Introdução à Teoria das Filas

“If the facts don't fit the theory, change the facts.”

--Albert Einstein

Notação

- **Processo de Chegada:** Se os usuários chegam nos instantes $t_1, t_2, \ldots, t_j$, então as variáveis aleatórias $\tau_j = t_j - t_{j-1}$ são chamadas de **intervalos entre chegadas**.
  - Assume-se, em geral, que a seqüência dos τ_j são v.a's IID.
  - Exemplo: chegadas de Poisson (intervalos entre chegadas são IID e exponencialmente distribuídos).
  - Outras distribuições: Erlang, Hiperexponencial, etc.

Notação

- **Distribuição do tempo de serviço:**
  - **Tempo de serviço:** tempo que o usuário gasta no servidor.
  - É comum assumir que os tempos de serviço sejam v.a's IID.
  - Distribuições utilizadas: **exponencial**, Erlang, hiperexponencial e geral.
- **Número de servidores:** para uma mesma fila.

Notação

- **Capacidade do sistema:** número máximo de usuários que podem ser acomodados no sistema (fila + servidor(es)).
  - Se a capacidade do sistema for grande, é mais fácil analisá-lo com a hipótese de que a fila seja infinita.
- **Tamanho da população:** número potencial de usuários.
  - Se o tamanho for grande, é mais fácil analisar o sistema com a hipótese de que a população seja infinita.

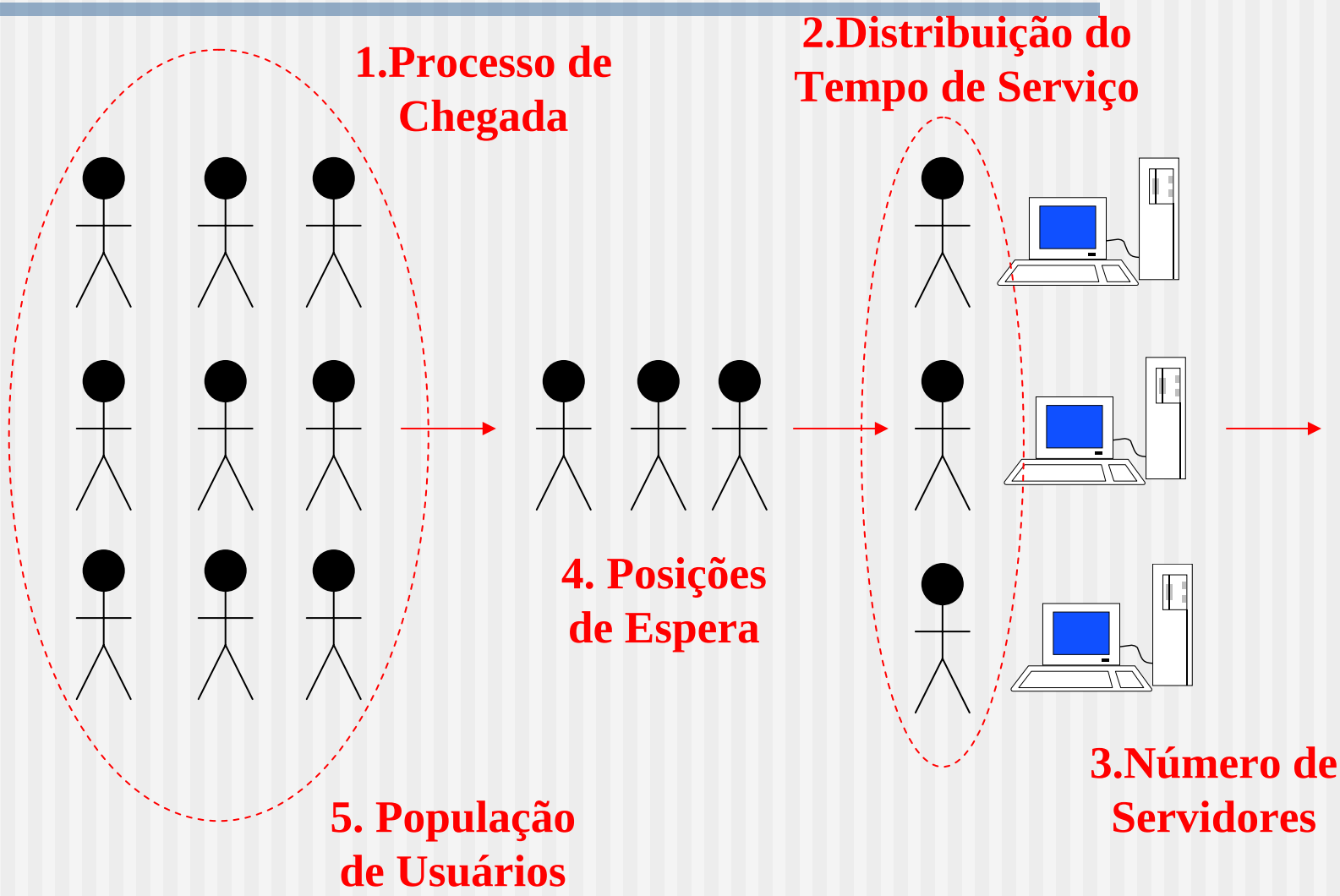
Notação

- **Disciplina de atendimento:** ordem na qual os usuários são atendidos. Exemplos:
  - FCFS: *First Come, First Served*
  - LCFS: *Last Come, First Served*
  - LCFS-PR: *Last Come, First Served with Preempt and Resume*
  - RR: *Round-Robin*
  - PS: *Processor Sharing*
  - IS: *Infinite Server*

Notação

- **Mais Disciplinas de Atendimento:**
  - SPT: Shortest Processing Time first
  - SRPT: Shortest Remaining Processing Time first
  - SEPT: Shortest Expected Processing Time first
  - SERPT: Shortest Expected Remaining Processing Time first
  - BIFS: Biggest In, First Served
  - LVFS: Loudest Voice, First Served

Componentes básicos de uma fila



Notação de Kendall

- **A / S / m / B / K / SD**
  - **A** é a distribuição do intervalo entre chegadas,
  - **S** é a distribuição do tempo de serviço,
  - **m** é o número de servidores,
  - **B** é o número de *buffers* (capacidade do sistema),
  - **K** é o tamanho da população e
  - **SD** é a disciplina de atendimento.

Representação das distribuições

- M Exponencial
- E_k Erlang com parâmetro k
- H_k Hiperexponencial com parâmetro k
- D Determinística
- G Geral
- $M^{[x]}$ Exponencial com chegada em bloco *(bulk)* de tamanho x

Exemplo: **M/M/3/20/1500/FCFS**

- O intervalo entre chegadas sucessivas é distribuído exponencialmente.
- Os tempos de serviço são exponencialmente distribuídos.
- Há três servidores.
- A fila possui buffers para 20 usuários. Isto é, 3 usuários em atendimento e 17 esperando por serviço. Enquanto o número de usuários estiver em seu valor máximo, 20, todos os usuários que chegarem serão perdidos até que o comprimento da fila diminua.
- Há um total de 1500 usuários que podem ser atendidos.
- A disciplina de atendimento é *first come, first served*.

Outros Exemplos

- $G/G/1/\infty/\infty/FCFS \Leftrightarrow G/G/1$
- $M^{[x]}/M/1$
- $M/G^{[x]}/m$

Regras válidas para todas as filas

- τ = intervalo entre duas chegadas sucessivas.
- λ = Taxa média de chegadas = $1/E[\tau]$. Em alguns sistemas, esta taxa pode ser uma função do estado do sistema. Por exemplo, ela pode depender do número de usuários que já se encontram no sistema.
- s = tempo de serviço por usuário.
- μ = taxa média de serviço por servidor = $1/E[s]$. A taxa total de serviço para m servidores é $m\mu$.

Regras válidas para todas as filas

- n = número de usuários no sistema. Também chamado de **comprimento da fila**. Note que inclui os usuários que já estão sendo atendidos, assim como os que estão esperando na fila.
- n_q = número de usuários esperando para serem atendidos. Este número é sempre menor do que n, dado que não inclui os usuários que estão sendo atendidos.
- n_s = número de usuários em atendimento.

Regras válidas para todas as filas

- r = tempo de resposta ou tempo no sistema. Inclui tanto o tempo de espera como o tempo em atendimento.
- w = tempo de espera, isto é, intervalo de tempo entre o instante de chegada e o instante em que iniciou a ser atendido.
- Todas estas variáveis, à exceção de λ e μ, são variáveis aleatórias.

Relacionamento entre as variáveis

- **Condição de estabilidade:** A taxa média de chegadas deve ser menor do que a taxa média de atendimento:
  $\lambda < m\mu$
  Válida apenas para filas infinitas. Com filas finitas o sistema nunca é instável.

Relacionamento entre as variáveis

- **Número no Sistema versus Número na Fila:**

  $n = n_q + n_s$

  $E[n] = E[n_q] + E[n_s]$

  Se a taxa de atendimento em cada servidor for independente do número de usuários na fila, temos:

  $\text{Cov}(n_q, n_s) = 0$ e

  $\text{Var}[n] = \text{Var}[n_q] + \text{Var}[n_s]$

Relacionamento entre as variáveis

- **Tempo no Sistema versus Tempo na Fila:**

  $r = w + s$

  $E[r] = E[w] + E[s]$

  Se a taxa de atendimento em cada servidor for independente do número de usuários na fila, temos:

  $Cov(w, s) = 0$

  e

  $Var[r] = Var[w] + Var[s]$

Relacionamento entre as variáveis

- **Número versus Tempo:** Se os usuários não forem perdidos por causa de buffers insuficientes, temos que:

  Número médio de usuários no sistema = taxa de chegada x tempo médio de resposta

  Número médio de usuários na fila = taxa de chegada x tempo médio de espera

  Lei de Little.

Lei de Little

- Número médio de usuários no sistema = taxa de chegada x tempo médio de resposta
- Provada pela primeira vez por LITTLE (1961), aplica-se a qualquer sistema ou parte do mesmo no qual o número de usuários que entram no sistema é igual ao número de usuários que terminam o atendimento.

Lei de Little

- **Prova:**
  Taxa de chegadas = total de chegadas/tempo total = N/T
  Tempo médio gasto no sistema = J/N
  onde J é o tempo total gasto no sistema por todos os usuários.

  Número médio no sistema $= \frac{J}{T}$

  $= \frac{N}{T} \times \frac{J}{N}$

  $=$ taxa de chegada $\times$

  tempo médio gasto no sistema

Processos Estocásticos

- **Processos estocásticos**: funções aleatórias dependentes do tempo. São úteis para representar o estado de sistemas de filas.

  Exemplos:

  - *n(t)* número de jobs na CPU de um sistema computacional.

Tipos comuns de processos estocásticos

- *Processos de Estado Discreto e de Estado Contínuo:* dependendo dos valores que as suas variáveis de estados podem assumir. Um processo estocástico de estado discreto é também chamado de **cadeia estocástica**.
- *Processos de Markov*: se os estados futuros do sistema forem independentes do passado e dependerem exclusivamente do estado atual. Um processo de Markov de estado discreto é também chamado de **cadeia de Markov**.

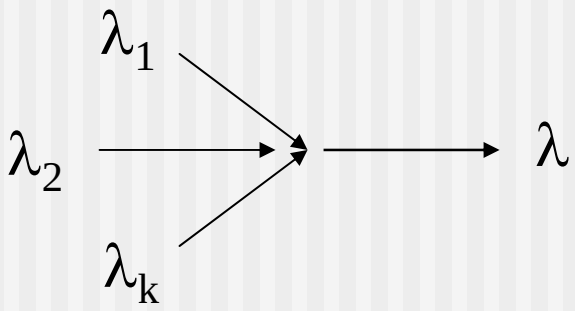
Tipos comuns de processos estocásticos

- *Processos de Nascimento e Morte:* processos de Markov de espaço discreto no qual as transições entre estados estão restritas a estados vizinhos.
- *Processos de Poisson*: para intervalos entre chegadas independentes e IID com distribuição exponencial.

Propriedades dos Fluxos de Poisson

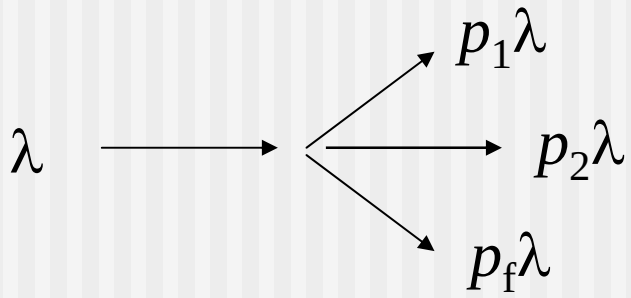
- **Superposição** de fluxos de Poisson:

$$\lambda = \sum_{i=1}^{k} \lambda_i$$



Propriedades dos fluxos de Poisson

- **Divisão**: se um fluxo de Poisson for dividido em *f* subfluxos com probabilidade p_i de um usuário seguir o subfluxo *i*, então, cada subfluxo é também um fluxo de Poisson com taxa média $p_i\lambda$.



Propriedades dos fluxos de Poisson

- Se as chegadas a uma fila com *m* servidores e tempos de serviço exponenciais forem Poisson com taxa média λ, então as partidas também são Poisson, com mesma média λ (desde que $\lambda < \Sigma\ \mu_i$).