

CDA6530: Performance Models of Computers and Networks

Chapter 6: Elementary Queuing Theory

Definition

- ❑ **Queuing system:**
 - ❑ a buffer (waiting room),
 - ❑ service facility (one or more servers)
 - ❑ a scheduling policy (first come first serve, etc.)
- ❑ **We are interested in what happens when a stream of customers (jobs) arrive to such a system**
 - ❑ throughput,
 - ❑ sojourn (response) time,
 - ❑ Service time + waiting time
 - ❑ number in system,
 - ❑ server utilization, etc.

Terminology

- **A/B/c/K queue**
 - A - arrival process, interarrival time distr.
 - B - service time distribution
 - c - no. of servers
 - K - capacity of buffer

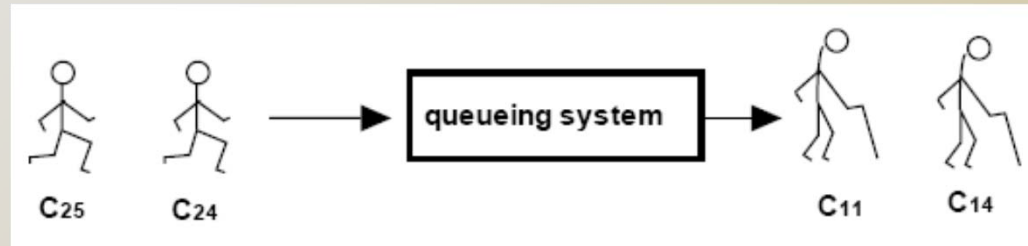
- Does not specify scheduling policy

Standard Values for A and B

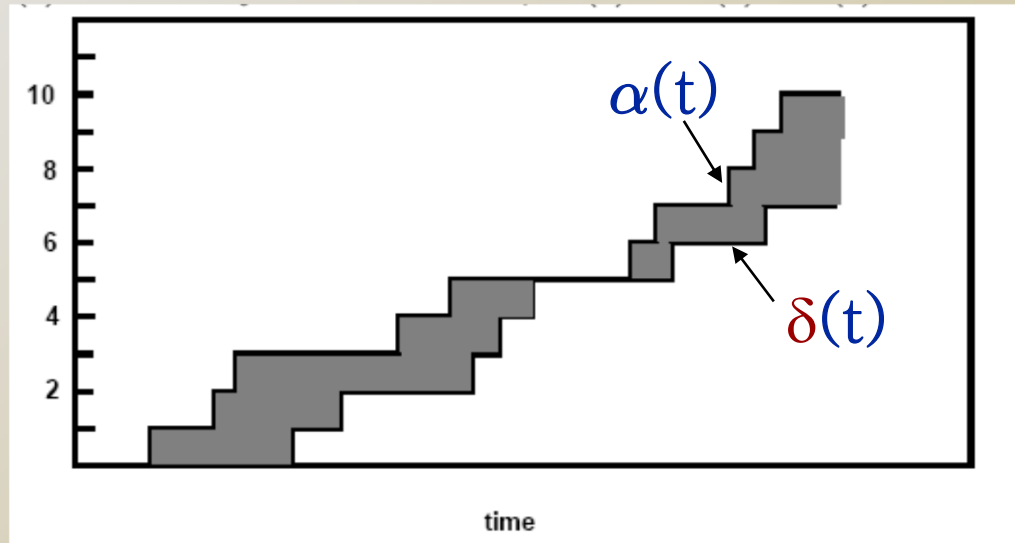
- ❑ M - exponential distribution (M is for Markovian)
- ❑ D - deterministic (constant)
- ❑ GI; G - general distribution

- ❑ M/M/1: most simple queue
- ❑ M/D/1: expo. arrival, constant service time
- ❑ M/G/1: expo. arrival, general distr. service time

Some Notations



- C_n : customer n , $n=1,2,\dots$
- a_n : arrival time of C_n
- d_n : departure time of C_n
- $\alpha(t)$: no. of arrivals by time t
- $\delta(t)$: no. of departure by time t
- $N(t)$: no. in system by time t
 - $N(t)=\alpha(t)-\delta(t)$



- Average arrival rate (from $t=0$ to now):
 - $\lambda_t = \alpha(t)/t$

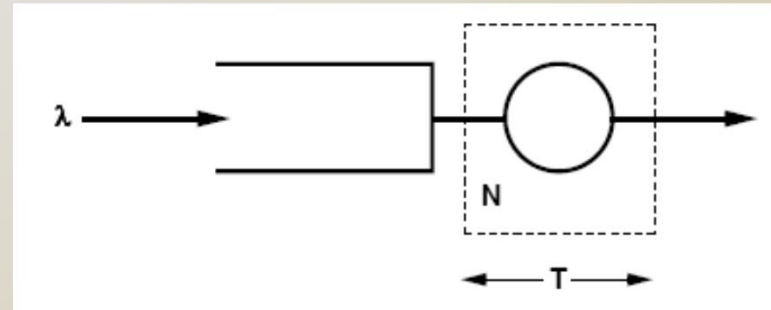
Little's Law

- $\gamma(t)$: total time spent by all customers in system during interval $(0, t)$

$$\gamma(t) = \sum_{n=1}^{\alpha(t)} \min\{d_n, t\} - a_n = \int_0^t N(s) ds$$

- T_t : average time spent in system during $(0, t)$ by customers arriving in $(0, t)$ $T_t = \gamma(t)/\alpha(t)$
- N_t : average no. of customers in system during $(0, t)$
 - $N_t = \gamma(t)/t$
- For a stable system, $N_t = \lambda_t T_t$
 - Remember $\lambda_t = \alpha(t)/t$
- For a long time and stable system
- $N = \lambda T$
- Regardless of distributions or scheduling policy

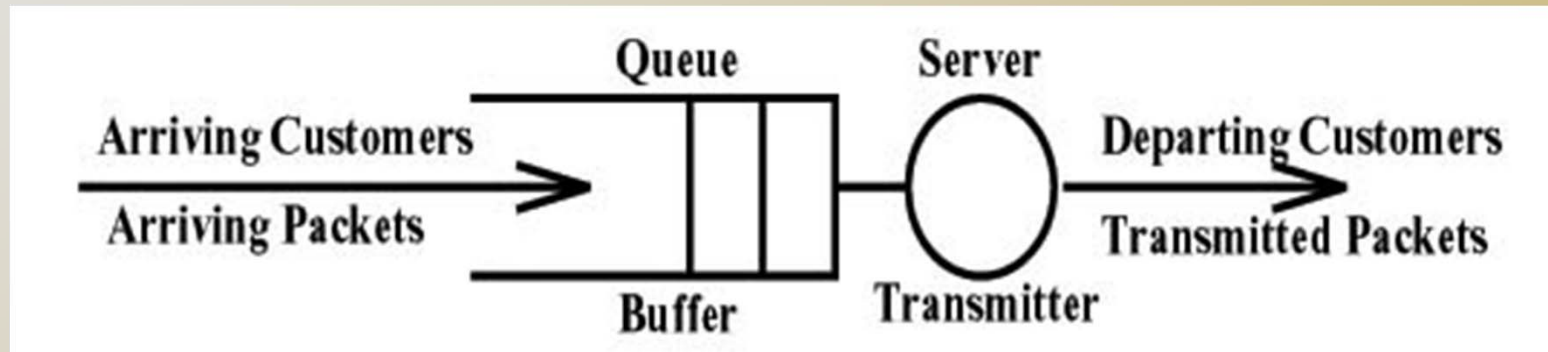
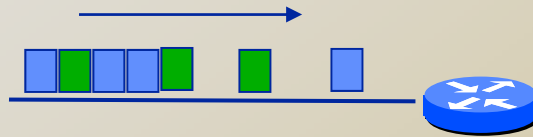
Utilization Law for Single Server Queue



- ❑ X : service time, mean $T = E[X]$
- ❑ Y : server state, $Y=1$ busy, $Y=0$ idle
- ❑ ρ : server utilization, $\rho = P(Y=1)$
- ❑ Little's Law: $N = \lambda E[X]$
- ❑ While: $N = P(Y=1) \cdot 1 + P(Y=0) \cdot 0 = \rho$
- ❑ Thus Utilization Law:
$$\rho = \lambda E[X]$$

Q: What if the system includes the queue?

Internet Queuing Delay Introduction



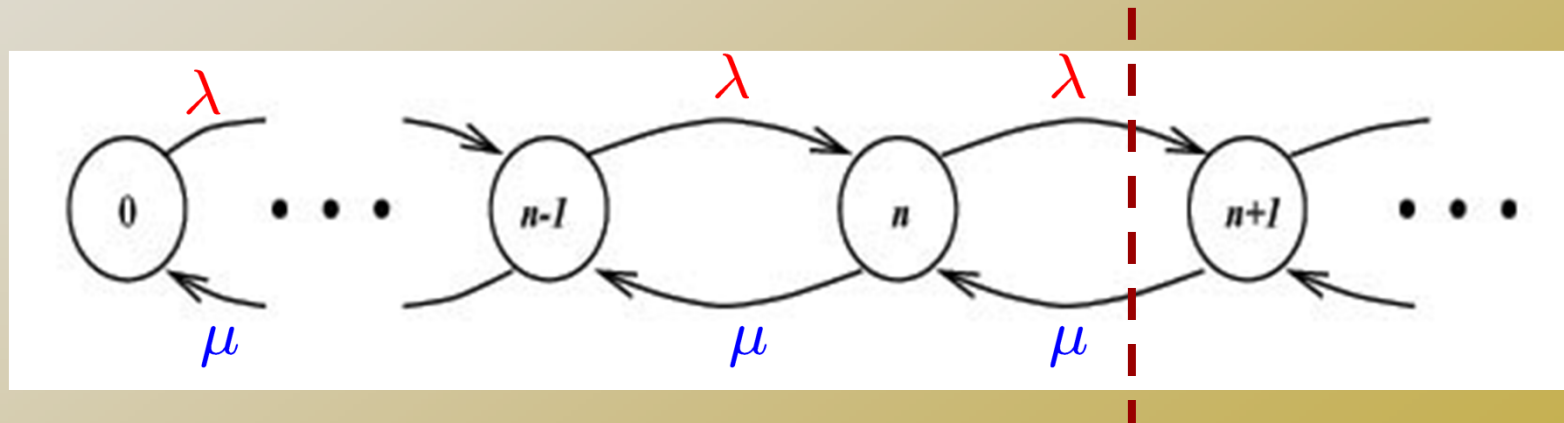
- ❑ How many packets in the queue?
- ❑ How long a packet takes to go through?

The M/M/1 Queue

- ❑ An M/M/1 queue has
 - ❑ Poisson arrivals (with rate λ)
 - ❑ Exponential time between arrivals
 - ❑ Exponential service times (with mean $1/\mu$, so μ is the “service rate”).
 - ❑ One (1) server
 - ❑ An infinite length buffer
- ❑ The M/M/1 queue is the most basic and important queuing model for network analysis

State Analysis of M/M/1 Queue

- N : number of customers in the system
 - (including queue + server)
 - Steady state
- π_n defined as $\pi_n = P(N=n)$
- $\rho = \lambda/\mu$: Traffic rate (traffic intensity)

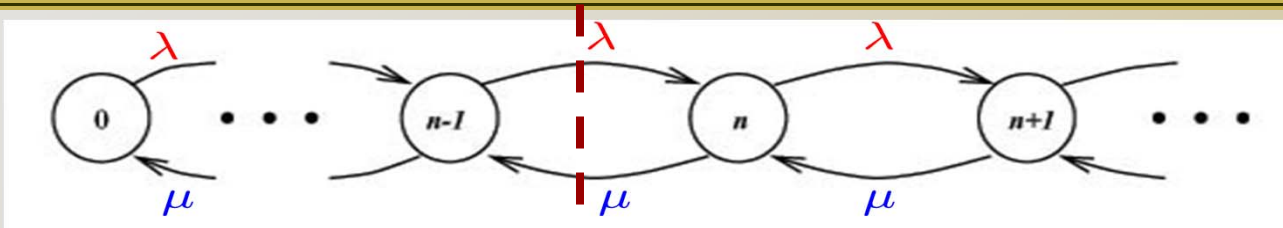


State transition diagram

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & \dots \\ \mu & -(\lambda + \mu) & \lambda & \dots \\ 0 & \mu & -(\lambda + \mu) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

- ❑ we can use $\pi Q = 0$ and $\sum \pi_i = 1$
- ❑ We can also use balance equation

State Analysis of M/M/1 Queue



□ # of transitions \rightarrow $\stackrel{!}{=}$ # of transitions \leftarrow

$$\pi_0 \lambda = \pi_1 \mu \quad \Rightarrow \quad \pi_1 = \rho \pi_0$$

$$\pi_1 \lambda = \pi_2 \mu \quad \Rightarrow \quad \pi_2 = \rho^2 \pi_0$$

\vdots

$$\pi_{n-1} \lambda = \pi_n \mu \quad \Rightarrow \quad \pi_n = \rho^n \pi_0$$

π_n are probabilities:

$$\sum_{i=0}^{\infty} \pi_i = 1$$

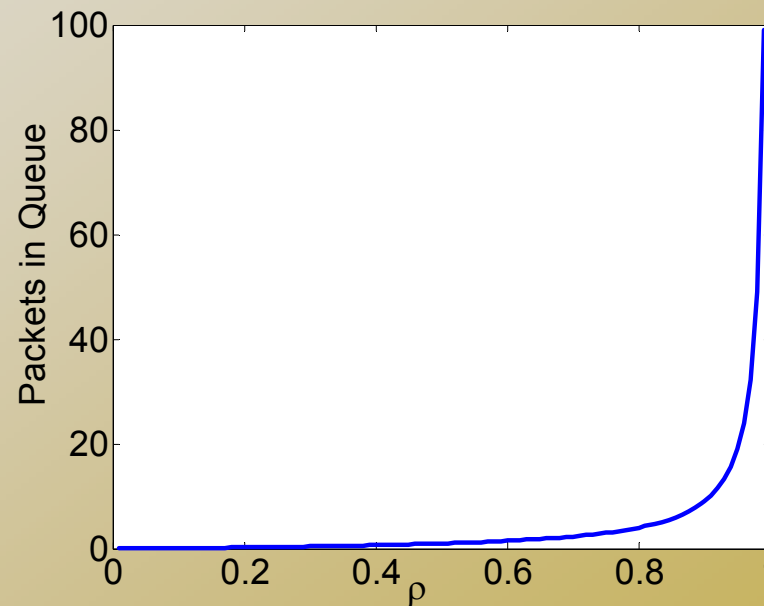
$$\Rightarrow \quad \pi_0 = 1 - \rho$$

$\rho = 1 - \pi_0$: prob. the server is working (why ρ is called “**server utilization**”)

State Analysis of M/M/1 Queue

- **N**: avg. # of customers in the system

$$E[N] = \sum_{k=1}^{\infty} k\pi_k = \pi_0 \sum_{k=1}^{\infty} k\rho^k = \frac{\rho}{1-\rho}$$



M/M/1 Waiting Time

- X_n : service time of n-th customer, $X_n \stackrel{st}{=} X$ where X is exponential rv
- W_n : waiting time of n-th customer
 - Not including the customer's service time
- T_n : sojourned time $T_n = W_n + X_n$
- When $\rho < 1$, steady state solution exists and $X_n, W_n, T_n \rightarrow X, W, T$
-
- Q: $E[W]$?

State Analysis of M/M/1 Queue

- **W: waiting time for a new arrival**

$$W = X_1 + X_2 + \cdots + X_{n-1} + R$$

X_i : service time of i-th customer

R : remaining service time of the customer in service
Exponential r.v. with mean $1/\mu$ due to **memoryless**
property of expo. Distr.

$$E[W] = E[(N - 1)X] + E[R] = E[N] \cdot E[X]$$

- **T: sojourn (response) time**

$$E[T] = \frac{1}{\mu} + E[W] = \frac{1}{\mu - \lambda}$$

Alternative Way for Sojourn Time Calculation

- ❑ We know that $E[N] = \rho/(1-\rho)$
- ❑ We know arrival rate λ
- ❑ Then based on Little's Law
- ❑ $N = \lambda T$

$$\rightarrow E[T] = E[N]/\lambda = 1/(\mu - \lambda)$$

M/M/1 Queue Example

- ❑ A router's outgoing bandwidth is 100 kbps
- ❑ Arrival packet's number of bits has expo. distr. with mean number of 1 kbits
- ❑ Poisson arrival process: 80 packets/sec
 - How many packets in router expected by a new arrival?
 - What is the expected waiting time for a new arrival?
 - What is the expected access delay (response time)?
 - What is the prob. that the server is idle?
 - What is $P(N > 5)$?
 - Suppose you can increase router bandwidth, what is the minimum bandwidth to support avg. access delay of 20ms?

Sojourn Time Distribution

- T's pdf is denoted as $f_T(t)$, $t \geq 0$
- $T = X_1 + X_2 + \dots + X_n + X$
 - Given there are $N=n$ customers in the system
 - Then, T is sum of $n+1$ exponential distr.
 - T is $(n+1)$ -order Erlang distr.
 - When conditioned on n , the pdf of X is denoted as $f_{T|N}(t|n)$

$$f_{T|N}(t|n) = \frac{\mu(\mu t)^n e^{-\mu t}}{n!}$$

Sojourn Time Distribution

- Remove condition $N=n$:

- Remember $P(N=n) = \pi_n = (1-\rho)\rho^n$

$$f_T(t) = f_{T|0}(t|0)\pi_0 + f_{T|1}(t|1)\pi_1 + \dots$$

$$f_T(t) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \frac{\mu(\mu t)^n e^{-\mu t}}{n!}$$

$$= (1-\rho)\mu e^{-\mu t} \sum_{n=0}^{\infty} (\rho\mu t)^n / n!$$

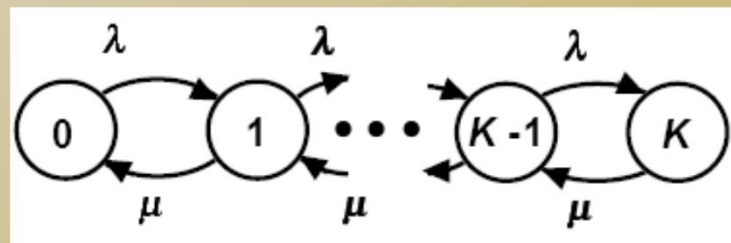
$$= (\mu - \lambda)e^{-\mu t} e^{\lambda t}$$

$$= (\mu - \lambda)e^{-(\mu-\lambda)t}$$

Thus, T is exponential distr. with rate $(\mu-\lambda)$

M/M/1/K Queue

- ❑ Arrival: Poisson process with rate λ
- ❑ Service: exponential distr. with rate μ
- ❑ Finite capacity of K customers
 - ❑ Customer arrives when queue is full is rejected
- ❑ Model as B-D process
 - ❑ $N(t)$: no. of customers at time t
 - ❑ State transition diagram



Calculation of π_0

- **Balance equation:**

- $\pi_i = \rho \pi_{i-1} = \rho^i \pi_0, i=0, \dots, K$

- **If $\lambda \neq \mu$:**

$$\sum_{i=0}^K \pi_i = \pi_0 \sum_{i=0}^K \rho^i = \pi_0 \frac{1 - \rho^{K+1}}{1 - \rho}$$

$$\sum_{i=0}^K \pi_i = 1 \Rightarrow \pi_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

- **If $\lambda = \mu$:** $\sum_{i=0}^K \pi_i = \pi_0 \sum_{i=0}^K \rho^i = (K + 1)\pi_0$

$$\pi_i = 1/(K + 1), i = 0, \dots, K$$

$E[N]$

□ If $\lambda \neq \mu$:

$$\begin{aligned} E[N] &= \sum_{i=0}^K i \pi_i \\ &= \frac{1 - \rho}{1 - \rho^{K+1}} \sum_{i=0}^K i \cdot \rho^i \end{aligned}$$

□ If $\lambda = \mu$:

$$\begin{aligned} E[N] &= \sum_{i=0}^K i \pi_i = \frac{1}{K+1} \sum_{i=0}^K i \\ &= \frac{1}{K+1} \frac{K(K+1)}{2} = \frac{K}{2} \end{aligned}$$

Throughput

□ Throughput?

□ When not idle = μ

□ When idle = 0

□ Throughput = $(1-\pi_0)\mu + \pi_0 \cdot 0$

□ When not full = λ (arrive pass)

□ When full = 0 (arrive drop)

□ Prob. Buffer overflow = π_K

□ Throughput = $(1-\pi_K)\lambda + \pi_K \cdot 0$

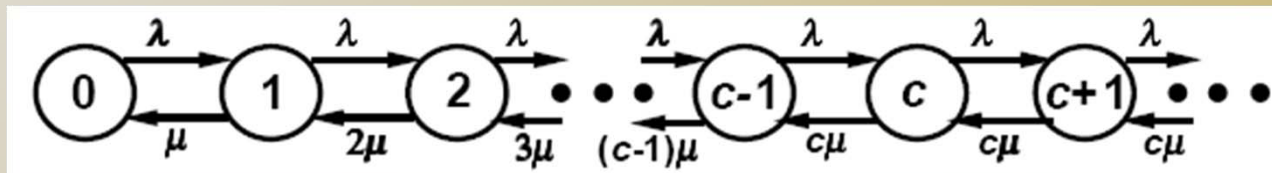
Sojourn Time

- One way: $T = X_1 + X_2 + \dots + X_n$ if there are n customers in ($n \leq K$)
 - Doable, but complicated
- Another way: Little's Law
 - $N = \lambda T$
 - The λ means *actual* throughput

$$E[T] = \frac{E[N]}{\text{throughput}} = \frac{E[N]}{(1 - \pi_0)\mu}$$

M/M/c Queue

- ❑ c identical servers to provide service
- ❑ Model as B-D process, $N(t)$: no. of customers
- ❑ State transition diagram:



- ❑ Balance equation:

$$\begin{cases} \lambda\pi_{i-1} = i\mu\pi_i, & i \leq c, \\ \lambda\pi_{i-1} = c\mu\pi_i, & i > c \end{cases}$$

-
-
- Solution to balance equation:

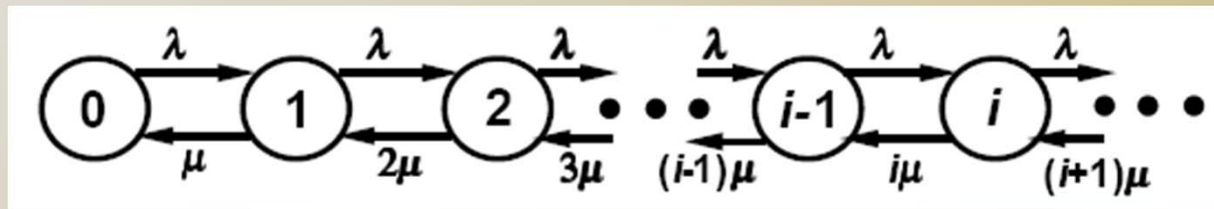
$$\pi_i = \begin{cases} \frac{\rho^i}{i!} \pi_0, & 0 \leq i \leq c, \\ \frac{\rho^i}{c! c^{i-c}} \pi_0, & c < i \end{cases}$$

- Prob. a customer has to wait (prob. of queuing)

$$P(\text{queuing}) = P(\text{wait}) = \sum_{n=c}^{\infty} \pi_n$$

M/M/∞ Queue

- Infinite server (delay server)
 - Each user gets its own server for service
 - No waiting time



- Balance equation:

$$\lambda\pi_{i-1} = i\mu\pi_i, \quad i = 0, 1, \dots$$

$$\pi_i = \frac{\rho^i}{i!}\pi_0 = \frac{\rho^i}{i!}e^{-\rho} \text{ why?}$$

$$\begin{aligned} E[N] &= \sum_{i=0}^{\infty} i\pi_i = \sum_{i=1}^{\infty} \frac{i\rho^i e^{-\rho}}{i!} \\ &= \rho e^{-\rho} \sum_{i=1}^{\infty} \frac{\rho^{i-1}}{(i-1)!} = \rho \end{aligned}$$

$$E[T] = \frac{E[N]}{\lambda} = \frac{1}{\mu} \quad \text{Why?}$$

PASTA property

- ❑ **PASTA: Poisson Arrivals See Time Average**
- ❑ Meaning: When a customer arrives, it finds the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time.
- ❑ $N(t)$: system state at time t
- ❑ Poisson arrival process with rate λ
- ❑ $M(t)$: system at time t given that an arrival occurs in the next moment in $(t, t+\Delta t)$

$$\begin{aligned}
P(M(t) = n) &= P(N(t) = n | \text{arrival in } (t, t + \Delta t)) \\
&= \frac{P(N(t) = n, \text{arrival in } (t, t + \Delta t))}{P(\text{arrival in } (t, t + \Delta t))} \\
&= \frac{P(N(t) = n)P(\text{arrival in } (t, t + \Delta t))}{P(\text{arrival in } (t, t + \Delta t))} \\
&= P(N(t) = n)
\end{aligned}$$

- If not Poisson arrival, then not correct