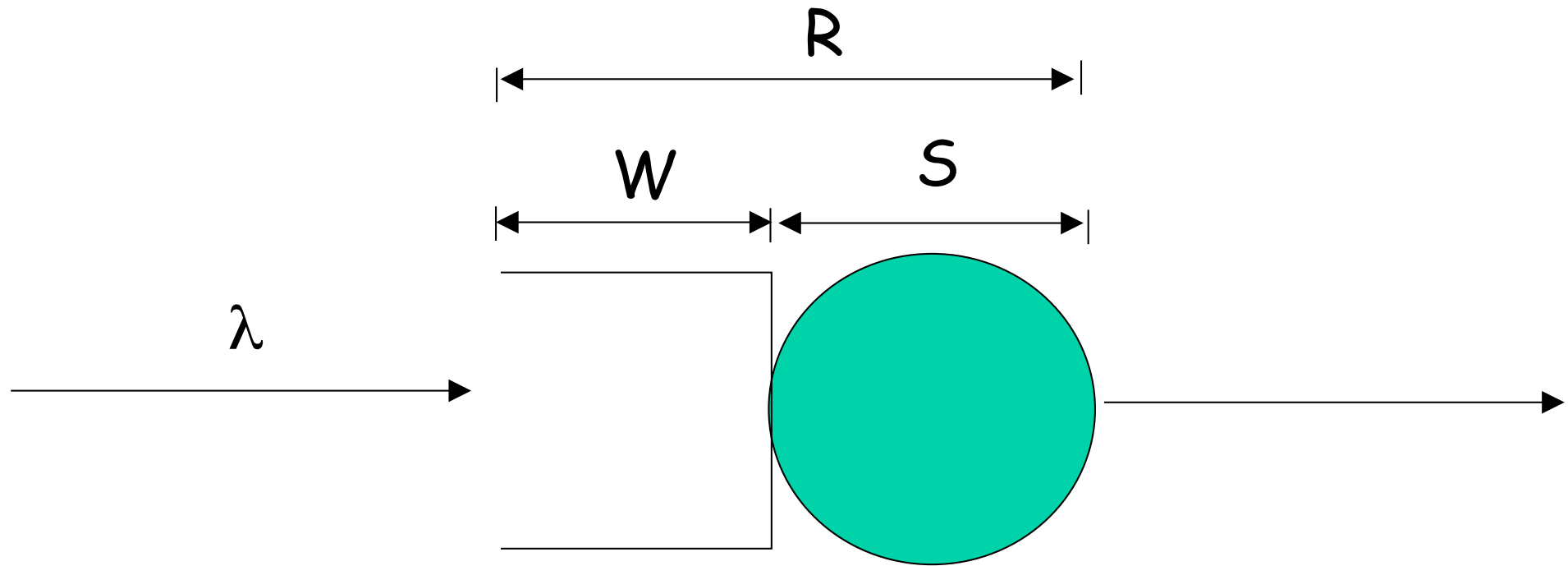


# Modelos Probabilísticos

## Filas $M/M/1$ , $M/G/1$

Profa. Jussara M. Almeida  
1º Semestre de 2014

# Modelos Probabilísticos de Filas



Notação de Kendall

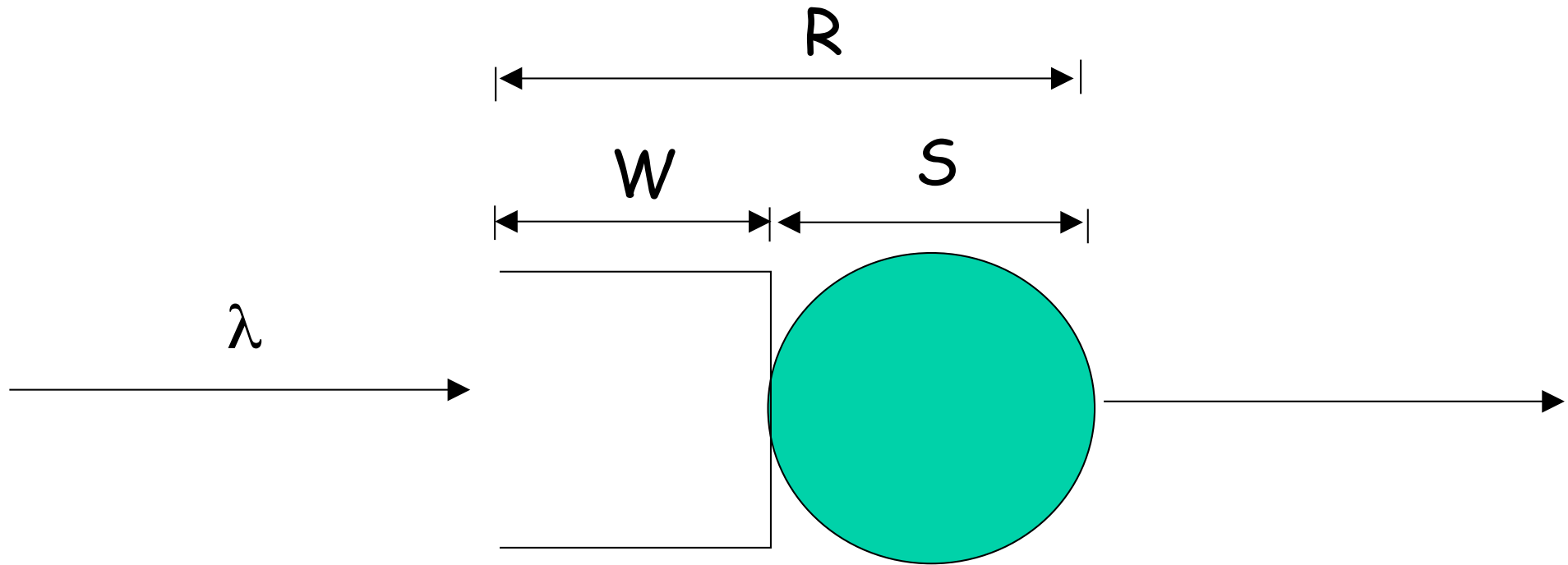
Fila  $G / G / 1$

1 = um único servidor

Distribuição dos tempos entre chegadas de clientes  
 $G$  = distribuição genérica

Distribuição do tempo de serviço  
 $G$  = distribuição genérica

# Filas G/G/1



$$E(N_w) = \lambda \times E(W)$$

$$E(N) = \lambda \times E(R)$$

$$E(N_s) = \rho = \lambda \times E(S)$$

$\rho$  = utilização do servidor

Probabilidade servidor ocioso  $p_0$ :

$$p_0 = 1 - \rho = 1 - \lambda \times E(S)$$

# Exemplo

O tempo entre chegadas de pacotes em um canal de comunicação de 100 Mbps segue uma distribuição Gamma com média de 5 mseg. A transmissão de um pacote pelo canal segue uma distribuição normal com parâmetros  $\mu = 3$  mseg e  $\sigma = 1$  mseg. Qual a largura de banda ocupada? Qual a probabilidade do canal estar vazio?

$$\lambda = 1 / \text{tempo médio entre chegadas} = 1/5 \text{ mseg}$$

$$E(S) = 3 \text{ mseg}$$

$$\rho = \lambda \times E(S) = 1/5 \times 3 = 0.6 = 60\%$$

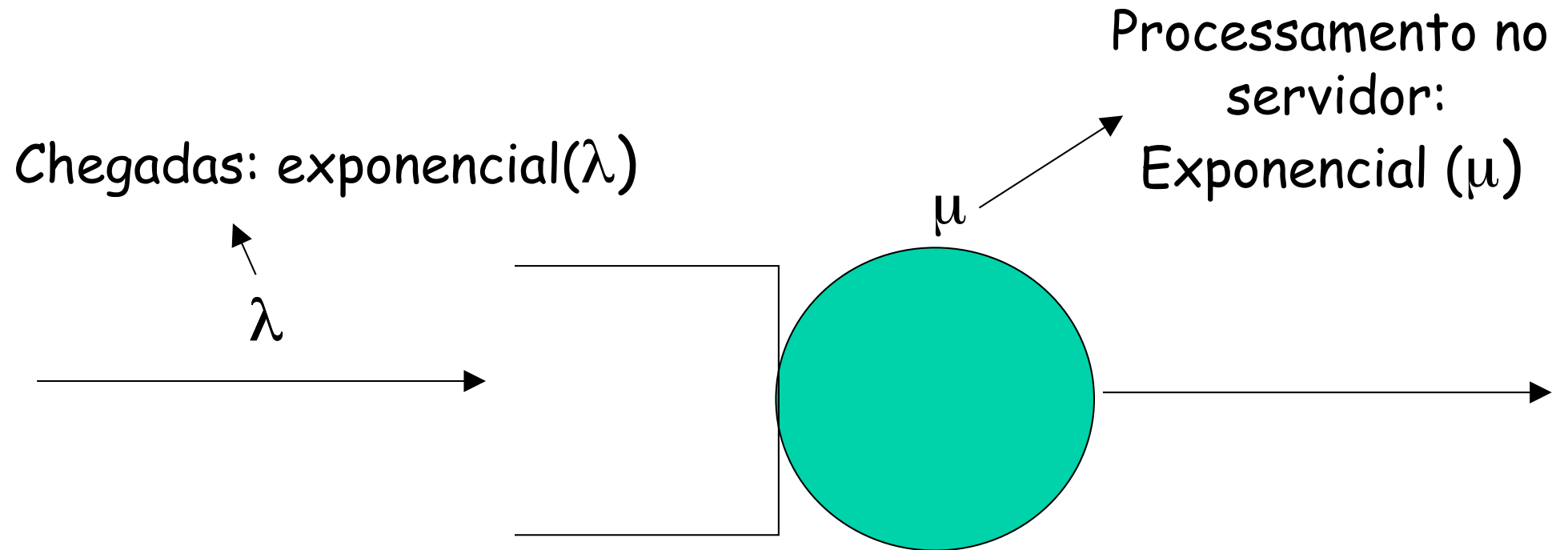
$$\text{Banda ocupada} = 0.6 \times 100 \text{ Mbps} = 60 \text{ Mbps.}$$

$$p_0 = 1 - \rho = 0.4 \Rightarrow \text{único resultado exato conhecido} \\ \text{para filas } G/G/1 \\ \text{(aproximações vide livro [MeAD04])}$$

# A Fila M/M/1

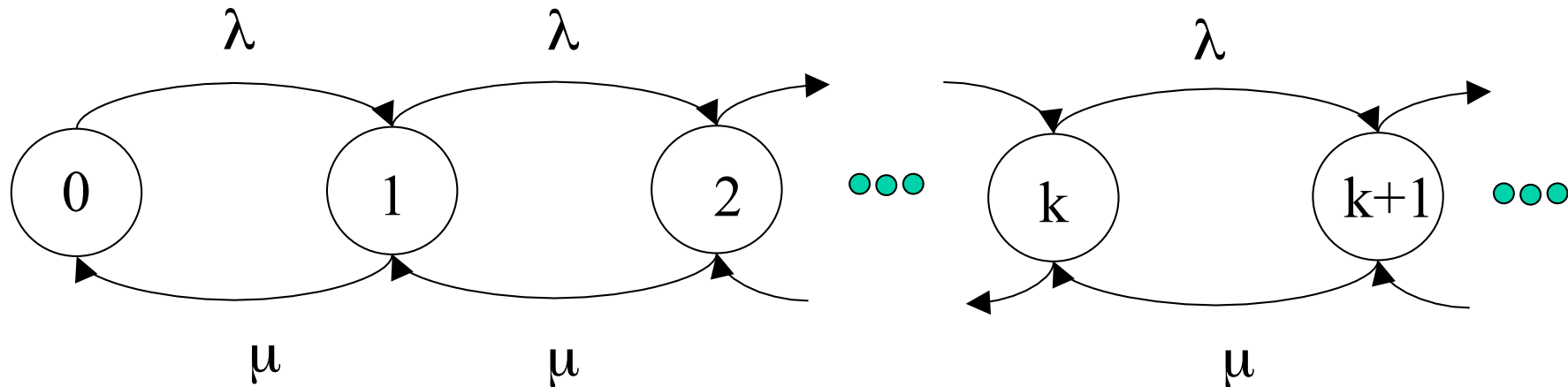
- Caso especial das filas G/G/1 mais comumente estudado
  - Caso clássico
- O "M" vem de Markoviano ou memoryless
  - Tempos entre chegadas de clientes: distr. exponencial ( $\lambda$ )
  - Tempos de serviço: distr. exponencial ( $\mu$ )
  - Um único servidor

# A Fila M/M/1



# A Fila M/M/1

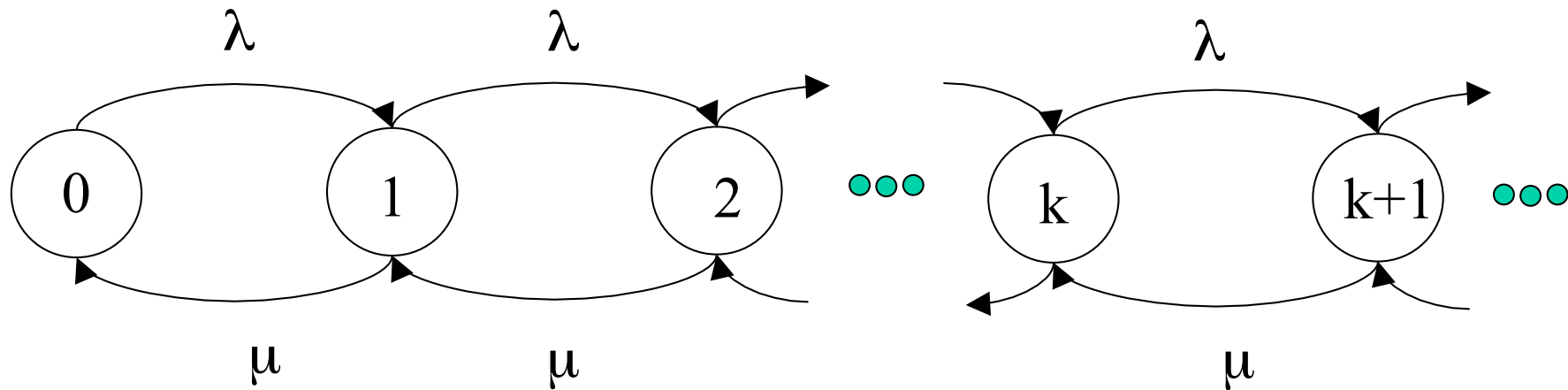
- Fila M/M/1 pode ser solucionada através de um modelo de Markov simples
  - Processo de Birth and Death:  $\lambda_k = \lambda$  e  $\mu_k = \mu$ ,  $k=0,1,\dots$



Clientes são servidos na ordem de chegada: política FCFS

Premissa importante: espaço na fila é infinito  
(premissa pode ser relaxada: buffer finito)

# A Fila M/M/1: Solução



$$\pi \times Q = 0$$

$$[\pi_0 \quad \pi_1 \quad \dots] \times \begin{bmatrix} -\lambda & \lambda & 0 & \dots \\ \mu & -(\lambda + \mu) & \lambda & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} = [0 \quad 0 \quad \dots]$$

$$\pi_0 + \pi_1 + \dots = 1$$



# A Fila M/M/1: Solução

$$\pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \quad k \geq 0$$

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^k}$$

⇒ Esta soma converge sse  $\lambda/\mu < 1$   
⇒ senão,  $\pi_0 = 0$  :  
não tem caso estado estacionário

$\lambda < \mu$ : condição necessária para cadeia ter estado estacionário

# A Fila M/M/1: Solução

Se  $\lambda < \mu$ :

$$\pi_0 = \frac{1}{1 + \frac{\lambda/\mu}{1 - \lambda/\mu}} = 1 - \lambda/\mu$$

$$\rho = \lambda/\mu \Rightarrow \pi_0 = 1 - \rho$$

$$\pi_k = (1 - \rho)\rho^k \Rightarrow \text{Distribuição Geométrica}$$

$\pi_k$  = probabilidade de haver k clientes na fila

$\pi_k$  depende de  $\lambda$  e  $\mu$  apenas através da razão  $\rho$  das duas taxas, isto é, da utilização do sistema

# A Fila M/M/1: Solução

$E(N)$  = número médio de clientes no sistema

$$\begin{aligned} E(N) &= \sum_{k=0}^{\infty} k \pi_k = (1 - \rho) \sum_{k=0}^{\infty} k \rho^k \\ &= (1 - \rho) \rho \frac{\partial}{\partial \rho} \sum_{k=0}^{\infty} \rho^k = (1 - \rho) \rho \frac{\partial}{\partial \rho} \frac{1}{1 - \rho} \\ &= (1 - \rho) \rho \frac{1}{(1 - \rho)^2} = \frac{\rho}{1 - \rho} \end{aligned}$$

$\sigma_N^2$  = variância do número de clientes no sistema

$$\sigma_N^2 = \sum_{k=0}^{\infty} (k - E(N))^2 \pi_k = \frac{\rho}{(1 - \rho)^2}$$

# A Fila M/M/1: Solução

Probabilidade de haver pelo menos k clientes no sistema

$$\begin{aligned}P[\geq k \text{ no sistema}] &= \sum_{i=k}^{\infty} \pi_i = \sum_{i=k}^{\infty} (1 - \rho) \rho^i \\&= (1 - \rho) \sum_{i=0}^{\infty} \rho^{i+k} \\&= (1 - \rho) \rho^k \sum_{i=0}^{\infty} \rho^i \\&= (1 - \rho) \rho^k \frac{1}{1 - \rho} = \rho^k\end{aligned}$$

# A Fila M/M/1: Solução

Tempo de resposta médio

$$E(R) = \frac{E(N)}{\lambda} = \left( \frac{\rho}{1 - \rho} \right) \times \frac{1}{\lambda} = \frac{\cancel{1}/\mu}{1 - \rho} = \frac{1}{\mu(1 - \rho)}$$

Tempo de resposta

$$R = S + W$$

$$S \sim \text{exponencial}(\mu) \Rightarrow P(S \leq s) = 1 - e^{-\mu \times s}$$

$$P(W \leq w) = 1 - \rho \times e^{-\mu(1-\rho)w}$$

$$P(R \leq r) = 1 - e^{-\mu(1-\rho)r}$$

# Exemplo

Um servidor de arquivos recebe requisições de um processo Poisson à taxa de 30 req/s. Dados medidos no servidor indicam que o coeficiente de variação do tempo de serviço de uma requisição é muito próximo de 1 e que o tempo médio de serviço é de 15 msec. Qual a utilização e tempo médio de resposta do servidor? Qual a probabilidade do tempo de resposta ser superior a 100 msec?

processo Poisson  $\rightarrow$  tempo entre chegadas : exponencial( $\lambda$ )

$CV = 1 \rightarrow$  tempo de serviço  $\sim$  exponencial ( $\mu$ )

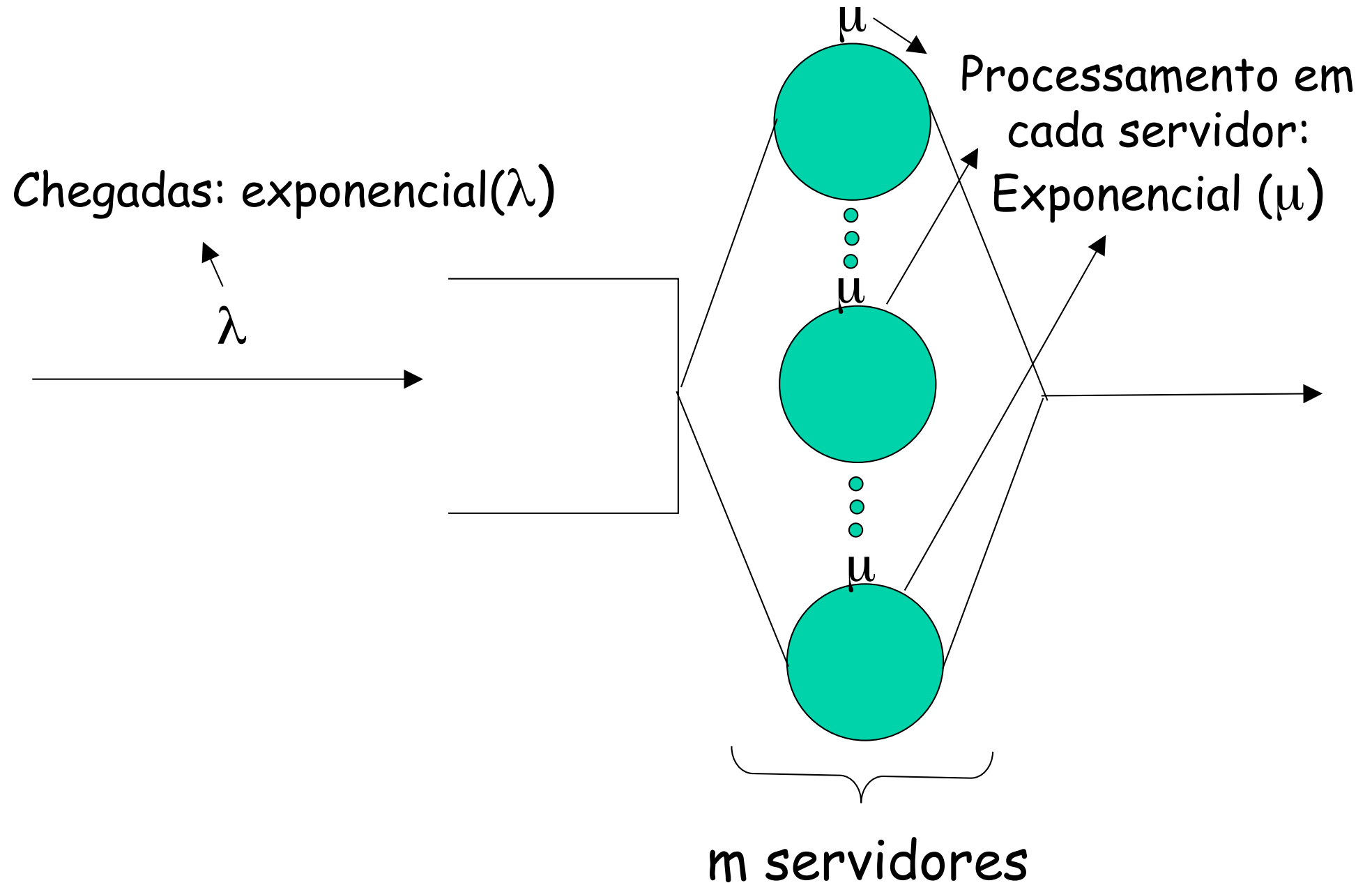
Fila M/M/1 com  $\lambda = 30$  req/s e  $\mu = 1/0.015 = 66.67$

Utilização  $\rho = \lambda/\mu = 30 \times 0.015 = 0.45 = 45\%$

$E(R) = 1/\mu(1 - \rho) = 0.027$  segundos

$\text{Prob}(R > 0.1) = 1 - P(R \leq 0.1) = 1 - (1 - e^{-\mu(1-\rho)0.1}) = e^{-\mu(1-\rho)0.1} = e^{-3.67}$

# A Fila M/M/m



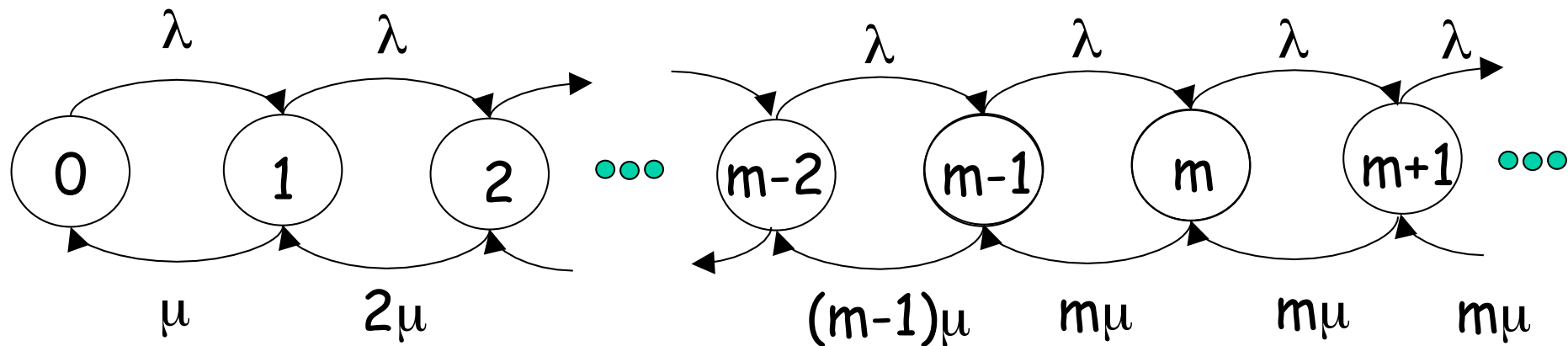
# A Fila M/M/m

Processo de Birth and Death:

$$\lambda_k = \lambda, \quad k = 0, 1, \dots$$

$$\mu_k = k \times \mu \quad 0 \leq k \leq m$$

$$m \times \mu \quad k > m$$



$\lambda < m\mu$ : condição necessária para cadeia ter estado estacionário

$\rho = \lambda/m\mu$  : utilização do sistema ( $\rho < 1$ )



# A Fila M/M/m: Solução

$$k \leq m : \pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!}$$

$$k > m : \pi_k = \pi_0 \prod_{i=0}^{m-1} \frac{\lambda}{(i+1)\mu} \prod_{j=m}^{k-1} \frac{\lambda}{m\mu} = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{m! m^{k-m}}$$

$$\pi_k = \begin{cases} \pi_0 \frac{(m\rho)^k}{k!} & k \leq m \\ \pi_0 \frac{(\rho)^k m^m}{m!} & k > m \end{cases} \quad \pi_0 = \frac{1}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right)}$$

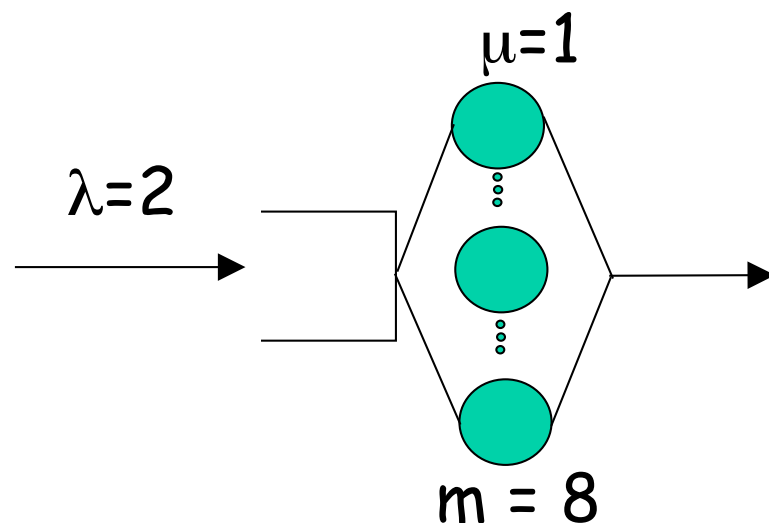
# A Fila M/M/m: Solução

$$\begin{aligned} \text{Pr ob}(\text{enf ileiramento}) &= \sum_{k=m}^{\infty} \pi_k = \sum_{k=m}^{\infty} \pi_0 \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}} \\ &= \frac{\left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right)}{\left[ \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right) \right]} \end{aligned}$$

# Exemplo

Um multiprocessador consiste de 8 processadores idênticos. Durante período de pico, pedidos de processamento chegam à taxa de 2 processos por hora. Cada processo executa, em média, por 1 hora e deixa o sistema. Após um cuidadoso estudo de caracterização de carga, você determinou que os tempos entre chegadas de novos processos e os tempos de processamento podem ser razoavelmente bem aproximados por distribuições exponenciais.

Sabendo que o escalonamento de processos é centralizado, qual a probabilidade de que haja pelo menos um processador ocioso.



$$\rho = \lambda / m\mu = 2 / 8 = 0.25$$

Utilização do sistema = 25%

$\rho < 1$ : existe estado estacionário

# Exemplo

$$\begin{aligned}\text{Prob}(\geq 1 \text{ ocioso}) &= \sum_{k=0}^{m-1} \pi_k = \sum_{k=0}^{m-1} \pi_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} \\ &= \pi_0 \sum_{k=0}^{m-1} (m\rho)^k \frac{1}{k!} \\ &= \frac{\sum_{k=0}^{m-1} (m\rho)^k \frac{1}{k!}}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right)}\end{aligned}$$

$\rho = 0.25$ ,  $m = 8 \rightarrow$  calcular  $\text{Prob}(\geq 1 \text{ processador ocioso})$

# A Fila M/M/1/K

Fila tem tamanho finito:

$K$  = número máximo de clientes que podem estar  
no sistema em qualquer instante

Ex :  $K-1$  = tamanho máximo do buffer de conexões

Processo de Birth and Death:

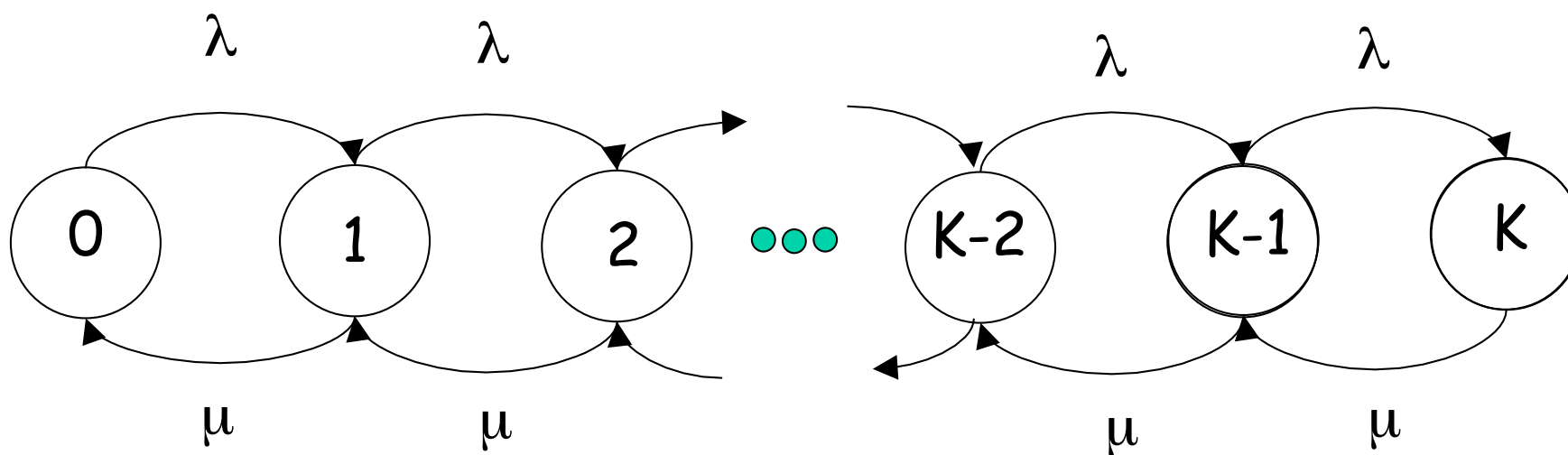
$$\lambda_k = \lambda, \quad 0 \leq k < K$$

$$0, \quad k \geq K$$

$$\mu_k = \mu \quad 1 \leq k \leq K$$

$\rho = \lambda/\mu < 1$ : condição para estado estacionário

# A Fila M/M/1/K



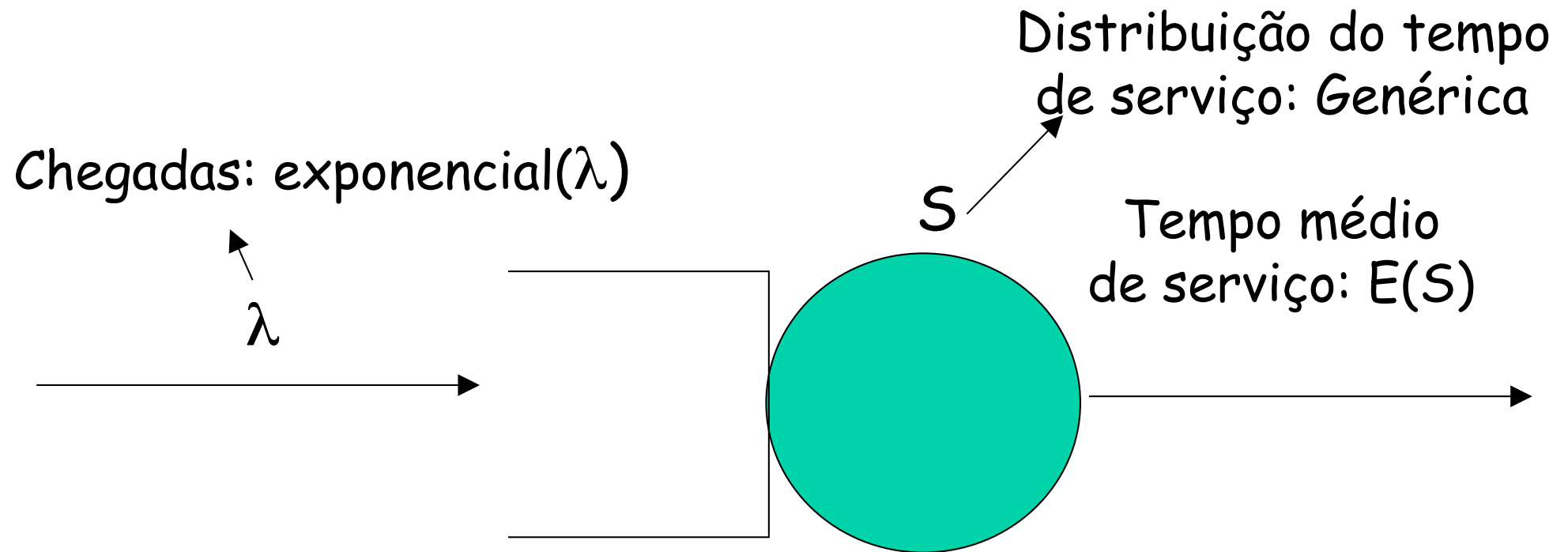
# A Fila M/M/1/K: Solução

$$k \leq K : \pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \quad k > K : \pi_k = 0$$

$$\pi_0 = \frac{1}{1 + \sum_{k=0}^K \left( \frac{\lambda}{\mu} \right)^k} = \frac{1 - \lambda/\mu}{1 - \left( \lambda/\mu \right)^{K+1}}$$

$$\pi_k = \begin{cases} \frac{1 - \rho}{1 - \rho^{K+1}} \rho^k & 0 \leq k \leq K \\ 0 & k > K \end{cases}$$

# A Fila M/G/1



$$\rho = \lambda / 1/E(S) = \lambda \times E(S)$$



# A Fila M/G/1: Solução

Resultados analíticos para fila M/G/1

Tempo médio de espera na fila:

$$E(W) = \frac{\rho \times E(S) \times (1 + C_s^2)}{2(1 - \rho)}$$

$C_s$  = coeficiente de variação da distribuição do tempo de serviço

$$\text{coeficiente de variação} = \frac{\text{desvio padrão}}{\text{média}}$$

Se  $C_s = 1 \Rightarrow S \sim \text{exponencial} \Rightarrow \text{fila M/M/1}$

# A Fila M/G/1: Solução

Resultados analíticos para fila M/G/1

Tempo médio de resposta

$$E(R) = E(S) + E(W) = E(S) + \frac{\rho \times E(S) \times (1 + C_s^2)}{2(1 - \rho)}$$

Número médio de clientes na fila de espera

$$N_w = \frac{\rho^2 (1 + C_s^2)}{2(1 - \rho)}$$

Número médio de clientes no sistema

$$N = \rho + \frac{\rho^2 (1 + C_s^2)}{2(1 - \rho)}$$

# Exemplo

Suponha que mensagens de e-mail cheguem a um servidor de e-mails a partir de um processo Poisson à taxa de 1.2 mensagens por segundo. Suponha também que 30% das mensagens sejam processadas em 100 ms, 50% em 300 ms e 20% em 2 s. Qual o tempo médio que uma mensagem espera na fila para ser processada? Qual o tempo médio de resposta? Qual o número médio de mensagens esperando na fila? Qual o número médio de mensagens no servidor de e-mails?

Tempo médio de processamento  $E(S)$

$$E(S) = 0.3 \times 0.1 + 0.5 \times 0.3 + 0.2 \times 2 = 0.030 + 0.15 + 0.4 = 0.58 \text{ s}$$

Utilização do servidor  $\rho = \lambda \times E(S) = 1.2 \times 0.58 = 0.696 = 69.6\%$

Qual o CV do tempo de serviço?

# Exemplo

Qual o CV do tempo de serviço?

$$CV = \sigma / E(S) \Rightarrow \sigma_s = \sqrt{E(S^2) - [E(S)]^2}$$

$$E(S^2) = 0.3 \times 0.1^2 + 0.5 \times 0.3^2 + 0.2 \times 2^2 = 0.848$$

$$\sigma_s = \sqrt{0.848 - 0.58^2} = \sqrt{0.848 - 0.336} = 0.715$$

$$CV = 0.715 / 0.58 = 1.233$$

Qual o tempo médio de espera?

$$E(W) = \frac{\rho E(S)(1 + CV^2)}{2(1 - \rho)} = \frac{0.696 \times 0.58 \times (1 + 1.233^2)}{2(1 - 0.696)} = 1.673s$$

# Exemplo

Qual o tempo médio de resposta?

$$E(R) = E(S) + E(W) = 0.58 + 1.673 = 2.253s$$

Qual o número médio de mensagens na fila de espera?

$$N_w = \lambda \times E(W) = 1.2 \times 1.63 = 2.008$$

Qual o número médio de mensagens no sistema?

$$N = N_w + \rho = 2.008 + 0.696 = 2.704$$

$$N = \lambda \times E(R) = 1.2 \times 2.253 = 2.704$$

# Exemplo

Qual a razão entre o tempo médio de espera em uma fila M/G/1 com tempo de serviço exponencialmente distribuído (M/M/1) e o tempo médio de espera em uma fila M/G/1 com tempos de serviço constantes (considere o mesmo tempo médio de serviço e mesma carga)?

M/G/1 com tempos de serviço exponenciais:  $CV = 1$

M/G/1 com tempos de serviço constantes:  $CV = 0$

$$E(W_{\text{exp}}) = \frac{\rho \times E(S) \times (1 + 1^2)}{2(1 - \rho)} = \frac{\rho}{(1 - \rho)} E(S)$$

$$E(W_{\text{constante}}) = \frac{\rho \times E(S) \times (1 + 0^2)}{2(1 - \rho)} = \frac{\rho}{2(1 - \rho)} E(S) = \frac{1}{2} E(W_{\text{exp}})$$

O tempo médio de espera na fila de um servidor com tempos exponenciais é, em média, duas vezes o tempo de espera em um servidor com tempo constante.

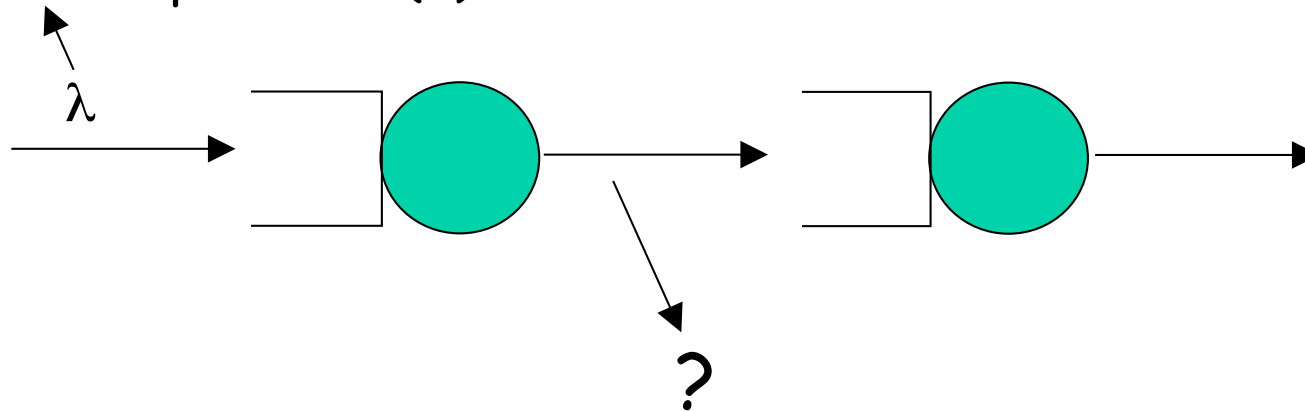
# Redes de Filas Markovianas

- Até agora, vimos como obter dados de desempenho de filas únicas
  - Cliente demanda um único serviço do sistema
  - Visão macro do sistema
- Precisamos agora estender resultados para **redes de filas**
  - Bem mais complexo
  - Questões chaves:
    - Estrutura topológica da rede
    - Natureza do fluxo estocástico de entrada em cada servidor (nó) do sistema

# Redes de Filas Markovianas

- Exemplo simples:

Chegadas: exponencial( $\lambda$ )

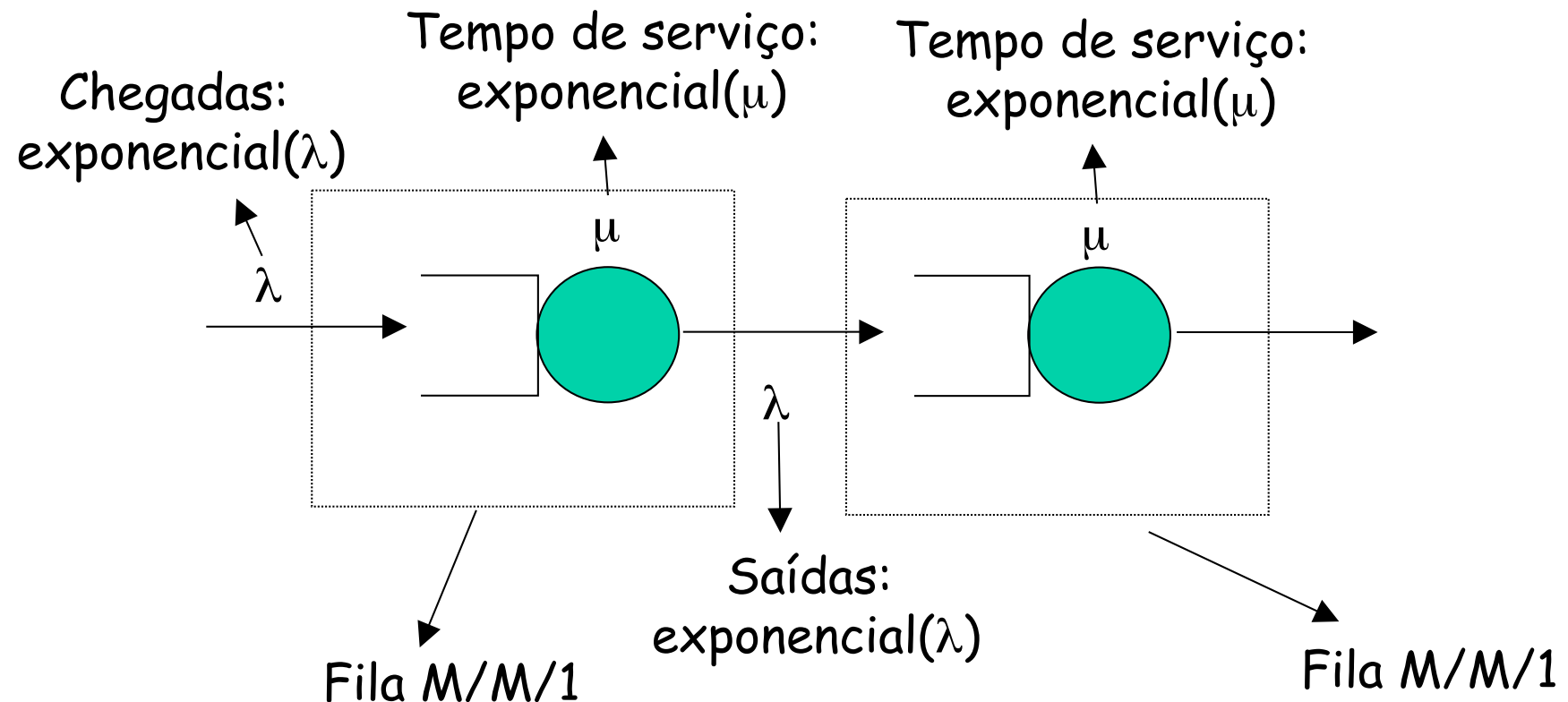




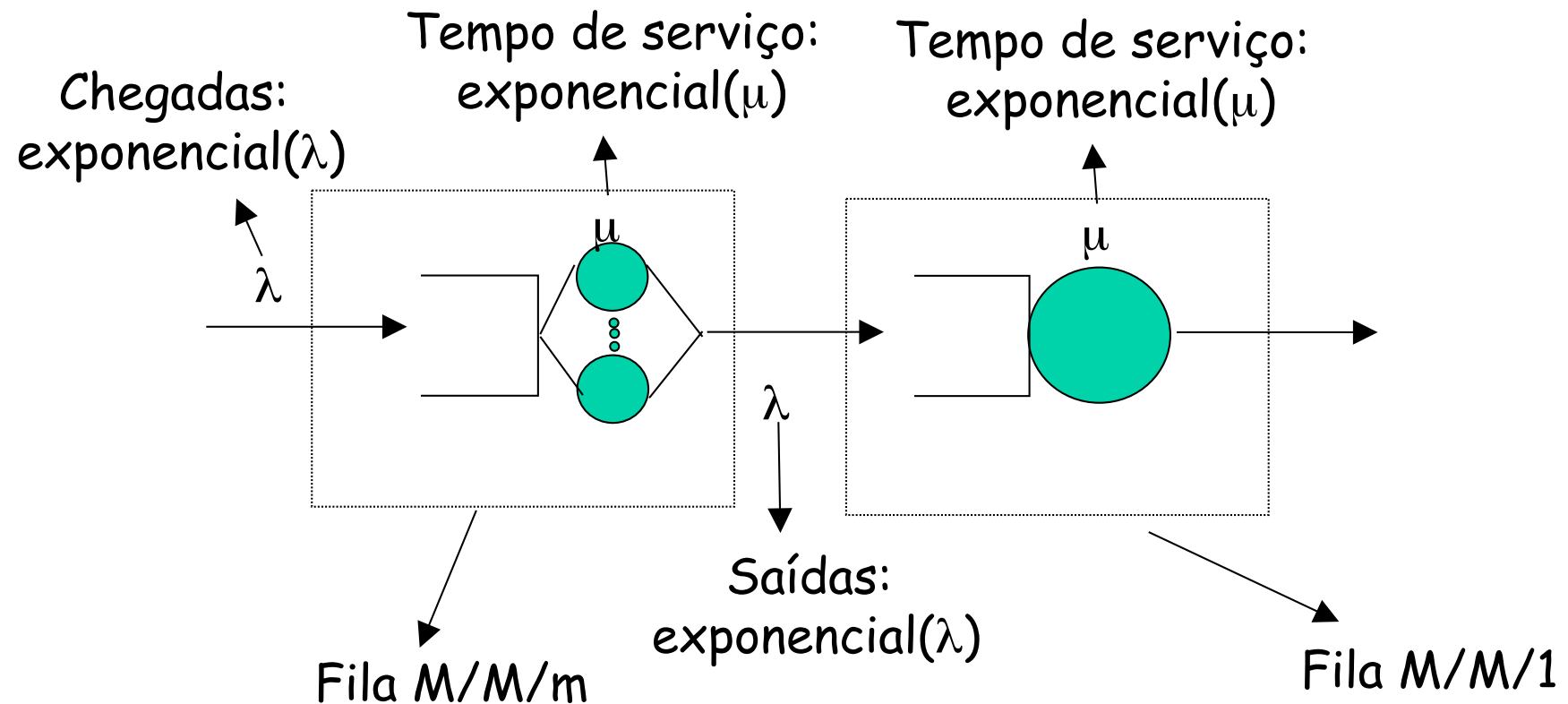
# Teorema de Burke

- Em um sistema de filas em estado estacionário, um processo Poisson alimentando um servidor com tempos exponenciais gera saídas em um processo Poisson.
  - Os tempos entre saídas de uma  $M/M/1$  seguem uma distribuição exponencial.
- Resultado mais geral: a saída de uma fila  $M/M/m$  estável em estado estacionário com parâmetro de entrada  $\lambda$  e parâmetro de tempos de serviço  $\mu$  para cada um dos  $m$  canais é também um processo Poisson com parâmetro  $\lambda$  (taxa de saída = taxa de chegada)
- Teorema válido para redes "feedforward" com filas  $M/M/m$

# Teorema de Burke



# Teorema de Burke



# Outros Teoremas

- Teorema de Burke generalizado para redes com feedbackward (retorno) por Jackson
- Teorema generalizado para redes com população finita (redes fechadas) por Gordon e Newell.

# Exercício 1

Você está projetando a infra-estrutura computacional de uma máquina de busca. Ele deve conter  $k$  computadores idênticos responsáveis pelo processamento das consultas e um computador front-end responsável pelo pré-processamento e roteamento das consultas (broker). Todos os computadores de processamento terão acesso total à base de dados. Logo o roteamento feito no broker consiste em enviar a primeira requisição pendente para a máquina que se tornar ociosa primeiro.

Assumindo que o tempo de processamento no broker é desprezível, se comparado ao tempo esperado para processamento das consultas, determine a configuração (valor de  $k$ ) com melhor custo-benefício se a probabilidade do broker ter que enfileirar consultas tiver que ser inferior a 0.001. As suas opções são 5, 10 ou 20 máquinas de processamento. Assuma que os tempos entre chegadas de consultas e os tempos de processamento de consultas sejam, ambos, exponencialmente distribuídos, com médias 0.1 e 0.2 segundos, respectivamente.

# Exercício 2

Para melhorar o desempenho de um servidor de arquivos você deseja realocar uma parte da memória correntemente usada para gerenciamento de conexões pendentes para a extensão do cache de blocos. Até então, você nunca teve problemas de recusa de conexão por falta de memória, motivando assim sua decisão.

Sabendo que as chegadas de requisições seguem um processo Poisson com taxa 10 reqs/hora e que os tempos de serviços são exponencialmente distribuídos com média 10 segundos, qual a probabilidade do servidor recusar conexões, caso a memória seja reduzida de tal maneira que somente até 4 requisições possam estar pendentes a qualquer instante. Ainda neste cenário, qual seria o número médio de requisições no sistema (em atendimento e pendentes).

Refaça os cálculos caso a carga estimada seja mais alta, com uma taxa de 5 requisições por minuto, em média.

# Exercício 3

Considere um storage frame composto por um disk array com 10 discos idênticos e sua controladora. Os arquivos são armazenados no array com um leiaute de faixas. Em outras palavras, cada bloco é dividido igualmente por todos os discos. O tempo médio de serviço em um disco para recuperar  $k$  blocos de 512 bytes do storage frame é  $0.004 + 0.000003k$  segundos. Além disto, devido ao overhead de sincronização, estima-se que o tempo de resposta medio do array seja em torno de 10% superior ao tempo de resposta medio de um único disco. Sabe-se que:

- o storage frame recebe requisições para blocos de 500KB
- os tempos entre chegadas de requisições ao storage frame são exponencialmente distribuídos com média 0.1 segundos,
- os tempos de serviço em cada disco tem CV de 1.5 e
- os tempos de serviço da controladora do array seguem uma distribuição exponencial com média 10 ms

Calcule o tempo médio de resposta do storage frame. Qual a probabilidade do tempo gasto na controladora ser superior a duas vezes o tempo médio?

Se os tempos de serviço em cada disco tiverem coeficiente de variação igual a 0.998, calcule a probabilidade de haver mais de 5 requisições pendentes em um disco e a probabilidade do tempo de resposta do disk array ser superior o tempo medio de resposta de 1 disco em mais de 50%.