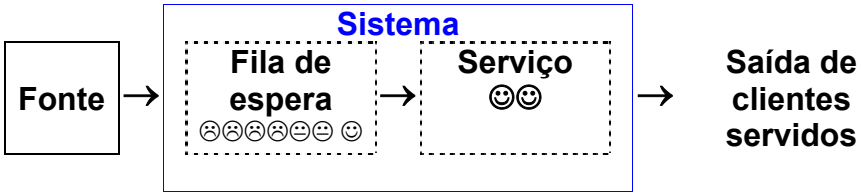
FILAS DE ESPERA

Notas baseadas em
“Introduction to Operations Research”
de Hillier e Lieberman.

ESTRUTURA BÁSICA DOS SISTEMAS DE FILAS DE ESPERA

Quando um determinado **serviço** é procurado por vários **clientes**, poder-se-ão formar **filas de espera**, já que o número de **servidores** e a **duração do serviço** de cada cliente usualmente não permite que cada cliente seja atendido assim que solicita o serviço.

Poderemos representar esquematicamente o processo de formação de filas de espera do modo seguinte:



Fonte (população) : **dimensão** (finita ou infinita);
processo de chegadas (distribuição estatística das chegadas, número de clientes por chegada);
atitude dos clientes (p.ex., possibilidade de recusa de um cliente aceder ao serviço ao constatar que a fila de espera é muito longa; possibilidade de desistência de um cliente que abandona o sistema sem ter sido servido depois de uma longa espera).

Sistema : **fila** (única; múltipla; comprimento limitado ou ilimitado; **disciplina**);
serviço (nº de servidores; distribuição estatística da duração de um atendimento; dimensão do serviço (o número de clientes que podem ser servidos simultaneamente))
capacidade do sistema (o número máximo de clientes que, num dado instante, podem estar no sistema, incluindo os clientes que aguardam na fila e os que estão a ser servidos).

Disciplina (isto é, a ordem pela qual os clientes são atendidos, destacando-se as disciplinas FIFO "first in, first out", ou seja, atendimento por ordem de chegada; LIFO "last in, first out", ou seja, a última entrada é processada primeiro; SIRO "service in random order", ou seja, serviço por ordem aleatória e PRI, correspondente a uma ordenação com prioridades).

A **NOTAÇÃO DE KENDALL v/w/x/y/z** é utilizada para caracterizar uma fila de espera: **v** caracteriza o processo de chegadas, representando a distribuição do intervalo de tempo entre chegadas consecutivas; **w** caracteriza a duração do serviço; **x** denota o número de servidores; **y** representa a capacidade do sistema ou a dimensão da fonte, e **z** especifica a disciplina da fila. Cada uma das especificações **v** e **w** poderá ser igual a **D**, **M**, $\mathbf{E}_k$, ou **G**, correspondendo a Determinístico, com Distribuição Exponencial (processo Markoviano), com distribuição Erlang-*k*, *k* = 1, 2, … (Gama), ou com qualquer outra distribuição, respectivamente. Muitas vezes não se especifica **y** e **z**, assumido-se que a capacidade do sistema é ilimitada e que a disciplina é FIFO.

Exemplos: **M/M/2/10/FIFO** (terá uma distribuição do intervalo de tempo entre chegadas consecutivas e uma distribuição da duração do serviço exponenciais, dois servidores, um limite máximo de 10 clientes no

interior do sistema e os clientes serão atendidos por ordem de chegada).

M/D/1 (processo de chegadas com uma distribuição do intervalo de tempo entre chegadas consecutivas exponencial, e uma duração determinística do serviço, um servidor, assumindo-se que o sistema tem uma capacidade ilimitada e que a disciplina será FIFO).

Consideremos dois exemplos de diferentes sistemas de filas de espera e identifiquemos as suas características:

♦ Os visitantes chegam ao átrio de entrada da Torre Panorâmica, onde poderão estar até 100 pessoas, aguardando que o único elevador disponível (com uma lotação de 20 pessoas) chegue para as levar ao miradouro panorâmico.

Assim, temos um sistema com capacidade limitada (100 pessoas), em que os clientes são os visitantes, com um único servidor (o elevador) e com uma dimensão de serviço igual a 20 (a lotação do elevador). Supõe-se que a disciplina da fila será FIFO.

♦ Numa fábrica de têxteis existem 15 teares que, quando se avariam, são reparados por dois técnicos de manutenção. Sabe-se que o intervalo de tempo entre duas avarias consecutivas se pode considerar com distribuição exponencial de média 5 horas e que a reparação de cada tear avariado tem uma duração que se pode considerar com distribuição exponencial de média 1 hora.

Parece aceitável assumir-se que o serviço será feito por ordem de ocorrência das avarias, ou seja, a disciplina da fila será FIFO. Assumindo que todas as máquinas avariadas serão reparadas, não há limitação na capacidade do sistema (no entanto, não será possível ter-se mais do que 15 máquinas avariadas), pelo que teremos um sistema M/M/2/15/FIFO alimentado por uma fonte com dimensão finita (15), já que os teares avariados serão os clientes. Os dois servidores são os técnicos de manutenção.

Exercício:

Numa olaria trabalham dois artesãos – um deles na produção das peças propriamente ditas, e o outro na sua decoração. As peças são fabricadas uma a uma, chegando ao artesão encarregado da decoração a um ritmo que se pode considerar constante de uma peça por cada 30 minutos. A duração da decoração de cada peça pode também ser assumida constante e igual a 45 minutos. O artesão decorador começa sempre pela última peça que acaba de receber da produção.

As actividades na olaria são iniciadas às 8:30 horas e a produção é interrompida das 12:30 às 13:30 e a decoração é interrompida das 12:45 às 13:45 para almoço dos artesãos. A produção diária termina às 15:30 horas e o artesão que se dedica à decoração mantém-se a trabalhar para escoar todas as peças produzidas nesse dia, pelo que nos inícios das manhãs não há peças a aguardar a decoração.

Simule manualmente o funcionamento do sector de decoração, determinando o valor médio de peças que aguardam a sua decoração durante a manhã (8:30 - 12:45 horas). Determine ainda a que horas o artesão decorador termina as suas actividades.

Pode-se considerar que o sector de decoração corresponde a um sistema D/D/1 com um tempo entre chegadas consecutivas igual a 30 minutos e uma duração de serviço igual a 45 minutos. As peças a decorar são os clientes e o artesão decorador é o servidor. A disciplina da fila de espera é LIFO, ou seja, atendimento por ordem inversa à da chegada.

No quadro seguinte procede-se à simulação manual do sistema.

T (minutos)	Chegada de cliente	Atendimento de cliente	Fila de espera
8:30	---	---	---
9:00	nº 1	nº 1	---
9:30	nº 2	nº 1	nº 2
9:45	---	nº 2	---
10:00	nº 3	nº 2	nº 3
10:30	nº 4	nº 4	nº 3
11:00	nº 5	nº 4	nº 3, nº 5
11:15	---	nº 5	nº 3
11:30	nº 6	nº 5	nº 3, nº 6
12:00	nº 7	nº 7	nº 3, nº 6
12:30	nº 8	nº 7	nº 3, nº 6, nº 8
12:45	---	---	nº 3, nº 6, nº 8
13:45	---	nº 8	nº 3, nº 6
14:00	nº 9	nº 8	nº 3, nº 6, nº 9
14:30	nº 10	nº 10	nº 3, nº 6, nº 9
15:00	nº 11	nº 10	nº 3, nº 6, nº 9, nº 11
15:15	---	nº 11	nº 3, nº 6, nº 9
15:30	nº 12	nº 11	nº 3, nº 6, nº 9, nº 12
16:00	---	nº 12	nº 3, nº 6, nº 9
16:45	---	nº 9	nº 3, nº 6
17:00	---	nº 6	nº 3
17:45	---	nº 3	---
18:30	---	---	---

Durante a manhã (8:30 – 12:45), há um período de 75 minutos sem peças a aguardar decoração, há 90 minutos com uma peça em espera, 75 minutos com duas peças em espera, e 15 minutos com três peças em espera. Assim, o número médio de peças a aguardar a decoração durante a manhã é igual a $(0 . 75 + 1 . 90 + 2 . 75 + 3 . 15) / 255 \approx 1{,}12$ peças.

O artesão encarregado da decoração das peças termina a sua actividade às 18:30, ficando prontas 12 peças por dia.

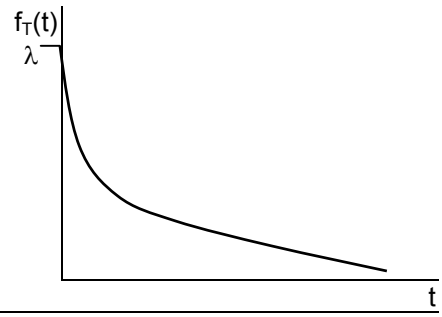
A DISTRIBUIÇÃO EXPONENCIAL

A distribuição exponencial é particularmente importante na caracterização dos sistemas de Filas de Espera, nomeadamente para descrever os intervalos de tempo entre chegadas consecutivas e as durações de serviço. Recordemos, então, algumas características desta distribuição:

Seja T uma variável aleatória com distribuição Exponencial Negativa, com parâmetro λ, isto é, $T \sim Exp(\lambda)$.

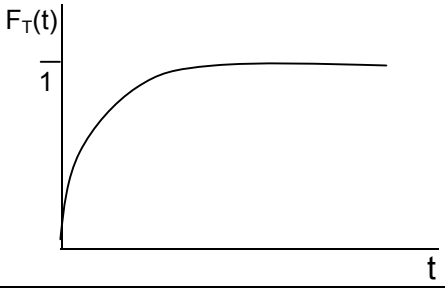
Função densidade de probabilidade:

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, t \geq 0 \\ 0, t < 0 \end{cases}$$



Função de distribuição acumulada:

$$F_T(t) = \begin{cases} 0, t < 0 \\ 1 - e^{-\lambda t}, t \geq 0 \end{cases}$$



μ = Valor Médio = $1 / \lambda$; σ = Desvio Padrão = $1 / \lambda$; $\gamma 1$ = Coef. Assim. = +2

Propriedade 1: **A função densidade de probabilidade da distribuição Exponencial é estritamente decrescente (para $t \geq 0$).**

Como consequência directa desta propriedade, poderemos escrever

$T \sim$ Exponencial, $P(0 \leq T \leq \mu) > P(\mu \leq T \leq 2\mu) > P(2\mu \leq T \leq 3\mu)$.

Com efeito, $P(0 \leq T \leq \mu) = P(T \leq \mu) = F_T(\mu) = 1 - e^{-\lambda\mu} = 1 - e^{-1} \approx$ **63,2 %** (Ou seja, **grande parte dos valores tomados por uma variável aleatória com distribuição Exponencial Negativa são inferiores ao respectivo valor médio**).

$P(\mu \leq T \leq 2\mu) = F_T(2\mu) - F_T(\mu) \approx$ **23, 3 %** (De notar que, por outro lado, só poucos valores são 'elevados': por exemplo, maiores do que o dobro do valor médio $P(T > 2\mu) = 1 - F_T(2\mu) = 1 - (1 - e^{-\lambda 2\mu}) = e^{-2} \approx 13{,}5\%$).

$P(2\mu \leq T \leq 3\mu) = F_T(3\mu) - F_T(2\mu) \approx$ **8,6 %**

Assim, se assumirmos que os intervalos de tempo entre chegadas consecutivas se distribuem exponencialmente, estaremos a assumir que a maior parte desses intervalos de tempo serão curtos, pelo que se irão formando filas de espera; só esporadicamente um intervalo de tempo será 'elevado', permitindo uma eventual regularização do sistema.

Se assumirmos que as durações do serviço (atendimento) se distribuem exponencialmente, estaremos a admitir que a maior parte dos clientes será atendida 'rapidamente' (mais rigorosamente, com durações de serviço inferiores à duração média), e que apenas um baixo número de clientes originarão durações de atendimento elevadas. Esta hipótese parece aceitável para situações em que o atendimento a diferentes clientes pode originar tarefas distintas, mas é menos adequada para situações em que o serviço a ser prestado a todos os clientes seja *idêntico*.

Propriedade 2: **A distribuição Exponencial não tem memória (Propriedade Markoviana).**

$$T \sim \text{Exponencial}, \quad P(T \leq a + b \mid T > a) = P(T \leq b)$$

Esta propriedade significa, quando T representa a distribuição dos intervalos de tempo entre chegadas consecutivas, que o intervalo de tempo até à próxima chegada é independente do instante que decorreu desde a última chegada – o que parece aceitável para a generalidade dos processos de chegadas dos clientes. É por este motivo que se diz que **a Distribuição Exponencial não tem memória**.

Propriedade 3: **O *mínimo* de várias variáveis aleatórias independentes com distribuição Exponencial é uma variável aleatória com distribuição Exponencial.**

Sejam $T_1, T_2, \ldots, T_n$ variáveis aleatórias independentes com distribuição Exponencial, de parâmetros, respectivamente, iguais a $\lambda_1, \lambda_2, \ldots, \lambda_n$ e

$$\mathbf{U = \text{mínimo}\ \{ T_1, T_2, \ldots, T_n \}}$$

Prova-se que $U \sim \text{Exponencial}(\lambda)$, com $\lambda = \sum_{i=1}^{n} \lambda_i$, ou seja, o mínimo de n v.a. Exponenciais é ainda Exponencial com parâmetro igual à soma dos parâmetros das n v.a..

Esta propriedade tem várias implicações:

i) supondo que há diferentes tipos de clientes, cada um dos quais tem um processo de chegadas com distribuição Exponencial com um parâmetro específico, pode concluir-se que o processo de chegadas *agregado*, correspondente aos vários tipos de clientes, ainda é descrito por uma distribuição Exponencial, com parâmetro igual à soma dos parâmetros *individuais*.

ii) se se tiver n servidores a assegurar o atendimento dos clientes, todos assegurando uma duração de serviço com distribuição Exponencial de parâmetro μ, num dado instante, o tempo necessário para o próximo final de serviço, de qualquer do n servidores, terá distribuição Exponencial com parâmetro $(n.\mu)$. Assim, o sistema comporta-se como se tivesse um único servidor, com duração de serviço com distribuição Exponencial de parâmetro $(n.\mu)$.

Esta propriedade muito útil no estudo dos modelos com múltiplos servidores.

Propriedade 4: **A distribuição Exponencial está relacionada com a distribuição de Poisson.**

Se o intervalo de tempo entre chegadas consecutivas tiver distribuição Exponencial, com parâmetro λ, então o número de chegadas por unidade de tempo t tem uma distribuição de Poisson, com parâmetro $m = \lambda t$. Refira-se, desde já, que o parâmetro **m** será igual ao valor médio da distribuição de Poisson.

λ representa a taxa média de ocorrências

Se para um processo de ocorrências, a distribuição do intervalo de tempo entre ocorrências consecutivas for Exponencial, e se os sucessivos intervalos de tempo entre ocorrências consecutivas forem independentes entre si, estaremos perante um **Processo de Poisson**.

A relação entre a distribuição Exponencial e a distribuição de Poisson é particularmente útil para se avaliar o *número de atendimentos efectuados* num dado intervalo de tempo t, admitindo que a duração do atendimento tem distribuição Exponencial com parâmetro μ. Assim, o número de atendimentos efectuados por um servidor no intervalo de tempo t pode caracterizar-se com uma distribuição de Poisson de média $m = \mu t$. No caso de termos um atendimento efectuado por n servidores, a distribuição de Poisson passará a ter média $m = n \mu t$.

Propriedade 5: **Na distribuição Exponencial de parâmetro λ, para <u>todos</u> os valores positivos de t, verifica-se que $P (T \leq t + \Delta t \mid T > t) \approx \lambda \Delta t$, para pequenos Δt.**

Se continuarmos a interpretar T como o intervalo de tempo que decorreu desde o último acontecimento (chegada, ou final de serviço), esta propriedade indica-nos que, qualquer que seja o tempo que já decorreu, a probabilidade de ocorrência de um novo acontecimento no próximo pequeno intervalo Δt é proporcional a Δt, sendo a constante de proporcionalidade λ (a taxa de ocorrências). De notar que o número esperado de ocorrências no intervalo Δt é exactamente igual a $\lambda . \Delta t$. A probabilidade de ocorrência, no entanto, pode diferir ligeiramente deste valor já que existe uma (baixíssima) probabilidade de que *mais do que um* acontecimento ocorra neste pequeno intervalo de tempo …

Propriedade 6: **A distribuição Exponencial não é afectada pela agregação, ou desagregação.**

Imaginemos que a um sistema chegam 3 tipos de clientes, segundo processos de Poisson independentes, com taxas de chegada λ_1, λ_2, λ_3. Para descrevermos o processo geral de chegadas dos clientes, poderemos *agregar* estes três processos num único processo de Poisson, com taxa de chegada $\lambda = \lambda_1+\lambda_2+\lambda_3$. Inversamente, se tivermos um processo de chegadas Poissoniano, com taxa de chegadas λ e, se existir uma probabilidade fixa de cada cliente pertencer a um determinado tipo (por exemplo, com 40 % de probabilidade será um cliente de tipo A e com 60 % de probabilidade um cliente de tipo B), poderemos *desagregar* o processo inicial em vários processos autónomos, <u>também de Poisson</u>, associados aos vários tipos de clientes (no nosso exemplo, com taxas de chegada, respectivamente iguais a $\lambda_A = 0{,}4 . \lambda$ e $\lambda_B = 0{,}6 . \lambda$).

Exercício:

Admita que o processo de chegadas de clientes a uma loja pode ser considerado um Processo de Poisson, com uma taxa de 5 chegadas por minuto. Caracterize a distribuição do intervalo de tempo entre duas chegadas consecutivas e a distribuição do número de chegadas por minuto.

Considerando o segundo como a unidade de tempo, o intervalo de tempo entre duas chegadas consecutivas poderá ser descrito por uma variável Exponencial de média 12 segundos (= 60 / 5), isto é com parâmetro $\lambda = 1/12$. O número de chegadas por minuto poderá ser descrito por uma distribuição de Poisson com parâmetro $m = \lambda\, t = (1/12)\, 60 = 5$.

Finalmente, recordemos que **a soma de k variáveis aleatórias, independentes e identicamente distribuídas, com distribuição Exponencial (λ), é uma variável aleatória Erlang-K** (Gama). Pelo Teorema do Limite Central, se k for muito elevado, a distribuição Erlang-k tenderá para a distribuição Normal.

Sejam X_i v.a. i.i.d, $X_i \sim$ Exponencial (λ),

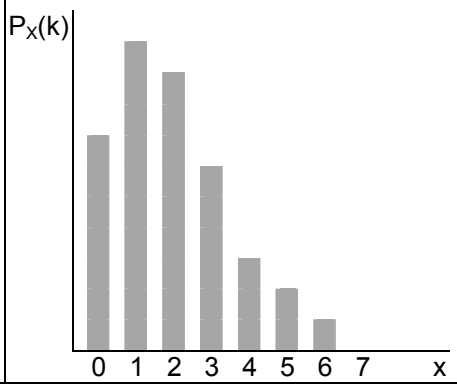
$$T \sim (X_1 + X_2 + \ldots + X_k) \sim \text{Erlang-k } (\lambda)$$

Como $E[X] = 1/\lambda$ e $Var[X] = 1/\lambda^2$, é fácil constatar que $E[T] = k/\lambda$ e $Var[T] = k/\lambda^2$.

Apresentemos agora, muito sumariamente, algumas características da **Distribuição de Poisson.** Seja X uma variável aleatória com distribuição de Poisson, com parâmetro m, isto é, $X \sim \text{Poisson}(m)$.

Função de distribuição de probabilidade:

$P_X(k) = P(X = k) = e^{-m}\, m^k / k!$, k = 0, 1, 2, …



μ = Valor Médio = m ; σ^2 = Variância = m

A Distribuição de Poisson é uma das (poucas) distribuições estatísticas que goza da **aditividade**, isto é, a soma de variáveis aleatórias independentes com distribuição de Poisson é ainda uma variável aleatória de Poisson (com parâmetro igual à soma dos parâmetros das variáveis que foram somadas).

Por outro lado, dado o Teorema do Limite Central (a soma de n variáveis independentes e identicamente distribuídas tende para a distribuição Normal, quando n se torna *elevado*), **poderemos aproximar a Distribuição de Poisson (m) da Distribuição Normal (com valor médio e variância iguais a m), quando m é *elevado*** (em termos práticos m maior do que 20). Esta aproximação permite efectuar o cálculo de probabilidades com maior facilidade … No entanto, há que ter em atenção o facto

de uma variável aleatória com distribuição de Poisson ser discreta, enquanto que a distribuição Normal descreve uma variável aleatória contínua … e fazer a chamada **correcção de continuidade** … (Remete-se o leitor mais esquecido para um compêndio de Estatística…).

Exercício FE01

Exercício FE02

O PROCESSO DE NASCIMENTO E MORTE

A maior parte dos modelos elementares de filas de espera baseia-se no **processo de nascimento e morte**. No contexto das filas de espera, um *nascimento* corresponde à chegada de um novo cliente e uma *morte* corresponde à partida de um cliente.

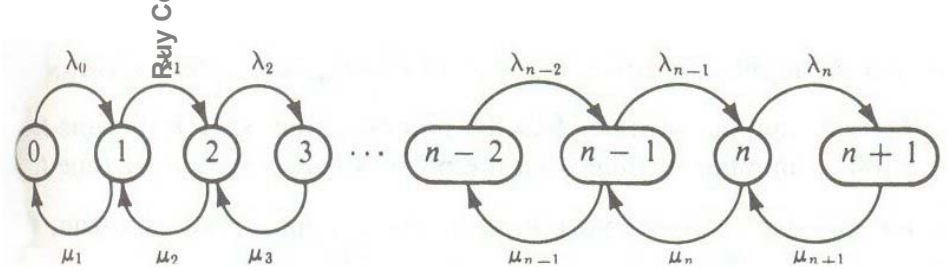
O **estado** do sistema no instante t é o número de clientes no sistema, denotado por **N(t)**. Um **processo de nascimento e morte** obedece a três hipóteses-base:

> **Hip.1:** Dado N(t) = n, a distribuição de probabilidade do tempo *restante* até ao próximo **nascimento** (chegada) é *Exponencial* com parâmetro λ_n (n = 0, 1, 2, …).
>
> **Hip.2:** Dado N(t) = n, a distribuição de probabilidade do tempo *restante* até à próxima **morte** (final de atendimento) é *Exponencial* com parâmetro μ_n (n = 0, 1, 2, …).
>
> **Hip.3:** Em cada instante só pode ocorrer ou um nascimento, ou uma morte.

As hipóteses 1 e 2 tornam o **processo de nascimento e morte** um tipo particular de Cadeias de Markov contínuas, o que facilita o tratamento das Filas de Espera que assim podem ser descritas. A hipótese 3 simplifica adicionalmente a análise.

Na figura seguinte esquematiza-se o **diagrama de transição** correspondente ao processo de nascimento e morte:

Estado:



λ_n: taxa média de chegadas (nº esperado de chegadas por unidade de tempo) de novos clientes quando **n** clientes estão no sistema.
μ_n: taxa média de serviço *global* * (nº esperado de atendimentos completados por unidade de tempo) quando **n** clientes estão no sistema.
[* *global* ⇔ taxa combinada relativa aos servitores ocupados]

As setas no diagrama representam as únicas transições possíveis no estado do sistema e os valores inscritos por cima, ou por baixo, de cada seta representam a respectiva taxa média para essa transição.

Depois de o sistema ter atingido o **estado de equilíbrio** (se tal for possível), o diagrama de transição facilita a determinação de resultados relevantes.

Há um **princípio fundamental** "**taxa de entrada = taxa de saída"**, que estipula que, **para qualquer estado do sistema (n = 0, 1, 2, …), a taxa média de entradas é igual à taxa média de saídas**. Este princípio permitirá escrever, para *todos* os estados, a respectiva **equação de equilíbrio**, em função das incógnitas P_n (probabilidades). A resolução do sistema de equações permitirá determinar o valor dessas probabilidades.

Ilustremos a utilidade do diagrama de transição para determinarmos as equações relativas aos estados 0 e 1:

Dado que só se pode entrar no estado 0, a partir do estado 1, a taxa média de entrada no estado 0 depende apenas da taxa média de entrada no estado 0 sabendo-se que o sistema está no estado 1, μ_1, e da probabilidade de ocorrência do estado 1, P_1, sendo igual a $\mu_1 . P_1$. Por outro lado, a taxa média de saída do estado 0 será igual a $\lambda_0 . P_0$. Assim, relativamente ao estado 0, poderemos escrever:

$$\mu_1 . P_1 = \lambda_0 . P_0$$

A entrada no estado 1 pode dar-se a partir do estado 0 (dependendo da taxa λ_0 e da probabilidade de ocorrência do estado 0, P_0), ou a partir do estado 2 (dependendo da taxa μ_2 e da probabilidade P_2), sendo, assim, a taxa média de entrada no estado 1 igual a $\lambda_0 . P_0 + \mu_2 . P_2$. Por outro lado, a saída do estado 1 pode dar-se para o estado 0 (dependendo da taxa μ_1 e da probabilidade P_1), ou para o estado 2 (dependendo da taxa λ_1 e da probabilidade P_1), sendo, assim, a taxa média de saída do estado 1 igual a $\mu_1 . P_1 + \lambda_1 . P_1$. Assim, relativamente ao estado 1, poderemos escrever:

$$\lambda_0 . P_0 + \mu_2 . P_2 = \mu_1 . P_1 + \lambda_1 . P_1$$

Raciocinado analogamente poderemos determinar as equações de equilíbrio para o processo de nascimento e morte:

Estado	**Equação de equilíbrio**
0	$\mu_1 . P_1 = \lambda_0 . P_0$
1	$\lambda_0 . P_0 + \mu_2 . P_2 = (\lambda_1 + \mu_1) . P_1$
2	$\lambda_1 . P_1 + \mu_3 . P_3 = (\lambda_2 + \mu_2) . P_1$
…	…
n	$\lambda_{n-1} . P_{n-1} + \mu_{n+1} . P_{n+1} = (\lambda_n + \mu_n) . P_n$
…	…

Como se pode observar, temos uma variável “a mais” em relação ao número de equações. Assim, dever-se-á resolver este sistema em função de uma das variáveis (em geral, P_0).

Resolvendo sequencialmente, e a partir da primeira equação, obtemos:

$$P_1 = \frac{\lambda_0}{\mu_1} P_0$$

$$P_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0$$

…

$$P_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} P_0$$

Para simplificar a notação, poderemos denotar, para n = 1, 2, … .

$$C_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n}$$

Assim, as probabilidades de equilíbrio são dadas por:

$$P_n = C_n P_0 , \quad \text{para } n = 1, 2, \ldots .$$

Como o somatório das probabilidades tem que igualar 1, obtém-se:

$$P_0 = 1 \ / \ (1 + \sum_{n=1}^{\infty} C_n)$$

Desde já poderemos avançar alguns **resultados gerais** aplicáveis a sistemas de filas de espera, baseados no processo de nascimento e morte:

- o **número médio de clientes no sistema** será:

$$L = \sum_{n=0}^{\infty} n \, . \, P_n$$

Se tivermos um sistema com s servidores, haverá s clientes que estarão a se atendidos, pelo que

- o **número médio de clientes a aguardar atendimento na fila (comprimento médio da fila de espera)** será:

$$L_q = \sum_{n=s}^{\infty} (n - s) \, . \, P_n$$

- o **tempo médio no sistema, por cliente** (incluindo a duração do atendimento) será:

$$W = L \ / \ \overline{\lambda} \qquad \textbf{Fórmula de Little}$$

$\overline{\lambda}$ designa a taxa média de chegadas, a longo prazo,

$$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n \, . \, P_n$$

• o **tempo médio a aguardar o atendimento, por cliente** (na fila de espera - exclui a duração do atendimento) será:

$$W_q = L_q \ / \ \overline{\lambda}$$

Embora algumas das expressões anteriores envolvam somatórios com um número infinito de termos, muitas vezes esses somatórios podem ser *resolvidos* analiticamente; noutros casos, poderão ser aproximados numericamente. De notar ainda que as expressões indicadas assumem que, com os valores assumidos pelos parâmetros λ_n e μ_n, se possa atingir o estado de equilíbrio – tal é sempre o caso quando existir um número finito n ($n > 0$) de estados; tal é sempre o caso quando $\rho = \lambda / (s \mu) < 1$; tal não será o caso se $\sum_{n=0}^{\infty} C_n = \infty$.

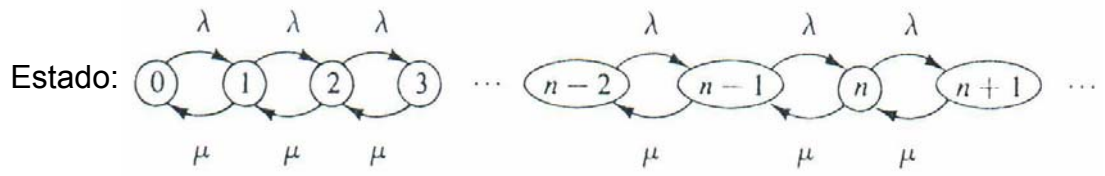
MODELOS DE FILAS DE ESPERA BASEADOS NO PROCESSO DE NASCIMENTO E MORTE (modelos com distribuições Exponenciais)

Os processos de nascimento e morte servem de base para modelar vários sistemas de Filas de Espera. Estes processos assumem um processo de chegadas Poissoniano (distribuição Exponencial para modelar o tempo *restante* até ao próximo nascimento, i.e., chegada) e um tempo de serviço (o tempo *restante* até à próxima morte, i.e., final de atendimento) também Exponencial. Estes modelos diferem essencialmente nas hipóteses assumidas para n, λ_n e μ_n.

Comecemos por considerar o

• Modelo M/M/1 com população infinita e fila ilimitada

Neste modelo assume-se que existe um único servidor, que os intervalos de tempo entre chegadas consecutivas são independentes e identicamente distribuídos, com distribuição Exponencial (λ), e que as durações dos serviços são independentes e identicamente distribuídas, com distribuição Exponencial (μ). Assim, neste caso, teremos $\lambda_n = \lambda$ (**taxa média de chegada** dos clientes) , para n = 0, 1, 2, … e $\mu_n = \mu$ (**taxa média de atendimento** dos clientes), para n = 1, 2, … . Assumiremos que a fila tem capacidade ilimitada que é *alimentada* por uma população infinita e que a disciplina praticada será FIFO. O **diagrama de transição** resultante será:



A partir das referidas taxas médias poderemos definir o **factor de utilização** ρ (por vezes designado por **intensidade de tráfego**):

$$\rho = \lambda / \mu$$

O factor de utilização representa o número esperado de chegadas durante um serviço médio. Assim, se $\rho > 1$, o ritmo das chegadas ultrapassa a capacidade de atendimento do servidor, pelo que *explodirá*, i.e., não se atingirá uma 'situação de equilíbrio'. **Se $\rho < 1$ o sistema poderá atingir uma 'situação de equilíbrio'**, ou seja o ritmo a que decorre o atendimento dos clientes é suficiente para dar vazão aos clientes que vão chegando.

Recordemos que λ representa a taxa de chegadas, e que estamos a assumir que é constante e independente do número de clientes já no sistema. As **Fórmulas de Little** permitem-nos relacionar L com W e L_q com W_q:

$$L = \lambda W \quad ; \quad L_q = \lambda W_q$$

De notar que se a taxa de chegada depender do estado do sistema, as Fórmulas de Little ainda são válidas desde que substituamos, nas expressões apresentadas, λ por $\overline{\lambda}$ (isto é, a taxa média de chegadas).

De notar que, para este modelo, $C_n = (\lambda / \mu)^n = \rho^n$. Assim, a probabilidade de estarem exactamente n pessoas no sistema, $\mathbf{P_n}$, será dada por:

$$P_n = \rho^n \, . \, P_0 = \rho^n \, (1 - \rho \,)$$

A **taxa de desocupação** do sistema, $\mathbf{P_0}$, isto é, a probabilidade de não haver clientes no sistema:

$$P_0 = 1 - \rho$$

A probabilidade de estarem mais do que K pessoas no sistema será dada por:

$$P(\, n > K \,) = \rho^{K+1}$$

O tempo médio de permanência de um cliente no sistema (W) e o tempo médio de espera na fila (W_q) podem ser relacionados facilmente se notarmos que $1/\mu$ corresponde ao tempo médio gasto no serviço:

$$W = W_q + \, 1 / \mu$$

Tendo em conta a expressão anterior e as Fórmulas de Little, poderemos escrever a relação entre o número médio de clientes no sistema (L), o comprimento médio da fila (L_q):

$$L = L_q + \, \lambda / \mu \, = L_q + \rho$$

De notar que, num sistema M/M/1, L_q não é igual a $L - 1$, mas sim a $L - \rho$!

A partir das expressões apresentadas poderemos deduzir os seguintes resultados:

$$L = \sum_{n=1}^{\infty} P_n \; = \frac{\rho}{1-\rho} \; = \frac{\lambda}{\mu - \lambda}$$

$$L_q = \frac{\rho^2}{1-\rho} \; = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

(cont.)

(continuação)

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

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

Quanto à **probabilidade de um cliente estar mais do que t unidades de tempo no sistema, ou na fila em espera** ($P(w > t)$, ou $P(w_q > t)$, respectivamente), poderemos obter:

$$P(w > t) = e^{-\mu(1-\rho)t} \quad \text{para } t \geq 0$$

$$P(w_q > t) = \rho\, e^{-\mu(1-\rho)t} \quad \text{para } t \geq 0$$

Aproveitamos para recordar que a **taxa de desocupação** do sistema, P_0, representa a probabilidade de não haver clientes no sistema, o que coincide com a probabilidade de um cliente não ter de esperar na fila, pois um cliente só é atendido assim que chega se não houver clientes no sistema. Assim,

$$P_0 = 1 - \rho = P(w_q = 0)$$

Em seguida, sintetizaremos os resultados válidos para um sistema M/M/1, alimentado por uma população infinita e sem limitações quanto ao comprimento máximo da fila de espera:

Sistema M/M/1, População = ∞; Fila máxima = ∞

Processo de **chegadas** Poissoniano com uma taxa de chegadas de λ clientes por unidade de tempo.

Duração do **serviço** com distribuição Exponencial Negativa – taxa de atendimento de μ clientes por unidade de tempo (pelo **único servidor**).

Disciplina da fila: FIFO (atendimento por ordem de chegada)

Taxa de **ocupação** $\rho = \lambda / \mu$ $(\rho < 1)$

Taxa de **desocupação** $= 1 - \rho = P_0 = P(w_q = 0)$

$$L = L_q + \lambda / \mu$$

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

$$L_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

$$W = W_q + 1 / \mu$$

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

$$W_q = L_q / \lambda = \frac{\rho}{\mu - \lambda}$$

$$P_0 = 1 - \rho = P(w_q = 0)$$

$$P_n = \rho^n P_0 = \rho^n (1 - \rho)$$

$$P(n > k) = \rho^{k+1}$$

$$P(w > t) = e^{-\mu(1-\rho)t} = e^{-t/W} \quad \text{para } t \geq 0$$

$$P(w_q > t) = \rho\, e^{-\mu(1-\rho)t} = \rho\, e^{-t/W} \quad \text{para } t \geq 0$$

Façamos agora um exercício de aplicação:

Exercício:

"O Docinho" é uma pequena pastelaria, sem lugares sentados, onde são vendidas especialidades regionais, pela sua única empregada. Pode-se considerar que as chegadas constituem um Processo de Poisson, com uma taxa de 15 chegadas por hora, estimando-se que a duração do atendimento de um cliente se possa considerar exponencialmente distribuído, com valor médio igual a 3 minutos.

1 - Determine:

a) a probabilidade de estar apenas um cliente na pastelaria;
b) a probabilidade de estarem, pelo menos, três clientes na pastelaria;
c) o comprimento médio da fila de espera;
d) o tempo médio de espera na fila;
e) a probabilidade de que um cliente esteja mais do que 5 minutos na pastelaria;
f) a probabilidade de que um cliente esteja mais do que 3 minutos à espera para começar a ser atendido.

2 – O proprietário d"O Docinho" está convencido de que seria possível diminuir o tempo médio de espera na fila para 6 minutos se a sua empregada aumentasse o ritmo de trabalho, diminuindo a duração média do atendimento de um cliente. Comente.

1 - $\lambda = 15\ h^{-1} = 15/60\ min^{-1}$; $\mu = 1/3\ min^{-1}$; $\rho = \lambda / \mu = 3/4$ ($\rho < 1$ ✓)

a) $P_1 = \rho\ (1 - \rho) = 3/4 \cdot 1/4 = 3/16 \approx$ **19 %**

b) $P_0 = 1 - \rho = 1/4$; $P_2 = \rho^2\ (1 - \rho) = 9/16 \cdot 1/4 = 9/64$. Assim, a probabilidade pedida é igual a $1 - P_0 - P_1 - P_2 = 27/64 \approx$ **42 %**

c) $L_q = \dfrac{\rho^2}{1-\rho} = 36/16 =$ **2,25 clientes**

d) $W_q = L_q / \lambda =$ **9 min**

e) $P(\mathcal{W} > 5) = e^{-\mu(1-\rho)\,5} \approx$ **66 %**

(ou, $W = \dfrac{1}{\mu - \lambda} = 12\ \text{min} \rightarrow P(\mathcal{W} > 5) = e^{-5/W} \approx 66\ \%$ ✓)

f) $P(\mathcal{W}_q > 3) = \rho\ e^{-\mu(1-\rho)\,3} \approx$ **58 %** (ou, $P(\mathcal{W}_q > 3) = \rho\ e^{-3/W} \approx 58\ \%$ ✓)

2 - $\lambda = 15/60 = 1/4\ min^{-1}$; $W_q = \dfrac{\frac{\lambda}{\mu}}{\mu - \lambda} = 6\ \text{min} \Leftrightarrow \mu^2 - 1/4\ \mu - 1/24 = 0$

$\Leftrightarrow \mu \approx 0{,}729\ min^{-1}$

$\Leftrightarrow$ Duração média do atendimento de 1 cliente = $1 / \mu \approx$ **1,37 min**.

Conclusão: Atingir este objectivo implica reduzir a duração média do serviço de 3,00 para 1,37 min ! Muito provavelmente, tal será difícil de atingir com uma única empregada!

Exercício FE03

E se, no exercício d' "O Docinho", tivéssemos duas empregadas, em vez de apenas uma ? O que aconteceria ? Para podermos responder a esta questão, apresentaremos em seguida o

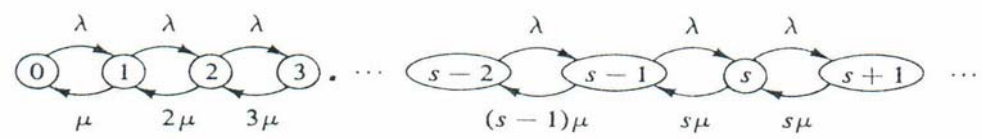
• Modelo M/M/S com população infinita e fila ilimitada

Agora temos *S* servidores, alimentado por uma população infinita e sem limitações quanto ao comprimento máximo da fila de espera. De notar que, se a taxa média de chegadas de clientes ao sistema continua a ser λ, independentemente do estado, a taxa média de serviço (que se assume ser μ por cada um dos *S* servidores), dependerá do estado do sistema:

$$\mu_n = \begin{cases} n\mu & ;n = 1,2,...,S \\ S\mu & ;n \geq S+1 \end{cases}$$

O **diagrama de transição** resultante será:

Estado:



Apresentaremos, em seguida, os resultados válidos para um sistema M/M/S, alimentado por uma população infinita e sem limitações quanto ao comprimento máximo da fila de espera:

Sistema M/M/S, População = ∞; Fila máxima = ∞

Processo de **chegadas** Poissoniano com uma taxa média de chegadas de λ clientes por unidade de tempo.

Duração do **serviço** com distribuição Exponencial Negativa com taxa média de μ clientes por unidade de tempo por cada um dos ***S* servidores**.

$$\mu_n = \begin{cases} n\mu & ;n = 1,2,...,S \\ S\mu & ;n \geq S+1 \end{cases}$$

Disciplina da fila: FIFO (atendimento por ordem de chegada)

Taxa de **ocupação** $\rho = \lambda / (S \mu)$ $(\rho < 1)$

Taxa de **desocupação** $= 1 - \rho$

$$L = L_q + \lambda / \mu$$

$$L_q = \frac{S^S \rho^{S+1} P_0}{S!(1-\rho)^2}$$

$$W = W_q + 1/\mu = L/\lambda$$

$$W_q = L_q / \lambda$$

$$P_0 = \left[\frac{S^S \rho^{S+1}}{S!(1-\rho)} + \sum_{n=0}^{S} \frac{(S\rho)^n}{n!} \right]^{-1}$$

$$P_n = \begin{cases} \frac{(S\rho)^n}{n!} P_0 & ;n = 1,...,S \\ \frac{S^S \rho^n}{S!} P_0 & ;n \geq S+1, \end{cases}$$

$$P(\mathcal{W} > t) = e^{-\mu.t} \left[1 + \frac{(S\rho)^S P_0 \left(1 - e^{-\mu.t(S-1-S\rho)}\right)}{S!(1-\rho)(S-1-S\rho)} \right] \quad \text{para } t \geq 0$$

$$P(\mathcal{W}_q > t) = \frac{(S\rho)^S P_0}{S!(1-\rho)} e^{-S\mu.t(1-\rho)} \quad \text{para } t \geq 0$$

$$P(\mathcal{W}_q = 0) = 1 - \frac{(S\rho)^S P_0}{S!(1-\rho)}$$

Façamos agora um exercício de aplicação:

Exercício:

Resolva a questão 1 do exercício d'"O Docinho", assumindo que há duas empregadas mantendo-se todas as outras características. Compare os resultados com os correspondentes quando há apenas uma mpregada, comentando.

$\lambda = 15\ h^{-1} = 15/60\ min^{-1}$; $\mu = 1/3\ min^{-1}$;

$S = 2$; $\rho = \lambda / (S\,\mu) = 3/8$ ($\rho < 1$✓)

a) a probabilidade de estar apenas um cliente na pastelaria

$$P_0 = \left[\frac{S^S \rho^{S+1}}{S!(1-\rho)} + \sum_{n=0}^{S} \frac{(S\rho)^n}{n!} \right]^{-1} = 0{,}45(45) \approx 45\ \%$$

$$P_1 = \frac{(S\rho)^1}{1!} P_0 \approx \mathbf{34\ \%}$$ (Compare-se com o valor obtido no Ex.4: 19 %)

b) a probabilidade de estarem, pelo menos, três clientes na pastelaria

$P_0 \approx 45\ \%$; $P_2 = \frac{(S\rho)^2}{2!} P_0 \approx 13\ \%$.

Assim, $1 - P_0 - P_1 - P_2 \approx$ **8 %** (Ex.4: 42 %)

De notar que se tivesse sido pedida a probabilidade de estarem exactamente três clientes na pastelaria, P_3, teríamos $\frac{S^2 \rho^3}{S!} P_0$, ou seja, $P_3 \approx 4{,}8\ \%$.

c) o comprimento médio da fila de espera

$$L_q = \frac{S^S \rho^{S+1} P_0}{S!(1-\rho)^2} = \mathbf{0{,}12\ clientes}$$ (Ex.4: 2,25 clientes)

d) o tempo médio na fila

$$W_q = \frac{L_q}{\lambda} = \mathbf{0{,}48\ min}$$ (Ex.4: 9,00 min)

e) a probabilidade de que um cliente esteja mais do que 5 minutos na pastelaria

$$P(\mathcal{W} > 5) = e^{-5\mu} \left[1 + \frac{(S\rho)^S P_0 \left(1 - e^{-5\mu(S-1-S\rho)}\right)}{S!(1-\rho)(S-1-S\rho)} \right] \approx \mathbf{5{,}3\ \%}$$ (Ex.4: 66 %)

f) a probabilidade de que um cliente esteja mais do que 3 minutos à espera para começar a ser atendido

$$P(\mathcal{W}_q > 3) = \frac{(S\rho)^S P_0}{S!(1-\rho)} e^{-3S\mu(1-\rho)} \approx \mathbf{5{,}9\ \%}$$ (Ex.4: 58 %)

Comentário: As diminuições esperadas do tempo médio de espera dos clientes e do comprimento médio da fila de espera (resultantes da introdução de uma segunda

empregada) são de tal modo significativas que indiciam que duas empregadas (a tempo inteiro) talvez sejam *de mais* …

Exercício FE04

E o que acontece se houver limitações físicas, que não permitam o desenvolvimento de uma fila de espera ilimitada ?

• Modelo M/M/1/K com população infinita e fila limitada

Consideremos que nas instalações onde decorre o serviço, não podem ser acomodados mais do que K clientes e, quando já estiverem K clientes no sistema e se verificar a chegada de um novo cliente, ser-lhe-á recusado o acesso ao sistema, i.e., trata-se de uma fila de espera com **capacidade finita**. De notar que os *potenciais* clientes com acesso vedado não poderão aguardar no exterior do sistema, para entrada posterior.

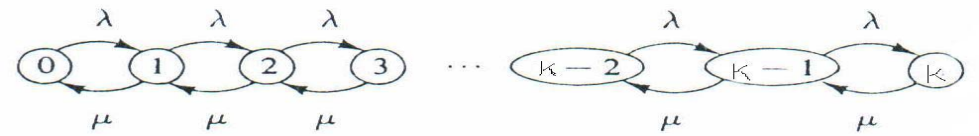
Neste caso a taxa média de entrada em cada estado será dependente do estado:

$$\lambda_n = \begin{cases} \lambda & ; n = 0,1,...,K-1 \\ 0 & ; n \geq K \end{cases}$$

De notar que se se está a admitir que a capacidade do sistema é limitada a um máximo de K clientes, os estados serão 0, 1, …, K –1, K.

O **diagrama de transição** resultante será:

Estado:



Caracterizaremos, em seguida, o sistema M/M/1/K com capacidade finita, assumindo que a população é ilimitada.

Sistema M/M/1/K, População = ∞; Fila máxima = K – 1

Número máximo de clientes no sistema = *K*

Processo de **chegadas** Poissoniano com uma taxa de chegadas de λ clientes por unidade de tempo. A taxa de **entradas** no sistema será dependente do estado *n* do sistema (isto é, do número *n* de clientes no sistema):

$$\lambda_n = \begin{cases} \lambda & ; n = 0,1,...,K-1 \\ 0 & ; n \geq K \end{cases} \quad ; \quad \bar{\lambda} = \lambda\,(1 - P_K)$$

Duração do **serviço** com distribuição Exponencial Negativa – taxa de atendimento de μ clientes por unidade de tempo (pelo **único servidor**).

Disciplina da fila: FIFO (atendimento por ordem de chegada)

Taxa de **pressão** $\rho = \lambda / \mu$

Taxa de **ocupação** $= \bar{\lambda}/\mu$

Taxa de **desocupação** $= 1 - \bar{\lambda}/\mu = P_0 = P(w_q = 0) = \dfrac{1-\rho}{1-\rho^{K+1}}$

$$L = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & ; \rho \neq 1 \\ \dfrac{K}{2} & ; \rho = 1 \end{cases}$$

$L_q = L - \bar{\lambda} / \mu$

$W = W_q + 1 / \mu$

$W = L / \bar{\lambda}$

$W_q = L_q / \bar{\lambda}$

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}} = P(w_q = 0)$$

$$P_n = \begin{cases} \rho^n P_0 & ; \rho \neq 1 \wedge n \leq K \\ 1/(K+1) & ; \rho = 1 \wedge n \leq K \\ 0 & ; n > K \end{cases}$$

E agora um exercício de aplicação ...

Exercício FE05

Em seguida, caracterizaremos o sistema M/M/S/K, com capacidade máxima para K clientes, S servidores, continuando-se a assumir que a população é ilimitada.

• Modelo M/M/S/K com população infinita e fila limitada

À semelhança do modelo anterior, não podem ser acomodados mais do que K clientes no sistema, i.e., trata-se de uma fila de espera com **capacidade finita** com S servidores.

Tal como no modelo anterior, a taxa média de entrada em cada estado será dependente do estado:

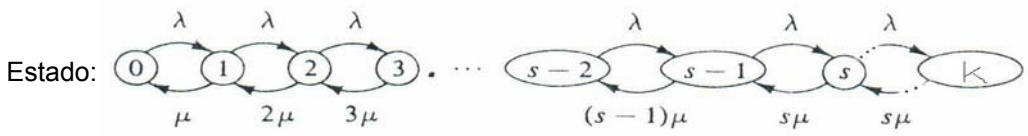
$$\lambda_n = \begin{cases} \lambda & ; n = 0,1,...,K-1 \\ 0 & ; n \geq K \end{cases}$$

A diferença relativamente ao modelo anterior reside na existência de S servidores, o que fará com que a taxa média de saída de cada estado também seja dependente do estado (à semelhança do que aconteceu no modelo M/M/S):

$$\mu_n = \begin{cases} n\mu & ; n = 1,2,...,S \\ S\mu & ; n \geq S+1 \end{cases}$$

À semelhança do modelo M/M/1/K, os estados serão 0, 1, …, K –1, K.

O **diagrama de transição** resultante será:



Caracterizemos, então, este modelo:

Sistema M/M/S/K, População = ∞; Fila máxima = K – S

$S \leq K$; Nº máximo de clientes no sistema = K; Nº de servidores = S

Processo de **chegadas** Poissoniano com uma taxa de chegadas de λ clientes por unidade de tempo. A taxa de **entradas** de clientes no sistema será dependente do estado n do sistema (isto é, do número n de clientes no sistema):

$$\lambda_n = \begin{cases} \lambda & ; n = 0,1,...,K-1 \\ 0 & ; n \geq K \end{cases} \quad ; \quad \overline{\lambda} = \lambda\,(1 - P_K)$$

Duração do **serviço** com distribuição Exponencial Negativa com taxa média de μ clientes por unidade de tempo por cada um dos ***S* servidores**.

$$\mu_n = \begin{cases} n\mu & ; n = 1,2,...,S \\ S\mu & ; n \geq S+1 \end{cases}$$

Disciplina da fila: FIFO (atendimento por ordem de chegada)

Taxa de **pressão** $\rho = \lambda / (S\,\mu)$

Taxa de **ocupação** = $\overline{\lambda} / (S\,\mu)$ $\qquad\qquad$ $\overline{\lambda} = \lambda\,(1 - P_K)$

Taxa de **desocupação** = $1 - \overline{\lambda} / (S\,\mu)$

$$P_0 = \begin{cases} \left[\dfrac{S^S \rho^{S+1}(1-\rho^{K-S})}{S!(1-\rho)} + \displaystyle\sum_{n=0}^{S} \frac{(S\rho)^n}{n!} \right]^{-1} & ; \rho \neq 1 \\ \left[\dfrac{S^S}{S!}(K-S) + \displaystyle\sum_{n=0}^{S} \frac{S^n}{n!} \right]^{-1} & ; \rho = 1 \end{cases}$$

$$P_n = \begin{cases} \dfrac{(S\rho)^n}{n!} P_0 & ; n = 1,...,S \\ \dfrac{S^S \rho^n}{S!} P_0 & ; n = S+1,...,K \\ 0 & ; n \geq K+1 \end{cases}$$

$$P(\mathcal{W}_q = 0) = \sum_{n=0}^{S-1} P_n$$

$$L_q = \frac{S^S \rho^{S+1} P_0}{S!(1-\rho)^2} \left[1 - \rho^{K-S} - (1-\rho)(K-S)\rho^{K-S} \right]$$

$W_q = L_q / \overline{\lambda}$

$W = W_q + 1/\mu$; $L = \overline{\lambda}\ W = L_q + \overline{\lambda} / \mu$

E agora um exercício de aplicação ...

Exercício FE05

• Modelo M/M/S/N com população finita e fila ilimitada

Imaginemos que numa fábrica de têxteis existem 15 teares que, quando se avariam, são reparados por dois técnicos de manutenção. Sabe-se que o intervalo de tempo entre duas avarias consecutivas se pode considerar com distribuição exponencial e que a reparação de cada tear avariado tem uma duração que também se pode considerar com distribuição exponencial.

Trata-se de um sistema M/M/2/15 alimentado por uma **fonte com dimensão finita** (15), ou **população finita**, onde os clientes serão os teares avariados e os dois servidores serão os técnicos de manutenção.

Em algumas aplicações industriais, tal como no exemplo apresentado, é muito importante considerar uma nova extensão dos sistemas M/M/1 e M/M/S: a **população com dimensão finita**, isto é uma fonte que possa gerar, no máximo, N clientes. Trata-se dos sistemas M/M/1/N e M/M/S/N com fonte com dimensão finita.

Neste modelo, a taxa média de entrada em cada estado será dependente do estado:

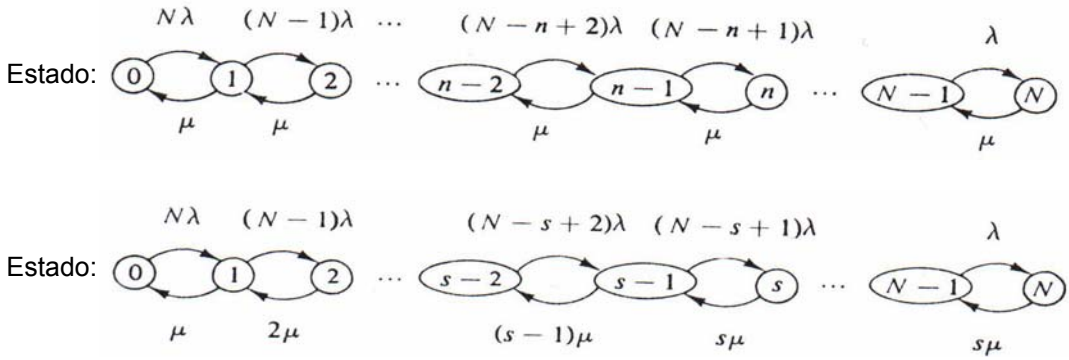
$$\lambda_n = \begin{cases} \lambda(N-n) & ; n = 0,1,...,N-1 \\ 0 & ; n \geq N \end{cases}$$

Se se estiver a admitir um único servidor, a taxa média de saída de cada estado será igual a μ, para todos os estados; caso o sistema tenha S servidores, a taxa média de saída de cada estado será dependente do estado (como sucedia no modelo M/M/S):

$$\mu_n = \begin{cases} n\mu & ; n = 1,2,...,S \\ S\mu & ; n \geq S+1 \end{cases}$$

De notar que se a população é finita, com dimensão N, os estados serão 0, 1, …, N-1, N.

Assim, os **diagramas de transição** correspondentes aos modelos M/M/1/N e M/M/S/N serão os seguintes:



Caracterizaremos, em seguida, o sistema M/M/S/N com fonte com dimensão finita, referindo alguma particularização decorrente da existência de um único servidor (S = 1).

Sistema M/M/*S*/*N*, População = *N* (Fila máxima = *N* – *S*)

$S \leq N$; Nº máximo de clientes no sistema = *N*; Nº de servidores = *S*

Processo de **chegadas** Poissoniano com uma taxa de chegadas de λ clientes por unidade de tempo. A taxa de **entradas** de clientes no sistema será dependente do estado *n* do sistema (isto é, do número *n* de clientes no sistema):

$$\lambda_n = \begin{cases} \lambda(N-n) & ;n = 0,1,...,N-1 \\ 0 & ;n \geq N \end{cases} \quad ; \quad \overline{\lambda} = \lambda\,(N - L)$$

Duração do **serviço** com distribuição Exponencial Negativa com taxa média de μ clientes por unidade de tempo por cada um dos ***S* servidores**.

$$\mu_n = \begin{cases} n\mu & ;n = 1,2,...,S \\ S\mu & ;n \geq S+1 \end{cases}$$

Disciplina da fila: FIFO (atendimento por ordem de chegada)

Taxa de **ocupação** = $\overline{\lambda}$ / ($S\,\mu$)

Taxa de **desocupação** = 1 – $\overline{\lambda}$ / ($S\,\mu$)

$$P_0 = \left[\sum_{n=0}^{S-1} \frac{N!}{(N-n)!n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=S}^{N} \frac{N!}{(N-n)!S!S^{n-S}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}$$

Caso particular **S = 1**:

$$P_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1} = \text{taxa de desocupação}$$

$$P_n = \begin{cases} \dfrac{N!}{(N-n)!n!} \left(\dfrac{\lambda}{\mu}\right)^n P_0 & ;n = 1,...,S \\ \dfrac{N!}{(N-n)!S!S^{n-S}} \left(\dfrac{\lambda}{\mu}\right)^n P_0 & ;n = S+1,...,N \\ 0 & ;n \geq N+1 \end{cases}$$

Caso particular **S = 1**:

$$P_n = \begin{cases} \dfrac{N!}{(N-n)!} \left(\dfrac{\lambda}{\mu}\right)^n P_0 & ;n = 1,...,N \\ 0 & n > N \end{cases}$$

$$P(w_q = 0) = \sum_{n=0}^{S-1} P_n$$

continua

continuação

Sistema M/M/*S*/*N*, População = *N* (Fila máxima = *N* – *S*)

$$L_q = \sum_{n=0}^{N} (n - S) P_n$$

Caso particular **S = 1**:

$$L_q = N - \frac{\lambda + \mu}{\lambda}(1 - P_0)$$

$W_q = L_q / \bar{\lambda}$

$W = W_q + 1/\mu \quad ; \quad L = \bar{\lambda}\, W = L_q + \bar{\lambda}/\mu$

Desde já poderemos observar que, para N *elevado*, se torna praticamente incomportável determinar P_0 e as restantes medidas de desempenho do sistema por cálculo manual !

E agora um exercício de aplicação ...

Exercício FE07

• Modelo com taxa de chegada e/ou taxa de serviço dependente do estado

Os modelos apresentados têm vindo a assumir uma taxa média de serviço constante, independente do número de clientes no sistema. No entanto, na realidade, quando os clientes são pessoas, muitas vezes o aumento do número de clientes no sistema vai *pressionando* o(s) servidor(es) que aumenta(m) a sua taxa média de serviço.

Analogamente, a taxa média de chegadas ao sistema pode não ser constante – os potenciais clientes ao verem o sistema com elevado número de clientes em espera, poderão optar por não entrar no sistema, fazendo com que a taxa média de chegadas seja, na realidade, decrescente com o número de clientes no sistema.

Descreveremos, em seguida, uma possível formulação destas situações, ainda a partir do processo de nascimento e morte. Por simplicidade, optaremos por fazê-lo **exclusivamente para o caso de um único servidor**, deixando ao leitor interessado na generalização a sugestão de uma consulta à Bibliografia.

Assumamos, então, que S = 1 e que

$$\mu_n = n^c \cdot \mu_1 , \quad \text{para } n = 1, 2, \ldots$$

n representa o número de clientes no sistema; μ_1 a taxa média de serviço, quando está apenas um cliente no sistema (situação sem *pressão*); μ_n a taxa média de serviço, quando estão n clientes no sistema (situação de *pressão*); e c é uma constante positiva, o *coeficiente de pressão*, que indica o grau de influência que o número de clientes tem sobre a taxa de serviço. De notar que nos modelos anteriormente apresentados se assumiu implicitamente c = 0.

Se assumirmos, adicionalmente, que o processo de chegadas é Poissoniano com $\lambda_n = \lambda$, para n = 0, 1, 2, … , poderemos obter os coefientes C_n correspondentes ao processo de nascimento e morte correspondente:

$$C_n = \frac{(\lambda / \mu_1)^n}{(n!)^c} , \quad \text{para } n = 1, 2, \ldots$$

Notemos que, desde que c > 0, se poderão atingir as condições de equilíbrio, pelo que poderão aplicar os resultados obtidos para o processo de nascimento e morte. Ainda que não exaustivamente, recordemos alguns desses resultados:

$$P_n = C_n P_0 , \quad \text{para } n = 1, 2, \ldots .$$

$$P_0 = 1 \ / \ (1 + \sum_{n=1}^{\infty} C_n)$$

$$L = \sum_{n=0}^{\infty} n \, . \, P_n$$

Infelizmente, neste modelo, não é possível obter expressões analíticas para os somatórios, que deverão ser determinados numericamente. Alguns livros apresentam alguns gráficos com a representação de algumas das relações, em função de alguns valores dos parâmetros envolvidos.

Exercício FE08

No modelo apresentado, modelou-se a pressão sobre a taxa média de serviço. Mas, como se referiu, pode ser importante assumir uma situação de pressão sobre a taxa média de chegadas ao sistema. Tal pode ser modelado do modo seguinte:

$$\lambda_n = (n + 1)^{-b} \cdot \lambda_0 , \quad \text{para } n = 0, 1, 2, \ldots$$

n, λ_0, λ_n e b são definidos de modo análogo ao anteriormente referido. Se se assumir que a taxa média de serviço é constante, $\mu_n = \mu$, teremos.

$$C_n = \frac{(\lambda_0 / \mu)^n}{(n!)^b}, \quad \text{para } n = 1, 2, \ldots$$

Poderemos utilizar estes coeficientes C_n nos resultados já referidos para a situação de equilíbrio do processo de nascimento e morte.

Exercício FE09

Um modelo mais geral permite uma modelação conjunta dos efeitos de pressão sobre as taxas médias de chegada e de serviço. Basta assumir-se conjuntamente:

$$\mu_n = n^a \cdot \mu_1 , \quad \text{para } n = 1, 2, \ldots$$
$$\lambda_n = (n + 1)^{-b} \cdot \lambda_0 , \quad \text{para } n = 0, 1, 2, \ldots$$

sendo a e b constantes de pressão definidas de modo análogo ao anteriormente referido. Ter-se-á, então

$$C_n = \frac{(\lambda_0 / \mu_1)^n}{(n!)^{a+b}}, \quad \text{para } n = 1, 2, \ldots$$

De notar que as três formulações se podem reduzir a uma única, sendo então necessário especificar o quociente que é elevado à potência n, bem como o coeficiente de pressão que é a potência à qual se eleva n!, o que permite utilizar os resultados gerais apresentados graficamente em alguns livros.

Só uma nota final para recordar a utilização de logaritmos, que pode ser útil, na resolução de problemas com este modelo. Suponha-se que, num dado exercício se assume que μ_1 = 2,5 clientes por hora e μ_6 = 5,0 clientes por hora. A determinação da constante de pressão faz-se facilmente: $5,0 = 6^c \cdot 2,5$.

$$\mathbf{5,0 = 6^c \cdot 2,5} \Leftrightarrow 6^c = 2 \Leftrightarrow \log_6(6^c) = \log_6(2) \Leftrightarrow c = [\ \ln(2) / \ln(6)\ ] \Leftrightarrow \mathbf{c = 0,3387}$$

Exercício FE10

Até agora apresentámos modelos de Filas de Espera, baseados no processo de nascimento e morte que, consequentemente, descreviam os intervalos de tempo entre chegadas consecutivas e as durações de atendimento com distribuições exponenciais.

Mas, se as chegadas forem previamente marcadas, ou, de algum modo, reguladas, o processo de chegadas perde o seu carácter Poissoniano. Analogamente, se as necessidades dos vários clientes, em termos de atendimento, forem idênticas, deixa de fazer sentido a utilização da distribuição Exponencial para descrever a duração do atendimento de cada cliente.

Neste casos, dever-se-á considerar modelos que envolvam distribuições não exponenciais.

MODELOS ENVOLVENDO DISTRIBUIÇÕES NÃO EXPONENCIAIS

• Modelo M/G/1 (Cadeias de Markov encaixadas)

No modelo M/G/1 assume-se um processo Poissoniano de chegadas de clientes, com taxa média fixa λ, que vão ser atendidos por um único servidor, não sendo postas quaisquer restrições à distribuição da duração do atendimento – é apenas necessário conhecer (ou estimar) o valor médio (que consideraremos ser igual a $1 / \mu$) e a variância σ^2 dessa distribuição (não especificada).

Neste caso, em geral, não será possível recorrer às equações de equilíbrio, uma vez que a duração dos atendimentos poderá apresentar "memória". Em alternativa, pode analisar-se o sistema imediatamente após a saída de cada cliente, constituindo uma cadeia de Markov que fica "encaixada" num processo não-Markoviano.

Se $\rho = \lambda / \mu < 1$ um tal sistema poderá eventualmente atingir o estado de equilíbrio, sendo então válidos os seguintes resultados:

$$P_0 = 1 - \rho$$

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

$$L = \rho + L_q$$

$$W_q = L_q / \lambda$$

$$W = W_q + 1 / \mu$$

É interessante notar que apesar deste modelo permitir qualquer distribuição para a duração do atendimento, se conseguem obter resultados analíticos, baseados na **Fórmula de Pollaczek-Khintchine** para a determinação de L_q:

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

Infelizmente, não foi possível determinar expressões analíticas para o caso de múltiplos servidores.

Se a duração do atendimento for Exponencial, $\sigma^2 = 1 / \mu^2$, obtendo-se os resultados já apresentados para o modelo M/M/1.

Uma última nota para o facto de, para um valor médio da duração do serviço constante, W, W_q, L e L_q aumentarem com o aumento da variância da distribuição da duração do serviço ! Tal mostra que o desempenho do servidor não é apenas relevante no que toca ao seu valor médio …

• Modelo M/D/1

Se continuarmos a assumir que o processo de chegadas é Poissoniano, que temos apenas um servidor e se o atendimento aos diferentes clientes consistir numa rotina relativamente idêntica, praticamente não se verificará qualquer variação na duração do serviço, sendo então útil o modelo M/D/1, que pode ser encarado como um caso particular do modelo M/G/1, fazendo $\sigma^2 = 0$.

Assim, a **Fórmula de Pollaczek-Khintchine** originará:

$$L_q = \frac{\rho^2}{2(1-\rho)}$$

É interessante notar que o valor indicado na fórmula acima é metade do correspondente valor para o modelo M/M/1: ou seja, se a distribuição do atendimento for Exponencial com parâmetro μ, o comprimento da fila de espera será duplo do que seria se todos os atendimentos fossem executados com duração determinística (igual a $1/\mu$). Fica assim patente a importância da variância da distribuição da duração do atendimento no desempenho do sistema !

Num sistema com múltiplos servidores (M/D/S) tudo se torna mais complicado – deixa-se ao leitor mais interessado a sugestão de uma leitura da Bibliografia. (Note-se que, no entanto, alguns livros apresentam alguns gráficos que representam as principais relações, em função dos valores dos parâmetros).

• Modelo M/E_k/1 (Método dos Estádios)

Como se referiu anteriormente, nos sistemas M/D/S assume-se que a duração do atendimento de um cliente é determinística ($\sigma = 0$) – uma situação teórica que raramente ocorre *rigorosamente* na prática. Num outro *extremo*, temos o s modelos M/M/S, em que se assume uma variação muito grande ($\sigma = 1 / \mu$ = duração média do atendimento de um cliente). Ora, na realidade, muitas vezes nem temos uma duração determinística, nem temos uma variação tão elevada – é para estes casos que se torna útil recorremos à **distribuição Erlang-k**.

Consideremos um sistema com um processo Poissoniano de chegadas, com taxa λ, e com a duração do atendimento de um cliente, T, com média $1/\mu$. Imaginemos que se sabe que a duração de cada atendimento **não** segue uma distribuição Exponencial e, que se assume que cada atendimento se pode decompor numa sequência de **k estádios** consecutivos, cada um deles com durações, T_i, independentes e identicamente distribuídas, com distribuição Exponencial de valor médio $1/(k\mu)$.

$T_1, T_2, \ldots, T_k$ v.a. i.i.d.

T_i ~ Exponencial de média igual a $1/(k\mu)$

$T \sim T_1 + T_2 + \ldots + T_k$

T ~ Erlang-k com valor médio igual a $1/\mu$ e variância igual a $1/(k\mu^2)$

A **distribuição Erlang-k** (mais rigorosamente, distribuição Erlang, com parâmetros k e μ), tem valor médio igual a $1/\mu$ e variância igual a $1/(k\mu^2)$. Assim, o **coeficiente de variação** da distribuição Erlang-k será $[\ 1/(\sqrt{k}\ \mu)\ ]\ /\ [\ 1/\mu\ ]$, ou seja, será igual a $1/\sqrt{k}$. De notar que como **k** influi directamente na variância da distribuição Erlang-k, costuma designar-se por **parâmetro de forma**.

De notar que o coeficiente de variação da distribuição Erlang-k é sempre menor, ou igual a 1 (a igualdade ocorre quando k = 1, i.e., quando a distribuição Erlang-k coincide com a distribuição Exponencial).

Assim, para utilizarmos o **método dos estádios**, começaremos por calcular o coeficiente de variação da distribuição da duração do atendimento de um cliente (que terá de ser inferior a 1, para que o método se possa utilizar). Imaginemos que o valor médio era igual a 15,00 minutos e que o desvio padrão era igual a 6,75 minutos – ter-se-ia, assim, o coeficiente de variação igual a 0,45. Fazendo $1/\sqrt{k}$ = 0,45, vem que k = 4,938, pelo que parece razoável adoptar-se k = 5. Adoptar-se-ia, então, a distribuição Erlang-k, com k = 5 e μ = 1/15 (adoptando o minuto como unidade de tempo).

O **Modelo $M/E_k/1$** pode ser caracterizado a partir da análise do correspondente diagrama de transição de estados (baseado no processo de nascimento e morte). De notar o cuidado inicial que se terá de ter na designação dos estados: 0; 1,k; 1,k-1; …; 1,2; 1,1; 2,k; …; 2, 1; 3,k; …;3, 1; …
No que diz respeito às chegadas, do estado 0 transita-se para o estado 1,1 com taxa média λ; … do estado 1,3 transita-se para o estado 2,3 com taxa média λ … No que diz respeito às finalizações de atendimento (ou, mais precisamente, de estádios de atendimento), passar-se-á do estado 2,k-1 para o estado 2,k, deste para o estado 1,1, deste para o estado 1,2, …, para o estado 1,k e, finalmente, para o estado 0, sendo cada taxa média de transição igual a $k\mu$.

Esta designação dos estados poderá ser, posteriormente, simplificada para uma designação unidimensional… Em seguida poderemos escrever as equações de equilíbrio para os vários estados e, após várias manipulações, deduzir alguns resultados.

No entanto, como já apresentámos o modelo mais geral M/G/1, poderemos encarar o modelo $M/E_k/1$ como um caso particular desse, com $\sigma^2 = 1/(k\mu^2)$. Assim, a fórmula de Pollaczek-Khintchine para a determinação de L_q será:

$$L_q = \frac{1+k}{2k}\,\frac{\lambda^2}{\mu(\mu-\lambda)}$$

Os demais parâmetros relevantes podem ser obtidos por aplicação das fórmulas apresentadas no modelo M/G/1, bem como alguns resultados gerais (p.ex., $L = \lambda W$).

Para os leitores mais interessados deixamos dois tópicos que poderão desenvolver com leituras complementares da Bibliografia:

1 - O método dos estádios também é aplicável aos sistemas $\mathbf{E_k/M/1}$ e $\mathbf{E_k/E_k/1}$.

2 – Se o coeficiente de variação da distribuição da duração do atendimento de um cliente (/distribuição das chegadas) for maior que 1, poder-se-á aplicar o método dos estádios, mas os estádios deverão desenvolver-se em paralelo (e não em série, como se apresentou).

Exercício FE11

• Modelos sem entradas Poissonianas

Os modelos M/ …/… assumem um processo de chegadas Poissoniano (intervalos de tempo entre chegadas consecutivas independentes e identicamente distribuídos, com distribuição Exponencial). No entanto, em certas situações, tal poderá não ser o mais adequado. Como proceder então?

Se a duração do atendimento de um cliente for Exponencial, com um parâmetro fixo, poderemos obter, de imediato, três modelos por *inversão* das distribuições assumidas para as chegadas e para os atendimentos, nos modelos M/G/1, M/D/1, $M/E_k/1$, obtendo então os modelos G/M/1, D/M/1 e $E_k/M/1$.

O modelo G/M/1 não impõe qualquer restrição à distribuição associada ao processo de chegadas; o modelo D/M/1 assume chegadas a intervalos regulares; o modelo $E_k/M/1$ permite modelar um processo de chegadas que, não sendo Poissoniano, também não é determinístico (intervalos de tempo constantes). Para alguns destes modelos, bem como as suas versões com múltiplos servidores, foi possível representar graficamente algumas relações com interesse.

MODELOS DE FILAS DE ESPERA COM DUISCIPLINA PRIORITÁRIA

Em certos sistemas de filas de espera o atendimento não é feito apenas por ordem de chegada, mas existe um sistema de prioridades, pelo que o atendimento de um cliente é feito pela respectiva prioridade.

O tratamento analítico de sistemas com prioridades é, obviamente, mais complicado do que o de sistemas sem prioridades. Como consequência, apenas se dispõe maioritariamente de resultados para o caso de um único servidor. Contudo, há um sistema com múltiplos servidores que apresenta resultados interessantes. Caracterizemo-lo:

- Assume-se que existem **N classes de prioridade** (a classe 1 com prioridade mais elevada e a classe N com mais baixa prioridade). Os clientes são atendidos por ordem das suas classes de prioridade e, dentro da cada classe, por ordem de chegada;

- Assume-se que o **processo de chegadas é Poiss**oniano, permitindo-se que a taxa de chegadas de clientes das várias classes possa ser diferente;

- Assume-se que as **durações de atendimento são Exponenciais** para cada classe, assumindo-se, adicionalmente, que a duração média de atendimento é igual para todas as classes.

De notar que, se se ignorar as prioridades, estaremos perante o modelo M/M/S. Assim, quando contabilizarmos o número **total** de clientes no sistema, poderemos considerar as distribuições limite apresentadas para o modelo M/M/S. Consequentemente, para um cliente seleccionado aleatoriamente, são válidas as expressões obtidas para L, L_q, W e W_q nesse modelo. O que muda é a distribuição do tempo de espera: num modelo com prioridades a variância da distribuição do tempo de espera aumenta – teremos clientes de prioridade mais elevada com tempos de espera mais baixos do que ocorreriam com a disciplina FIFO sem prioridades e, como se esperaria, clientes de prioridade mais baixa com tempos de espera mais elevados … O que não é de estranhar já que se pretende melhorar o desempenho do sistema no que diz respeito aos clientes de mais elevada prioridade, à custa de um pior desempenho para os clientes de mais baixa prioridade.

Assim, é importante calcular o **tempo de espera médio para um cliente de cada classe de prioridade**.

Assumamos que **prioridades "não absolutas"** (*nonpreemptive priorities*), i.e., um cliente que está a ser atendido, não vê o seu atendimento interrompido pela chegada de um cliente com mais elevada prioridade.

Assumindo **prioridades "não absolutas"**, $\mathbf{W_k}$, o **tempo de espera médio para um cliente da classe de prioridade k** (incluindo a duração do atendimento) será dado por:

$$W_k = \frac{1}{A.B_{k-1}.B_k} + \frac{1}{\mu}, \quad \text{para k = 1, 2, …, N}$$

Com $A = S!\left(\frac{S\mu - \lambda}{r^S}\right)\sum_{j=0}^{S-1}\frac{r^j}{j!} + S.\mu$,

$B_0 = 1$,

$$B_k = 1 - \frac{\sum_{i=1}^{k}\lambda_i}{S\mu}, \quad \text{para k = 1, 2, …, N,}$$

e S = número de servidores,
μ = taxa média de serviço por cada servidor ocupado,
λ_i = taxa média de chegadas da classe de prioridade i, i = 1, 2, …, N
$\lambda = \sum_{i=1}^{N} \lambda_i$ e

$r = \lambda / \mu$

Estes resultados assumem que $\sum_{i=1}^{k} \lambda_i < S . \mu$, de modo a que a classe de prioridade k possa atingir um estado de equilíbrio. Para cada classe de prioridade aplica-se a Fórmula de Little, pelo que o **número esperado de clientes da classe de prioridade k no sistema (incluindo os que estão a ser atendidos)** será

$$L_k = \lambda_k . W_k, \quad \text{para k = 1, 2, … N.}$$

Notas: 1) Para a classe k, o **tempo médio de espera a aguardar atendimento**, será igual a $W_k - 1 / \mu$ para k = 1, 2, … N;

2) O **comprimento médio da fila de espera** correspondente à classe k será igual a $\lambda_k . (W_k - 1 / \mu)$, para k = 1, 2, … N.

3) Se S = 1, $A = \mu^2 / \lambda$.

Assumindo **prioridades "absolutas"** (*preemptive priorities*), i.e., o atendimento de um cliente será interrompido (e re-enviado para a fila de espera) pela chegada de um cliente com mais elevada prioridade, e mantendo as demais hipóteses já referidas, **W_k,** o **tempo de espera médio para um cliente da classe de prioridade k (incluindo a duração do atendimento)** será, **para um único servidor**, dado por:

$$W_k = \frac{1/\mu}{B_{k-1}.B_k}, \quad \text{para } k = 1, 2, \ldots, N.$$

$$L_k = \lambda_k \, . \, W_k \, , \quad \text{para } k = 1, 2, \ldots N.$$

Para o caso de múltiplos servidores, dever-se-á adoptar m processo iterativo (para os leitores mais interessados, recomenda-se a consulta de Hillier e Lieberman).

Os correspondentes resultados para a fila de espera (excluindo os clientes que estão a ser atendidos) obtêm-se a partir de W_k e L_k com se referiu para as prioridades "não absolutas".

Refira-se, finalmente, que dado que a distribuição Exponencial não tem memória, as interrupções de atendimento não afectam, em média, o processo de atendimento: a duração média total do atendimento continua a ser igual a $1 / \mu$. Quando um cliente com atendimento interrompido voltar a ser atendido, a distribuição da duração do atendimento *restante* continuará a mesma. De notar que tal não ocorre para nenhuma outra distribuição da duração de um atendimento !

Exercício FE12

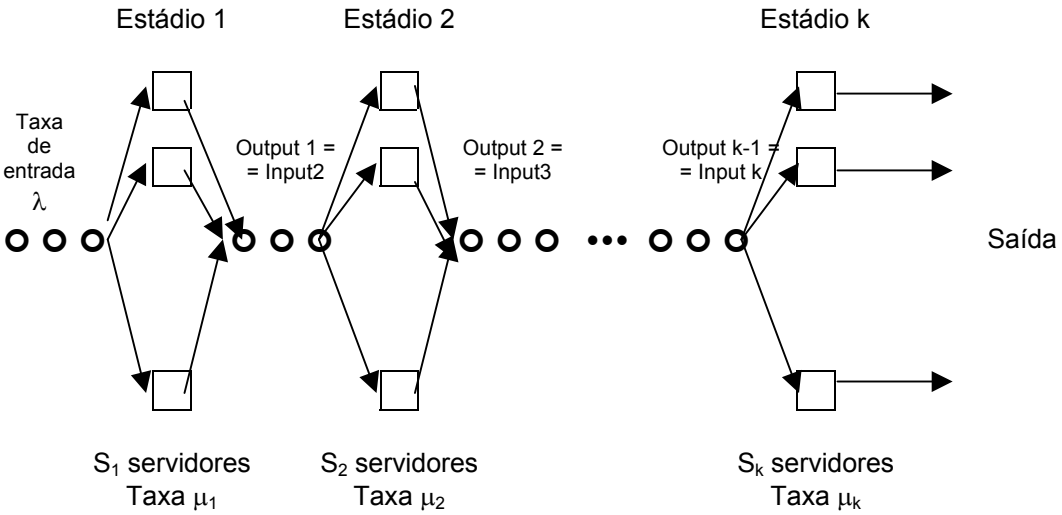
REDES DE FILAS DE ESPERA

Até agora temos vindo a considerar sistemas de Filas de Espera com um único local de atendimento (ainda que com um ou mais servidores). No entanto, em muitas situações reais, um cliente tem de passar por uma sequência de filas de espera (seguindo, ou não, uma determinada ordem) – o *output* de algumas dessas filas será o *input* de outras. Estaremos, assim, perante um sistema de **redes de filas de espera**. Quando tal ocorre, é importante estudar globalmente a rede para determinar o tempo total de espera, ou o número total de clientes no sistema.

Dada a dificuldade de modelação destes sistemas, a maior parte dos modelos divulgados assume processos Poissonianos de chegada e durações de atendimento exponenciais.

• Filas ilimitadas em série

Consideremos um sistema constituído por k filas (sem limite de capacidade), em série), como se esquematiza em seguida:



O importantíssimo **Teorema de Jackson**, garante-nos que:

> Se
> (1) o processo de chegadas dos clientes a um sistema de espera for Poissoniano com taxa λ,
> (2) as durações dos atendimentos dos servidores em cada estádio forem exponenciais, com parâmetro μ_i, e
> (3) cada estádio permitir a formação de uma fila ilimitada (modelo M/M/S), com $S.\mu > \lambda$,
>
> então o processo de saídas dos clientes de cada estádio do sistema de espera é Poissoniano, com taxa λ.

Façamos alguns comentários breves ao Teorema de Jackson:

i) De notar que **não é feita qualquer restrição à disciplina** das filas !

ii) Se o processo de saídas dos clientes de cada estádio do sistema de espera é Poissoniano, com taxa λ, então o processo de chegadas a cada estádio é Poissoniano, com taxa λ.

iii) Em condições de equilíbrio, **cada estádio k poderá ser analisado independentemente dos outros**: será *alimentado* por um processo Poissoniano de chegadas, terá S_k servidores, com duração de atendimento exponencial, com taxa μ_k: assim, poderá ser tratado com um modelo M/M/S_k (com, ou sem prioridades).

iv) Infelizmente, na realidade, nem sempre é possível garantir a formação de filas ilimitadas em todos os estádios, pelo que o Teorema de Jackson nem sempre poderá ser invocado.

Realcemos, pela sua extraordinária importância, o terceiro comentário acima apresentado. A possibilidade de se utilizar um modelo M/M/S para cada estádio, independentemente dos outros, é uma enorme simplificação. Por exemplo, a probabilidade conjunta de se ter n_1 clientes no estádio 1, n_2 clientes no estádio 2, …, n_k clientes no estádio k poderá, assim, obter-se simplesmente como o produto das probabilidades individuais:

$$P(N_1 = n_1 \wedge N_2 = n_2 \wedge \ldots \wedge N_k = n_k) = P_{n1} . P_{n2} . \ldots . P_{nk}$$

Esta forma da solução designa-se por **forma de produto** (*product form*). Os sistemas com filas com capacidade limitada não apresentam soluções na forma de produto !

Exercício FE13

• Redes de Jackson

A utilização dos modelos M/M/S independentemente para cada estádio ocorre também noutros contextos, para além das filas ilimitadas em série. As **Redes de Jackson** também permitem essa abordagem. Nas filas ilimitadas em série os clientes têm de percorrer obrigatoriamente todos os estádios, em sequência (estádio1, estádio2, …, estádio k); nas redes de Jackson os clientes podem nem visitar todos os estádios, poderão visitá-los por qualquer ordem e, para cada estádio, os clientes poderão ser provenientes quer de outros estádios, quer do exterior (segundo um processo Poissoniano). Resumamos, então, as características deste tipo de sistema:

Uma **Rede de Jackson** é um sistema de k estádios, onde o estádio i (i = 1, 2, …, k) tem:

1) uma fila ilimitada;
2) os clientes chegam do exterior do sistema de acordo com um processo Poissoniano com parâmetro a_i e
3) S_i servidores, que asseguram uma distribuição de atendimento exponencial, com parâmetro μ_i.

Um cliente que deixe o estádio i segue para outro estádio j (j = 1, 2, …, k e $j \neq i$) com probabilidade p_{ij}, ou partirá do sistema com probabilidade $q_i = 1 - \sum_{\substack{j=1 \\ j \neq i}}^{k} p_{ij}$.

Text

Em situação de equilíbrio, **cada estádio j de uma rede de Jackson (j = 1, 2, …, k) comporta-se como se fosse um sistema M/M/S independente**, com taxa de chegadas λ_i:

$$\lambda_j = a_j + \sum_{\substack{i=1 \\ i \neq j}}^{k} \lambda_i . p_{ij} \,, \qquad \text{com } S_j . \mu_i > \lambda_j$$

Intuitivamente poderemos compreender este resultado, recordando-nos que o carácter Poissoniano de um processo de Poisson não é afectado pela desagregação do processo, ou pela sua agregação a outros processos Poissonianos. Ora se sabemos que o processo de saídas de cada estado é Poissoniano, se sabemos que o processo de chegadas do exterior é Poissonano (que se desagregará, em contribuições para entradas em cada estádio directamente do exterior), pode concluir-se que o processo de entrada em cada estádio é uma agregação de contribuições Poissonianas e, assim, é um processo Poissoniano.

Exercício (Hillier e Lieberman)**:**

Considere uma Rede de Jackson, com os dados seguintes:

Estádio j	S_j	μ_j	a_j	Pij		
				i = 1	i = 2	i = 3
j = 1	1	10	1	–	0,1	0,4
j = 2	2	10	4	0,6	–	0,4
j = 3	1	10	3	0,3	0,3	–

a) Determine as taxas de entrada nos diferentes estádios.
b) Determine o número total de clientes no sistema.
c) Determine o tempo total esperado de permanência no sistema por cliente.

a) Escrevamos as equações $\lambda_j = a_j + \sum_{\substack{i=1 \\ i \neq j}}^{k} \lambda_i . p_{ij}$:

$$\begin{cases} \lambda_1 = 1 + 0{,}1\,\lambda_2 + 0{,}4\,\lambda_3 \\ \lambda_2 = 4 + 0{,}6\,\lambda_1 + 0{,}4\,\lambda_3 \\ \lambda_3 = 3 + 0{,}3\,\lambda_1 + 0{,}3\,\lambda_2 \end{cases}$$

Resolvendo o sistema, obtém-se $\lambda_1 = 5$, $\lambda_2 = 10$, $\lambda_3 = 7{,}5$.

Assim, podemos considerar que, para cada estádio i, se tem um sistema M/M/S cm taxa de entradas λ_i, taxa de serviço μ_i e S_i servidores.

b)
Estado 1: fila M/M/1, com $\lambda_1 = 5$ e $\mu_1 = 10$.
$\rho = 5/10 = 0{,}5$; $P_n = \rho^n . P_0$ e $P_0 = 1 - \rho$, ou seja, $\mathbf{P_{n1} = 0{,}5^{n+1}}$
$L = \lambda/(\mu-\lambda)$, ou seja $\mathbf{L_1 = 1}$.

Estado 2: fila M/M/2, com $\lambda_1 = 10$ e $\mu_1 = 10$.

$\rho = 10/(10 . 2) = 0{,}5$; $P_0 = 1/3$, ou seja, $\mathbf{P_{n2}} =$ $\begin{cases} \mathbf{1/3,} & \mathbf{para\ n_2 = 0} \\ \mathbf{1/3,} & \mathbf{para\ n2 = 1} \\ \mathbf{(1/3)).(1/2)^{n2+1}} & \mathbf{para\ n2 \geq 2} \end{cases}$

$L = L_q + \lambda/\mu = P_0 . (\lambda/\mu)^S . \rho / (S! . (1-\rho)^2) + \lambda/\mu$, ou seja $\mathbf{L_2 = 4/3}$.

Estado 3: fila M/M/1, com $\lambda_1 = 7{,}5$ e $\mu_1 = 10$.
$\rho = 7{,}5/10 = 0{,}75$; $P_n = \rho^n . P_0$ e $P_0 = 1 - \rho$, ou seja, $\mathbf{P_{n3} = 0{,}75^n . 0{,}25}$
$L = \lambda/(\mu-\lambda)$, ou seja $\mathbf{L_3 = 3}$.

Assim, a função de probabilidade conjunta (não pedida), pode escrever-se na forma produto:

$$P (N_1 = n_1 \wedge N_2 = n_2 \wedge N_3 = n_3) = P_{n1} . P_{n2} . P_{n3}$$

Quanto ao número de clientes no sistema, teremos $L = L_1 + L_2 + L_3 = 5{,}33(3)$.

c) A determinação do tempo total esperado de permanência no sistema por cliente não pode ser feita tão imediatamente. Com efeito, como nem todos os clientes são obrigados a ir a todos os estádios, não poderemos somar os tempos correspondentes a cada estádio. No entanto poderemos ainda utilizar a Fórmula de Little, considerando que a taxa global de chegadas de clientes vindos do exterior é $\lambda = a_1 + a_2 + a_3 = 8$. Assim, **$W = L / \lambda = 2/3$** (unidades de tempo).

Exercício FE14

Naturalmente, as Redes de Filas de Espera não se esgotam nos dois modelos apresentados. As **filas de espera com bloqueio** (devido à limitação da capacidade das filas em série, quando uma fila a jusante atinge o seu limite de capacidade, produz-se um bloqueio nas filas a montante, impedindo aí o processamento de clientes), as **redes fechadas de Jackson** (que consideram uma população limitada a N clientes, que continuamente *re-alimentam* o sistema) e as **filas cíclicas** (rede fechada em que o output da 1ª fila é o input da 2ª fila, o output da 2ª fila é o input da 3ª fila, …, o output da k-ésima fila é o input da 1ª fila) são algumas extensões das Redes de Filas de Espera. O leitor interessado nalgum destes tópicos é remetido para a Bibliografia (em particular, Gross, D. and C. Harris, “Fundamentals of Queueing Theory”, J. Wiley, New York (1985)).

CONCLUSÃO

A **importância das Filas de Espera** é evidente no dia-a-dia e em variados contextos. Assim, é evidente que a gestão adequada de um sistema de filas de espera tem repercussões na qualidade de vida e na produtividade.

Na modelação matemática de sistemas de filas de espera a **distribuição Exponencial** tem um papel fulcral, ainda que em determinadas situações possa ser útil considerar outras distribuições, nomeadamente a **Erlang-k**.

De referir ainda a necessidade de, em certos sistemas, se tornar necessário separar os clientes em diferentes classes, cada uma das quais com um nível de **prioridade** distinto.

Quando um cliente precisa de recorrer a vários serviços, num mesmo sistema, torna-se útil modelá-lo como uma **rede de filas de espera**.

Refira-se, finalmente, que quando há particularidades especiais, não contempladas em qualquer modelo conhecido, poderemos recorrer à **simulação de filas de espera**.

BIBLIOGRAFIA ESPECÍFICA

- Gross, D. and C. Harris, “Fundamentals of Queueing Theory”, J. Wiley, New York (1985).
- Hillier, F. and G. Lieberman, “Introduction to Operations Research”, McGraw-Hill Int. Editions (5 ed., 1990);
- Winston, W, “Operations Research – Applications and Algorithms”, Duxbury Press (1994)