# Introduction to modeling and analysis of queues

B.J. Prabhu
LAAS-CNRS, Toulouse

Grenoble, 04 July 2016

# Questions of interest

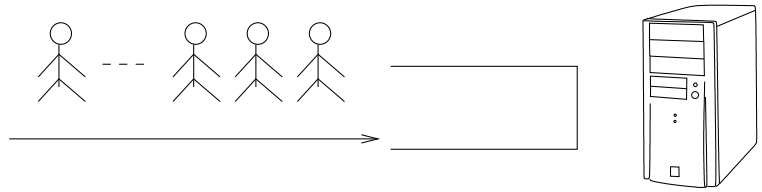
- ✔ Customer point of view
  - Waiting time, Sojourn time
  - Blocking probability
  - Which queue to join?

# Questions of interest

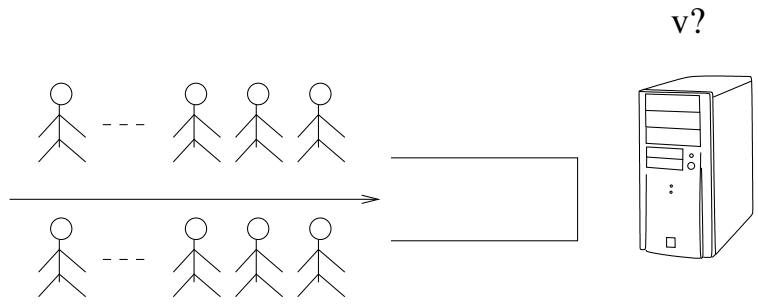
- ✔ Customer point of view
  - Waiting time, Sojourn time
  - Blocking probability
  - Which queue to join?
- ✔ System designer point of view
  - Dimensioning of the system
  - Service discipline
  - Load balancing policy

# Three example problems

# Dimensioning for rush hour



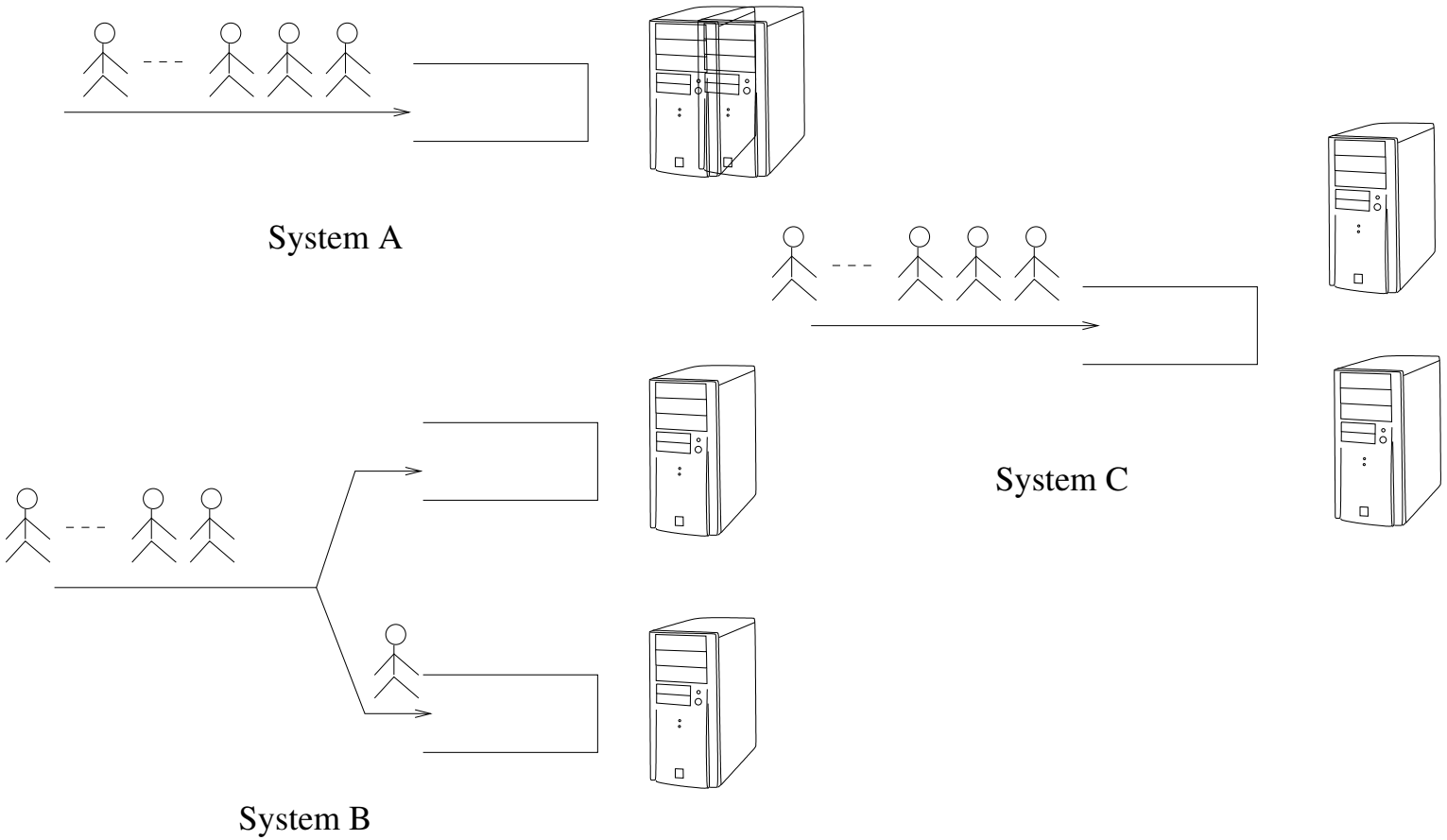
Normal period




Rush hour

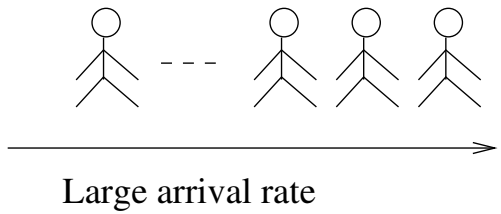
- How fast should the server be during the rush hour?

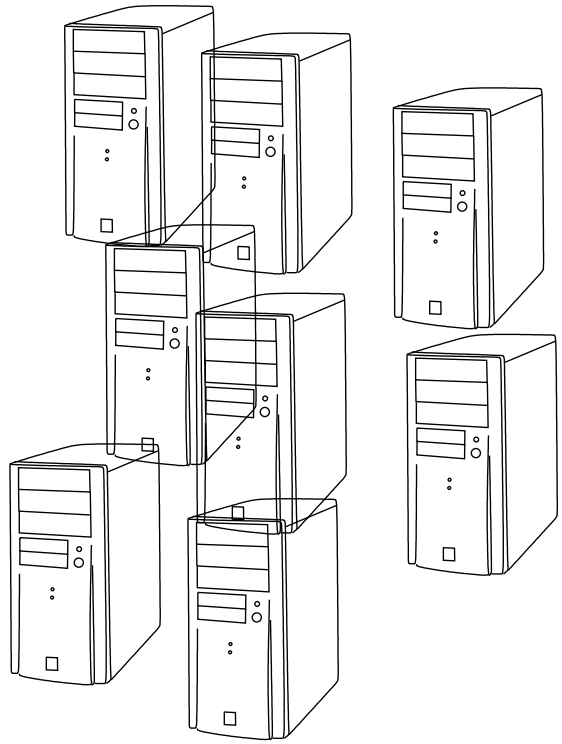
# Comparison of systems




- Which is better for mean sojourn time?

# Dimensioning call/data centers



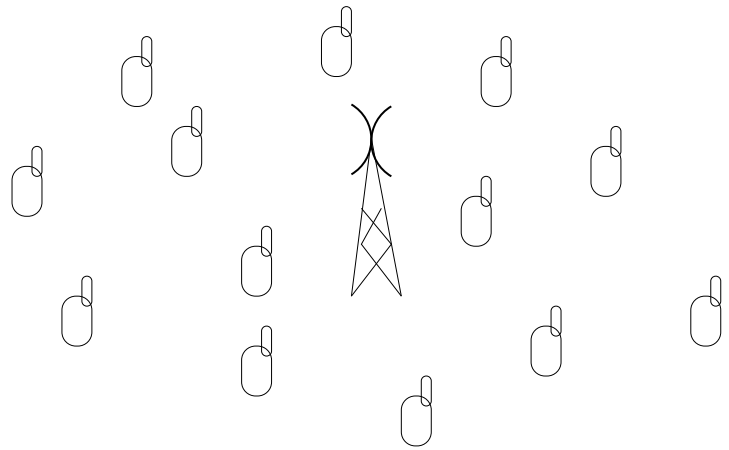
- How many servers?

# Modelling of queues

# Building blocks of a queue

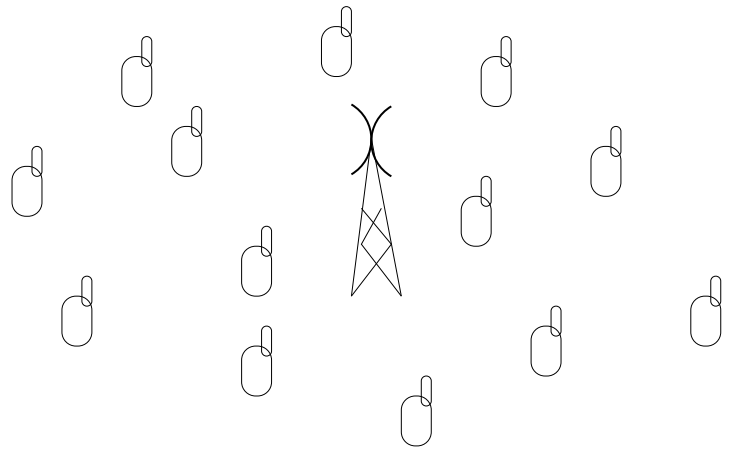
- Arrival process: Poisson process, long-range dependent process,...
- Service time distribution: deterministic, uniform, exponential, Pareto,...
- Number of servers
- System capacity
- Service discipline
- Load balancing
- Retrial, Abandonments,...

# Arrival process: Poisson process



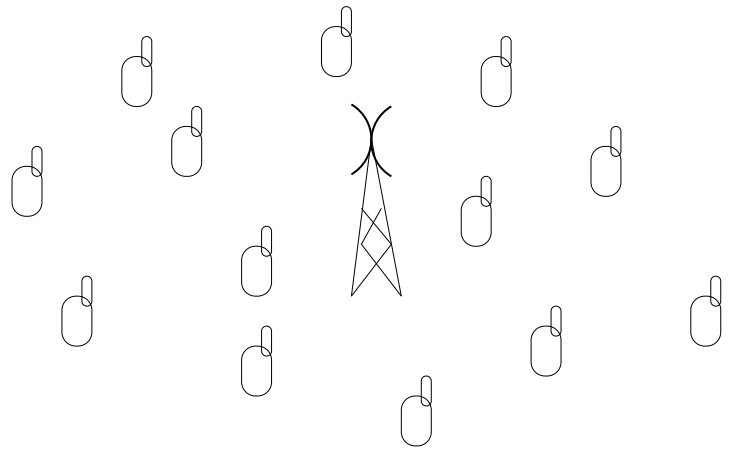
- $N$ nodes, each transmits with probability $p$.

# Arrival process: Poisson process



- $N$ nodes, each transmits with probability $p$.
- Number of transmissions?

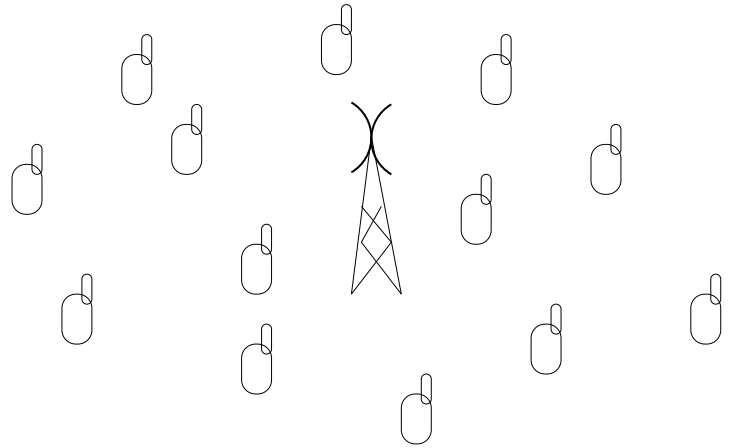
# Arrival process: Poisson process



- $N$ nodes, each transmits with probability $p$.
- Number of transmissions?

$$\mathbb{P}(X = k) = \binom{N}{k} p^k (1-p)^k \tag{1}$$

# Arrival process: Poisson process

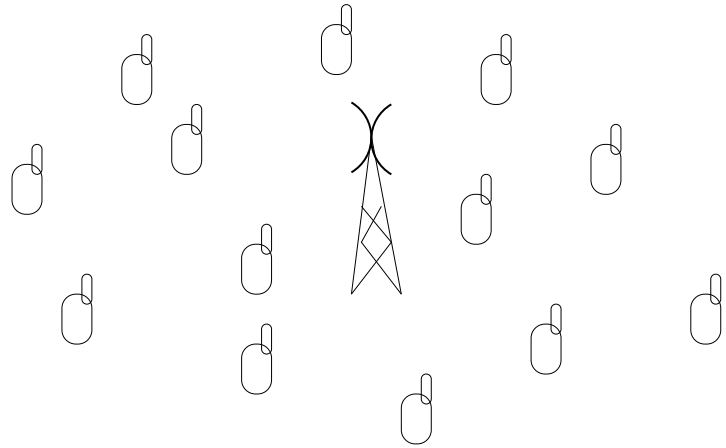


- $N$ nodes, each transmits with probability $p$.
- Number of transmissions?

$$\mathbb{P}(X = k) = \binom{N}{k} p^k (1-p)^k \tag{1}$$

- Expected number of transmissions?

# Arrival process: Poisson process



- $N$ nodes, each transmits with probability $p$.
- Number of transmissions?

$$\mathbb{P}(X = k) = \binom{N}{k} p^k (1-p)^k \tag{1}$$

- Expected number of transmissions? $Np$

# Arrival process: Poisson process

- Set $p = \lambda/N$
- As $N \to \infty$

$$\lim_{N\to\infty} \binom{N}{k} p^k (1-p)^k \to e^{-\lambda} \frac{\lambda^k}{k!} \tag{2}$$

# Arrival process: Poisson process

- Set $p = \lambda/N$
- As $N \to \infty$

$$\lim_{N\to\infty} \binom{N}{k} p^k (1-p)^k \to e^{-\lambda} \frac{\lambda^k}{k!} \tag{2}$$

- Finite number of transmissions. Expected number of transmission is $\lambda$.

# Arrival process: Poisson process

- Set $p = \lambda/N$
- As $N \to \infty$

$$\lim_{N\to\infty} \binom{N}{k} p^k (1-p)^k \to e^{-\lambda} \frac{\lambda^k}{k!} \tag{2}$$

- Finite number of transmissions. Expected number of transmission is $\lambda$.

## Definition (Poisson distribution)

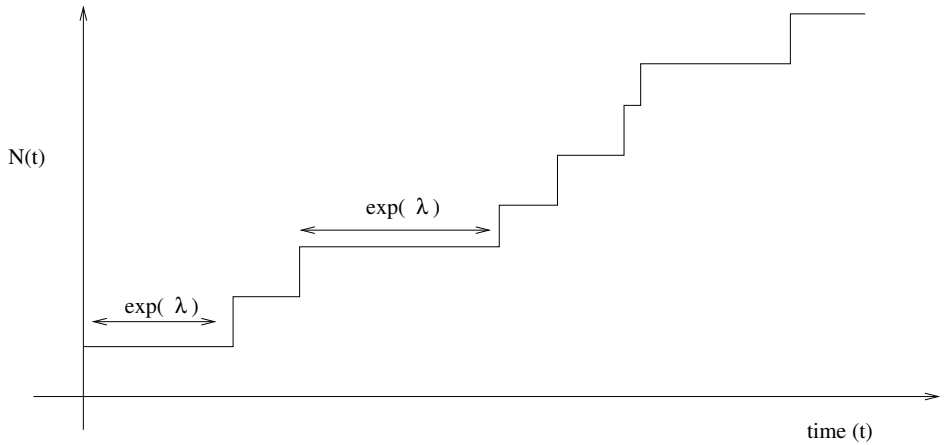
A random variable $Y$ has Poisson distribution if

$$\mathbb{P}(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!} \tag{3}$$

✔ Poisson distribution is the superposition of large number of variables of small size.

# Arrival process - Poisson process

## Definition (Poission process)



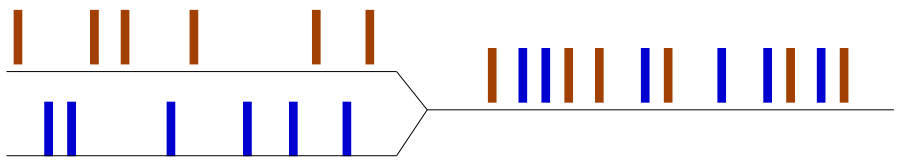

A stochastic process $N(t)$ is a Poisson process if:

- $N(t) \in \mathbb{Z}$
- $N(t)$ has increasing sample paths
- $\mathbb{P}(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$
- Time between jumps in exponentially distributed with rate $\lambda$

# Poisson process:properties

- Superposition: Poisson process is stable by aggregation. If $N_1(t) \sim Poi(\lambda_1)$ and $N_2(t) \sim Poi(\lambda_2)$ then

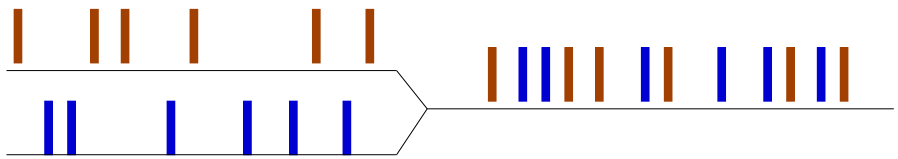
$$N_1(t) + N_t(t) = Poi(\lambda_1 + \lambda_2)$$



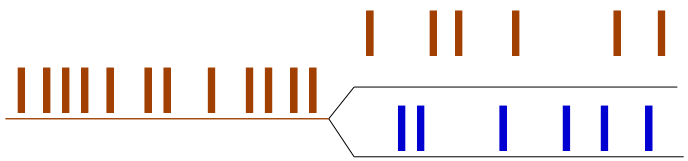
# Poisson process:properties

- Superposition: Poisson process is stable by aggregation. If $N_1(t) \sim Poi(\lambda_1)$ and $N_2(t) \sim Poi(\lambda_2)$ then

$$N_1(t) + N_t(t) = Poi(\lambda_1 + \lambda_2)$$



- Splitting: independent splitting of Poisson process are again Poisson



Does not hold for round-robin.

# Service time distribution

- Exponential: $\mathbb{P}(X > x) = e^{-\mu x}$

# Service time distribution

- Exponential: $\mathbb{P}(X > x) = e^{-\mu x}$
- Pareto: $\mathbb{P}(X > x) \sim x^{-\alpha}$ task size observed in seen in computer systems ($1.1 \leq \alpha \leq 1.3$)

# Service time distribution

- Exponential: $\mathbb{P}(X > x) = e^{-\mu x}$
- Pareto: $\mathbb{P}(X > x) \sim x^{-\alpha}$ task size observed in seen in computer systems $(1.1 \leq \alpha \leq 1.3)$
  80-20 rule, infinite variance $\alpha < 2$

# Service time distribution

- Exponential: $\mathbb{P}(X > x) = e^{-\mu x}$
- Pareto: $\mathbb{P}(X > x) \sim x^{-\alpha}$ task size observed in seen in computer systems $(1.1 \leq \alpha \leq 1.3)$
  80-20 rule, infinite variance $\alpha < 2$
- Deterministic, uniform,...

# Other blocks

- Number of servers $\in \mathbb{N}$
- System capacity $\in \mathbb{N}$
- Service discipline: FIFO, PS, Random, LCFS, SRPT, Priority
- Load balancing: Round robin, Random, JSQ,...

# Kendall notation for queues

- ✔ $A/S/N/C/D/L$
  - $A$: arrival process. $A = M$ if Poisson; $A = D$; deterministic, $A = G$ if General
  - $S$: service time. $S = M$ if exponential; $S = D$; deterministic, $S = G$ if General
  - $N \in \mathbb{N}$
  - $C$: default $\infty$
  - $D$: default FIFO

# Kendall notation for queues

✔ $A/S/N/C/D/L$

- $A$: arrival process. $A = M$ if Poisson; $A = D$; deterministic, $A = G$ if General
- $S$: service time. $S = M$ if exponential; $S = D$; deterministic, $S = G$ if General
- $N \in \mathbb{N}$
- $C$: default $\infty$
- $D$: default FIFO

Examples:

- $M/M/1$, $M/Pareto/2/10$, $M/D/3/JSQ$

# Markov chains

# Discrete-time Markov chains

## Definition (DTMC)

A process $(X_n)_{n\geq 0}$ is a DTMC if

- $X_n$ has countable state space.
- $\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1}, X_{n-2}, ...) = P(X_{n+1} = j | X_n = i)$

## Example

$X_n \in \{No\ rain = 0, Rain = 1\}$

# Discrete-time Markov chains

## Definition (DTMC)

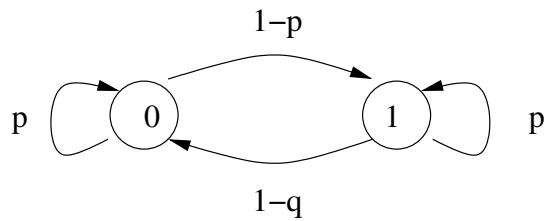
A process $(X_n)_{n\geq 0}$ is a DTMC if

- $X_n$ has countable state space.
- $\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1}, X_{n-2}, ...) = P(X_{n+1} = j | X_n = i)$

## Example

$X_n \in \{No\ rain = 0, Rain = 1\}$

$$P = \begin{array}{c} \\ \text{Current} \end{array} \begin{array}{c} \begin{array}{cc} \text{Next state} & \\ 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \end{array}$$



# Discrete-time Markov chains

## Definition (DTMC)

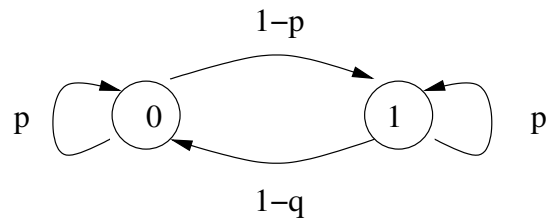
A process $(X_n)_{n\geq 0}$ is a DTMC if

- $X_n$ has countable state space.
- $\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1}, X_{n-2}, ...) = P(X_{n+1} = j | X_n = i)$

## Example

$X_n \in \{No\ rain = 0, Rain = 1\}$

$$P = \begin{array}{c} \\ \text{Current} \end{array} \begin{array}{c} \begin{array}{cc} \text{Next state} & \\ 0 & 1 \end{array} \\ \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \end{array} \begin{array}{c} \\ \begin{array}{c} 0 \\ 1 \end{array} \end{array}$$




✔$P$ is called the transition probability matrix.

# Discrete-time Markov chains

- Long term behavior? $\lim_{n\to\infty} X_n$

## Theorem (Kolmogorov)

*Let $X_n \in \mathcal{S}$. $\lim_{n\to\infty} X_n$converges in distribution to a random variable with distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ which is the solution of*

$$\begin{aligned} \pi P &= \pi \\ \sum_{i\in\mathcal{S}} \pi_i &= 1 \end{aligned}$$

# Discrete-time Markov chains

- Long term behavior? $\lim_{n\to\infty} X_n$

## Theorem (Kolmogorov)

*Let $X_n \in \mathcal{S}$. $\lim_{n\to\infty} X_n$converges in distribution to a random variable with distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ which is the solution of*

$$\begin{aligned} \pi P &= \pi \\ \sum_{i\in\mathcal{S}} \pi_i &= 1 \end{aligned}$$

## Example

$$P = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \Rightarrow (\pi_0, \pi_1) = \left( \frac{1-q}{2-p-q}, \frac{1-p}{2-p-q} \right)$$

# Discrete-time Markov chains

- Long term behavior? $\lim_{n\to\infty} X_n$

## Theorem (Kolmogorov)

*Let $X_n \in \mathcal{S}$. $\lim_{n\to\infty} X_n$converges in distribution to a random variable with distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ which is the solution of*

$$\begin{aligned} \pi P &= \pi \\ \sum_{i\in\mathcal{S}} \pi_i &= 1 \end{aligned}$$

## Example

$$P = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \Rightarrow (\pi_0, \pi_1) = \left( \frac{1-q}{2-p-q}, \frac{1-p}{2-p-q} \right)$$

✔ The vector $\pi$ represents the fraction of time spent in each state

# Continous Time Markov Chain

## Definition (CTMC)

A process $(X(t))_{t\geq 0}$ is a CTMC if

- $X(t)$ has countable state space.
- $\mathbb{P}(X(t+s)|X(s), s \leq t) = P(X(t+s)|X(t))$
- Time spent in each state is exponential

## Example

$X(t) \in \{Down = 0, Up = 1\}$. Time to repair $\sim exp(\lambda)$; Time to failure $\sim exp(\mu)$

# Continous Time Markov Chain

## Definition (CTMC)

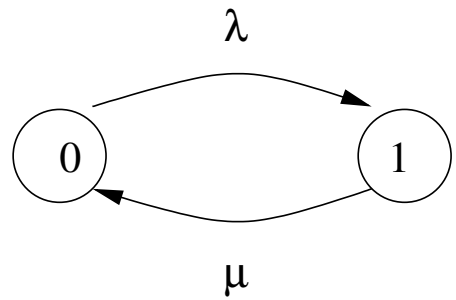
A process $(X(t))_{t\geq 0}$ is a CTMC if

- $X(t)$ has countable state space.
- $\mathbb{P}(X(t+s)|X(s), s \leq t) = P(X(t+s)|X(t))$
- Time spent in each state is exponential

## Example

$X(t) \in \{Down = 0, Up = 1\}$. Time to repair $\sim exp(\lambda)$; Time to failure $\sim exp(\mu)$

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \end{array}$$



# Continous Time Markov Chain

## Definition (CTMC)

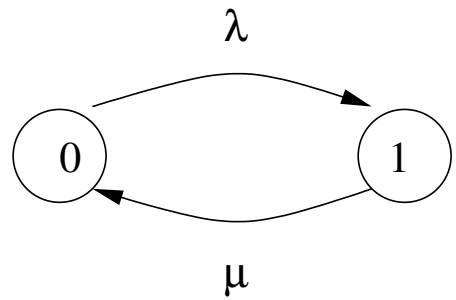
A process $(X(t))_{t\geq 0}$ is a CTMC if

- $X(t)$ has countable state space.
- $\mathbb{P}(X(t+s)|X(s), s \leq t) = P(X(t+s)|X(t))$
- Time spent in each state is exponential

## Example

$X(t) \in \{Down = 0, Up = 1\}$. Time to repair $\sim exp(\lambda)$; Time to failure $\sim exp(\mu)$

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \end{array}$$




✔$Q$ is called the rate matrix.

# Discrete-time Markov chains

- Long term behavior? $\lim_{t\to\infty} X(t)$

## Theorem (Kolmogorov)

*Let $X_n \in \mathcal{S}$. $\lim_{t\to\infty} X(t)$converges in distribution to a random variable with distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ which is the solution of*

$$\begin{aligned} \pi Q &= 0 \\ \sum_{i\in\mathcal{S}} \pi_i &= 1 \end{aligned}$$

# Discrete-time Markov chains

- Long term behavior? $\lim_{t\to\infty} X(t)$

## Theorem (Kolmogorov)

*Let $X_n \in \mathcal{S}$. $\lim_{t\to\infty} X(t)$converges in distribution to a random variable with distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ which is the solution of*

$$\begin{aligned} \pi Q &= 0 \\ \sum_{i\in\mathcal{S}} \pi_i &= 1 \end{aligned}$$

## Example

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \Rightarrow (\pi_0, \pi_1) = \left( \frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right)$$

# Discrete-time Markov chains

- Long term behavior? $\lim_{t\to\infty} X(t)$

## Theorem (Kolmogorov)

*Let $X_n \in \mathcal{S}$. $\lim_{t\to\infty} X(t)$converges in distribution to a random variable with distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ which is the solution of*

$$\begin{aligned} \pi Q &= 0 \\ \sum_{i\in\mathcal{S}} \pi_i &= 1 \end{aligned}$$

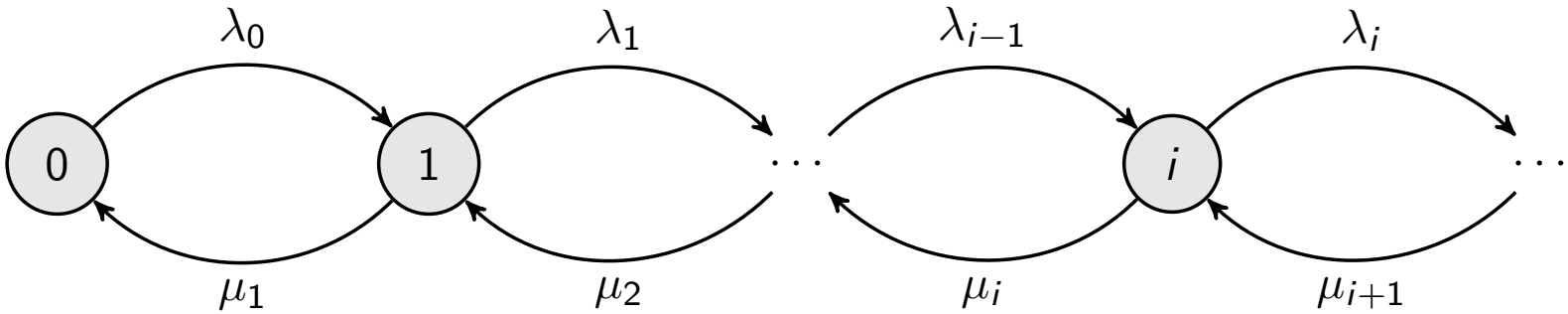
## Example

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \Rightarrow (\pi_0, \pi_1) = \left( \frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right)$$

✔ The vector $\pi$ represents the fraction of time spent in each state

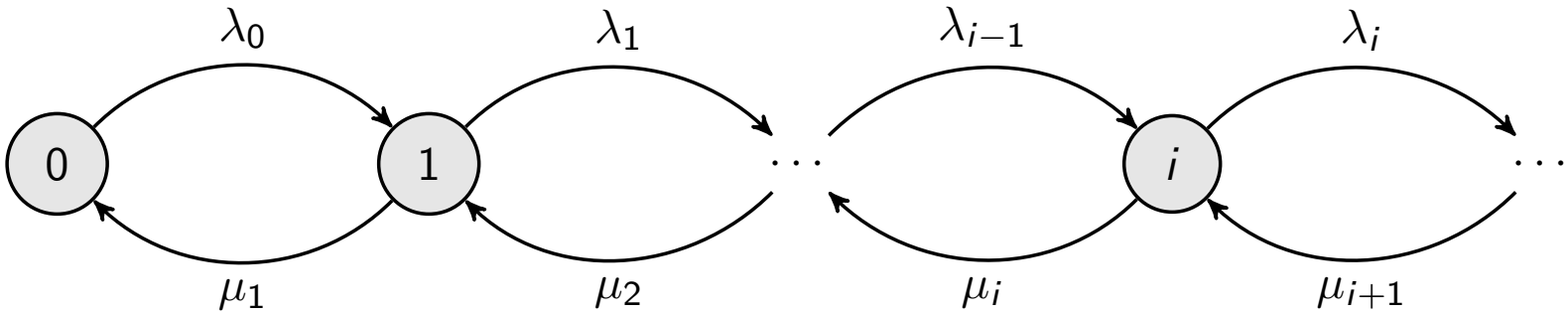
# Birth-death processes

- A special case of CTMC with jumps restricted to neighboring states



# Birth-death processes

- A special case of CTMC with jumps restricted to neighboring states




- Simple form for steady state probabilities

$$\pi_i = \pi_0 \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}$$

$$\pi_0 = \frac{1}{\sum_{j=0}^{\infty} \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}}$$

# Analysis of queues: single server case

# M/M/1

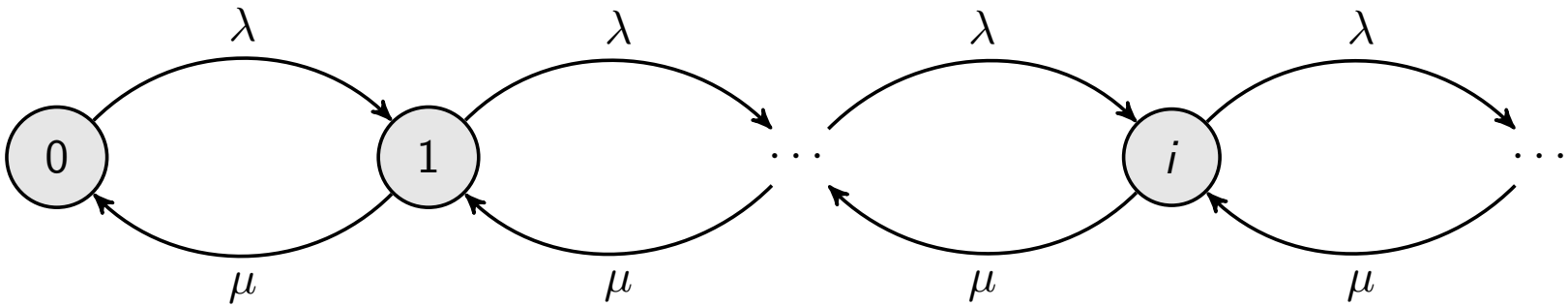
- Arrivals $\sim Poi(\lambda)$; service times $\sim exp(\mu)$; FIFO

# M/M/1

- Arrivals $\sim Poi(\lambda)$; service times $\sim exp(\mu)$; FIFO
- $X(t)$: number in the system at time $t$ is a CTMC

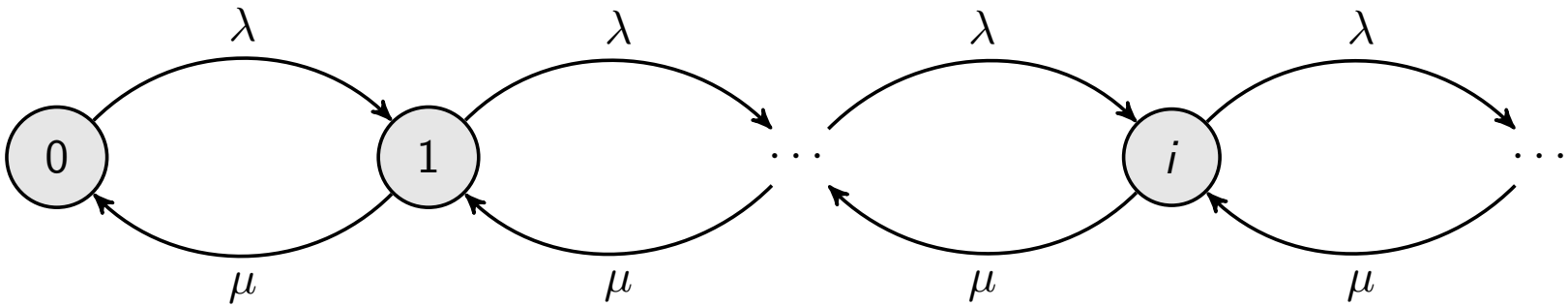
# M/M/1

- Arrivals $\sim Poi(\lambda)$; service times $\sim exp(\mu)$; FIFO
- $X(t)$: number in the system at time $t$ is a CTMC



# M/M/1

- Arrivals $\sim Poi(\lambda)$; service times $\sim exp(\mu)$; FIFO
- $X(t)$: number in the system at time $t$ is a CTMC




- Stationary distribution

$$\pi_i := \mathbb{P}(X(\infty) = i) = (1 - \rho)\rho^i,$$

where $\rho = \frac{\lambda}{\mu} < 1$ is the load on the system

- $\rho < 1$ is the (natural) stability condition

# M/M/1: Performance measures

- Expected number in system:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho}$$

# M/M/1: Performance measures

- Expected number in system:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho}$$

- Expected number in queue:

$$\mathbb{E}[X_q] = \sum_{i=1}^{\infty} (i-1)\pi_i = \frac{\rho^2}{1-\rho}$$

# M/M/1: Performance measures

- Expected number in system:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho}$$

- Expected number in queue:

$$\mathbb{E}[X_q] = \sum_{i=1}^{\infty} (i-1)\pi_i = \frac{\rho^2}{1-\rho}$$

- Note that $E[X] = \rho + E[X_q]$

# M/M/1: Performance measures

- Expected number in system:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho}$$

- Expected number in queue:

$$\mathbb{E}[X_q] = \sum_{i=1}^{\infty} (i-1)\pi_i = \frac{\rho^2}{1-\rho}$$

- Note that $E[X] = \rho + E[X_q]$
- Fraction of time server is busy $1 - \pi_0 = \rho$

# M/M/1: Performance measures (expected times)

## Theorem (Little's Law)

$$\textit{Expected time spent in a system} = \frac{\textit{Expected number in a system}}{\textit{Arrival rate}}$$

# M/M/1: Performance measures (expected times)

## Theorem (Little's Law)

$$\textit{Expected time spent in a system} = \frac{\textit{Expected number in a system}}{\textit{Arrival rate}}$$

- Expected sojourn time:

$$E[T] = \frac{1}{\lambda} \frac{\rho}{1-\rho}$$

# M/M/1: Performance measures (expected times)

## Theorem (Little's Law)

$$\textit{Expected time spent in a system} = \frac{\textit{Expected number in a system}}{\textit{Arrival rate}}$$

- Expected sojourn time:

$$E[T] = \frac{1}{\lambda}\frac{\rho}{1-\rho}$$

- Expected waiting time

$$\mathbb{E}[W] = \frac{1}{\lambda}\frac{\rho^2}{1-\rho}$$

# Dimensioning for rush period

☞ Arrival rate is twice that in normal period. What should be increase in the server speed in order to have

1. the same sojourn time?
2. the same waiting time?

# Dimensioning for rush period

☞ Arrival rate is twice that in normal period. What should be increase in the server speed in order to have

1. the same sojourn time?
2. the same waiting time?

- $\lambda_r = 2\lambda_n$; $\mu_n$:normal period;
- $\mu_r$?

# Dimensioning for rush period

☞ Arrival rate is twice that in normal period. What should be increase in the server speed in order to have

1. the same sojourn time?
2. the same waiting time?

- $\lambda_r = 2\lambda_n$; $\mu_n$:normal period;
- $\mu_r$?
- Expected sojourn time:

$$\frac{1}{\lambda_n}\frac{\rho_n}{1-\rho_n} = \frac{1}{\lambda_r}\frac{\rho_r}{1-\rho_r}$$

# Dimensioning for rush period

☞ Arrival rate is twice that in normal period. What should be increase in the server speed in order to have

1. the same sojourn time?
2. the same waiting time?

- $\lambda_r = 2\lambda_n$; $\mu_n$:normal period;
- $\mu_r$?
- Expected sojourn time:

$$\frac{1}{\lambda_n}\frac{\rho_n}{1-\rho_n} = \frac{1}{\lambda_r}\frac{\rho_r}{1-\rho_r}$$

- Let $\mu_r = v\mu_n$.

$$v = 1 + \rho_n$$

# Dimensioning for rush period

☞ Arrival rate is twice that in normal period. What should be increase in the server speed in order to have

1. the same sojourn time?
2. the same waiting time?

- $\lambda_r = 2\lambda_n$; $\mu_n$:normal period;
- $\mu_r$?
- For the same expected waiting time

$$v = \rho_n + \sqrt{(1-\rho_n)^2 + 1}$$

# M/G/1

✗ $X(t)$ is not Markov in general

# M/G/1

✗ $X(t)$ is not Markov in general

☞ Use mean value analysis for sojourn (waiting) time

# M/G/1

✗ $X(t)$ is not Markov in general

☞ Use mean value analysis for sojourn (waiting) time

- For waiting time

$$W = \sum_{i=1}^{N_q} S_i + R$$

$R$ is the residual service time

# M/G/1

✗ $X(t)$ is not Markov in general

☞ Use mean value analysis for sojourn (waiting) time

- For waiting time

$$\mathbb{E}[W] = \mathbb{E}[N_q]\mathbb{E}[S_i] + \mathbb{E}[R]$$

# M/G/1

✗ $X(t)$ is not Markov in general

☞ Use mean value analysis for sojourn (waiting) time

- For waiting time

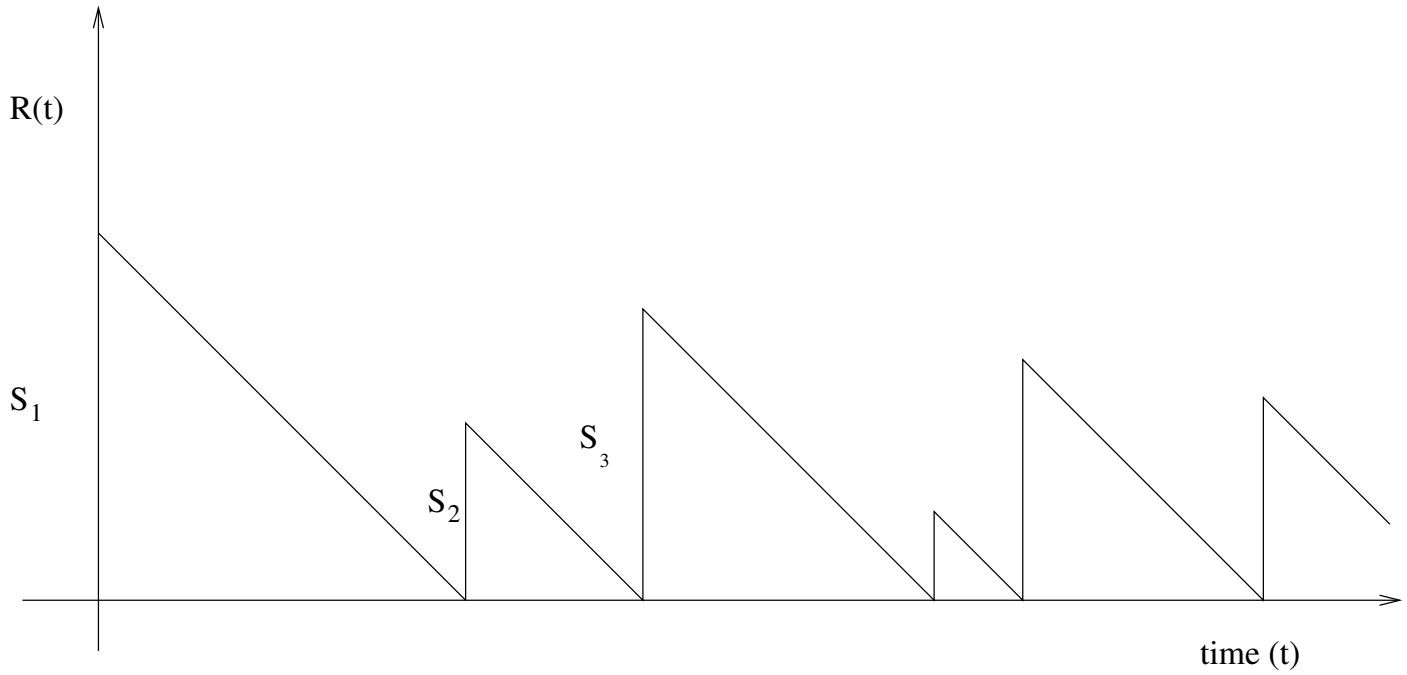
$$\mathbb{E}[W] = \mathbb{E}[N_q]\mathbb{E}[S_i] + \mathbb{E}[R]$$

Using Little's Law

$$\mathbb{E}[W] = \frac{\mathbb{E}[R]}{1-\rho}$$

- For exponential service times residual service time is also exponentially distributed (memoryless property)

# M/G/1




$$\begin{aligned}\mathbb{E}[R] &= \lim_T \frac{1}{T}\int_0^T R(t)dt \\ &= \lim_T \frac{1}{T}\int_0^T \frac{1}{2}S(t)^2 dt \\ &= \frac{1}{2}\mathbb{E}[S^2]\end{aligned}$$

# M/G/1

$\Rightarrow$

$$\mathbb{E}[W] = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)}$$
$$= \frac{1+c_v^2}{2} \frac{\rho}{1-\rho} \mathbb{E}[S]$$

$c_v^2 = \frac{V[S]}{E[S]^2}$ is the squared coefficient of variation, and

$$\mathbb{E}[T] = \mathbb{E}[W] + \mathbb{E}[S]$$

# M/G/1

$\Rightarrow$

$$\mathbb{E}[W] = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)}$$
$$= \frac{1+c_v^2}{2} \frac{\rho}{1-\rho} \mathbb{E}[S]$$

$c_v^2 = \frac{V[S]}{E[S]^2}$ is the squared coefficient of variation, and

$$\mathbb{E}[T] = \mathbb{E}[W] + \mathbb{E}[S]$$

☞Mean in not sufficient! Variance plays a role

# M/G/1

$\Rightarrow$

$$\mathbb{E}[W] = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)}$$
$$= \frac{1+c_v^2}{2} \frac{\rho}{1-\rho} \mathbb{E}[S]$$

$c_v^2 = \frac{V[S]}{E[S]^2}$ is the squared coefficient of variation, and

$$\mathbb{E}[T] = \mathbb{E}[W] + \mathbb{E}[S]$$

☞Mean in not sufficient! Variance plays a role

- Larger the variance, larger is the waiting time
- Tasks get blocked behind large tasks

# Influence of service distribution

☞Deterministic has the least sojourn (waiting) time

- Exponential vs. Deterministic ($c_v^2 = 1, c_v^2 = 0$)

$$\frac{\mathbb{E}[W]_d}{\mathbb{E}[W]_e} = 0.5$$

# Influence of service distribution

☞Deterministic has the least sojourn (waiting) time

- Exponential vs. Deterministic ($c_v^2 = 1, c_v^2 = 0$)

$$\frac{\mathbb{E}[W]_d}{\mathbb{E}[W]_e} = 0.5$$

- Exponential vs Pareto ($c_v^2 = 1, c_v^2 = \frac{1}{\alpha(\alpha-2)}$)

$$\frac{\mathbb{E}[W]_p}{\mathbb{E}[W]_d} = \begin{cases} \frac{(\alpha-1)^2}{2\alpha(\alpha-2)} & \alpha > 2 \\ \infty & \alpha \leq 2 \end{cases}$$

# Influence of service distribution

☞Deterministic has the least sojourn (waiting) time

- Exponential vs. Deterministic ($c_v^2 = 1, c_v^2 = 0$)

$$\frac{\mathbb{E}[W]_d}{\mathbb{E}[W]_e} = 0.5$$

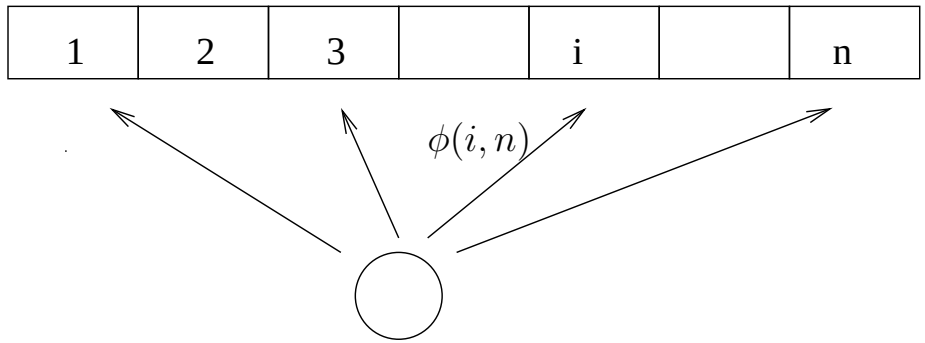
- Exponential vs Pareto ($c_v^2 = 1, c_v^2 = \frac{1}{\alpha(\alpha-2)}$)

$$\frac{\mathbb{E}[W]_p}{\mathbb{E}[W]_d} = \begin{cases} \frac{(\alpha-1)^2}{2\alpha(\alpha-2)} & \alpha > 2 \\ \infty & \alpha \leq 2 \end{cases}$$

☞ How can the influence of variance be reduced?

# Symmetric service disciplines

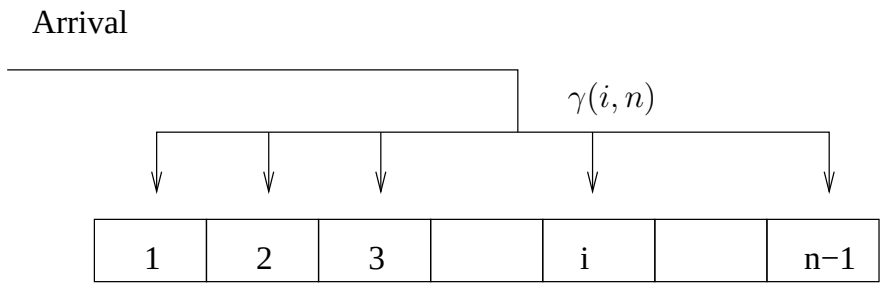
- $\phi(i, n)$: fraction of service rate give to the task $i$th position




  - FIFO: $\phi(1, n) = 1$ and $\phi(i, n) = 0$ for $i \geq 2$
  - PS: $\phi(i, n) = 1/n \ \forall i$
  - LCFS-PR:$\phi(n, n) = 1$ and and $\phi(i, n) = 0$ for $i < n$

# Symmetric service disciplines

- $\gamma(i, n)$: probability the an incoming task joins in position $i$. Tasks in position $i$ move back one spot




  - FIFO: $\gamma(n, n) = 1$ and $\gamma(i, n) = 0$ for $i < 2$
  - PS: $\phi(i, n) = 1/n \ \forall i$
  - LCFS-PR:$\gamma(n, n) = 1$ and and $\gamma(i, n) = 0$ for $i < n$

# Symmetric service disciplines & insensitivity

## Definition (Symmetric discipline)

A service discipline is symmetric if $\phi(i, n) = \gamma(i, n)$

- ✗ FIFO is not symmetric
- ✓ PS, LCFS-PR are symmetric

# Symmetric service disciplines & insensitivity

## Definition (Symmetric discipline)

A service discipline is symmetric if $\phi(i, n) = \gamma(i, n)$

- ✗ FIFO is not symmetric
- ✓ PS, LCFS-PR are symmetric

☞Symmetric service disciplines are *insensitive*. Their performance measures depend only upon the mean service duration and not the distribution

# Symmetric service disciplines & insensitivity

## Definition (Symmetric discipline)

A service discipline is symmetric if $\phi(i, n) = \gamma(i, n)$

- ✗ FIFO is not symmetric
- ✓ PS, LCFS-PR are symmetric

☞Symmetric service disciplines are *insensitive*. Their performance measures depend only upon the mean service duration and not the distribution

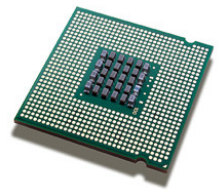
- no dependence of variance
- Steady state distribution of $X(t)$ is same as that of the exponential case

# A Markov model for energy efficiency

# Power consumption of a Processor

For CMOS processors

$$P = P_{static} + cV^2 f,$$

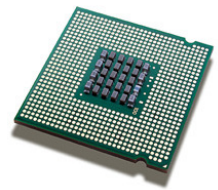


$c$ : effective switch capacitance
$V$ : supply voltage
$f$ : frequency

# Power consumption of a Processor

For CMOS processors

$$P = P_{static} + cV^2 f,$$



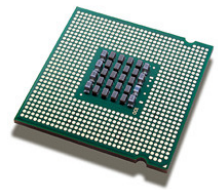
$c$ : effective switch capacitance
$V$ : supply voltage
$f$ : frequency

$$f = c_1 V$$

# Power consumption of a Processor

For CMOS processors

$$P = P_{static} + cV^2 f,$$



$c$ : effective switch capacitance
$V$ : supply voltage
$f$ : frequency

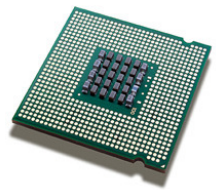
$$f = c_1 V$$

$$\Rightarrow P_{dynamic} = cf^3$$

# Power consumption of a Processor

For CMOS processors

$$P = P_{static} + cV^2 f,$$



$c$ : effective switch capacitance
$V$ : supply voltage
$f$ : frequency

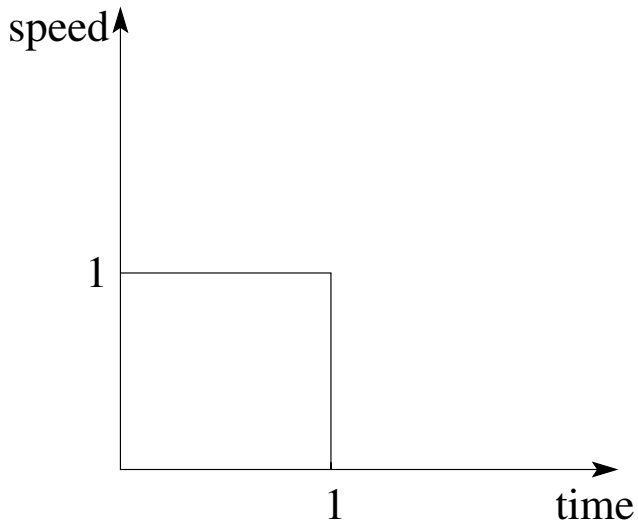
$$f = c_1 V$$

$$\Rightarrow P_{dynamic} = cf^3$$

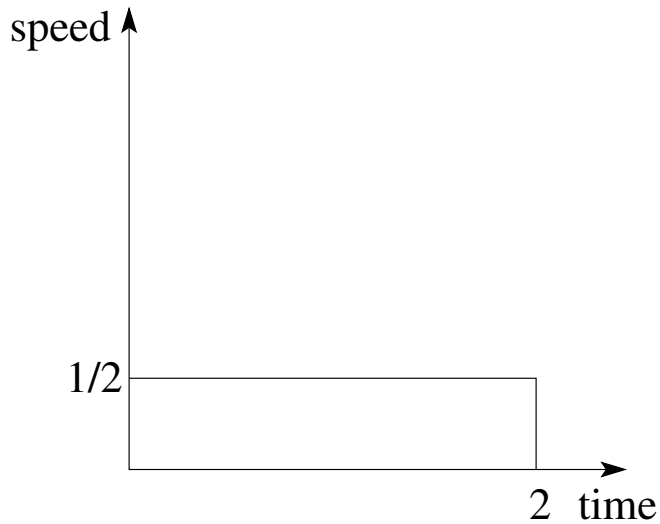
Power-saving strategies : Dynamic Frequency Scaling (DFS), Dynamic Voltage Frequency Scaling (DVFS)

# Energy vs. Performance

$$P \propto f^{\alpha} \quad \text{but} \quad T \propto \frac{1}{f}$$



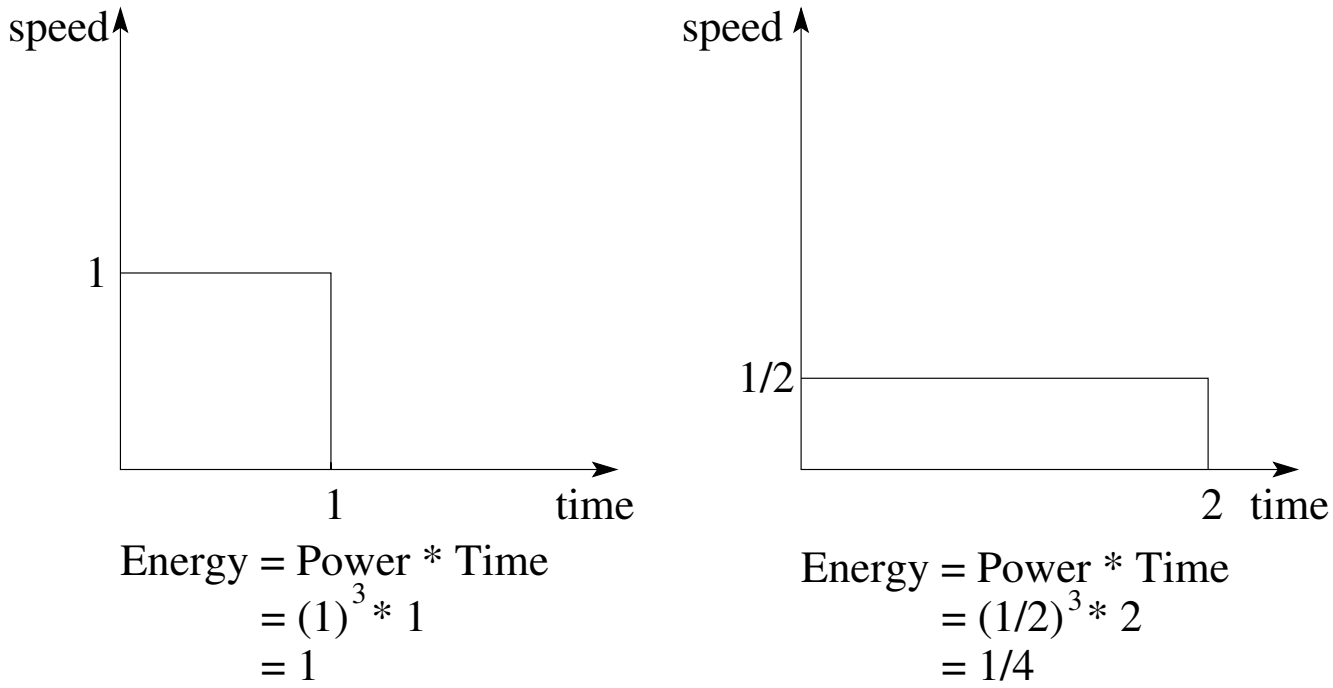

Energy = Power * Time
$= (1)^3 * 1$
$= 1$




Energy = Power * Time
$= (1/2)^3 * 2$
$= 1/4$

# Energy vs. Performance

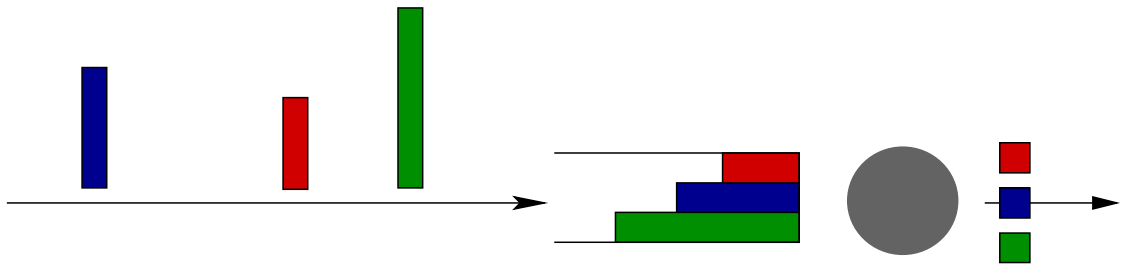
$$P \propto f^{\alpha} \quad \text{but} \quad T \propto \frac{1}{f}$$




- reducing frequency can reduce energy consumption but increase the processing time.

$\Rightarrow$ vary frequency to strike the right balance

# Processor model



- Poisson arrivals at rate $\lambda$
- Tasks sizes have mean 1 unit; can have any distribution
- $v_n \in [0, V]$ : processor speed when $n$ tasks are present.
- Processor Sharing (PS) discipline.

# Objective

- Objective function

  $$\text{Mean processing time} + \beta \cdot \text{mean energy consumption},$$

  $\beta$ is a weighting factor.

# Objective

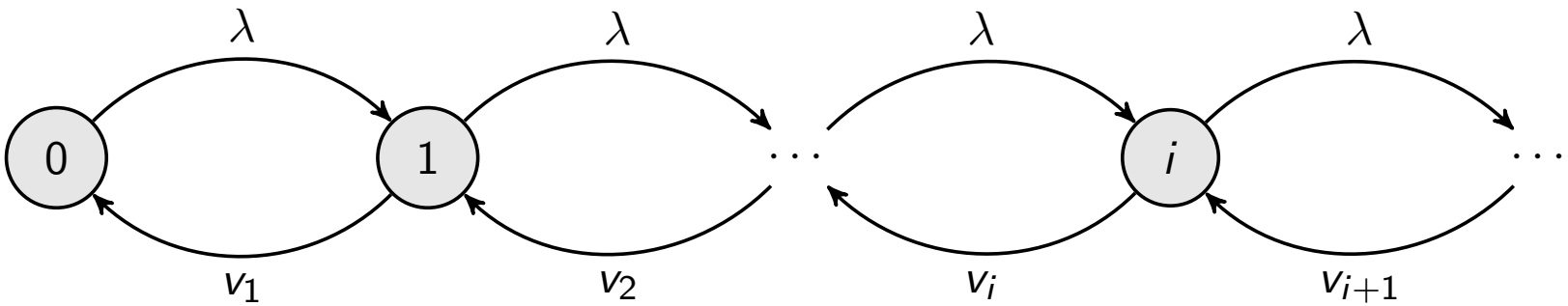
- Objective function

  $$\text{Mean processing time} + \beta \cdot \text{mean energy consumption},$$

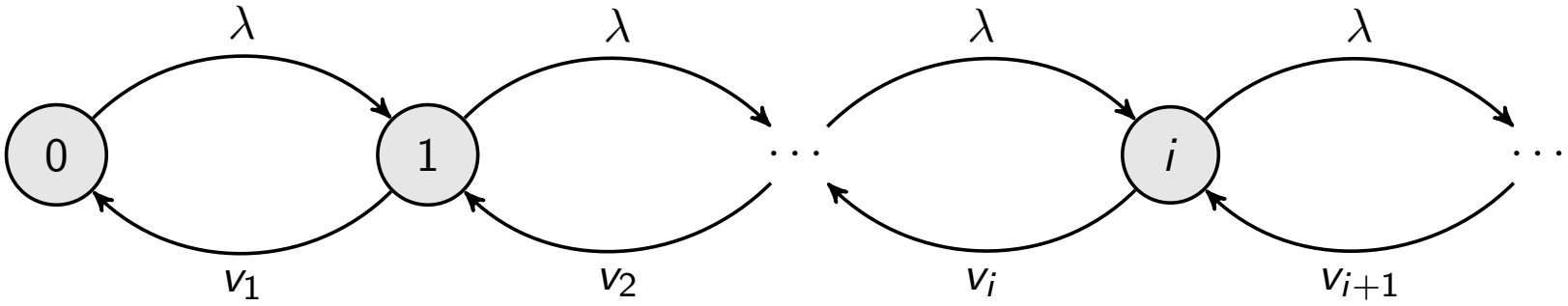
  $\beta$ is a weighting factor.

☞ Determine $v_n$, the optimal energy consumption and the sojourn time

# Markov chain model



# Markov chain model




From the formula of birth-death processes,

$$\pi_n = \pi_0 \prod_{j=1}^{n} \frac{\lambda}{\nu_j},$$

where $\pi_0 = \left(1 + \sum_{n \geq 1} \prod_{j=1}^{n} \frac{\lambda}{s_j}\right)^{-1}$.

# Optimization problem

Find $\mathbf{s}^* \equiv (s_n^*)_{n \geq 0}$ which

$$\begin{aligned} \text{minimize} \quad & \sum_{n \geq 0} n\pi_n(\mathbf{s}) + \beta \sum_{n=0}^{\infty} \pi_n(\mathbf{s}) s_n^{\alpha} \qquad \text{(OPT:D)} \\ \text{subject to} \quad & \mathbf{s} \geq 0. \end{aligned}$$

# Optimization problem

Find $\mathbf{s}^* \equiv (s_n^*)_{n\geq 0}$ which

$$\begin{aligned} \text{minimize} \quad & \sum_{n\geq 0} n\pi_n(\mathbf{s}) + \beta \sum_{n=0}^{\infty} \pi_n(\mathbf{s}) s_n^{\alpha} \\ \text{subject to} \quad & \mathbf{s} \geq 0. \end{aligned} \tag{OPT:D}$$

which is equivalent to

$$\begin{aligned} \text{minimize} \quad & \sum_{i\geq 0} n\pi_n + \beta \sum_{n=0}^{\infty} \pi_n s_n^{\alpha} \\ \text{subject to} \quad & \pi_n \lambda = \pi_{n+1} s_{n+1} \\ & \sum_{n\geq 0} \pi_n = 1, \mathbf{s} \geq 0, \pi \geq 0. \end{aligned} \tag{OPT:DM}$$

# Optimal speed

## Theorem

*Let* $s_n^*$ *be solution of* (OPT:DM). *Then,*

1. $$(1-\alpha)(s_n^*)^\alpha + \beta n + \alpha(s_{n+1}^*)^{\alpha-1}\lambda + \nu = 0,$$

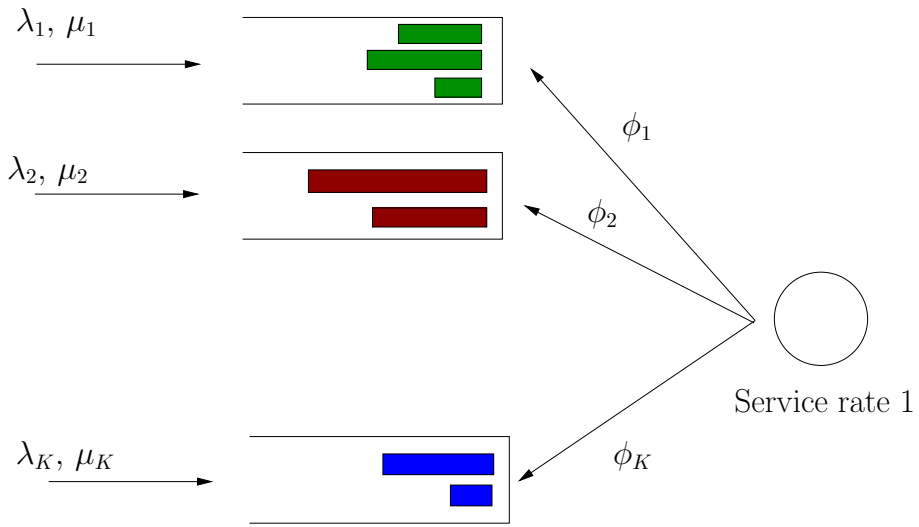
2. $$\lim_{n\to\infty} \frac{s_n^*}{n^{1/\alpha}} = c_1,$$

   *where* $c_1 > 0$

- (*cf.* Bansal *et al.* for the static case and Wierman *et al.* for the dynamic case).
- Different from DVFS currently implemented in processors.

# Scheduling in multi-class queues

# Scheduling




- $K$ classes.
- class $i$ has a holding cost of $c_i$
- fraction $\phi_i$ is given to class $i$.

# Scheduling

- ☞ Minimize weighted mean sojourn time

$$\sum_i \lambda_i c_i \mathbb{E}[T_i]$$

# Scheduling

☞ Minimize weighted mean sojourn time

$$\sum_i \lambda_i c_i \mathbb{E}[T_i]$$

- Only mean service times are known
  $\Rightarrow$ give priority to class with smallest $c_i \mu_i$ ($c\mu$ rule, see Cox and Smith)

# Scheduling

☞ Minimize weighted mean sojourn time

$$\sum_i \lambda_i c_i \mathbb{E}[T_i]$$

- Only mean service times are known
  $\Rightarrow$ give priority to class with smallest $c_i \mu_i$ ($c\mu$ rule, see Cox and Smith)
- Service requirement is known for each task
  $\Rightarrow$ shortest remaining service time (SRPT) is optimal even for $G/G/1$ (see Schrage)

# Scheduling

☞ Minimize weighted mean sojourn time

$$\sum_i \lambda_i c_i \mathbb{E}[T_i]$$

- Only mean service times are known
  $\Rightarrow$ give priority to class with smallest $c_i \mu_i$ ($c\mu$ rule, see Cox and Smith)
- Service requirement is known for each task
  $\Rightarrow$ shortest remaining service time (SRPT) is optimal even for $G/G/1$ (see Schrage)
- Distribution is known: Depends on the hazard rate (Poisson arrivals)

# Scheduling

☞ Minimize weighted mean sojourn time

$$\sum_i \lambda_i c_i \mathbb{E}[T_i]$$

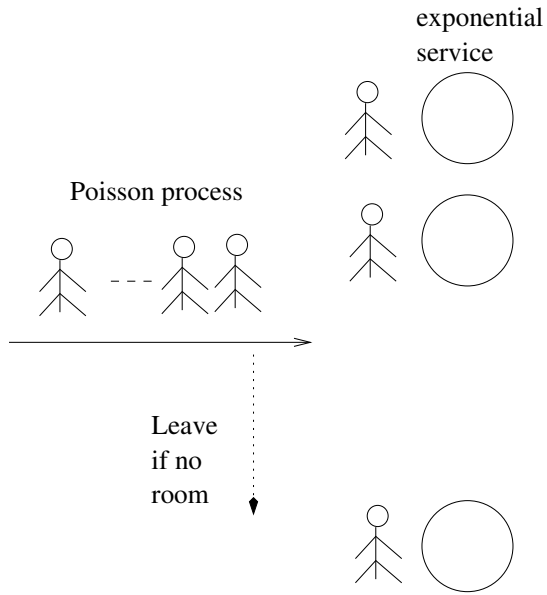
- Only mean service times are known
  $\Rightarrow$ give priority to class with smallest $c_i \mu_i$ ($c\mu$ rule, see Cox and Smith)
- Service requirement is known for each task
  $\Rightarrow$ shortest remaining service time (SRPT) is optimal even for $G/G/1$ (see Schrage)
- Distribution is known: Depends on the hazard rate (Poisson arrivals)

☞ Which service discipline minimizes the mean sojourn time?

Answer depends upon on the model

# Analysis of queues: multi server case

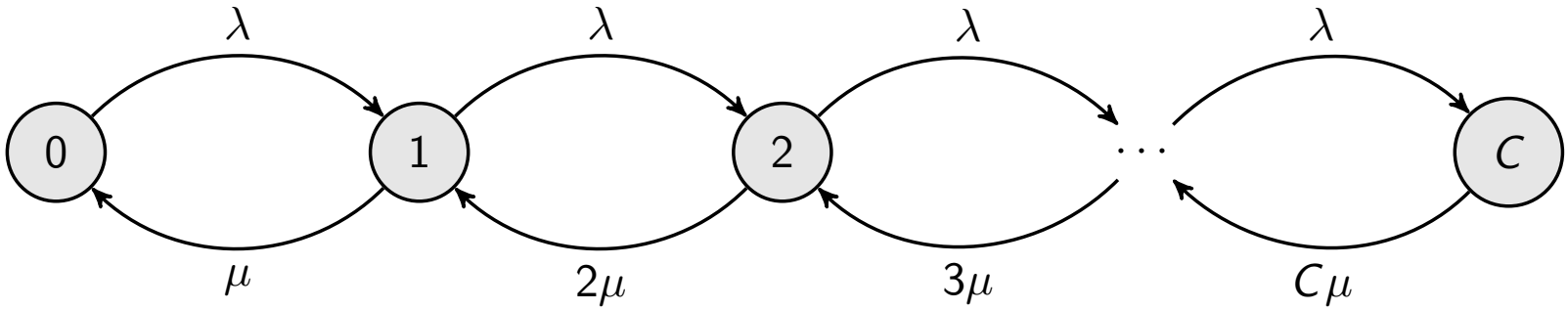
# M/M/C/C or loss system



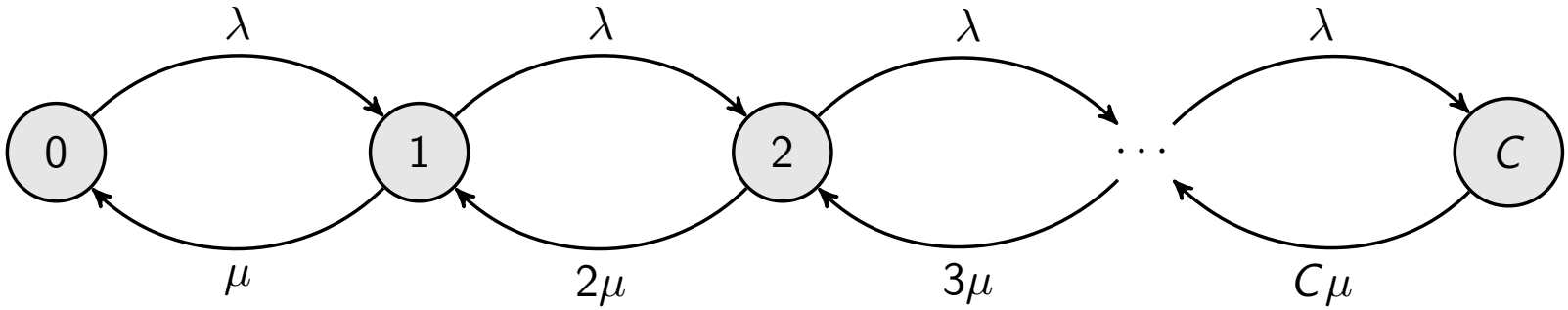

- C servers and no waiting.
- Used for dimensioning telephone network.

☞ Performance measure: blocking probability

# M/M/C/C or loss system



# M/M/C/C or loss system

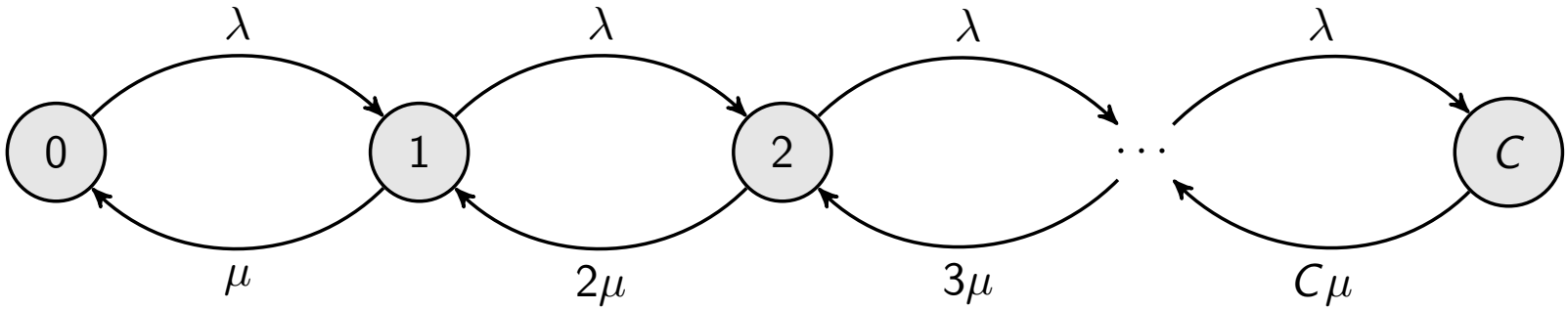



- Steady-state probabilities: $\pi_i = \pi_0 \frac{\rho^i}{i!}$
- Blocking probability:

$$\pi_C = \pi_0 \frac{\rho^C}{C!}$$

Known as the Erlang-B formula.

# M/M/C/C or loss system




- Steady-state probabilities: $\pi_i = \pi_0 \frac{\rho^i}{i!}$
- Blocking probability:

$$\pi_C = \pi_0 \frac{\rho^C}{C!}$$

Known as the Erlang-B formula.

☞ Is insensitive!

# Asymtotics of the Erlang-B formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$\pi_C \sim e^{-\beta C}$$

# Asymtotics of the Erlang-B formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$\pi_C \sim e^{-\beta C}$$

High quality but low efficiency

# Asymtotics of the Erlang-B formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$\pi_C \sim e^{-\beta C}$$

  High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$

$$\pi_C \sim 1 - \gamma$$

# Asymtotics of the Erlang-B formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$\pi_C \sim e^{-\beta C}$$

  High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$

$$\pi_C \sim 1 - \gamma$$

  Low quality but high efficiency

# Asymtotics of the Erlang-B formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$\pi_C \sim e^{-\beta C}$$

High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$

$$\pi_C \sim 1 - \gamma$$

Low quality but high efficiency

- $C = \lambda + \gamma\sqrt{\lambda}$

$$\pi_C \sim \frac{1}{\sqrt{C}}$$

# Asymtotics of the Erlang-B formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$\pi_C \sim e^{-\beta C}$$

High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$

$$\pi_C \sim 1 - \gamma$$

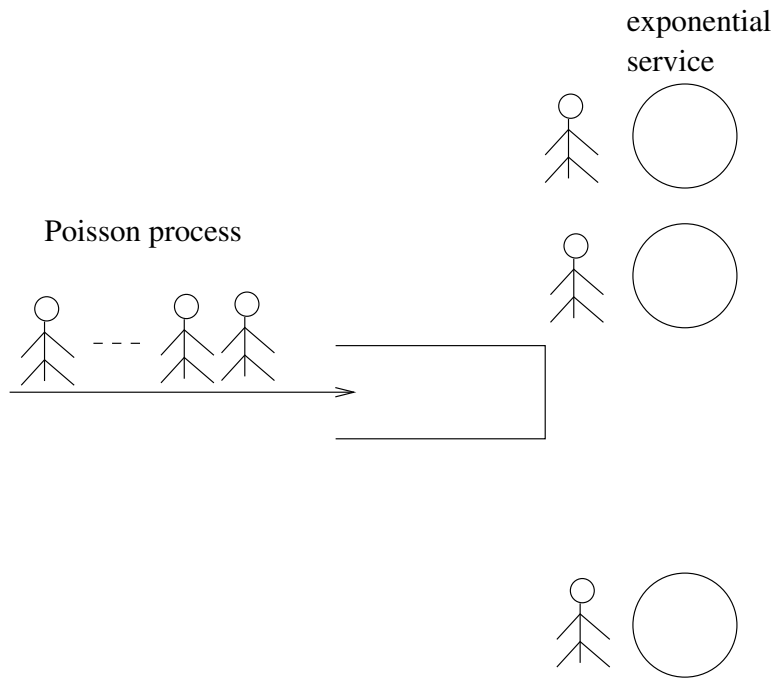
Low quality but high efficiency

- $C = \lambda + \gamma\sqrt{\lambda}$

$$\pi_C \sim \frac{1}{\sqrt{C}}$$

Compromise between quality and efficiency. (Erlang, Jagerman)

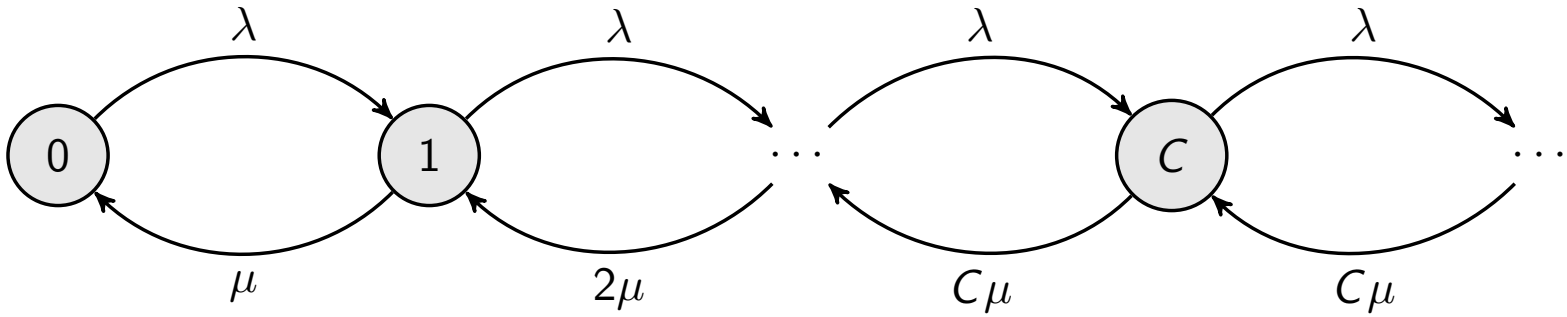
# M/M/C



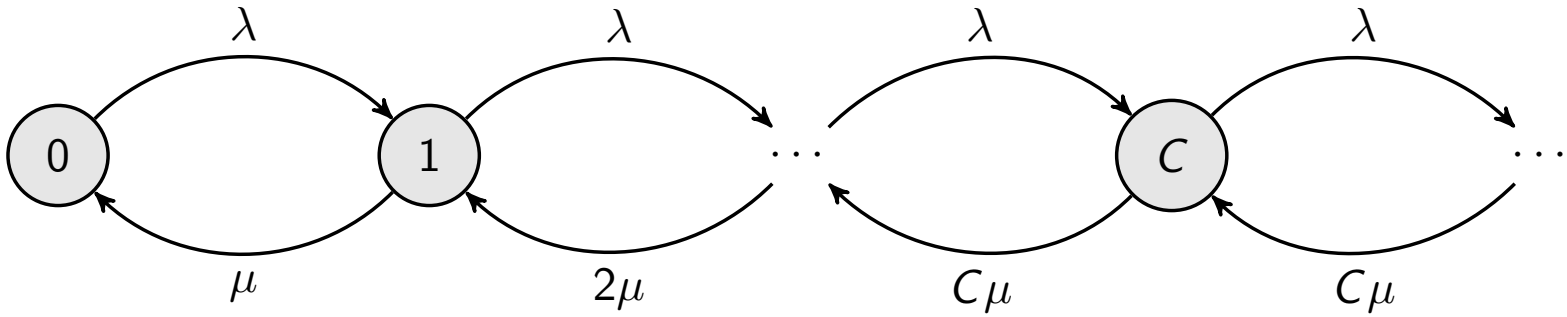

- C servers with single waiting queue.
- Call centers, airports.

☞ Performance measure: waiting time, probability of waiting

# M/M/C



# M/M/C

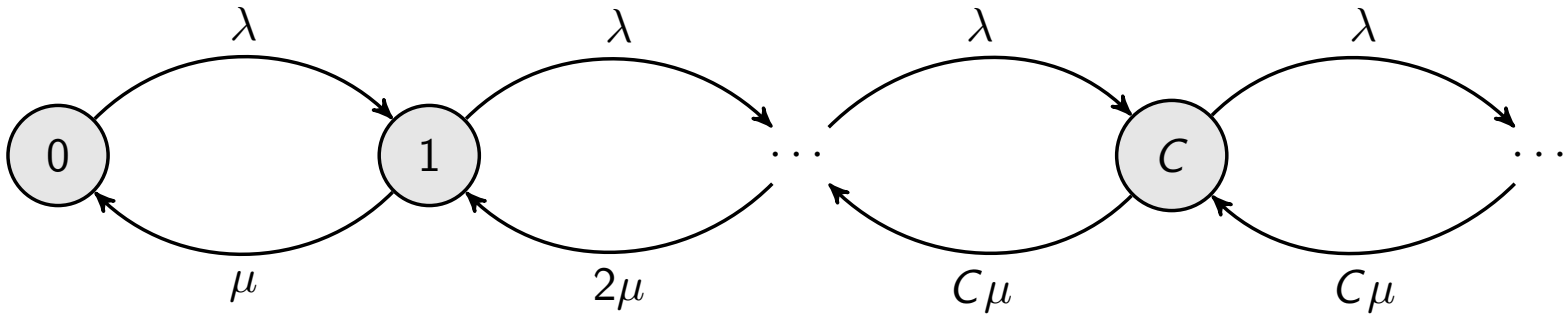



- Probability of waiting:

$$P_w = \sum_{i \geq C} \pi_i$$

Known as the Erlang-C formula. (Erlang)

# M/M/C




- Probability of waiting:

$$P_w = \sum_{i \geq C} \pi_i$$

  Known as the Erlang-C formula. (Erlang)

☞ Is **not** insensitive

# Asymtotics of the Erlang-C formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$P_w \sim e^{-\beta C}$$

# Asymtotics of the Erlang-C formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$P_w \sim e^{-\beta C}$$

High quality but low efficiency

# Asymtotics of the Erlang-C formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$P_w \sim e^{-\beta C}$$

High quality but low efficiency

# Asymtotics of the Erlang-C formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$P_w \sim e^{-\beta C}$$

  High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$ is unstable

# Asymtotics of the Erlang-C formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$P_w \sim e^{-\beta C}$$

  High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$ is unstable
- $C = \lambda + \gamma\sqrt{\lambda}$

$$P_w > 0$$

# Asymtotics of the Erlang-C formula

☞ How many trunk lines when $\lambda \to \infty$.

- $C = \gamma\lambda$, $\gamma > 1$

$$P_w \sim e^{-\beta C}$$

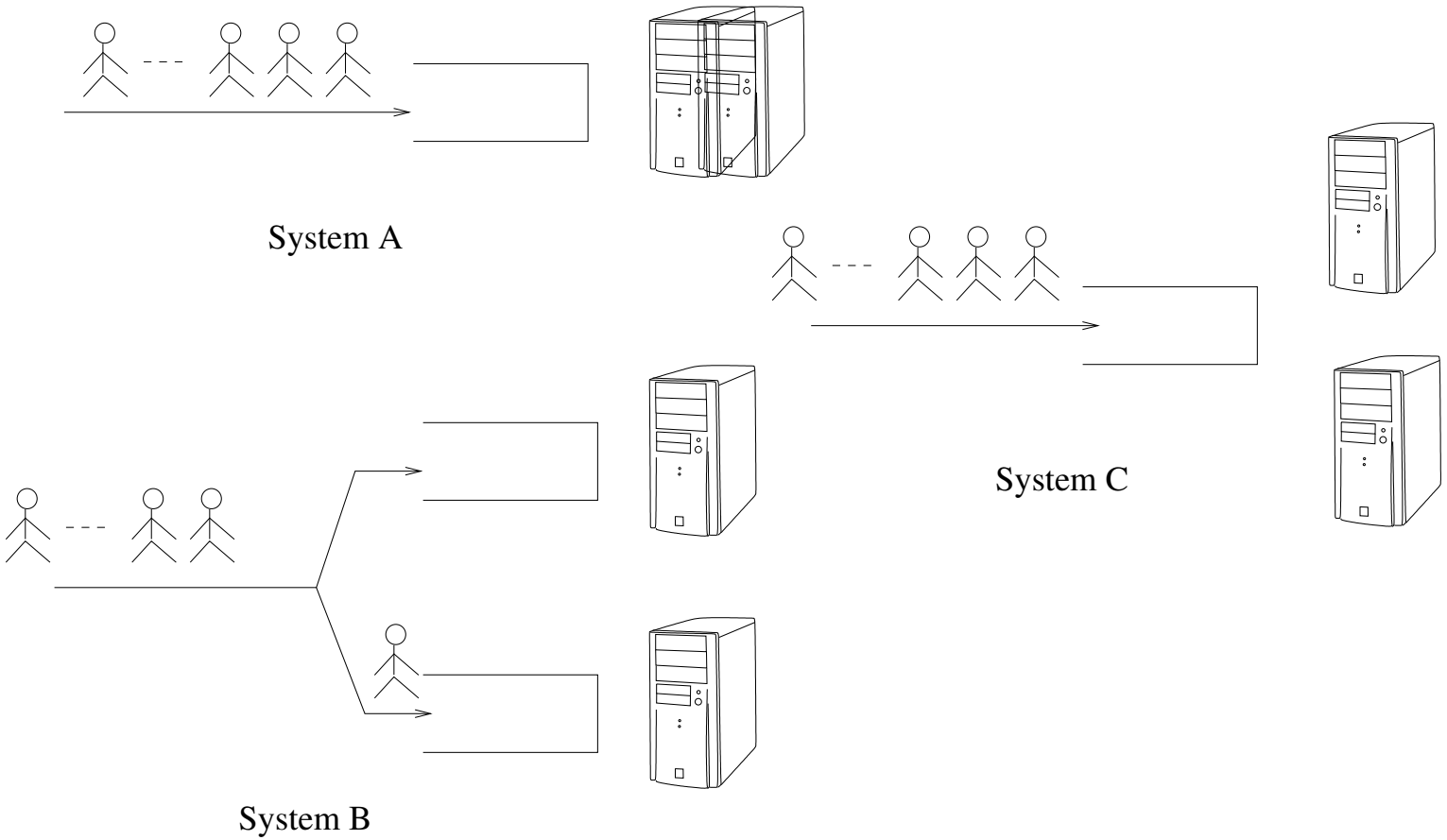
  High quality but low efficiency

- $C = \gamma\lambda$, $\gamma < 1$ is unstable
- $C = \lambda + \gamma\sqrt{\lambda}$

$$P_w > 0$$

  Compromise between quality and efficiency. (Halfin-Whitt regime)

# Comparison of systems




- Which is better for mean sojourn time?

# Comparison of systems

- Assume: Poisson arrivals rate $2\lambda$. Exponential service times of rate 1.

# Comparison of systems

- Assume: Poisson arrivals rate $2\lambda$. Exponential service times of rate 1.
- System A: $M/M/1$ with Poission arrivals of rate $2\lambda$ and service rate $2 \Rightarrow \rho_A = \lambda$

$$\mathbb{E}[T_A] = \frac{1}{2\lambda}\frac{\rho_A}{1-\rho_A} = \frac{1}{2}\frac{1}{1-\lambda}$$

# Comparison of systems

- Assume: Poisson arrivals rate $2\lambda$. Exponential service times of rate 1.
- System A: $M/M/1$ with Poission arrivals of rate $2\lambda$ and service rate $2 \Rightarrow \rho_A = \lambda$

$$\mathbb{E}[T_A] = \frac{1}{2\lambda}\frac{\rho_A}{1-\rho_A} = \frac{1}{2}\frac{1}{1-\lambda}$$

- System B: 2 $M/M/1$ queues each with Poission arrivals of rate $\lambda$ and service rate $1 \Rightarrow \rho_B = \lambda$

$$\mathbb{E}[T_B] = \frac{1}{\lambda}\frac{\rho_B}{1-\rho_B} = \frac{1}{1-\lambda}$$

# Comparison of systems

- Assume: Poisson arrivals rate $2\lambda$. Exponential service times of rate 1.
- System A: $M/M/1$ with Poission arrivals of rate $2\lambda$ and service rate 2 $\Rightarrow \rho_A = \lambda$

$$\mathbb{E}[T_A] = \frac{1}{2\lambda}\frac{\rho_A}{1-\rho_A} = \frac{1}{2}\frac{1}{1-\lambda}$$

- System B: 2 $M/M/1$ queues each with Poission arrivals of rate $\lambda$ and service rate 1 $\Rightarrow \rho_B = \lambda$

$$\mathbb{E}[T_B] = \frac{1}{\lambda}\frac{\rho_B}{1-\rho_B} = \frac{1}{1-\lambda}$$

- $\mathbb{E}[T_A] = \frac{1}{2}\mathbb{E}[T_B]$.

# Comparison of systems

- Assume: Poisson arrivals rate $2\lambda$. Exponential service times of rate 1.
- System A: $M/M/1$ with Poission arrivals of rate $2\lambda$ and service rate $2 \Rightarrow \rho_A = \lambda$

$$\mathbb{E}[T_A] = \frac{1}{2\lambda}\frac{\rho_A}{1-\rho_A} = \frac{1}{2}\frac{1}{1-\lambda}$$

- System C: $M/M/2$ queue with Poission arrivals of rate $2\lambda$ and service rate 1 per server

$$\mathbb{E}[T_A] = \frac{1+\rho_A}{2}\mathbb{E}[T_C]$$

# Comparison of systems

- Assume: Poisson arrivals rate $2\lambda$. Exponential service times of rate 1.
- System A: $M/M/1$ with Poission arrivals of rate $2\lambda$ and service rate $2 \Rightarrow \rho_A = \lambda$

$$\mathbb{E}[T_A] = \frac{1}{2\lambda}\frac{\rho_A}{1-\rho_A} = \frac{1}{2}\frac{1}{1-\lambda}$$

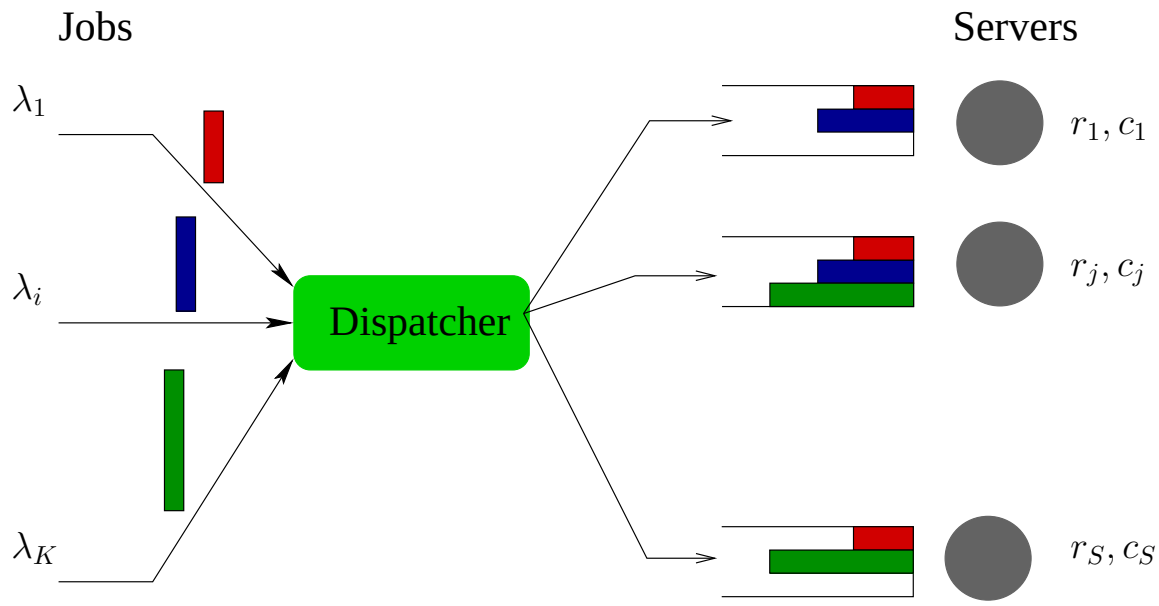
- System C: $M/M/2$ queue with Poission arrivals of rate $2\lambda$ and service rate 1 per server

$$\mathbb{E}[T_A] = \frac{1+\rho_A}{2}\mathbb{E}[T_C]$$

- $\mathbb{E}[T_A] < \mathbb{E}[T_C]$.

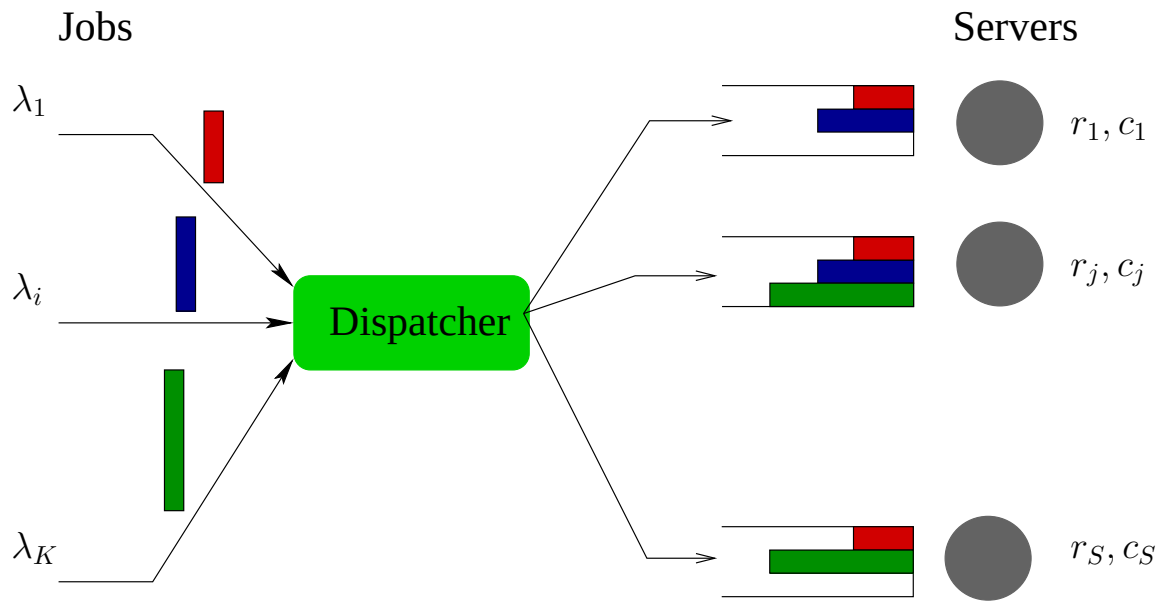
# Load balancing

# Insensitive load balancing




- $K$ classes. Poisson arrivals of rate $\lambda_i$. $\sigma_i$: fixed service time
- $S$ servers. PS discipline
  - $r_j$: speed of server $j$
  - $c_j$: cost rate of server $j$

# Insensitive load balancing




- $K$ classes. Poisson arrivals of rate $\lambda_i$. $\sigma_i$: fixed service time
- $S$ servers. PS discipline
  - $r_j$: speed of server $j$
  - $c_j$: cost rate of server $j$
- Bernoulli routing: route an arrival of class $i$ to server $j$ with probability $p_{i,j}$

# Load balancing

☞ Objective: minimize the weighted sojourn time

$$\min_{\mathbf{p}} \sum_{j} c_j \frac{\rho_j}{1-\rho_j}$$

$$s.t. \quad \sum_{j=1}^{C} p_{ij} = 1, \quad i = 1, 2, ..., K;$$

$$p_{ij} \geq 0, \quad \forall i, j;$$

$$\rho_j = \sum_{i=1}^{K} p_{ij} \frac{\lambda_i \sigma_i}{r_j}$$

# Load balancing

☞ Objective: minimize the weighted sojourn time

$$\min_{\mathbf{p}} \sum_j c_j \frac{\rho_j}{1-\rho_j}$$

$$s.t. \quad \sum_{j=1}^{C} p_{ij} = 1, \quad i = 1, 2, ..., K;$$

$$p_{ij} \geq 0, \quad \forall i, j;$$

$$\rho_j = \sum_{i=1}^{K} p_{ij} \frac{\lambda_i \sigma_i}{r_j}$$

- Important assumptions: Poisson arrivals, PS discipline, Bernoulli routing

# Insensitive load balancing

☞structure of optimal policy

- Order the servers in increasing order of $\frac{c_j}{r_j}$

$$\frac{c_1}{r_1} \leq \frac{c_2}{r_2} \cdots \leq \frac{c_S}{r_S}$$

# Insensitive load balancing

☞structure of optimal policy

- Order the servers in increasing order of $\frac{c_j}{r_j}$

$$\frac{c_1}{r_1} \leq \frac{c_2}{r_2} \cdots \leq \frac{c_S}{r_S}$$

- There exists a $s^* \leq S$ such that $\rho_j > 0 \Leftrightarrow j \leq s^*$

# Insensitive load balancing

☞structure of optimal policy

- Order the servers in increasing order of $\frac{c_j}{r_j}$

$$\frac{c_1}{r_1} \leq \frac{c_2}{r_2} \cdots \leq \frac{c_S}{r_S}$$

- There exists a $s^* \leq S$ such that $\rho_j > 0 \Leftrightarrow j \leq s^*$
- The optimal loads on server $j \leq s^*$ is

$$\rho_j^* = 1 - \sqrt{\frac{c_j}{r_j}}\gamma$$

# Insensitive load balancing

☞structure of optimal policy

- Order the servers in increasing order of $\frac{c_j}{r_j}$

$$\frac{c_1}{r_1} \leq \frac{c_2}{r_2} \cdots \leq \frac{c_S}{r_S}$$

- There exists a $s^* \leq S$ such that $\rho_j > 0 \Leftrightarrow j \leq s^*$
- The optimal loads on server $j \leq s^*$ is

$$\rho_j^* = 1 - \sqrt{\frac{c_j}{r_j}}\gamma$$

- There exists an optimal class-independent routing

$$p_{i,j}^* = \frac{\rho_j^* r_j}{\sum_{k=1}^{K} \lambda_k \sigma_k}$$

# Efficient load balancing

- ☞ Can we do better? Round robin is better for homogenous servers
  - If whole state is known: JSQ is optimal but for homogenous servers
  - Partial state is known: JSQ(d) (Supermarket model)
  - Pull based mechanisms: JIQ
- ☞ Difficult to analyze. Use of large server asymptotics.