Aalto University
School of Electrical
Engineering

ELEC-E7450
Performance Analysis

ELEC-C7210 recap

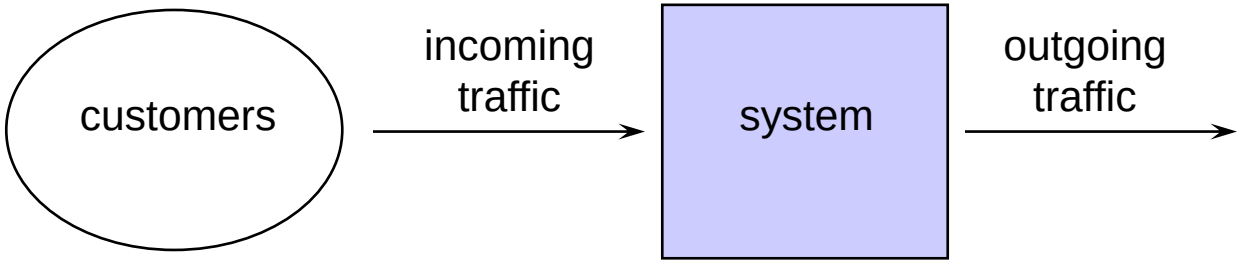
Samuli Aalto
Department of Communications and Networking

Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Stochastic service system

- System has service capacity that is used to serve customers
- Service discipline determines how the service capacity is shared among the customers
- Customers arrive randomly (incoming traffic), have random service requirements (service times), and depart after service (outgoing traffic)
- Traffic load depends both on the arrivals and the service requirements
- System performance depends on service capacity, service discipline and traffic load

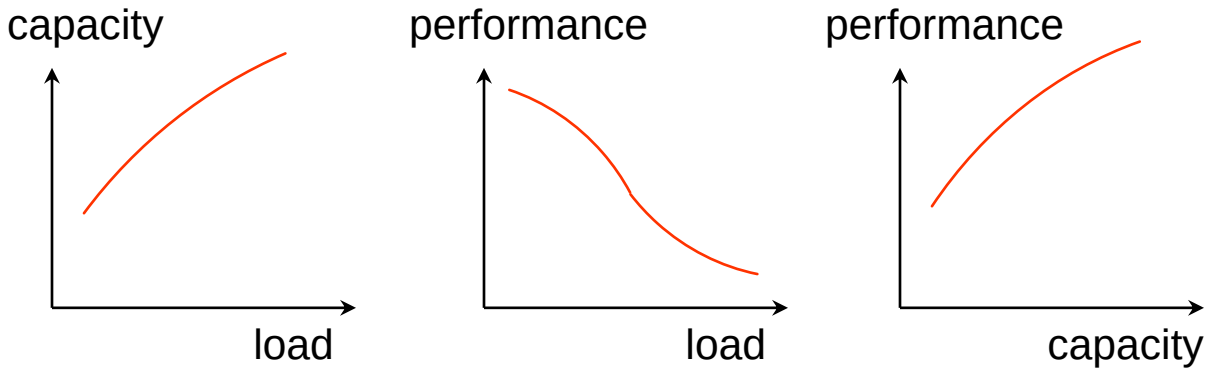


Interesting questions

- Design:
  Given the traffic load and the required system performance,
  what is the required service capacity (for a given service discipline)?
- Operation:
  Given the service capacity and the required system performance,
  what is the maximum traffic load allowed?
- Analysis:
  Given the service capacity, the service discipline, and the traffic load,
  what is the system performance?
- Optimization:
  Given the service capacity and the traffic load,
  what is the optimal service discipline to maximize the performance?

Relationships between different factors

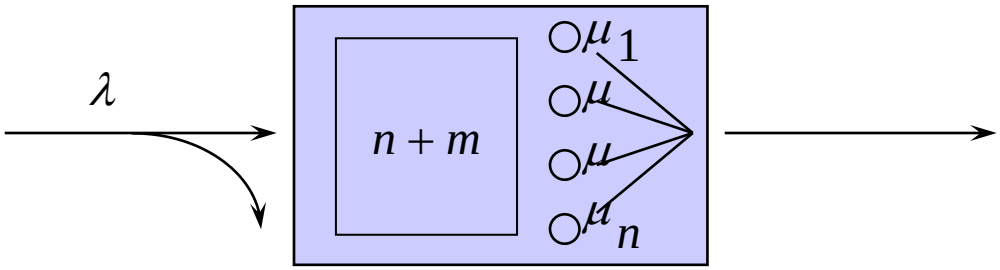
- Relationships between capacity, load, and performance for any reasonable discipline are easy to understand qualitatively
- However, stochastic queueing models are needed to describe the relationships quantitatively



Contents

Basic queueing model

- Customers arrive at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time (in time units)
- Customers are served by n parallel servers
- When busy, each server serves at rate μ (customers per time unit)
  - $1/\mu$ = average service time (in time units)
- There are $p = n + m$ customer places in the system
  - If all $n + m$ customer places are occupied when a new customer arrives, the customer is not served but lost



Kendall's notation: A/B/n/p/k

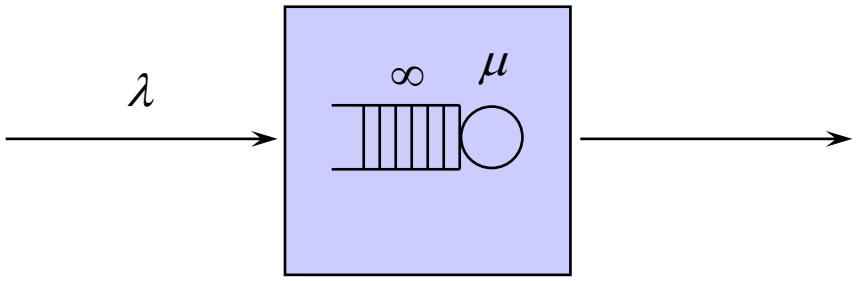
- A refers to the arrival process.
  Default: IID interarrival times.
- B refers to the service process.
  Default: IID service times.
- Distribution codes for A and B:
  - M = exponential (memoryless)
  - D = deterministic
  - G = generally distributed
- n = nr of (parallel) servers
- p = nr of customer places
- k = size of customer population
- Default values (usually omitted):
  - $p = \infty, k = \infty$

IID = independently and identically distributed

- Examples:
  - M/M/1 (Single-server queue)
  - M/D/1 (Single-server queue)
  - M/G/1 (Single-server queue)
  - M/M/n (Multi-server queue)
  - M/G/n (Multi-server queue)
  - M/G/∞ (Infinite-server system)
  - M/M/∞ (Poisson model)
  - M/M/n/n (Erlang loss system)
  - M/M/k/k/k (Binomial model)
  - M/M/n/n/k (Engset loss system)

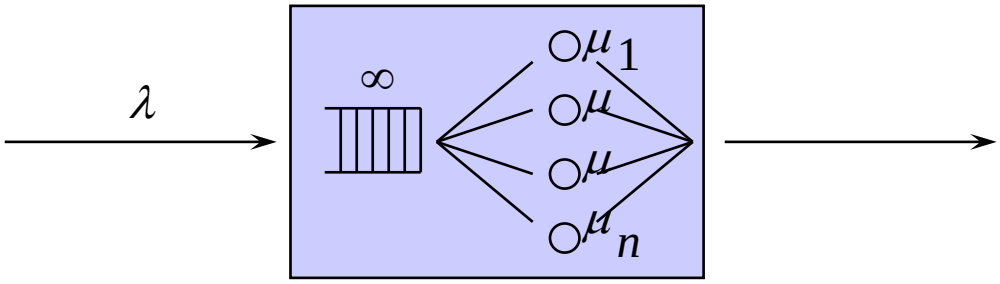
Single-server queue G/G/1

- Customers arrive at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are served by 1 server
- When busy, the server serves at rate μ (customers per time unit)
  - $1/\mu$ = average service time
- There is an infinite number of customer places
  - No customers are lost, but they may be delayed and they may accumulate in the system if the traffic load is too high (causing delays and even instability)



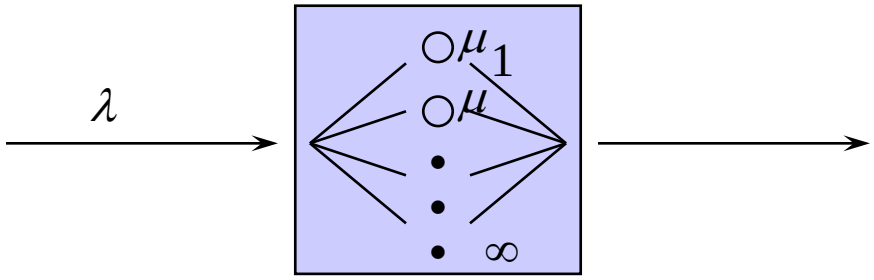
Multi-server queue G/G/n

- Customers arrive at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are served by n parallel servers
- When busy, each server serves at rate μ (customers per time unit)
  - $1/\mu$ = average service time
- There is an infinite number of customer places
  - No customers are lost, but they may be delayed and they may accumulate in the system if the traffic load is too high (causing delays and even instability)



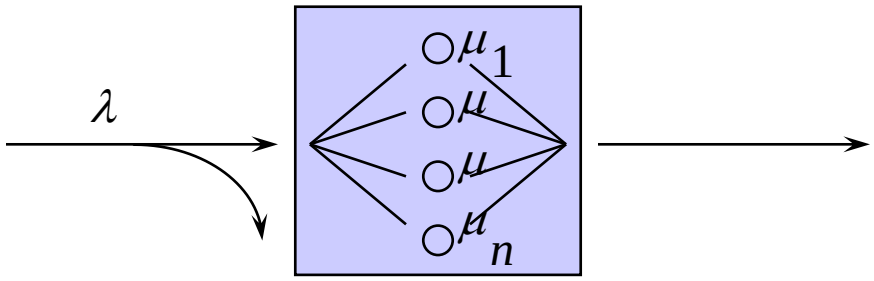
Infinite-server system G/G/∞

- Customers arrive at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are served by an infinite number of parallel servers
- When busy, each server serves at rate μ (customers per time unit)
  - $1/\mu$ = average service time
- There is an infinite number of customer places
  - No customers are lost nor delayed (i.e., an ideal system)



Loss system G/G/n/n

- Customers arrive at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are served by n parallel servers
- When busy, each server serves at rate μ (customers per time unit)
  - $1/\mu$ = average service time
- There are no waiting places
  - No customers are delayed, but they may be blocked if the traffic load is too high (causing losses)



Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Little's formula

- Consider a service system where
  - new customers arrive at rate λ
- Assume stability:
  - Every now and then, the system is empty
- Consequence:
  - Customers are not accumulated in the system so that they depart at rate λ
- Let
  - $\overline{N}$ = average number of customers in the system
  - $\overline{T}$ = average time a customer spends in the system = average delay

- **Theorem (Little's formula):**

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

Proof (1)

- $N(t)$ = the number of customers in the system at time t
- $A(t)$ = the number of customers arrived in the system by time t
- $B(t)$ = the number of customers departed from the system by time t
- T_i = the time customer i spends in the system = its delay
- Always

$$A(t) \geq B(t)$$

- Due to the stability assumption, we have, as $t \to \infty$ and $n \to \infty$,

$$\frac{1}{t} A(t) \to \lambda, \quad \frac{1}{t} B(t) \to \lambda \tag{1}$$

$$\frac{1}{t} \int_0^t N(s) ds \to \overline{N}, \quad \frac{1}{n} \sum_{i=1}^{n} T_i \to \overline{T} \tag{2}$$

Proof (2)

- We may assume that
  - the system is empty at time $t = 0$
- Then (see the figures in the following slides)

$$\sum_{i=1}^{B(t)} T_i \le \int_0^t N(s)ds \le \sum_{i=1}^{A(t)} T_i$$

- Thus,

$$\frac{B(t)}{t}\frac{1}{B(t)}\sum_{i=1}^{B(t)} T_i \le \frac{1}{t}\int_0^t N(s)ds \le \frac{A(t)}{t}\frac{1}{A(t)}\sum_{i=1}^{A(t)} T_i$$

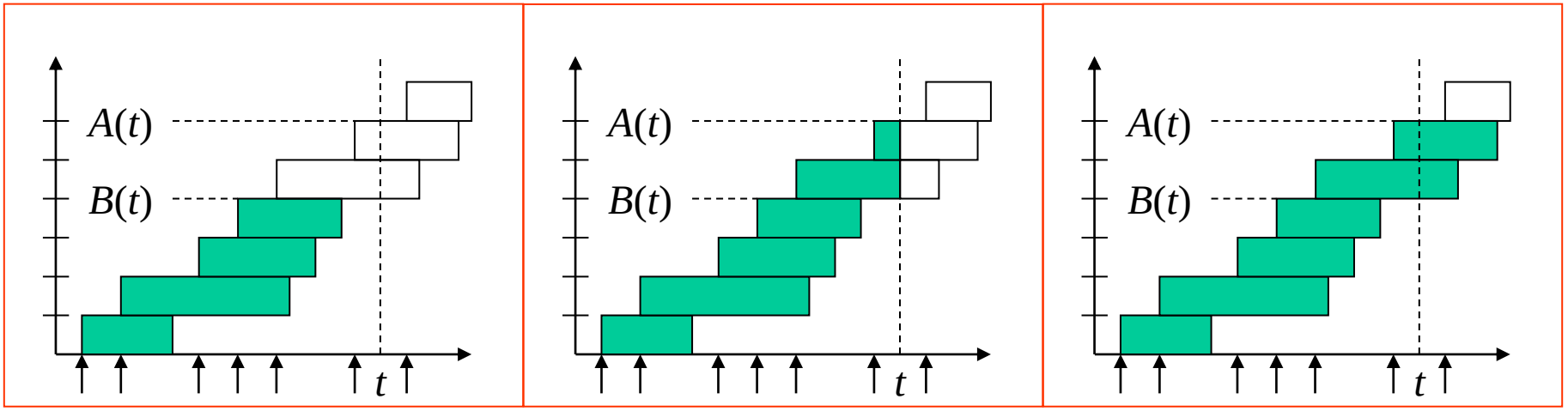
- As $t \to \infty$, we have, by (1) and (2),

$$\lambda \overline{T} \le \overline{N} \le \lambda \overline{T}$$

- Q.E.D.

Proof (3)

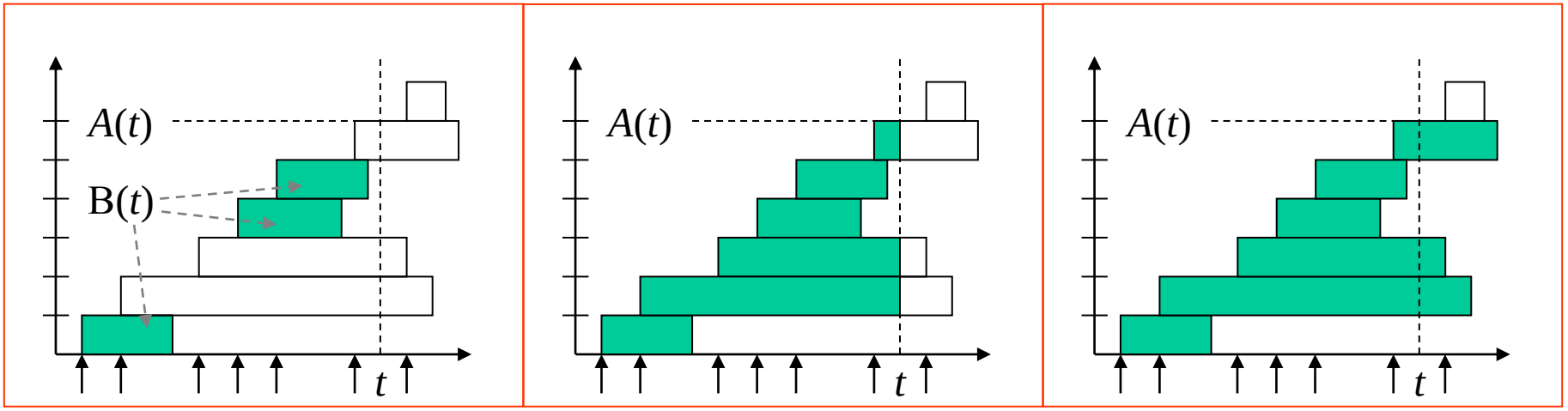
- Single server and FIFO service discipline:



$$\sum_{i=1}^{B(t)} T_i \leq \int_0^t N(s)ds \leq \sum_{i=1}^{A(t)} T_i$$

Proof (4)

- Any number of servers and any service discipline:



$$\sum_{i \in B(t)} T_i \leq \int_0^t N(s)ds \leq \sum_{i=1}^{A(t)} T_i$$

Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Exponential distribution

$$X \sim \mathrm{Exp}(\mu), \quad \mu > 0$$

- continuous counterpart of the geometric distribution ("failure" prob. $\approx \mu dt$)
- μ = intensity (of an exponential phase)
- $P\{X \in (t,t+h] \mid X > t\} = \mu h + o(h)$, where $o(h)/h \to 0$ as $h \to 0$

- Value space: $S_X = (0,\infty)$
- PDF and CDF:

$$f_X(x) = \mu e^{-\mu x}, \quad x > 0$$

$$F_X(x) := P\{X \le x\} = 1 - e^{-\mu x}$$

Moments

$$X \sim \mathrm{Exp}(\mu), \quad \mu > 0$$

- Mean value: $E[X] = \int_0^\infty \mu x\, e^{-\mu x}\, dx = 1/\mu$
- Second moment: $E[X^2] = \int_0^\infty \mu x^2\, e^{-\mu x}\, dx = 2/\mu^2$
- Variance: $D^2[X] = E[X^2] - E[X]^2 = 1/\mu^2$
- Standard deviation: $D[X] = \sqrt{D^2[X]} = 1/\mu$
- Coefficient of variation: $C[X] = D[X]/E[X] = 1$

Memoryless property

- Exponential distribution has the so called memoryless property

$$P\{X - x > y \mid X > x\} = P\{X > y\}, \quad x, y > 0$$

since

$$P\{X - x > y \mid X > x\} = \frac{P\{X > x+y\}}{P\{X > x\}} = \frac{e^{-\mu(x+y)}}{e^{-\mu x}} = e^{-\mu y} = P\{X > y\}$$

- In fact, only the exponential distribution has this property (among the continuous distributions)
- It follows that the mean residual lifetime (MRL) remains the same:

$$\mathrm{MRL}(x) := E[X - x \mid X > x] = \frac{1}{\mu}$$

Minimum of exponential random variables

- Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim \text{Exp}(\mu_i)$. Then

$$X^{\min} := \min\{X_1, \ldots, X_n\} \sim \text{Exp}(\mu_1 + \ldots + \mu_n)$$

since

$$P\{X^{\min} > x\} = P\{X_1 > x\} \ldots P\{X_n > x\} = e^{-(\mu_1 + \ldots + \mu_n)x}$$

- In addition, we have

$$E[X^{\min}] = \frac{1}{\mu_1 + \ldots + \mu_n}, \quad P\{X^{\min} = X_i\} = \frac{\mu_i}{\mu_1 + \ldots + \mu_n}$$

Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Stochastic processes (1)

- **Definition**: A (real-valued) stochastic process $X = (X_t \mid t \in I)$ is a collection of random variables X_t
  - taking values in some (real-valued) set S, $X_t(\omega) \in S$, and
  - indexed by a real-valued (time) parameter $t \in I$.
- Stochastic processes are also called random processes
  - The index set $I \subset \Re$ is called the parameter space of the process
  - The value set $S \subset \Re$ is called the state space of the process
- As a shorthand notation, X_t is used to refer to the whole stochastic process (in addition to a single random variable related to time t)

Stochastic processes (2)

- Each (individual) random variable X_t is a mapping from the sample space Ω into the real values $\Re$:

$$X_t : \Omega \to \Re, \quad \omega \mapsto X_t(\omega)$$

- Thus, a stochastic process X can be seen as a mapping from the sample space Ω into the set of real-valued functions $\Re^I$ (with $t \in I$ as an argument):

$$X : \Omega \to \Re^I, \quad \omega \mapsto X(\omega)$$

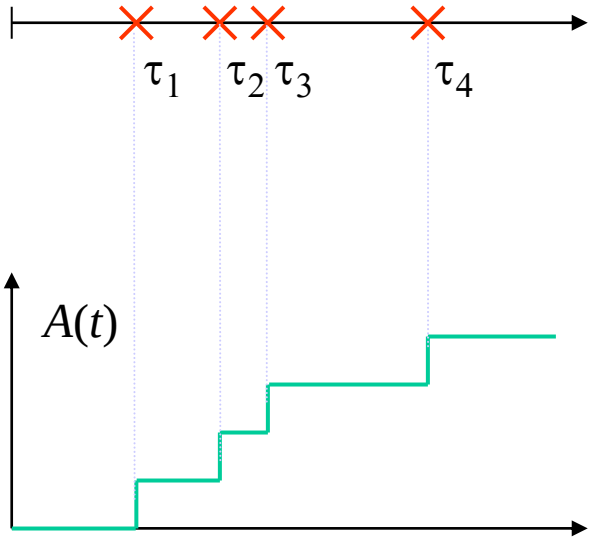
- Each sample point $\omega \in \Omega$ is associated with a real-valued function $X(\omega)$. Function $X(\omega)$ is called a realization of the process.

Stochastic processes (3)

- Given the sample point $\omega \in \Omega$
  - $X(\omega) = (X_t(\omega) \mid t \in I)$ is a real-valued function (of $t \in I$)
- Given the time index $t \in I$,
  - $X_t = (X_t(\omega) \mid \omega \in \Omega)$ is a random variable
- Given the sample point $\omega \in \Omega$ and the time index $t \in I$,
  - $X_t(\omega)$ is a real value

Examples

- Point process τ_n
  - model for random events in time (e.g., arrivals in a queueing system)
  - τ_n = the occuring time of the *n*-th event
  - discrete-time, $I = \{1,2,\ldots\}$
  - continuous-state, $S = (0,\infty)$
- Counter process $A(t)$
  - counting random events as time evolves (e.g., arrivals in a queueing system)
  - $A(t)$ = the number of arrivals until time t
  - continuous-time, $I = [0,\infty)$
  - discrete-state, $S = \{0,1,\ldots\}$



Poisson distribution

$$X \sim \text{Poisson}(a), \quad a > 0$$

- limit of binomial distribution as $n \to \infty$ and $p \to 0$ in such a way that $np \to a$

- Value space: $S_X = \{0,1,\ldots\}$
- Point probabilities:

$$P\{X = i\} = \frac{a^i}{i!} e^{-a}$$

- Mean value: $E[X] = a$
- Second moment: $E[X^2] = a^2 + a$
- Variance: $D^2[X] = E[X^2] - E[X]^2 = a$

Poisson process

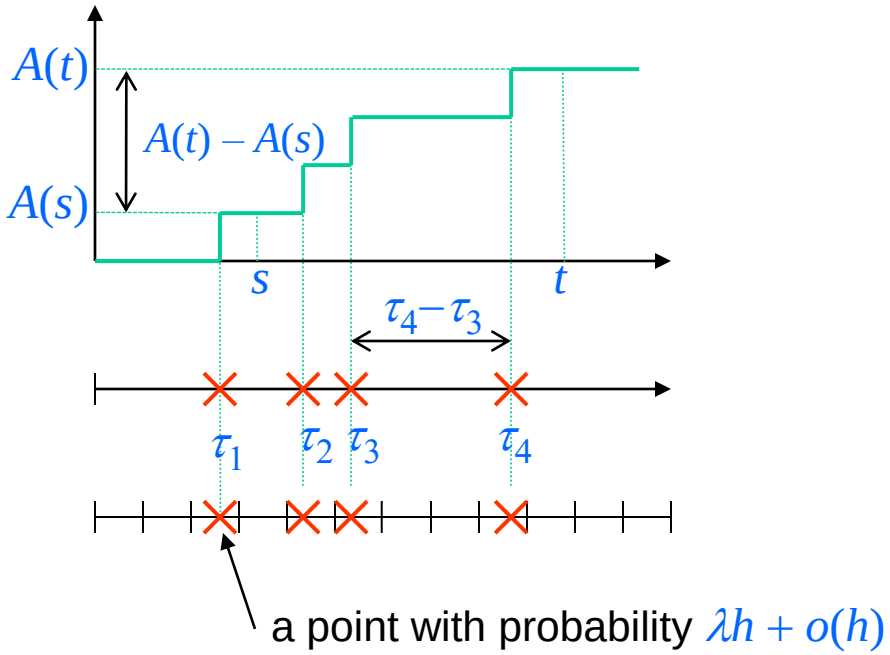
- **Definition 1**:
  A point process $(\tau_n \mid n = 1,2,\ldots)$ is a Poisson process with intensity λ if the intervals $\tau_n - \tau_{n-1}$ are IID with distribution $\mathrm{Exp}(\lambda)$

- **Remark**:
  The probability that there is a point during a short time interval $(t, t+h]$ is $\lambda h + o(h)$ independently of the other time intervals

- **Definition 2**:
  A counter process $(A(t) \mid t \geq 0)$ is a Poisson process with intensity λ if its increments $A(t) - A(s)$ in disjoint intervals $(s, t]$ are independent obeying the Poisson distribution with parameter $\lambda(t-s)$

Three characterizations

- It is possible to show that the different characterizations for a Poisson process are, indeed, equivalent



Uniformly distributed points

- **Property 1 (Uniform distribution)**:
  Let $A(t)$ be a Poisson process with intensity λ. If $A(t) - A(s) = n$, these n points are independently and uniformly distributed within the interval $(t, t+s]$

  – **Proof**:
  Follows from the discrete-time construction.

Sum process

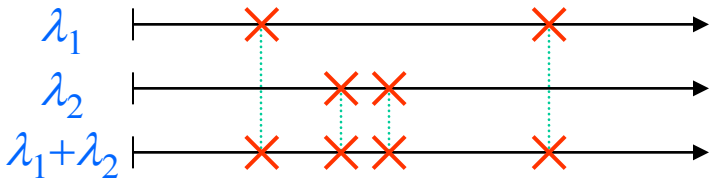
- **Property 2 (Sum)**:
  Let $A_1(t)$ and $A_2(t)$ be two independent Poisson processes with intensities λ_1 and λ_2. Then the sum (superposition) process $A_1(t) + A_2(t)$ is a Poisson process with intensity $\lambda_1 + \lambda_2$.

  – **Proof**:
  Probability that there are no points in an interval $(t, t+h]$ the superposition is

$$(1-\lambda_1 h+o(h))(1-\lambda_2 h+o(h)) = 1-(\lambda_1+\lambda_2)h+o(h)$$

On the other hand, the probability that there is exactly one point is

$$(\lambda_1 h+o(h))(1-\lambda_2 h+o(h))+(1-\lambda_1 h+o(h))(\lambda_2 h+o(h)) = (\lambda_1+\lambda_2)h+o(h)$$



Random sampling

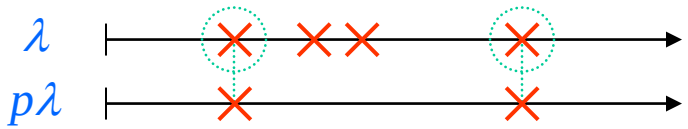
- **Property 3 (Random sampling)**:
  Let τ_n be a Poisson process with intensity λ. Denote by σ_n the point process resulting from a random and independent sampling (with probability p) of the points of τ_n. Then σ_n is a Poisson process with intensity $p\lambda$.

  – **Proof**:
  Probability that there are no points in an interval $(t, t+h]$ the superposition is

$$(1-\lambda h+o(h))+(1-p)(\lambda h+o(h))=1-p\lambda h+o(h)$$

On the other hand, the probability that there is exactly one point is

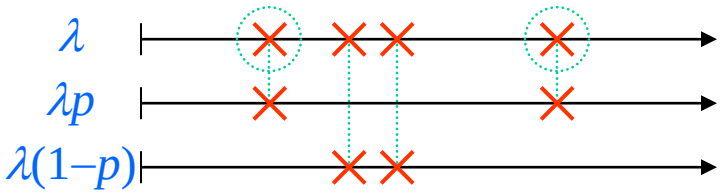
$$p(\lambda h+o(h))=p\lambda h+o(h)$$



Random sorting

- **Property 4 (Random sorting)**:
  Let τ_n be a Poisson process with intensity λ. Denote by σ_n the point process resulting from a random and independent sampling (with probability p) of the points of τ_n. Denote by ρ_n the point process resulting from the remaining points. Then σ_n and ρ_n are independent Poisson processes with intensities λp and $\lambda(1-p)$, respectively.

  – **Proof**:
  Due to Property 2, it is enough to prove that the resulting two processes are independent. This boils down to show that the number of points during an interval in the two processes are independent, which follows from the discrete-time construction.



PASTA

- **Property 5 (PASTA)**:
  Consider a stable service system with Poisson arrivals. Let $X(t)$ denote the state of system at time t (continuous-time process) and X_n^* the state of the system seen by the nth arriving customer (embedded discrete-time process). Then the limiting distribution of $X(t)$ is the same as the limiting distribution of X_n^*. Thus, we can say that arriving customers see the system in the steady state.
  - PASTA = "Poisson Arrivals See Time Averages"

- The PASTA property is only valid for Poisson arrivals (but not for other arrival processes)
  - Consider e.g. your own laptop. Whenever you start a new session, the system is always idle. In continuous time, however, the system is not only idle but also busy (when you use it).

Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Markov process

- Consider a continuous-time and discrete-state stochastic process $X(t)$ with a countable state space S
- **Definition**:
  Process $X(t)$ is a Markov process with transition rates $q_{ij} \geq 0$, $j \neq i$, if
  - holding times T_i in state i are independent and exponentially distributed with intensity $q_i := \Sigma_{j \neq i} q_{ij}$
  - when jumping, $X(t)$ jumps from state i to state $j \neq i$ with probability $p_{ij} := q_{ij}/q_i$ (independently of everything else)

- **Remark:**
  Each Markov process has the following Markov property

$$P\{X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1, \ldots, X(t_n) = x_n\} = P\{X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n\}$$

Transition matrix

- State transitions in a Markov process $X(t)$ are time-homogeneous:

$$P\{X(s+t) = j \mid X(s) = i\} = P\{X(t) = j \mid X(0) = i\}$$

- The transition rates q_{ij}, $j \neq i$, and q_i satisfy

$$q_{ij} = \lim_{h \downarrow 0} \frac{1}{h} P\{X(h) = j \mid X(0) = i\}$$

$$q_i = \lim_{h \downarrow 0} \frac{1}{h} P\{X(h) \neq i \mid X(0) = i\}$$

- Denote the corresponding transition matrix by Q,

$$Q := (q_{ij}; i, j \in S)$$

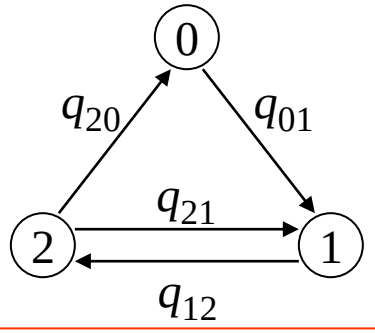
with diagonal elements $q_{ii} := -q_i$

Transition diagram

- A Markov process can be represented by a transition diagram, which is a directed graph where
  - nodes correspond to states and
  - one-way links correspond to potential state transitions: there is a link from node i to j if and only if $q_{ij} > 0$

- **Example**:
  Markov process with three states, $S = \{0,1,2\}$

$$Q = \begin{pmatrix} -q_{01} & q_{01} & 0 \\ 0 & -q_{12} & q_{12} \\ q_{20} & q_{21} & -(q_{20}+q_{21}) \end{pmatrix}$$



Global balance equations (1)

- **Definition**:
  Let $\pi = (\pi_i \mid \pi_i \geq 0,\ i \in S)$ be a distribution defined on the state space S,

$$\sum_{i \in S} \pi_i = 1 \qquad \text{(N)}$$

  It is an equilibrium distribution of the Markov process $X(t)$ if the following global balance equations (GBE) are satisfied for each $i \in S$:

$$\sum_{j \neq i} \pi_i q_{ij} = \sum_{j \neq i} \pi_j q_{ji} \qquad \text{(GBE)}$$

- It is possible that no equilibrium distribution exists, or there may be multiple of them.
- However, if a unique equilibrium distribution π exists, it is the same as the limiting distribution:

$$\pi_i = \lim_{t \to \infty} P\{X(t) = i\}$$

Global balance equations (2)

- The global balance equations (GBE) are linearly dependent:
  - any one fixed equation is automatically satisfied if the other ones are satisfied
- Therefore, the normalization condition (N) is always needed when solving the equilibrium distribution

Irreducibiliity

- **Definition**:
  There is a path from state i to state j ($i \to j$) if there is a directed path from state i to state j in the corresponding transition diagram.
  - In this case, starting from state i, the process visits state j with a positive probability (sometimes in the future)

- **Definition**:
  States i and j communicate ($i \leftrightarrow j$) if $i \to j$ and $j \to i$.

- **Definition**:
  Markov process is irreducible if all states $i, j \in S$ communicate

Equilibrium distribution (1)

- **Theorem**:
  Assume that an irreducible Markov process $X(t)$ has an equilibrium distribution π.
  (i) Then the equilibrium distribution is unique, and equals the limiting distribution of the process,

$$\pi_i = \lim_{t\to\infty} P\{X(t) = i\}$$

  (ii) If additionally the equilibrium distribution is the initial distribution, then the process is stationary, and the equilibrium distribution equals the stationary distribution,

$$\pi_i = P\{X(t) = i\} \quad \text{for all } t$$

Equilibrium distribution (2)

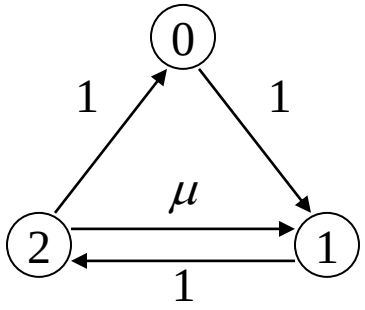
- **Proposition**:
  If an irreducible Markov process $X(t)$ does not have an equilibrium distribution, then

$$\lim_{t\to\infty} P\{X(t)=i\}=0 \quad \text{for all } i$$

- **Proposition**:
  An irreducible Markov process $X(t)$ with a finite state space has always a unique equilibrium distribution π.

Example

$$Q = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & \mu & -(1+\mu) \end{pmatrix}$$



$$\pi_0 + \pi_1 + \pi_2 = 1 \qquad \text{(N)}$$

$$\begin{aligned} \pi_0 \cdot 1 &= \pi_2 \cdot 1 \\ \pi_1 \cdot 1 &= \pi_0 \cdot 1 + \pi_2 \cdot \mu \\ \pi_2 \cdot (1+\mu) &= \pi_1 \cdot 1 \end{aligned} \qquad \text{(GBE)}$$

$$\Rightarrow \quad \pi_0 = \frac{1}{3+\mu}, \quad \pi_1 = \frac{1+\mu}{3+\mu}, \quad \pi_2 = \frac{1}{3+\mu}$$

Detailed balance equations

- **Proposition**:
  Let $X(t)$ be a Markov process defined by transition rates $q_{i,j}$ and $\pi = (\pi_i \mid \pi_i \geq 0,\ i \in S)$ a distribution defined on the state space S:

$$\sum_{i \in S} \pi_i = 1 \qquad \text{(N)}$$

  If the detailed balance equations (DBE)

$$\pi_i q_{ij} = \pi_j q_{ji} \qquad \text{(DBE)}$$

  are satisfied for each pair i,j, then π is an equilibrium distribution of $X(t)$.

  - **Proof**:
    (GBE) follows from (DBE) by summing over all $j \neq i$

Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Birth-death process

- **Definition**:
  A one-dimensional Markov process $X(t)$ is a birth-death process if the state transitions are possible only between neighbouring states,

$$|i-j|>1 \quad \Rightarrow \quad q_{ij}=0$$

- In this case, we denote

$$\lambda_i := q_{i,i+1} \geq 0$$

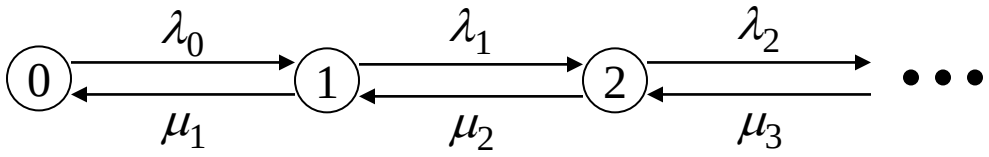
$$\mu_{i+1} := q_{i+1,i} \geq 0$$

- These are called the birth and death rates, respectively.

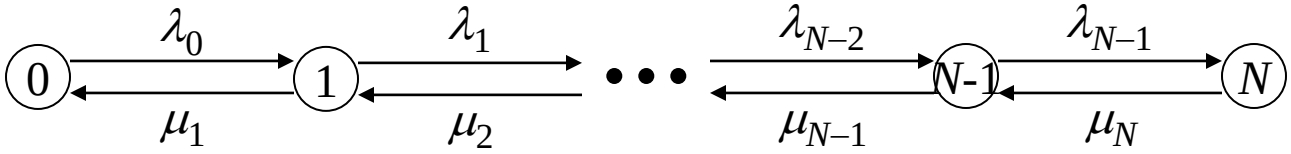
Irreducibility

- **Remark**:
  A birth-death process is clearly irreducible if and only if
  $\lambda_i > 0$ and $\mu_{i+1} > 0$ for all i

- State transition diagram of an infinite-state irreducible BD process:



- State transition diagram of a finite-state irreducible BD process:



Equilibrium distribution (1)

- Consider an irreducible birth-death process $X(t)$.
- The equilibrium distribution π (if it exists) can be derived using the detailed balance equations (DBE):

$$\pi_i \lambda_i = \pi_{i+1} \mu_{i+1} \qquad \text{(DBE)}$$

- Thus we get the following recursive formula:

$$\pi_{i+1} = \frac{\lambda_i}{\mu_{i+1}} \pi_i \quad \Rightarrow \quad \pi_i = \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}$$

- The normalizing condition (N) gives

$$\sum_{i \in S} \pi_i = \pi_0 \sum_{i \in S} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} = 1 \qquad \text{(N)}$$

Equilibrium distribution (2)

- Thus, the equilibrium distribution exists if and only if

$$\sum_{i \in S} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} < \infty$$

- Finite state space: The sum above is always finite, and the equilibrium distribution is

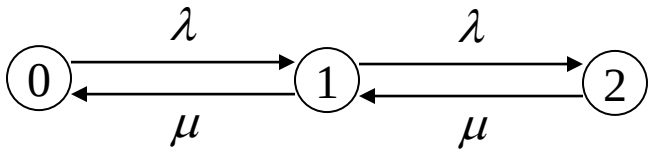
$$\pi_i = \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}, \quad \pi_0 = \left(1 + \sum_{i=1}^{N} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \right)^{-1}$$

- Infinite state space: If the sum above is finite, the equilibrium distribution is

$$\pi_i = \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}, \quad \pi_0 = \left(1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \right)^{-1}$$

Example (M/M/1/2)

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 \\ \mu & -(\lambda+\mu) & \lambda \\ 0 & \mu & -\mu \end{pmatrix}$$



$$\pi_i \lambda = \pi_{i+1} \mu$$

$$\Rightarrow \quad \pi_{i+1} = \rho \pi_i \quad (\rho := \lambda / \mu) \qquad \text{(DBE)}$$

$$\Rightarrow \quad \pi_i = \pi_0 \rho^i$$

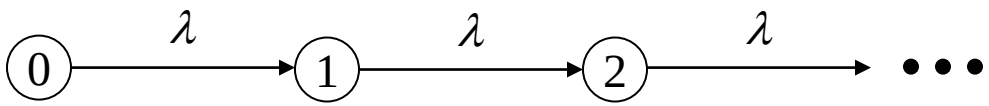
$$\pi_0 + \pi_1 + \pi_2 = \pi_0 (1 + \rho + \rho^2) = 1 \qquad \text{(N)}$$

$$\Rightarrow \quad \pi_i = \frac{\rho^i}{1+\rho+\rho^2}$$

Pure birth process

- **Definition**:
  A birth-death process is a pure birth process if $\mu_i = 0$ for all $i \in S$

- **Example**:
  Poisson process is a pure birth process with a constant birth rate $\lambda_i = \lambda$ for all $i \in S = \{0,1,\ldots\}$

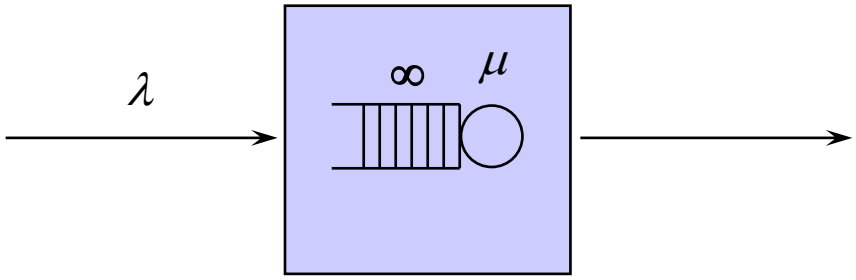


Contents

- Stochastic service system
- Basic queueing models
- Little’s formula
- Exponential distribution
- Poisson process
- Markov processes
- Birth-death processes
- M/M/1 queue

Single-server queue M/M/1

- Customers arrive according to a Poisson process at rate λ
  - $1/\lambda$ = average inter-arrival time
  - IID inter-arrival times obeying the exponential distribution with intensity λ
- Customers are served by 1 server
- When busy, the server serves at rate μ
  - $1/\mu$ = average service time
  - IID service times obeying the exponential distribution with intensity μ
- There is an infinite number of customer places



Service discipline

- **Definition**:
  Service discipline determines the way the customers are served
  - It specifies whether the customers are served one-by-one or simultaneously
  - If the customers are served one-by-one, it specifies the order in which they are taken to service
  - If the customers are served simultaneously, it specifies how the service capacity is shared among them

- Service discipline is also called as queueing discipline, or scheduling discipline

- **Definition**:
  A service discipline is work-conserving if customers are served whenever the system is non-empty

Work-conserving service disciplines

- First In First Out (FIFO)
  - customers are served one-by-one until completion
  - service in the arrival order (“ordinary queue”)
  - the customer that arrived first is served with rate μ
  - also known as First Come First Served (FCFS)
- Processor Sharing (PS)
  - customers are served simultaneously
  - the service capacity is shared evenly among all customers (“fair queue”)
  - when i customers are in the system, each of them is served with rate μ/i
  - ideal version of the Round Robin (RR) service discipline

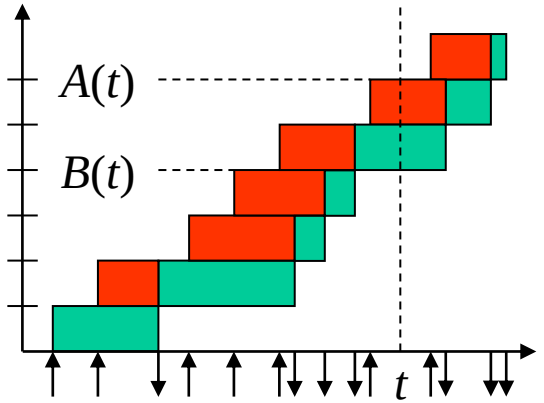
- **Definition**:
  A service discipline is non-anticipating
  if the service decisions are based only on the history of the system (e.g., on the attained service times, but not on the remaining service times)

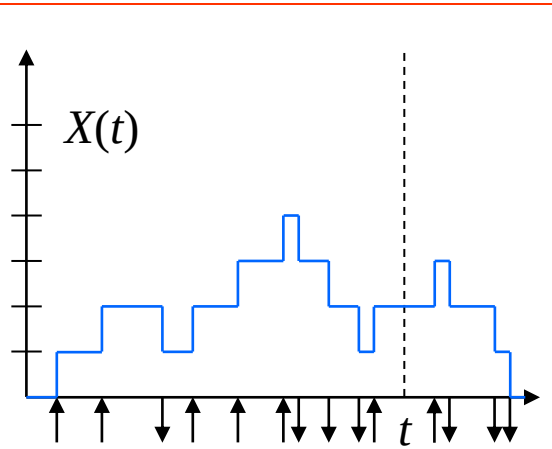
Queue length process

- $A(t)$ = the number of arrivals until time t
  = arrival process
- $B(t)$ = the number of departures until time t
  = departure process
- $X(t)$ = the number of customers at time t
  = queue length process

$$X(t) = A(t) - B(t)$$

$$X := X(\infty) := \lim_{t \to \infty} X(t)$$





Related random variables

- X = the number of customers in the system in equilibrium
  = queue length in equilibrium
- X^* = the number of customers in the system at a typical arrival time
  = queue length seen by an arriving customer
- W = the waiting time of a typical customer
- S = the service time of a typical customer
- $T = W + S$ = the total time in the system of a typical customer = delay

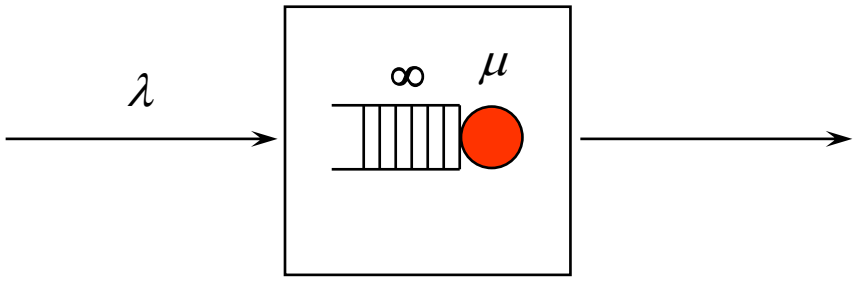
Traffic load in M/M/1

- **Definition**:
  Traffic load ρ is defined by

$$\rho := \frac{\lambda}{\mu}$$

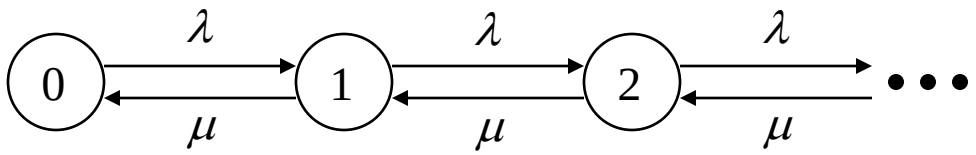
- Assume $\rho < 1$. By applying Little's formula to the subsystem consisting of just the server, we get

$$P\{X > 0\} = E[1_{\{X>0\}}] \overset{\text{Little}}{=} \lambda E[S] = \frac{\lambda}{\mu} = \rho, \quad P\{X = 0\} = 1 - \rho$$



State transition diagram

- Let $X(t)$ denote the number of customers in the system at time t
  - Assume that $X(t) = i$ at some time t, and consider what happens during a short time interval $(t, t+h]$
  - With probability $\lambda h + o(h)$, a new customer arrives (state transition $i \to i+1$)
  - If $i > 0$, then, with probability $\mu h + o(h)$, a customer leaves the system (state transition $i \to i-1$)
- Process $X(t)$ is clearly a Markov process with transition diagram



- Note: $X(t)$ is an irreducible birth-death process with an infinite state space $S = \{0,1,2,...\}$

Equilibrium distribution (1)

- Detailed balance equations (DBE):

$$\pi_i \lambda = \pi_{i+1} \mu \qquad \text{(DBE)}$$

$$\Rightarrow \quad \pi_{i+1} = \frac{\lambda}{\mu} \pi_i = \rho \pi_i$$

$$\Rightarrow \quad \pi_i = \rho^i \pi_0, \quad i = 0,1,2,\ldots$$

- Normalizing condition (N):

$$\sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \rho^i = 1 \qquad \text{(N)}$$

$$\Rightarrow \quad \pi_0 = \left(\sum_{i=0}^{\infty} \rho^i \right)^{-1} = \left(\frac{1}{1-\rho} \right)^{-1} = 1 - \rho, \quad \text{if } \rho := \frac{\lambda}{\mu} < 1$$

Equilibrium distribution (2)

- Thus, for a stable system ($\rho < 1$), the unique equilibrium distribution for the queue length X is a geometric distribution:

$$\rho < 1 \;\Rightarrow\; X \sim \mathrm{Geom}(\rho)$$

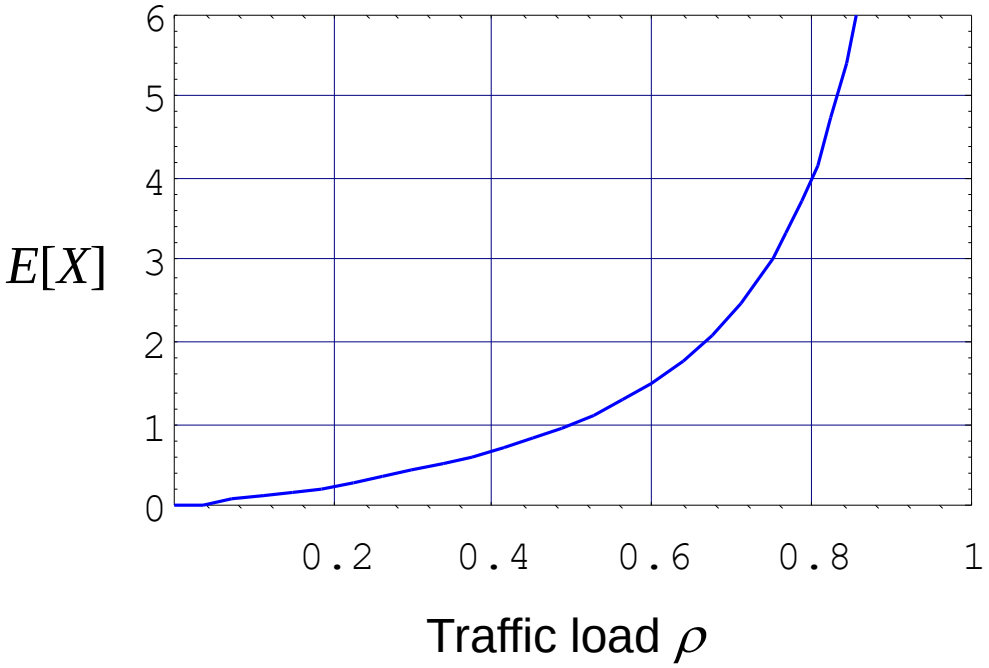
$$P\{X = i\} = (1-\rho)\rho^i, \quad i = 0,1,2,\ldots$$

$$E[X] = \frac{\rho}{1-\rho}, \quad D^2[X] = \frac{\rho}{(1-\rho)^2}$$

- This result is valid for any work-conserving and non-anticipating discipline (incl. FIFO and PS)

- Note on insensitivity:
  - For FIFO, this result is not insensitive to the service time distribution
  - However, for any symmetric service discipline (such as PS) the result is, indeed, insensitive to the service time distribution

Mean queue length vs. traffic load



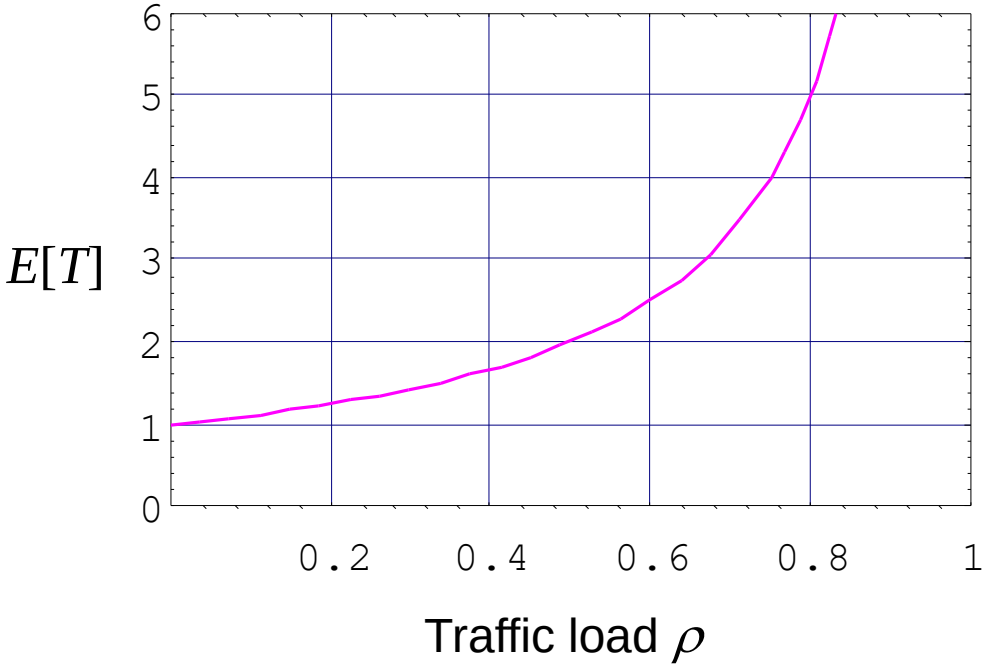
Mean delay for M/M/1

- Let $T = W + S$ denote the delay of a typical customer
  - including both the waiting time W and the service time S
- Applying Little' formula for a stable system ($\rho < 1$):

$$E[T] \overset{\text{Little}}{=} \frac{E[X]}{\lambda} = \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho} = \frac{E[S]}{1-\rho} = \frac{1}{\mu - \lambda}$$

- The mean delay $E[T]$ is the same for any work-conserving and non-anticipating discipline (incl. FIFO and PS)
- But the variance $D^2[T]$ and the delay distribution $P\{T \leq t\}$ depend on the discipline

Mean delay vs. traffic load



Mean waiting time for M/M/1

- Let W denote the waiting time of a typical customer
- Since $W = T - S$, we have for a stable system ($\rho < 1$):

$$E[W] = E[T] - E[S] = \frac{1}{\mu} \cdot \frac{1}{1-\rho} - \frac{1}{\mu} = \frac{1}{\mu} \cdot \frac{\rho}{1-\rho}$$

- The mean waiting time $E[W]$ is the same for any work-conserving and non-anticipating discipline (incl. FIFO and PS)
- But the variance $D^2[W]$ and the waiting time distribution $P\{W \leq t\}$ depend on the discipline

Summary

- Stochastic service system
  - service capacity, service discipline, traffic load, system performance
- Basic queueing models
  - Kendall's notation, M/M/1, M/G/1, M/G/n, M/G/∞
- Little's formula
- Exponential distribution
  - memoryless property
- Poisson process
  - IID exponential intervals, merging/splitting, PASTA
- Markov processes
  - exponential holding times, transition matrix, transition diagram, GBE, equilibrium distribution, irreducibility
- Birth-death processes
  - DBE
- M/M/1 queue