Algoritmos para Projeto de Redes de Telecomunicações: Introdução a Teoria de Filas

Programa de Pós-Graduação em Informática PPGIA - PUCPR

Prof. Marcelo E. Pellenz
http://www.ppgia.pucpr.br/~marcelo
marcelo@ppgia.pucpr.br

Introdução a Teoria de Filas

- A teoria de filas é uma importante área de aplicação da teoria de probabilidades, que estuda o fenômeno da formação de filas de solicitantes de serviços, fornecidos por um determinado recurso.
- As filas surgem porque a demanda de serviço é maior que a capacidade de atendimento do sistema.
- A teoria de filas é utilizada para análise e dimensionamento de redes de comunicações e sistemas computacionais.
- Permite estimar importantes medidas de desempenho de um sistema a partir de propriedades mensuráveis das filas.
- Permite dimensionar um determinado sistema segundo a demanda dos seus clientes, evitando desperdícios ou gargalos.
- As filas apresentam comportamento *estocástico*.

Introdução a Teoria de Filas

- Exemplos
  - Compartilhamento de tempo em computadores:
    - Clientes: programas (processos)
    - Servidores: CPU, disco, dispositivos de I/O, barramento
  - Multiplexadores estatísticos baseados em pacotes:
    - Clientes: pacotes (células)
    - Servidores: Enlaces (links)
  - Comutadores por Circuito:
    - Clientes: Chamadas
    - Servidores: Canais

Introdução a Teoria de Filas

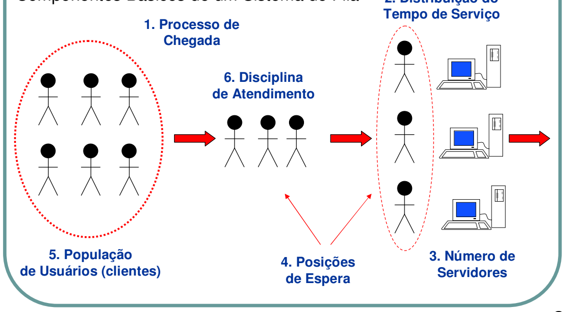
- Exemplos
  - Redes de Múltiplo Acesso (Ethernet LAN / Wireless LAN)
    - Clientes: pacotes (frames)
    - Servidores: Meio Físico (fibra/par trançado/RF)
  - Servidores Web
    - Clientes: requisições
    - Servidor

Introdução a Teoria de Filas

- Sistemas de fila em redes de comunicações:
  - Fila de pacotes aguardando por transmissão.
  - Fila de pacotes aguardando por roteamento/comutação.
  - Fila de pacotes recebidos na placa de rede de um terminal.
  - Fila de chamadas telefônicas aguardando por linha em um PABX.
  - Fila de amostras de voz recebidas em um telefone IP.

Introdução a Teoria de Filas

Componentes Básicos de um Sistema de Fila



Introdução a Teoria de Filas

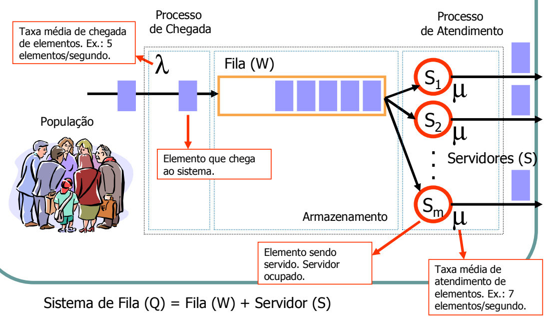
- Um sistema de filas (Q – *Queuing System*) é um sistema composto por:
  - Uma ou mais filas (W – *Waiting Line*) onde são armazenados os elementos que aguardam por atendimento.
  - Um ou mais servidores (S – *Servers*) que atendem os elementos.
  - Um processo de chegada, que define como os elementos chegam ao sistema.
  - Um processo de atendimento, que define como os elementos são atendidos pelo sistema.
  - O tamanho da população que gera os elementos.

7

Introdução a Teoria de Filas

Autor: Prof. Antonio M. Alberti - INATEL

- Sistema de Filas com 1 Fila e vários Servidores

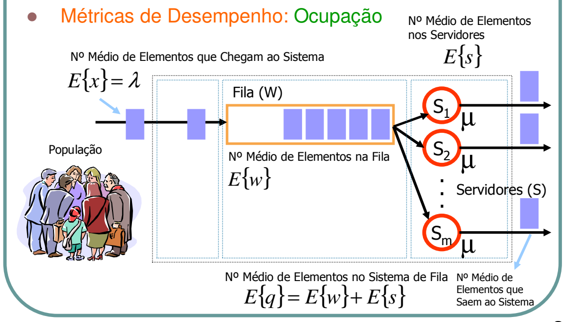


Sistema de Fila (Q) = Fila (W) + Servidor (S)

Introdução a Teoria de Filas

Autor: Prof. Antonio M. Alberti - INATEL

- Métricas de Desempenho: Ocupação



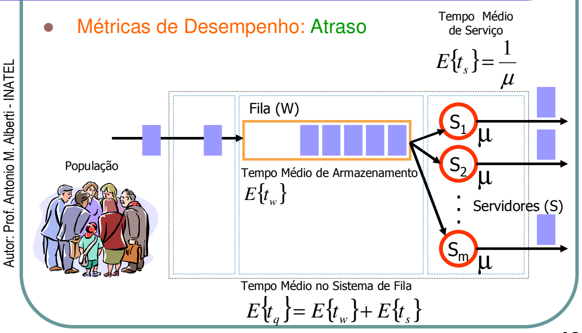
$E\{y\}=\mu$

9

Introdução a Teoria de Filas

Autor: Prof. Antonio M. Alberti - INATEL

- Métricas de Desempenho: Atraso



10

Introdução a Teoria de Filas

- A notação de Kendall foi desenvolvida em 1951 para descrever o comportamento de um sistema de fila em uma única frase:

$$A / B / m / K / S / X$$

- Processo de Chegada
- Processo de Atendimento
- Número de Servidores
- Nº Total de Elementos no Sistema
- Tamanho da População
- Disciplina de Serviço

11

Introdução a Teoria de Filas

- Notação de Kendall Expandida:

$$A / B / m / J / K / S / X$$

- Processo de Chegada
- Processo de Atendimento
- Número de Servidores
- Número Máximo de Elementos na Fila
- Nº Total de Elementos no Sistema
- Tamanho da População
- Disciplina de Serviço

12

Introdução a Teoria de Filas

- Processo de Chegada (A)
  - Descreve o processo que modela as chegadas de elementos ao sistema.
  - Modelos clássicos utilizados:
    - **M – Markoviano**
      - Intervalo de tempo entre chegadas é exponencial.
    - **D – Determinístico**
      - Intervalo de tempo entre chegadas é constante.
    - **G – Genérico**
      - Intervalo de tempo entre chegadas é tratado de forma genérica, independente da distribuição.

13

Introdução a Teoria de Filas

- Processo de Chegada (A)
  - Apresenta um comportamento *estocástico*.
  - Instantes de chegada: $t_1, t_2, \ldots, t_j$
  - Intervalo entre chegadas: $\tau_j = t_j - t_{j-1}$
  - É geralmente assumido que os intervalos entre chegadas formam uma seqüência de variáveis aleatórias independentes e identicamente distribuídas (iid).
  - O processo de chegada mais comum é o processo de Poisson, que significa que os intervalos entre chegadas são iid e exponencialmente distribuídos.

14

Introdução a Teoria de Filas

- Processo de Atendimento (B)
  - Descreve o processo que modela o atendimento de elementos no sistema.
  - Modelos clássicos utilizados:
    - **M – Markoviano**
      - O tempo de serviço de um elemento é exponencial.
    - **D – Determinístico**
      - O tempo de serviço de um elemento é constante.
    - **G – Genérico**
      - O tempo de serviço de um elemento é tratado de forma genérica, independente da distribuição.

15

Introdução a Teoria de Filas

- Processo de Atendimento (B)
  - É comum assumir que os tempos de serviço dos clientes também são variáveis aleatórias independentes e identicamente distribuídas (iid).
    - Atendimentos Simples / Lote
    - Serviço Independente/Dependente do Estado

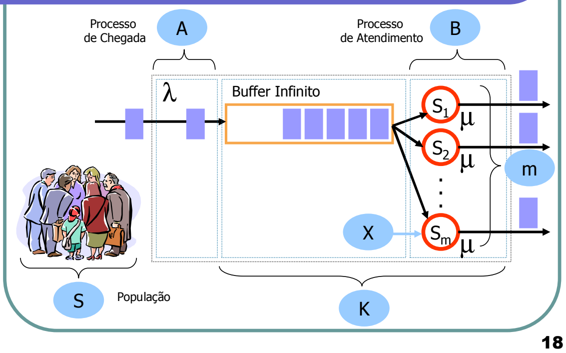
16

Introdução a Teoria de Filas

- Tamanho da População (S)
  - Descreve o tamanho da população que gera elementos para o sistema. Tipicamente é considerada como infinito.
- Disciplina de Serviço (X)
  - Os elementos que aguardam por serviço na fila podem ser selecionados de acordo com uma regra chamada disciplina de serviço. Dentre as principais disciplinas estão:
    - **FCFS** – ***First Come First Served*** **(FIFO)**
      - Primeiro elemento que chega é o primeiro a ser atendido.
    - **LCFS** – ***Last Come First Served*** **(LIFO)**
      - Último elemento que chega é o primeiro a ser atendido.
    - **SIRO** – ***Service In a Random Order***
      - Elementos são atendidos em ordem aleatória.
    - **RR** – ***Round Robin***
    - ***WRR – Weighted Round Robin***

17

Introdução a Teoria de Filas

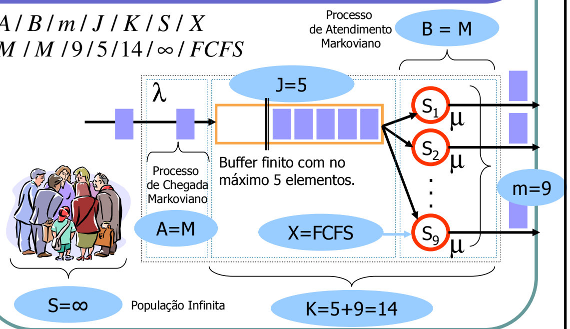


18

Introdução a Teoria de Filas

$A / B / m / J / K / S / X$

$M / M / 9 / 5 / 14 / \infty / FCFS$



Introdução a Teoria de Filas

- Exemplo - M/M/3/20/1500/FCFS
  1. O tempo entre chegadas sucessivas é exponencialmente distribuído (M)
  2. Os tempos de serviço são exponencialmente distribuídos (M)
  3. Existem três servidores (3)
  4. A fila possui um buffer para 20 clientes. Isso consiste de três lugares para clientes sendo servidos e 17 lugares para clientes a espera de serviço (20)
  5. Existem um total de 1500 clientes que podem ser servidos (1500)
  6. A disciplina de serviço é FIFO (FCFS)

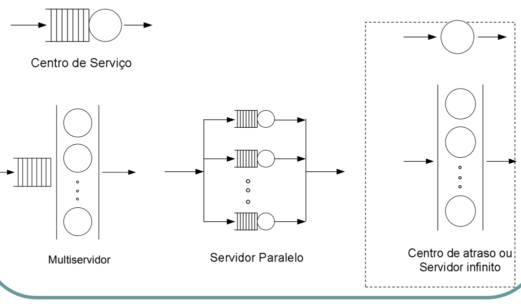
Introdução a Teoria de Filas

- Sistemas Simples de Filas: são em geral utilizados para representar recursos compartilhados de rede.
- Redes de Filas: são geralmente utilizadas para representar outros tipos de redes.
  - Redes de Processos
  - Redes de Comutação

Introdução a Teoria de Filas

- Fontes de Atraso na Rede
  - Atraso de Processamento
    - Assuma que a velocidade de processamento não é uma restrição.
  - Atraso de Filas
    - Tempo de espera no *buffer* para transmissão.
  - Atraso de Transmissão
  - Atraso de Propagação
    - Tempo gasto no enlace – transmissão do sinal elétrico.
    - Independente do tráfego transportado pelo enlace.
- Foco de Estudo: Filas + Atraso de Transmissão

Introdução a Teoria de Filas



Distribuições de Probabilidade

- Determinística (D)
  - Implica que os tempos entre as chegadas ou de serviço são constantes.
  - Não há variância estatística.
  - Pelo menos uma das distribuições (chegada ou serviço) precisa ser aleatória, caso contrário o sistema de filas terá baixa aplicabilidade no mundo real.

Distribuições de Probabilidade

- Exponencial (M)
  - Para Chegadas (A): o intervalo entre uma chegada e a próxima é completamente independente do período anterior.
  - Para Tempos de Serviço (B): o tempo de serviço atual é independente do tempo de serviço anterior.
  - Esses processos são ditos "sem memória", pois seus intervalos não estão correlacionados no tempo. Portanto, podem ser caracterizados por uma distribuição exponencial.

25

Distribuição Exponencial

- Uma V.A. contínua X segue a distribuição exponencial com parâmetro μ, se a sua função densidade de probabilidade é:

$$f_X(x) = \begin{cases} a \cdot e^{-ax} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Função Distribuição (Acumulada) de Probabilidade:

$$F_X(x) = P\{X \leq x\} = \begin{cases} 1 - e^{-ax} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Demonstração:

$$F_X(x) = \int_0^x a\, e^{-ay} dy \qquad \int e^u\, du = e^u \qquad \int e^{ax}\, dx = \frac{e^{ax}}{a}$$

26

Distribuição Exponencial

- Média e Variância:

$$E[X] = \frac{1}{a} \qquad \sigma_X^2 = Var[X] = \frac{1}{a^2}$$

- Demonstração:

$$E[X] = \int_0^\infty x\, f_X(x) dx = \int_0^\infty x\, a\, e^{-ax}\, dx =$$

$$= -x\, e^{-ax}\Big|_0^\infty + \int_0^\infty e^{-ax}\, dx = \frac{1}{a}$$

$$E[X^2] = \int_0^\infty x^2\, a\, e^{-ax}\, dx = -x^2 e^{-ax}\Big|_0^\infty + 2\int_0^\infty x\, e^{-ax}\, dx = \frac{2}{a} E[X] = \frac{2}{a^2}$$

$$Var[X] = E[X^2] - (E[X])^2 = \frac{2}{a^2} - \frac{1}{a^2} = \frac{1}{a^2}$$

27

Distribuição Exponencial

- Integração:

$$\int x \cdot e^{ax}\, dx = \frac{e^{ax}}{a}\left(x - \frac{1}{a}\right)$$

$$\int e^{ax}\, dx = \frac{e^{ax}}{a}$$

$$\int x^n \cdot e^{ax}\, dx = \frac{x^n e^{ax}}{a} - \frac{n}{a}\int x^{n-1} e^{ax} dx$$

28

Propriedade da Ausência de Memória

- A história passada não influencia no futuro:

$$P\{X > x + t \mid X > t\} = P\{X > x\}$$

- Prova:

$$P\{X > x + t \mid X > t\} = \frac{P\{X > x + t, X > t\}}{P\{X > t\}} = \frac{P\{X > x + t\}}{P\{X > t\}}$$

$$= \frac{e^{-\mu(x+t)}}{e^{-\mu t}} = e^{-\mu x} = P\{X > x\}$$

- A distribuição exponencial é a *única* distribuição contínua com a propriedade da ausência de memória.

29

Distribuições de Probabilidade

- Uniforme (U)
  - Os tempos de chegada estão limitados por algum valor finito ($a \leq x \leq b$)
  - A probabilidade de *X* assumir qualquer dos valores do intervalo é a mesma → Média = (a + b)/2
- Arbitrária ou Geral (G)
  - Não é especificada uma distribuição de probabilidade para os tempos de chegada e serviço.
  - Resultados são válidos para todas as distribuições.

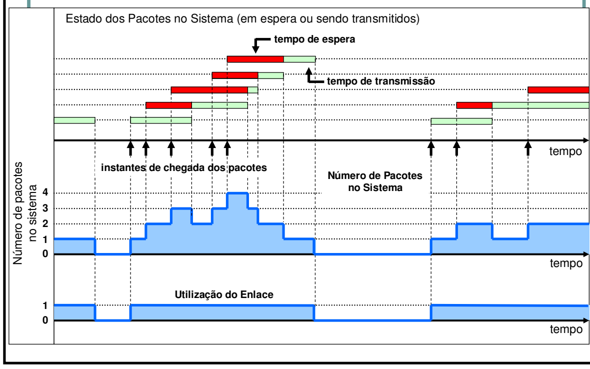
30

Processos Estocásticos

- Os *processos estocásticos* são funções ou seqüências aleatórias dependentes do tempo.
- Exemplos:
  - *n(t)* – Número de processos rodando na CPU de um computador.
  - *w(t)* – Tempo de espera em uma fila.
- Os *processos estocásticos* são úteis para representar o estado de sistemas com filas.

31

Processo de Tráfego de Dados



Tipos de Processos Estocásticos

1. **Processos de Estados Discretos e Estados Contínuos:**

- Discreto: Se o número de valores de estado possíveis é finito ou contável. Estes processos também são chamado de *cadeia estocástica*. Exemplo: O número de processos, *n(t)*, em uma CPU pode apenas assumir valores discretos 0,1,2,...

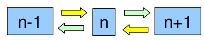
  Contínuo: O estados podem assumir qualquer valor real. Exemplo: O tempo de espera na fila, *w(t)*

2. **Processos de Markov:** Se os estados futuros de um processo são independentes do passado e dependem exclusivamente do estado atual. Isso torna o processo mais fácil de ser analisado pois não é necessário relembrar toda a trajetória passada. Um processo de Markov de estados discretos é chamado de *Cadeia de Markov*

32

Tipos de Processos Estocásticos

3. **Processos de Nascimento e Morte:** processos de Markov de estados discretos em que as transições entre estados estão restritas a estados vizinhos.
   Exemplo: Número de processos em um servidor único com chegadas individuais.

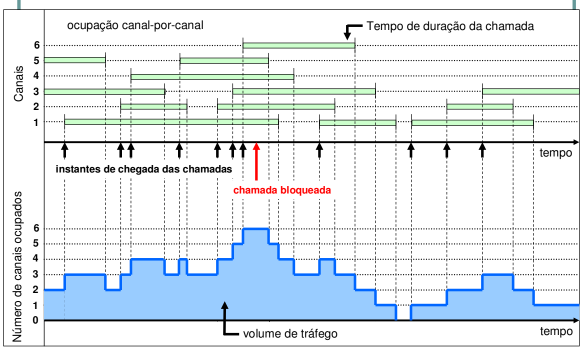


4. **Processos de Poisson:** Se os intervalos entre chegadas for *iid* e exponencialmente distribuído, o número de chegadas, *n*, em um dado intervalo (*t,t+x*) terá uma distribuição de Poisson. O processo de chegadas é chamado de *Processo de Poisson*.
   As chegadas são "sem memória", pois o tempo entre as chegadas é *iid* e exponencialmente distribuído.

Exemplo Tráfego de Vídeo MPEG

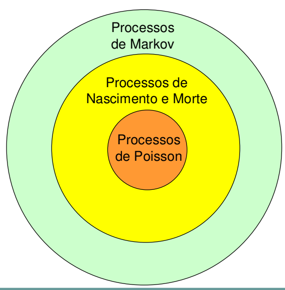
35

Processo de Tráfego Telefônico



Tipos de Processos Estocásticos

- Relação entre Processos Aleatórios



36

Distribuição de Poisson

- Uma V.A. discreta *X* segue a distribuição de Poisson com parâmetro λ se a sua função densidade de probabilidade for:

$$P\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0,1,2,...$$

- Média e Variância

$$E[X] = \lambda, \quad \mathrm{Var}(X) = \lambda$$

- Ampla aplicabilidade na modelagem do número de eventos aleatórios que ocorrem durante um dado intervalo de tempo – *O Processo de Poisson*:
  - Clientes que chegam num posto de serviço durante um dia.
  - Chamadas telefônicas recebidas por engano durante uma semana.
  - Pacotes que chegam a um roteador da rede.

Aproximação de Poisson por Binomial

- Distribuição Binomial com parâmetros *(n,p)*

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

- Como $n \to \infty$ e $p \to 0$, com $np = \lambda$ moderado, a distribuição binomial converge para Poisson com parâmetro λ

Soma de Variáveis Aleatórias de Poisson

- X_i , $i = 1, 2, \ldots, n$, são variáveis aleatórias independentes
- X_i segue a distribuição de Poisson com parâmetro λ_i
- Soma parcial definida como:

$$S_n = X_1 + X_2 + ... + X_n$$

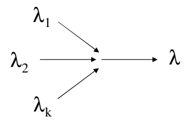
- S_n segue a distribuição de Poisson com parâmetro λ

$$\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n$$

Propriedades dos Fluxos de Poisson

- Superposição de Fluxos

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



- A superposição de fluxos de Poisson dá como resultado um novo fluxo de Poisson cuja taxa de chegada é o somatório das taxas dos fluxos originais.

Amostrando uma V.A. de Poisson

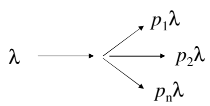
- Considere um processo de Poisson, *X*, com taxa média λ
- Se o processo (fluxo) *X* for dividido em n sub-processos, onde a probabilidade do fluxo seguir o i-ésimo sub-fluxo é p_i, $i = 1, 2, \ldots, n$

$$p_1 + p_2 + \ldots + p_n = 1$$

- X_i denota o número de chegadas do tipo i
- $X_1, X_2, \ldots X_n$ são independentes
- X_i segue a distribuição de Poisson com taxa média $\lambda_i = \lambda\, p_i$

Propriedades dos Fluxos de Poisson

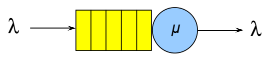
- Divisão de Fluxos



- Se um fluxo de Poisson for dividido em *n* sub-fluxos com probabilidade p_i de um usuário seguir o sub-fluxo *i*, então cada sub-fluxo é também um fluxo de Poisson com taxa média $p_i\lambda$

Propriedades dos Fluxos de Poisson

- Se as chegadas a uma fila com um servidor único e tempo de serviço exponencial forem Poisson com taxa média λ, então as partidas também serão Poisson, com a mesma taxa λ (desde que $\lambda < \mu$)

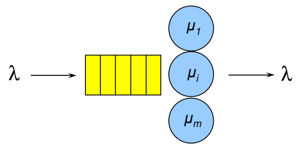


- $\lambda < \mu$: condição de equilíbrio do sistema

43

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 serão Poisson, com a mesma taxa λ (desde que $\lambda < \Sigma\, \mu_i$)



44

Processo de Poisson

- {A(t): t≥0} Processo de Contagem
  - A(t) é o número de eventos (chegadas) que ocorreram desde instante 0 – quando A(0)=0 – até o instante t.
  - A(t)-A(s) número de chegadas no intervalo (s, t].
- O número de chegadas em intervalos disjuntos é independente.
- Número de chegadas em qualquer intervalo (t, t+τ] de tamanho τ.
  - Depende apenas do tamanho τ.
  - Segue a distribuição de Poisson com parâmetro λτ.

$$P\{A(t+\tau)-A(t)=n\}=e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!},\ n=0,1,...$$

- Número médio de chegadas λτ; λ é a taxa de chegada.

45

Intervalo Entre Chegadas

- Os intervalos entre chegadas para um processo de Poisson são independentes e seguem a distribuição exponencial com parâmetro λ
- t_n: tempo da *n*-ésima chegada;
- $\tau_n = t_{n+1} - t_n$: *n*-ésimo intervalo de chegada

$$P\{\tau_n \le s\} = 1 - e^{-\lambda s},\ s \ge 0$$

Prova:

- Função distribuição de probabilidade

$$P\{\tau_n \le s\} = 1 - P\{\tau_n > s\} = 1 - P\{A(t_n + s) - A(t_n) = 0\} = 1 - e^{-\lambda s}$$

- A independência vem da independência do número de chegadas em intervalos disjuntos

46