Introdução à teoria de filas

Prof. Renato Pimentel

2025/2

Sumário

1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Filas

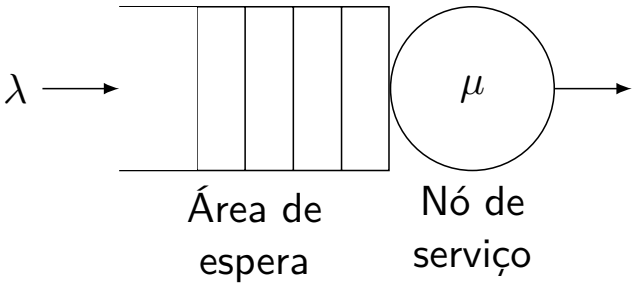
- Virtualmente **impossível** eliminar tempo de espera (sem custos extraordinários);
- Dimensionar o problema a níveis aceitáveis de custo $\times$ espera.

Aplicações

- Otimização do fluxo de pessoas em serviços (bancos, mercados, hospitais, serviços *online*);
- Controle de tráfego;
- Estimativa de performance computacional;
- *Design* de sistemas de manufatura;
- Redes e telecomunicações.

Aspectos relevantes

- Distribuição de entrada (*input*);
- Distribuição de saída/serviço (*output*);
- Número de canais de serviço (*servers*);
- Número máximo de clientes no sistema;
- Disciplina de serviço;
- *Calling source* (fonte dos clientes).



Sumário

1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Notação de Kendall

- Notação clássica para descrever fila, proposta por Kendall (1953).
- Normalmente no formato $A/S/c/K/N/D$:
  - A: *Arrival process* – distribuição das **chegadas**;
  - S: *Service distribution* – distribuição do **serviço**/atendimento;
  - c: número de **servidores** (atendentes);
  - K: Capacidade da fila;
  - N: **População** (*calling source size*);
  - D: **Disciplina** da fila.

Exemplos de usos

- A e S: M (*memoryless* ou markoviana – i.e. **exponencial**),
  BMAP (*batch markovian*, generalização do modelo markoviano para **chegadas em lote**),
  D (**determinística** ou degenerada),
  G (geral ou arbitrária – i.e. não especificada);
- K: finita ou infinita;
- N: finita ou infinita;
- D: FIFO/FCFS – *first-in(-come), first-out(-served)*. Ex.: atendimento bancário;
  LIFO/LCFS – *last-in(-come), first-out(-served)*. Ex.: estocagem (como carregamentos de contêineres em navios);
  SIRO – *service-in-random-order* (**aleatório**). Ex.: contemplação de consórcios;
  PQ – *priority queue* (fila de **prioridades**). Ex.: pronto atendimento hospitalar.

- $K/N/D$ são **opcionais** (valores padrão: ∞/∞/FIFO);
- Tipo mais clássico de fila: $M/M/1$.

Sumário

1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Comportamento de clientes

- ***Jockeying***: Mudar para fila menor;
- ***Balking***: Desistir de entrar no sistema;
  - Isto é, **taxa de entrada** é função do **tamanho da fila**.
- ***Reneging***: Possibilidade de cliente abandonar a fila no meio do processo.

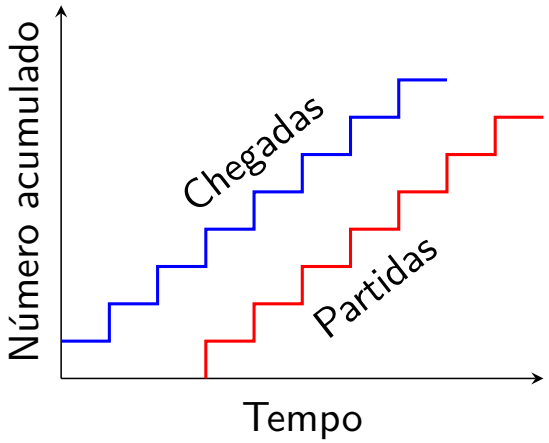
Quantidades de interesse

- L: número médio (esperado) de clientes no sistema;
- L_Q ou L_0: número médio de clientes na fila;
- W: tempo de espera médio de clientes no sistema;
- W_Q ou W_0: tempo de espera médio de clientes na fila.

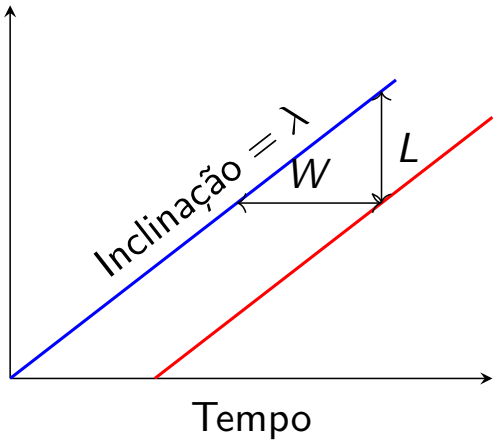
Fórmula de Little:

$$L = \lambda W \tag{1}$$

(a) Chegadas e partidas aleatórias



(b) Valores suavizados



- Semelhantemente:

$$L_Q = \lambda W_Q$$

- Essas fórmulas valem para quase todos os modelos de fila.

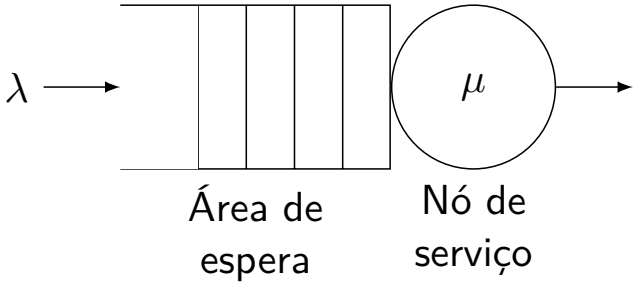
Exemplo 1

Navios cargueiros chegam em um cais de acordo com um processo de Poisson com taxa $\lambda = 2$ navios por dia. Registros diários mostram que há em média três navios descarregando ou esperando descarregar em um dado instante qualquer. Em média, quanto tempo um navio gasta no porto? Assuma que o navio saia imediatamente após a descarga.

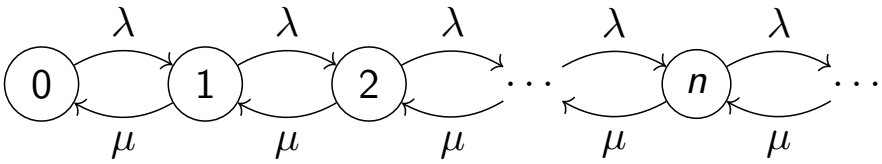
Sumário

1 Teoria de filas
- Introdução
- Notação de Kendall
- Caracterização
- Fila $M/M/1$
- Fila $M/M/\infty$
- Fila $M/M/c$
- Processos de nascimento e morte
- Redes de filas (*queue networks*)

Fila $M/M/1$



- Diagrama de taxa de transição:



- Como calcular a distribuição limitante?

- A fila $M/M/1$ tem matriz geradora:

$$\boldsymbol{G} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \mu & -(\mu+\lambda) & \lambda & \\ & & & & \\ & & & & \end{bmatrix}$$

.

- Precisamos calcular $\boldsymbol{\pi} = [\pi_0,\ \pi_1,\ \pi_2,\ \ldots]$ tal que $\boldsymbol{\pi}\boldsymbol{G} = \boldsymbol{0}$.
- Resolvendo-se o sistema, obtém-se

$$\pi_j = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^j, \quad j = 0, 1, \ldots.$$

- A distribuição acima oferece outros resultados, como a **média de clientes no sistema**,

$$L = \frac{\lambda}{\mu - \lambda}. \tag{2}$$

Fator de utilização

- Definimos $\rho = \lambda/\mu$ como o **fator de utilização**, ou **intensidade de tráfego**.
- Note que, pela distribuição estacionária:

$$\pi_0 = 1 - \frac{\lambda}{\mu}\,.$$

- π_0 representa a probabilidade de a fila estar vazia a longo prazo.

Tempos gastos

- Além disso, da fórmula de Little (1) e de (2), obtemos que:

$$W = \frac{1}{\mu - \lambda}\,.$$

- Quanto desse tempo W é gasto na fila (W_q):

$$W_Q = W - E[S] \tag{3}$$

- Onde $E(S)$ é o tempo esperado de serviço. Se o tempo de serviço é exponencialmente distribuído (**markoviano**), temos: $E(S) = 1/\mu$.

- Substituindo-se $E(S)$ em (3):

$$\begin{aligned} W_Q &= W - E[S] \\ &= \frac{1}{\mu - \lambda} - \frac{1}{\mu} \\ &= \frac{\lambda}{\mu(\mu - \lambda)} \,. \end{aligned}$$

- Logo, o **número médio de clientes na fila** (i.e. seu **tamanho médio**) é dado por

$$\begin{aligned} L_Q &= \lambda W_Q \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)} \,. \end{aligned}$$

Exemplo 2

Clientes chegam a um banco de acordo com um processo de Poisson com taxa de chegada $\lambda = 5$ por hora. Há um único caixa aberto, e os tempos de serviço são distribuídos exponencialmente com média de 10 minutos.

1. A longo prazo, qual a probabilidade de que há duas pessoas ou mais no banco sendo atendidas, ou esperando atendimento?
2. Qual o fator de utilização do sistema?
3. Qual a probabilidade de que um cliente terá que esperar antes de ser atendido?
4. Qual o tempo médio de espera de um cliente?

Exemplo 3

Clientes chegam no caixa do supermercado segundo um processo de Poisson com taxa $\lambda = 1$ por minuto. A gerência do mercado quer saber se deve ou não contratar um empacotador. Os tempos de *checkout* são distribuídos exponencialmente, e com empacotador o tempo médio é 30 s, enquanto sem empacotador o tempo médio aumenta para 50 s. Compare o tamanho esperado das filas com e sem empacotador.

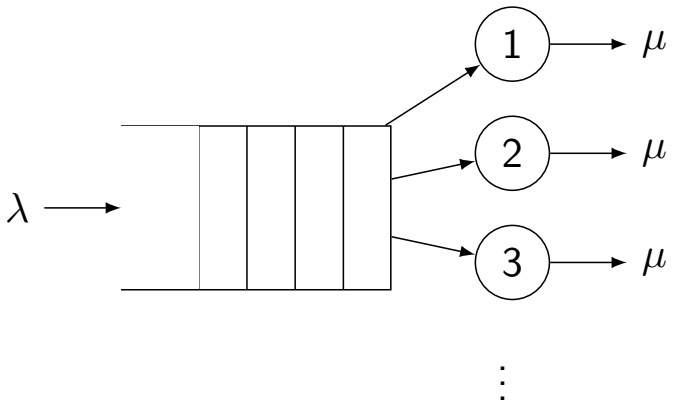
Exemplo 4

Um McDonald's usa em média uma tonelada de batata por semana. A quantidade média de batatas na lanchonete em um dado momento é 500 kg. Em média, quanto tempo uma batata fica na lanchonete antes de ser usada?

Sumário

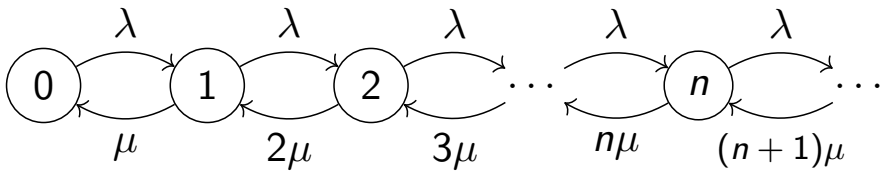
1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Fila $M/M/\infty$



- Como fica o tempo de espera se temos **infinitos** servidores disponíveis?
  - Todo cliente que chega é atendido imediatamente.

- Diagrama de taxa de transição:



- A fila $M/M/\infty$ tem matriz geradora:

$$\boldsymbol{G} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & \\ & & 3\mu & -(3\mu+\lambda) & \lambda \\ & & & & \\ & & & & \end{bmatrix}.$$

$\ddots \quad \ddots$

- A distribuição limitante é dada por

$$\pi_j = \frac{\left(\frac{\lambda}{\mu}\right)^j \mathrm{e}^{-\lambda/\mu}}{j!}, \quad j = 0, 1, \ldots$$

- Além disso, a média de clientes no sistema é

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

- Ao passo que o tempo médio de espera no sistema é

$$W = \frac{1}{\mu}$$

- Como ficam, neste caso, L_Q e W_Q?

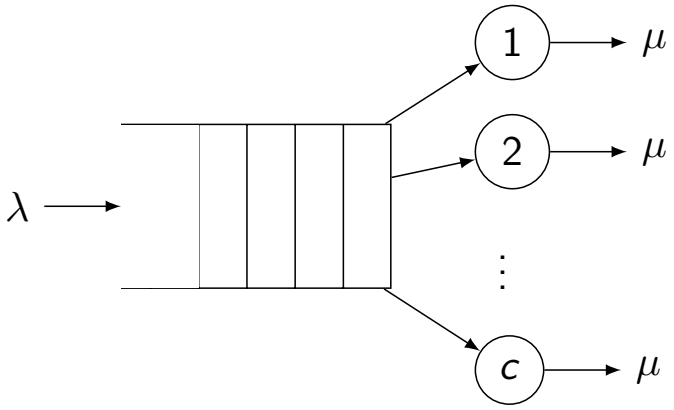
Exemplo 5

Chamadas chegam aleatoriamente a um *call center* com taxa de 140 por hora. Se há um número muito grande de linhas disponíveis e as chamadas duram em média 3 minutos, qual é o número médio de linhas em uso?

Sumário

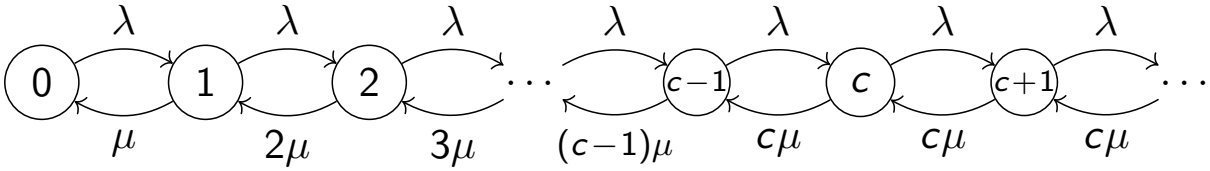
1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Fila M/M/c



- Como fica o tempo de espera se temos c servidores simultâneos, ao invés de apenas um?
- Com múltiplos processos de Poisson, somamos as taxas de chegada.
  - Taxa de serviço possível no $M/M/c$, portanto, é de até $c\mu$.

- Diagrama de taxa de transição:



- E a matriz geradora é

$$\boldsymbol{G} = \begin{array}{c} \begin{bmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & & & \\ & & & & & & \\ & & & c\mu & -(c\mu+\lambda) & \lambda & \\ & & & & c\mu & -(c\mu+\lambda) & \lambda \\ & & & & & & \\ & & & & & & \end{bmatrix} \end{array} \begin{array}{l} 0 \\ 1 \\ 2 \\ \\ c \\ c+1 \\ \\ \\ \end{array}.$$

Sendo o fator de utilização dado por

$$\rho = \frac{\lambda}{c\mu}\,, \tag{4}$$

temos que

$$\pi_0 = \left[\left(\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!}\right) + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}\right]^{-1},$$

$$\pi_j = \begin{cases} \pi_0 \dfrac{(c\rho)^j}{j!}, & \text{se } 0 \leq j < c\,; \\[2ex] \pi_0 \dfrac{(c\rho)^j c^{c-j}}{c!}, & \text{se } c \leq j\,. \end{cases} \tag{5}$$

- Ou, substituindo-se (4) em (5):

$$\pi_0 = \left[1 + \sum_{k=1}^{c-1} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k + \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{1}{1-\frac{\lambda}{c\mu}}\right]^{-1}$$

$$\pi_j = \begin{cases} \dfrac{1}{j!}\left(\dfrac{\lambda}{\mu}\right)^j \pi_0, & \text{se } 0 \leq j < c; \\[2ex] \dfrac{1}{c!}\left(\dfrac{\lambda}{\mu}\right)^c \left(\dfrac{\lambda}{c\mu}\right)^{j-c} \pi_0, & \text{se } c \leq j. \end{cases} \tag{6}$$

- É possível encontrar as outras quantidades de interesse, a partir da distribuição. Temos que

$$L_Q = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{\frac{\lambda}{c\mu}}{\left(1-\frac{\lambda}{c\mu}\right)^2}\pi_0 \,.$$

Logo, $W_Q = \dfrac{L_Q}{\lambda}$ (fórmula de Little), e

$$W = W_Q + \frac{1}{\mu}\,, \qquad L = \lambda W = \lambda\left(W_Q + \frac{1}{\mu}\right) = L_Q + \frac{\lambda}{\mu}\,.$$

Probabilidade de um cliente ter de esperar na fila

- Neste tipo de fila, tal quantidade pode ser de interesse, e é dada por – a partir de (6):

$$P_{\text{espera}} = P[j \geq c] = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{1}{1-\frac{\lambda}{c\mu}}\pi_0 \,.$$

  - Corresponde à probabilidade de todos os c servidores estarem ocupados.
- É possível estimar W_Q a partir de P_{espera}:

$$W_Q = \frac{P_{\text{espera}}}{c\mu - \lambda}\,.$$

Exemplo 6

Determine o tempo de espera médio (W) em uma fila $M/M/2$ com $\lambda = 2$ e $\mu = 1,2$. Compare com o tempo de espera (W) em uma fila $M/M/1$ com taxa de chegada $\lambda = 1$ e taxa de serviço $\mu = 1,2$.
Antes de calcular, tente imaginar como vai ser o resultado!

Exemplo 7

Um mercado tem dois caixas operando os pagamentos. Se o tempo de serviço é exponencial com média de quatro minutos, e as pessoas chegam ao mercado em um processo de Poisson com taxa 10/h e ficam no mercado um tempo exponencial antes de se dirigirem ao caixa, responda:

1. Qual a probabilidade de ter de esperar para ser servido no caixa?
2. Qual é o tempo ocioso de cada operador de caixa?
3. Se o cliente precisa esperar na fila, qual é o tempo que ele passa no sistema?

Exemplo 8

Há três tradutores no escritório de uma embaixada. Cada tradutor traduz seis e-mails por hora. Sabendo-se que e-mails chegam para ser traduzidos com uma taxa de 15 por hora, responda:

1. Que fração do tempo todos os tradutores estarão ocupados?
2. Qual a quantidade média de e-mails esperando para serem traduzidos?

Às vezes as regras da fila nos obrigam a usar um espaço de estados diferente de simplesmente "número de clientes no sistema". Vamos considerar um exemplo desse tipo:

Exemplo 9

Considere um *spa* que consiste de duas cadeiras: uma de massagem (0), e uma de acupuntura (1), cada qual com seu atendente.
Suponha que um cliente que chega sempre vai à cadeira de massagem. Quando a massagem termina, ele vai para a cadeira de acupuntura se ela estiver vazia, ou espera na cadeira de massagem até que a cadeira de acupuntura fique livre. Suponha também que um cliente em potencial somente entra no *spa* se a cadeira de massagem (0) estiver vazia.
Se a chegada de clientes segue Poisson com taxa λ, e as duas cadeiras têm taxas exponenciais de serviço distintas μ_0 e μ_1, responda:

1. Qual a proporção de clientes que entra no *spa*?
2. Qual a média de clientes dentro do *spa*?
3. Qual o tempo médio que um cliente passa no *spa*?
4. Qual é μ_b, a fração de clientes que tem que esperar entre a massagem e a acupuntura?

Sumário

1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Todos os modelos de filas vistos são **processos de nascimento e morte** (***birth and death***).

Processo de nascimento e morte

Cadeia de Markov contínua no tempo com estados $\{0, 1, 2, \dots\}$ onde:

- Somente é possível transicionar do estado j para o estado $j-1$ ou $j+1$;
- $v_0 = \lambda_0; \quad v_j = \lambda_j + \mu_j, j > 0.$
- $p_{01} = 1; \quad p_{j,j+1} = \dfrac{\lambda_j}{\lambda_j + \mu_j}, j > 0; \quad p_{j,j-1} = \dfrac{\mu_j}{\lambda_j + \mu_j}, j > 0\,.$
- $\{\lambda_j\}_{j=0}^{\infty}$ e $\{\mu_j\}_{j=0}^{\infty}$ são, respectivamente, as **taxas de nascimento** (ou **chegada**) e **de morte** (ou **saída**).

- **Fila** $M/M/1$: processo de nascimento e morte onde as taxas de nascimento e morte são, respectivamente:

$$\lambda_j = \lambda, j \geq 0, \qquad \mu_j = \mu, j \geq 1\,.$$

- **Fila** $M/M/c$: processo de nascimento e morte onde as taxas de nascimento e morte são, respectivamente:

$$\lambda_j = \lambda, j \geq 0, \qquad \mu_j = \begin{cases} j\mu, & 1 \leq j \leq c\,; \\ c\mu, & j > c\,. \end{cases}$$

- **Fila** $M/M/\infty$: exercício.

- Outro exemplo: **Processo de Poisson** homogêneo.
- Processo de nascimento e morte onde

$$\lambda_j = \lambda,\ j \geq 0, \qquad\qquad \mu_j = 0,\ j \geq 0\,.$$

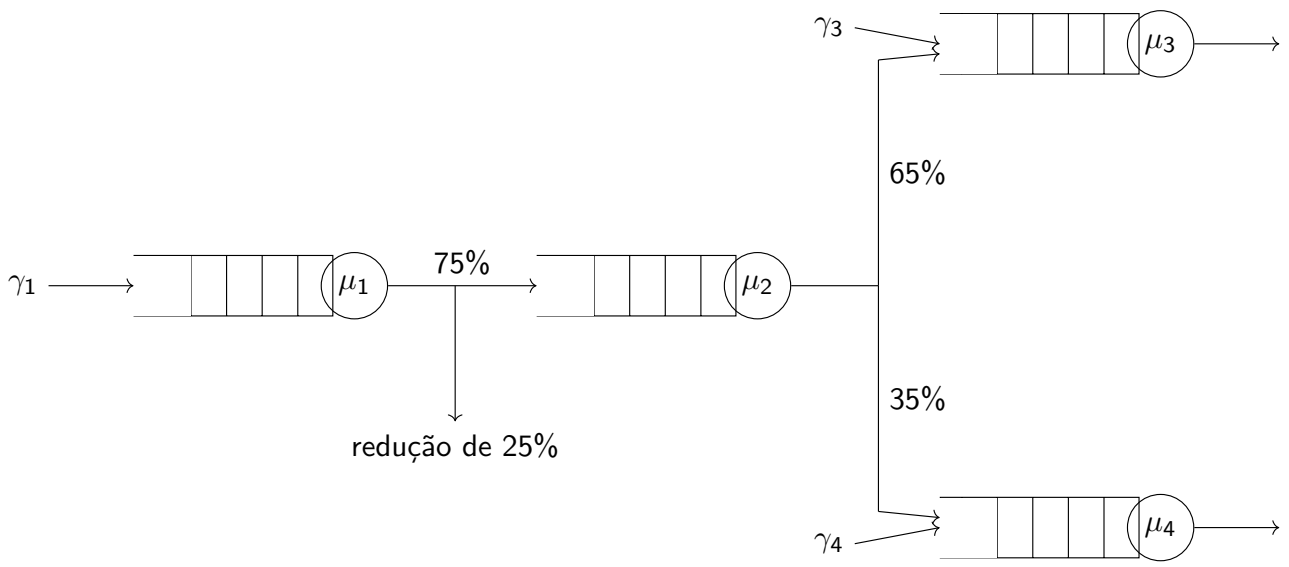
- i.e. partidas (mortes) nunca ocorrem; e chegadas ocorrem a intervalos de tempo com média exponencial $1/\lambda$, constante.
- Como $\mu_j = 0$, para qualquer j, o processo de Poisson é chamado **processo de nascimento puro**.

Sumário

1. Teoria de filas
   - Introdução
   - Notação de Kendall
   - Caracterização
   - Fila $M/M/1$
   - Fila $M/M/\infty$
   - Fila $M/M/c$
   - Processos de nascimento e morte
   - Redes de filas (*queue networks*)

Redes de filas (*queue networks*)

E se temos filas que alimentam outras filas?
Exemplos reais?



Exemplo de rede de filas aberta. Adaptado de: https://www.grotto-networking.com/DiscreteEventPython.html#NetQueues

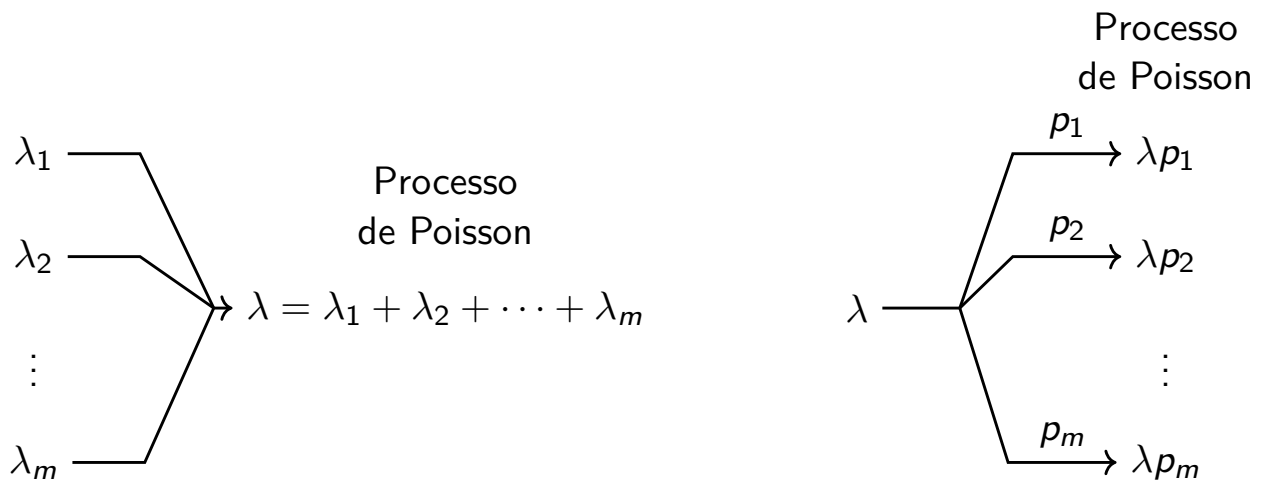
Redes jacksonianas

Requisitos:

- Chegadas externas seguem Poisson;
- Filas são sempre FIFO;
- Após serviço, um cliente:
  a) sai do sistema; ou
  b) se junta a uma nova fila, seguindo uma probabilidade determinada;
- Fator de utilização das filas menor que 1 (i.e. $\lambda_i < \mu_i$).

James R. Jackson (1924-2011).
By Not stated - "James R. Jackson". Production and Operations Management (Wiley) 17 (6): i-ii. 2008.
http://onlinelibrary.wiley.com/doi/10.1111/poms.2008.17.issue-6/issuetoc

Relembrando: mesclagem e divisão de processos de Poisson



Teorema de Jackson

- A probabilidade de um estado da rede de filas é equivalente ao produto das probabilidades dos estados de cada uma das filas individualmente.
  - Isto é:

$$\pi(j_1, j_2, ..., j_m) = \pi_1(j_1) \cdot \pi_2(j_2) \cdots \pi_m(j_m) \,.$$

  - Tratamos as filas como $M/M/c$ independentes, cada uma com suas taxas de chegada e serviço próprias.

Medida de performance

Uma vez que cada fila i é $M/M/c$ com fator de utilização $\rho_i = \dfrac{\lambda_i}{c\mu_i}$:

- Todas as medidas vistas para rede $M/M/c$ se aplicam. Se $c = 1$ para todas, por exemplo, usam-se as medidas vistas para $M/M/1$:

$$L_i = \frac{\rho_i}{1-\rho_i}; \qquad \pi_i(j) = (1-\rho_i)\rho_i^j;$$
$$W_i = \frac{1}{\mu_i - \lambda_i}; \qquad W_{Qi} = \frac{\rho_i}{\mu_i - \lambda_i}.$$

- Para cada fila i, λ_i corresponde à **taxa de chegadas externas** à mesma, γ_i, somada das taxas oriundas de outras redes, ponderadas pelas respectivas probabilidades:

$$\lambda_i = \gamma_i + \sum_{j=1}^{m} \lambda_j P_{ji}\,.$$

- Se todas as filas são $M/M/1$, para a rede na totalidade: o **número médio de clientes na rede**, L_N, é:

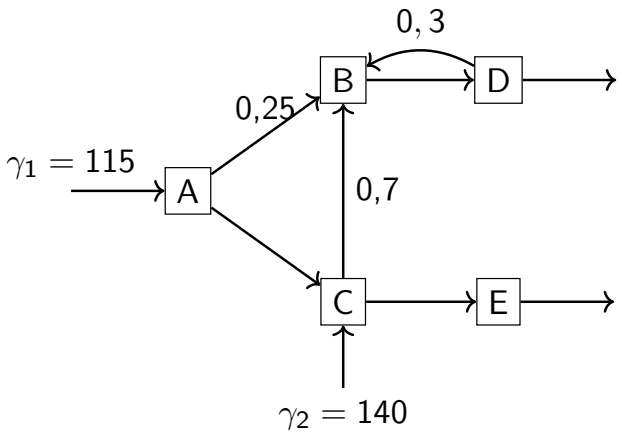
$$L_N = \sum_{i=1}^{m} L_i = \sum_{i=1}^{m} \frac{\rho_i}{1-\rho_i} \tag{7}$$

- Também é possível estimar W_N, o tempo médio que se permanece na rede, a partir de L_N. Temos que considerar a fórmula de Little $L_N = \lambda_N W_N$, onde $\lambda_N = \sum_i \gamma_i$ (somente as taxas externas). Logo, a partir de (7):

$$W_N = \frac{1}{\lambda_N}\sum_{i=1}^{m} L_i = \frac{1}{\lambda_N}\sum_{i=1}^{m} \frac{\rho_i}{1-\rho_i}\,.$$

Exemplo 10

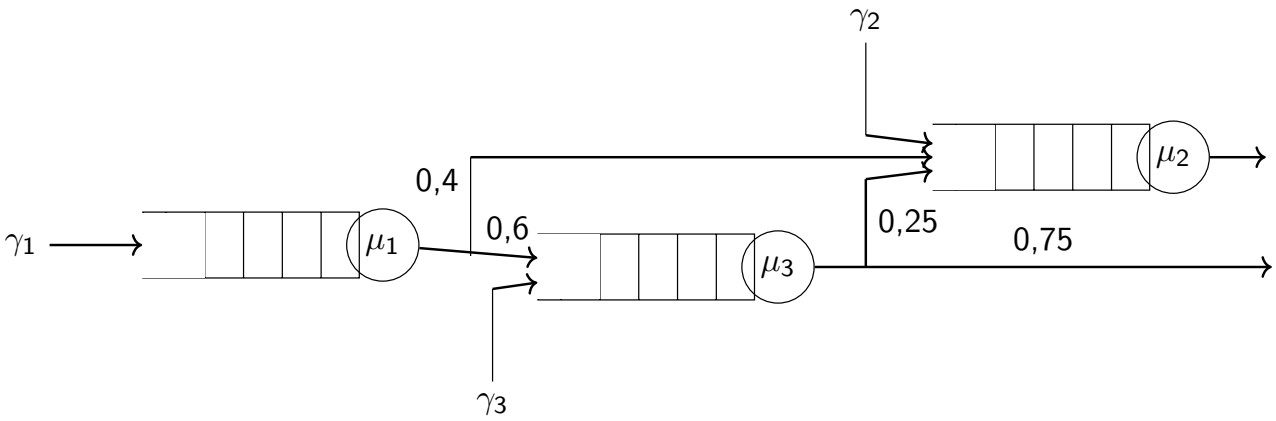
Vamos analisar esta rede de filas (veja que uma rede jacksoniana pode ter até laços e diversas entradas, contanto que as filas satisfaçam os critérios vistos anteriormente).



Fila	A	B	C	D	E
Núm. servidores (c_i)	1	3	1	4	1
Taxa de serviço (μ_i)	150	120	240	80	80

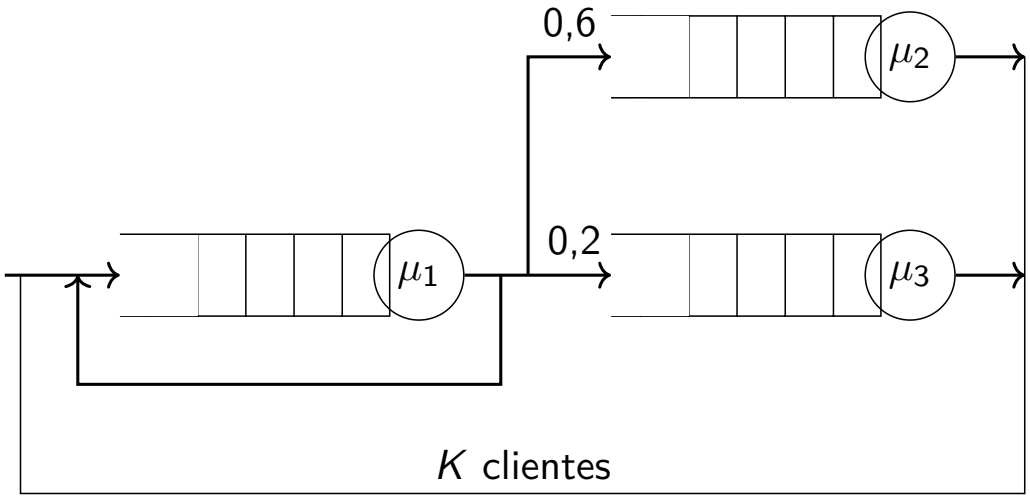
Exemplo 11

Vamos analisar esta rede de filas (Veja que uma rede jacksoniana pode ter até laços e diversas entradas, contanto que as filas satisfaçam os critérios vistos anteriormente).



Exemplo 12

Vamos analisar esta rede de filas com limite de K clientes.



Referências

1. ROSS, S. *Introduction to Probability Models*, 13. ed. (Academic Press), 2024.
2. https://www.probabilitycourse.com/

Os materiais de parte desta seção foram gentilmente cedidos por Leandro N. Couto (FACOM/UFU).
Adaptações e LaTeXagem: Renato Pimentel, FACOM/UFU