systems

**M/M/1 queuing system**

— arrival process is a Poisson process (or the inter-arrival time is exponentially distributed)

— service process is also a Poisson process (or the service time is exponentially distributed)

*advantage:* a mathematically tractable model with solutions applicable to a wide variety of situations.

A counting process \( \{N(t), t \geq 0\} \) representing the # of events that have occurred up to time t.
Poisson process with an average arrival rate $\lambda$: $\lambda$ is the proportionality constant

$$\begin{array}{c|c|c|c|c|c|c} 
 & t & \Delta t & \Delta t & \Delta t & \Delta t & \cdots \\
\hline
\hline
\end{array}$$

time

$Pr(\text{exactly 1 arrival in } [t, t+\Delta t]) = \lambda \Delta t$

$Pr(\text{no arrivals in } [t, t+\Delta t]) = 1 - \lambda \Delta t$

Pr\left\{ \begin{array}{l}
1 \text{ event occurs at } t + \Delta t \\
\text{event does not occur at } t
\end{array} \right\}

= 1 - e^{-\lambda \Delta t}

= 1 - \left\{ 1 + (-\lambda \Delta t) + \frac{(-\lambda \Delta t)^2}{2!} + \cdots \right\}

\approx \lambda \Delta t

Analogy:

Coin flipping: results of coin flips are independent

Arrivals are also independent
Let $P_n(t) \equiv P(\# \text{ of arrivals} = n \text{ at time } t)$

$P_{ij}(\Delta t) \equiv \text{the prob. of going from } i \text{ arrivals to } j \text{ arrivals in a time interval of } \Delta t \text{ seconds}$

$P_0(t + \Delta t) = P_0(t)P_{0,0}(\Delta t) = \lambda \Delta t$

$P_n(t + \Delta t) = P_n(t)P_{n,n}(\Delta t) + P_{n-1}(t)P_{n-1,n}(\Delta t) \equiv \lambda \Delta t$

$P_n(t + \Delta t) - P_n(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$

$P_0(t + \Delta t) - P_0(t) = -\lambda P_0(t)$

$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t)$

$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t)$

\[ \begin{array}{c}
0 \quad \lambda \quad 1 \quad \lambda \quad 2 \quad \lambda \quad 3 \quad \lambda \quad \ldots
\end{array} \]
Let $\Delta t \to 0$ then

\[
\frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t) \quad \text{(1)}
\]

\[
\frac{dP_0(t)}{dt} = -\lambda P_0(t) \quad \text{(2) solution} \Rightarrow P_0(t) = e^{-\lambda t}
\]

From (1)

\[
\frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda e^{-\lambda t} \quad \therefore P_1 = \lambda te^{-\lambda t}
\]

\[
\frac{dP_2(t)}{dt} = -\lambda P_2(t) + \lambda^2 te^{-\lambda t} \quad \therefore P_2 = \frac{\lambda^2 t^2}{2} e^{-\lambda t}
\]
Continuing, by induction,

\[ P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \] — Poisson distribution

meaning: prob. of n arrivals in an interval of t seconds

\[ \sum_{n=0}^{\infty} P_n(t) = 1 \]

\[ \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} = 1 \]

\[ \sum_{n=0}^{\infty} \frac{X^n}{n!} = e^X \]

The Poisson distribution

Ex: \( \lambda = 100 \) arrivals/min., what is the prob. of no arrivals in 5 sec.?

\[ P_0(5 \text{ sec.}) = e^{-100 \left( \frac{1}{12} \right)} = 0.00024 \]
* the mean & the variance of the Poisson dist. are both equal to $\lambda t$.

**Mean:**

\[
n(t) = \sum_{n=1}^{\infty}nP_n(t) = \sum_{n=1}^{\infty} n \frac{(\lambda t)^n e^{-\lambda t}}{n!} = e^{-\lambda t} \cdot \sum_{n=1}^{\infty} \frac{(\lambda t)^n}{(n-1)!}
\]

**Variance:**

\[
[\sigma_n(t)]^2 = \sum_{n=0}^{\infty} n^2 P_n(t) - [n(t)]^2 = \lambda t
\]

The inter-arrival time $T$ is a random variable

- Inter-arrival time cumulative dist. function $t$ = P (time between arrivals ≤ $t$)
- $= 1$ - P (time between arrivals > $t$)  \( \because \) no arrivals in a time interval of $t = P_0(t)$
- $= 1 - P_0(t)$
- $= 1 - e^{-\lambda t}$

\[d(1-e^{-\lambda t}) = \lambda e^{-\lambda t}\]

\[\therefore \ T\ is\ an\ exponentially\ distributed\ r.v.\]

\[\therefore \ T\ has\ a\ memory-less\ property\]
e.g., an average of mean interarrival time $= 20$ min.\[ \frac{1}{\lambda} = \text{mean interarrival time} \]

The last train arrived 19 minutes ago. What is the expected time until the next train arrives?

- 20 min. $\times$
- prob. yes $\lambda\Delta t$
- no $1-\lambda\Delta t$

**Coin flipping explanation in $\Delta t**

Memory-less property $\equiv$ Markov property

$$P(T > t_0 + t | T > t_0) = P(T > t)$$

* Definition: a **Markov chain** is a Markov process with a discrete state space.
Probability flux (or flow):
\[(\text{probability of a state}) \ast (\text{transition rate originating from the state})\]

physical meaning: \# of times per second the event corresponding to the transition occurs.

\[P_n(t + \Delta t) = P_n(t)P_{n,n}(\Delta t) + P_{n-1}(t)P_{n-1,n}(\Delta t) + P_{n+1}(t)P_{n+1,n}(\Delta t)\]

\[P_0(t + \Delta t) = P_0(t)P_{0,0}(\Delta t) + P_1(t)P_{1,0}(\Delta t)\]

\[\frac{dP_n(t)}{dt} = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)\]

\[\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t)\]
\[
\frac{dP_n(t)}{dt} = -\left(\lambda + \mu\right)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)
\]
flows out of state \(n\)  flows into state \(n\)
flows out of state \(n\)  flows into state \(n\)

Study:
— transient behavior \(\frac{dP_n(t)}{dt} \neq 0\)
— equilibrium behavior \(\frac{dP_n(t)}{dt} = \frac{dP_{n-1}(t)}{dt} = \ldots = \frac{dP_0(t)}{dt} = 0\)

Global balance equations: a set of linear equations for \(t \rightarrow \infty\)
Equilibrium state probabilities  
\[ \sum_{i=0}^{\infty} P_i(\infty) = 1 \]  

conservation of probability:  

normalization equation  

Use local balance equations to solve the global balance equations  
1. Local satisfies global  
2. Local allows us to relate \( P_n \) with a reference state, e.g., \( P_0 \)  

Definition of local balance:  
“the probability flow into a state due to an arrival to a queue equals the probability flow out of the same state due to a departure from the same queue”
\[ P_0: \quad P_0 \]
\[ P_1: \quad P_0 \lambda = P_1 \mu \quad \Rightarrow \quad P_1 = \left( \frac{\lambda}{\mu} \right) P_0 \]
\[ P_2: \quad P_1 \lambda = P_2 \mu \quad \Rightarrow \quad P_2 = \left( \frac{\lambda}{\mu} \right) P_1 \]
\[ P_{n-1} \lambda = P_n \mu \quad \Rightarrow \quad P_n = \left( \frac{\lambda}{\mu} \right) P_{n-1} \]
\[ P_n = \left( \frac{\lambda}{\mu} \right)^n P_0 \]

Applying the normalization equation

\[ P_0 \sum_{n=0}^{\infty} \left( \frac{\lambda}{\mu} \right)^n = 1 \quad \sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad 0 \leq x < 1 \]

\[ P_0 = \frac{1}{\sum_{n=0}^{\infty} \left( \frac{\lambda}{\mu} \right)^n} = 1 - \left( \frac{\lambda}{\mu} \right) \quad \text{if} \quad 0 \leq \frac{\lambda}{\mu} < 1 \]

\[ \therefore P_n = \left( \frac{\lambda}{\mu} \right)^n \left( 1 - \frac{\lambda}{\mu} \right) \]
Utilization: prob. that an M/M/1 queuing system is nonempty

Let \( \frac{\lambda}{\mu} = \rho \), \( P_n = \rho^n (1 - \rho) \)

\( P_n \) for M/M/1 system when \( \frac{\lambda}{\mu} = \rho = \frac{1}{2} \)

* for a lightly loaded system, there are usually less than 4 customers in the system.

<table>
<thead>
<tr>
<th>( P(\mathrm{0} \leq n \leq 3) )</th>
<th>(light) ( \rho=0.1 )</th>
<th>( \rho=0.5 )</th>
<th>(heavy) ( \rho=0.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^{-4} )</td>
<td>0.0586</td>
<td>0.2256</td>
<td></td>
</tr>
<tr>
<td>( 10^{-8} )</td>
<td>0.0366</td>
<td>0.1480</td>
<td></td>
</tr>
<tr>
<td>( 10^{-12} )</td>
<td>0.000249</td>
<td>0.2825</td>
<td></td>
</tr>
</tbody>
</table>
check \( \rho = 1 - P_0 \)?

\[
\rho = 1 - P_0 = 1 - (1 - \frac{\lambda}{\mu}) = \frac{\lambda}{\mu}
\]

* \( \rho \leq 1 \) otherwise \( \lambda > \mu \) and the queuing system would no longer be in equilibrium \( \rightarrow \) i.e., unstable.

Q1: throughput?

\[
x = \sum_{n=1}^{\infty} P_n \cdot \mu
\]

\[
= \mu \cdot \sum_{n=1}^{\infty} P_n
\]

\[
= \mu (1 - P_0)
\]

\[
= \mu \cdot \frac{\lambda}{\mu} = \lambda
\]

because when there is no customer, there is no contribution to throughput.

Q2: Average # of customers in the queuing system?

\[
\overline{n} = \sum_{n=0}^{\infty} n \cdot P_n = \sum_{n=0}^{\infty} n \cdot \rho^n (1 - \rho) = (1 - \rho) \sum_{n=0}^{\infty} n \cdot \rho^n = (1 - \rho) \cdot \rho \sum_{n=0}^{\infty} n \cdot \rho^{n-1}
\]

\[
= (1 - \rho) \cdot \rho \cdot \frac{d}{d\rho} \sum_{n=0}^{\infty} \rho^n = (1 - \rho) \cdot \rho \cdot \frac{d}{d\rho} \left( \frac{1}{1 - \rho} \right) = (1 - \rho) \cdot \rho \cdot \frac{1}{(1 - \rho)^2}
\]

\[
= \frac{\rho}{1 - \rho}
\]

\[
\overline{n} = \frac{\rho}{1 - \rho}
\]

e.g.,

\[
\lambda = \frac{1}{2} \mu \quad \lambda = \frac{2}{3} \mu
\]

\[
\therefore \rho = \frac{\lambda}{\mu} \quad \therefore \rho = \frac{2}{3}
\]

\[
= \frac{\frac{1}{2} \mu}{\mu} = \frac{1}{2} \quad \overline{n} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2
\]

\[
\overline{n} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1
\]

76
Let R be the mean response time per customer

Q3: R?

since \( n = \lambda R \) by little’s law (to be discussed later)

\[
R = \frac{n}{\lambda} = \frac{\rho}{(1 - \rho)\lambda} = \frac{1/\mu}{(1 - \frac{\lambda}{\mu})}
\]

\[\text{When } \rho = 1 \text{ system is unstable}\]

M/M/1: average # of customers \( \bar{n} \) as a function of \( \rho \)

M/M/1 service time waiting time

\[R = (1 + n) \cdot D\]

\[= (1 + n) \cdot \frac{1}{\mu}\]

\[= \left(1 + \frac{\rho}{1 - \rho}\right) \cdot \frac{1}{\mu}\]

\[= \frac{1}{1 - \rho} \cdot \left(\frac{1}{\mu}\right)\]
M/M/1/N Queuing system: the finite buffer case

Following the previous derivation for M/M/1/∞,

\[
P_n = \left(\frac{\lambda}{\mu}\right)^n P_0
\]

\&

\[
P_0 = \frac{1}{\sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} = \frac{1 - \left(\frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}}
\]

\[
\therefore P_n = \frac{1 - \left(\frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} \cdot \left(\frac{\lambda}{\mu}\right)^n
\]

an arriving customer is “lost” or “turned away” when there are already N customers in the system.

no restriction on the range of \(\frac{\lambda}{\mu}\)

\[
N \not\in P_N \not\in \lambda /\mu \not\in P_N \not\in
\]
Q1: the prob. that the queuing system is full? \( P_N \)
Q2: how fast are customers lost? \( P_N \times \lambda \)

\[
\frac{\lambda}{\mu} \quad P_N = P_5 \quad \text{(blocking probability)}
\]

when \( N=5 \)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>9*10^{-6}</td>
</tr>
<tr>
<td>0.5</td>
<td>0.016</td>
</tr>
<tr>
<td>0.75</td>
<td>0.072</td>
</tr>
<tr>
<td>1.00</td>
<td>0.166</td>
</tr>
<tr>
<td>2.00</td>
<td>0.508</td>
</tr>
<tr>
<td>5.00</td>
<td>0.800</td>
</tr>
</tbody>
</table>

\[
\lim_{x \to 1} \frac{x^5 - x^6}{1 - x^6} = \lim_{x \to 1} \frac{5x^4 - 6x^5}{-6x^5} = 0.166
\]

applying L’Hopital’s rule

Q3: population?

\[
\bar{n} = \sum_{n=0}^{N} n \cdot P_n
\]

Q4: throughput?

\[
x = \sum_{n=1}^{N} \mu \cdot P_n = \mu \cdot \sum_{n=1}^{N} P_n = \mu (1 - P_0) = \mu \rho < \lambda
\]

Q5: Utilization?

\[
\rho = 1 - P_0 = 1 - \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} < \frac{\lambda}{\mu}
\]

\( \rho \) is utilization
Variations of M/M/1

**M/M/∞:**

- infinite # of servers

\[ \lambda = \sum_{n=1}^{\infty} n \cdot P_n = \sum_{n=1}^{\infty} n \frac{\left( \frac{\lambda}{\mu} \right)^n}{n!} = \frac{\lambda}{\mu} \]

\[ P_n = \left( \prod_{i=1}^{n} \frac{\lambda}{i\mu} \right) \cdot P_0 = \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n P_0 \]

\[ P_0 = \frac{1}{\sum_{n=0}^{\infty} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n} = \frac{1}{e^{\frac{\lambda}{\mu}}} = e^{-\frac{\lambda}{\mu}} \]

\[ P_n = \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n e^{-\frac{\lambda}{\mu}} \]

**Q1:** throughput? \( \lambda \)

**Q2:** response time? \( \frac{1}{\mu} \)

**Q3:** population? \( \frac{\lambda}{\mu} \)

by Little’s law

\(-\)

Q1: throughput? \( \lambda \\
Q2: response time? \ 1/\mu \\
Q3: population? \ \lambda /\mu \\
by Little’s law

\[ P_n = \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n e^{-\lambda/\mu} \]
**M/M/m**

![M/M/m Diagram](image)

- **m servers**
- e.g., a system with m processors

**Solution:**

\[
P_n = \frac{1}{m! m^{n-m}} \left( \frac{\lambda}{\mu} \right)^n \cdot P_0
\]

- can be obtained by considering 2 cases separately
- \( \mu_n = \begin{cases} \mu_n & 0 \leq n \leq m \\ m \mu & n \geq m \end{cases} \)

- \( P_0 = \left[ 1 + \sum_{n=1}^{m-1} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n + \frac{1}{m!} \left( \frac{\lambda}{\mu} \right)^m \left( \frac{1}{1 - \rho} \right) \right]^{-1} \)
  - where \( \rho = \frac{\lambda}{m \mu} \)

**Q1:** what is the probability that all servers are busy?  
**Ans:** \( \sum_{n=m}^{\infty} P_n \)

**Q2:** throughput?  
**Ans:** \( \lambda \)

**Q3:** response time?
M/M/m/m

m servers with a single queue having a buffer space of m (when all servers are busy, a customer walks away), e.g., a telephone switching system.

\[ P_n = \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n P_0 \quad \therefore P_0 = \frac{1}{\sum_{n=0}^{m} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n} \quad \& \quad P_n = \frac{\frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n}{\sum_{i=0}^{m} \frac{1}{i!} \left( \frac{\lambda}{\mu} \right)^i} \]

Q1. Prob. that all m servers are busy (e.g., in a telephone switch company)? \( P_m \) The expression for \( P_m \) is called Erlang’s B formula.

Q2. Mean # of calls turned away per time unit? \( P_m \times \lambda \)
A Client-Server System

Request arrival rate per user: \( \lambda \)

Service rate of the server system with one server: \( \mu \)

Response time: the time spent by a user at the system between submitting the request & the return of the response

State Description: one state component representation

- \( n \): a number representing the number of users in the server system
- \( \therefore \) number of users still thinking (i.e., not issuing requests) = \( m - n \)

\[
\begin{array}{cccccc}
& 0 & m & (m-1) & (m-2) & \cdots & m-1 \\
\mu & m\lambda & \mu & (m-1)\lambda & \mu & \cdots & \mu \\
\end{array}
\]
Recall in M/M/1

\[ P_n = \left( \frac{\lambda}{\mu} \right)^n P_0 \]

\[ P_n = \left( \prod_{i=1}^{n} \frac{\lambda(m-i+1)}{\mu} \right) P_0 \]

or \[ P_n = \left( \frac{\lambda}{\mu} \right)^n \frac{m!}{(m-n)!} P_0 \]

where \[ P_0 = \frac{1}{\sum_{n=0}^{m} \left[ \left( \frac{\lambda}{\mu} \right)^n \frac{m!}{(m-n)!} \right]} \]

Q1: Avg. # of users in the server system? \( \bar{n} = \sum_{n=1}^{m} n \cdot P_n \)

Q2: Avg. # of users still thinking (not issuing requests)? \( m - \bar{n} \)

Q3: System throughput? \( x = \sum_{n=1}^{m} \mu \cdot P_n = \left( 1 - P_0 \right) \cdot \mu \)

Q4: Response time per user?
What happens if the server system has \( m \) servers, each with a service rate of \( \mu \)?

**Q1: Throughput?**

\[
x = \sum_{n=1}^{m} (P_n \cdot n\mu)
\]

**Q2: Response time?**
Fundamental Laws: algebraic relationships among performance measurement quantities.

\[ \lambda = \text{arrival rate} = \frac{A}{T} \quad \text{e.g.,} \quad \frac{100 \text{ arrivals}}{1 \text{ hr}} \]

\[ C = \# \text{ of completions} \]

\[ x = \frac{C}{T} \quad \text{throughput} \]

\[ B = \text{total system busy time} \]

\[ D = \frac{B}{C} \quad \text{average service time per request} \]

\[ \rho = \frac{B}{T} \quad \text{utilization of the system} \]

mathematically

\[
\frac{B}{T} = \frac{C}{T} \times \frac{B}{C}
\]

\[ \rho = x \times D \]

utilization law
Little’s law

Consider response time per customer

# of arrivals $A(t)$
or
# of departures $D(t)$

Consider # of customers per time unit

* A meaning of $W$ is the total time spent by all customers in the system. $\therefore R = W/C$

* Another meaning of $W$ is the total population accumulated (in queue & in service) over $T$ time units. $\therefore \bar{n} = \frac{W}{T}$

Algebraically $\bar{n} = \frac{W}{T} = \frac{W}{C} \cdot \frac{C}{T} = R \cdot x$ $\therefore \bar{n} = R \cdot x$