Markovské řetězce se spojitým parametrem II

Přednášející:
Mgr. Rudolf B. Blažek, Ph.D.

Katedra aplikované matematiky, Fakulta informačních technologií
České vysoké učení technické v Praze

Statistika pro informatiku

MI-SPI, LS 2015/16, Přednáška 15

Continuous-time Markov Chains II

Lecturer:
Mgr. Rudolf B. Blažek, Ph.D.

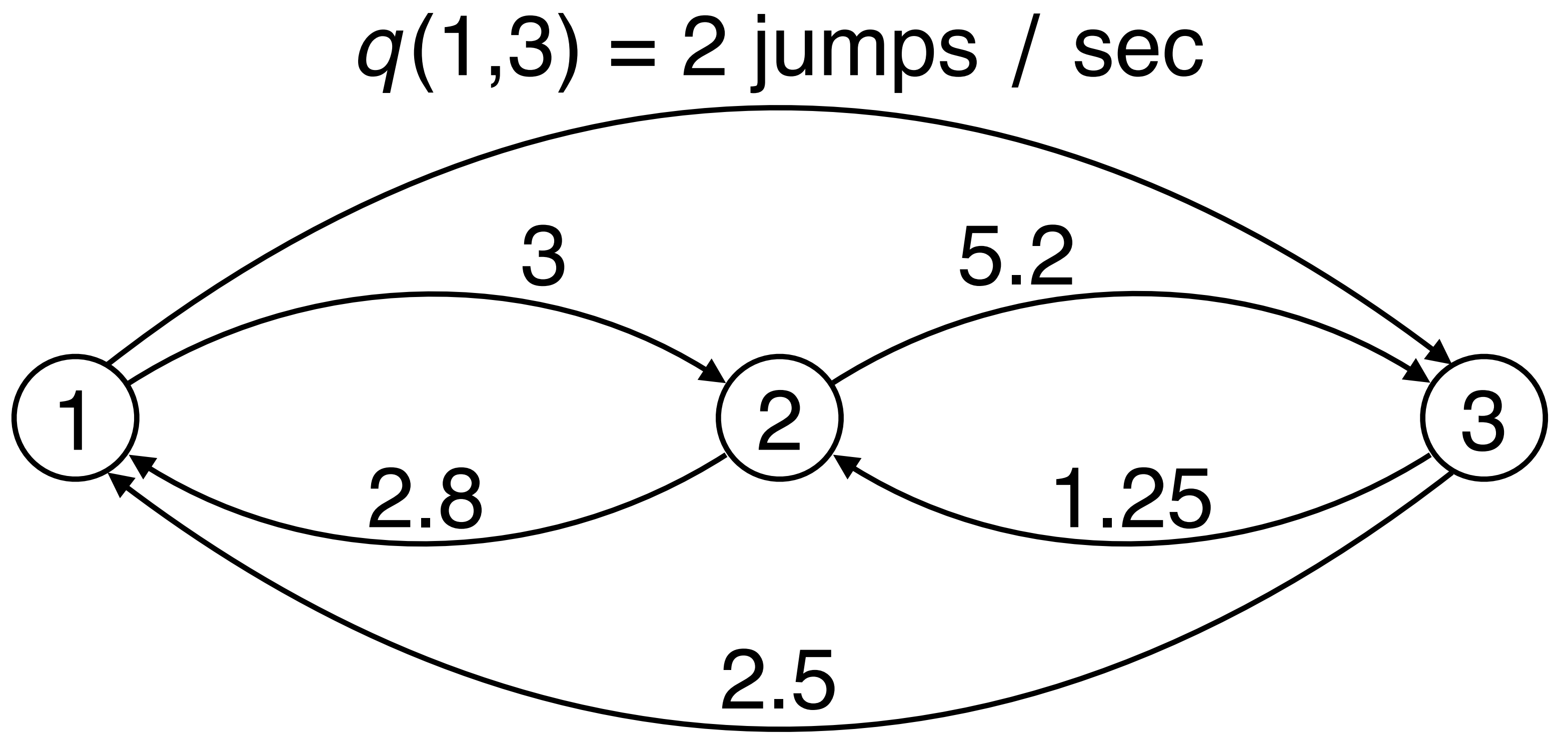
Department of Applied Mathematics, Faculty of Information Technology
Czech Technical University in Prague

Statistics for Informatics

MIE-SPI, LS 2015/16, Lecture 15

Review

From Jump Rates To a Markov Chain



$\lambda_1 = 2 + 3 = 5$ jumps away from ① / second

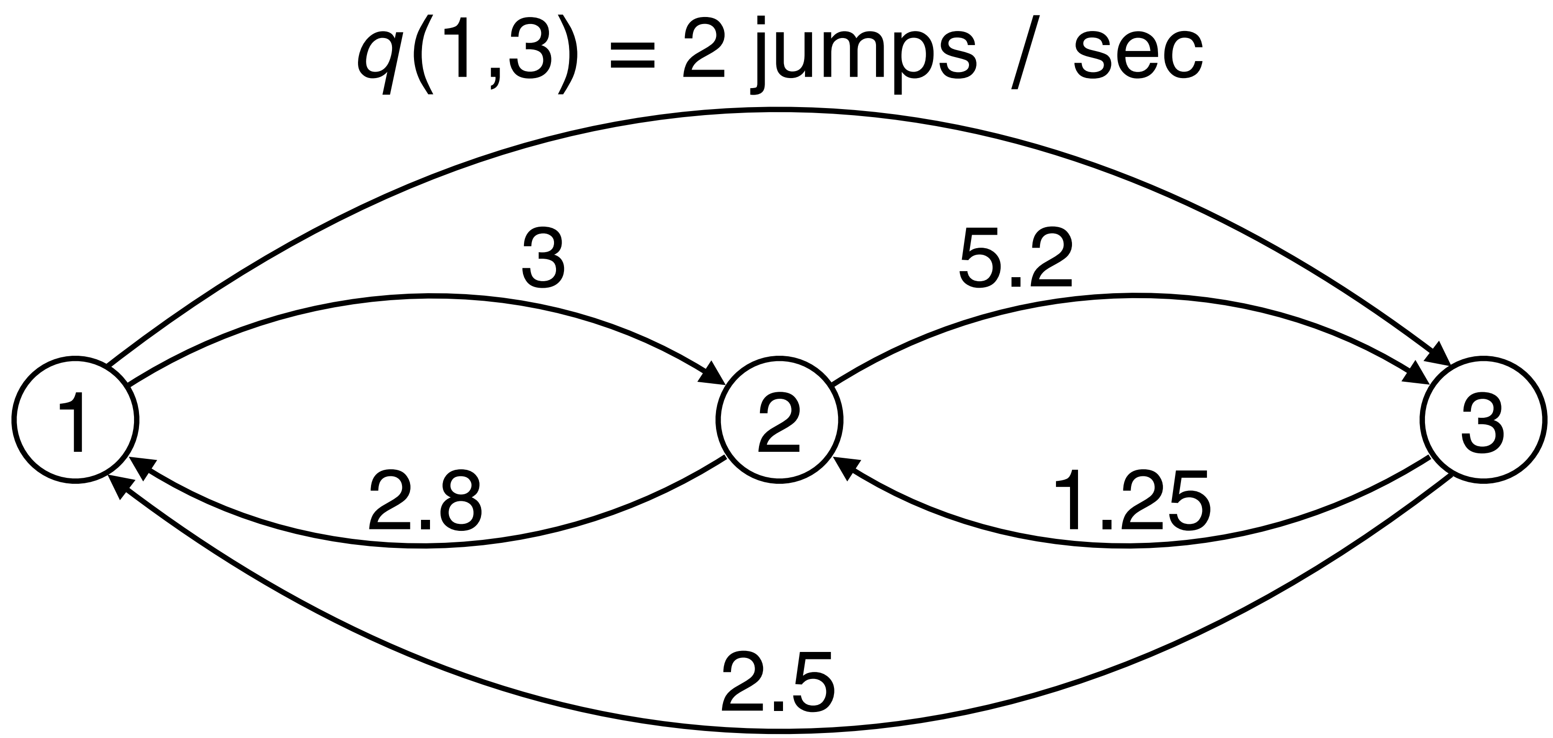
$r(1,2) = q(1,2)/\lambda_1 = 3/5 = P(Y_{n+1} = ② \mid Y_n = ①)$

$r(1,3) = q(1,3)/\lambda_1 = 2/5$

$\lambda_2 = 2.8 + 5.2 = 8$ $\quad r(2,1) = 2.8/8 \quad r(2,3) = 5.2/8$

$\lambda_3 = 1.25 + 2.5 = 3.75$ $\quad r(3,1) = 2.5/3.75 \quad r(3,2) = 1.25/3.75$

From Jump Rates To a Markov Chain



$\lambda_1 = 2 + 3 = 5$ $\lambda_2 = 2.8 + 5.2 = 8$ $\lambda_3 = 1.25 + 2.5 = 3.75$

$\lambda_{\max} = \max\{5, 8, 3.75\} = 8$ jumps / sec (on average, random times)

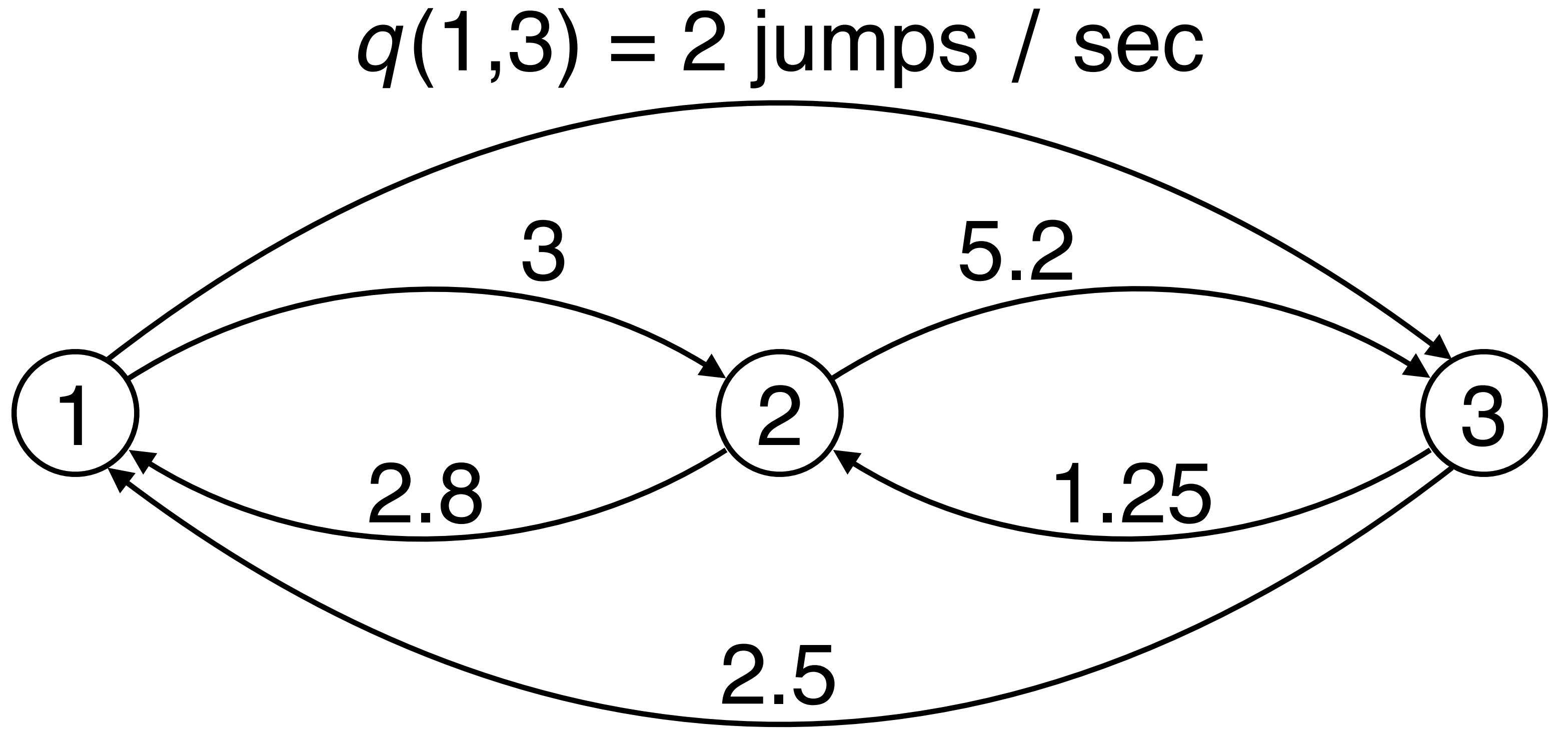
$u(1,2) = q(1,2)/\lambda_{\max} = 3/8 = P(Y_{n+1} = ② \mid Y_n = ①)$

$u(1,3) = q(1,3)/\lambda_{\max} = 2/8$ $u(1,1) = 1 - 5/8 = 3/8$

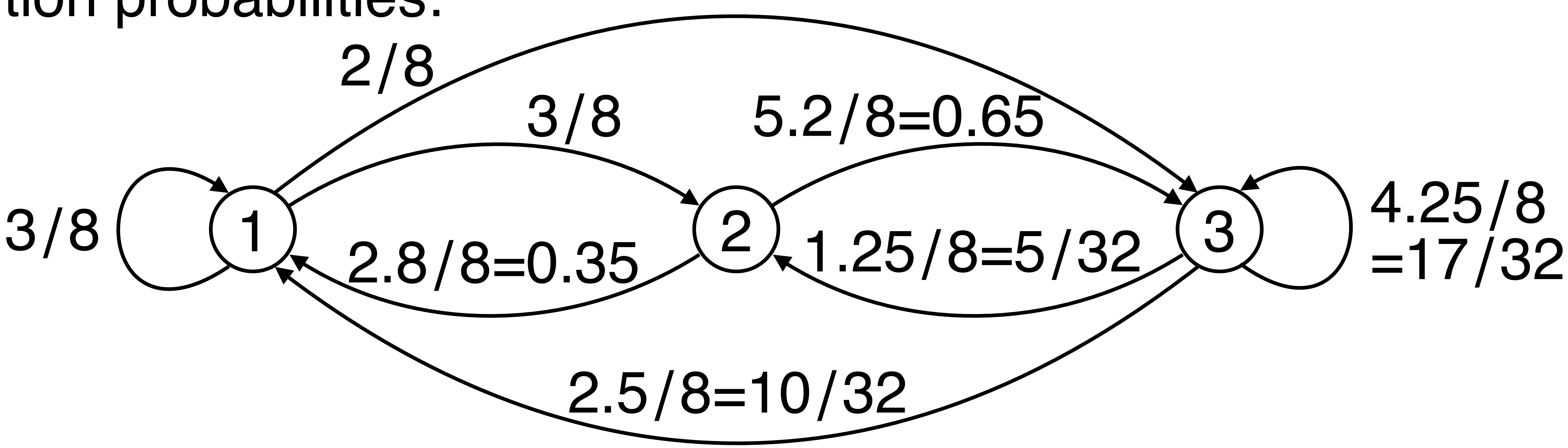
$u(2,1) = 2.8/8$ $u(2,3) = 5.2/8$ $u(2,2) = 0$

$u(3,1) = 2.5/8$ $u(3,2) = 1.25/8$ $u(3,3) = 1 - 3.75/8 = 4.25/8$

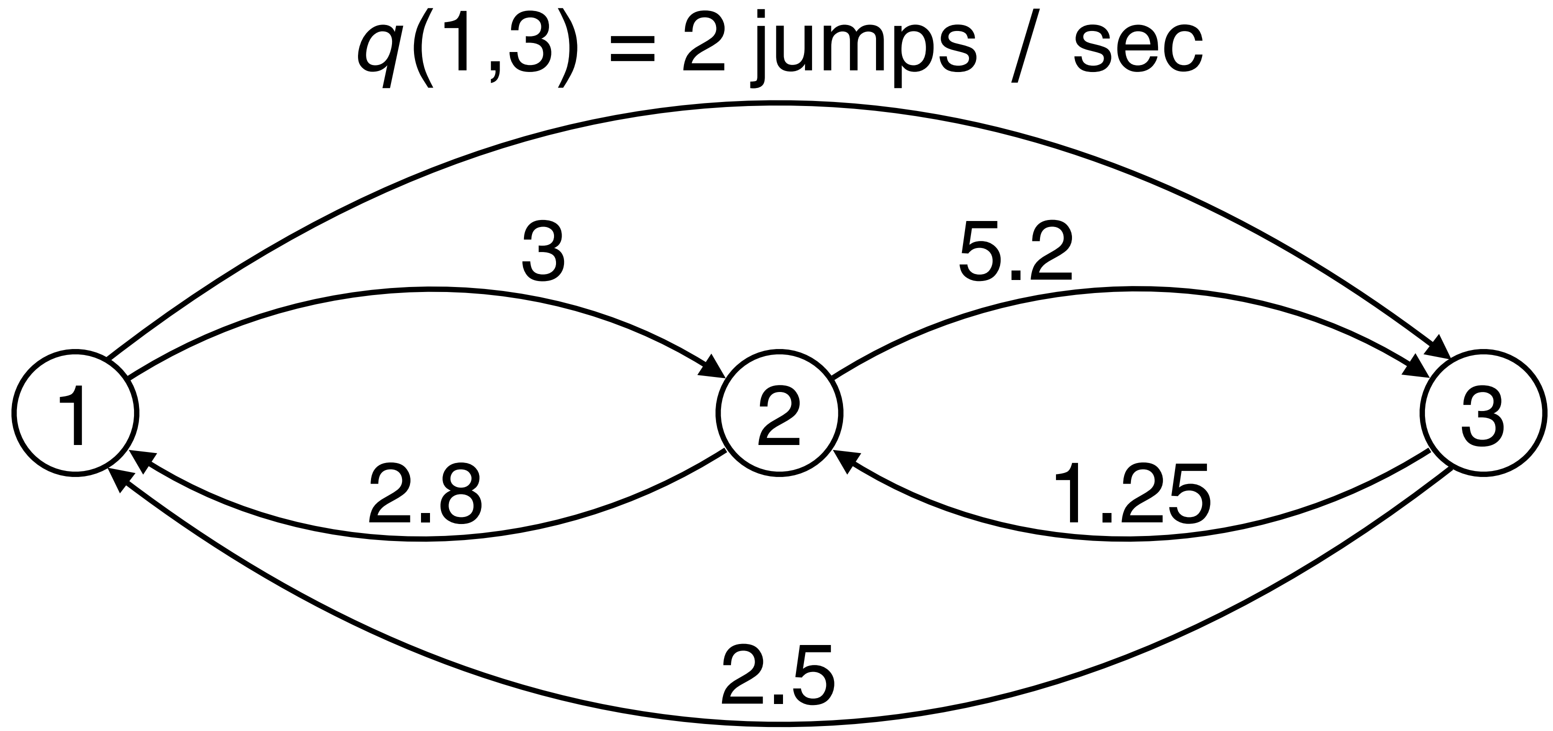
From Jump Rates To a Markov Chain



Poisson process: $\lambda_{\max} = \lambda_2 = 8$ jumps/sec (on average, random times)
Transition probabilities:

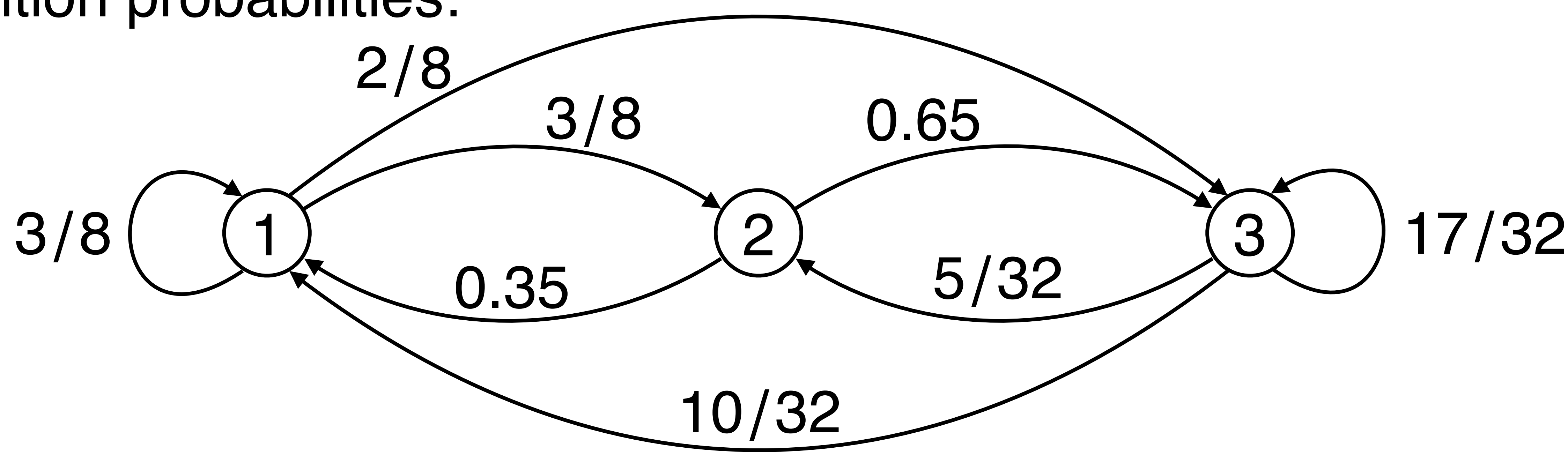


From Jump Rates To a Markov Chain



Poisson process: $\lambda_{\max} = \lambda_2 = 8$ jumps/sec (on average, random times)

Transition probabilities:



Stationarity and Limiting Behavior

Limiting Behavior

Theorem

If a continuous-time Markov Chain X_t is irreducible and has a stationary distribution π, then

$$\lim_{t \to \infty} p_t(i,j) = \pi(j)$$

For discrete-time MC π is a stationary distribution if

$$\pi \, \boldsymbol{P} = \pi \qquad (\boldsymbol{P} \text{ is the transition matrix})$$

For continuous-time there is no “first transition”. We need

$$\pi \, p_t = \pi \quad \text{for all } t > 0 \qquad \text{(stronger condition)}$$

Irreducible: $\exists$ finite path between all states with $q(x,y) > 0$

Stationary Distribution

Definition

For a continuous-time Markov chain, π is a **stationary distribution** if

$$\pi\, p_t = \pi \quad \text{for all } t > 0$$

Theorem

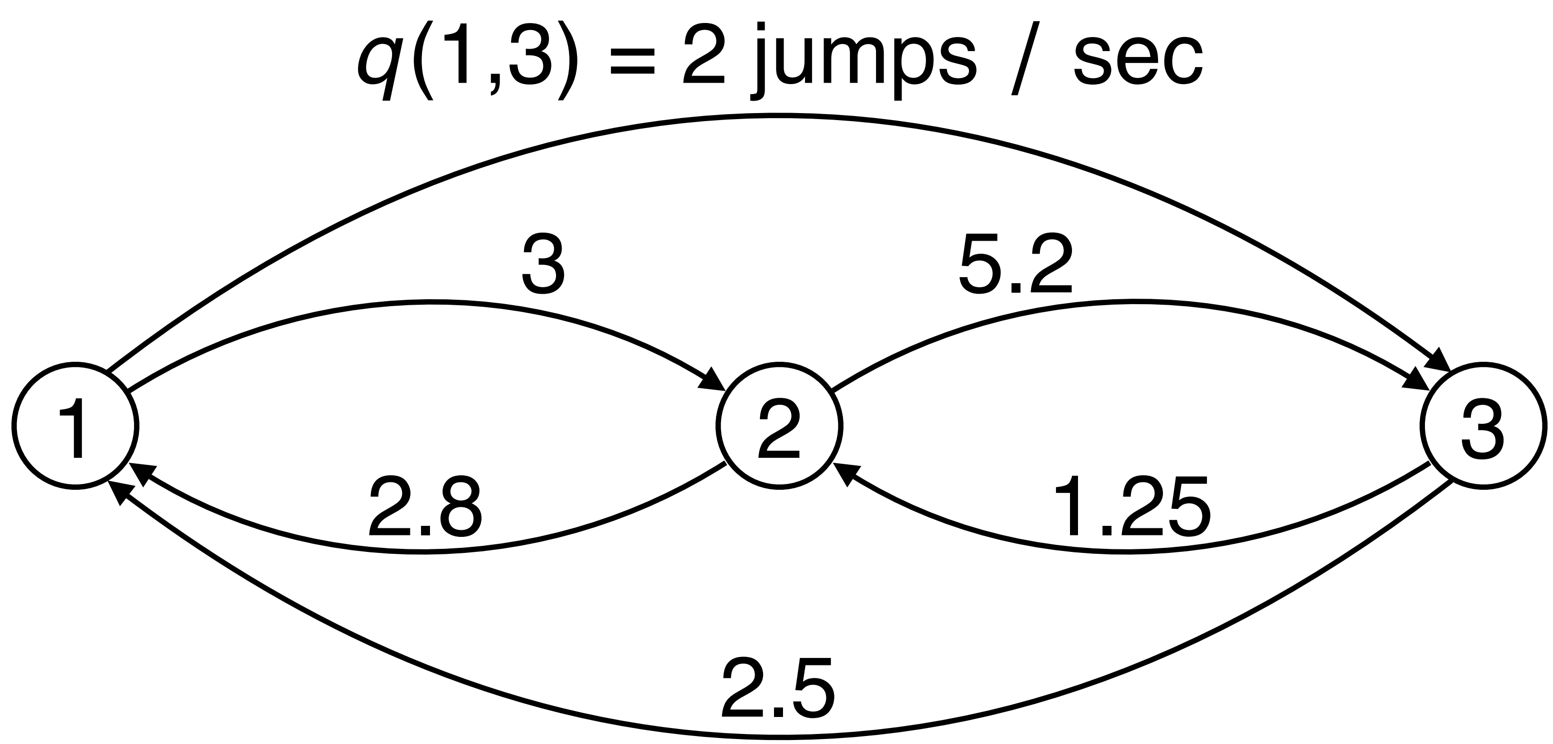
π is a stationary distribution if and only if

$$\pi\, Q = 0$$

Recall:

$$Q(i,j) = \begin{cases} q(i,j) & \text{if } j \neq i \\ -\lambda_i & \text{if } j = i. \end{cases}$$

Stationary Distribution: Continuous Time



π is a stationary distribution if and only if $\pi\, \boldsymbol{Q} = 0$

$\lambda_1 = 2 + 3 = 5$ $\qquad \lambda_2 = 2.8 + 5.2 = 8$ $\qquad \lambda_3 = 1.25 + 2.5 = 3.75$

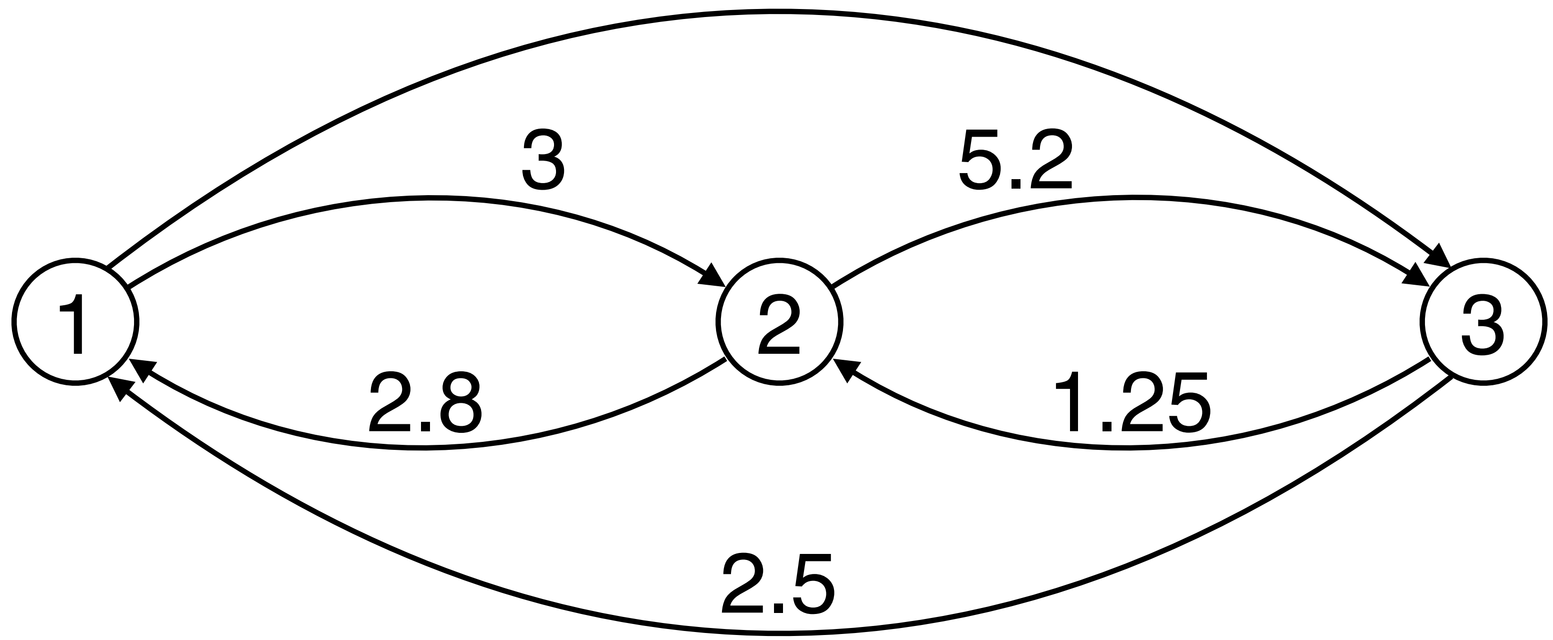
$$\boldsymbol{Q} = \begin{array}{c} \\ ① \\ ② \\ ③ \end{array} \begin{array}{c} \begin{array}{ccc} ① & ② & ③ \end{array} \\ \begin{pmatrix} -5 & 3 & 2 \\ 2.8 & -8 & 5.2 \\ 2.5 & 1.25 & -3.75 \end{pmatrix} \end{array}$$

$$\begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix} \begin{pmatrix} -5 & 3 & 2 \\ 2.8 & -8 & 5.2 \\ 2.5 & 1.25 & -3.75 \end{pmatrix} = 0$$

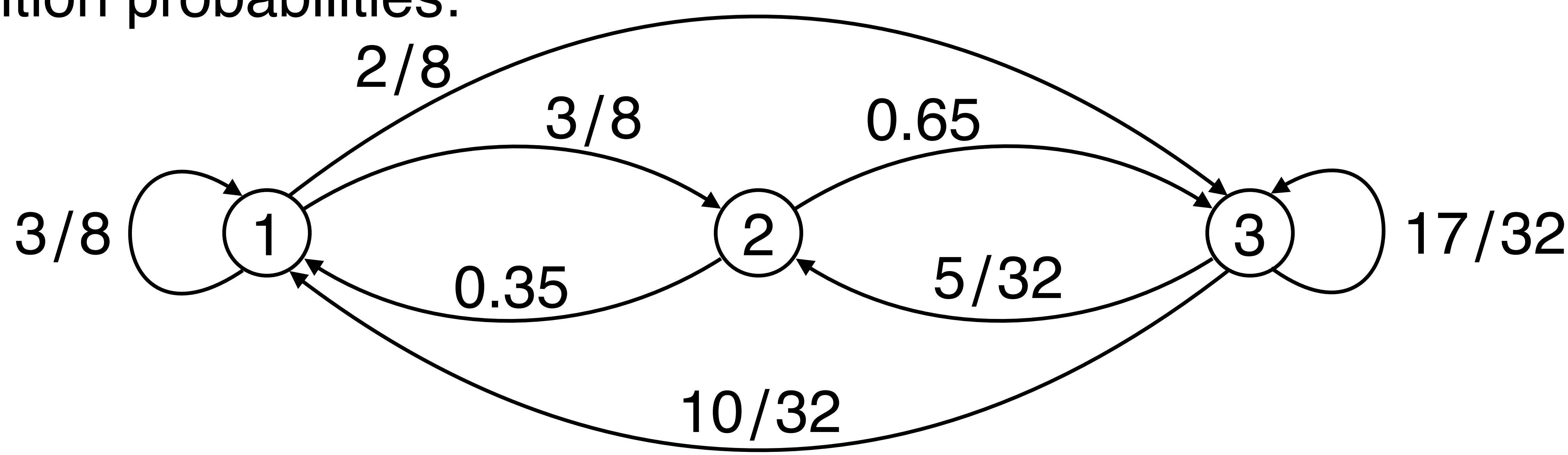
$$\pi = (0.341322,\ 0.19971,\ 0.458969)$$

From Jump Rates To a Markov Chain

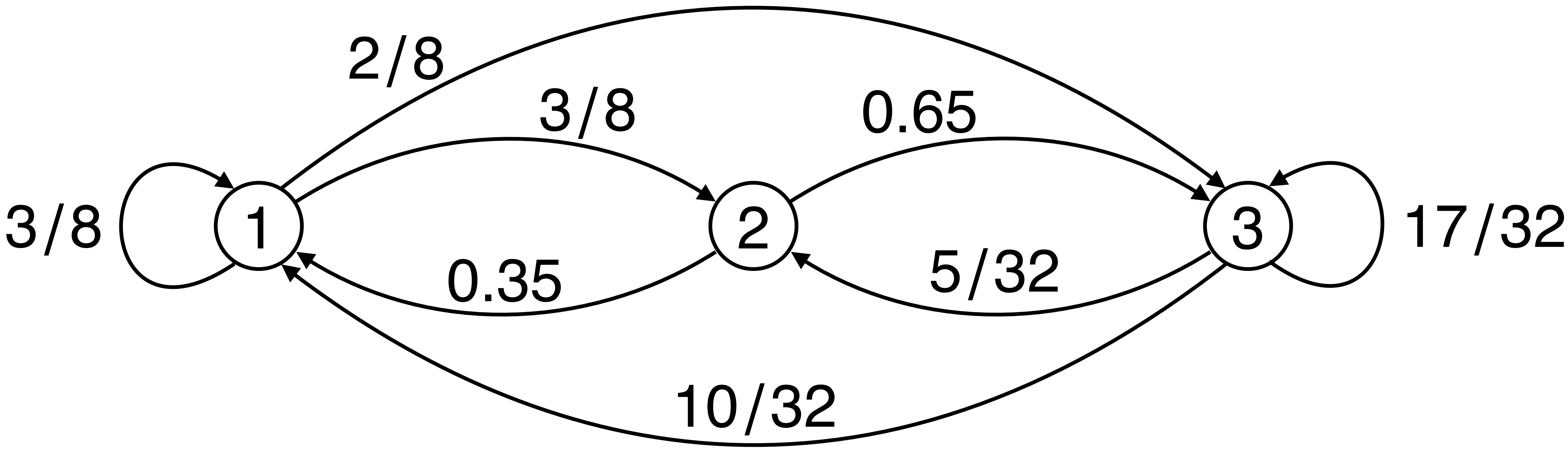
$q(1,3) = 2$ jumps / sec



Poisson process: $\lambda_{\max} = \lambda_2 = 8$ jumps/sec (on average, random times)
Transition probabilities:



Stationary Distribution: Discrete Time



π is a stationary distribution if and only if $\pi\,\boldsymbol{P} = \pi$

$$\begin{pmatrix}\pi_1 & \pi_2 & \pi_3\end{pmatrix}\begin{pmatrix}3/8 & 3/8 & 2/8\\ 0.35 & 0 & 0.65\\ 10/32 & 5/32 & 17/32\end{pmatrix} = \begin{pmatrix}\pi_1 & \pi_2 & \pi_3\end{pmatrix}$$

$$\boldsymbol{P} = \begin{array}{c c} & \begin{array}{ccc} ① & ② & ③ \end{array} \\ \begin{array}{c} ① \\ ② \\ ③ \end{array} & \begin{pmatrix}3/8 & 3/8 & 2/8\\ 0.35 & 0 & 0.65\\ 10/32 & 5/32 & 17/32\end{pmatrix}\end{array}$$

$\pi = (0.341322, 0.19971, 0.458969)$

... the same as before

Detailed Balance Condition

Definition

For a continuous-time Markov chain X_t, a distribution π is said to satisfy the **detailed balance condition** (DBC) if

$$\pi(k)q(k,j) = \pi(j)q(j,k)$$

(The MC is “reversible”)

Theorem

If a distribution π satisfies the detailed balance condition, then π is a stationary distribution of the Markov chain.

Recall: Detailed Balance Condition for Discrete-time Markov Chains

Definition

The probability distribution π satisfies condition of detailed balance if

podmínka detailní rovnováhy

$$\pi_i \boldsymbol{P}_{ij} = \pi_j \boldsymbol{P}_{ji} \ \forall i, j.$$

We say that the chain $(X_n)_{n\geq 0}$ with the starting distrinution π is reversible .

Recall Our Example

$N(t)$ is a Poisson Process (λ); Y_n = discrete-time MC with transition prob. $u(i,j)$; $N(t)$ is indep. of Y_n. Then $X_t = Y_{N(t)}$ has jump rates $q(i,j) = \lambda\, u(i,j)$.

Detailed Balance for Y_n:

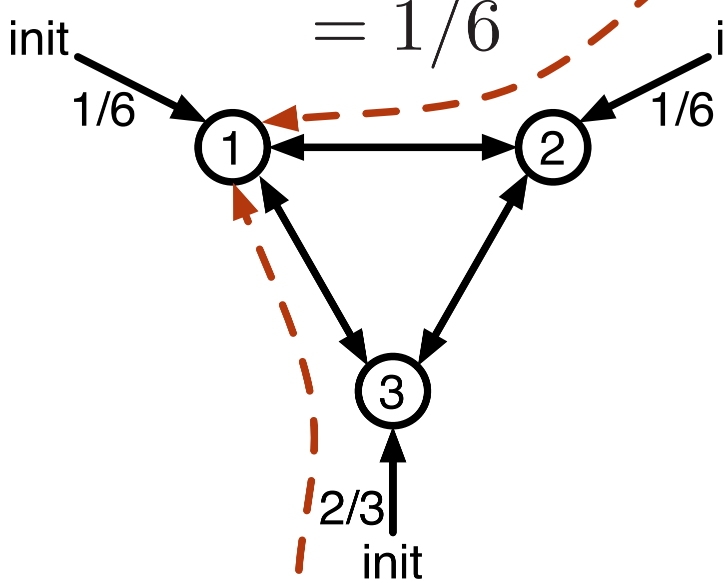
Detailed Balance for X_t:

$$\pi(i)\, u(i,j) = \pi(j)\, u(j,i) \xleftrightarrow[\times\lambda]{} \pi(i)\, \lambda\, u(i,j) = \pi(j)\, \lambda\, u(j,i)$$

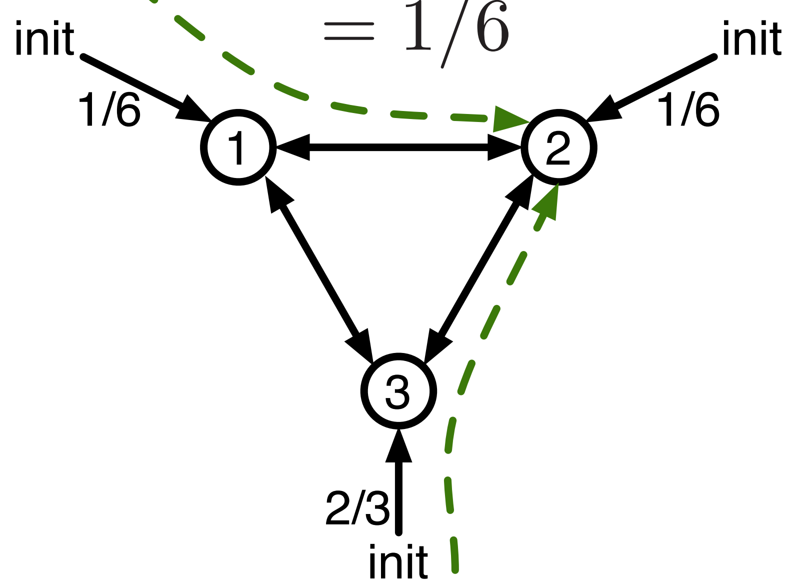
$$\pi(i)\, q(i,j) = \pi(j)\, q(j,i)$$

Stationarity: $\pi \boldsymbol{P} = \pi$

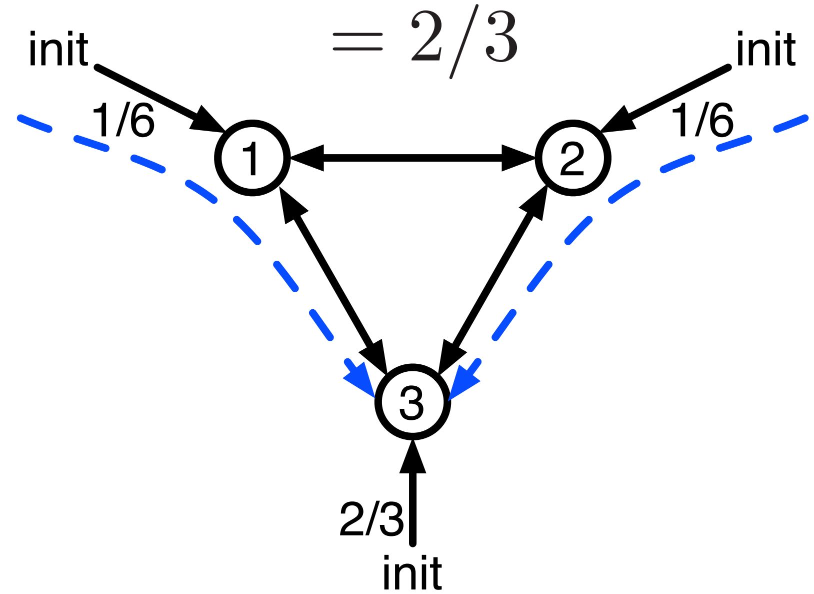
$$P(\bullet \to 2 \to 1) + P(\bullet \to 3 \to 1) = 1/6$$



$$P(\bullet \to 1 \to 2) + P(\bullet \to 3 \to 2) = 1/6$$

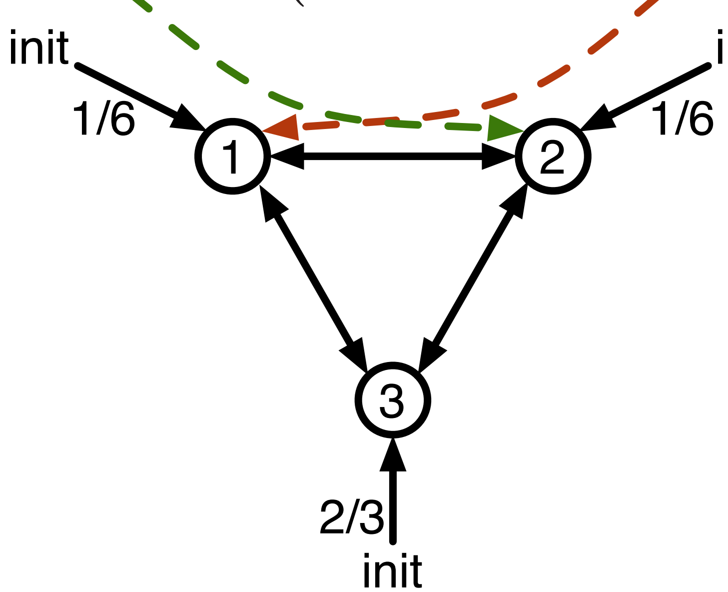


$$P(\bullet \to 1 \to 3) + P(\bullet \to 2 \to 3) = 2/3$$

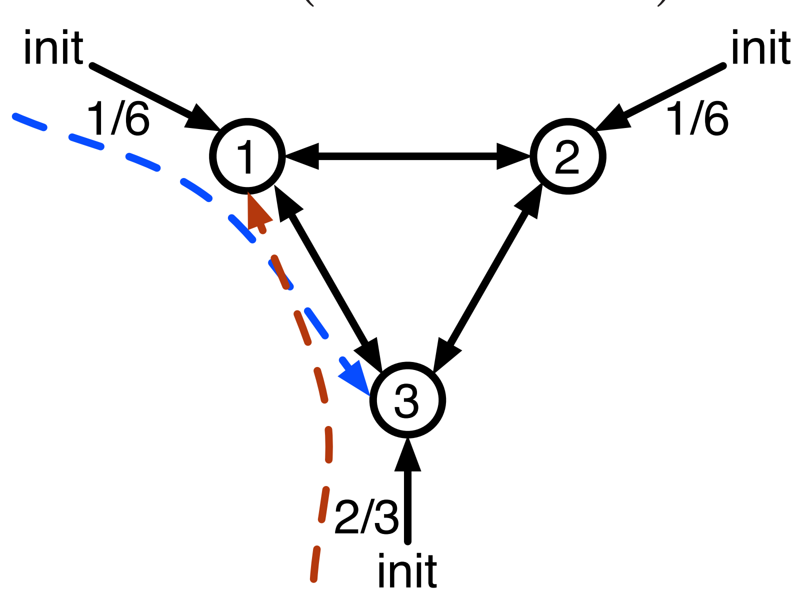


Detailed Balance: $\pi_i \boldsymbol{P}_{ij} = \pi_j \boldsymbol{P}_{ji} \ \forall i, j$

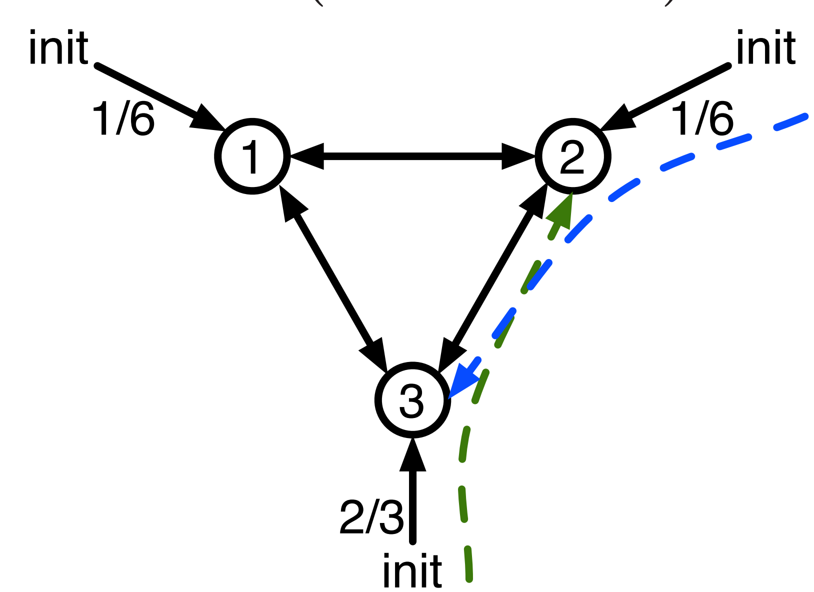
$$P(\bullet \to 1 \to 2) = P(\bullet \to 2 \to 1)$$



$$P(\bullet \to 1 \to 3) = P(\bullet \to 3 \to 1)$$



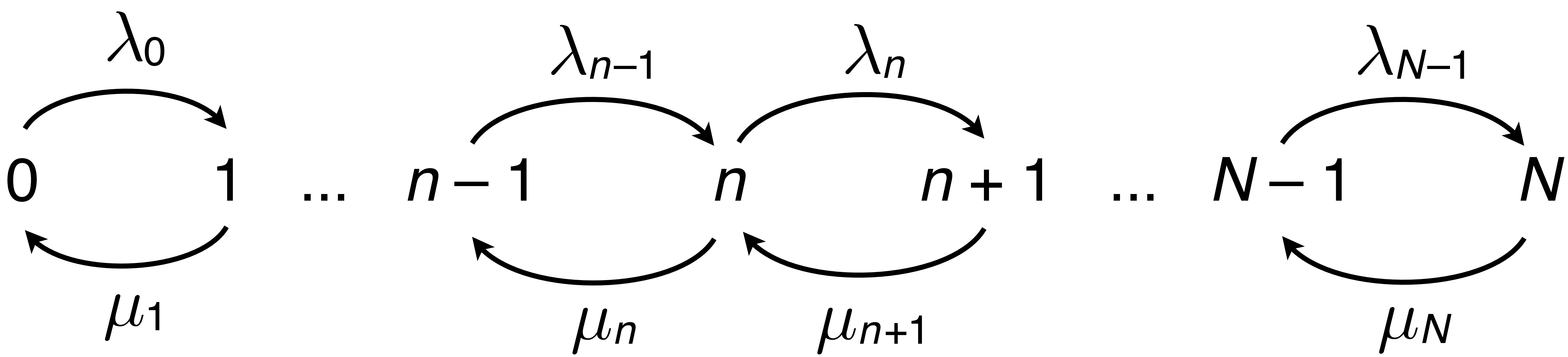
$$P(\bullet \to 2 \to 3) = P(\bullet \to 3 \to 2)$$



Birth & Death Chains (B&D Chains)

X(t) = Number of customers in a queueing system

Jump rates



$$q(n, n+1) = \lambda_n, \quad \text{for } 0 \leq n < N$$

$$q(n, n-1) = \mu_n, \quad \text{for } 0 < n \leq N$$

Detailed Balance Condition for Birth & Death Chains

Recall a Theorem

If a distribution π satisfies the detailed balance condition, then π is a ***stationary distribution*** of the Markov chain.

The DBC $$\pi(k)q(k,j) = \pi(j)q(j,k)$$

The DBC for Birth & Death Chains

$$\begin{aligned}
\pi(n-1)q(n-1,n) &= \pi(n)q(n,n-1)\\
\pi(n-1)\lambda_{n-1} &= \pi(n)\mu_n\\
\pi(n) &= \frac{\lambda_{n-1}}{\mu_n}\,\pi(n-1)
\end{aligned}$$

Detailed Balance Condition for Birth & Death Chains

The DBC for Birth & Death Chains

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n}\,\pi(n-1)$$

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n} \cdot \frac{\lambda_{n-2}}{\mu_{n-1}} \cdot \pi(n-2)$$

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n} \cdot \frac{\lambda_{n-2}}{\mu_{n-1}} \cdots \frac{\lambda_0}{\mu_1} \cdot \pi(0)$$

π is a ***stationary distribution*** of the Markov chain

Stationary Distribution for Birth & Death Chains

Theorem

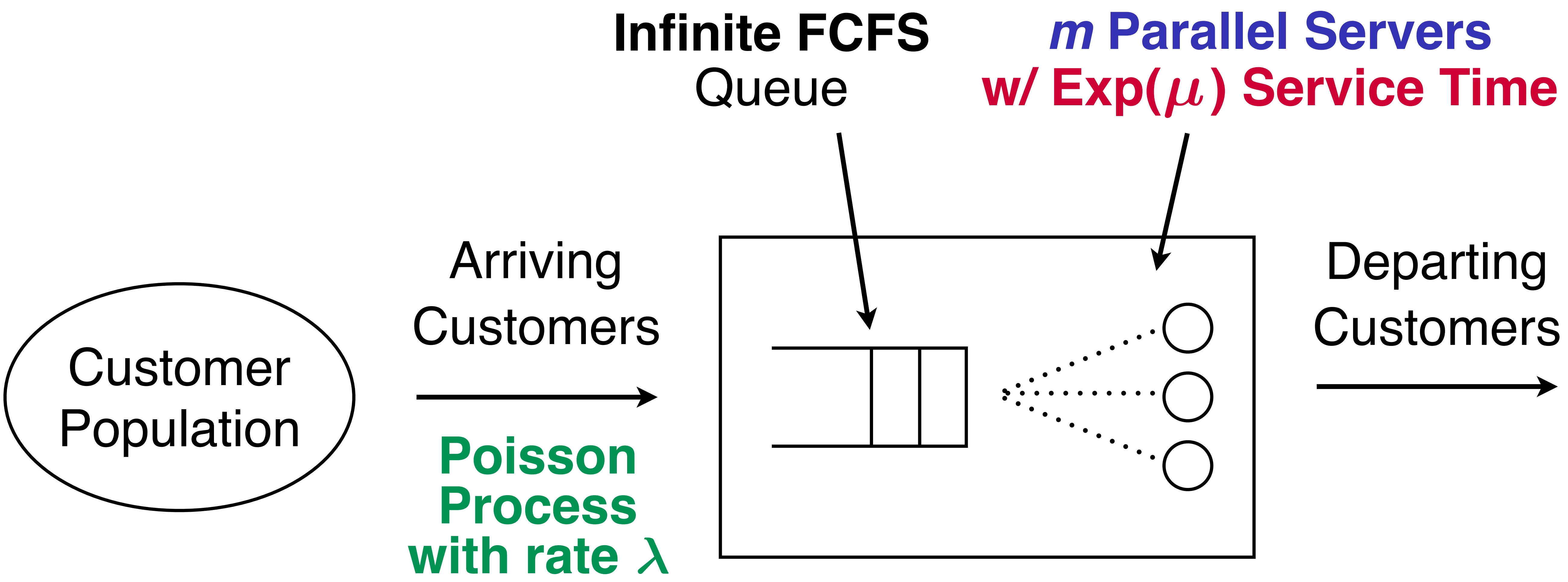
The stationary distribution of the B&D Markov chain is

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n} \cdot \frac{\lambda_{n-2}}{\mu_{n-1}} \cdots \frac{\lambda_0}{\mu_1} \cdot \pi(0)$$

provided that

$$\sum_{n=0}^{\infty} \pi(n) = 1 \quad \text{and} \quad \pi(n) \geq 0 \text{ for all } n \geq 1$$

Queueing System M / M / *m*



Resource Utilization

Definition

Resource utilization of a queueing system is the fraction of time the system is used

$$\rho = \frac{\text{Time a server is occupied}}{\text{Time available}}$$

Later we will show that for M/M/m systems $\rho = \dfrac{\lambda}{m\mu}$

For M/M/1 systems $\rho = \dfrac{\lambda}{\mu}$

Queueing System M / M / *m*

Example — M / M / *m* Queue

Consider load-balancing *m* replicated database servers.
A request is routed to the next available server.
Requests line-up in a single queue if all servers are busy.
Requests arrive at times of a Poisson Process w/ rate λ:

$$q(n, n+1) = \lambda \qquad \text{for all } n \geq 0$$

Service times are random independent ~ Exp(μ):

$$q(n, n-1) = n\mu \qquad \text{if } 0 \leq n \leq m$$

$$q(n, n-1) = m\mu \qquad \text{if } n \geq m$$

Stationary Distribution for the Number of Customers in M/M/1 Queueing Systems

Example — M / M / 1 Queue

Requests arrive at times of a Poisson Process w/ rate λ:

$$q(n, n{+}1) = \lambda \qquad \text{for all } n \geq 0$$

Service times are random independent ~ Exp(μ):

$$q(n, n{-}1) = \mu \qquad \text{for all } n \geq 1$$

The stationary distribution π:

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n} \cdot \frac{\lambda_{n-2}}{\mu_{n-1}} \cdots \frac{\lambda_0}{\mu_1} \cdot \pi(0)$$

$$\pi(n) = \left(\frac{\lambda}{\mu}\right)^n \cdot \pi(0) = \rho^n \, \pi(0)$$

Stationary Distribution for the Number of Customers in M/M/1 Queueing Systems

Example — M / M / 1 Queue

The stationary distribution: $\pi(n) = (\lambda/\mu)^n \, \pi(0) = \rho^n \, \pi(0)$

For $|\rho| < 1$ we have

$$1 = \sum_{n=0}^{\infty} \pi(n) = \sum_{n=0}^{\infty} \rho^n \, \pi(0) = \frac{\pi(0)}{1-\rho}$$

The stationary distribution becomes:

$$\pi(n) = (1-\rho)\,\rho^n$$

π+1 has Geometric distribution (1st head toss #) with parameter $1 - \rho = 1 - \lambda/\mu$.

Introduction to Queueing Theory

Základy teorie front

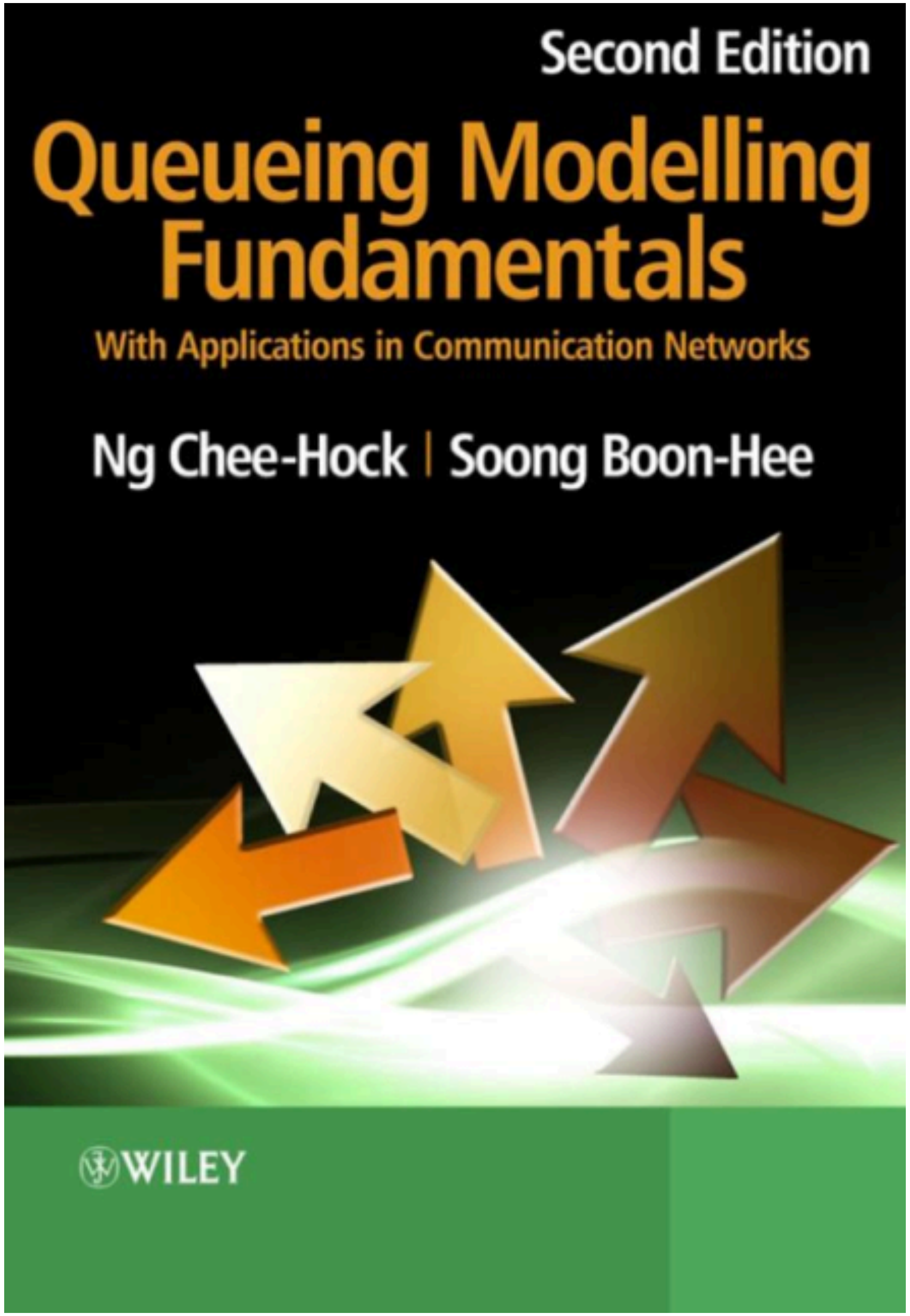
Alternative Spelling: Queuing

Textbook

Chee-Hock Ng & Soong Boon-Hee

Queueing Modelling Fundamentals
With Applications in Communication Networks

John Wiley and Sons, Inc.,
2 edition, 2008



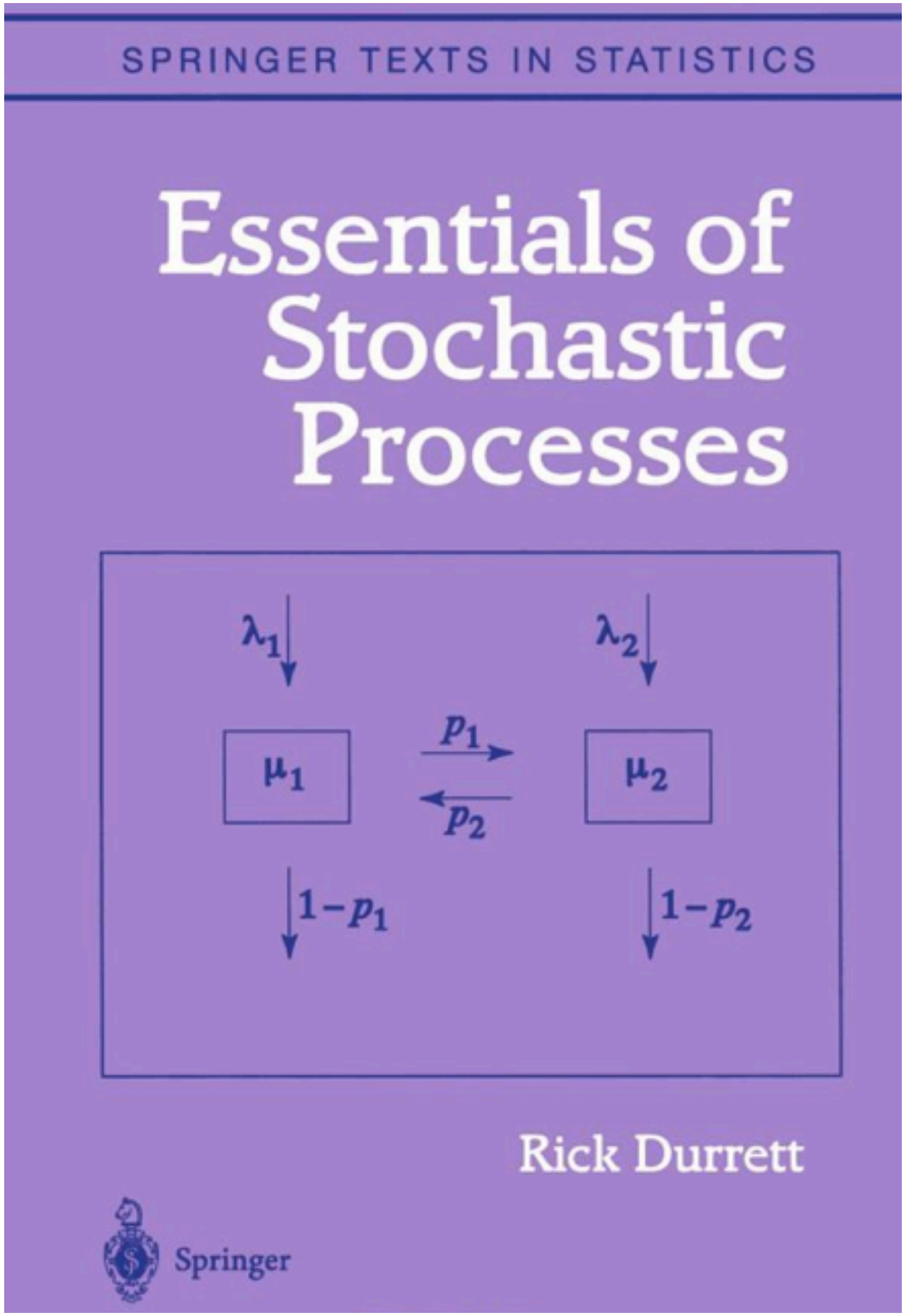
Textbook

Richard Durrett

Essentials of Stochastic Processes

Springer Texts in Statistics

1st ed. 1999
2nd ed. 2010



Queueing Systems

Systémy hromadné obsluhy

Activities today are highly interdependent and intertwined

- Sharing of resources is common in all walks of life
- Sharing leads to waiting for resources in queues

In data communications (e.g. the Internet)

- Data packets are queued in switch/router buffers for transmission

In computer systems

- Jobs are queued for processing by CPU or I/O devices

Queueing Theory

Originally developed for telephone networks

- A.K. Erlang (Danish engineer, published a paper in 1909)
- D.G. Kendall (British statistician, Oxford, Cambridge, introduced notation in 1953)

In modern packet switching networks

- L. Kleinrock (American engineer, MIT, UCLA, main work in early 1960s)

Poisson Process

- The most common model for customer arrivals

Basic Principles and Terminology

Customers

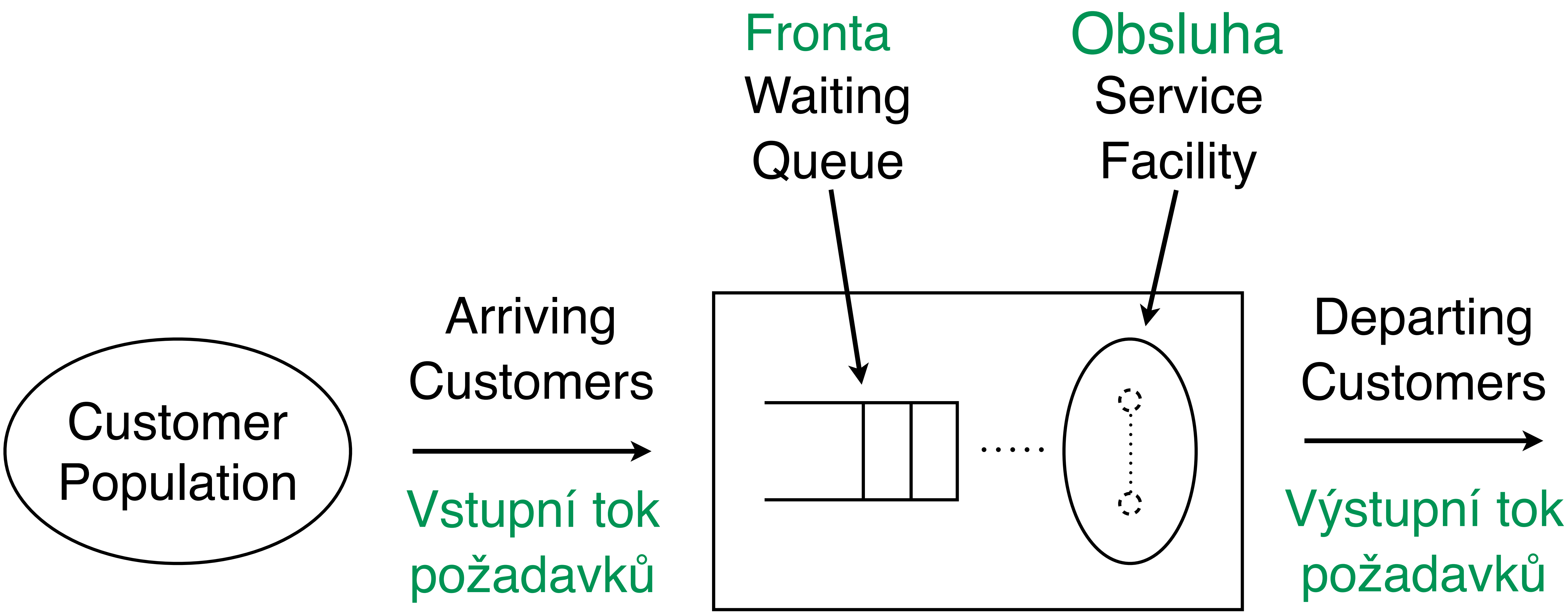
- Arrive according to an “arrival process”
- Want to obtain service from a “service facility”
- Networking: Data packets, data frames, ...
- Computing: Jobs, transactions, user requests, ...

Service facility

- Contains one or more servers
- Each server can serve one customer at a time

Customers join a queue if all servers are occupied

Queueing System Diagram



Prvky systému hromadné obsluhy

Basic Principles and Terminology

Queueing System

- Arrival process
- Service Facility
- Waiting Queue

Systém hromadné obsluhy

- Vstupní tok požadavků
- Obsluha a její režim
- Fronta a její režim

We need a mathematical description of

- The input process
- The system structure
  - Queueing policy; Service policy
- The output process

Characteristics of the Input Process

Charakteristiky vstupního toku požadavků

(i) The size of arriving population

- Finite or infinite
- Infinite is easier to solve – arrival rate not affected by size
- We often assume infinite arriving population
- It approximates a “very large” population

(ii) Behavior of arriving customers

- May leave forever if the queue is full
- May leave randomly if the queue is too long

Characteristics of the Input Process

Charakteristiky vstupního toku požadavků

(iii) Arriving patterns

M: Markovian or Memoryless $\Rightarrow$ Poisson Process
(I.e. exponential & independent interarrival times)

D: Deterministic, constant interarrival times

E_k: Erlang distribution of order k of interarrival times

G: General probability distribution of interarrival times

GI: General & Independent distribution of interarrival times

Default Assumption: Poisson Process

Characteristics of the System Structure

(i) Physical number and layout of servers

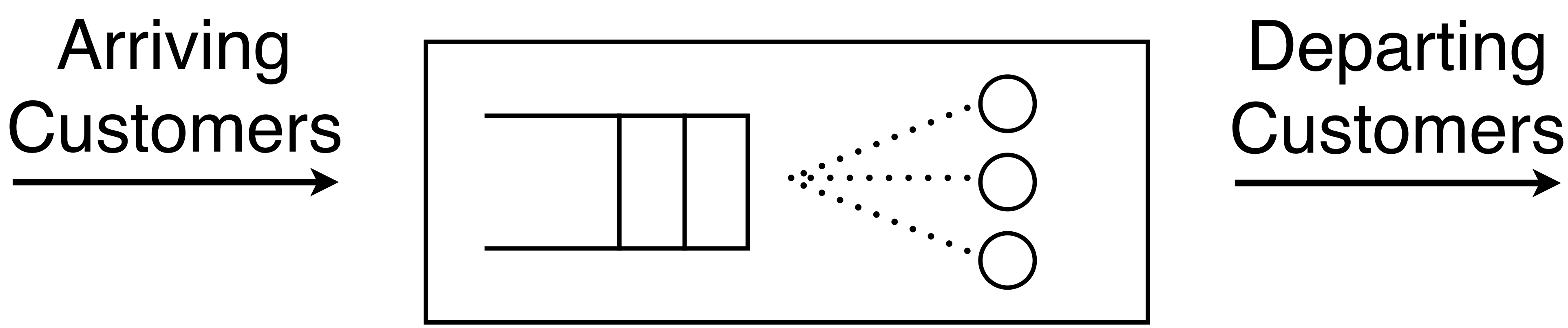
- One or more; <u>identical</u> or different; serial or <u>parallel</u>
- We will focus on parallel identical servers
- I.e. customers can go to any free server and then leave

(ii) The system capacity

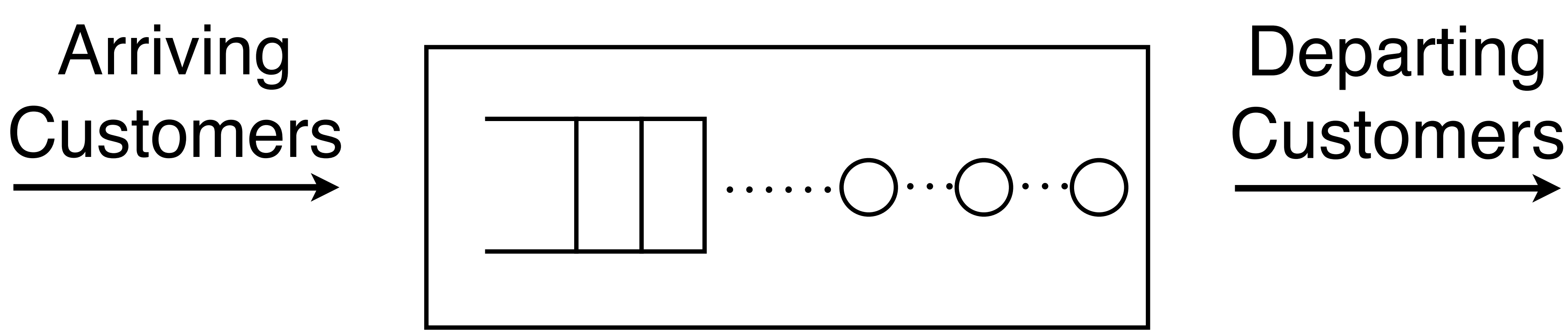
System capacity = waiting customers + served customers

- Non-blocking system – infinite queue
  - Easier to solve – often we assume this
- Blocking system – finite queue
  - Customers leave forever if the queue is full

Physical Layout of Servers



Parallel Servers



Serial Servers

Characteristics of the Output Process

Charakteristiky výstupního toku požadavků

(i) Queueing discipline or serving discipline

Obsluha a její režim / Fronta a její režim

- First-come-first-served (FCFS) / First-in-first-out (FIFO)
- Last-come-first-served (LCFS) / First-in-last-out (FILO)
- Priority based (preemptive or non-preemptive)
- Processor sharing (1/k of time to each of k customers)
- Random

Default Assumption: FCFS/FIFO

Characteristics of the Output Process

Charakteristiky výstupního toku požadavků

(ii) Service time distribution

M: Markovian or Memoryless $\Rightarrow$ exponential service times

D: Deterministic, constant service times

E_k: Erlang distribution of order k of service times

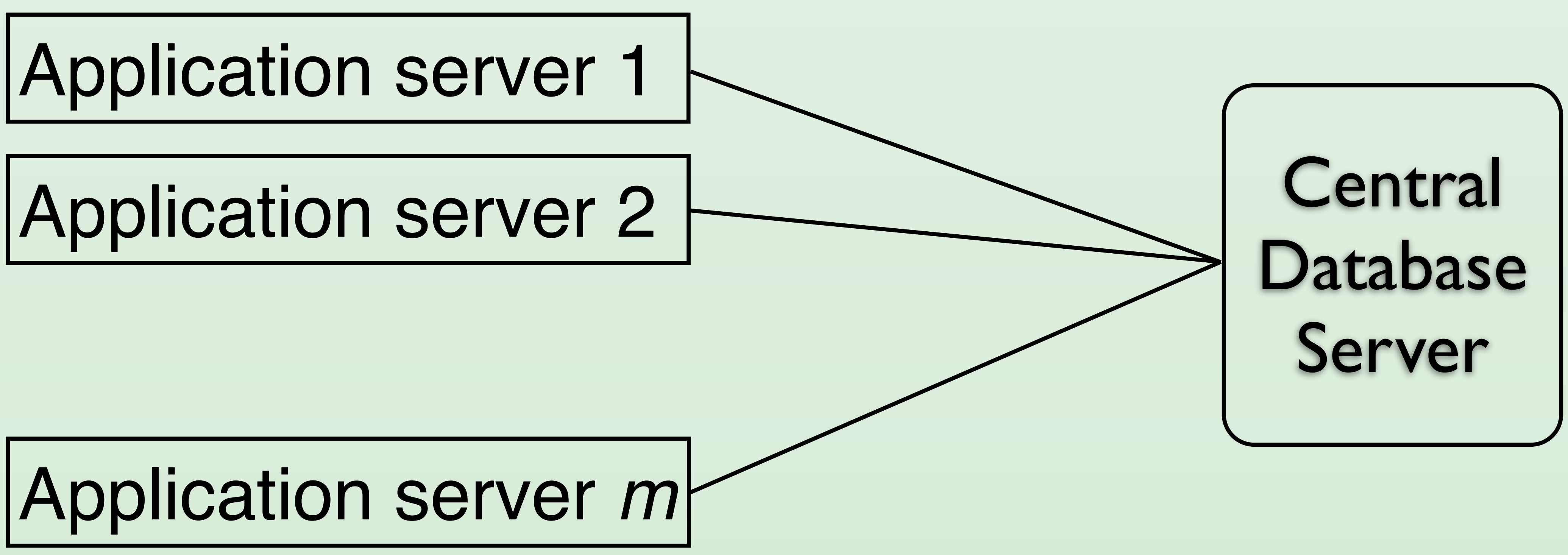
G: General probability distribution of service times

Default Assumption: Exponential service times

Web and Database Servers

Example

Pool of m application servers (e.g. Tomcat) submits a job to a central database server



Web and Database Servers

Example

We assume the Poisson arrival process for the requests. So we obtain a state-dependent Poisson arrival process with the rate

$$\lambda(k) = \begin{cases} (m-k)\lambda & k < m \\ 0 & k \geq m \end{cases}$$

Kendall Notation

$$A \,/\, B \,/\, X \,/\, Y \,/\, Z$$

A: Customer arrival pattern (Interarrival time distribution)
B: Service pattern (Service time distribution)
X: Number of parallel servers
Y: System capacity
Z: Queueing discipline

Default values: $Y = \infty$, $Z = \text{FCFS}$

Example: $M \,/\, M \,/\, 1 = M \,/\, M \,/\, 1 \,/\, \infty \,/\, \text{FCFS}$
(Poisson arrivals, Exp. service times, 1 server)

Plan of Study

We will focus on M/M/m systems

- We must therefore study
  - The Exponential Distribution (interarrival & service times)
  - The Poisson Process (interarrival times are Exponential)
  - Birth & Death Markov chains with continuous time (the number of customers in the system)

We will also look at a M/G/∞ system

- Poisson arrivals, General service time, ∞ many servers

But we will look at some general results first...

Little’s Theorem and Related Results

Ergodicity

Two ways of calculating the average value of a process:

Time average

$$\overline{X}_t = \frac{1}{t}\int_0^t X(\tau)\,d\tau$$

Expected value

$$EX_t = \sum_{k=0}^{\infty} k\ P(X_t = k)$$

Definition

A random process X is **ergodic** if the two averages are equal (have the same limit) as $t \to \infty$

Little's Theorem

Theorem

$$N = \lambda\, T$$

where

N = average number of customers in the system

λ = average arrival rate of customers

T = average time a customer spends in the system

Very general – no assumptions about

- Interarrival and service time distribution
- Queueing policy
- Number of parallel servers

Illustration of Little's Theorem

Example

Consider a deterministic system:

2 customers arrive at the start of every minute

One stays 30s, the other 1 minute (avg. stay $T = 45$s)

How many people will be in the system, on average?

1st half of every minute: 2 customers

2nd half of every minute: 1 customer ... avg. $N = 1.5$

Little's Theorem:

$$N = \lambda\, T = 2 \text{ cust/min} \times 3/4 \text{ min} = 3/2 \text{ customers}$$

Proof of Little's Theorem

Assume FCFS (FIFO) policy

Define

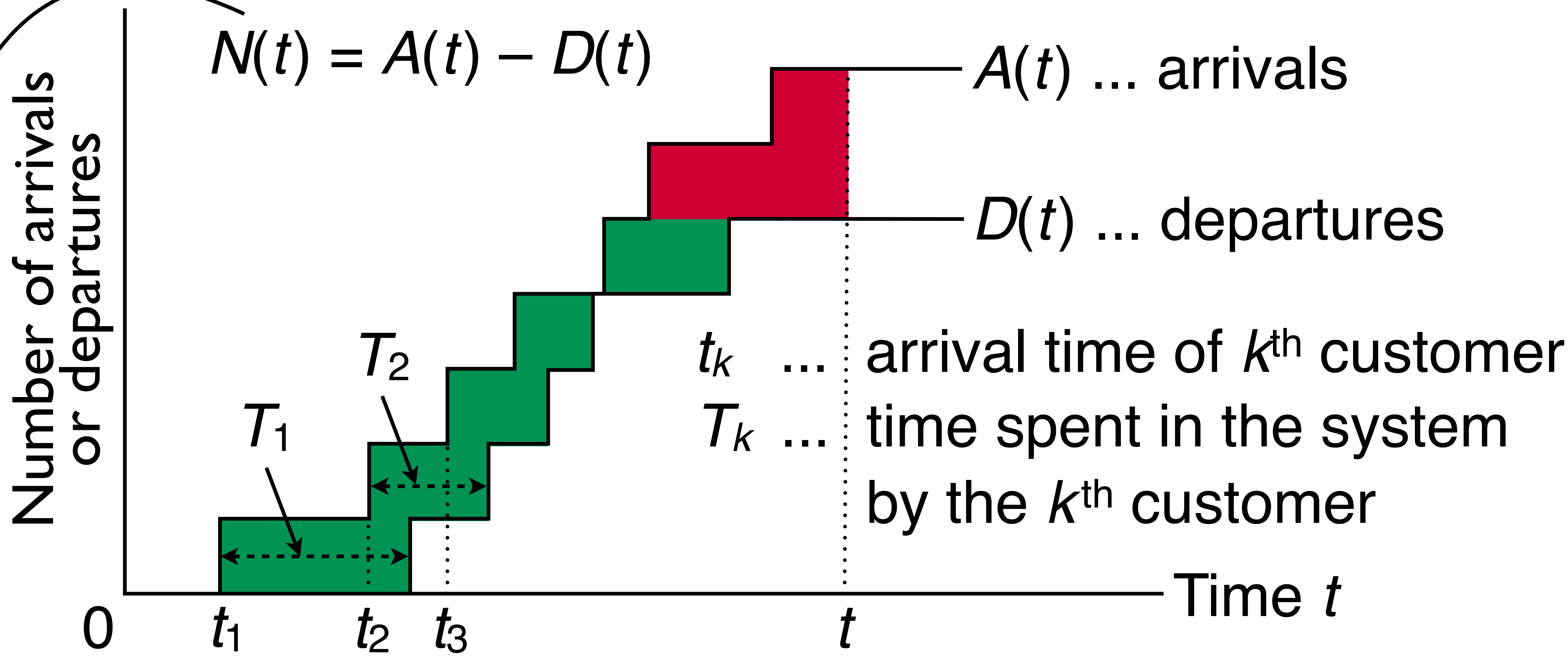
$$\begin{aligned} A(t) &= \text{Number of arrivals during time interval } (0,t) \\ D(t) &= \text{Number of departures during } (0,t) \\ N(t) &= \text{Number of customers in the system at time } t \end{aligned}$$

Assume we start wit empty system at time 0.

Then the number of people in the system at time t is

$$N(t) = A(t) - D(t)$$

Proof of Little's Theorem



Area under $N(t)$ up to time t equals the area between the curves:

departed

in the system

$$\int_0^t N(\tau)\,d\tau = \int_0^t [A(\tau) - D(\tau)]\,d\tau = \sum_{k=1}^{D(t)} T_k \times 1 + \sum_{k=D(t)+1}^{A(t)} (t - t_k) \times 1$$

We got the total number of customers up to time t

$$\int_0^t N(\tau)\,d\tau = \int_0^t [A(\tau) - D(\tau)]\,d\tau = \sum_{k=1}^{D(t)} T_k \times 1 + \sum_{k=D(t)+1}^{A(t)} (t - t_k) \times 1$$

The time average of the number of customers N_t is

$$N_t = \frac{1}{t}\int_0^t N(\tau)\,d\tau = \left[\sum_{k=1}^{D(t)} T_k + \sum_{k=D(t)+1}^{A(t)} (t - t_k)\right] \times \frac{1}{t}$$

$$= \underbrace{\left[\sum_{k=1}^{D(t)} T_k + \sum_{k=D(t)+1}^{A(t)} (t - t_k)\right] \times \frac{1}{A(t)}}_{T_t} \times \underbrace{\frac{A(t)}{t}}_{\lambda_t}$$

$$N_t = T_t \times \lambda_t$$

T_t: time average of the time spent by a customer

λ_t: time average of the arrival rate

Proof of Little's Theorem

We have

$$N_t = T_t \times \lambda_t$$

N_t ... time average of the number of customers
T_t ... time average of the time spent by a customer
λ_t ... time average of the arrival rate

For ergodic processes the time average

$N_t \rightarrow N =$ Expectation in steady state, as $t \rightarrow \infty$

Let $t \rightarrow \infty$ to get

$$N = T \times \lambda$$

Note: Most queueing systems are ergodic.

Resource Utilization

Definition

Resource utilization of a queueing system is the fraction of time the system is used

$$\rho = \frac{\text{Time a server is occupied}}{\text{Time available}}$$

Resource Utilization

Note

With m servers, N customers during $(t, t{+}T)$, arrival rate λ, each server serves on average $N\ /\ m = (\lambda T/\ m)$ customers.

If the average service time is $(1/\mu)$ then

$$\rho = \frac{\text{Time a server is occupied}}{\text{Time available}}$$

$$= \frac{(\lambda T/m) \times (1/\mu)}{T} = \frac{\lambda}{m\mu}$$

μ ... average number served per unit of time

Traffic Intensity

Definition

Traffic Intensity (offered load) of a queueing system is the product of the average arrival rate λ and the average service time ($1/\mu$):

$$\alpha = \frac{\lambda}{\mu}$$

Flow Conservation Law

Proposition

For a stable queueing system

$$\text{rate of customers leaving} = \text{rate of customers entering}$$

$$\lambda_{out} = \lambda_{in}$$