Basic queues

COMPSCI 655
Lecture 8
\( X(t) = \# \) customers in system at \( t \) (buffer + servers)

- \( \lim_{t \to \infty} P(X_{M/M/1}(t) = n) = (1-\rho)^n \) \( n=0,1,\ldots \) provided \( \rho < 1 \)
  \( \to \) Geometric distribution
  \( E[X_{M/M/1}] = \rho / (1-\rho) \)

- \( \lim_{t \to \infty} P(X_{M/M/K}(t) = n) = (1-\rho)^n / (1-\rho^{K+1}) \), \( n = 0,1,\ldots,K \)
  holds for any \( \rho > 0 \)

- \( \lim_{t \to \infty} P(X_{M/M/c}(t) = n) = x_0 \rho^n / n! \) \( n=0,1,\ldots,c \)
  = \( x_0 (\rho/c)^n c^c / n! \) \( n > c \)

provided \( \rho < c \) with
\[
x_0 = \left[ \sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \left( \frac{1}{1-\rho/c} \right) \right]^{-1}
\]
$M/M/c/c$ queue

- Poisson
- Lost when full
- No waiting room

1

2

$c$
$M/M/c/c$ queue (con't)

$X(t) \in S = \{0,1,\ldots,c\}$ # customers in system at time $t$, including the one in service, if any

$\{X(t)\}_t$ CMTC as construction rule applies with

- if $X(t) = 0$: $\eta_{0,1} = \lambda$, $\eta_{0,j} = 0$ for $j \neq 1$
- if $1 \leq X(t) = i \leq c-1$: $\eta_{i,i+1} = \lambda$, $\eta_{i,i-1} = i\mu$, $\eta_{i,j} = 0$ for $j \neq i-1, i+1$
- if $X(t) = c$: $\eta_{c,c-1} = c\mu$, $\eta_{c,j} = 0$ for $j \neq i-1$

Balance eqns:

\[
\begin{align*}
\lambda \pi_0 &= \mu \pi_1 \\
(\lambda + i\mu) \pi_i &= \lambda \pi_{i-1} + (i+1)\mu \pi_{i+1}, \quad i = 1, \ldots, c-1 \\
c\mu \pi_c &= \lambda \pi_{c-1}
\end{align*}
\]
$\mathcal{M}/\mathcal{M}/c/c$ queue (con't)

Balance eqns: \[
\begin{align*}
\lambda x_0 &= \mu x_1 \\
(\lambda + i \mu) x_i &= \lambda x_{i-1} + (i+1)\mu x_{i+1}, \quad i = 1, \ldots, c-1 \\
c\mu x_c &= \lambda x_{c-1}
\end{align*}
\]

Unique normalized sol. : \[x_i = \rho^i / \sum_{0 \leq j \leq c} \rho^j / j!, \quad i = 0, 1, \ldots, c\]

Irreducibility + CTMC Thm imply \[\pi_i = \lim_{t \to \infty} P(X(t)=i) = \rho^i / \sum_{0 \leq j \leq c} \rho^j / j!, \quad i = 0, 1, \ldots, c\]

System always stable

- Prob. customer lost = \[\pi_c = \rho^c / \sum_{0 \leq j \leq c} \rho^j / j!\]

Erlang loss formula (1920)
M/M/c/c queue (con’t)

- Prob. customer lost = \( \pi_c = \frac{\rho^c}{\sum_{0 \leq j \leq c} \rho^j / j!} \)

Why does Erlang’s formula so famous in telephony?

Q: Do call durations exponential in general?

No!

A: Insensitivity to service time distr. proved by Erlang = \# of busy servers in steady-state does not depend on service time distribution, depends only on its mean!

Consequence:

\[ \pi_i = \frac{\rho^i}{\sum_{0 \leq j \leq c} \rho^j / j!} \quad , \quad i = 0, 1, \ldots, c \]

holds for any service time distribution
$M/M/c/c$ queue (con’t)

However:

$\{X(t)\}$ is **not** a CTMC if service time distribution not exponential
About insensitivity of queueing results to service time distribution

Key feature: any customer admitted into the system starts receiving service upon arrival.

Case of LIFO, PS  ...
Regarding results on $M/M/1$, $M/M/1/K$, $M/M/c$ and $M/M/c/c$ queues

We have not used any feature of scheduling policies

Previous results hold for any work-conserving scheduling policy, i.e., a policy which

1 - does not keep a server idles if there are customers waiting

2 - does not create load
M/D/1 queue

Same as M/M/1 queue except that service times are all constant and equal to D

\[ X(t) = \# \text{ customers in system} \]

Q: is \{X(t), t\geq0\} CTMC?

Assume \( X(t) > 0 \). Can you predict what will happen next?

A: No!

Next departure time cannot be predicted. Lack of information

\[ Y(t) := \text{time before next departure when } X(t) > 0. \]

\( \Rightarrow \{(X(t), Y(t)), t\geq0\} \) (hybrid) Markov process
PASTA property

PASTA = Poisson Arrivals See Time Averages

Consider general system with Poisson arrival

- \( M(t) = \) system state at \( t \) given 1 arrival in \((t, t+h)\)
- \( N(t) = \) system state at \( t \)

\[
P(M(t) = n) = P(N(t) = n \mid \text{one arrival in } (t, t+h))
\]

\[
= \frac{P(N(t) = n, \text{ one arrival in } (t, t+h))}{P(\text{one arrival in } (t, t+h))}
\]

Bayes formula

\[
= \frac{P(N(t) = n)P(\text{one arrival in } (t, t+h))}{P(\text{one arrival in } (t, t+h))}
\]

Poisson assumption

\[
= P(N(t) = n)
\]
In words, if objects arrive to a system according to a Poisson process, upon arrival an object sees the system in its stationary regime.

Not true in general:

- Arrivals at $t_n = n$ sec. for $n=0,1,\ldots$ in single-server queue
- Customer $n$ requires service time $s_n = 0.999$ sec., $n= 0,1,\ldots$

If system empty at $t = 0$ any arriving customer will see system empty.

However, since server is always working in $(n,n+0.999)$ the probability to find it empty if one arrives at an arbitrary time in $(n,n+1)$ cannot be zero (it is actually 1).
M/G/1/FIFO queue

- Poisson arrivals \( \{t_i\}_{i \geq 1} \), rate \( \lambda > 0 \)
- Arbitrary distributed i.i.d. service times \( \{\sigma_i\}_{i \geq 1} \)
  \[ \frac{1}{\mu} = \mathbb{E}[\sigma_i], \quad \sigma^{(2)} = \mathbb{E}[\sigma_i^2] \quad [P(\sigma_i < x) \text{ arbitrary}] \]

2nd order moment of service time
M/G/1 FIFO queue

- Poisson arrivals \( \{t_i\}_{i \geq 1} \), rate \( \lambda > 0 \)
- Arbitrary distributed i.i.d. service times \( \{\sigma_i\}_{i \geq 1} \)
  
  \[
  \frac{1}{\mu} = E[\sigma_i], \quad \sigma^{(2)} = E[\sigma_i^2]
  \quad \text{[}P(\sigma_i < x) \text{ arbitrary]} \]

\( W_i \) = waiting of \( i \)th arriving customer \( (W_1 = 0) \)

\( \overline{W} \) = expected waiting time in steady-state

\[
= \lim_{i \to \infty} \left( \frac{1}{i} \sum_{1 \leq j \leq i} W_j \right) \quad \text{a.s. when limit exists}
\]

\( X(t) \) = nb. customers in waiting room at time \( t \)

\( R(t) \) = residual service time of customer in server at time \( t \), if any (i.e. if \( X(t) > 0 \))
**M/G/1 FIFO queue**

**Convention:**

\[ X(t_i) = \text{nb. customers in waiting room just before arrival of ith customer. In particular, } X(t_1) = 0. \]

\[ W_i = R(t_i) + \sum_{j=i-X(t_i)}^{i-1} \sigma_j \]

- \( W_i \) is the time before cust. in server served
- \( R(t_i) \) is the time needed to serve the \( X(t_i) \) customers queued in front of cust. \( i \)
M/G/1 FIFO queue

Convention:
$X(t_i) = \text{nb. customers in waiting room just before arrival of } i\text{th customer. In particular, } X(t_1) = 0$

$$W_i = R(t_i) + \sum_{j=i-X(t_i)}^{i-1} \sigma_j$$

Only FIFO assumption used in above equation

Taking expectation in both sides gives
\[
E[W_i] = E[R(t_i)] + \sum_{j=1-X(t_i)}^{i-1} \sigma_j
\]
\[
E[Z] = \sum_n E[Z|Y=n]P(Y=n)
\]

sum on \( k \) goes from 0 to \( i-2 \) since no more that \( i-1 \) customers can be in the buffer at time \( t_i \)

\[
= E[R(t_i)] + \sum_{k=0}^{i-2} \left[ \sum_{j=i-k}^{i-1} \sigma_j | X(t_i) = k \right] P(X(t_i) = k)
\]

\[
= E[R(t_i)] + \sum_{k=0}^{i-2} \sum_{j=i-k}^{i-1} E[\sigma_j | X(t_i) = k] P(X(t_i) = k)
\]
Claim: independent of \( \{\sigma_{i-k}, \ldots, \sigma_{i-1}\} \)

Depends only on \( \{\sigma_1, \ldots, \sigma_{i-k-1}\} \) but \( \{\sigma_1, \ldots, \sigma_{i-k-1}\} \) and \( \{\sigma_{i-k}, \ldots, \sigma_{i-1}\} \) independent rvs by assumption

\[
E[W_i] = E[R(t_i)] + \sum_{k=0}^{i-2} \sum_{j=i-k}^{i-1} E[\sigma_j | X(t_i) = k] P(X(t_i) = k)
\]

\[
= E[R(t_i)] + \frac{1}{\mu} \sum_{k=0}^{i-2} kP(X(t_i) = k) = E[R(t_i)] + \frac{1}{\mu} E[X(t_i)]
\]
We have shown that $E[W_i] = E[R(t_i)] + \frac{1}{\mu} E[X(t_i)]$

Result intuitively true ... but to prove it needs rvs $\{\sigma_n\}_n$ to be mutually independent with same mean $1/\mu$ [+ arrivals and service times independent]

Assumptions used so far: FIFO, $\{\sigma_n\}_n$ iid, arrivals and service times independent

Poisson assumption for arrivals not used so far.
We have shown that \( E[W_i] = E[R(t_i)] + \frac{1}{\mu} E[X(t_i)] \)

Define \( \bar{R}_a = \lim_{i \to \infty} E[R(t_i)] \) : mean residual time

\( \bar{X}_a = \lim_{i \to \infty} E[X(t_i)] \) : mean nb. customers at arrival epochs.

Give \( \bar{W} = \bar{R}_a + \bar{X}_a / \mu \)

Define \( \bar{R} = \lim_{t \to \infty} \frac{1}{t} \int_{t_0}^{t} R(s) \, ds \), \( \bar{X} = \lim_{t \to \infty} \frac{1}{t} \int_{t_0}^{t} X(s) \, ds \)

PASTA : \( \bar{R} = \bar{R}_a \), \( \bar{X} = \bar{X}_a \)

Therefore \( \bar{W} = \bar{R} + \bar{X} / \mu \)
Little’s formula

Result: The average number of customers in a stable system, denoted by $\bar{N}$, is equal to the average effective arrival rate, denoted by $\lambda$, multiplied by the average time that a customer spends in the system, denoted by $\bar{T}$, namely

$$\bar{N} = \lambda \bar{T}$$

Intuition: $\bar{N} / \bar{T}$ = departure rate = arrival rate $\lambda$ at equilibrium

A proof will be sketched later on in this class
\[ \bar{W} = \bar{R} + \bar{X} / \mu \]

Apply Little’s result to waiting room (assuming it is stable):

\[ \bar{X} = \lambda \bar{W} \]

so that

\[ \bar{W} = \bar{R} + \frac{\lambda}{\mu} \bar{W} = \bar{R} + \rho \bar{W} \]

and

\[ \bar{W} = \frac{\bar{R}}{1 - \rho} \]

providing \( \rho := \frac{\lambda}{\mu} < 1 \)

\( \rho < 1 \) stability condition of M/G/1 queue
\( \bar{W} = \frac{\bar{R}}{1 - \rho} \)

Remains to find \( \bar{R} \), expected residual service time
When queue is empty (we will assume that queue empties infinitely often - true if $\rho<1$). The number of served customers in $(0, \tau)$ is denoted by $Y(\tau)$. Here $Y(\tau)=4$. The equation $\overline{W} = \frac{\overline{R}}{1 - \rho}$ is stated, which is derived from the diagram. The diagram shows the function $R_{\tau}(t)$, with $\sigma_1$, $\sigma_2$, and $\sigma_3$ representing the served customers. The equation $\overline{R}_{\tau} = \frac{1}{\tau} \sum_{j=1}^{Y(\tau)} \frac{1}{2} \sigma_j^2$ is also given.
\[ R_\tau = \frac{1}{\tau} \sum_{j=1}^{1} \frac{1}{2} \sigma_j^2 \]

\[ R = \lim_{\tau \to \infty} R_\tau = \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{j=1}^{Y(\tau)} \frac{1}{2} \sigma_j^2 \]

\[ = \lim_{\tau \to \infty} \frac{Y(\tau)}{\tau} \sum_{j=1}^{Y(\tau)} \frac{1}{2} \sigma_j^2 = \lim_{\tau \to \infty} \frac{Y(\tau)}{\tau} \lim_{\tau \to \infty} \frac{1}{Y(\tau)} \sum_{j=1}^{Y(\tau)} \frac{1}{2} \sigma_j^2 \]

\[ = \lambda \lim_{k \to \infty} \frac{1}{k} \sum_{j=1}^{k} \frac{1}{2} \sigma_j^2 = \lambda \frac{E[\sigma^2]}{2} = \lambda \frac{\sigma^{(2)}}{2} \]

\[ R = \lambda \frac{\sigma^{(2)}}{2} \]
\[
\overline{W} = \frac{\overline{R}}{1 - \rho} \quad \text{with} \quad \overline{R} = \frac{{\lambda \sigma^{(2)}}}{2}
\]
M/G/1/FIFO queue (cont')

\[ \bar{W} = \frac{\lambda \sigma^{(2)}}{2(1 - \rho)} \]

P-K formula (Pollaczek-Khinchin)

Mean response time

\[ \bar{T} = \frac{1}{\mu} + \frac{\lambda \sigma^{(2)}}{2(1 - \rho)} \]

Expected nb. customers \( \bar{N} = \lambda \bar{T} \) by Little's result

Hence

\[ \bar{N} = \rho + \frac{\lambda^2 \sigma^{(2)}}{2(1 - \rho)} \]
M/G/1/FIFO queue (cont')

Recap.

\[
\begin{align*}
\overline{W}_{M/G/1} &= \frac{\lambda \sigma^{(2)}}{2(1 - \rho)} \\
\overline{T}_{M/G/1} &= \frac{1}{\mu} + \frac{\lambda \sigma^{(2)}}{2(1 - \rho)} \\
\overline{N}_{M/G/1} &= \rho + \frac{\lambda^2 \sigma^{(2)}}{2(1 - \rho)}
\end{align*}
\]

with \( \rho < 1 \)

If exp. service times then \( \sigma^{(2)} = \frac{2}{\mu^2} \) we retrieve M/M/1 results:

\[
\begin{align*}
\overline{W}_{M/M/1} &= \frac{\rho}{\mu(1 - \rho)} \\
\overline{T}_{M/M/1} &= \frac{1}{\mu - \lambda} \\
\overline{N}_{M/M/1} &= \frac{\rho}{1 - \rho}
\end{align*}
\]
About the stability condition $\rho < 1$

$$\rho = \lambda \cdot \frac{1}{\mu}$$

= Expected nb. of arrivals per time unit
  $\times$ Expected work brought by an arrival

= **Expected load** brought to system per unit of time

Server can handle at most one unit of work per unit of time, therefore $\rho < 1$ stability condition
(not a proof)

$\rho$ is called the **traffic intensity**
Result:

$\rho := \frac{\lambda}{\mu} < 1$ stability condition of any $G/G/1$ queue with stationary and ergodic arrivals $\{t_n\}_{n \geq 1}$ and service times $\{\sigma_n\}_{n \geq 1}$, i.e.

- $P(t_{n+k+1} - t_{n+k} < x) = P(t_{n+1} - t_n < x)$ and $P(\sigma_{n+k} < x) = P(\sigma_n < x)$ for any $n, k$ (stationarity)

- Limits $\lambda := \lim_N \frac{1}{N} \sum_{1 \leq n \leq N} (t_{n+1} - t_n)$ and $\mu^{-1} := \lim_N \frac{1}{N} \sum_{1 \leq n \leq N} \sigma_n$ exit almost surely (ergodicity)