Avaliação de desempenho de Sistemas de Informação
Teoria das Filas

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

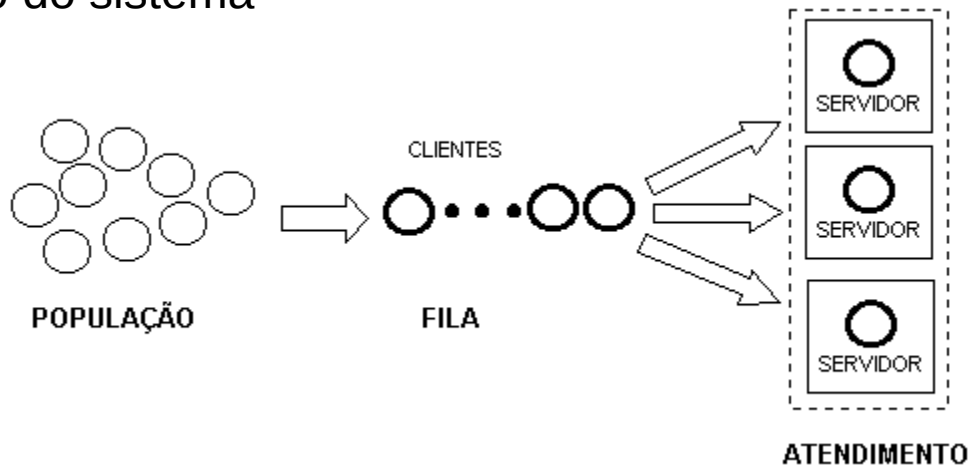
Prof. Sergio Nascimento

sergio.onascimento@sp.senac.br

Filas

Todos já passaram pelo aborrecimento de ter que esperar em uma fila: Fila de ônibus, banco, restaurante, trânsito, etc.

Em sistemas computacionais há filas em vários lugares:
Filas de impressão, pacotes em roteadores, processos aguardando processamento da CPU, etc.
As filas surgem porque a demanda de serviço é maior que a capacidade de atendimento do sistema



Teoria das Filas

- Ramo da probabilidade que estuda o fenômeno da formação de filas em solicitações de serviços.

- Permite criar modelos do sistema estudado para prever o seu comportamento.

  - Pode-se dimensionar um determinado sistema segundo a demanda de seus clientes, evitando desperdícios ou gargalos.

- Utilizada para modelar sistemas onde:
  - Clientes chegam para ser atendidos.
  - Esperam a sua vez de ser atendidos.
  - São atendidos e vão embora.

Sistemas de Fluxo

- Sistema no qual alguma “comodity” flui, se move ou é transferida por meio de um ou mais canais com capacidade limitada, de forma a ir de um ponto a outro no sistema.

- Sistemas transferem as comodities à uma taxa finita.

- Classificação:

  - Estáveis (Determinísticos)
  - Instáveis (Estocástico)

- A **Fila** é gerada pelos clientes à espera de atendimento (Não inclui o(s) cliente(s) em atendimento).

- **Serviço** ou atendimento é constituído por um ou mais postos de atendimento.

Fila + Serviço = Sistema

Sistemas de Fluxo – Estáveis (Determinísticos)

Sistemas onde o fluxo passa de forma previsível.

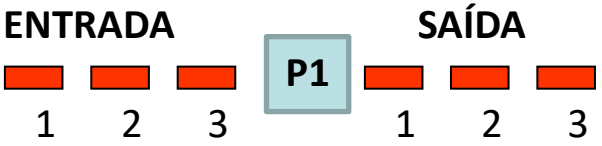
- Quantidade de fluxo pelo canal é constante e conhecida durante o período de observação.

- **Taxa de Serviço** (C) e **Taxa de Chegadas** (R)
  - Se C > R não existe sobrecarga.
    - Atendido por Sistemas Estáveis de Canal Único (*single channel*)
  - Se C < R existe congestionamento (Fila).
    - Atendido por Sistemas Estáveis com Redes de Canais (*multiple channel*).

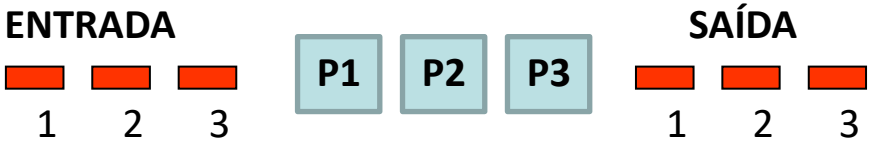
- Número de clientes no sistema (Em cada instante) = Estado do sistema.

Modelos de Filas

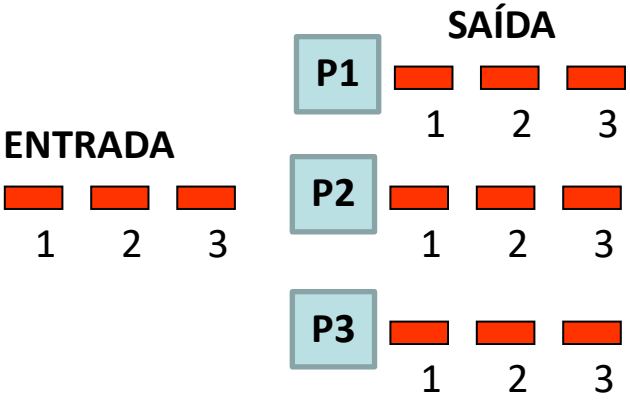
SINGLE-CHANNEL/SINGLE PHASE



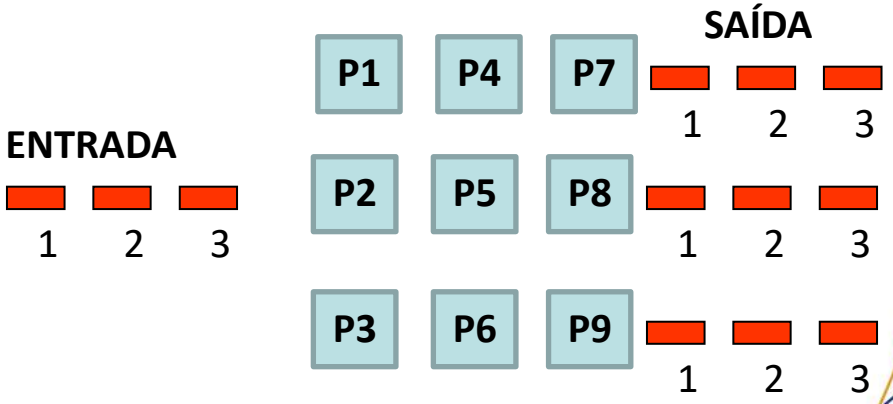
SINGLE-CHANNEL/MULTI PHASE



MULTI-CHANNEL/SINGLE PHASE



MULTI-CHANNEL/MULTI PHASE



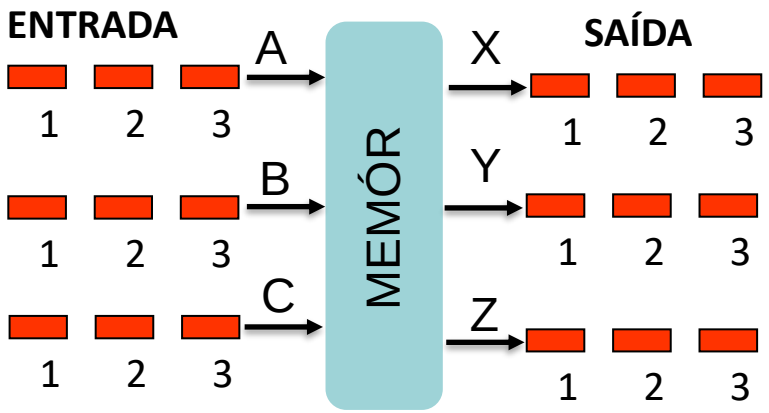
Sistemas de Fluxo – Instáveis (Estocástico)

Sistemas onde o fluxo passa de forma aleatória (chegada é aleatória).

- Demanda de uso do canal é imprevisível.
  - Exemplo: internet, telefonia IP, etc.

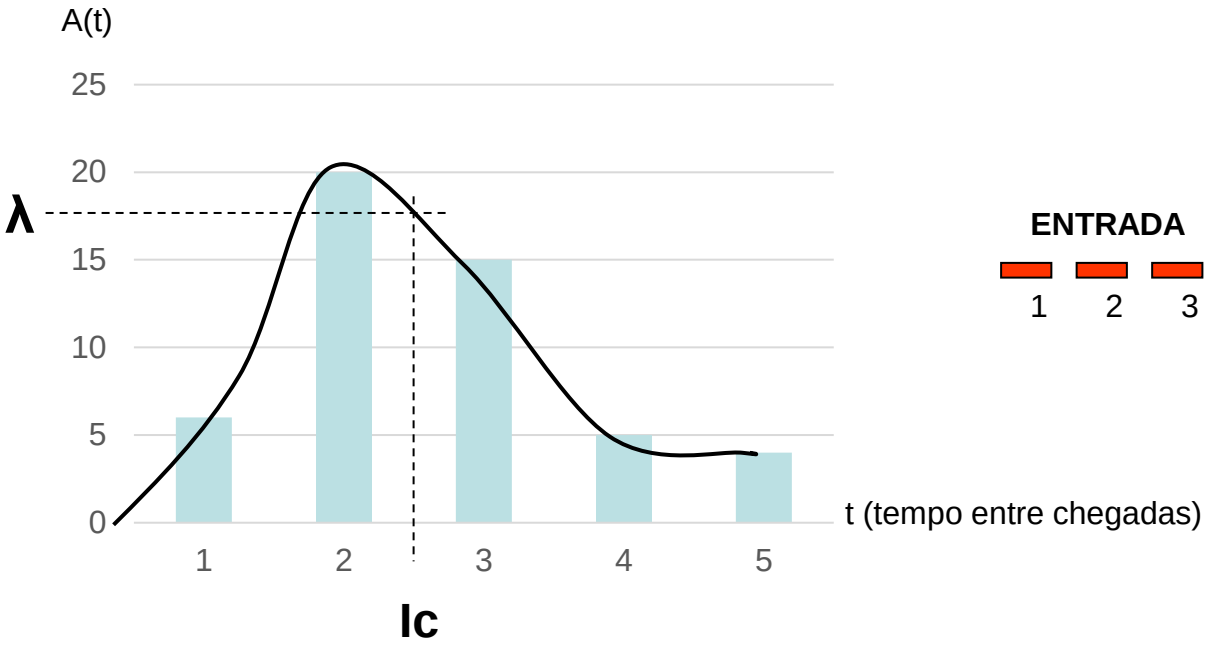
- Como os sistemas estáveis também pode apresentar canal único ou rede de canais.

- A Teoria da Filas se propõe a resolver problemas de sistemas de fluxo aleatório (estocásticos).



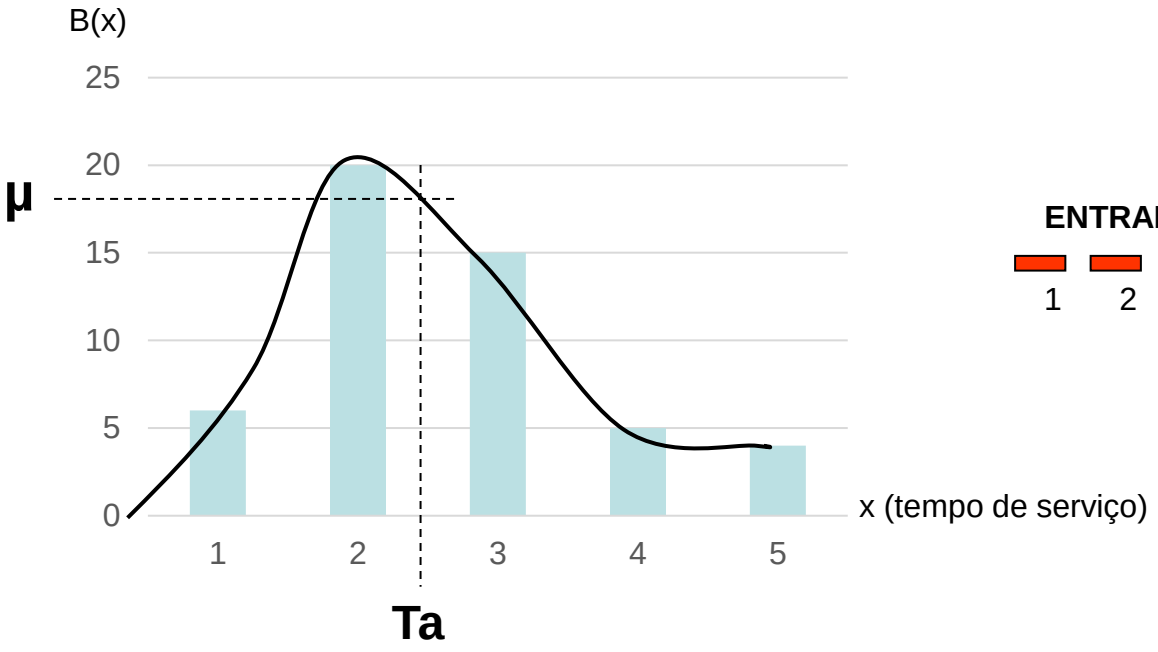
Processos Estocásticos

- Taxa de Chegada [A(t)]
  - A(t) = P [tempo entre chegadas ≤ t] Probabilidade que o tempo entre chegadas seja menor ou igual a t, onde 0<A(t)<1.
  - Ritmo médio de chegada: **λ**
  - Intervalo médio entre chegadas: Ic



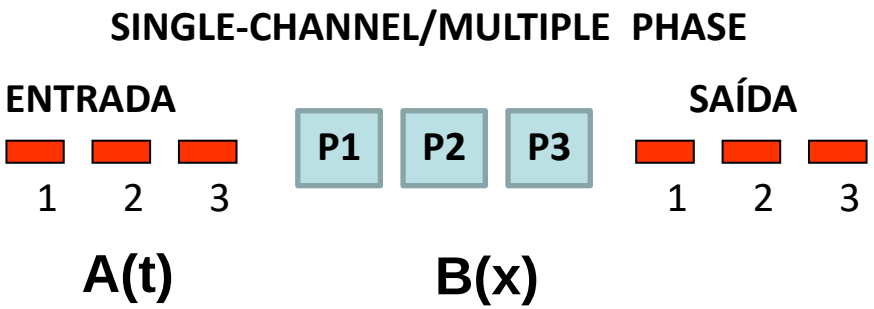
Processos Estocásticos

- Taxa de Serviço [B(x)]
  - B(x) = P [tempo de serviço ≤ x] Probabilidade que o tempo de serviço seja menor ou igual a x.
  - Ritmo médio de atendimento: **μ**
  - Tempo médio de atendimento: Ta



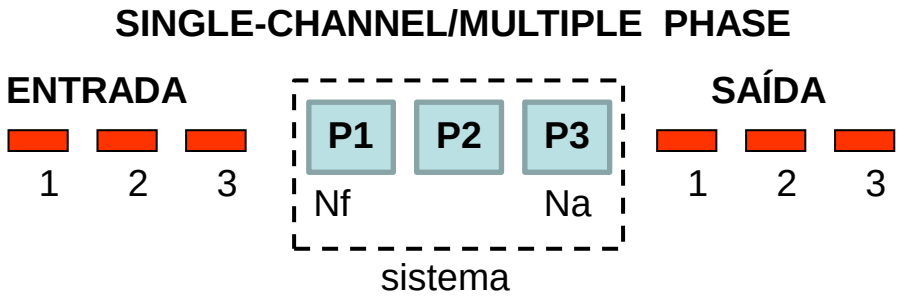
Sistema de Filas - Caracterização

- Linha de entrada – chegam clientes (solicitações, pacotes, jobs, etc.) com tempo de chegada aleatório [**A (t)**].
- Clientes se encaminham para o sistema de filas, onde são atendidos (servidos) por um tempo aleatório [**B (x)**].
- Depois de atendidos são liberados em uma linha de saída Sistemas podem ter m servidores (P1, P2, P3).



Variáveis

- Variáveis referentes ao sistema
  - Ts = Tempo médio de permanência no sistema.
  - Ns = Número médio de clientes no sistema.
- Variáveis referentes ao processo de chegada
  - **λ =Rítmo médio de chegada**
  - Ic = Intervalo médio entre chegadas = 1/ **λ**
  - Nc = Número médio de chegadas no sistema.
- Variáveis referentes à fila
  - Tf = Tempo médio de permanência na fila.
  - Nf = Número médio de clientes na fila.
- Variáveis referentes ao processo de atendimento
  - **μ = Rítmo médio de atendimento**
  - Ta =Tempo médio de atendimento ou serviço = 1/ **μ**
  - Na =Número médio de clientes sendo atendidos
  - c = Capacidade de atendimento



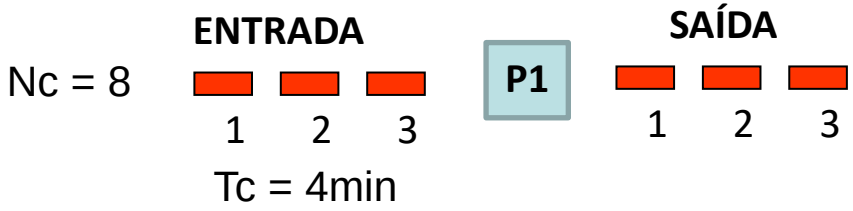
Variáveis de chegada

Ritmo médio de chegada $(\boldsymbol{\lambda_t}) = \frac{Nc}{Ic}$ onde

$(\boldsymbol{\lambda_t})$ Ritmo médio de chegada de clientes por unidade de tempo

- Ic = Intervalo médio entre chegadas

Ex: Um sistema recebe 8 chegadas de clientes em um período de 4 minutos. Qual o Ritmo médio de chegada?



$\boldsymbol{\lambda_t} = \frac{Nc}{Ic} = \frac{8}{4} =$ 2 chegadas/min

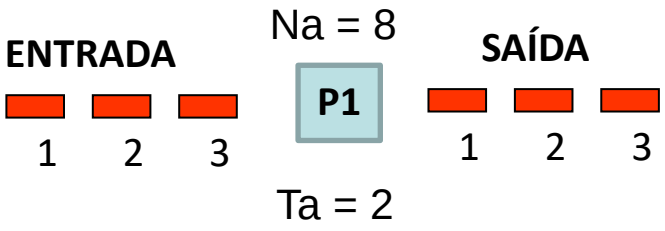
Variáveis de serviço

Ritmo médio de Serviço (μ_t) = $\frac{Na}{Ta}$ onde

(μ_t) Ritmo médio de serviço por cliente no servidor.

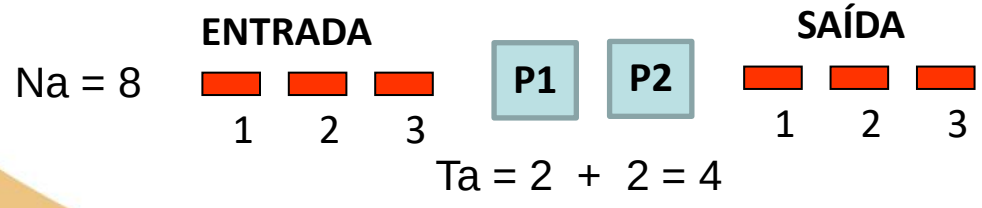
- Ta = Tempo médio de atendimento ou serviço.
- Para c servidores (em uma mesma fila), a taxa total é c μ

Ex: Cada sistema demora 2 segundos para processar 8 pacotes IP. Qual o Ritmo médio de serviço?



$\mu = \frac{Na}{Ta} = \frac{8}{2} = 4$ pacotes/s

Ex: Para dois sistemas a taxa seria de:



$\mu = \frac{Na}{cTa} = \frac{8}{4} = 2$ pacotes/s

Variáveis – Relações Básicas

Número de usuário no sistema: Ns = Nf + Na

Tempo de permanência no sistema: Ts = Tf + Ta

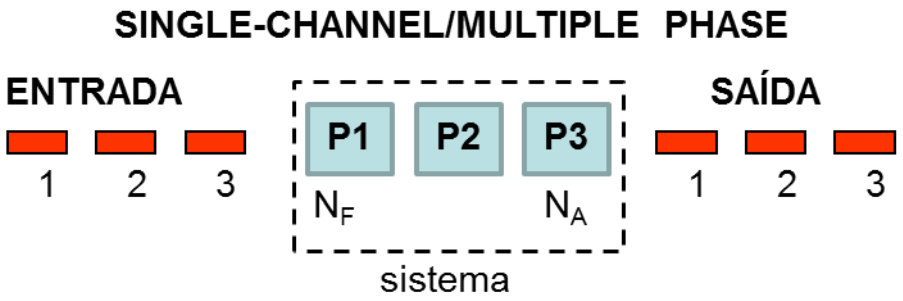
Pode-se também demonstrar que:

Na = λ / μ = Ta /Ic

Portanto:

Ns = Nf + Na

Ns = Nf + (λ / μ) = Nf + (Ta/Ic)



Lei de Little

E um dos resultados mais importantes e amplamente usados em desempenho de qualquer sistema. O numero medio de clientes (N) em um sistema estavel (em algum intervalo (T), é igual a taxa media de chegada λ, multiplicada pelo tempo medio de servico no sistema T, ou seja, J.D.C Little demonstrou que, para um sistema estável* de filas, temos:

$Nf = \lambda \times Tf$

- Número médio de clientes na fila é igual à taxa de chegadas multiplicada pelo tempo médio de permanência na fila.

$Ns = \lambda \times Ts$

- Número médio de clientes no sistema é igual à taxa de chegadas multiplicada pelo tempo de permanência no sistema.

❑ **Taxa de Utilização: U(t)** $= t/T = \lambda / \mu$

Onde t = tempo de serviço e T = tempo de observação

- Fração de tempo no qual o servidor está ocupado
  - U(t) = 0 – sistema vazio (ocioso)
  - U(t) > 0 – sistema ocupado
  - U(t) = 1 – sistema completamente ocupado (saturado)

* Sistema estável onde num intervalo grande de observação, o número de saídas é igual ao número de chegadas do sistema.

Medidas e Relacionamentos entre as variáveis

Condição de Estabilidade

$\lambda = c.\mu$ ($c \geq 1$) válido para filas infinitas

Intensidade de Tráfego (ρ) – mede o congestionamento de sistemas de filas

$\rho = \frac{\text{taxa de chegadas}}{\text{taxa de serviço}} = |\lambda / \mu| = |Ta / Tc|$ ou $\lambda / c \,.\, \mu$; onde $\rho \leq 1$

Vazão (Throughput) – média das solicitações processadas por unidade de tempo (taxa de saída do sistema).

$X = N/T$

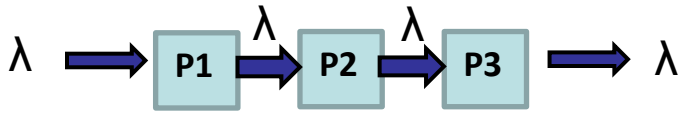
Equilíbrio: Taxa de saída é igual à taxa de chegada do sistema

$$\lambda = c \,.\, \mu \,.\, \rho$$

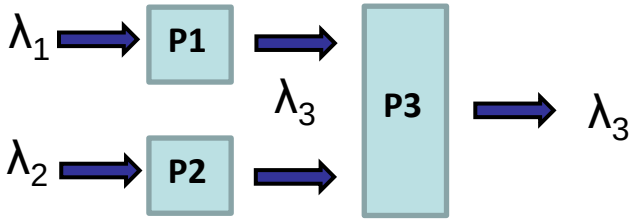
(Throughput para sistema equilibrado)

Postulados Básicos

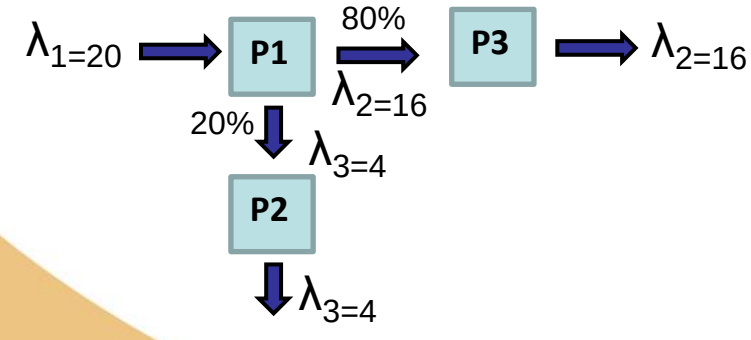
1) Em qualquer sistema estável, o fluxo que entra é igual ao fluxo que sai.



2) Em qualquer sistema estável, o fluxo de entrada se mantem nas diversas seções do sistema, desde que não exista junção ou desdobramento.



3) Em qualquer sistema estável, a junção de fluxos equivale às suas somas aritméticas, ou seja $\lambda_3 = \lambda_1 + \lambda_2$.



4) Em qualquer sistema estável, o desdobramento percentual de um fluxo é igual ao desdobramento aritmético do mesmo fluxo. Assim, se após a estação P1 temos 80% do fluxo se deslocando para P3, então o ritmo de chegada em P3 é de 0,8 x 20 = 16 clientes/minuto

Notação de Kendall

A / S / NS / B / K / SD

- **A / S** = distribuição de tempo de chegada e tempo de serviço]
  - **M** (Exponencial, Poisson), **Ek** (Erlang), **Hk** (Hiperexponencial), **D** (Determinístico), **G** (Geral – todas as distribuições)

- **NS** = Número de servidores
- **B** = Número de buffers (lugares na fila)
- **K** = Tamanho da população
- **SD** = Disciplina do serviço FIFO, LIFO, etc.

Padrão: **B** = ∞ , **K** = ∞, **SD** = FIFO

- **M / M / 1** = **M / M / 1 / ∞ / ∞ / FIFO** (chegadas Poisson, tempo de serviço exponencial, 1 servidor, buffer infinito, FIFO).

- **M / M / c** (idem anterior, com c servidores)

Exercícios – Teoria das filas

1. Se um sistema apresenta um tempo de serviço de 3 segundos e atende 5 solicitações durante o intervalo de observação, por quanto tempo esse servidor esteve ocupado?

2. Observamos que 2000 solicitações de serviços chegaram a um determinado componente do sistema, durante um intervalo de 20 minutos. Qual a taxa de chegada de solicitações (em solicitações por segundo)?

3. Observamos 50 solicitações sendo atendidas em um intervalo de observação de 4 minutos. Qual foi o throughput observado (em solicitações por segundo)?

4. Durante um período de 3 horas foi observado um servidor. Ele apresenta uma utilização de 65%. Qual foi o tempo que ele esteve ocupado durante a observação?

5. Um servidor de arquivos está conectado à rede da empresa. Ele recebe 150 solicitações de arquivos e atende 45 dessas solicitações durante um intervalo de observação de 30 minutos. O servidor gasta 5 segundos para atender cada solicitação. Calcule os valores de λ, X e U e o tempo que o servidor esteve ocupado.

Exercícios – Teoria das filas

1.Se um sistema apresenta um tempo de serviço de 3 segundos e atende 5 solicitações durante o intervalo de observação, por quanto tempo esse servidor esteve ocupado?
Ta = 3 seg ; Na = 5 → t = 3 x 5 = 15 seg

2.Observamos que 2000 solicitações de serviços chegaram a um determinado componente do sistema, durante um intervalo de 20 minutos. Qual a taxa de chegada de solicitações (em solicitações por segundo)?
Nc = 2000 ; Tc = 20 min => λ = 2000/20/60 = 1,67 solicitações/seg.

3.Observamos 50 solicitações sendo atendidas em um intervalo de observação de 4 minutos. Qual foi o throughput observado (em solicitações por segundo)?
Na = 50 ; T = 4 min => X = Na/T = 50/4/60 = 0,21 solicitações/seg

4.Durante um período de 3 horas foi observado um servidor. Ele apresenta uma utilização de 65%. Qual foi o tempo que ele esteve ocupado durante a observação?
T = 3 horas ; U = 0,65 => U =t/T => t = T x U = 3 x 0,65 = 1,95 horas = 1h 57 min

Exercícios – Teoria das filas

5. Um servidor de arquivos está conectado à rede da empresa. Ele recebe 150 solicitações de arquivos e atende 45 dessas solicitações durante um intervalo de observação de 30 minutos. O servidor gasta 5 segundos para atender cada solicitação. Calcule os valores de λ, μ , U e o tempo que o servidor esteve ocupado.

Nc = 150 solicitações ; T = 30 min

Na = 45 ; Ta = 5 seg

λ = Nc/T = 150/30 = 5 solicitações/min

μ = N/T = 45/30 x 60 = 0,025 solicitações/seg

U = t/T

t = 45 x 5 = 225 seg

T = 30 x 60 = 1800

U = 225/1800 = 0,125 OU 12,5 %.

Exercícios – Teoria das filas

6.Em um determinado sistema observamos que 10 solicitações foram atendidas durante o tempo de observação e que esse servidor esteve ocupado por 200 segundos durante esse mesmo período de observação. Qual o tempo médio de serviço observado desse sistema?

7.Um servidor apresenta uma taxa de chegada de 10 solicitações/segundo. Se observarmos o servidor por 10 minutos, quantas solicitações de serviço chegaram durante o período de observação?

8.O servidor analisado esteve ocupado por 19 minutos durante um período de observação de 30 minutos. Qual é a utilização desse servidor?

9.Um sistema computacional foi observado durante sete dias e verificou-se que, em média, o sistema estava sendo utilizado por 16 horas em cada dia. Qual a utilização do sistema durante esses sete dias?

Exercícios – Teoria das filas

6.Em um determinado sistema observamos que 10 solicitações foram atendidas durante o tempo de observação e que esse servidor esteve ocupado por 200 segundos durante esse mesmo período de observação. Qual o tempo médio de serviço observado desse sistema?

Na = 10 ; Ta = 200 seg => μ = Na/Ta ; μ=10/200 seg => Ta = 1/ μ = 200/10 = 20 seg.

7.Um servidor apresenta uma taxa de chegada de 10 solicitações/segundo. Se observarmos o servidor por 10 minutos, quantas solicitações de serviço chegaram durante o período de observação?

λ = 10 solicitações/seg ; T=10min
λ = Nc/Tc => Nc = λ x Tc = 10x10x60 = 6000 solicitações/seg

8.O servidor analisado esteve ocupado por 19 minutos durante um período de observação de 30 minutos. Qual é a utilização desse servidor?

U = t/T = 19/30 = 0,63 OU 63%

9.Um sistema computacional foi observado durante sete dias e verificou-se que, em média, o sistema estava sendo utilizado por 16 horas em cada dia. Qual a utilização do sistema durante esses sete dias?

U = t/T = 16/24 = 0,67 OU 67%