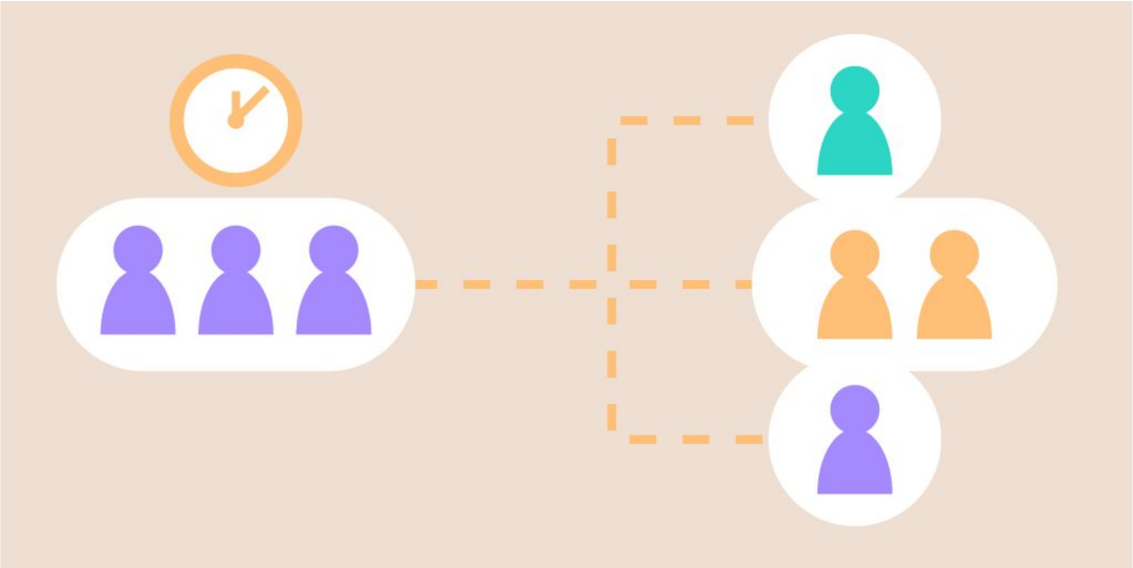
Modelagem e Simulação

Aula 6 - Filas

Recapitulando…

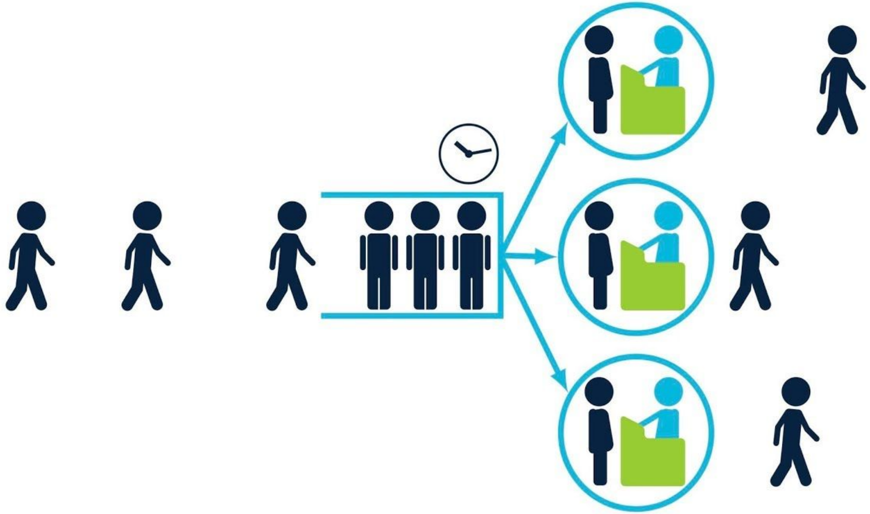
Com a cadeia de Markov contínua (CTMC) temos as ferramentas necessárias para falar de teoria de filas (queueing theory)



Filas

Não é problema de otimização!

- Virtualmente impossível eliminar tempo de espera (sem custos extraordinários)
- Dimensionar o problema a níveis aceitáveis de custo X espera

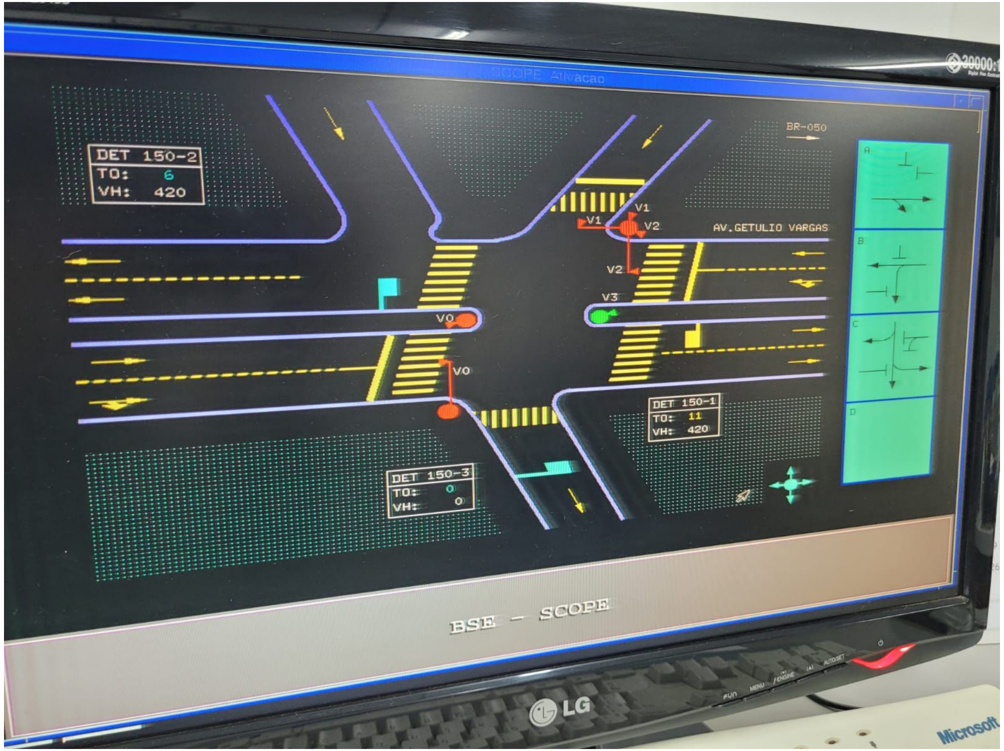


Aplicações

- 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

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



Características importantes

- 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)

Notação de Kendall

Notação clássica para descrever fila.

Normalmente no formato **A/S/c/K/N/D**

A: Arrival process

S: Service distribution

c: Número de servidores

K: Capacidade da fila

N: População (calling source size)

D: Disciplina da fila

Notação de Kendall

Notação clássica para descrever fila.

Exemplos de usos

A e S: M (markovian), BMAP (batch markovian), D (deterministic), G (general, ou arbitrária)

K: infinita ou finita

N: infinita ou finita

D: FIFO/FCFS, LIFO/LCLS, SIRO (random), PQ (priority)

Notação de Kendall

Notação clássica para descrever fila.

Normalmente no formato **A/S/c/K/N/D**

- K/N/D são opcionais (default: infinito/infinito/FIFO)
- Tipo mais clássico de fila: **M/M/1**

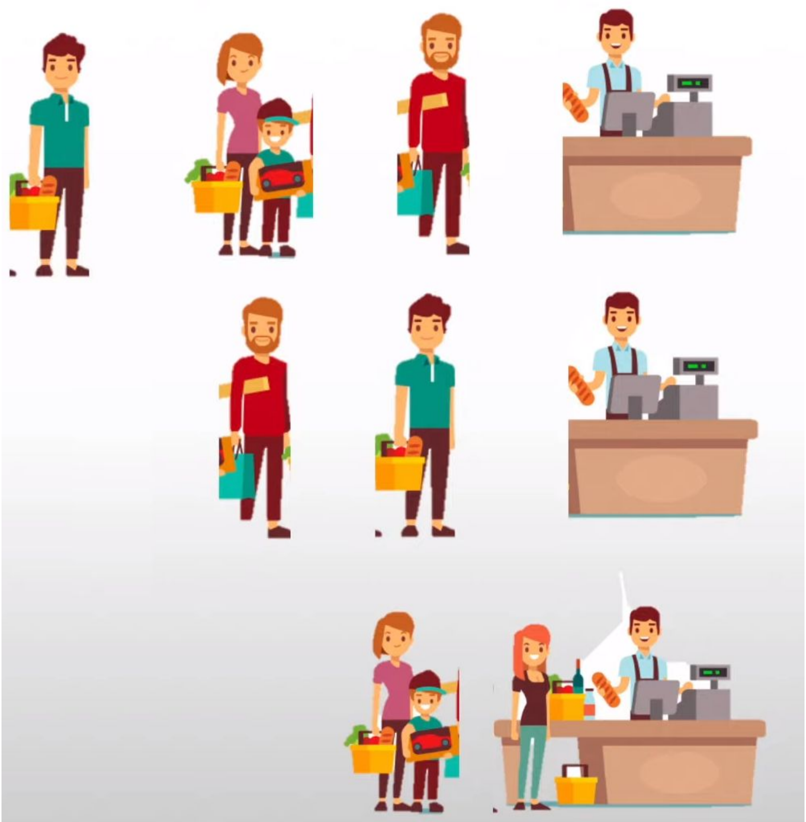
Comportamento de clientes

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



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

Rollercoaster Tycoon 2

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

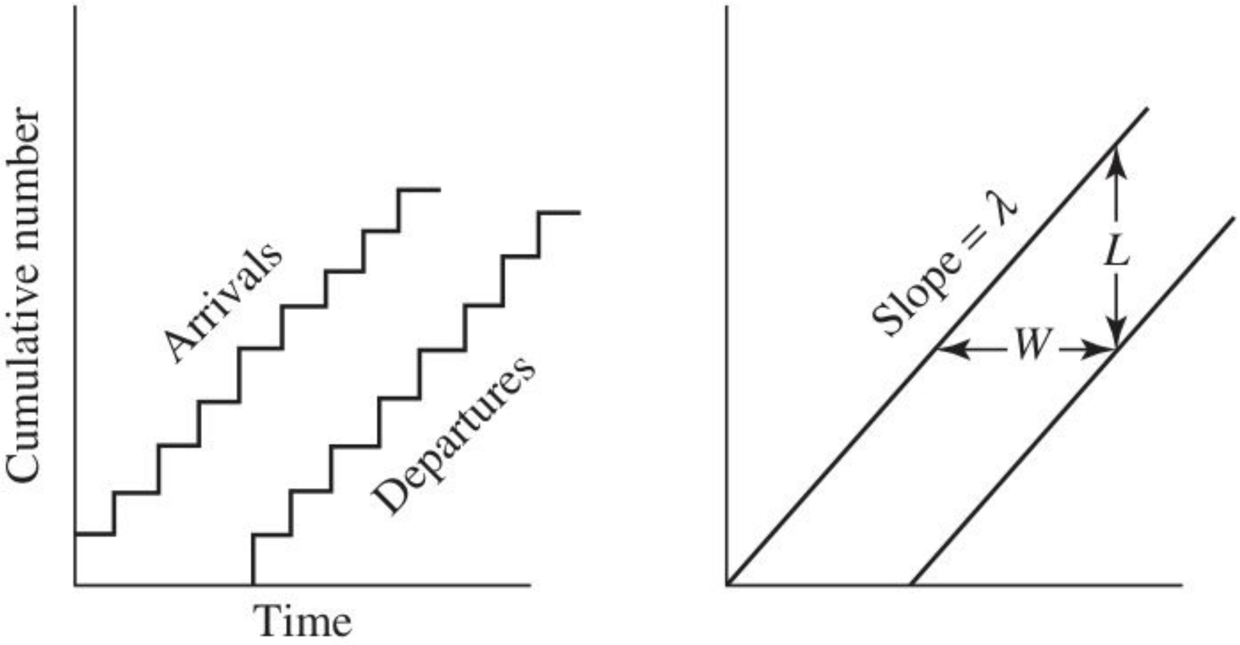
W_Q ou W_0: tempo de espera médio de clientes na fila

Teoria das Filas

A fórmula da fila:

$L = \lambda W$

Também chamada de fórmula de Little



(a) Random arrivals, departures

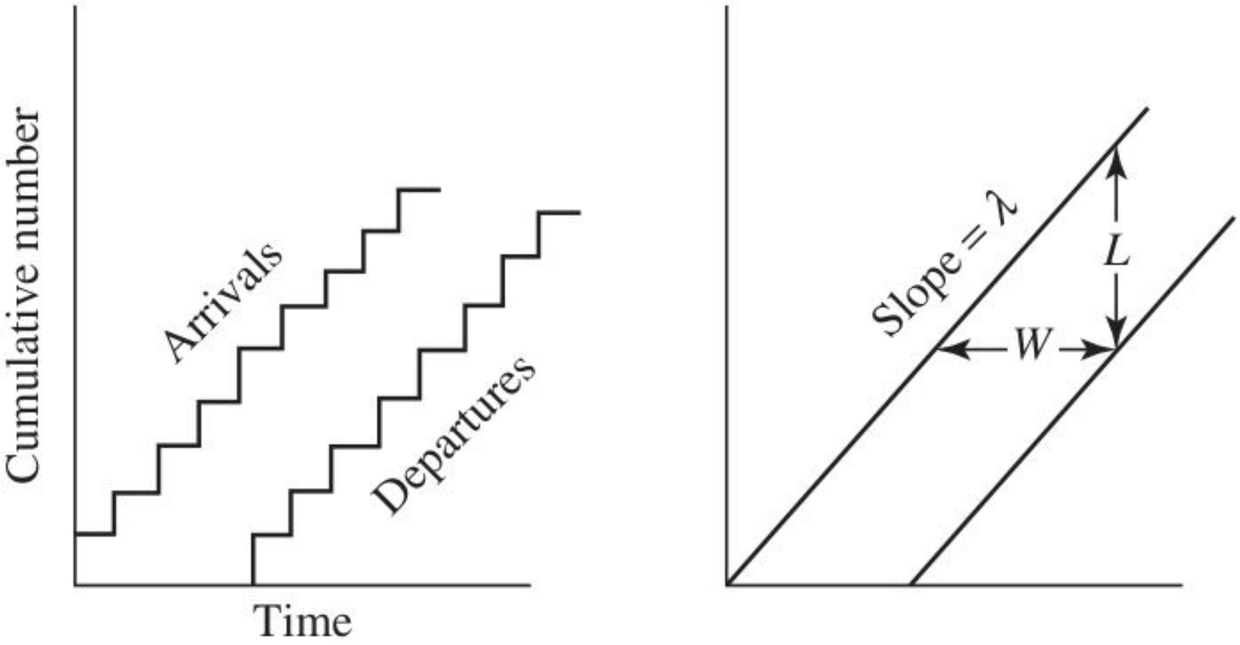
(b) Smoothed values

Teoria das Filas

Semelhantemente:

$$L_Q = \lambda W_Q$$

Essas fórmulas valem para quase todos os modelos de fila!



(a) Random arrivals, departures (b) Smoothed values

Teoria das Filas

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 3 navios descarregando ou esperando descarregar em um dado instante qualquer. Em média, quanto tempo um navio gasta no porto? Assuma que o navio sai imediatamente após a descarga.

Teoria das Filas - M/M/1

Vamos falar especificamente da fila M/M/1

Teoria das Filas - M/M/1

Como calcular a distribuição de equilíbrio/estacionária?

A fila M/M/1 tem matriz geradora:

$$G = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \cdots \\ 0 & \mu & -(\mu+\lambda) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

Teoria das Filas - M/M/1

Como calcular a distribuição de equilíbrio/estacionária?

Precisamos calcular π = [π0 π1 π2 …. πk] tal que π*G = 0

$$\pi_k = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k$$

Teoria das Filas - M/M/1

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

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

Teoria das Filas - M/M/1

Definimos **$\rho = \lambda/\mu$** como o **fator de utilização**, ou **intensidade de tráfego**

Note que, pela distribuição estacionária:

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

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

Teoria das Filas - M/M/1

Além disso, das 2 fórmulas de L que apresentamos, obtemos que:

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

Quanto desse tempo W é gasto na fila ($\mathbf{W_Q}$)?

Teoria das Filas - M/M/1

Além disso, das 2 fórmulas de L que apresentamos, obtemos que:

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

Quanto desse tempo W é gasto na fila (**W_Q**)?

W_Q = W - E(S)

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

Teoria das Filas - M/M/1

Fórmulas para W_Q e L_Q:

$$\begin{aligned} W_Q &= W - E[S] \\ &= W - \frac{1}{\mu} \\ &= \frac{\lambda}{\mu(\mu - \lambda)}, \\ L_Q &= \lambda W_Q \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)} \end{aligned}$$

Teoria das Filas

Exemplo 2

Clientes chegam em 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.

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

Teoria das Filas

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 30s, enquanto sem empacotador o tempo médio aumenta para 50s. Compare o tamanho esperado das filas com e sem empacotador.

Teoria das Filas

Exemplo 4

Um McDonalds usa em média 1 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?

λ = mean number of arrivals per time period

μ = mean number of people or items served per time period

L_s = average number of units (customers) in the system (waiting and being served)

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

W_s = average time a unit spends in the system (waiting time plus service time)

$$= \frac{1}{\mu - \lambda}$$

L_q = average number of units waiting in the queue

$$= \frac{\lambda^2}{\mu(\mu - \lambda)}$$

W_q = average time a unit spends waiting in the queue

$$= \frac{\lambda}{\mu(\mu - \lambda)}$$

ρ = utilization factor for the system

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

P_0 = probability of 0 units in the system (that is, the service unit is idle)

$$= 1 - \frac{\lambda}{\mu}$$

$P_{n>k}$ = probability of more than k units in the system, where n is the number of units in the system

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

Teoria das Filas

Fila M/M/∞

Como fica o tempo de espera se temos **infinitos** servidores disponíveis?

Todo cliente que chega é atendido imediatamente!

Como fica a matriz geradora/matriz de taxa de transição?

Teoria das Filas

Fila M/M/∞

$$Q = \begin{pmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & \\ & & 3\mu & -(3\mu+\lambda) & \lambda \\ & & & & \ddots \end{pmatrix}$$

Teoria das Filas

Fila M/M/∞

Distribuição estacionária:

$$\pi_k = \frac{(\lambda/\mu)^k e^{-\lambda/\mu}}{k!}$$

Média de clientes (L):

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

Tempo de espera (W)?

Teoria das Filas

Fila M/M/∞

Distribuição estacionária:

$$\pi_k = \frac{(\lambda/\mu)^k e^{-\lambda/\mu}}{k!}$$

Média de clientes (L):

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

(E quanto a L_Q e W_Q?)

Tempo médio de espera (W)? $W = 1/\mu$

Veja como verifica a fórmula de Little!

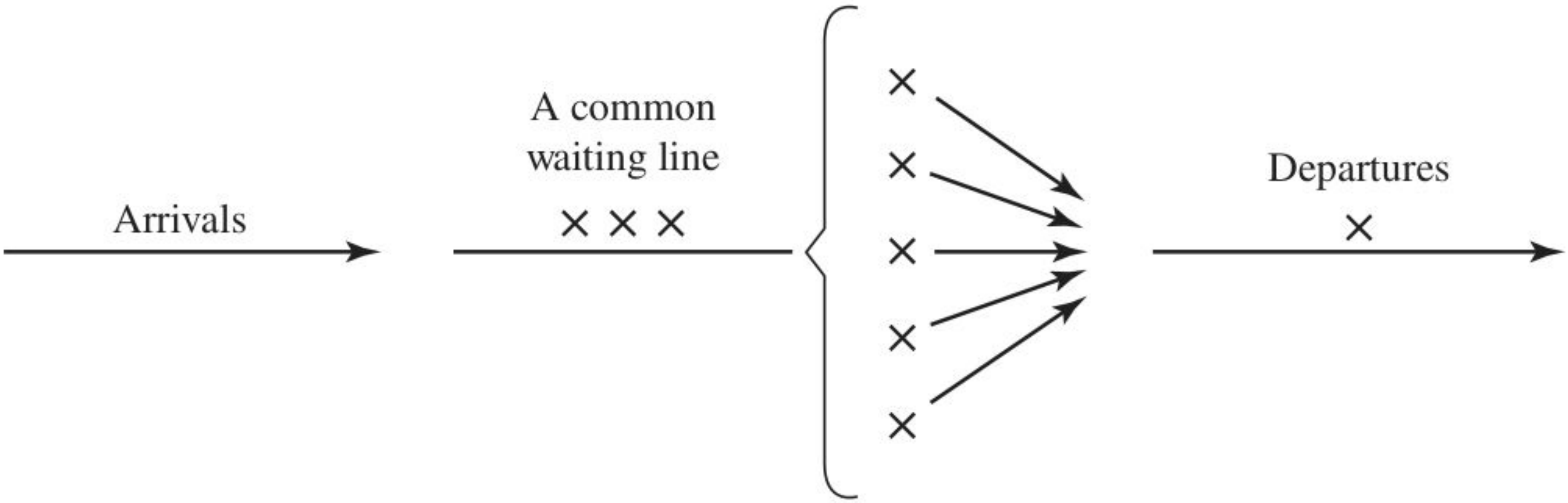
Teoria das Filas

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?

Teoria das Filas

Fila M/M/c



Teoria das Filas

Fila M/M/c

Como fica o tempo de espera se temos **c** servidores simultâneos ao invés de apenas 1?

Lembre-se que com múltiplos processos de Poisson, somamos as taxas de chegada. Taxa de serviço no M/M/c, portanto é **μ*c**.

E a matriz geradora G (ou Q)?

Teoria das Filas

Fila M/M/c

$$Q = \begin{pmatrix} -\lambda & \lambda & & & & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & & & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & & & & & \\ & & 3\mu & -(3\mu+\lambda) & \lambda & & & & \\ & & & & \ddots & & & & \\ & & & & c\mu & -(c\mu+\lambda) & \lambda & & \\ & & & & & c\mu & -(c\mu+\lambda) & \lambda & \\ & & & & & & c\mu & -(c\mu+\lambda) & \\ & & & & & & & & \ddots \end{pmatrix}$$

Teoria das Filas

Fila M/M/c (ou M/M/s ou M/M/k)

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

$$\pi_0 = \left[\left(\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!}\right) + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}\right]^{-1} \quad \pi_k = \begin{cases} \pi_0 \dfrac{(c\rho)^k}{k!}, & \text{if } 0 < k < c \\ \pi_0 \dfrac{(c\rho)^k c^{c-k}}{c!}, & \text{if } c \le k \end{cases}$$

Teoria das Filas

Fila M/M/c (ou M/M/s ou M/M/k)

$$\pi_k = \begin{cases} \dfrac{1}{k!}\left(\dfrac{\lambda}{\mu}\right)^k \pi_0 & \text{for } k = 0, 1, \ldots, s, \\ \dfrac{1}{s!}\left(\dfrac{\lambda}{\mu}\right)^s \left(\dfrac{\lambda}{s\mu}\right)^{k-s} \pi_0 & \text{for } k \geq s. \end{cases}$$

$$W_0 = \frac{L_0}{\lambda},$$

$$W = W_0 + \frac{1}{\mu},$$

$$L_0 = \frac{\pi_0}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{(\lambda/s\mu)}{(1-\lambda/s\mu)^2}.$$

$$L = \lambda W = \lambda\left(W_0 + \frac{1}{\mu}\right) = L_0 + \frac{\lambda}{\mu}$$

Teoria das Filas

Fila M/M/c (ou M/M/s ou M/M/k)

$$P[n \geq C] = \frac{(\lambda / \mu)^{c} C\mu}{C![(\mu - \lambda)]} P_0$$

Teoria das Filas

Exemplo 4

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!

Teoria das Filas

Exemplo 4

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!

(Resolva o mesmo exercício considerando uma fila M/M/1 com taxa de chegada $\lambda = 2$ e taxa de serviço $\mu = 2.4$)

Teoria das Filas

Exemplo 5

Um mercado tem 2 caixas operando os pagamentos. Se o tempo de serviço é exponencial com média 4 minutos, e as pessoas chegam no mercado em um processo de Poisson com taxa 10/h e ficam no mercado um tempo exponencial antes de se dirigirem pro caixa, responda:

a) Qual a probabilidade de ter que esperar pra ser servido no caixa?
b) Qual é o tempo ocioso de cada operador de caixa?
c) Se o cliente precisa esperar na fila, qual é o tempo que ele passa no sistema?

Teoria das Filas

Exemplo 6

Há 3 tradutores no escritório de uma embaixada. Cada tradutor traduz 6 e-mails por hora. Sabendo-se que e-mails chegam para ser traduzidos com uma taxa de 15 por hora, responda:

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

	$M/M/1$	$M/M/c$
$p(0)$	$1-\rho$	$\left[\frac{(c\rho)^c}{c!(1-\rho)}+\sum_{n=0}^{c-1}\frac{(c\rho)^n}{n!}\right]^{-1}$
L_q	$\frac{\rho^2}{1-\rho}$	$\frac{\rho(c\rho)^c p(0)}{c!(1-\rho)^2}$
L	$\frac{\rho}{1-\rho}$	$L_q+\frac{\lambda}{\mu}$
W_q	$\frac{\rho}{\mu(1-\rho)}$	$\frac{(c\rho)^c p(0)}{c!\,c\mu(1-\rho)^2}$
W	$\frac{1}{\mu(1-\rho)}$	$w_q+\frac{1}{\mu}$

Teoria das Filas

Exemplo 7

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:

Teoria das Filas

Exemplo 7

Considere um Spa que consiste de 2 cadeiras, 1 de massagem (1) e 1 de acupuntura (2). Suponha que um cliente que chega sempre vai na cadeira de massagem. Quando a massagem termina, ele vai pra cadeira de acupuntura se ela estiver vazia, ou espera na cadeira de massagem até a cadeira de acupuntura ficar livre. Suponha que um cliente em potencial só entra no Spa se a cadeira de massagem (1) estiver vazia. Supondo chegada de clientes seguindo Poisson com taxa λ, e que as duas cadeiras tem taxas de serviço distintas $\mu 1$ e $\mu 2$, responda:

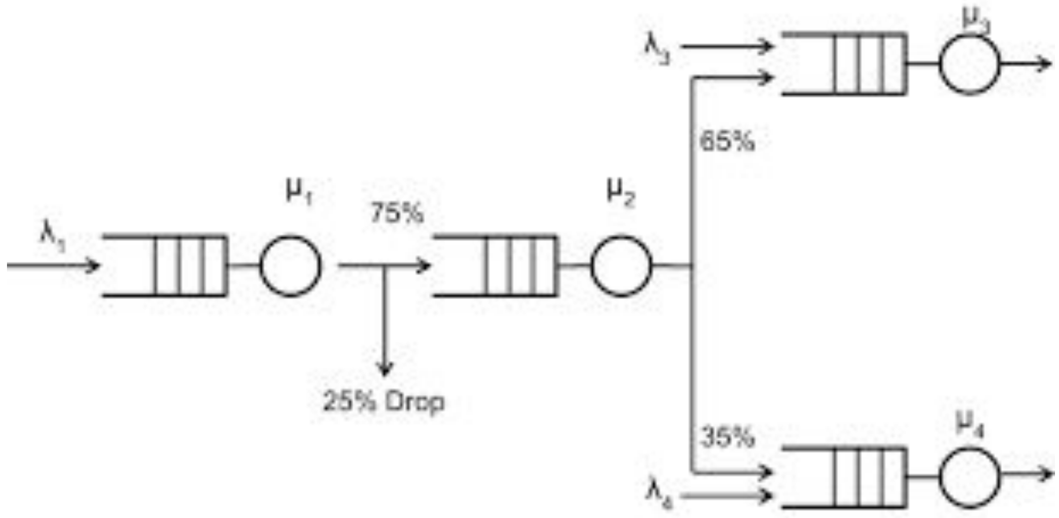
a) Qual a proporção de clientes que entra no Spa?
b) Qual a média de clientes dentro do Spa?
c) Qual o tempo médio que um cliente passa no Spa?
d) Qual é π_b, a fração de clientes que tem que esperar depois de receber massagem e antes da acupuntura?

Redes de Filas (Queue Networks)

Redes de filas

E se temos filas que alimentam outras filas?

Exemplos reais?



Redes de Filas (Queue Networks)

Redes Jacksonianas

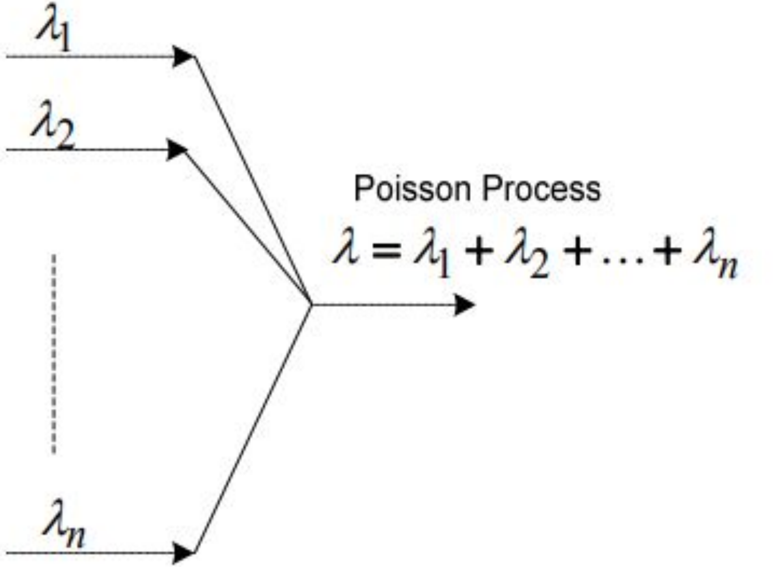
Requisitos:

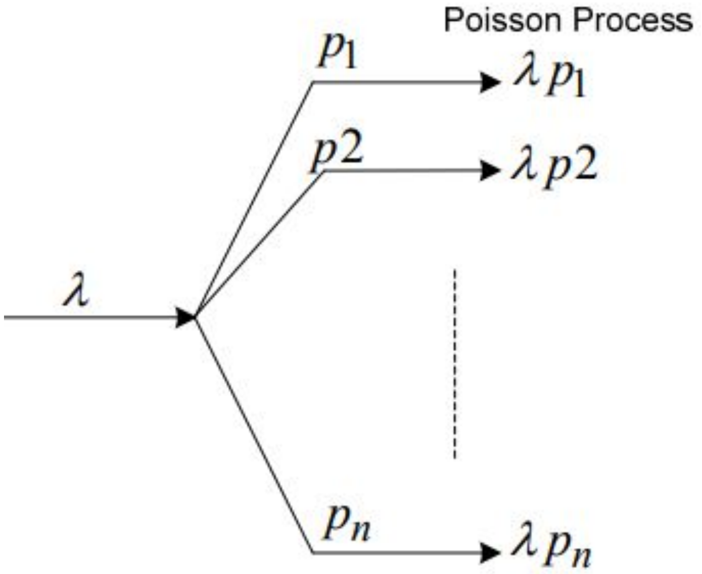
- Chegadas externas seguem Poisson
- Filas são sempre FIFO
- Após serviço, um cliente ou a) sai do sistema, ou b) se junta a uma nova fila seguindo uma probabilidade determinada
- Taxa de utilização das filas menor que 1

James R. Jackson

Redes de Filas (Queue Networks)

Redes Jacksonianas





Redes de Filas (Queue Networks)

Teorema de Jackson

A probabilidade do estado da rede de filas é o produto da probabilidade do estado de cada fila individualmente

Isto é, os estados das filas são independentes!

$$\pi(k_1, k_2, \ldots, k_N) = \pi_1(k_1) \cdot \pi_2(k_2) \cdot \ldots \cdot \pi_N(k_N)$$

Redes de Filas (Queue Networks)

Redes Jacksonianas

Tratamos redes como redes M/M/1 independentes, cada uma com sua taxa de chegada e serviço…

Lembrando que:

taxa que entra no estado N = taxa de sai do estado N

Since each queue i is a M/M/1 queue with ρ_i

$$L_i = \frac{\rho_i}{1-\rho_i} \qquad \pi_{n_i} = (1-\rho_i)\rho_i^{n_i}$$

$$W_i = \frac{1}{\mu_i - \lambda_i} \qquad W_{q_i} = \frac{\rho_i}{\mu_i - \lambda_i}$$

$\Rightarrow$ all M/M/1 measures apply (e.g., percentile of delay distribution, etc.)

For the network as a whole

LN – Average number of customers in network.

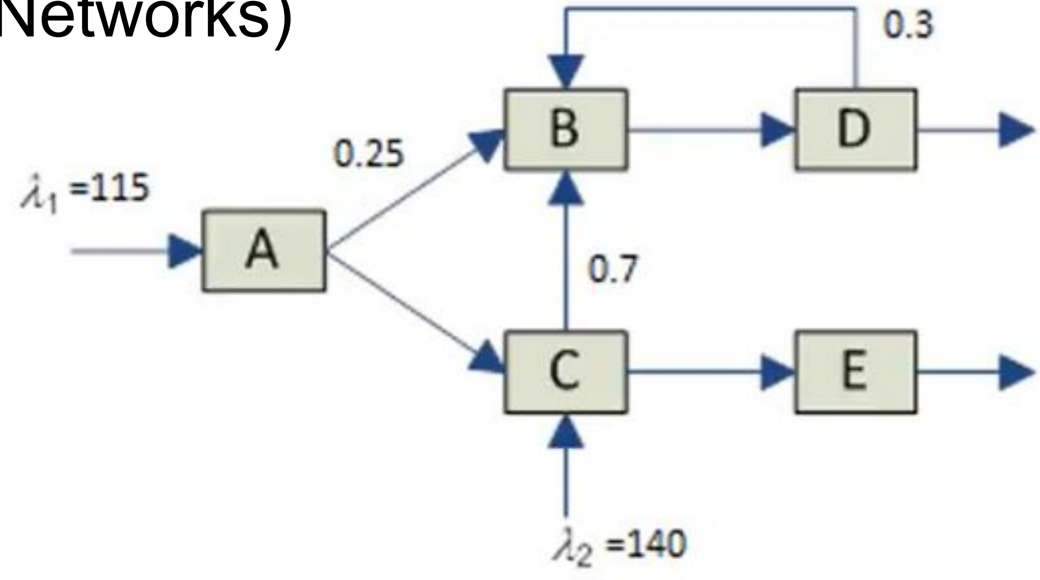
$$LN = \sum_{i=1}^{m} L_i = \sum_{i=1}^{m} \frac{\rho_i}{1-\rho_i}$$

Redes de Filas (Queue Networks)

Exemplo 8

- Vamos analisar esta rede de filas

  (Veja que uma rede Jacksoniana pode ter até loops e diversas entradas, contanto que as filas satisfaçam os critérios do slide anterior)



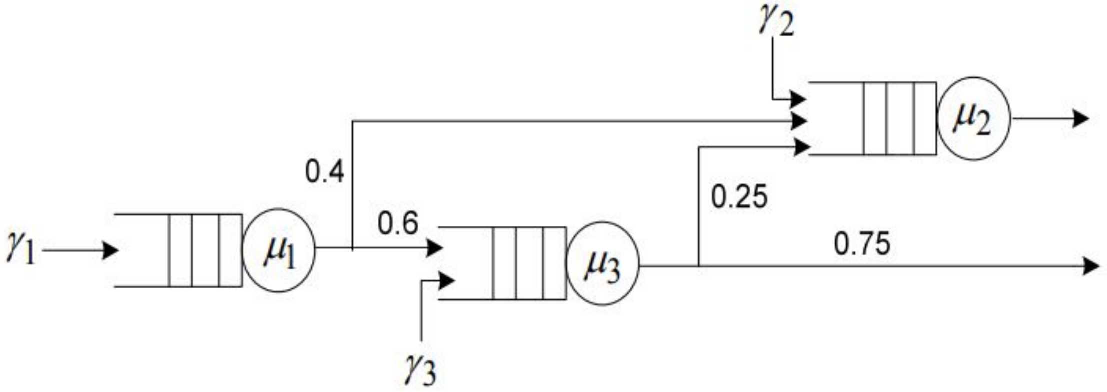
	Station				
	A	***B***	***C***	***D***	***E***
# servers (*c*)	1	3	1	4	1
Service rate (μ)	150	120	240	80	80

Redes de Filas (Queue Networks)

Exemplo 9

- Vamos analisar esta rede de filas

  (Veja que uma rede Jacksoniana pode ter até loops e diversas entradas, contanto que as filas satisfaçam os critérios do slide anterior)



Redes de Filas (Queue Networks)

Exemplo 10

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