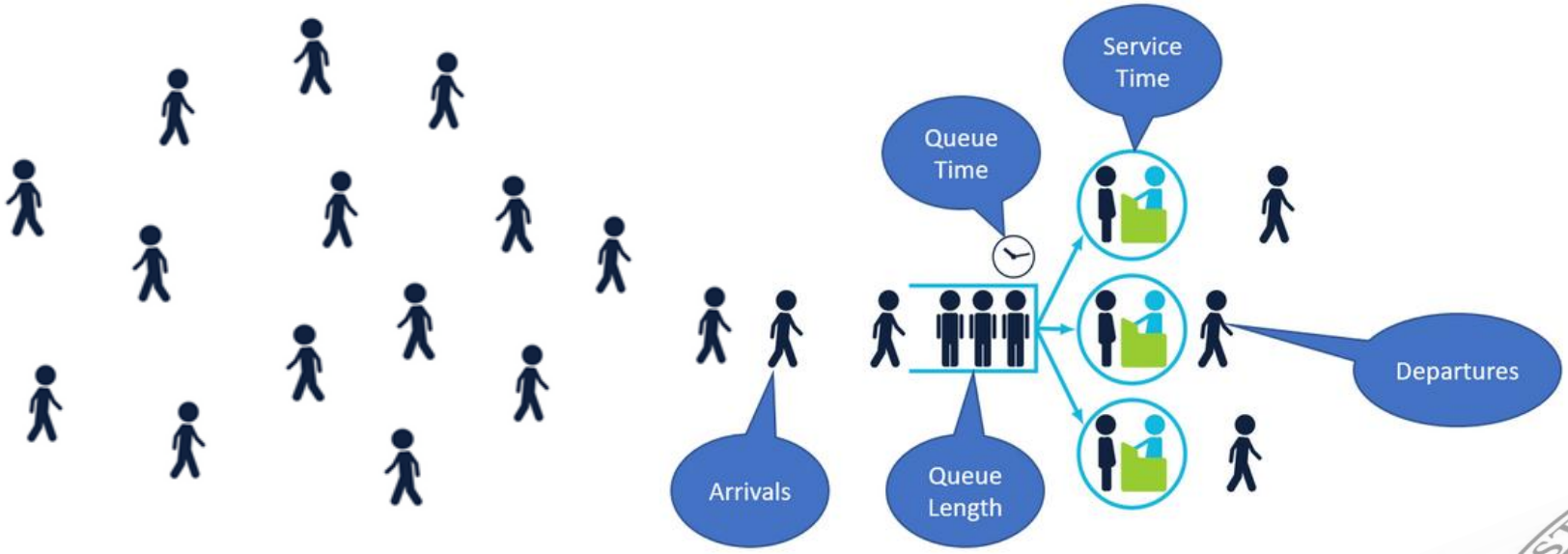
UNIVERSITY OF CAGLIARI

DIEE - Department of Electrical and Electronic Engineering

STOCHASTIC MODELS

-

Queueing theory

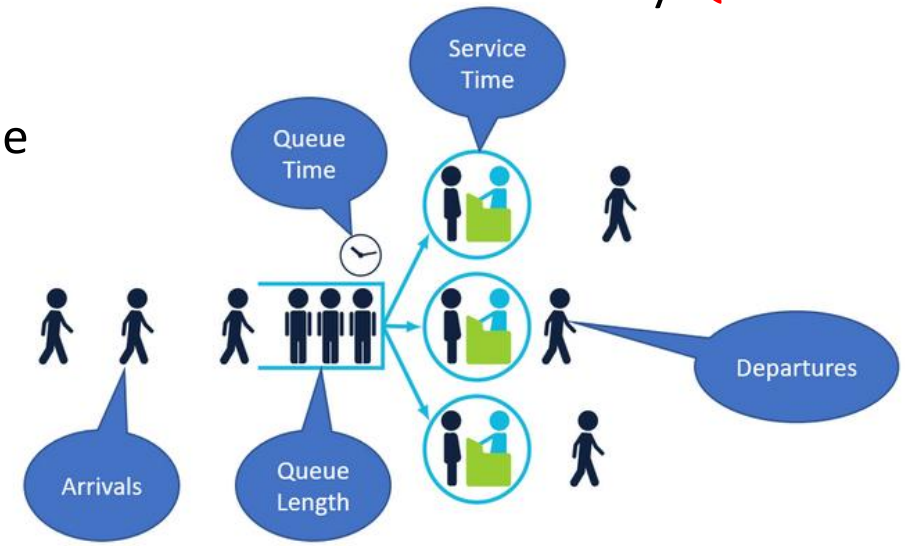


alessandro.pilloni@unica.it

Queueing theory

- Queueing theory is the mathematical study of waiting lines, or queues.
- It finds origin thanks to A. K. Erlang, who modeled the Copenhagen telephone exchange.
- Finds applications in many areas: ICT, traffic, industrial and manufacturing engineering, project management of shops/offices/hospitals, biology etc…
- Gives tool for the dimensioning/sizing/designing of service systems with shared resources (client-server systems) because it provides tools to evaluate many QoS metrics:

a) Mean time spent by costumers in the queue
b) Mean queue/buffer occupancy
c) Mean server utilization factor
d) Throughput (or productivity)
e) Etc.



Queues

- A queue, or "queueing node" is a complex object consisting of:

a) Many jobs/tasks or customers that arrive with a given arrival rate

b) One or more servers that, once free, can be paired each with a job

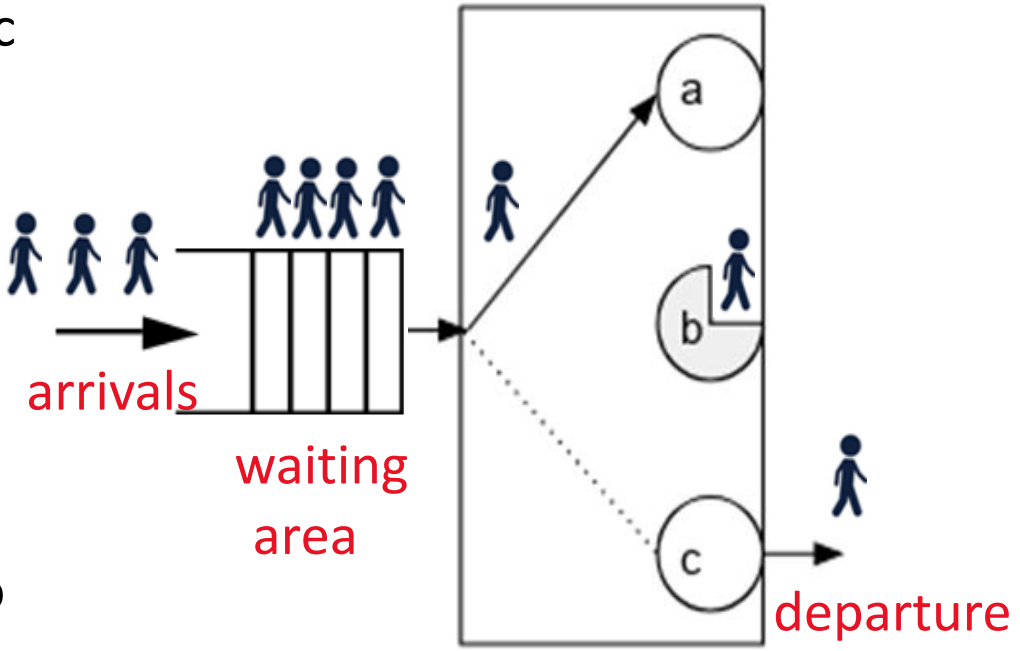
c) A waiting area/buffer where the taks wait their turn, according with a given policy, until a server is free

Example 1: A queue with 3 servers a, b, c

a. is "idle", thus an arrival is given to it to process

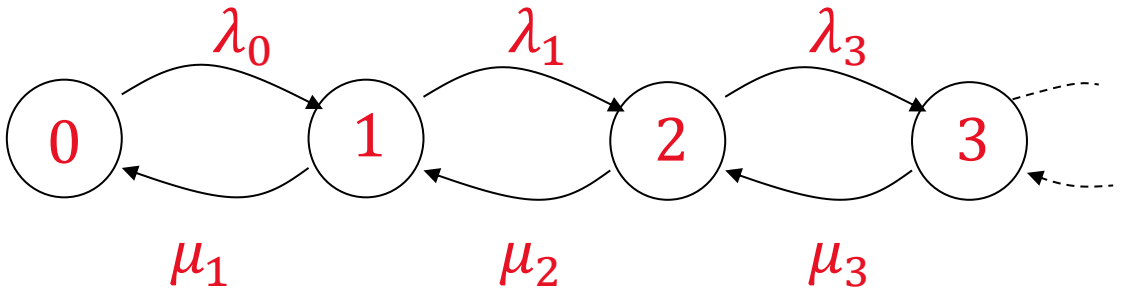
b. is "busy" and will take time before complete the service

c. completed a service and is "free" to receive the next job.



Queues and CT-BDP

- A single queue resource can be described by a CT-BDP, where the birth and death rates describe the arrivals and departures rates λ_i and μ_i, while the state of the CT-BDP indicates the number of costumers in the system.



$$\mathrm{E}[X_\infty] = \frac{\rho}{1-\rho}$$

$$\mathrm{Var}[X_\infty] = \frac{2\rho}{(1-\rho)^2}$$

- Remark: Often it is of interest the study the performance of a "queueing network", namely a composition of queues with a given "customer routing"

- Example 2: The Internet is a packet switching system. There many packets coming from different sources and with different destinations arrive to routers.
- In the router they wait up to be routed through their destination, or they may be lost, if the buffer is full (see traffic congestion).

Examples

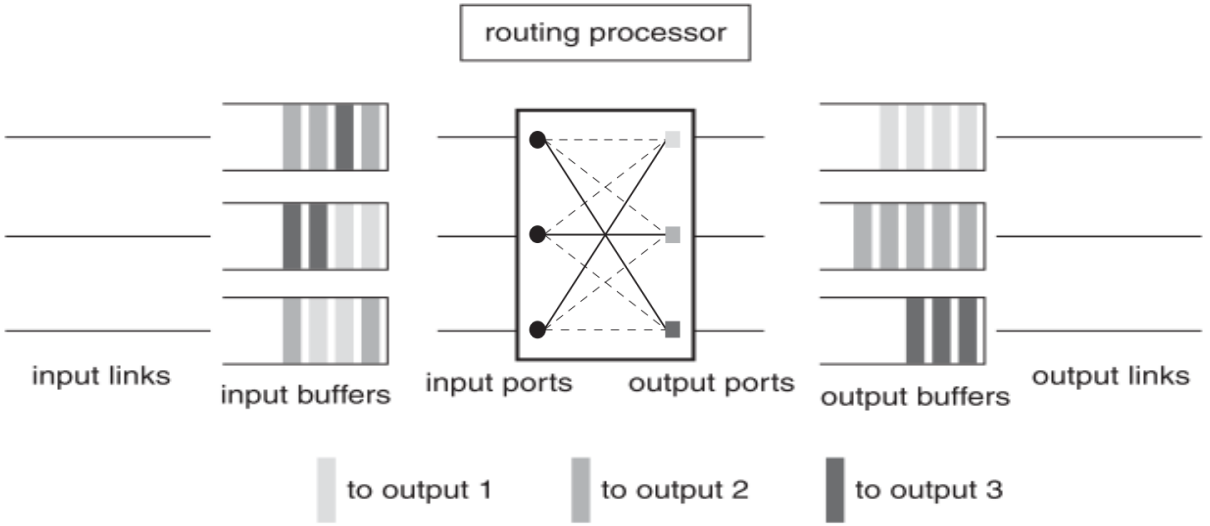
- Example of queuing network: **A router**. It consist of 4 major components:

a) Input ports

b) Output ports

c) Switch fabric

d) Routing processor



- Each I/O port maintains an I/O buffer
- The switch physically connect an input link to an output link
- The routing processor maintains the routing table and makes routing decisions

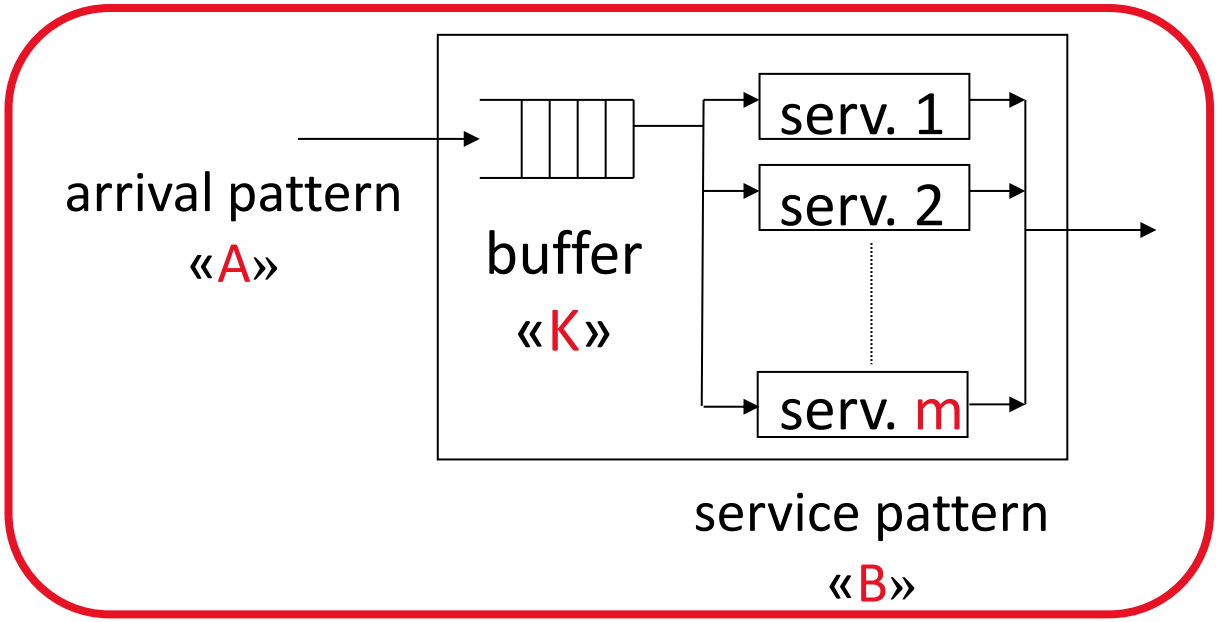
The performance of most of the ICT devices can be evaluated by mean of **Queues** and **Network of Queues**

Characteristics of queues and the Kendall Notation

- A **queue** (or **node**) is usually described by using the Kendall Notation
- It provides a compact description by means of a 6 fields string

A / B / m / K / N / ω

- Each field describes a different characteristics of the queue
  - ✓ A : arrival pattern
  - ✓ B : service pattern
  - ✓ m : number of servers
  - ✓ K : buffer size
  - ✓ N : population size
  - ✓ ω : queue's discipline



Remark: The last 3 fields are ommitted in the case of K = N = ∞ and ω = FIFO.

A : arrival pattern

- The "arrival pattern" defines the way customers enter the resource (or queue)
- It describes the **arrival SP** $\{X_A(t_k), t_k \in [0, \infty)\}$ where

$$t_0 < t_1 < t_2 < \cdots < t_k$$

- It is often expressed in term of inter-event time between 2 subsequent arrivals

$$\Delta t_k = t_k - t_{k-1}$$

- The "inter-arrival time" cab either deterministic or stochastic
- In the Kendall notation $A \in \{D, M, G, E_k\}$, where:
  - ✓ D : deterministic, e.g., $\Delta t_k = 1$sec., $\forall\, k$
  - ✓ M : markovian, which implies $\Delta t_k \sim Exp(\lambda)$ (implying Poisson arrivals)
  - ✓ G : inter-arrival times with general distribution
  - ✓ E_k : erlang with k as the shape parameter
  - ✓ Etc..

B : service pattern

- The "service pattern" describes how services are scheduled/implemented as the time passes

- The "service time" can be either deterministic or stochastic

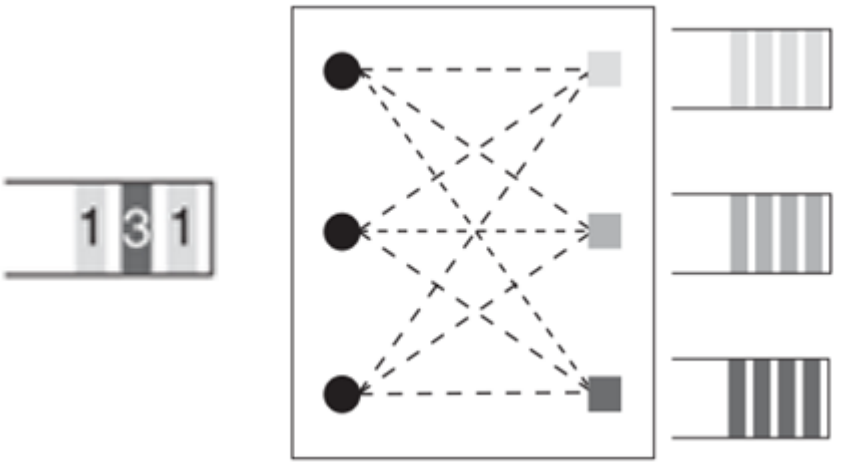
- If the service process is Poisson distributed, then the service time is exponential distributed, and the mean service time is denoted by

$$\mathrm{E}[\Delta t] = \frac{1}{\mu}$$ where μ is the mean service rate

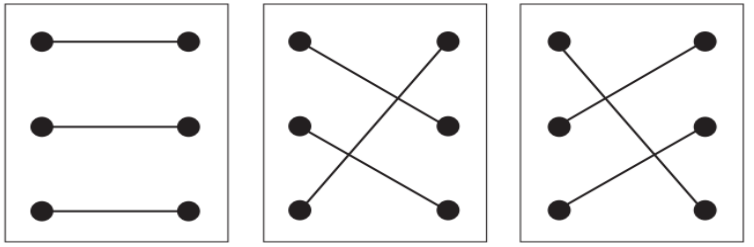
- In the Kendall notation it can be $A \in \{D, M, G, E_k\}$, where:
  - ✓ D : deterministic, e.g., $\Delta t_k = 1\text{sec.}, \forall\, k$
  - ✓ M : markovian, which implies $\Delta t_k \sim Exp(\mu)$
  - ✓ G : inter-arrival times with general distribution
  - ✓ E_k : erlang with k as the shape parameter
  - ✓ Etc.

m : number of servers

- The number of servers (or "channels") defines the maximun number of costumers (or packets) can be routed at the same time by the resource
- **In any case, each server can only serve one user at a time.**
- In the Kendall notation it can be
  - ✓ $m = 1$: single server
  - ✓ $m > 1$: multiple servers
  - ✓ $m = \infty$: infinite servers



- Example 6: Consider a switch with 1 Input and 3 Output ports
- Since up to 3 streams channels can be active simultaneusly, then $m = 3$.
- **Remark:** If the switch has more Input ports, each with its own buffer, its behaviour could more easily be modeled as a Queueing Network where each queue may have a specific queue discipline (see later).



Sequence of 3 matchings

K : buffer size

- The "buffer size" defines the max number of costumers can be accommodated in the waiting area.

- Thus, it accounts only the capacity of the waiting room.

- In the Kendall notation it can be:

  ✓ $K \in \mathbb{N}$: finite capacity

  ✓ $K = \infty$: infinite capacity

- **Remark 1:** The term "queue" or "queuing node" is refer to the entire resource (or node) while with "buffer" we refers only to the waiting area
- **Remark 2:** At most $K + m$ costumers can be in the resource at the same time.
- **IMPORTANT:** Some books denotes K as the max number of costumers in the resource, yielding $K = m + k$ where k is the buffer size, and m the servers number.

- **Remark 3:** If the buffer is full and a job arrives that **job** is **discharged**

N : population size

- The population size defines the **max number** of **potential costumers**
- **Costumers** can be seen as exogenous entities **"living outside"** the resource

- **Note:** Clearly a small population significantly affect the arrival rate

- Indeed, if N is finite and small, and most of the jobs are queued now, then fewer arrivals would be expected next, (cf. later Closed VS Open Queuing Networks),

- In the Kendall notation it can be:
  - ✓ $N = \mathbb{N}^+$: limited population (see closed queuing networks)
  - ✓ $N = \infty$: unlimited population

- Remark 1: It is common consider the population sufficiently large w.r.t K

- Remark 2: If $N = \infty$, this information is often omitted in the Kendall Notation

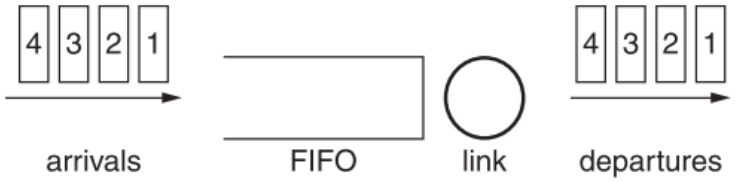
ω : queue's discipline

- The queue discipline or "service discipline" describe the under which priority order jobs in the waiting line are served
- In the Kendall notation it can be:

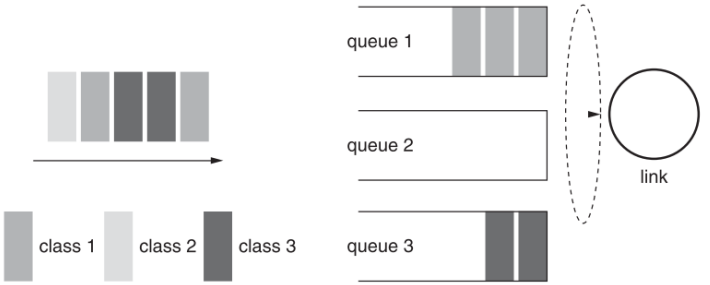
Symbol	Name	Description
FIFO	"First In First Out"	Jobs are served in the order they arrived in (if omitted is FIFO by default).
LIFO	"Last in First Out"	Jobs are served in the reverse order to the order they arrived in.
SIRO	"Service In Random Order"	Jobs are served in a random order with no regard to arrival order.
PQ	"Priority service"	**Many options:** Preemptive Priority, Non Preemptive, Class Based Weighted Fair Queuing, Weighted Fair Queuing
PS	"Processor Sharing"	

Example: FIFO vs Weighted Fair Queueing

- The advantage of a FIFO w.r.t the Weighted Fair Queueing (WFQ) is the simplicity and the fact that in many cases it properly describe the system needs.
- It **works** well when the **jobs population** is sufficiently homogeneous.
- When jobs are not homogeneous the whole system performance may decrease in the FIFO policy due to bottleneck
- In the Weighted Fair Queueing discipline jobs are sorted into classes, and treated as a separate class and a separate queue is maintained for each class



FIFO



Weighted Fair Queueing

- Remark: A different policy than FIFO do not enjoy the results of CT-BDP, and thus they require an ad-hoc modelization by means of CT-MCs

Quantity of interest of a queueing node

- $x(t) \in X = \mathbb{N}^+$: the number of costumers in the queue at time t
- $\pi_i(t) \in [0,1]$: probability of having exactly i costumers in the queue at time t
- $\pi_0(t)$: the probability the system is empty time t, called *idle probability*
- $\bar{x}(t) \in \mathbb{R}_{\geq 0}$: Mean number of costumer in the queue at time t

$$\bar{x}(t) = \mathsf{E}[x(t)] = \sum_{i=0}^{\infty} i \cdot \pi_i(t) \qquad \bar{x}(t) = -\left.\frac{dM(z,t)}{dz}\right|_{z=1}$$

- $\bar{x}_b(t) \in \mathbb{R}_{\geq 0}$: mean number of buffered costumers in at time t

$$\bar{x}_b(t) = \mathsf{E}[x_b(t)] \quad \text{where} \quad x_b(t) = \begin{cases} x_b(t) = 0 & \text{if} \quad x(t) \leq m \\ x_b(t) = x(t) - m & \text{if} \quad x(t) > m \\ x_b(t) \leq K \end{cases}$$

Quantity of interest of a queueing node (cont'd)

- Let $\lambda \in \mathbb{R}_{\geq 0}$ be the mean arrival rate (mean #arrivals per time unit)

$$\Rightarrow \text{the mean inter-arrival time is } \frac{1}{\lambda}$$

- Let $\mu \in \mathbb{R}_{\geq 0}$ be the mean service rate (mean #services completed in a time unit)

$$\Rightarrow \text{the mean service time is } \frac{1}{\mu}$$

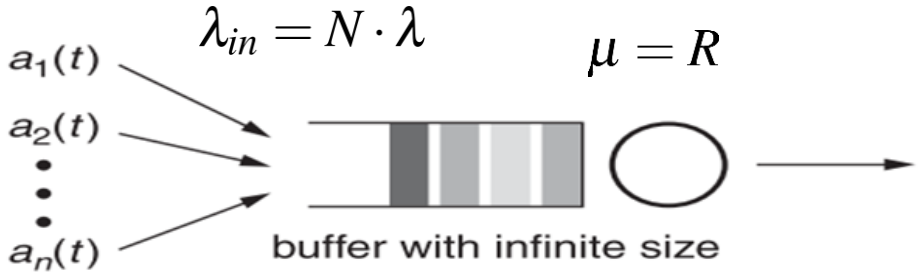
- $\bar{\theta} \in \mathbb{R}_{\geq 0}$: mean time spent in the queuing node

- $\bar{\theta}_b \in \mathbb{R}_{\geq 0}$: mean time spent in the waiting area

- … and others that we'll see later.

Traffic intensity

- Consider a data-link shared by N transmitters (sources)
- Each source injected a traffic $a_i(t)\ [bit/s]$ where $\mathrm{E}[a_i(t)] = \lambda$
- Assume the **channel bit-rate** is "**exactly**" $R[bit/s]$ **while** $\mathrm{E}[a_i(t)]N < R$
- Then due to the independence of $a_i(t)$ the **link** can be modelled by a M/D/1



$$\rho = \frac{N\lambda}{R} < 1$$

the traffic intensity is the ratio between offered and carried traffic

- **NOTE:** The number of buffered packets is $x_b(t) \in \mathbb{N}^+$ however this is a random variable because of $a_i(t)$ is random as well

$$x_b(t+dt) = \max\left\{0\ ,\ x_b(t) + \sum_{i=1}^{n} a_i(t) - R\right\}, \qquad x_b(0) \geq 0$$

The concept of "Traffic intensity"

- For a data-link the traffic intensity can be seen as the **ratio** between the **offered traffic** and the **carried traffic**, i.e.,

$$\rho = \frac{\lambda_{tot}}{\mu}$$

It is equivalent to the utilization factor of a M/M/1

$$\Pr(x(t) > 0) = 1 - \Pr(x(t) = 0) = 1 - (1 - \rho) = \rho$$

- It is measured in Erlangs (traffic units).

- Example 10: In a shared digital link, the traffic intensity can be computed as:

$$\rho = \frac{\lambda}{\mu} = \frac{a \cdot L}{R}$$

- ✓ $a = \sum_i E[a_i]$ [packet/s] is the mean arrival rate
- ✓ L [bit/packet] is the average packet length
- ✓ R [bit/s] is the transmission rate

Examples

- A wireless telephone exchange system can assign a channel for each VOIP call

  ⇒ thus the system can be modelled by M/M/∞

- Each user makes on average 1 call/hour whose average duration is 3 minutes

- Suppose there are N=100 users making call. Compute the **actual traffic intensity**

$$\lambda = \frac{100 \text{ call}}{\text{h}} = \frac{100 \text{ call}}{60 \text{ min}}$$

$$\frac{1}{\mu} = 3 \text{ min}$$

$$\rho = \lambda \cdot \frac{1}{\mu} = \frac{100 \text{ call}}{60 \text{ min}} \cdot 3 \text{ min} = 5 \text{ Erlang}$$

- Remark 1: For M/M/1 and M/M/∞ queues (see later) the quantity λ/μ takes meaning of traffic intensity
- Moreover, for M/M/∞, it corresponds to the **mean number** of **busy servers** $\bar{x}_s$ (*see later*)
- Remark 2: For a M/M/m the traffic intensity ρ is expressed "*per server*", i.e, $\rho = \frac{\lambda}{m\mu}$

The concept of "Traffic volume"

- In telecommunications, it is common refer also to the traffic volume, that is

$$V_T = \rho \times \Delta T \qquad \rho = \frac{V_T}{\Delta T}$$

- It is measured in erlang-hour (or erlang-minute, erlang-sec, etc…)
- It is a measure of the traffic processed during a time period.

- In telecommunication, service providers are vitally interested in traffic intensity and traffic volume, as it dictates the amount of equipment they must supply.

- Example: In the last 3hours are arrived 120calls, each on average of 3minutes, then the traffic volume is:

$$V_T = 120\text{call} \times 3\text{minute} = 360\ \text{Erlang-minute} = 6\ \text{Erlang-hour}$$

$$\rho = \frac{V_T}{\Delta T} = \frac{6\ \text{Erlang-hour}}{3\ \text{hour}} = 2\ \text{Erlang} \quad \equiv \quad \rho = \frac{\lambda}{\mu} = \frac{\frac{120\ \text{call}}{3\ \text{hours}}}{\frac{1\ \text{call}}{3\ \text{min}} \times \frac{60\ \text{min}}{1\ \text{hours}}} = 2$$

Little's Law

- The Little's Law is a result of the MIT instructor John Little which states that:

Little's Law

- Let λ be the long-term average arrival rate within the system
- Let $\bar{\theta}$ be the average time spent by a customer in the system.
- Let $\bar{x}$ be the the long-term average number of customers within the system
- The Little's Law says that

$$\bar{x} = \lambda \cdot \bar{\theta}$$

- Remark 1: **This result is very general**. It is **independent by either by the arrival**, and/or the **service process distribution**, and/or the **priority discipline**, or else!!

The Little Law holds for every G/G/.

- Remark 2: It can also be extended to **open networks** but **not** to **closed networks** because each queue cannot behave as an independent node with respect to the others (*see later*)

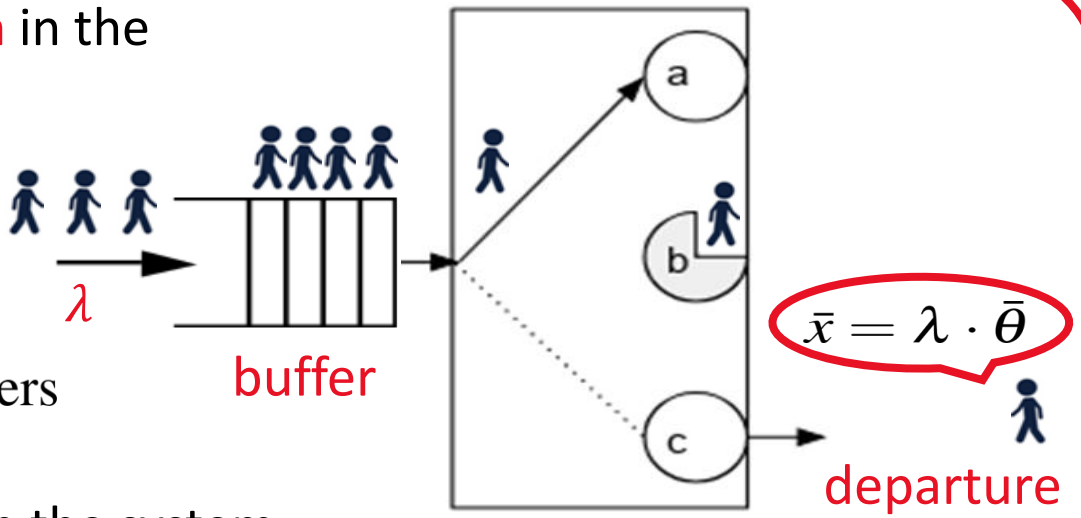
Intuitive explanation of the Little's Law

- Consider a queue at the system steady-state
- Consider a customer that is leaving the system after being served
- If that **costumer looks back**, s/he will **see on average**, a #customers $\bar{x}$ equal to the **#customers who arrived while** s/he **was in the system**, namely

$$\bar{x} = \lambda \cdot \bar{\theta}$$

- Example 9: A costumer spent 10 min in the G/G/3 queue before its departure.
- Let $\lambda = \frac{4\text{costumers}}{5\text{ min}}$

$$\bar{x} = \frac{4\text{ costumers}}{5\text{ min}} \cdot 10\text{ min} = 8\text{ costumers}$$



- This means on average 8 costumer in the system

- Moreover, it further results that:

$$\bar{x} = \lambda \cdot \bar{\theta} \quad \Longrightarrow \quad \bar{x}_s = \lambda \cdot \bar{\theta}_s = \lambda \cdot \frac{1}{\mu} \quad , \quad \bar{x}_b = \bar{x} - \bar{x}_b = \lambda \cdot \bar{\theta}_b$$

Deterministic queues

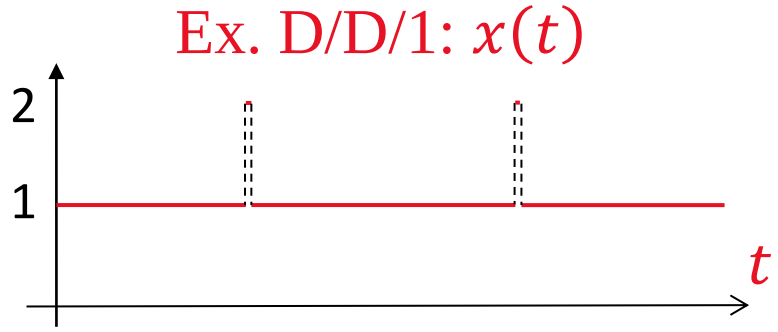
- Let the arrivals and service rate be deterministic and constant such that

$$\lambda(t) = \lambda \quad , \quad \mu(t) = \mu \qquad \forall\, t$$

- **Operating assumption:** If an arrival and a departure time coincide the departure occurs first.

- This is due to theoretical and practical reasons:

1. A server can host only 1 task at a time
2. For D/D/m, if m is large undesired and large instantaneous changes on $x(t)$ may occur



- Remark: In M/M/. queues (*see later*) due to the orderliness property that assumption is not necessary because **2 events cannot occur at the same time**

- To determine the long-term performance of a D/D/$\cdot$ we must analyze $x(t)$ until a stationary response is observed (if there exists)

D/D/1

- Consider a single server queue with deterministic arrivals and services

✓ If $\lambda > \mu$ the queue is called unstable $\rightarrow \lim_{t\to\infty} x(t) = \infty \Rightarrow \bar{x} \to \infty, \bar{\theta} \to \infty$

✓ If $\lambda < \mu$ the queue is called stable $\rightarrow \lim_{t\to\infty} x(t) = \bar{x} < \infty \Rightarrow \bar{\theta} = \frac{\bar{x}}{\lambda} < \infty$

✓ If $\lambda = \mu$, the departures' priority makes the queue stable as well (cf. with M/M/.)

- Let us now focus on the queue with $\lambda < \mu$ where:

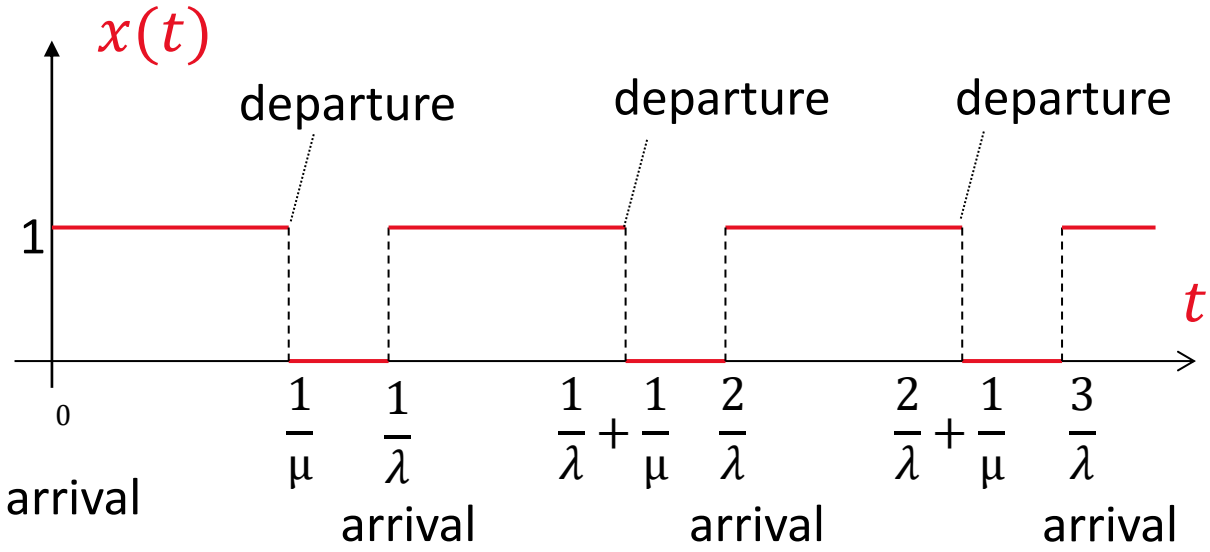
✓ 1st arrival at $t = 0$

✓ 1st departure at $t = \frac{1}{\mu}$

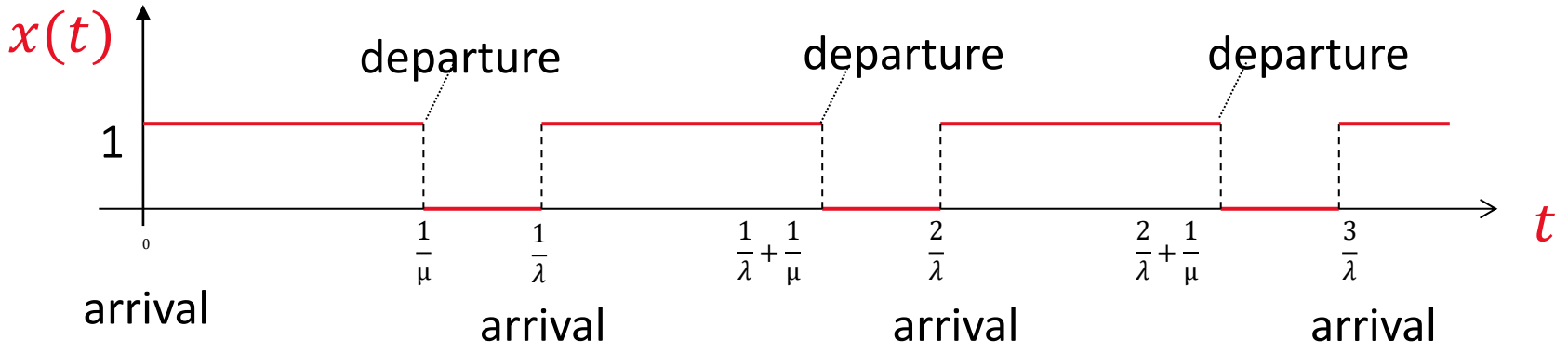
✓ 2nd arrival at $t = \frac{1}{\lambda}$

✓ 2nd departure at $t = \frac{1}{\lambda} + \frac{1}{\mu}$

✓ Etc…



D/D/1 (cont'd)



- **NOTE 1:** $x(t)$ exhibits a periodic steady-state response with period $T = \frac{1}{\lambda}$

- The proportion of time the resource is busy, called also utilization factor, is

Utilization factor

$0 < \hat{U} \leq 1$

$$\hat{U} = \frac{\frac{1}{\mu}}{\frac{1}{\lambda}} = \frac{\lambda}{\mu} = \rho$$

- **NOTE 2:** If $\lambda \leq \mu$ the **buffer is not necessary**, indeed it results that

Idle probability

$$\overbrace{\pi_0(\infty)} = 1 - \hat{U} \qquad \pi_1(\infty) = \frac{\frac{1}{\mu}}{\frac{1}{\lambda}} = \rho = \hat{U} \qquad \pi_j(\infty) = 0 \ \forall\, j \geq 2$$

- Thus it further results $\bar{x} = 0 \cdot \pi_0(\infty) + 1 \cdot \pi_1(\infty) = \rho \quad \equiv \quad \bar{x} = \lambda \cdot \bar{\theta} = \rho$

Little's Law

D/D/m

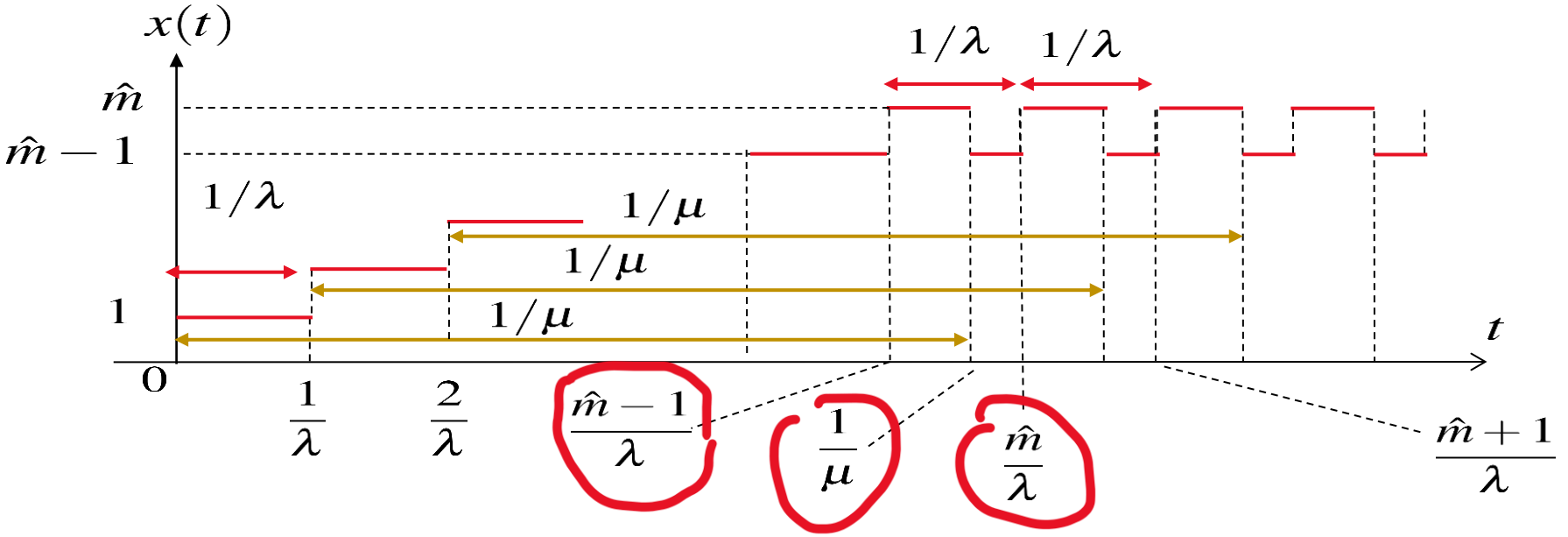
- In D/D/m queue both arrivals and services rate are deterministic

✓ If $\lambda > m\mu$ the queue is called unstable

✓ If $\lambda \leq m\mu$ the queue is called stable (due to the departure's priority)

NOTE: If $\lambda \leq m\mu$, $\exists\ \hat{m} \leq m$ such that a steady-state it results: $\frac{\hat{m}-1}{\lambda} < \frac{1}{\mu} \leq \frac{\hat{m}}{\lambda}$

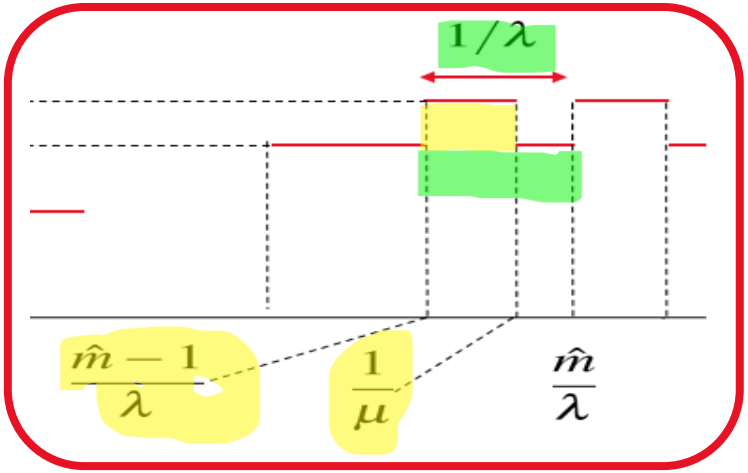
…graphically…



- **Remark:** A D/D/m allows to account m times slower servers than a D/D/1.

D/D/m (cont'd)

- The long-term evolution of the queue length $x(t)$ has a period $T = 1/\lambda$
- At the steady-state $x(\infty)$ may be either $\hat{m} - 1$ or $\hat{m}$ with probabilities:



$$\pi_{s,\hat{m}} = \Pr(x(\infty) = \hat{m}) = \frac{\frac{1}{\mu} - \frac{\hat{m}-1}{\lambda}}{\frac{1}{\lambda}} = \frac{\lambda}{\mu} - (\hat{m} - 1)$$

$$\pi_{s,\hat{m}-1} = \Pr(x(\infty) = \hat{m} - 1) = 1 - \pi_{s,\hat{m}} = \hat{m} - \frac{\lambda}{\mu}$$

$$\pi_j(\infty) = 0 \;\; \forall \; j \neq \{\hat{m} - 1, \hat{m}\}$$

- From which, one obtain:

$$\bar{x} = \sum_{i=0}^{\infty} i \cdot \pi_{s,i} = (\hat{m} - 1) \cdot \pi_{s,\hat{m}-1} + \hat{m} \cdot \pi_{s,\hat{m}} = \frac{\lambda}{\mu}$$

Equivalently by the Little's Law $\bar{x} = \lambda \cdot \bar{\theta} = \lambda \cdot \frac{1}{\mu}$

D/D/m (cont'd)

- Finally, the single server utilization, namely, the proportion of time a generic server is working (or equivalently the probability that a server is busy) is

Single server utilization: $\tilde{\rho} = \frac{\frac{1}{m \cdot \mu}}{\frac{1}{\lambda}} = \frac{\lambda}{m \cdot \mu}$

Mean number of busy servers: $\bar{x}_s = (\hat{m} - 1) \cdot \pi_{\hat{m}-1} + \hat{m} \cdot \pi_{\hat{m}} = \frac{\lambda}{\mu} \equiv m \cdot \tilde{\rho}$

D/D/∞

- They are used to model systems where $\bar{\theta} \approx \bar{\theta}_s = \frac{1}{\mu}$ independently by λ
- By mean of the Little's Law one has that $\bar{x} = \lambda \cdot \bar{\theta} = \lambda / \mu$
- Because there are ∞ servers, while arrivals are finite the number of busy servers is finite as in a M/M/m, and equal to $\bar{x}$, so the single-server utilization is zero.

Single server utilization of a D/D/∞: $\tilde{\rho} = \frac{\frac{1}{\infty \cdot \mu}}{\frac{1}{\lambda}} = \frac{\lambda}{\infty \cdot \mu} = 0$

There is always an infinity of servers free

Markovian queues

- In most of practical applications, it is of interest considering queues which arrivals and services are not deterministic but random

- If arrivals (and services) are independent, these system cancan be modelled by Markovian queues

- The following queues are of particular interest

  - ✓ M/M/1 (classic single server resource)
  - ✓ M/M/1/K (with finite buffer)
  - ✓ M/M/m (with m servers)
  - ✓ M/M/∞ (with ∞ servers)
  - ✓ M/M/1 with discouraged arrivals

- In the following the conditions under which these queue are ergodic will be discussed, and their main QoS metrics will be introduced.

PASTA (Poisson Arrivals See Time Averages) Property

- PASTA Property: From the point of view of a job/costumer outside the queue, let's call it "*Poisson arrival*", the probability to find n costumers at time t_k in the system **equals** the steady-state provability $\pi_n(\infty)$, i.e.,

$$\pi_n(t_k) = \Pr(x(t_k) = n) \equiv \pi_n(\infty) \quad \forall\ t \geq 0$$

- **Interpretation:** Upon arrival at a station, a costumer observes the system as if in steady state at an arbitrary instant for the system without that job.
- The expectation of every queue's parameter at time t_k equals its corresponding steady-state expected value, i.e.,

✓ $E[x(t_k)] \equiv E[x(\infty)] = \bar{x} = \lambda \cdot \bar{\theta}$ (mean number of customers at t_k)

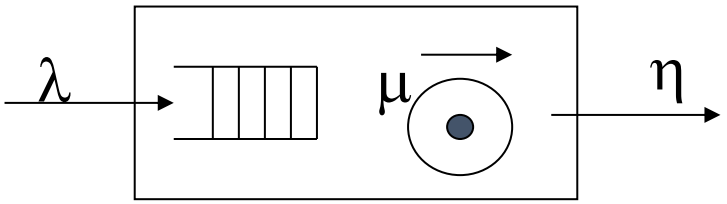
✓ $\Pr(x(t_k) \geq 1) \equiv \sum_{i=1}^{\infty} \pi_i(\infty) = \hat{U}$ (utilization factor at t_k)

✓ $\pi_0(t_k) \equiv \pi_0(\infty) = 1 - \hat{U}$ (idle probability at t_k)

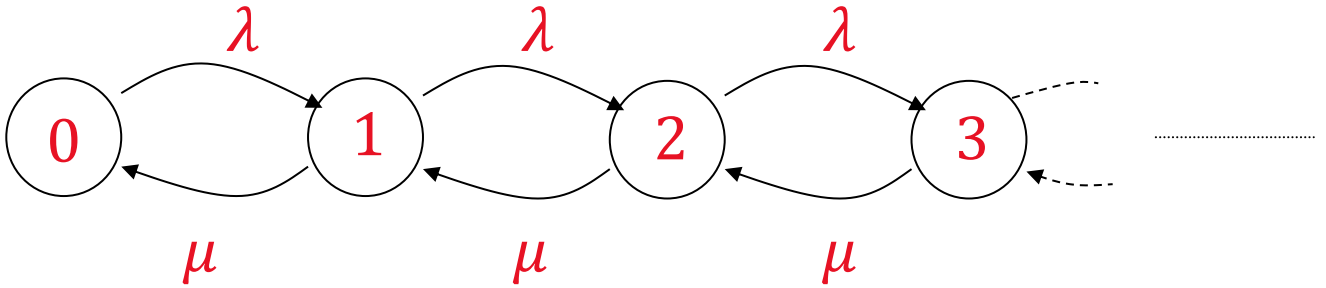
✓ $E[x_b(t_k)] \equiv E[x_b(\infty)] = \bar{x} - \bar{x}_s$ (mean number of buffered costumers)

M/M/1

- A M/M/1 describes a queue with a single server, Poisson arrivals with rate λ and exponential service times with rate μ.



- A M/M/1 is described by a uniform time-homogeneous CT-BDP



- If λ and μ are constant and state-independent this MC is aperiodic and irreducible
- However, to be be ergodic, it is required that

$$\rho = \frac{\lambda}{\mu} < 1 \Rightarrow \begin{cases} \Pi_s \cdot Q = 0 \\ \sum_i \pi_{s,i} = 1 \end{cases} \Rightarrow \pi_{s,i} \equiv \pi_i(\infty) = \rho^i \cdot (1-\rho) \;\; \forall\, i \geq 0$$

M/M/1 : Quantities of interest

- From

$$\pi_i(\infty) = \rho^i \cdot (1-\rho) \quad \forall\, i = 0,1,2,\ldots \qquad\qquad \rho = \frac{\lambda}{\mu} < 1$$

- The following quantities of interest can be determined

- Probability to find i costumers in the system is

$$\pi_i(\infty) = \rho^i \cdot (1-\rho) \quad \forall\, i \geq 1\ldots$$

- Idle probability, namely, the probability the system is empty

$$\pi_0(\infty) = (1-\rho)$$

- Utilization factor of the resource (and in this case of the server as well), i.e. , the average occupancy of a resource during a specified time-period is

$$\hat{U} \;=\; \Pr(\, x(\infty) \geq 1\,) \;=\; 1 - \pi_0(\infty) = \rho$$

M/M/1 : Quantities of interest (cont'd)

- **Burke's theorem:** If a queue is **ergodic** and **arrivals** are **Poisson with rate λ**, then at steady state the **departures are Poisson** with **the same rate λ**.

- Throughput: From the Burke's theorem, the throughput is λ, but it can be derived also as

$\eta =$ **utilization factor** $\times$ **service rate** $\qquad \eta = \hat{U} \cdot \mu = \frac{\lambda}{\mu} \cdot \mu = \lambda < \mu$

- Mean service time: Since the service rate is Poisson with rate μ, the service time $\theta_s \sim Exp(\mu)$. Thus, its expectation is

$$\bar{\theta}_s = \mathrm{E}(\theta_s) = \frac{1}{\mu}$$

- Mean number of costumer in the resource at steady state

$$\bar{x} = \mathrm{E}[x(\infty)] = \frac{\rho}{1-\rho} \qquad \text{(note that if } \rho \to 1 \Rightarrow \bar{x} \to \infty\text{)}$$

See module of CT-BDP.

M/M/1 : Quantities of interest (cont'd)

- Mean time spent in the system by a customer is

(From the Little's Law) $$\bar{\theta} = \frac{\bar{x}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}$$

- Mean number of busy servers, which coincides with the utilization factor, is

(From the Little's Law) $$\bar{x}_s = \lambda \cdot \theta_s = \lambda \cdot \frac{1}{\mu} = \rho$$

- Mean number of buffered costumers is:

$$\bar{x}_b = \bar{x} - \bar{x}_s = \bar{x} - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$

- Mean time spent in the buffer by a costumer

(From the Little's Law) $$\bar{\theta}_b = \frac{\bar{x}_b}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu-\lambda}$$

- Remark: As the traffic intensity $\rho \to 1$, then $\bar{x}$, and thus also $\bar{\theta}$ and $\bar{\theta}_b \to \infty$.

M/M/1 Dimensioning – Problem 1 "find the min service time"

- Recalling that $\bar{\theta} = \frac{1}{\mu - \lambda}$
- A **QoS specification** may refers, e.g., to a **percentile** of the **delay distribution $\bar{\theta}$**.
- Example 14: Requires that no more than $\alpha_{\%} = 1\%$ of packets will experience a delay greater than $\Delta t = 100\text{msec}$
- Problem 1: Determine the minimal $\mu^* : \Pr(\theta > \Delta t) \leq \alpha$
- Since arrival and services are assumed Poissonian then also $\theta \sim Exp\left(\frac{1}{\mu - \lambda}\right)$

$$\Pr(\theta \leq \Delta t) = 1 - e^{-(\mu^* - \lambda)\Delta t}$$

- For any $\alpha \in [0,1]$, and $\Delta t = 100\text{msec} > 0$

$$\Pr(\theta > \Delta t) = e^{-(\mu^* - \lambda)\Delta t} \leq \alpha$$

Solution: $\mu^* = \lambda - \frac{\ln(\alpha)}{\Delta t}$

M/M/1 Dimensioning – Problem 2 "find the max arrival rate tollerated"

- Alternatively, the dimensioning problem can be expressed in term of the arrival's mean rates
- Problem 2: Let $\Delta t = 100\text{msec} > 0$ and $\alpha_{\%} = 1\%$.

- Determine the maximal λ^* : $\Pr(\theta > \Delta t) < \alpha$

$$\Pr(\ \theta > \Delta t\ ) = \mathrm{e}^{-(\mu-\lambda^*)\Delta t} < \alpha$$

Solution: $\lambda^* = \mu + \dfrac{\ln(\alpha)}{\Delta t} < \mu$ indeed $\alpha \in (0,1) \Rightarrow \dfrac{\ln(\alpha)}{\Delta t} < 0$

- Remark: The solution λ^* to be feasible must also satisfy that

$$0 < \lambda^* = \mu + \underbrace{\frac{\ln(\alpha)}{\Delta t}}_{<0 \text{ beacuse } \alpha \in (0,1)} < \mu \quad \Rightarrow \quad \mu > -\frac{\ln(\alpha)}{\Delta t}$$

Multiplexing on M/M/1 resourses

- **Multiplexing techniques** are fundamental in many fields
- For instance they allow reducing latency of multiple traffic streams sharing a common telecommunication channel
- Operating assumption: **Packets** are **generated** in **a Poisson fashion**, implying the **packet length** is **exp. distributed**. **Example:** ETH payload is between 46 e 1500 byte
- If so, and because we are considering a shared communication channel, then the telecommunication system can be modelled as a M/M/1, with a given ρ, where

$$\pi_i(\infty) = \rho^i(1-\rho)$$

- while the delay statistics $\bar{\theta}$, and $\bar{\theta}_b$ depends by the inverse of $(\mu - \lambda)$, e.g.,

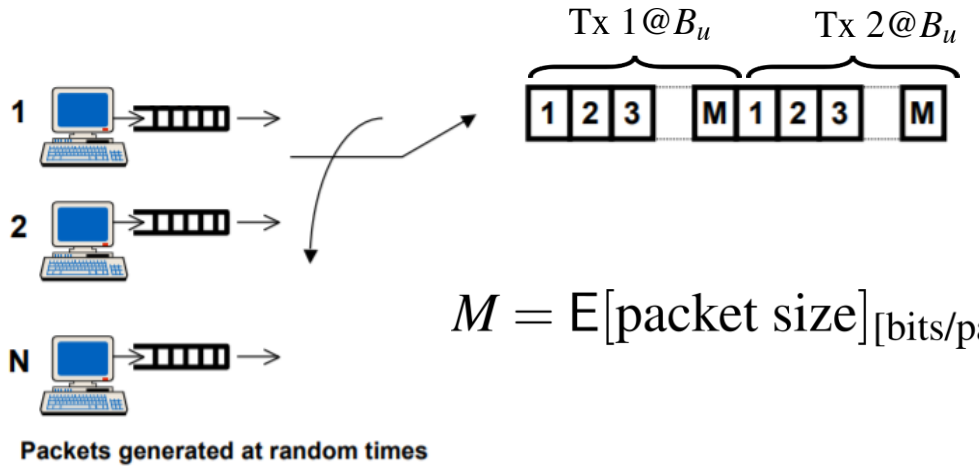
$$\bar{\theta} = \frac{\bar{x}}{\lambda} = \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}$$

- **MAIN IDEA:** If the **arrival** and **service rates** increase by N, it results ρ is unchanged (same traffic intensity) but the delay statistics improves a lot (smaller delay)

$$\bar{\theta} = \frac{1}{N \cdot (\mu - \lambda)}$$

Multiplexing: Time division multiple access (TDMA)

- In the TDMA each user has **one** or **more time-slots assigned** to **transmit**
- If a **source** has **no traffic**, these **time-slots** are **filled** with **empty packets**
- Let Ru[bit/sec] be the **source's demand bit rate**, and M the **mean packet-size**
- In **TDMA** each source transmits only in their **time-slots** with a **bit rate Bu [bit/sec]**, **meanwhile** the **remaining packets are buffered in their local buffers**



$$\lambda \left[\frac{\text{packet}}{\text{s}}\right] = \frac{R_u \left[\frac{\text{bits}}{\text{s}}\right]}{M[\text{bits/packet}]}$$

$$M = \text{E}[\text{packet size}]_{[\text{bits/packet}]}$$

$$\mu \left[\frac{\text{packet}}{\text{s}}\right] = \frac{B_u \left[\frac{\text{bits}}{\text{s}}\right]}{M[\text{bits/packet}]}$$

- Although the total demand is $N\lambda$, since **each source** transmits only within its **time-slot**, the system behaves as N many **M/M/1** in parallel, properly scheduled. Thus

$$\rho_{TDMA} = \frac{\lambda}{\mu} = \frac{R_u}{B_u} \Rightarrow \bar{x}_{TDMA} = \frac{\rho}{1-\rho} \qquad \bar{\theta}_{TDMA} = \frac{1}{\mu - \lambda}$$

Multiplexing: Full multiplexing (FMUX)

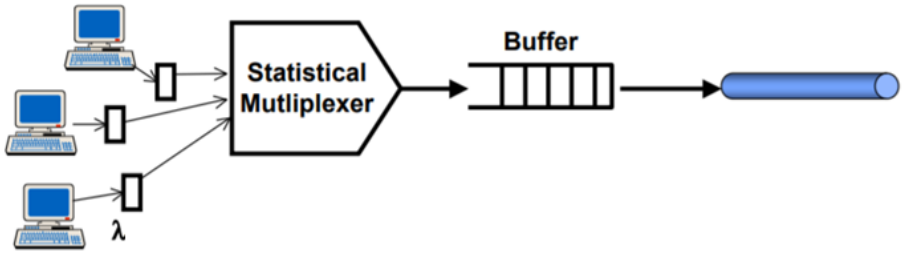
- Consider a FMUX access method (known also as statistical multiplexing, e.g. CSMA)
- Sources send their traffic to a **statistical multiplexer** stacking all the packets in a **common buffer**, in a **random way**, to avoid source's starvation.

$$\Rightarrow \text{total arrival rate is } N\lambda$$

- The transmitter forwards the packets to the destination with a N times faster rate, i.e. $N\mu$.

$$\rho_{FMUX} = \rho_{TDMA} \quad \rightarrow \quad \pi_i^{FMUX} = \pi_i^{TDMA}$$

- The main difference is that the traffic of all the sources is stored in 1 shared buffer



Interpretation: In place of n slow queues there is 1 n times faster queue

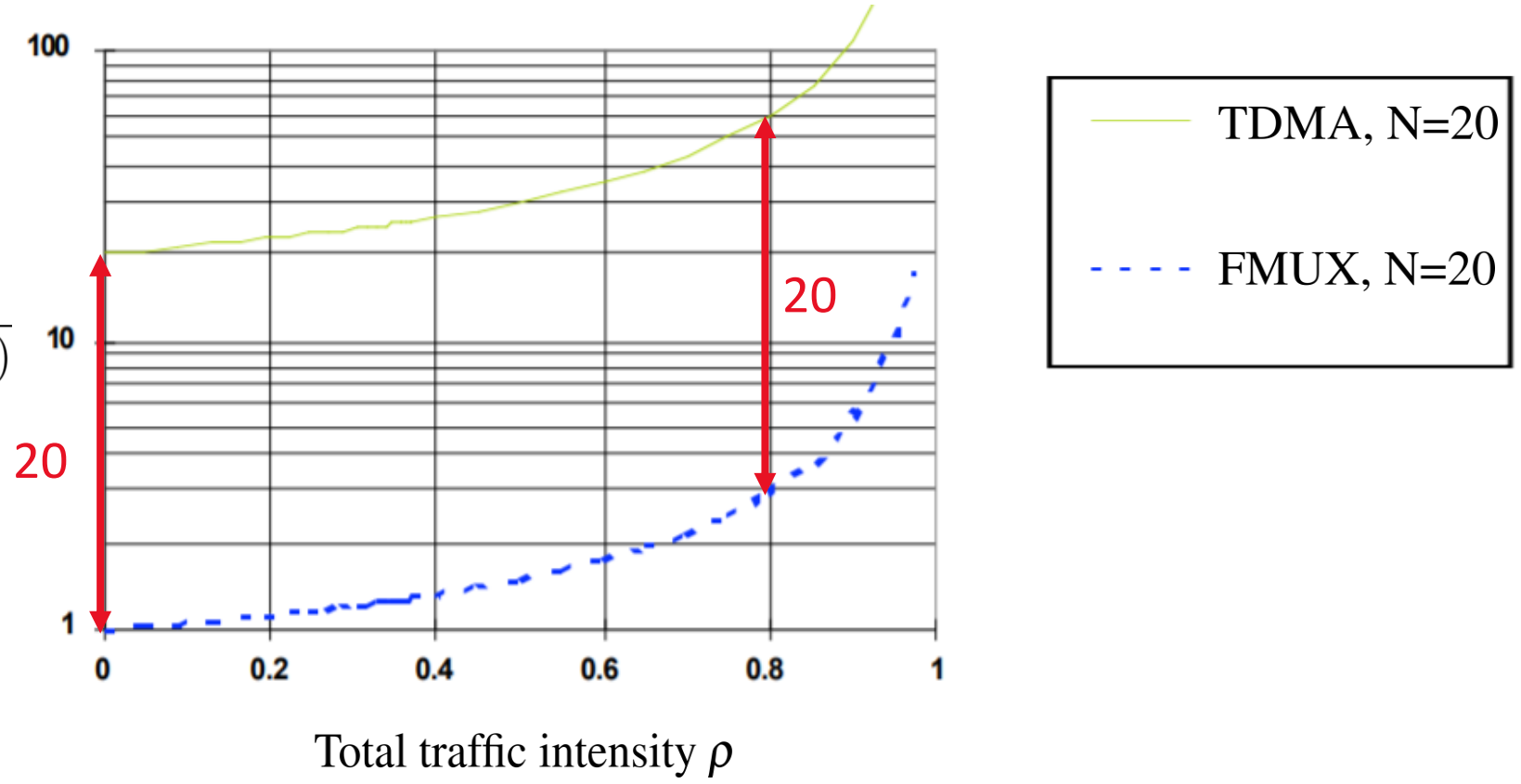
- Although TDMA and FMUX has the same ρ, the FMUX has much better latency

$$\rho_{FMUX} = \frac{N\lambda}{N\mu} = \rho_{TDMA} \quad \Rightarrow \quad \bar{x}_{FMUX} = \frac{\rho}{1-\rho} \qquad \bar{\theta}_{FMUX} = \frac{1}{N(\mu-\lambda)} \ll \bar{\theta}_{TDMA}$$

Multiplexing comparison

Average Packet Service Time: TDMA vs FMUX

$$\bar{\theta} = \frac{1}{N(1-\rho)}$$



Multiplexing dimensioning – Problem

- Problem: A service provider want to upgrade its telecommunication infrastructure from TDMA to a statistical access mechanism.
- Let N be the number of sources for the TDMA
- Suppose the latency QoS for the current scheme is $\bar{\theta}^* = \frac{1}{\mu - \lambda}$
- **Allocate the optimal service rate for the server**

$$\mu^* \ : \ \bar{\theta} = \bar{\theta}^*$$

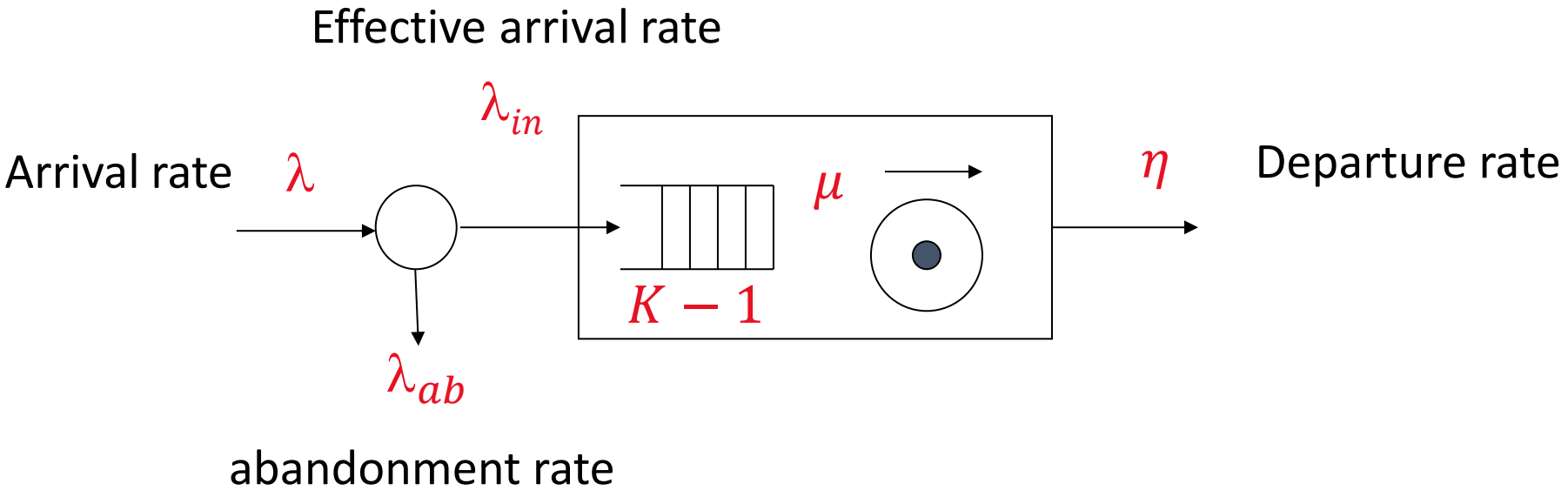
- In the case of FMUX, both λ and μ increase by N thus $\bar{\theta}^{FMUX} = \frac{1}{N\mu - N\lambda} \ll \bar{\theta}^*$

- However, to satisfy the given specification $\bar{\theta}^*$ it suffices find

$$\mu^* \ : \ \bar{\theta} = \frac{1}{\mu^* - N \cdot \lambda} = \frac{1}{\mu - \lambda} \quad \Rightarrow \quad \mu^* - N\lambda = \mu - \lambda$$

$$\Rightarrow \quad \mu^* = \mu + (N-1)\lambda \ \leq \ \overbrace{N \cdot \mu}^{\mu_{FMUX}}$$

M/M/1/K

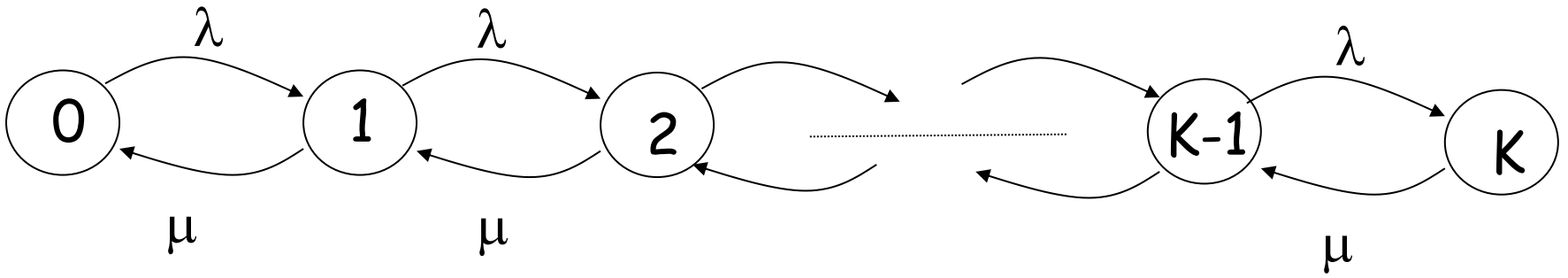


- Arrivals and services follows a Poisson process with a rate λ and μ respectively
- The M/M/1/K is a single server resource with a limited buffer

  Since the buffer has finite lenght, some arrivals may be lost due to overflow

- Remark: In the Kendall notation K often denotes the buffer size only.
- Here K denotes the overall number of costumer in the system, thus the buffer size is $K-1$.

M/M/1/K (cont'd)



- The M/M/1/K can be modelled by a finite-state CT-BDP

a) the birth rate is state dependent and decreases as

$$\lambda_i = \begin{cases} \lambda & \text{if} \quad i < K \\ 0 & \text{if} \quad i \geq K \end{cases}$$

b) The death rate μ is constant

Since the CT-BDP has a finite number of states it is ergodic!

No more required that $\rho = \frac{\lambda}{\mu} < 1$

M/M/1/K (cont'd)

- Since the system is ergodic, we can evaluates its limiting distribution by solving the following linear system

$$\begin{cases} \Pi_s \cdot Q = 0 \\ \sum_i \pi_{i,s} = 1 \end{cases}$$

$$\begin{cases} \pi_{s,i+1} = \dfrac{\lambda_i}{\mu_{i+1}} \pi_{s,i} \quad \forall\, i \geq 0 \qquad \rho = \dfrac{\lambda}{\mu} \\ \displaystyle\sum_{i=0}^{k} \pi_{s,i} = 1 \end{cases}$$

$$\begin{cases} \pi_{s,1} = \rho\, \pi_{s,0} \\ \pi_{s,2} = \rho\, \pi_{s,1} = \rho^2 \pi_{s,0} \\ \quad \vdots \\ \pi_{s,\mathrm{K}} = \rho\, \pi_{s,\mathrm{K}-1} = \rho^{\mathrm{K}} \pi_{s,0} \\ \\ \pi_{s,0}(1 + \rho + \rho^2 + \cdots + \rho^{\mathrm{K}}) = 1 \end{cases}$$

$$\pi_{s,0} \cdot \sum_{i=0}^{\mathrm{K}} \rho^i = \pi_{s,0} \cdot \frac{1 - \rho^{\mathrm{K}+1}}{1 - \rho} = 1$$

$$\pi_{s,0} = \frac{1 - \rho}{1 - \rho^{\mathrm{K}+1}}$$

$$\pi_{s,i} = \frac{\rho^i (1 - \rho)}{1 - \rho^{\mathrm{K}+1}} \qquad i \leq \mathrm{K}$$

M/M/1/K : Quantities of interest

- Probability of having exactly i costumers in the system is

$$\pi_i(\infty) = \frac{\rho^i(1-\rho)}{1-\rho^{K+1}} \qquad i \leq K \quad , \quad \forall\, \rho > 0$$

- Idle rate (or Idle probability), i.e, the probability the resource is empty

$$\pi_0(\infty) = \frac{(1-\rho)}{1-\rho^{K+1}}$$

- Blocking probability, due to the PASTA property, at any time t the probability a costumer is rejected "**blocked outside the system**", is

$$\pi_K(\infty) = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$$

- Abandonment rate at steady state

$$\lambda_{ab} = \lambda \cdot \pi_K(\infty) = \lambda \cdot \frac{\rho^K(1-\rho)}{1-\rho^{K+1}}$$

M/M/1/K : Quantities of interest

- Utilization factor, namely, the average occupancy of the resource

$$\hat{U} = 1 - \pi_0(\infty) = 1 - \frac{(1-\rho)}{1-\rho^{K+1}} = \frac{\rho(1-\rho^K)}{1-\rho^{K+1}}$$

- Throughput (or departure flow) at steady state (or resource productivity)

$$\eta = (1 - \pi_0(\infty)) \cdot \mu = \hat{U} \cdot \mu = \frac{\lambda(1-\rho^K)}{1-\rho^{K+1}} \quad \Longrightarrow \quad \eta = \lambda - \lambda_{ab} = \lambda_{in}$$

- Mean arrivals rate within the resource at steady state

- From the Burke's theorem
at steady state $\eta = \lambda_{in}$ $\Rightarrow \lambda_{in} = \eta = \frac{\lambda(1-\rho^K)}{1-\rho^{K+1}}$ $\Leftrightarrow$ $\lambda_{in} = \lambda - \lambda \cdot \pi_K(\infty)$

M/M/1/K : Quantities of interest

- By means of the moment generating function one has

$$\Pi(z,t) = \sum_{i=0}^{K} \pi_i(t) \cdot z^{-i} = \sum_{i=0}^{K} \frac{\rho^i(1-\rho)}{(1-\rho^{K+1})} \cdot z^{-i} = \frac{1-\rho}{1-\rho^{K+1}} \sum_{i=0}^{K} \left(\frac{\rho}{i!}\right)^i$$

$$= \frac{1-\rho}{1-\rho^{K+1}} \cdot \frac{1-\left(\frac{\rho}{z}\right)^{K+1}}{1-\frac{\rho}{z}}$$

- Mean number of costumer in the resource at steady state

$$\bar{x} = -\left.\frac{d\Pi(z,t)}{dz}\right|_{z=1} = \frac{\rho(1-(K+1)\rho^K + K\rho^{K+1})}{(1-\rho^{K+1})(1-\rho)}$$

Notice that:

$$\begin{cases} \lim_{\rho \to 0} \bar{x}(\rho) = 0 \\ \lim_{\rho \to \infty} \bar{x}(\rho) = K \end{cases}$$

M/M/1/K : Quantities of interest

- The mean time spent in the system by a customer is

$$\bar{\theta} = \frac{\bar{x}}{\lambda_{in}} = \frac{(1-(\mathrm{K}+1)\rho^{\mathrm{K}}+\mathrm{K}\rho^{\mathrm{K}+1})}{\mu(1-\rho)(1-\rho^{\mathrm{K}})}$$

$$\begin{cases} \lim_{\rho\to 0} \bar{\theta}(\rho) = \dfrac{1}{\mu} \\ \lim_{\rho\to \infty} \bar{\theta}(\rho) = \dfrac{\mathrm{K}}{\mu} \end{cases}$$

- The mean service time at steady state remains constant and equal to

$$\bar{\theta}_s = \frac{1}{\mu}$$

- The mean time spent in the buffer by a costumer at steady state

$$\bar{\theta}_b = \bar{\theta} - \bar{\theta}_s = \bar{\theta} - \frac{1}{\mu} = \frac{\rho(1-\mathrm{K}\rho^{\mathrm{K}-1}+(\mathrm{K}-1)\rho^{\mathrm{K}})}{\mu(1-\rho)(1-\rho^{\mathrm{K}})}$$

M/M/1/K : Quantities of interest

- Mean number of busy servers at steady state (coincides with the utilization factor because of it as 1 server only)

$$\bar{x}_s = \lambda_{in} \cdot \bar{\theta}_s = (\lambda - \lambda \pi_K(\infty)) \cdot \bar{\theta}_s = \frac{\rho(1-\rho^{\mathrm{K}})}{1-\rho^{\mathrm{K}+1}}$$

- Mean number of buffered costumers at steady state

$$\bar{x}_b = \lambda_{in} \cdot \bar{\theta}_b = \frac{\lambda(1-\rho^{\mathrm{K}})}{1-\rho^{\mathrm{K}+1}} \cdot \frac{\rho(1-\mathrm{K}\rho^{\mathrm{K}-1}+(\mathrm{K}-1)\rho^{\mathrm{K}})}{\mu(1-\rho)(1-\rho^{\mathrm{K}})}$$

$$= \frac{\rho^2(1-\mathrm{K}\rho^{\mathrm{K}-1}+(\mathrm{K}-1)\rho^{\mathrm{K}})}{(1-\rho)(1-\rho^{\mathrm{K}+1})}$$

- The single server utilization at steady state, since $m = 1$, becomes

$$\tilde{\rho} = \left.\frac{\bar{x}_s}{m}\right|_{m=1} = \hat{U} = \frac{\rho(1-\rho^{\mathrm{K}})}{1-\rho^{\mathrm{K}+1}}$$

M/M/m

- The M/M/m is the generalization with m independent servers of a M/M/1

- A M/M/m can be modeled by a CT-BDP where:

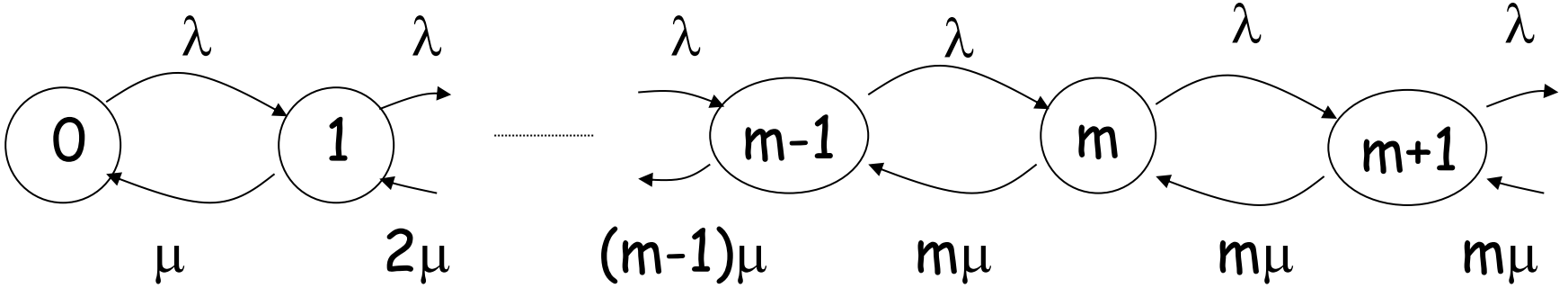
- 1. The birth rate is constant

$$\lambda_i = \lambda \qquad \forall\, i$$

- 2. The death rate depends on the number of costumer in the resource

$$\mu_i = \begin{cases} i \cdot \mu & \text{if} \quad i \leq m \\ m \cdot \mu & \text{if} \quad i > m \end{cases}$$

- where μ denotes the service rate.



M/M/m (cont'd)

- Since the CT-BDP has an infinite states and m servers the process is ergodic if,

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

- If the ergodic condition is satisfied, its limiting distribution is the solution of

$$\begin{cases} \Pi_s \cdot Q = 0 \\ \sum_{i=0}^{\infty} \pi_{s,i} = 1 \end{cases}$$

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

$$\begin{cases} \pi_{s,i+1} = \dfrac{\lambda_i}{\mu_{i+1}} \pi_{s,i} \quad \forall\, i \geq 0 \\ \displaystyle\sum_{i=0}^{\infty} \pi_{s,i} = 1 \end{cases}$$

$$\begin{cases} \pi_{s,1} = \frac{\lambda}{\mu}\pi_{s,0} = m\rho\,\pi_{s,0} \\ \pi_{s,2} = \frac{\lambda}{2\mu}\pi_{s,1} = \frac{(m\rho)^2}{2}\pi_{s,0} \\ \pi_{s,3} = \frac{\lambda}{3\mu}\pi_{s,2} = \frac{(m\rho)^3}{3!}\pi_{s,0} \\ \quad\vdots \\ \pi_{s,m} = \frac{\lambda}{m\mu}\pi_{s,m-1} = \frac{(m\rho)^m}{m!}\pi_{s,0} \\ \\ \pi_{s,m+1} = \frac{\lambda}{m\mu}\pi_m = \rho\frac{(m\rho)^m}{m!}\pi_{s,0} \\ \pi_{s,m+2} = \frac{\lambda}{m\mu}\pi_{m+1} = \rho^2\frac{(m\rho)^m}{m!}\pi_{s,0} \\ \quad\vdots \\ \sum_{i=0}^{\infty}\pi_{s,i} = 1 \end{cases}$$

M/M/m (cont'd)

$$\begin{cases} \pi_{s,1} = \frac{\lambda}{\mu}\pi_{s,0} = m\rho\,\pi_{s,0} \\ \pi_{s,2} = \frac{\lambda}{2\mu}\pi_{s,1} = \frac{(m\rho)^2}{2}\pi_{s,0} \\ \pi_{s,3} = \frac{\lambda}{3\mu}\pi_{s,2} = \frac{(m\rho)^3}{3!}\pi_{s,0} \\ \quad\vdots \\ \pi_{s,m} = \frac{\lambda}{m\mu}\pi_{s,m-1} = \frac{(m\rho)^m}{m!}\pi_{s,0} \\ \\ \pi_{s,m+1} = \frac{\lambda}{m\mu}\pi_m = \rho\frac{(m\rho)^m}{m!}\pi_{s,0} \\ \pi_{s,m+2} = \frac{\lambda}{m\mu}\pi_{m+1} = \rho^2\frac{(m\rho)^m}{m!}\pi_{s,0} \\ \quad\vdots \\ \sum_{i=0}^{\infty}\pi_{s,i} = 1 \end{cases}$$

$$\begin{cases} \pi_{s,i} = \frac{(m\rho)^i}{i!}\pi_{s,0} & i \le m \\ \pi_{s,i} = \rho^i\frac{(m\rho)^m}{m!}\pi_{s,0} & i > m \\ \pi_{s,0}\cdot\left(\sum_{i=0}^{m-1}\frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!}\sum_{i=m}^{\infty}\rho^i\right) = 1 \end{cases}$$

$$\pi_{s,0}\cdot\left(\sum_{i=0}^{m-1}\frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!}\overbrace{\sum_{i=0}^{\infty}\rho^i}^{\frac{1}{(1-\rho)}}\right) = 1$$

$$\pi_{s,0} = \frac{1}{\sum_{i=0}^{m-1}\frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)}}$$

- Remark 1: If it was considered a M/M/m/m, instead of a M/M/m, $\pi_{s,i} = 0, \forall\ i > m$

- Remark 2: The blocking prob. of an M/M/m/m is called **Erlang-B loss formula:**

$$\pi_m(\infty) = \frac{(m\rho)^m}{m!}\cdot\pi_0(\infty) = \frac{\frac{(m\rho)^m}{m!}}{\sum_{i=0}^{m-1}\frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!}}$$

See: https://en.wikipedia.org/wiki/Erlang_(unit)

M/M/m : Quantities of interest

- The Idle rate (or Idle probability) is

$$\pi_0(\infty) = \frac{1}{\sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)}}$$

- Probability of having exactly i costumers in the system at steady state is

$$\begin{cases} \pi_i(\infty) = \frac{(m\rho)^i}{i!}\pi_0(\infty) & i \leq m \\ \pi_i(\infty) = \rho^i \frac{(m\rho)^m}{m!}\pi_0(\infty) & i > m \end{cases}$$

- Waiting Probability: due to the PASTA property this is the proportion of time an incoming customer will have to wait before being served

Erlang-C Formula

$$\Pr(x(\infty) \geq m) = \Pr(\ \theta_b > 0\ ) = \sum_{i=m}^{\infty} \pi_i(\infty) = 1 - \sum_{i=0}^{m-1} \pi_i(\infty) = \frac{\frac{(m\rho)^m}{m!(1-\rho)}}{\sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)}}$$

See: https://en.wikipedia.org/wiki/Erlang_(unit)

M/M/m : Quantities of interest (cont'd)

- Mean service time:

$$\bar{\theta}_s = \frac{1}{\mu}$$

- Mean number of busy servers at steady state

$$\bar{x}_s = \sum_{i=0}^{m-1} i \cdot \pi_{s,i} + m \cdot \Pr(x \geq m)$$

Little's Law

$$\bar{x}_s = \lambda \cdot \bar{\theta}_s = \frac{\lambda}{\mu} = m \cdot \rho$$

- The utilization of a single server at steady state becomes

$$\tilde{\rho} = \frac{\bar{x}_s}{m} = \frac{m\rho}{m} = \rho$$

- Throughput at steady state, namely, the long term productivity of a server is

$$\eta = \lambda$$ (from the Burke's theorem)

M/M/m : Quantities of interest (cont'd)

- Mean number of costumer in the resource at steady state

(given without proof) $$\bar{x} = m \cdot \rho + \frac{m^m \rho^{m+1}}{m!(1-\rho)^2} \pi_0(\infty)$$

- Mean number of costumer in the buffer at steady state

$$\bar{x}_b = \bar{x} - \bar{x}_s = \bar{x} - m \cdot \rho$$

- Mean time spent in the system by a customer at steady state

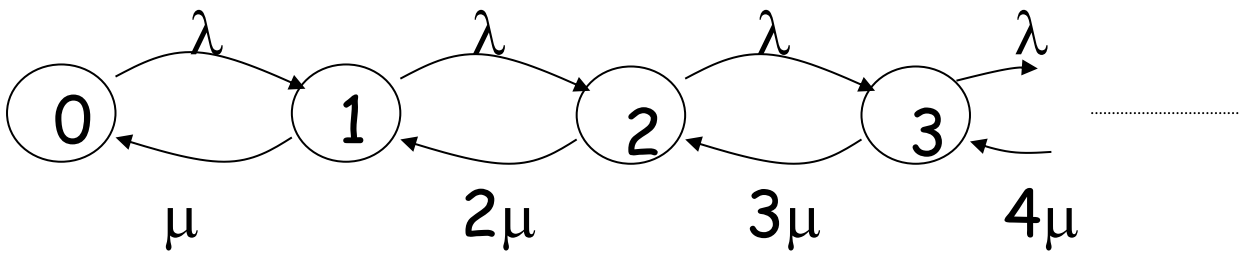
(from the Little's Law) $$\bar{\theta} = \frac{\bar{x}}{\lambda}$$

- Mean time spent in the buffer by a costumer at steady state

$$\bar{\theta}_b = \bar{\theta} - \bar{\theta}_s = \frac{\bar{x}}{\lambda} - \frac{1}{\mu}$$

M/M/∞

- Because the number of servers is unlimited all customers go immediately into the service area upon their arrival
- The M/M/∞ can be modelled by CT-BDP as follows



1) Costumer arrive in a Poisson fashion with constant rate

$$\lambda_i = \lambda \quad \forall\, i$$

2) The death rate increases with the number of busy servers

$$\mu_i = i \cdot \mu \quad \forall\, i$$

where μ denote the service rate of each server

Clearly ergodic

$$\forall\, \mu > 0, \forall\, \lambda > 0$$

M/M/∞ (cont'd)

- Because ergodic its limiting distribution is the solution of

$$\begin{cases} \Pi_s \cdot \boldsymbol{Q} = \mathbf{0} \\ \sum_i \pi_{i,s} = 1 \end{cases}$$

$$\begin{cases} \pi_{s,i+1} = \dfrac{\lambda_i}{\mu_{i+1}}\pi_{s,i} = \dfrac{\lambda}{(i+1)\mu}\pi_{s,i} = \dfrac{\rho}{(i+1)}\pi_{s,i} \quad \forall\, i \geq 0 \\ \displaystyle\sum_{i=0}^{\infty} \pi_{s,i} = 1 \end{cases} \qquad \rho = \frac{\lambda}{\mu}$$

$$\sum_{i=0}^{\infty} \frac{\rho^i}{i!}\pi_{s,0} = \pi_{s,0} \cdot \sum_{i=0}^{\infty} \frac{\rho^i}{i!} = \pi_{s,0} \cdot \mathrm{e}^{\rho} = 1$$

$$\begin{cases} \pi_{s,1} = \rho\,\pi_{s,0} \\ \pi_{s,2} = \frac{\rho}{2}\pi_{s,1} = \frac{\rho^2}{2}\pi_{s,0} \\ \quad \vdots \\ \pi_{s,i+1} = \frac{\rho}{i}\pi_{s,i} = \frac{\rho^i}{i!}\pi_{s,0} \\ \sum_{i=0}^{\infty} \pi_{s,i} = 1 \end{cases}$$

$$\pi_{s,0} = \mathrm{e}^{-\rho}$$

$$\pi_{s,i} = \frac{\rho^i}{i!} \cdot \mathrm{e}^{-\rho}$$

M/M/∞ : Quantities of interest

- Probability of having exactly i costumers in the system at steady state is

$$\pi_i(\infty) = \frac{\rho^i}{i!} \cdot \pi_0(\infty) = \frac{\rho^i}{i!} \cdot \mathrm{e}^{-\rho} \quad i \geq 0$$

- Idle probability

$$\pi_0(\infty) = \mathrm{e}^{-\rho}$$

- Utilization factor, namely, the average occupancy of the resource during a specified period of time

$$\hat{U} = \Pr(x \geq 1) = 1 - \pi_{s,0} = 1 - \mathrm{e}^{-\rho}$$

M/M/∞ : Quantities of interest (cont'd)

- Mean number of costumer in the resource at steady state (see slides 48)

$$\bar{x} = \rho$$

- Remark: Notice that as the traffic intensity increase, the mean number of costumer in the system increase as well.

- Throughput at steady state, namely, the long term resource productivity

$$\eta = \lambda$$ (From the Burke's theorem)

- From the Little's Law the mean time spent in the system by a customer is

$$\bar{\theta} = \frac{\bar{x}}{\lambda} = \frac{\rho}{\lambda} = \frac{1}{\mu}$$

- The mean service time at steady state remains constant and equal to

$$\bar{\theta}_s = \frac{1}{\mu} = \bar{\theta}$$

M/M/∞ : Quantities of interest (cont'd)

- Mean time spent by a customer in the buffer

$$\bar{\theta}_b = \bar{\theta} - \bar{\theta}_s = 0$$

- Mean number of costumer in the buffer at steady state

$$\bar{x}_b = \bar{x} - \bar{x}_s = \rho - \rho = 0$$

- Mean number of busy servers at steady state

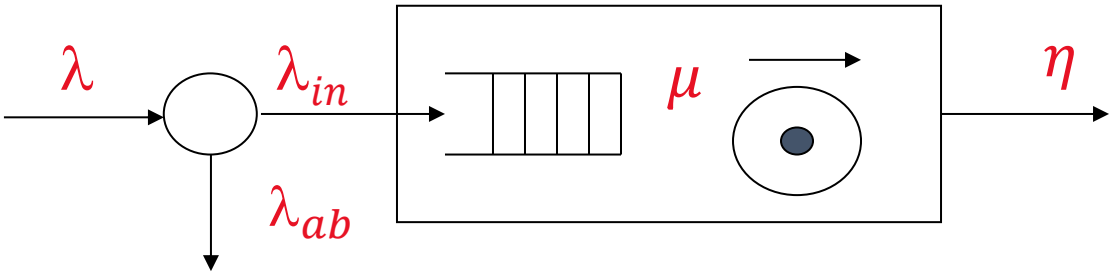
$$\bar{x}_s = \lambda \cdot \bar{\theta}_s = \lambda \cdot \frac{1}{\mu} = \rho$$

- The utilization of a single server at steady state, since $m = \infty$, becomes

$$\tilde{\rho} = \left. \frac{\bar{x}_s}{m} \right|_{m=\infty} = 0$$

M/M/1 with discouraged arrivals

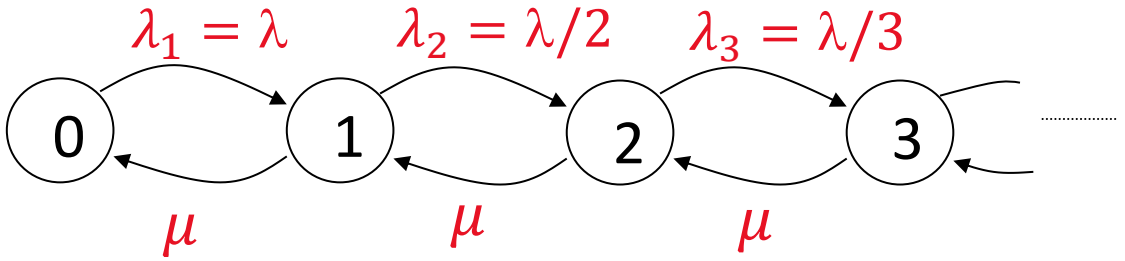
- The arrival flow λ is subject to a splitting process where:
- λ_{ab} is the abandonment rate



- In this queue system the **effective arrival flow** is Poisson where

$$\lambda_{in} = \lambda_i = \frac{\lambda}{(i+1)}$$

- the mean input rate λ_{in} depends on the actual number i of customer in the system
- The service time are, as usual, IID and exponentially distributed with rate μ
- It can be modelled by a time-homogenous, but not non-uniform, where the births decrease according to that hyperbolic law, while deaths are uniform



- Although different, it results that

$$\pi_{s,i} = \frac{\rho^i}{i!} \cdot \pi_{s,0} = \frac{\rho^i}{i!} \cdot e^{-\rho}, \quad \bar{x} = \rho$$

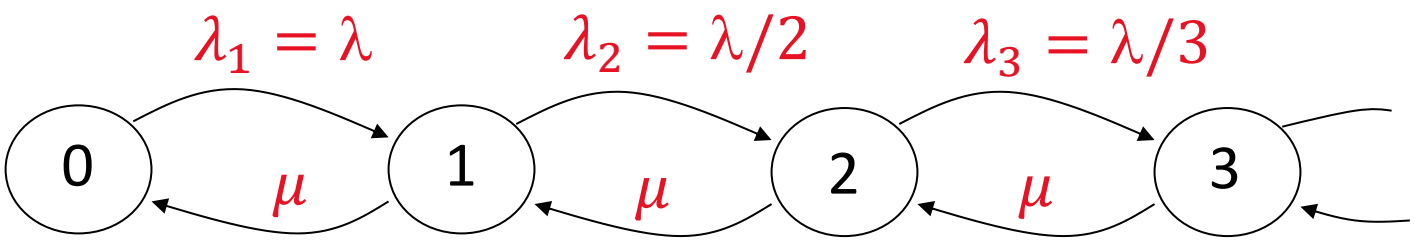
take the same value of the M/M/∞

M/M/1 with discouraged arrivals VS M/M/∞

- Remark 1: Althought the steady state probability of a M/M/∞ resource coincides with that of a M/M/1 with discouraged arrivals its functioning is completely different.

$$\pi_{s,0} = e^{-\rho}$$

$$\pi_{s,i} = \frac{\rho^i}{i!} \cdot \pi_{s,0}$$



- This fact would emphasize the following aspect:

The probability distribution of a given queue, seen individually, is not representative of how the resource works

To fully characterize a queue, we must consider all the fields of the associated Kendall Notation and the resulting queue model

M/M/1 with discouraged arrivals: Quantities of interest

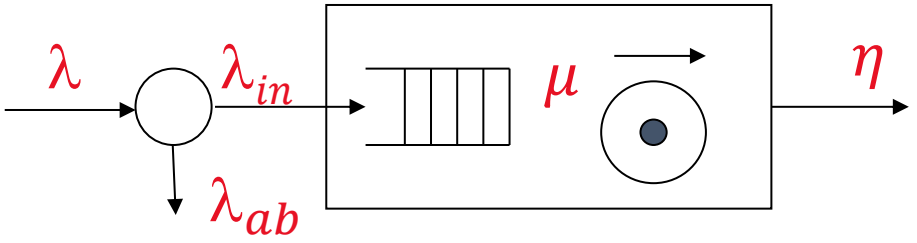
- Throughput (or departure flow) at steady state,

$$\eta = (1-\pi_{s,0})\cdot\mu = \hat{U}\cdot\mu = (1-\mathrm{e}^{-\rho})\cdot\mu \qquad \text{with} \qquad \pi_{s,0} = \mathrm{e}^{-\rho}$$

- Rate of arrivals within the resource at steady state

$$\lambda_{in} = \eta = (1-\mathrm{e}^{-\rho})\mu$$

- From the Burke's theorem at steady state $\eta = \lambda_{in}$



- Rate of abandonment at steady state

$$\lambda_{ab} = \lambda - \lambda_{in} = \lambda - (1-\mathrm{e}^{-\rho})\mu = \mu\cdot(\rho - 1 + \mathrm{e}^{-\rho})$$

- **Remark:** All other quantities of interest can be straightforwardly derived as made for the M/M/∞ queue and by exploiting the Little's Law.
- E.g. let as for the M/M/∞ $\bar{x} = \rho$ then one derives that:

$$\bar{\theta} = \bar{x}/\lambda_{in} = \rho/(\mu(1-\mathrm{e}^{-\rho}))$$

M/M/1 with discouraged arrivals : Quantities of interest

- From the Little's Law the mean time spent in the system by a customer is

$$\bar{x} = \lambda_{in} \cdot \bar{\theta} \quad \Rightarrow \quad \bar{\theta} = \frac{\bar{x}}{\lambda_{in}} = \frac{\rho}{\mu(1-e^{-\rho})}$$

- From the Little's Law the mean number of busy servers at steady state, which coincides with the utilization factor

$$\bar{x}_s = \lambda_{in} \cdot \bar{\theta}_s = \mu(1-e^{-\rho}) \cdot \frac{1}{\mu} = (1-e^{-\rho})$$

- Mean number of buffered costumers at steady state

$$\bar{x} = \bar{x}_b + \bar{x}_s \quad \Rightarrow \quad \bar{x}_b = \bar{x} - (1-e^{-\rho}) = \rho - 1 + e^{-\rho}$$