3.36pt

Teoria das Filas

André Rodrigues da Cruz

Centro Federal de Educação Tecnológica de Minas Gerais
Departamento de Computação

Otimização II

Teoria das Filas

- Estudo das **filas de espera**.
- *Exemplos*: Filas para compra, caixa eletrônico, máquinas para reparo, aviões decolarem (pousarem), veículos para serem descarregados, ...
- **Modelos de filas**: representação de *sistemas de filas*.
- Operação eficiente de filas.
- Fornecer capacidade de atendimento.
- Minimizar tempo de espera.
- Buscar equilíbrio entre custo de serviço e tempo de espera.

Exemplo Protótipo

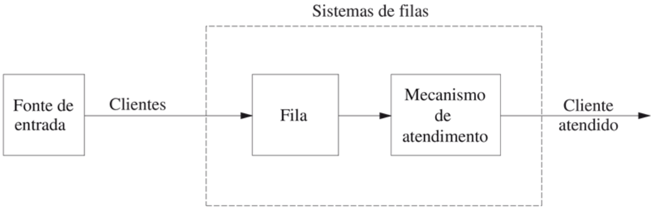
Um pronto de socorro de um hospital atende emergências. A qualquer momento há um médico de plantão. O hospital possui um aumento anual contínuo no número de atendimentos no pronto de socorro. É comum pacientes chegarem no horário de pico (início da noite) e ter de esperar atendimento médico.

Por esta razão, foi elaborada uma proposta de alocar um segundo médico para o pronto de socorro no horário de pico. Assim, duas emergências poderiam ser atendidas ao mesmo tempo.

O administrador do hospital foi designado para estudar esta questão.

Estrutura Básica dos Modelos de Fila

- Processo de filas básico:
  1. Clientes são gerados por uma *fonte de entrada*.
  2. Clientes entram no *sistema de filas* (pega a fila).
  3. Em certos momentos, um cliente é selecionado para atendimento.
     - Uma regra (*disciplina da fila*) conhecida é utilizada.
  4. Atendimento do cliente realizado por um *mecanismo de atendimento*.
  5. Cliente deixa a fila.



Fonte de Entrada

- **População solicitante**.
- Tamanho finito ou infinito.
- Fonte de entrada limitada ou ilimitada.
- Quando população é muito grande e finito, considera-se infinito.
  - Hipótese implícita, caso não afirme o contrário.
- Geração de clientes ao longo do tempo:
  - Padrão estatístico deve ser informado.
  - Número de clientes até dado momento:
    - Hipótese comum: *processo de Poisson*.
  - **Tempo entre chegadas** consecutivas:
    - Hipótese comum: distribuição *exponencial*.
- Especificação de hipóteses incomuns. *Exemplo*: recusa.

Fila

- Local onde clientes aguardam atendimento.
- Número máximo de clientes permitidos.
- Filas finitas ou infinitas.
- Filas finitas muito grandes: consideradas infinitas.
  - Facilidade dos cálculos.
  - Comportamento parecido.

Disciplina da Fila

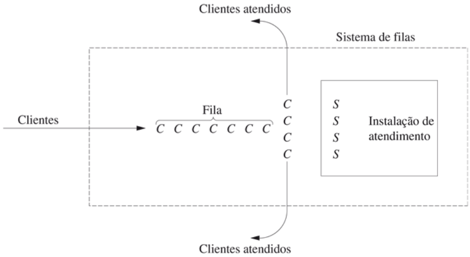
- Ordem de seleção dos integrantes da fila para atendimento.
- *Exemplos*:
    - Primeiro a chegar, primeiro atendido.
    - Aleatório.
    - Com prioridade.
    - ...

Mecanismo de Atendimento

- Conjunto de *instalações de atendimento*.
- Canais de atendimento paralelos: **atendentes**.
- Em uma instalação:
  - Clientes podem ser atendidos em série.
- Mais de uma instalação de atendimento:
  - Cliente em algum canal de atendimento paralelo.
  - Cliente é totalmente atendido por um atendente.
- Modelo de fila específica:
  - Disposição das instalações.
  - Números de atendentes (canais paralelos) de cada instalação.
- Comumente se tem uma instalação com um ou mais atendentes.
- **Tempo de atendimento** (tempo de permanência):
  - Possui distribuição probabilística para cada atendente.
  - Comum supor mesma distribuição para todos atendentes.
  - Distribuições comuns: exponencial, constante e Erlang (Gamma).

Processo de Fila Elementar

- Uma fila de espera única.
- Possui um ou mais atendentes.
- Cada cliente gerado por uma fonte de entradas.
- Cada cliente atendido por um atendente.
- Cliente aguarda um tempo para ser atendido.



- O exemplo do hospital é uma fila elementar.

Processo de Fila Elementar

Observações:

- *Um atendente não precisa ser um indivíduo*.
  Pode ser uma equipe, uma máquina, um veículo, um dispositivo eletrônico, . . .
- *Os clientes não precisam ser pessoas*.
  Podem ser peças que aguardam uma operação, carros que aguardam no pedágio, . . .
- *Não é necessário uma fila física*.
  A fila pode ser formada por máquinas em diferentes áreas aguardando reparo, jogadores online aguardando a vez de jogar, . . .

Processo de Fila Elementar

Observações:

- É fornecido número médio que aguardam, tempo de espera médio, ...
- Tempos entre **chegadas** e de **atendimentos** são independentes e distribuídos de forma idêntica.
- Identificação convencional: $\mathcal{A}/\mathcal{B}/\mathcal{C}$
  1. $\mathcal{A}$: Distribuição de tempos entre atendimentos.
  2. $\mathcal{B}$: Distribuição de tempos de atendimentos.
  3. $\mathcal{C}$: Número de atendentes.
- Distribuições comuns:
  - $M =$ distribuição exponencial (markoviana).
  - $D =$ distribuição degenerada (tempos constantes).
  - $E_k =$ distribuição de Erlang (parâmetro de forma k).
  - $G =$ distribuição geral (qualquer distribuição arbitrária).
- *Exemplo*: $M/M/s$, tempos entre atendimento e de atendimento com distribuição exponencial com s atendentes.

Terminologia e Notação

- **Estado do sistema**: número de clientes no sistema de filas.
- **Comprimento da fila**: número de clientes que aguardam atendimento.
  - Estado menos o número de clientes em atendimento.
- $N(t)$: número de clientes no sistema de fila no instante $t \geq 0$.
- $P_n(t)$: probabilidade de exatamente n clientes no sistema de filas no instante t, dado o número no instante 0.
- s: número de atendentes no sistema de fila.
- λ_n: taxa média de chegada de novos clientes quando n clientes se encontram no sistema.
- μ_n: taxa média de atendimento para o sistema global quando n clientes se encontram no sistema.

Terminologia e Notação

- Quando λ_n constante para todo n, representa-se por λ.
- Quando taxa média de atendimento por atendente ocupado é constante para todo $n \geq 1$, tem-se μ.
  - $\mu_n = s\mu$ quando $n \geq s$.
- $1/\lambda$: tempo esperado entre atendimentos.
- $1/\mu$: tempo de atendimento esperado.
- $\rho = \lambda/(s\mu)$: fator de utilização (fração de tempo esperada em que cada atendente se encontra ocupado).

Terminologia e Notação

Resultados de estado estável:

- **Condição transiente**: logo após o sistema iniciar operação, o estado do sistema é afetado pelo estado inicial e pelo tempo decorrido.
- **Condição de estado estável**: após tempo suficiente, o sistema se torna independente do estado inicial e do tempo decorrido.
  - $P_n =$ probabilidade de n clientes se encontrarem no sistema de filas.
  - $L =$ número de clientes esperado no sistema de fila $= \sum_{n=0}^{\infty} nP_n$.
  - $L_q =$ comprimento esperado da fila $= \sum_{n=s}^{\infty}(n-s)P_n$.
  - $\mathcal{W} =$ tempo de espera no sistema para cada cliente[1].
  - $W = E(\mathcal{W})$.
  - $\mathcal{W}_q =$ tempo de espera na fila para cada cliente[2].
  - $W_q = E(\mathcal{W}_q)$.

[1] Inclui atendimento.

[2] Exclui atendimento.

Relações entre L, W, L_q e W_q

Suponha $\lambda_n = \lambda$ constante para todo n. Em um processo de fila com estado estável:

- $L = \lambda W$.
- $L_q = \lambda W_q$
- Se λ_n não constante, substituir λ por $w\overline{\lambda}$ (taxa média de chegada a longo prazo).

Suponha que o tempo médio de atendimento, $1/\mu$, constante para todo $n \geq 1$:

- $W = W_q + \dfrac{1}{\mu}$

Distribuição Exponencial

- Suponha T variável aleatória que represente tempos entre chegadas ou tempos de atendimentos.
- Distribuição exponencial com parâmetro α:

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{para } t \geq 0 \\ 0 & \text{para } t < 0 \end{cases}$$

- Probabilidade cumulativas (dado $t \geq 0$):

$$\begin{aligned} P(T \leq t) &= 1 - e^{-\alpha t} \\ P(T > t) &= e^{-\alpha t} \end{aligned}$$

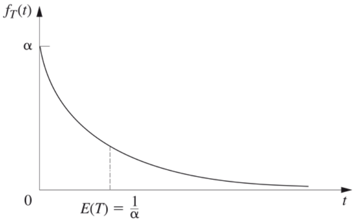
- Valor esperado e variância:

$$\begin{aligned} E(T) &= 1/\alpha \\ var(T) &= 1/\alpha^2 \end{aligned}$$

Distribuição Exponencial

Propriedade 1: $f_T(t)$ é decrescente para $t \geq 0$.

- $P(0 \leq T \leq \Delta t) > P(t \leq T \leq t + \Delta t)$ para quaisquer $t > 0$ e $\Delta t > 0$.
- T assume valor mais próximo de zero.
- Tempos próximos da média.
- Tempos de/entre atendimentos normalmente são breves.
- Tempos prolongados são mais raros.
- *Exemplos*: pronto socorro, caixa de supermercado, caixa eletrônico.



Distribuição Exponencial

Propriedade 2: Ausência de memória.

- $P(T > t + \Delta t \mid T > \Delta t) = P(T > t)$ para quaisquer $t > 0$ e $\Delta t > 0$:

$$\begin{aligned} P(T > t + \Delta t \mid T > \Delta t) &= \frac{P(T > \Delta t, T > t + \Delta t)}{P(T > \Delta t)} \\ &= \frac{P(T > t + \Delta t)}{P(T > \Delta t)} \\ &= \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}} \\ &= e^{-\alpha t} \\ &= P(T > t). \end{aligned}$$

- *Tempos entre atendimentos*: tempo até a próxima chegada não sofre influência de quando ocorreu última chegada.
- *Tempos de atendimentos*: situação na qual operações de atendimento diferem.

Distribuição Exponencial

Propriedade 3: O mínimo de diversas variáveis aleatórias segue distribuição exponencial.

- Sejam $T_1, T_2, \ldots, T_n$ variáveis aleatórias exponenciais independentes com parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_n$.
- Façamos $U = \min(T_1, T_2, \ldots, T_n)$, o mínimo dos valores assumidos por $T_1, T_2, \ldots, T_n$.
- Se T_i representa o tempo até que certo evento ocorra, então U representa o tempo até que o primeiro dos n eventos ocorra. Para $t \geq 0$:

$$
\begin{aligned}
P(U > t) &= P(T_1 > t, T_2 > t, \ldots, T_n > t) \\
&= P(T_1 > t)P(T_2 > t) \ldots P(T_n > t) \\
&= e^{-\alpha_1 t} e^{-\alpha_2 t} \ldots e^{-\alpha_n t} \\
&= \exp\left(-\sum_{i=1}^{n} \alpha_i t\right)
\end{aligned}
$$

- U possui distribuição exponencial com parâmetro $\alpha = \sum_{i=1}^{n} \alpha_i$.

Distribuição Exponencial

Propriedade 3: O mínimo de diversas variáveis aleatórias segue distribuição exponencial.

- Implicação para **tempos entre atendimento**: Suponha n tipos de clientes distintos, com tempos entre atendimento exponenciais T_i de parâmetros α_i para cada tipo de cliente $i \in \{1, \ldots, n\}$. U é o tempo entre atendimento do sistema, como um todo, com parâmetro $\alpha = \sum_{i=1}^{n} \alpha_i$.
- Ignora-se a distinção de clientes.

Distribuição Exponencial

Propriedade 3: O mínimo de diversas variáveis aleatórias segue distribuição exponencial.

- Implicação para **tempos de atendimento**: Suponha n atendentes com tempos de atendimentos T_i exponenciais com parâmetro $\alpha_i = \mu$, $i \in \{1, \ldots, n\}$. U é o tempo de atendimento do sistema, como um todo, com parâmetro $\alpha = n\mu$.
- Mesmo comportamento de uma fila com um atendente com parâmetro $n\mu$.
- Probabilidade do atendente $j \in \{1, \ldots, n\}$ terminar um atendimento primeiro, entre os n atendentes:

$$P(T_j = U) = \frac{\alpha_j}{\sum_{i=1}^{n} \alpha_i}$$

Distribuição Exponencial

Propriedade 4: Relação com a distribuição de Poisson.

- Suponha que o tempo entre um evento consecutivo siga distribuição exponencial com parâmetro α.
- Determinação da probabilidade do número de vezes que esse evento ocorreu ao longo do tempo.
- Seja $X(t)$ o número de ocorrências no tempo $t \geq 0$. Para $n = 0, 1, \ldots$, tem-se que:

$$P(X(t) = n) = \frac{(\alpha t)^n e^{-\alpha t}}{n!}$$

- $X(t)$ segue distribuição de Poisson com parâmetro αt.
- Para $n = 0$, $P(X(t) = 0) = e^{-\alpha t}$, probabilidade da distribuição exponencial para o primeiro evento após o tempo t.

Distribuição Exponencial

Propriedade 4: Relação com a distribuição de Poisson.

- Média: $E(X(t)) = \alpha t$.
- α é a taxa média de ocorrência do evento.
- **Número de términos de atendimento**:
  - Tempos de atendimento com distribuição exponencial com $\alpha = \mu$.
  - $X(t)$: número de atendimentos realizados por um atendente até t.
  - Para modelos com n atendentes, $\alpha = n\mu$.
  - *Processo de Poisson* com taxa α: eventos contados de forma contínua.
- **Número de chegadas**:
  - Tempos entre atendimentos com distribuição exponencial com $\alpha = \lambda$.
  - $X(t)$: número de chegadas no tempo decorrido t.
  - *Processo de Poisson* com parâmetro λ (taxa de chegada).
  - *Entrada de Poisson*.
  - Todo período de duração fixa tem a mesma chance de ter uma chegada independentemente de quando ocorreu a última chegada.

Distribuição Exponencial

Propriedade 5: Para $t > 0$, $P(T \leq t + \Delta t | T > \Delta t) \approx \alpha \Delta t$, para Δt pequeno.

- É sabido que $e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$.
- Para um valor pequeno $\Delta t > 0$:

$$
\begin{aligned}
P(T \leq t + \Delta t | T > t) &= P(T \leq \Delta t) \\
&= 1 - e^{-\alpha \Delta t} \\
&= 1 - 1 + \alpha \Delta t - \sum_{n=2}^{\infty} \frac{(-\alpha \Delta t)^n}{n!}
\end{aligned}
$$

- Com $\alpha \Delta t \approx 0$, os termos do somatório se tornam relativamente desprezíveis.

Distribuição Exponencial

Propriedade 6: Não se afeta por agregação ou desagregação

- Agregação:
  - Suponha n tipos distintos de clientes que chegam em processos independentes de Poisson com parâmetro λ_i, $i \in \{1, 2, \ldots, n\}$.
  - Processo de entrada agregado é um processo de Poisson com parâmetro (taxa média de chegada) $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.
- Desagregação:
  - Seja um processo de Poisson agregado com parâmetro λ e n tipos distintos de clientes, com probabilidade fixa de chegada p_i, para $i \in \{1, 2, \ldots, n\}$ e $\sum_{i=1}^{n} p_i = 1$.
  - A taxa de chegada de cada cliente será $\lambda_i = p_i \lambda$.

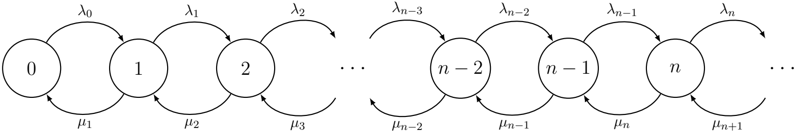
- *Exemplo*: Dois tipos de clientes chegam em uma fila em um processo de Poisson com parâmetro λ. Cada cliente tem probabilidade fixa p de recusar a entrar na fila. E $1 - p$ de entrar na fila. Cada tipo de cliente chega em um processo de Poisson com parâmetros $p\lambda$ e $(1 - p)\lambda$. O desempenho da fila pode ser analisado através do parâmetro $(1 - p)\lambda$.

Processo de Nascimento e Morte

- Nascimento (novo cliente) e Morte (cliente atendido).
- Estado é o número de clientes $N(t)$ no tempo $t \geq 0$.
- Descrição probabilística de $N(t)$.
- Nascimentos e mortes individuais ocorrem aleatoriamente, cujas taxas dependem apenas do estado atual do sistema.
- Hipóteses:
  1. Dado $N(t) = n$, $(n = 0, 1, 2, \ldots)$, segue a distribuição exponencial com parâmetro:
     - λ, a variável tempo remanescente até a próxima chegada.
     - μ, a variável tempo remanescente até a próxima saída.
  2. As variáveis aleatórias para o tempo até a próxima chegada e para o tempo até a próxima saída são independentes. A transição ocorre por unidades $(n \to n + 1$ ou $n \to n - 1)$.

Análise do Processo de Nascimento e Morte

- Cadeia de Markov em tempo contínuo:



- λ_n: taxa média de chegada estando o sistema no estado n.
- μ_n: taxa média de saída estando o sistema no estado n.
- P_n: proporção de tempo que sistema fica no estado n.
- Condição de estado estável: taxa que entra igual a taxa que sai.
- Equação de equilíbrio par o estado n.
  - Para o estado 0: $\mu_1 P_1 = \lambda_0 P_0$.
  - Para o estado 1: $\mu_1 P_1 + \lambda_1 P_1 = \lambda_0 P_0 + \mu_2 P_2$.
  - Para o estado 2: $\mu_2 P_2 + \lambda_2 P_2 = \lambda_1 P_1 + \mu_3 P_3$.
  - $\ldots$
  - Ao final $\sum_{i=0}^{\infty} P_i = 1$.

Análise do Processo de Nascimento e Morte

Estado:

$$
\begin{aligned}
0: &\quad P_1 = \tfrac{\lambda_0}{\mu_1} P_0 \\
1: &\quad P_2 = \tfrac{\lambda_1}{\mu_2} P_1 + \tfrac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) \quad = \tfrac{\lambda_1}{\mu_2} P_1 \quad = \tfrac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0 \\
2: &\quad P_3 = \tfrac{\lambda_2}{\mu_3} P_2 + \tfrac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) \quad = \tfrac{\lambda_2}{\mu_3} P_2 \quad = \tfrac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0 \\
&\vdots
\end{aligned}
$$

De maneira geral, para $n \in \{0, 1, 2, \ldots\}$, tem-se que:

- $C_n = \dfrac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n \mu_{n-1}\ldots\mu_1}$ para $n > 0$ e $C_0 = 1$
- $P_n = C_n P_0$

Análise do Processo de Nascimento e Morte

- $\sum_{n=0}^{\infty} P_n = 1$ implica $\left(\sum_{n=0}^{\infty} C_n\right) P_0 = 1$ e $P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1}$.
- Para filas que se baseiam no processo de nascimento e morte:
  - $L = \sum_{n=0}^{\infty} nP_n$
  - $L_q = \sum_{n=s}^{\infty} (n-s)P_n$
  - $W = L/\overline{\lambda}$
  - $W_q = L_q/\overline{\lambda}$
  - $\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$ (taxa média de chegada a longo prazo)

Análise do Processo de Nascimento e Morte

- Somatórios possuem solução analítica em casos especiais:
  - $\sum_{n=0}^{N} x^n = (1 - x^{N+1})/(1 - x)$.
  - $\sum_{n=0}^{\infty} x^n = 1/(1 - x)$, para $|x| < 1$.
- Resultados válidos sob hipótese de que parâmetros λ_n e μ_n alcancem condição de estado estável.
- Hipótese sempre válida quando:
  - $\lambda_n = 0$ para $n > 0$ (número finito de estados).
  - λ e μ são definidos de forma que $\rho = \lambda/(s\mu) < 1$.
- Hipótese não válida quando $\sum_{n=0}^{\infty} C_n = \infty$.

Modelo M/M/s

- Tempos entre atendimentos e Tempos de atendimento são distribuídos de forma independente e idêntica de acordo com a distribuição exponencial.
- Número de atendentes é o inteiro positivo s.
- Caso especial do processo de nascimento e morte.
- Taxa **média** de chegada (λ) e de atendimento por atendente ocupado (μ) são constantes.

Modelo M/M/s

- Quando $s = 1$ (único atendente):
  - $\lambda_n = \lambda, n = 0, 1, 2, \ldots$
  - $\mu_n = \mu, n = 1, 2, 3, \ldots$
- Quando $s > 1$ (vários atendentes):
  - $\mu_n = \begin{cases} n\mu & \text{quando } n \leq s. \\ s\mu & \text{quando } n > s. \end{cases}$
  - Quando $s\mu$ excede λ, tem-se

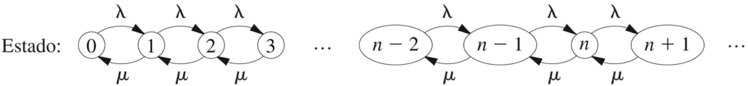
$$\rho = \frac{\lambda}{s\mu} < 1$$

.
  - Nesta condição atinge-se estado estável.

Modelo M/M/s

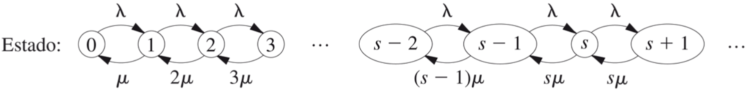
(*a*) *Um único atendente* ($s = 1$)

$$\lambda_n = \lambda, \quad \text{para } n = 0, 1, 2, \ldots$$
$$\mu_n = \mu, \quad \text{para } n = 1, 2, \ldots$$



(*b*) *Diversos atendentes* ($s > 1$)

$$\lambda_n = \lambda, \quad \text{para } n = 0, 1, 2, \ldots$$
$$\mu_n = \begin{cases} n\mu, & \text{para } n = 1, 2, \ldots, s \\ s\mu, & \text{para } n = s, s+1, \ldots \end{cases}$$



Diagramas de taxas para o modelo *M*/*M*/*s*.

Resultados para M/M/1

- $P_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1} = 1-\rho$
- $C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n$ para $n = 0, 1, 2, \ldots$
- $P_n = \rho^n P_0 = (1-\rho)\rho^n$ para $n = 0, 1, 2, \ldots$
- $L = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \ldots = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$
- $L_q = \sum_{n=1}^{\infty} (n-1)P_n = L - 1(1-P_0) = \frac{\lambda^2}{\mu(\mu-\lambda)}$

Resultados para M/M/1

- Quando $\lambda \geq \mu$ fila cresce sem limites a longo prazo.
- Supondo $\lambda < \mu$, deriva-se distribuição de probabilidades do tempo de espera total no sistema, $\mathcal{W}$.
- Atendimentos em ordem de chegada.
- Se na chegada encontra-se n clientes, deve-se contabilizar $n+1$ tempos de atendimento $T_1, T_2, \ldots, T_n, T_{n+1}$ que são variáveis aleatórias que seguem distribuição exponencial, com parâmetro μ.
- Seja S_{n+1} o tempo de espera condicional dado n clientes no sistema:

$$S_{n+1} = T_1 + T_2 + \ldots + T_n + T_{n+1}, \quad \text{para } n = 0, 1, 2, \ldots$$

- S_{n+1} segue distribuição de Erlang (gama).

Resultados para M/M/1

- $P(\mathcal{W} > t) = \sum_{n=0}^{\infty} P_n P(S_{n+1} > t) = e^{-\mu(1-\rho)t}$ para $t \geq 0$.
- $\mathcal{W}$ segue distribuição exponencial com parâmetro $\mu(1-\rho)$.
- $W = E(\mathcal{W}) = \dfrac{1}{\mu(1-\rho)} = \dfrac{1}{(\mu - \lambda)}$.
- $P(\mathcal{W}_q = 0) = P_0 = 1 - \rho$, com nenhum cliente no sistema.
- $P(\mathcal{W}_q > t) = \sum_{n=1}^{\infty} P_n P(S_n > t) = \ldots = \rho e^{-\mu(1-\rho)t}$ para $t \geq 0$.
- $W_q = E(\mathcal{W}_q) = \dfrac{\lambda}{\mu(\mu - \lambda)}$.

Resultados para M/M/s com $s > 1$

- $P_0 = 1/\left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\lambda/(s\mu)}\right)$
- $C_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} & \text{para } n = 1, 2, \ldots, s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} & \text{para } n = s, s+1, \ldots \end{cases}$
- $P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & \text{para } 0 \leq n \leq s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0 & \text{para } n \geq s \end{cases}$
- $L_q = \sum_{n=s}^{\infty} (n-s) P_n = \ldots = \frac{P_0 (\lambda/\mu)^s \rho}{s!(1-\rho)^2}$
- $W_q = \frac{L_q}{\lambda}$
- $W = W_q + \frac{1}{\mu}$
- $L = \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}$

Resultados para M/M/s com $s > 1$

- Quando $\lambda \geq s\mu$ fila cresce indefinitivamente a longo prazo.
- Resultados não se aplicam.
- Supondo $\lambda < s\mu$, para $t \geq 0$:
  - $P(\mathcal{W} > t) = e^{-\mu t}\left[1 + \dfrac{P_0(\lambda/\mu)^s}{s!(1-\rho)}\left(\dfrac{1-e^{-\mu(s-1-\lambda/\mu)}}{s-1-\lambda/\mu}\right)\right].$
  - $P(\mathcal{W}_q = 0) = \sum_{n=0}^{s-1} P_n.$
  - $P(\mathcal{W}_q > t) = (1 - P(\mathcal{W}_q = 0))e^{-s\mu(1-\rho)t}.$

Exemplo do Hospital

O administrador do hospital verificou que a chegada de atendimentos e tempo o gasto para atendê-los seguem distribuições exponenciais. Assim, ele optou por uma modelagem $M/M/s$.

Inicialmente, ele estimou que os pacientes chegam em um taxa média de 1 a cada $1/2$ hora e que um médico precisa em média de 20 minutos para atender cada paciente. Portanto, utilizando uma hora como unidade de tempo, tem-se:

- $1/\lambda = 1/2$ hora por cliente ($\lambda = 2$ clientes por hora).
- $1/\mu = 1/3$ hora por cliente ($\mu = 3$ clientes por hora).

Exemplo do Hospital

São considerados ter um ($s = 1$) ou dois médicos ($s = 2$). Assim:

	$s = 1$	$s = 2$
ρ	$2/3$	$1/3$
P_0	$1/3$	$1/2$
P_1	$2/9$	$1/3$
P_n, $n \geq 2$	$\frac{1}{3}\left(\frac{2}{3}\right)^n$	$\left(\frac{1}{3}\right)^n$
L_q	$4/3$	$1/12$
L	2	$3/4$
W_q	$2/3$ hora	$1/24$ hora
W	1 hora	$3/8$ hora
$P(\mathcal{W}_q > 0)$	$0{,}667$	$0{,}167$
$P(\mathcal{W}_q > 1/2)$	$0{,}404$	$0{,}022$
$P(\mathcal{W}_q > 1)$	$0{,}245$	$0{,}003$
$P(\mathcal{W}_q > t)$	$(2/3)e^{-t}$	$(1/6)e^{-4t}$
$P(\mathcal{W} > t)$	e^{-t}	$(1/2)e^{-3t}(3 - e^{-t})$

Um único médico seria inadequado!