Universidade Estadual do Ceará (UECE)
Centro de Ciências e Tecnologia (CCT)
Centro Federal de Educação Tecnológica do Ceará (CEFET-CE)
Diretoria de Pesquisa e Pós-Graduação – DIPPG

MESTRADO INTEGRADO PROFISSIONALIZANTE EM COMPUTAÇÃO APLICADA-UECE/CEFET

TEORIA DAS FILAS COMO FERRAMENTA PARA ANÁLISE DE DESEMPENHO DE SISTEMAS DE ATENDIMENTO: ESTUDO DO CASO DE UM SERVIDOR DA UECE

Candidato: Edwin Carrión
Orientador: Prof. Dr. Antônio Clecio Fontelles Thomas

EDWIN ARTURO CARRIÓN

TEORIA DAS FILAS COMO FERRAMENTA PARA ANÁLISE DE DESEMPENHO DE SISTEMAS DE ATENDIMENTO: ESTUDO DO CASO DE UM SERVIDOR DA UECE

Dissertação submetida à Coordenação de Mestrado Profissional em Computação da Universidade Estadual do Ceará como requisito parcial para obtenção do título de mestre.

FORTALEZA – CEARÁ
2007

FICHA CATALOGRÁFICA

C316t Carrión, Edwin Arturo

Teoria das filas como ferramenta para análise de desempenho de sistemas de atendimento: estudo do caso de um servidor da UECE/ Edwin Arturo Carrión.--Fortaleza, 2007

80p.

Orientador: Prof. Dr. Clecio Antônio Fontelles Thomaz

Dissertação (Mestrado Integrado Profissionalizante em Computação Aplicada) - Universidade Estadual do Ceará, Diretoria de Pesquisa e Pós-Graduação.

1. Congestionamento. 2. Teoria das Filas. 3. Distribuições de Probabilidade. I. Universidade Estadual do Ceará. Diretoria de Pesquisa e Pós-Graduação.

CDD:001.6

Agradecimentos

Primeiro agradecer a Deus pela saúde e serenidade para cada dia da minha vida ter disposição para terminar esta dissertação.

Ao meu orientador, Prof. Clecio Thomaz, por estar sempre disponível em responder minhas duvidas referente ao trabalho.

Aos professores pelas orientações para melhorar o trabalho, em particular ao Prof. Guilherme Ellery pela entrega e compromisso para com o curso.

Dedicatória

Dedico este trabalho a toda minha família pelo apoio dado durante este longo período de estudo.

RESUMO

O excesso de mensagens que chegam a um servidor gera congestionamento. Esta pesquisa propõe otimizar o congestionamento do servidor da UECE usando a teoria das filas como ferramenta. Para entender o problema de otimizar um servidor, o trabalho é baseado no estudo dos últimos avanços feitos nesta área mediante leitura de artigos científicos publicados e literatura bibliográfica de filas de espera. Chega-se a descobrir que devido à complexidade de predizer o comportamento do tráfego digital, este tipo de congestionamento é extremadamente difícil de otimizar. Por tanto, se assume que as chegadas e saídas das mensagens no servidor da UECE são governadas pela distribuição exponencial. Foi elaborada uma pesquisa de campo para obter os dados de congestionamento do servidor da UECE e com os dados se calculou as variáveis envolvidas numa fila de espera gerada pelo congestionamento no servidor usando fórmulas apropriadas do modelo que se adapte melhor ao tráfego gerado no servidor da UECE. Analisam-se os resultados obtidos para otimizar o congestionamento gerado no servidor da UECE.
Também se descrevem os diferentes tipos de modelos da teoria das filas usando exemplos. Mostra-se o comportamento do tráfego comparado com o comportamento de congestionamentos conhecidos como o de chamadas telefônicas na camada de ligação de dados do modelo de interconexão de sistemas abertos. Finalmente, se identificam as distribuições apropriadas para este tipo de congestionamento (distribuições de cauda longa) e identificam-se métodos que servem para encontrar de forma aproximada a transformada de Laplace destas distribuições. Mas estes métodos não são analisados aqui devido a seus elevados conhecimentos de matemática avançada.

Palavras-chave: Congestionamento,Teoria das Filas, Distribuições de Probabilidade

ABSTRACT

The excess of messages that get into a computer server generates congestion. The purpose of this study is to use the theory of queues as an instrument to optimize the server of UECE. To understand the problem how to optimize the server, study is done on published articles and bibliography on queues. As a result, it is discovered that this type of queue is extremely complicated to optimize because of the lack of understanding in predicting the behavior of this kind of traffic. Because of that, it is assumed that arrivals and processing times of messages that get into the server are activities governed by the exponential distribution. There were questions to be asked to the people in the computer lab of UECE to get information related to the traffic in the server. Based on that, it was calculated all the variables of the queue generated in the server. The congestion has to fit a specific type of model of the theory of queues. The results gotten from the formulas are analyzed with the intention to optimize the traffic generated in the server of UECE.

In addition, this project describes the types of queues in general and compares the behavior of the traffic with the traffic created by human voice such as phone calls, which is a known type of traffic. This is done in the link layer of the OSI model. Finally, there are some types of distributions (heavy tail distributions) that govern better the behavior of this traffic. These distributions are mentioned in this study. It mentions some methods created by scientists to approximate the Laplace transform of these distributions. But these methods are not analyzed here because of its advanced mathematical concepts.

Key words: Theory of Queues, Probability Distributions, Congestion.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

CA	Custo médio de atendimento
CE	Custo médio de espera
CT	Custo médio total no sistema de filas
D	Atividades descritas por uma distribuição determinística (variância = 0)
E_k	Atividades descritas pela distribuição de Erlang de parâmetro R = λk e com forma k
FCFS	Primeiro a chegar, primeiro a ser atendido.
GD	Disciplina geral de uma fila de espera.
G	Atividades descritas por uma distribuição geral de probabilidade.
GI	Atividades de chegada descritas por uma distribuição geral de probabilidade.
GPSSH	Programa de Simulação
K	Número máximo de clientes no sistema.
L	Número médio de clientes no sistema.
L_q	Número médio de clientes na fila de espera.
L_s	Número médio de clientes em atendimento.
LCFS	Último a chegar, primeiro em ser atendido.
M	Atividades descritas por distribuições de Poisson ou Exponencial
S	Número de atendentes no sistema
TMM	Método Comparativo de Transformação.
TAM	Método de Aproximação da Transformação
TRM	Método TAM de Recursão
W	Tempo médio de um cliente no sistema.
W_q	Tempo médio de um cliente na fila de espera.
W_s	Tempo médio de um cliente em atendimento.

SUMÁRIO

1. INTRODUÇÃO

O objetivo da teoria das filas é otimizar o desempenho de um sistema, reduzindo seus custos operacionais. Para otimizar o desempenho dos modelos de filas de espera, é necessário analisar os resultados gerados por fórmulas apropriadas a um modelo específico. Resultados que permitam realizar a análise de uma situação particular, onde eles podem ser gerados manualmente substituindo os dados de entrada nas fórmulas; ou podem ser obtidos através de um programa de computador como por exemplo um *Add-in* em Excel, escrito no linguagem de programação Visual Basic, ou programas chamados *applets* (pequenos programas desenvolvidos em Java).

Em geral todo sistema de filas tem diferentes características, mas suas formas de funcionamento são similares. Ou seja, existem formas de chegada e formas de atendimento. Para obter resultados de um modelo é fundamental ter alguns dados de entrada para alimentar as fórmulas de uma fila de espera, como por exemplo a razão de chegada, a razão de atendimento etc. Também é necessário conhecer outras características de uma fila de espera, tais como: se um sistema de filas de espera é Markoviano ou não Markoviano. Todas essas características serão explicadas no transcurso deste estudo do sistema da teoria de filas de espera.

As atividades de chegada e atendimento dos pacotes de informação são governadas por uma distribuição de probabilidade. O tipo de distribuição de probabilidade é fundamental para estudar o congestionamento no servidor. Logo as distribuições escolhidas devem ter uma maneira para encontrar seus momentos (valor esperado, variância). Os momentos das distribuições de probabilidade são necessários nos estudos de teoria das filas como ferramenta de otimização. No desenvolvimento do trabalho serão identificadas várias classes de distribuições de probabilidade apropriadas para este tipo de congestionamento. No caso de encontrar dificuldade em encontrar os momentos destas distribuições, será necessário encontrar métodos aproximados para obter seus momentos (transformada de Laplace). Estes métodos são identificados no desenvolvimento do trabalho, mas não serão analisados aqui.

O congestionamento do tráfego de mensagens no servidor da UECE é uma aplicação moderna da teoria das filas. Usando esta teoria, se procura otimizar

este congestionamento. Por razões matemáticas, o congestionamento no servidor é tratado aqui como um modelo markoviano. Ou seja, a distribuição exponencial descreve as chegadas e também os tempos de processamento das mensagens. Nessas condições, assumindo constantes as razões de chegada e atendimento, como dados de entrada se deve ter: a razão de chegada das mensagens, a razão de atendimento do servidor, e sua capacidade de guardar mensagens no *buffer*. Logo se aplica a teoria das filas e assim se encontra uma possibilidade de otimizar seu desempenho. De outra forma, torna-se muito complicado otimizar o desempenho do servidor porque que os cientistas ainda não puderam predizer de forma clara o comportamento deste tipo de congestionamento. O principal desafio é que o comportamento deste tipo de tráfego é totalmente diferente que os congestionamentos que existem em diversas situações, por exemplo, o tráfego de chamadas telefônicas. Por tanto, existe grande dificuldade em encontrar uma solução clara a este tipo de problema. No desenvolvimento desta pesquisa também será discutido o comportamento do tráfego no servidor da UECE. Este comportamento é analisado na camada de ligação de dados do modelo de interconexão de sistemas abertos.

1.1 Objetivos

Geral

Analisar a possibilidade de otimizar o congestionamento do servidor da UECE usando a teoria das filas como ferramenta.

Específicos

- Estudar algumas distribuições de probabilidades fundamentais para conhecer o funcionamento de uma fila de espera;
- Explicar o que significa um processo estocástico e informar o significado das cadeias de Markov e o processo de Markov;
- Analisar os resultados obtidos usando as fórmulas da teoria das filas aplicadas no servidor com o objetivo de otimizar seu desempenho;
- Descrever os diferentes tipos de filas de espera, usando exemplos em cada modelo e identificar custos no desenho de uma fila de espera no modelo MM1;
- Apresentar as distribuições apropriadas para este congestionamento, identificando-se alguns métodos para obter os momentos de forma aproximada das distribuições que descrevem o congestionamento no servidor da UECE; e,

finalmente, explicar o comportamento do tráfego de mensagens no servidor da UECE.

1.2 Delimitação do problema

Como otimizar o desempenho no sistema de atendimento do provedor UECE/NPTEC?

1.3 Metodologia

Segundo Stake (1994) o estudo de caso não é um método, mas a escolha de um objeto a ser estudado. O estudo de caso pode ser único ou múltiplo e a unidade de análise pode ser um ou mais indivíduos, grupos, organizações, eventos, paises, ou regiões.

Esta pesquisa usa o estudo de caso como forma de estudar o comportamento do tráfego de mensagens no servidor da UECE utilizando a teoria das filas como ferramenta de otimização.

Segundo Severino (2000) existe vários métodos de pesquisa: pesquisa bibliográfica, pesquisa de campo, pesquisa experimental, pesquisa documental, pesquisa histórica etc.

Para entender o funcionamento de um sistema de filas foi realizada uma pesquisa bibliográfica (estudo sistematizado desenvolvido com base a material publicado em livros, revistas, jornais, redes eletrônicas), incluindo artigos e documentos publicados. O levantamento bibliográfico explica os diferentes tipos de modelos de um sistema de filas. Os artigos publicados analisam o congestionamento num servidor em geral. Nestas publicações buscaram-se os últimos avanços para otimizar o congestionamento no servidor da UECE.

Foi feito um levantamento de contribuições no laboratório de computação da UECE. Para isto, foi feita uma pesquisa de campo (informação coletada mediante entrevistas) no laboratório de computação da UECE para obter dados do servidor. Foram formuladas as seguintes perguntas ao pessoal dessa dependência.

- Como obter os dados de chegada das mensagens que entram ao servidor e os dados das mensagens enviados pelo o servidor?
- Que tipo de software é usado para obter os dados das mensagens que entram e saem do servidor da UECE?

A coleta de respostas e contribuições foi feita pelo pesquisador com a participação do orientador. Obtiveram-se os dados das mensagens de chegada e

saída do servidor (mensagens 1 mês). Os dados levantados na entrevista foram coletados para logo utilizar a teoria das filas como ferramenta para ter uma possibilidade de otimizar o servidor.

1.4 Estrutura Interna do Trabalho

O trabalho inicia no capitulo 1 com a introdução. Mencionam-se os objetivos a serem alcançados e a delimitação do problema. Também neste capitulo se descreve a metodologia usada para o desenvolvimento do trabalho.

No capitulo 2, descreve-se a historia das filas de espera e brevemente se faz um enfoque não matemático das filas de espera. Primeiro é explicado a nomenclatura de um modelo de filas de espera. A seguir se inicia a parte matemática definindo os momentos de uma distribuição de probabilidade de forma geral. São descritas as distribuições: Poisson, Exponencial e Erlang. Logo se define o processo estocástico descrevendo o processo quando o parâmetro tempo é discreto e quando é continuo. Uma vez explicados estes conceitos matemáticos, se aborda à teoria das filas. Primeiro o se descreve as formas de chegada e atendimento. Segundo é introduzido o processo de nascimento e morte como ferramenta para encontrar as fórmulas dos modelos markovianos de um sistema de filas.

No capitulo 3 se apresenta os modelos das teorias das filas. Estes modelos são divididos em modelos markovianos e modelos não markovianos. As fórmulas são usadas, incluindo o teorema de Little, para otimizar diversos modelos mostrados mediante exemplos específicos. Também se identificam os custos numa fila de espera do modelo MM1 e se analisa o comportamento transitório de um congestionamento.

Uma vez que se familiariza com a parte teórica da teoria das filas é introduzido o capitulo 4. Aqui se inclui algumas aplicações onde os especialistas no assunto utilizam a teoria das filas como ferramenta de otimização. Logo o trabalho é focado no tema principal desta pesquisa que é a otimização do congestionamento do servidor da UECE. Também se analisa o congestionamento num servidor de forma gráfica.

No capitulo 5 se descreve os resultados alcançados. Logo se faz os comentários finais e recomendações. Por último se mostra a bibliografia consultada em ordem alfabética.

2. FUNDAMENTAÇÃOTEÓRICA

2.1 **História das Filas de Espera**

“Agner Krarkup Erlang (Janeiro 1, 1878-Fevereiro 3, 1929) foi o matemático, estatístico, engenheiro dinamarquês que idealizou pela primeira vez os conceitos de Engenharia de Tráfego (traffic engineering) e de Teoria das Filas.

A teoria das filas, como tal, foi desenvolvida para prover modelos matemáticos que predizem o comportamento de sistemas que tentam providenciar atendimento às demandas em contínuo crescimento aleatório.

Trabalhando na empresa “Copenhagen Telephone Company”, foi que Erlang teve que resolver um clássico problema de determinar quantos circuitos são necessários para providenciar um atendimento aceitável nas chamadas telefônicas. Mas seu raciocínio o ajudou a perceber que a matemática resolveria outro problema, isto é: quantos operadores de telefone são necessários para atender um número de chamadas telefônicas determinadas previamente. Nessa época a maioria das centrais telefônicas, usava trabalhadores como operadores para gerenciar as chamadas telefônicas, conectando os fios telefônicos nas tomadas elétricas das placas com circuitos.

Existiam avanços nas aplicações telefônicas, mas na teoria das filas não teve um avanço semelhante. Logo, a partir da década dos anos 50, foi quando as aplicações em áreas além dos sistemas de telefone começaram a evoluir.

Erlang trabalhou no desenvolvimento da área de tráfego nos sistemas de chamadas telefônicas e publicou o seguinte:

- Em 1909, “A Teoria das Probabilidades e as Conversações Telefônicas” em que provou que a distribuição de Poisson se aplica ao tráfego aleatório de chamadas telefônicas.
- Em 1917, “Soluções de Alguns Problemas na Teoria de Probabilidades de Importância nas Chamadas Automáticas de Telefone” em que inclui sua fórmula clássica de tempo de espera e tempo perdido.“ (CHELST, 2006, P.84)

2.2 **Psicologia de uma Fila de Espera**

“A experiência de esperar em uma fila é influenciada pelo ambiente na sala de espera e a expectativa do tempo de espera. Imagine ficar esperando parado numa fila de espera pelo dentista por 20 minutos, sabendo que um paciente está gritando numa sala de atendimento. Agora imagine uma espera alternativa numa

cadeira confortável, com acesso às mais modernas revistas com temas interessantes. Para seu filho de dez anos, existe uma máquina de vídeo game e a sala é à prova de som.

Muitas empresas (Disney é um exemplo) se tornaram espertas em entender a psicologia de esperar numa fila. Ficar numa fila de espera que tenha movimento é menos tedioso que numa fila de espera sem mover-se. Monitores de TV com imagens interessantes ajudam aos usuários a esquecer a hora no relógio. Somado a isto se os usuários podem ver e escutar aos que completaram seu tempo de espera, a antecipação cresce e a espera parece que vale a pena. Se os usuários que ingressam na fila de espera são informados que têm que esperar 15 minutos, pelo menos eles podem decidir se ficam ou não. Se ficassem e a demora é menor que os 15 minutos, eles ficariam gratamente surpreendidos.

Outra dimensão da psicologia de esperar é o sentimento de justiça. Pode ser muito desagradável ver alguém chegar depois e ser atendido mais rápido. Isto poderia acontecer se existissem duas filas separadas. Você pode ficar parado quando o cliente que está sendo atendido demora muito. Como resultado, fica esperando mais tempo que na outra fila. Muitas empresas sabem disso e como resultado criam apenas uma fila para os clientes. Assim, qualquer pessoa, chegando depois, deve ficar atrás e deve ser atendida somente depois que você termine de ser atendido.” (CHELST, 2006, P.84)

As filas de espera são estudadas usando determinadas distribuições de probabilidade. Estas distribuições de probabilidade governam as atividades de chegada e atendimento de um cliente num sistema de filas.

2.3 **Notação**

Para descrever um sistema de filas de espera, Kendall (1951) inventou a seguinte notação: Cada sistema de filas de espera é descrito por 6 características: 1/2/3/4/5/6.

A primeira característica especifica a natureza do processo de chegada. As seguintes abreviações são usadas:

M = Chegadas com tempos de chegadas independentes, variáveis randômicas identicamente distribuídas (iid) descritas por distribuições exponenciais. Ou chegadas por unidade de tempo independentes, variáveis aleatórias identicamente distribuídas (iid) descritas por distribuições de Poisson.

D = Chegadas com tempo de chegada são iid e determinísticas (variância = 0)

E_k= Chegadas com tempo de chegada são iid e descritas pela distribuição de Erlang de parâmetro R = λk e com forma k

GI = Chegadas com tempo de chegada são iid e controladas por uma distribuição geral.

A segunda característica especifica a natureza dos tempos de atendimento.

M = Tempos de atendimento independentes, variáveis randômicas identicamente distribuídas (iid) e distribuídas exponencialmente.

D = Tempos de atendimento iid e determinísticas

E_k = Tempo de atendimento são iid descritas pela distribuição de Erlang de parâmetro R = kλ com forma k

G = Tempos de atendimento são iid e seguindo uma distribuição geral

A terceira característica é o número de atendentes em paralelo.

A quarta característica descreve a disciplina da fila de espera:

FCFS = primeiro a chegar, primeiro a ser atendido.

LCFS = último a chegar, primeiro em ser atendido.

SIRO = Atendimento em ordem randômico.

GD = Disciplina geral de uma fila de espera.

A quinta característica denota o máximo número permitido de clientes no sistema, incluindo clientes que estão na fila de espera e clientes que estão sendo atendidos pelo servidor.

A sexta característica significa o tamanho da população. A população geralmente é considerada infinita.

Em muitos modelos importantes, as três últimas variáveis são omitidas quando seus valores respectivos são GD/∞/∞. A continuação se mostra exemplos com a notação de Kendall e seu significado:

- M/M/1:chegadas com distribuição de Poisson, atendimento com distribuição exponencial, um atendente, disciplina geral, infinito número de clientes no

sistema e infinita população. Aqui as três últimas variáveis são omitidas mas com significado implícito.

- M/E_2/8/FCFS/10/∞: pode representar uma clínica de saúde com oito doutores, tempos de chegada exponenciais, tempo de atendimento Erlang com duas faces, a disciplina da fila de espera é FCFS (o primeiro em chegar é atendido) e uma capacidade total de dez pacientes

2.4 DESCRIÇÃO DAS DISTRIBUIÇÕES DE PROBABILIDADE

2.4.1 Definição de Valor Esperado

O valor esperado, escrito como E(X), calcula a média de todos os valores possíveis que toma uma variável aleatória X. Cada valor de X tem uma probabilidade de ocorrência P.

Para valores discretos de X a fórmula é:

$$\mathrm{E}(X) = \sum_{i=0}^{i=n} X_i \otimes \mathrm{P}(X = X_i),\ \text{onde}\ \mathrm{P}(\mathrm{X} = X_i) \geq 0 \quad \text{e} \quad \sum_{\mathrm{i}=0}^{\mathrm{i}=\mathrm{n}} \mathrm{P}(\mathrm{X} = \mathrm{X}_\mathrm{i}) = 1, \quad \mathrm{i} = 0,1,2\ldots,\mathrm{n}$$

Para valores contínuos de X a fórmula é:

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x \otimes f(x)dx, \text{onde} \int_{-\infty}^{+\infty} f(x)dx = 1$$

Onde f(x), representa uma função densidade de probabilidade

2.4.2 Definição de Variância

A variância de uma variável aleatória X, Var(X), em média, mede a dispersão de X com respeito a seu valor esperado E(X).

Ou seja, $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2]$

2.4.3 Definição de Desvio Padrão

O desvio padrão de uma variável aleatória X, escrita como $\sigma(X)$, indica em média, quanto os valores da variável aleatória X, se desviam com respeito a seu valor esperado E(X). A fórmula é $\sigma(X) = [\mathrm{Var}(X)]^{1/2}$.

2.4.4 Distribuição de Poisson

Definição: Seja X uma variável aleatória discreta, tomando os seguintes valores: 0, 1, 2,....Quando: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, (k= 0, 1, 2…..)

X tem distribuição de Poisson, com parâmetro λ > 0.

Demonstração

Para verificar que a expressão acima representa uma legítima distribuição de probabilidade, basta observar que $\sum_{k=0}^{\infty} P(X=k)$ = $\sum_{k=0}^{\infty}\left(\frac{e^{-\lambda}\lambda^{k}}{k!}\right)$ = $e^{-\lambda}$. e^{λ} = 1.

$\sum_{k=0}^{k=n}\frac{\lambda^{k}}{k!}$ = 1 + λ + $\frac{\lambda^{2}}{2!}$ +..... $\frac{\lambda^{n}}{n!}$ = e^{λ}

2.4.4.1 **Valor Esperado**

Se X tiver distribuição de Poisson com parâmetro λ então E (X) = λ. Este valor esperado indica a média do número de eventos num intervalo de tempo.

Demonstração: Para n = 0, 1, 2, ...; P(X= n) = a_n ; $|z| \leq 1$ e usando transformações de Z : $P^{T}{}_{x}(z) = E(z^{x}) = \sum_{n=0}^{n=\infty} a_n z^{n}$ A primeira derivada é:

$$\left(\frac{\mathrm{d}\,\mathrm{P}^{\mathrm{T}}x(z)}{d\ z}\right)_{z=1} = \left(\sum_{n=0}^{n=\infty} a_n n z^{n-1}\right)_{z=1} = \sum_{n=0}^{n=\infty} n a_n = E(x).$$ Derivando pela segunda vez:

$$\left(\frac{\mathrm{d}^2\ \mathrm{P}^{\mathrm{T}}x(z)}{dz^{2}}\right) = \left(\sum_{n=0}^{n=\infty} a_n n(n-1) z^{n-2}\right)_{z=1} = \sum_{n=0}^{n=\infty} (n^2 - n) a_n = E(x^2) - E(x)$$

Então: $P^{T}{}_{x}(z) = E(z^{x}) = \sum_{k=0}^{k=\infty}\left(\frac{e^{-\lambda}\lambda^{k}}{k!}\right) Z^{k} = \sum_{k=0}^{k=\infty}\frac{e^{-\lambda}(\lambda Z)^{k}}{k!}$

$\sum_{k=0}^{k=\infty}\frac{(\lambda z)^{k}}{k!} = e^{z\lambda}$ Por tanto, $P^{T}{}_{X}(z) = e^{-\lambda}e^{z\lambda} = e^{z\lambda-\lambda} = e^{\lambda(z-1)}$

$$\text{Derivando a equação acima, fica : } \left(\frac{\mathrm{d}\,\mathrm{P}^{\mathrm{T}}x(z)}{d\ z}\right)_{z=1} = \left(e^{\lambda(z-1)}\lambda\right)_{z=1} = \lambda = E(x)$$

Portanto: E (X) = λ

2.4.4.2 **Variância**

A variância é igual o seu valor esperado. Ou seja, V(X) = λ

Demonstração

$$\text{Derivando pela segunda vez, fica : } \left(\frac{\mathrm{d}^2\ \mathrm{P}^{\mathrm{T}}x(z)}{dz^{2}}\right) = \left(e^{\lambda(z-1)}\lambda\lambda\right)_{z=1} = \lambda^{2} = E(x^2) - E(x)$$

Então: $E(X^2) = \lambda^2 + \lambda$

Variância é igual a ,

$V(X) = E(x^2) - [E(X)]^2$

Por tanto: $V(X) = (\lambda^2 + \lambda) - \lambda^2 = \lambda$. (Cf. Winston, 2004, p.85)

2.4.4.3 **Gráfico**

Figura 1

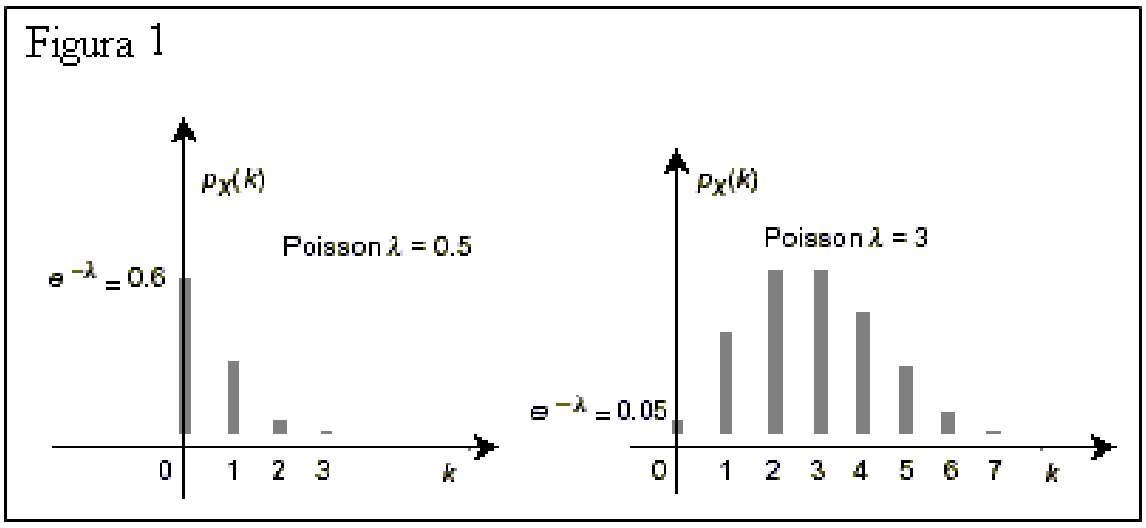


Fig 1: Distribuição de Poisson

Na figura 1 o gráfico da distribuição de Poisson mostra diferentes valores de *λ*. Se *λ = 0,5* logo a distribuição de Poisson é monotomicamente decrescente, mas se *λ = 3*, então a distribuição primeiro cresce e em seguida decresce à medida que o valor de *k* aumenta. (Cf. Bertsekas,Tsitsiklis, 2002, P.84)

Também é importante notar na figura 1 que quando k > *λ* a distribuição decresce gradualmente, mas quando k < *λ* a distribuição decresce rapidamente.

Figura 2

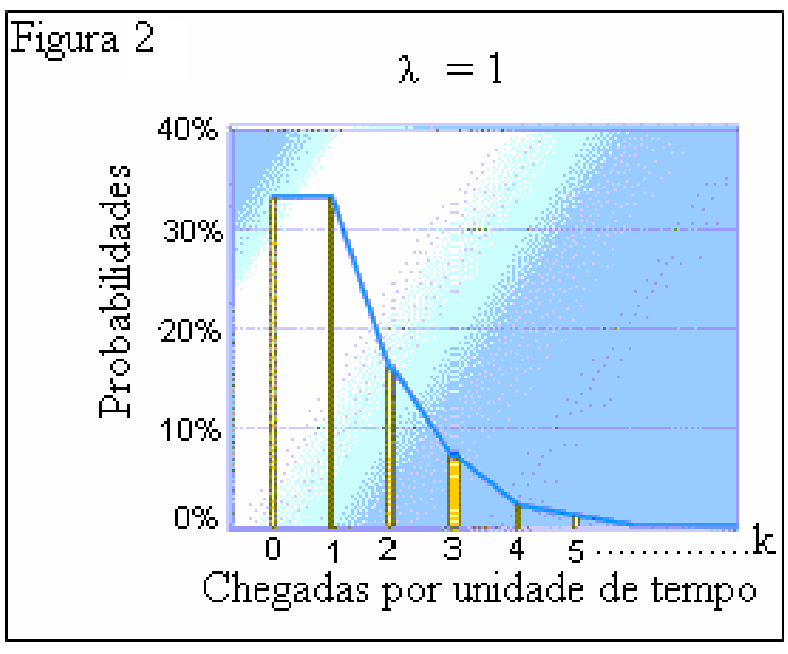


Fig 2: Distribuição de Poisson

Na figura 2, o valor esperado de chegadas por unidade de tempo é 1 e sua máxima probabilidade é sempre perto deste valor (λ), ou seja, zero ou uma chegada tem probabilidade de 33% de ocorrer. À medida que as chegadas sejam

maiores, a probabilidade diminui. Por exemplo, a probabilidade de ocorrer cinco chegadas em um unidade de tempo é perto de 0%. (Cf. Kalinski, 2001, P.84).

A distribuição de Poisson com parâmetro *λ* é uma boa aproximação da distribuição binomial com parâmetros *n* e *p,* dados *λ* = *np*, com *n* muito grande, e p muito pequeno, logo:

$$e^{-\lambda}\frac{\lambda^k}{k!} \approx \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}, \quad k = 0,1,2.....n$$

Exemplo

Seja n = 100 e p = 0,01. Logo a probabilidade de k = 5 sucessos em 100 tentativas é calculada usando a distribuição binomial. Então:

$$\frac{100!}{95!\ 5!}0{,}01^5(1-0{,}01)^{95} = 0{,}00290$$

Usando Poisson com *λ* = *np* = (100) (0.01) = 1, esta probabilidade é aproximadamente igual a:

$$e^{-1}\frac{1}{5!} = 0{,}00306$$

Neste caso os dois valores são aproximados. (Cf. Bertsekas,Tsitsiklis, 2002, P.84)

2.4.5 Distribuição Exponencial

Definição: Uma variável aleatória contínua X, que assume todos os valores não negativos, terá uma *distribuição exponencial* com parâmetro μ > 0, se sua função densidade de probabilidade (*fdp)* for dada por

$$f(X) = \begin{cases} \mu e^{-\mu X}, X \geq 0 \\ 0, X < 0 \end{cases}$$

Demonstração

Por definição a probabilidade de uma variável aleatória contínua é sempre igual à área do gráfico da curva da f.d.p. Neste caso:

$$P(x \geq 0) = \int_0^{+\infty} f(x)dx = \int_0^{\infty} \mu e^{-\mu X}\, dx = -\int_0^{\infty} e^{-\mu X}(-\mu)\, dx$$

$$\int e^{f(x)}\, f'(x)\, dx = e^{f(x)} + C \text{ , onde C é uma constante qualquer}$$

Então:

$$- \int_0^{\infty} e^{-\mu X}(-\mu)\, dx = - e^{-\mu X} \Big|_0^{\infty} = -e^{-\infty} + e^{0} = 0+1 = 1$$

Como $P(X \geq x) = \int_x^{\infty} \mu e^{-\mu X}\, dx = - e^{-\mu X} \Big|_x^{\infty} = - e^{-\mu \infty} + e^{-\mu X} = 0 + e^{-\mu X} = e^{-\mu X}$

Também $P(X \geq x) = e^{-\mu X}$

2.4.5.1 **Valor esperado**

Indica a média do tempo de duração dos eventos descritos pela distribuição exponencial. Seu valor é E(x) = $\frac{1}{\mu}$

Demonstração

$$E(x) = \int_0^{\infty} x \,.\, \mu e^{-\mu X}\, dx$$

Substituindo u(x) = x, u´(x) = 1, v(x) = $-e^{-\mu X}$, v´(x) = $\mu e^{-\mu X}$; fica:

$$E(x) = \int_0^{\infty} u(x) \,.\, v'(x)\, dx$$; logo, integrando por partes:

$$\int_0^{\infty} u(x) \,.\, v'(x)\, dx = \left[u(x) \,.\, v(x)\right]_0^{\infty} - \int_0^{\infty} v(x) \,.\, u'(x)\, dx$$

Esta fórmula é o resultado do seguinte conceito matemático:

A derivada do produto de duas funções quaisquer, u(x) e v(x), que sejam diferenciáveis é $\frac{d[u(x).v(x)]}{dx} = u'(x).v(x) + u(x).v'(x)$. Logo, integrando ambos os membros da igualdade, fica: $u(x).v(x) = \int \left(u'(x).v(x) + u(x).v'(x)\right)dx$

A parte esquerda da igualdade fica assim porque a derivada e seu integral indefinida se anulam. Logo, fica:

$\int_0^{\infty} u(x) \,.\, v'(x)\, dx = \left[u(x) \,.\, v(x)\right]_0^{\infty} - \int_0^{\infty} v(x) \,.\, u'(x)\, dx$. Aqui se usa a integral definida de zero até infinito.

Então, $E(x) = \int_0^{\infty} x \,.\, \mu e^{-\mu X}\, dx = (x \,.\, -e^{-\mu X}) \Big|_0^{\infty} - \int_0^{\infty} (-1)\, e^{-\mu X} \,.\, 1\, dx$

$E(x) = -\infty \,.\, e^{-\mu \infty} + 0 \,.\, e^{-\mu.0} + \int_0^{\infty} e^{-\mu X}\, dx = 0 + 0 + \int_0^{\infty} e^{-\mu X}\, dx$

$E(x) = \int_0^{\infty} e^{-\mu X}\, dx$

Por definição: $\int e^{f(x)} f'(x)\, dx = e^{f(x)}$. Então, fazendo $f(x) = -\mu x$

$E(x) = -\frac{1}{\mu} \int_0^{\infty} (-\mu) e^{-\mu . X}\, dx = -\frac{1}{\mu} e^{-\mu x} \Big|_0^{\infty} = -\frac{1}{\mu} (0 - 1) = \frac{1}{\mu}$, logo

$E(x) = \frac{1}{\mu}$ (Cf. Winston, 2004, p.85)

2.4.5.2 **Variância**

A variância é igual ao quadrado de seu valor esperado. Ou seja, $Var(x) = \frac{1}{\mu^2}$

Demonstração

$E(x^2) = \int_0^{\infty} x^2 \,.\, \mu e^{-\mu X}\, dx$; fazendo $u(x) = x^2$ e $v(x) = -e^{-\mu X}$, que integrando por partes:

$E(x^2) = \int_0^{\infty} u(x) \,.\, v'(x)\, dx$. Integrando por partes do mesmo modo que foi feito nas prévias demonstrações, fica:

$E(x^2) = -x^2 \cdot e^{-\mu X} \Big|_0^{\infty} + 2\int_0^{\infty} (x)\, e^{-\mu X}\, dx = -\infty^2 \,.\, e^{-\mu \infty} + 0^2 e^{-\mu 0} + 2\int_0^{\infty} (x)\, e^{-\mu X}\, dx$

$E(x^2) = 0 + 2\int_0^{\infty} (x)\, e^{-\mu X}\, dx = 2\int_0^{\infty} (x)\, e^{-\mu X}\, dx = 2\,(\frac{1}{\mu})\int_0^{\infty} (x)\, \mu\, e^{-\mu X}\, dx = 2\frac{1}{\mu}\ E(x)$

$E(x^2) = 2\ \frac{1}{\mu}\ \frac{1}{\mu} = \frac{2}{\mu^2}$

$var(x) = E(x^2) - [E(x)]^2$

Logo, Var(x) = $\frac{2}{\mu^2}$ - $\frac{1}{\mu^2}$ = $\frac{1}{\mu^2}$ Var(x) = $\frac{1}{\mu^2}$

2.4.5.3 **Gráfico**

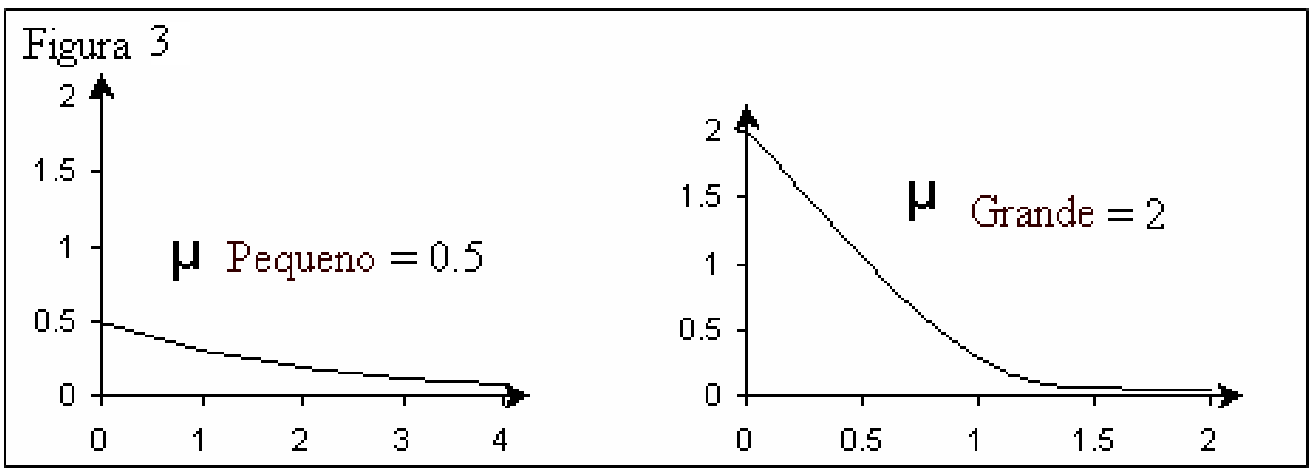


Fig 3: Distribuição Exponencial

Na figura 3, mostra-se o gráfico da função densidade de probabilidade (fdp) de uma distribuição exponencial com parâmetro μ. O primeiro é com μ pequeno e o segundo com μ grande.

A figura 4 mostra que quando (t >1/μ) a probabilidade diminui, e quando (t <1/μ) então a probabilidade aumenta. Na figura também observar que o tempo de atendimento vai para infinito, tendo em conta que a probabilidade de isso acontecer é muito pequena. (Cf. Kalinski, 2000, P.84)

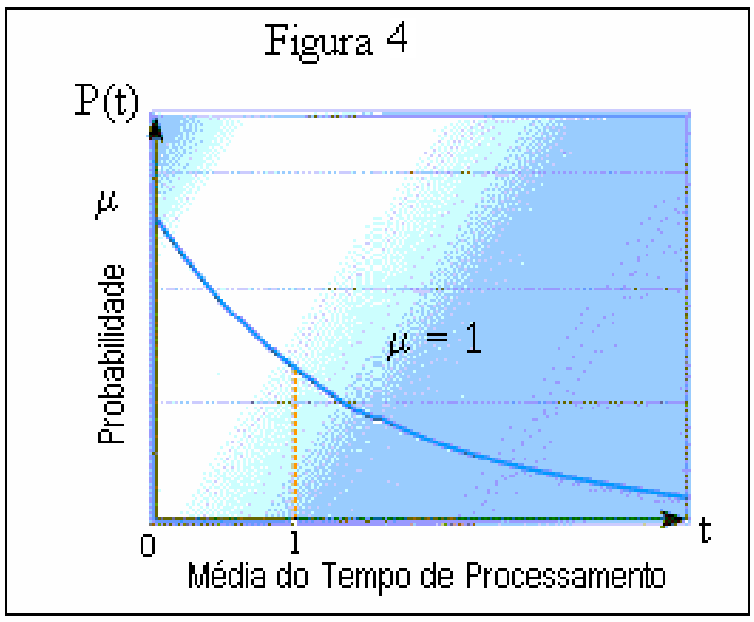


Fig 4: Distribuição Exponencial

Exemplo

O tempo em que um meteorito atinge a Terra no deserto é modelado como uma distribuição exponencial com valor esperado de dez dias. O horário atual é meia-noite. Qual é a probabilidade de que o meteorito chegue entre seis horas e 18 horas no primeiro dia?.

Seja X o tempo transcorrido até o evento acontecer, medido em dias. Logo X descreve uma distribuição exponencial com média de 1/ μ = 10, onde μ = 1/10.

Então como X está em horas e se precisa de X em dias, fazendo a conversão, fica:

$$6h\frac{1\,dia}{24\,horas}=\frac{1}{4}dia \quad \text{Logo}\,18h\frac{1\,dia}{24\,horas}=\frac{3}{4}dia$$

Usar a fórmula **P**($X \geq x$) = **P**($X > x$) = $e^{-\mu x}$.

Então a probabilidade buscada é:

P(1/4 ≤ X ≤ 3/4) = **P**($X \geq 1/4$) - **P**($X > 3/4$) = $e^{-1/40} - e^{-3/40}$ = 0.0476,

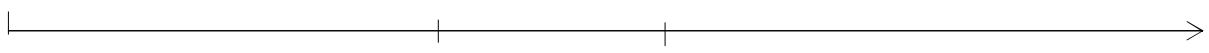
Var(x) = 1/ $(0,1)^2$ = 100 dias. (Cf. Bertsekas, Tsitsiklis, 2002, P.84)

2.4.5.4 **Propriedade da Não-Memória da Distribuição Exponencial**

O tempo de um cliente para ser atendido não depende de quanto tempo já passou desde que o último cliente teve atendimento concluído. Ou seja, não depende do passado, mas somente do futuro. Em termos matemáticos,

$$P\left(\frac{X > t+h}{X \geq t}\right) = P\,(X > h)$$

Demonstração



$$P(X>t+h) = \int_{t+h}^{\infty} \lambda e^{-\lambda t}\,dt = -\int_{t+h}^{\infty} -\lambda e^{-\lambda t}\,dt$$

$\int e^{f(x)} f'(x)\,dx = e^{f(x)}$. Então, fazendo f(x) = -λx fica:

$$-\int_{t+h}^{\infty} -\lambda e^{-\lambda t}\,dt = -e^{-\lambda t}\Big|_{t+h}^{\infty} = -e^{-\lambda\infty} + e^{-\lambda(t+h)} = -\frac{1}{e^{\infty}} + e^{-\lambda(t+h)} = 0 + e^{-\lambda(t+h)} = e^{-\lambda(t+h)}$$

Assim mesmo:

$$P(x \geq t) = \int_{t}^{\infty} \lambda e^{-\lambda t}\,dt = -e^{-\lambda t}\Big|_{t}^{\infty} = -e^{-\lambda\infty} + e^{-\lambda t} \quad \text{Logo,}\; -\frac{1}{e^{\infty}} + e^{-\lambda t} = 0 + e^{-\lambda t}$$

Portanto P(x ≥ t) = $e^{-\lambda t}$

Logo: $P\left(\frac{X > t+h}{X \geq t}\right) = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = \frac{e^{-\lambda t-\lambda h}}{e^{-\lambda t}} = \frac{e^{-\lambda t}e^{-\lambda h}}{e^{-\lambda t}} = e^{-\lambda h}$

Agora: P(X > h) = $\int_{h}^{\infty} \lambda e^{-\lambda h} = -\int_{h}^{\infty} -e^{-\lambda h} = -e^{-\lambda h}\Big|_{h}^{\infty} = e^{-\lambda h}$

Portanto, P$\left(\frac{X > t+h}{X \geq t}\right)$= P (X > h). (Cf. Winston, 2004, p.85)

Exemplo

Considerar esperar um táxi numa estação. Assuma que um táxi, em termo médio, chega à estação a cada 30 segundos. Isto é, o intervalo de tempo médio entre chegadas é de $\frac{1}{\lambda}$= 30 segundos. Supondo que você chegue à estação num instante qualquer. Em termo médio (valor esperado), quanto tempo você espera até a chegada do seguinte táxi?. A maioria responde 15 segundos. Esta resposta é correta se o táxi chega exatamente em 30 segundos (sem variância). Se existe variância, a resposta é sempre maior que 15 segundos. Pode ser demonstrado que se você examina o sistema em qualquer instante, logo:

$$\text{Valor esperado da primeira chegada} = \frac{1}{2}\left[\frac{1}{\lambda} \oplus \lambda \otimes (\text{variância de tempos de chegada})\right]$$

Então, se o termo da variância de tempos de chegada é positivo, logo o valor esperado da primeira chegada é sempre maior que $\frac{1}{2}\left(\frac{1}{\lambda}\right)$. Notar que quando as chegadas são descritas pela distribuição exponencial, sua variância é $\frac{1}{\lambda^2}$, logo o valor esperado da primeira chegada é 30 segundos ($\frac{1}{\lambda}$), propriedade da não memória, e quando a variância é maior que $\frac{1}{\lambda^2}$, o valor esperado da primeira chegada é maior que o intervalo de tempo médio entre chegadas. (Cf. Harvey, 1969, P.85)

2.4.6 Distribuição de Erlang

Se os tempos de atendimento não obedecem a uma distribuição exponencial, então eles são modelados por uma distribuição de Erlang. A distribuição Erlang é uma variável randômica contínua com parâmetro de razão R e parâmetro k. O parâmetro k tem que ser inteiro ≥ 1 e também controla a forma do gráfico da função densidade de probabilidade (fdp) Erlang. A função densidade de probabilidade é a seguinte:

$$f(t) = \frac{R(Rt)^{k-1} e^{-Rt}}{(k-1)!} \qquad (t \geq 0),\ K \geq 1),\ R = k\,\mu$$

Fazendo k = 1, f(t) se converte na distribuição exponencial com R = μk = μ

A distribuição Erlang é igual ao somatório de um número k de variáveis randômicas exponenciais independentes. Ou seja: $E_k = A_1 + A_2 + A_3 + ...+A_k$, onde A_i é uma variável randômica exponencial com parâmetro R = μk. O número k é chamado número das fases da distribuição Erlang. Por exemplo, representando o tempo de atendimento de Erlang, a figura 5 mostra:

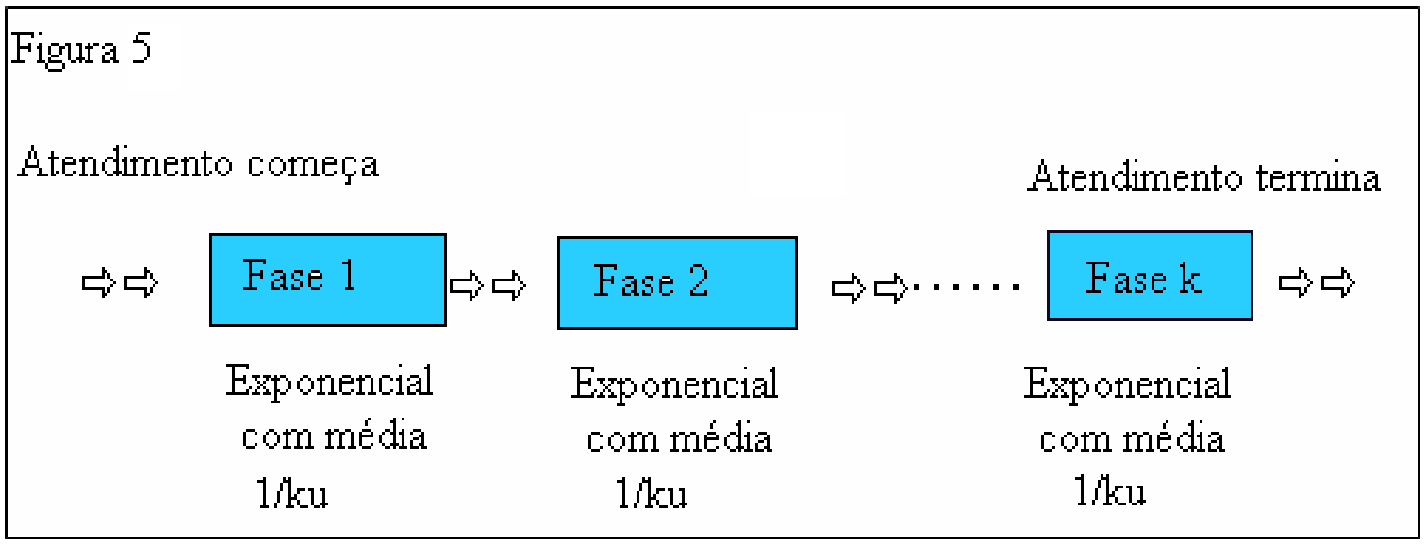


Fig 5: Tempo de Atendimento Erlang

2.4.6.1 **Valor esperado**

Indica a média do tempo de duração dos eventos descritos pela distribuição de Erlang. Seu valor é E(T) = $\frac{k}{R} = \frac{k}{k\mu} = \frac{1}{\mu}$

Demonstração

$$E(T) = \int_0^{\infty} t.f(t).dt$$

$$E(T) = \int_0^{\infty} t.\frac{R(Rt)^{k-1} e^{-R.t}}{(k-1)!}.dt \text{ , } t \geq 0,\ k \geq 1$$

$$E(T) = \int_0^{\infty} \frac{(tR)^1 (Rt)^{k-1} e^{-R.t}}{(k-1)!} dt = \int_0^{\infty} \frac{(Rt)^k e^{-R.t}}{(k-1)!} dt = \frac{1}{(k-1)!} \int_0^{\infty} (\mu kt)^k e^{-\mu kt} dt = \frac{(\mu k)^k k}{k!} \int_0^{\infty} t^k e^{-(\mu k)t} dt$$

Integrando por partes:

f(t) = $-\frac{1}{\mu k}e^{-\mu k t}$, f'(t)= $e^{-\mu\kappa t}$, g(t) = t^k, g'(t) = kt^{k-1}

$\int_0^\infty$ g(t) . f´(t) dt = f(t) . g(t) - $\int_0^\infty$ f(t) . g´(t) dt, substituindo:

E(t) = $\frac{e^{-\mu k t}}{-\mu k} t^k \Big|_0^\infty + \frac{k}{\mu k}\int_0^\infty t^{k-1}e^{-\mu k t}dt = 0 + \frac{(\mu k)^k}{\mu(k-1)!}\int_0^\infty t^{k-1}e^{-\mu k t}dt = \frac{(\mu k)^{k-1}k}{(k-1)!}\int_0^\infty t^{k-1}e^{-(\mu k)t}dt$

$$E(t) = \frac{(\mu k)^{k-1}k}{(k-1)!}\int_0^\infty t^{k-1}e^{-\mu k t}dt \text{ , k≥1}$$

Integrando novamente por partes: $E(t) = \frac{(\mu k)^{k-2}k}{(k-2)!}\int_0^\infty t^{k-2}e^{-\mu k t}dt$ para $k \geq 1$

Se novamente se integra por partes pela terceira vez:

$$E(t) = \frac{(\mu k)^{k-3}k}{(k-3)!}\int_0^\infty t^{k-3}e^{-\mu k t}dt \text{ para } k \geq 1$$

Se continuar assim até derivar por partes pela k-ésima vez:

$$E(t) = \frac{(\mu k)^{k-k}k}{(k-k)!}\int_0^\infty t^{k-k}e^{-\mu k t}dt \text{ para } k \geq 1 \quad E(t) = k\int_0^\infty e^{-\mu k t}dt, k \geq 1$$

E(t) = $k\int_0^\infty e^{-\mu k t}dt = \frac{1}{-\mu}\int_0^\infty (-\mu k)e^{-\mu k t}dt = -\frac{1}{\mu}e^{-\mu k t}\Big|_0^\infty = -\frac{1}{\mu}(0-1) = \frac{1}{\mu}$, Logo E(T) $= \frac{1}{\mu}$

2.4.6.2 **Variância**

É igual ao quadrado de seu valor esperado dividido para o número de fases k. Seu valor é Var(t) = $\frac{k}{R^2} = \frac{k}{(k\mu)^2} = \frac{1}{k\mu^2}$. Se $K \to \infty$. Logo $Var(t) \to 0$.

Demonstração

Por demonstração anterior, Var (t) = $E(t^2) - [E(t)]^2$

$E(t^2)$ = $\int_0^\infty t^2 \frac{R(Rt)^{k-1}e^{-R.t}}{(k-1)!}dt = \frac{1}{R(k-1)!}\int_0^\infty (Rt)^{K+1}e^{-R.t}dt$. Integrando por partes:

f(t) = $-\frac{1}{R}e^{-Rt}$, g(t) = $(Rt)^{k+1} = R^{k+1}(t^{k+1})$, f'(t) = e^{-Rt}, g'(t) = $R^{k+1}[(k+1).\ t^k] = R.(Rt)^k(k+1)$

$$E(t^2) = \frac{1}{R(k-1)!}\left[-\frac{(tR)^{k+1}e^{-R.t}}{R}\Big|_0^{\infty} + \frac{R}{R}(k+1)\int_0^{\infty}(Rt)^k e^{-R.t}dt\right]$$

$$E(t^2) = \frac{1}{R(k-1)!}\left(0 + (k+1)(k-1)!E(t)\right) = \frac{(k+1)\left(E(t)\right)}{R}$$

Var(t) = $\frac{(k+1)\left(E(t)\right)}{R}$ - $\left(\frac{k}{R}\right)^2$ = $\frac{k}{R^2}(k+1) - \frac{k^2}{R^2}$ = $\frac{k}{R^2}$. Portanto: Var(t) = $\frac{1}{k\mu^2}$

2.4.6.3 **Gráfico**

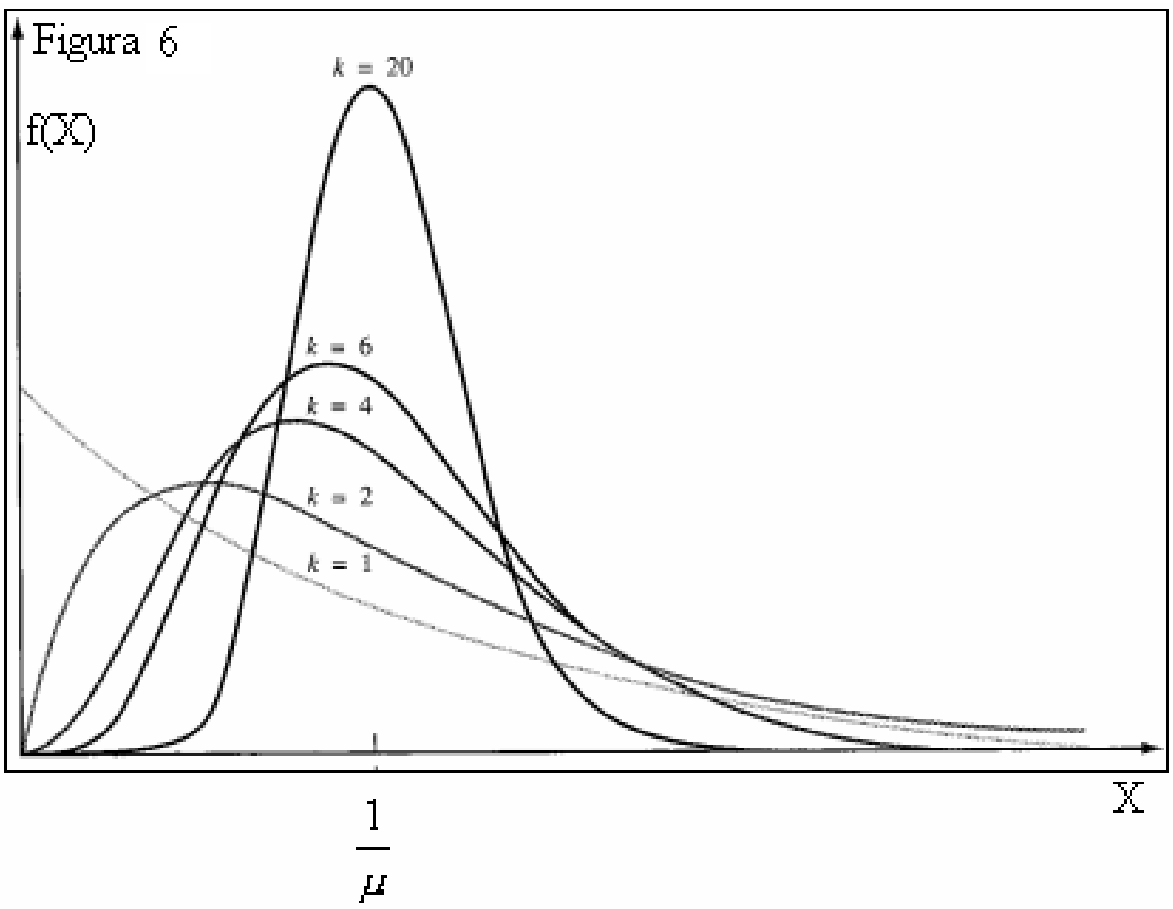


Fig 6: Distribuição de Erlang

Exemplo

Considere um sistema com o modelo M/G/1/GD/∞/∞ no qual a média de chegadas é de dez por hora. Supondo que o tempo de atendimento segue uma distribuição de Erlang com parâmetro R=1 clientes por minuto e parâmetro de forma k = 4.

a) Encontrar o número médio de clientes na fila de espera

b) Encontrar o tempo médio do cliente no sistema

c) Qual será a fração do tempo que o atendente fica ocioso?

R = μ k = 1, 1 = 4 μ → μ = ¼ V = $\frac{1}{k\mu^2} = \frac{1}{4} * 16 = 4$. λ = 10 * 1/60 = 1/6 por minuto.

$$\rho = \frac{\lambda}{\mu} = \frac{1/6}{1/4} = \frac{2}{3}$$

a) $L_q = \frac{\lambda^2 V \oplus \rho^2}{2(1-\rho)} = \frac{(\frac{1}{6})^2(4)+\frac{4}{9}}{2(1-\frac{2}{3})} = \frac{\frac{4}{36}+\frac{4}{9}}{2(\frac{1}{3})} = \frac{20}{36} * \frac{3}{2} = \frac{5}{6}$

b) $W = W_s + W_q = \frac{1}{\mu} + \frac{Lq}{\lambda} = 4 + 5 = 9\,\text{min}$

c) $\pi_0 = 1-\rho$ $\pi_0 = 1 - 2/3 = 1/3$ (Cf. Winston, 2004, p.85)

2.5 PROCESSO ESTOCÁSTICO

Em muitas situações práticas, os atributos de um sistema mudam de forma aleatória com o tempo, como por exemplo: o número de clientes numa fila de espera, o congestionamento no trânsito, o número de itens num depósito, ou o valor de uma ação financeira, entre outras. Em algumas circunstâncias, é possível descrever os fundamentos do processo que explica como a mudança ocorre. Quando as características do processo são governadas pela teoria da probabilidade, se tem um **processo estocástico**.

O primeiro passo para modelar um processo dinâmico é definir o conjunto de estados que pode alcançar e descrever os mecanismos que governam suas transições. Um **estado** é como um *snapshot* (foto instantânea) do sistema em um tempo determinado. É uma abstração da realidade que descreve os atributos de um sistema que interessa. O **tempo** é uma medida linear através da qual o sistema se movimenta, e pode ser visto como um parâmetro. Devido à existência do tempo, existe: passado, presente e futuro. Usualmente se sabe qual foi a trajetória que o sistema tomou para chegar ao estado atual. Usando esta informação, o objetivo é antecipar o futuro comportamento do sistema em termos básicos de um conjunto de atributos. Aqui são mostradas uma série de técnicas teóricas disponíveis para este propósito.

Por razões de modelagem, o estado e o tempo podem ser tratados de forma contínua e discreta. Mas, por razões computacionais e considerações teóricas, o estado somente será considerado em forma discreta. O tempo terá forma contínua ou discreta.

Para obter uma computação tratável, assumir que o processo estocástico satisfaz a propriedade de Markov. Isto é, o caminho que o processo segue no futuro depende só do estado atual e não da seqüência de estados visitados previamente

ao estado atual. Um tempo discreto no sistema induze ao modelo das **Cadeias de Markov**. Para um tempo contínuo no sistema existe um modelo denominado de **Processo de Markov.**

Um modelo de um processo estocástico descreve atividades que terminam em eventos. Os eventos geram a transição de um estado a outro. Assumindo que a duração de uma atividade é uma variável aleatória contínua, eventos ocorrem na continuidade do tempo.

2.5.1 Nomenclatura de um Processo Estocástico

Processo Estocástico

Dada uma variável aleatória, {*X*(*t*)}, onde *t* é um índice de tempo que toma valores de um conjunto dado **T**. **T** pode ser discreto ou continuo. *X*(*t*) é escalar que pode assumir valores discretos ou contínuos. Considera-se somente processos estocásticos discretos finitos.

Tempo

O parâmetro de um processo estocástico.

Estado

Um vetor que descreve atributos de um sistema em um tempo qualquer. O vetor estado tem *m* componentes. $\mathbf{S} = (S_1, S_2 \ldots S_m)$. *X*(*t*) descreve algum atributo do estado.

Conjunto de Estados

Coleção de todos os estados possíveis.

Atividade

Uma atividade começa em um tempo determinado, tem uma duração e termina em um evento. Geralmente a duração de uma atividade é uma variável aleatória com uma distribuição de probabilidade conhecida.

Evento

É a finalização de uma atividade. Um evento tem o potencial de mudar o estado do processo.

Calendário

O conjunto de eventos que podem ocorrer em um estado dado, *Y*(s)

Próximo evento

Num estado qualquer, um ou mais eventos podem ocorrer. O próximo que ocorre é chamado de próximo evento. Começando pelo tempo atual, o tempo do próximo evento é o tempo mínimo que em termos matemáticos ficaria assim:

$$t_x = \text{Minimum}\{t_e | e \in \mathbf{Y}(\mathbf{s})\}$$

O próximo evento é o valor de *x* que corresponde ao tempo mínimo. Quando as durações de eventos são variáveis aleatórias, ambos, o próximo evento e o tempo do próximo evento, são variáveis aleatórias.

Transição

A função que determina o próximo estado, **s**', baseado no estado atual **s**, e o evento, x. O número de elementos da função de transição é o mesmo do número de elementos do vetor estado. **s'** = $T(\mathbf{s}, x)$.

Rede de Estados de Transição

Na representação gráfica da figura, os estados são representados por nodos; e eventos, representados por arcos. Uma transição é mostrada em forma de arco ou seja que tem direção e vai de um nodo a outro.

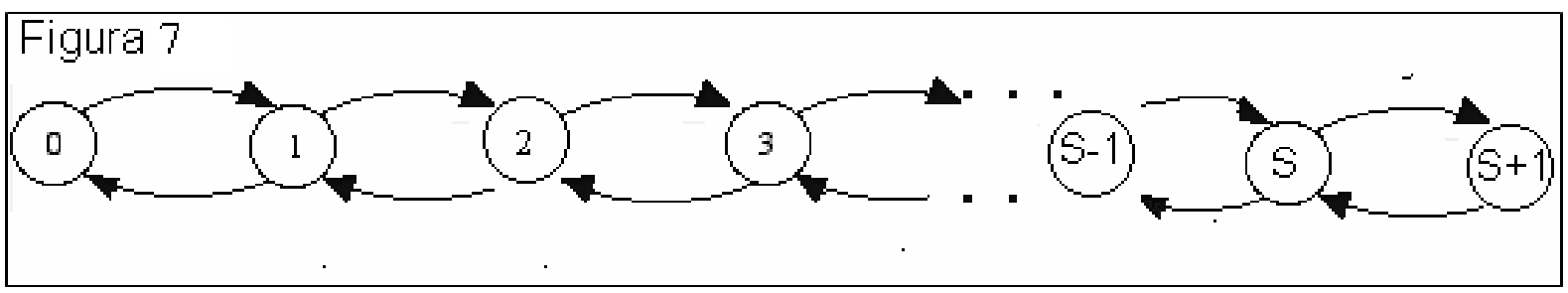


Fig 7: Rede de Transição de Estados

Propriedade de Markov.

Dado o estado atual conhecido, a probabilidade condicional do próximo estado é independente dos estados prévios ao estado atual.

2.5.2 Cadeias de Markov em Tempo Discreto

É um processo estocástico que satisfaz a propriedade de Markov e tem parâmetro de tempo discreto. Algumas vezes este processo é chamado simplesmente de *Cadeias de Markov*.

Se um sistema se movimenta de um estado i durante um período a um estado j durante o seguinte período, uma transição de i a j ocorreu. As

probabilidades $P_{i,,j}$ são referidas como **probabilidade de transições** das cadeias de Markov.

$$P(x_{t+1\,=\,j} \mid x_{t=I}) = P_{i,j} \qquad (1)$$

Significa que t pode ser um segundo, uma hora, ou um dia etc.

A equação 1 implica que a regra de probabilidade relacionada com o estado do próximo período ao estado atual não muda com o tempo. Por esta razão a equação 1 é chamada de **suposição estacionária**. Qualquer cadeia de Markov que satisfaz 1 é chamada da **cadeia de Markov estacionária**.

$P(x_0 = i) = q_i$, onde q_i é a probabilidade que a cadeia esteja no estado i no tempo 0. Chamar o vetor $q = [q_1\ q_2 \ldots\ldots q_s]$ de **distribuição de probabilidade inicial** da cadeia de Markov. As probabilidades de transições são mostradas como uma S_xS **matriz de probabilidades de transições** P.

Dado que o estado no tempo t é i, o processo deve estar em algum lugar no tempo t+1. Isto significa que para cada estado i.

$$\sum_{j=1}^{j=s} P(x_{t+1} = j \mid P(x_t = i)) = 1 \qquad \sum_{j=1}^{j=s} P_{i,j} = 1$$

Cada entrada na matriz P deve ser positiva e as entradas de cada fila na matriz devem somar 1. P pode ser escrita como:

$$P = \begin{bmatrix} P_{11} & P_{12} \cdots\cdots P_{1S} \\ P_{21} & P_{22} \cdots\cdots P_{2S} \\ \vdots & \vdots\cdots\cdots\vdots \\ \vdots & \vdots\cdots\cdots\vdots \\ P_{S1} & P_{S2} \cdots\cdots P_{SS} \end{bmatrix}$$

2.5.3 Cadeias de Markov em Tempo Contínuo

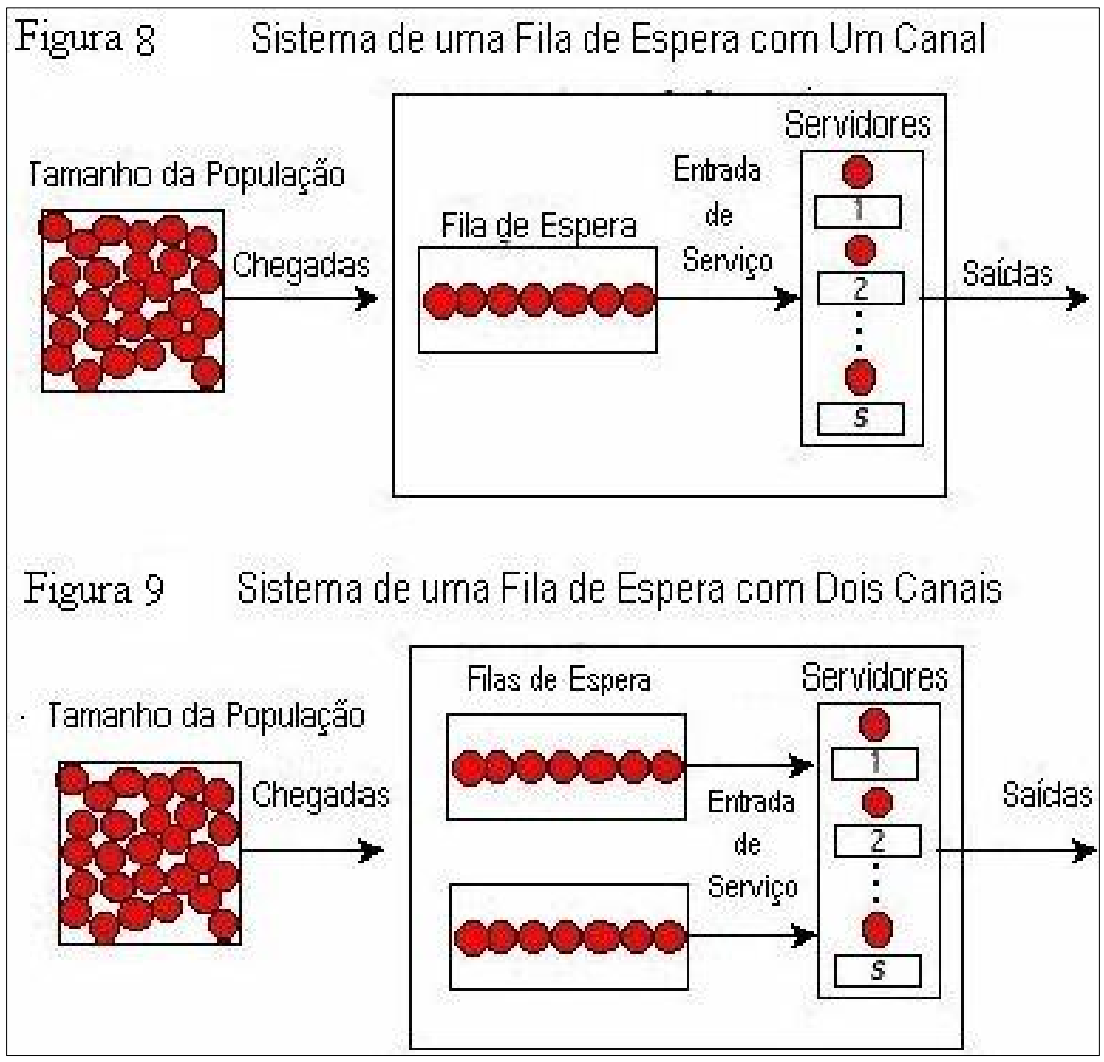
É um processo estocástico que satisfaz a propriedade de Markov e tem um parâmetro de tempo contínuo. Algumas vezes este processo é chamado *Processo de Markov*.

2.6 DESCRIÇÃO DE UM SISTEMA DE FILAS

2.6.1 Características de um Sistema de Filas

A teoria das filas é uma aplicação do processo estocástico de tempo contínuo chamado também de Processo de Markov.

As figuras 8 e 9 mostram os componentes básicos de um sistema de filas. Uma com duas filas e a outra com uma. Os clientes potenciais e atuais são representados por círculos pequenos e podem ser pessoas, máquinas, partes ou qualquer outro ente. Os atendentes são representados por retângulos numerados e podem ser qualquer tipo de fonte, tais como, pessoas, máquinas, oficina de reparos, que executam uma função.



Figuras 8 e 9: Sistema de Filas

Os clientes que chegam ao sistema entram em atendimento imediatamente se algum dos atendentes está ocioso. Se todos os atendentes estão ocupados, o cliente espera na fila até que um atendente esteja livre. Depois de um período finito de tempo, o cliente sai do sistema. Os detalhes do processo dependem dos valores dos parâmetros e suposições adotadas pelos componentes do sistema.

A **fonte de *input***, também conhecida como população de chamada, é um grupo de clientes potenciais que podem precisar dos serviços oferecidos pelo sistema. A fonte de input está caracterizada por seu tamanho N, que geralmente é assumido infinito por motivos de modelagem e a distribuição de probabilidade, descrevendo os tempos de chegada.

A **fila de espera** é o número de clientes esperando ser atendidos, e podem estar concentrados num lugar fixo como num banco ou podem estar

distribuídos no tempo e espaço como aviões preparados para aterrissar. A disciplina da fila de espera define as regras pelas quais os clientes são selecionados para atendimento.

O mecanismo de **atendimento** é o processo pelo qual os clientes são atendidos. A suposição geral é que o atendimento é providenciado por um ou mais atendentes idênticos operando em paralelo. No caso de ter uma rede de filas de espera, várias configurações serão consideradas. As características do atendimento são o número de atendentes S, e a distribuição de probabilidade do tempo de atendimento.

Um **sistema de filas** é a combinação das filas de espera e os atendentes. O número de clientes no sistema é a primeira medida para analisar o sistema de filas de espera. Seu número representa o **estado do sistema**.

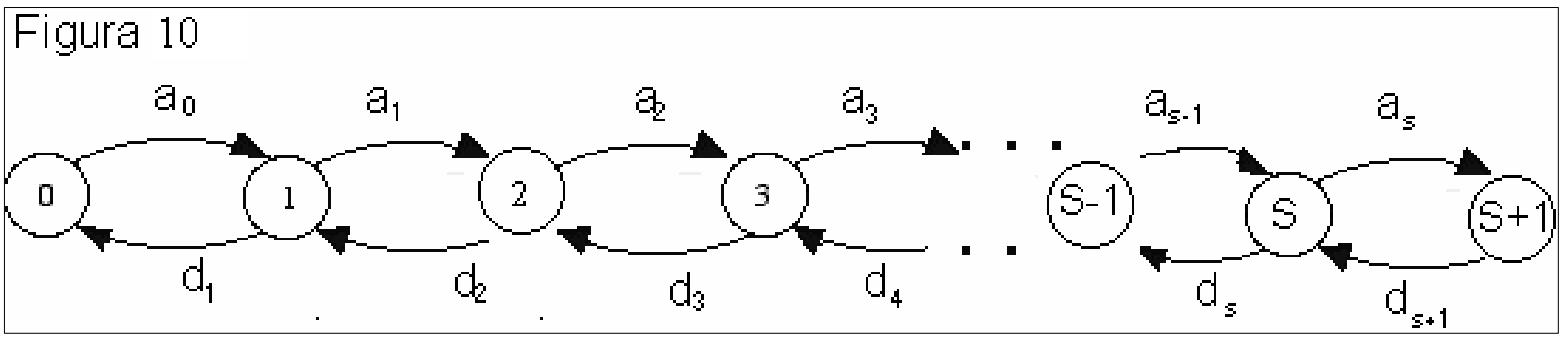


Fig 10: Rede de Transição de Estados de um Sistema de Filas

Na figura 10, os estados estão representados por pequenos círculos com um número indicando a quantidade de clientes nesse estado. O estado zero é um estado vazio quando não existem clientes e todos os atendentes estão ociosos. Nos estados de 1 até S, todos os clientes estão sendo atendidos e ninguém está na fila de espera. Para estados maiores que S, todos os atendentes estão ocupados e alguns clientes estão na fila de espera (k-s). Os arcos ou setas representam **eventos**. Uma chegada denominada pela letra *a*, causa ao sistema aumentar em um; enquanto que uma saída denominada por *d*, causa o número do sistema decrescer em um. Usa-se subscritos nestas denominações para indicar que o processo associado pode depender do estado do sistema.

Posto que ambos os tempos de chegada e atendimento sejam variáveis randômicas, **o estado do sistema é um processo estocástico**. Para sistemas estáveis, existem probabilidades em estados estáveis do número de clientes no sistema. Chamar a probabilidade em estado estável de n clientes π_n. As probabilidades em estado estável têm dois significados: Probabilidade de encontrar

o sistema no estado *n* num tempo determinado selecionado aleatoriamente, ou a quantidade determinada de tempo em que o sistema está no estado *n*.

A teoria das filas envolve fórmulas para calcular as probabilidades em estado estável de diferentes configurações do sistema de filas. A maioria delas requer que os tempos de chegada e os tempos de atendimento sejam governados pela distribuição exponencial de probabilidade. Resultados aproximados existem quando as distribuições não são exponenciais. Dadas as probabilidades em estado estável, se calcula uma variedade de variáveis que interessam ao desenhista ou operador de um sistema de filas. Estas englobam o valor esperado do número de clientes no sistema, o valor esperado do tempo que um cliente fica no sistema, a eficiência dos atendentes etc. Os termos *average* são sinônimos com os termos estatísticos *média* ou *valor esperado*. O valor esperado de número e tempo também podem ser calculados para a fila de espera e para o atendimento.

2.6.1.1 **Nomenclatura de um Sistema de Filas**

Estado

O número total de clientes no sistema. Ou seja, o número de clientes na fila de espera mais o número de clientes que estão sendo atendidos.

K = O número máximo de pessoas no sistema. Quando o número máximo é finito e este número é alcançado, a chegada do próximo cliente não entra na fila de espera. Ele desiste de entrar.

π_n = Probabilidade de *n* clientes no sistema quando o estado do sistema é estável.

L = O valor esperado do número de clientes no sistema quando o sistema é estável.

W = O valor esperado do tempo dos clientes no sistema quando o sistema é estável. (Cf. Jensen, 2006, P.84).

Um sistema de filas é descrito basicamente por três características:

- Processo de chegada
- Disciplina da fila: SIRO (atendimento em ordem randômica), FCFS (o primeiro a chegar é o primeiro a ser atendido), LCFS (o último a chegar é o primeiro a ser atendido pelo servidor) etc.

- Processo de atendimento

2.6.2 **Características das Formas de Chegadas**

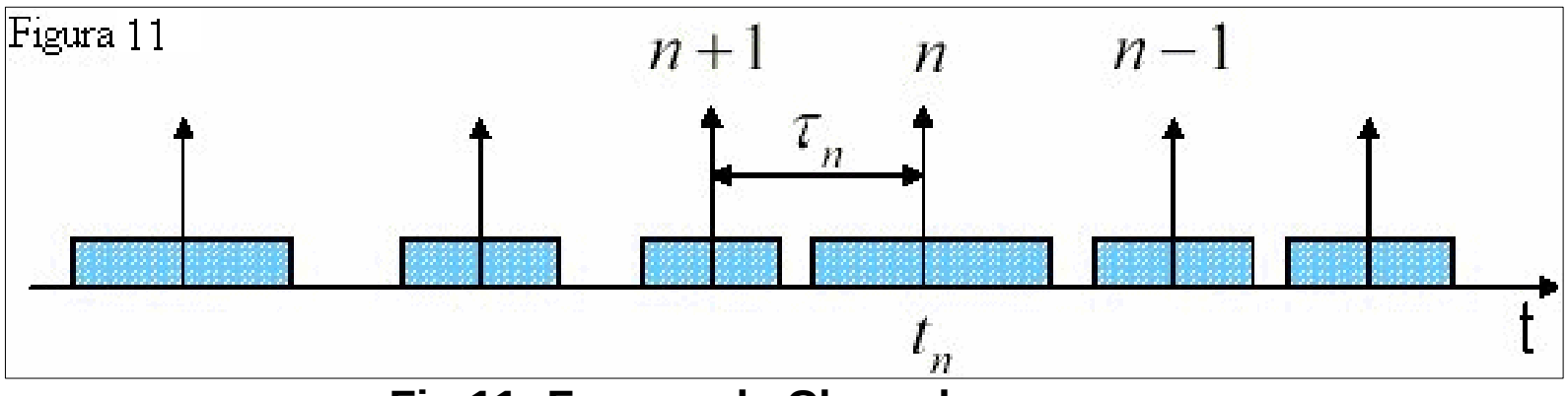


Fig 11: Formas de Chegadas

A figura 11mostra:

τ_n = Tempo de chegada entre o cliente *n* e o cliente *n*+1 é uma variável aleatória

(τ_n, n ≥ 1) é um processo estocástico.

Os tempos de chegada têm distribuições idênticas e têm o mesmo valor esperado.

O valor esperado é o seguinte: $E(\tau_n) = E(\tau) = \frac{1}{\lambda}$

λ é chamada a razão de chegada. (Cf.Korilis, 2001, P.84)

2.6.2.1 **Nomenclatura das Formas de Chegadas**

N = O tamanho da população.

λ_n= O valor esperado do número de chegadas por unidade de tempo quando *n* clientes estão no sistema.

λ = A razão de chegada quando o estado do sistema não afeta a razão de chegada dos clientes. O valor esperado do tempo entre chegadas é $\frac{1}{\lambda}$

$\overline{\lambda}$ = O valor esperado do número de chegadas por unidade de tempo quando o estado do sistema afeta a razão de chegada.

A forma de chegada está definida pela probabilidade de distribuição do tempo entre sucessivos eventos de chegada. Os dois extremos de possibilidades de formas de chegada podem ser: chegadas aleatórias ou chegadas predeterminadas. As formas de chegadas podem depender das formas de atendimento.

2.6.3 **Características de uma Fila de Espera**

2.6.3.1 **Disciplina da fila**

Regra na qual clientes são selecionados da fila de espera para receber atendimento. Elas podem ser: FCFS (primeiro em chegar primeiro em ser atendido, LIFO (último em chegar primeiro a ser atendido), e prioridade de atendimento, obedecendo a uma regra predeterminada. É assumido que os clientes formam somente uma fila ainda que existam vários atendentes.

2.6.3.2 **Nomenclatura de uma Fila de Espera**

K – *s* = O número máximo no sistema menos o número de atendentes no sistema é igual ao número máximo de clientes numa fila de espera.

L_q = O valor esperado do número de clientes na fila de espera num sistema estável.

W_q = O valor esperado do tempo dos clientes na fila de espera num sistema estável. (Cf. Jensen, 2006, P.84).

2.6.4 **Características das Formas de Atendimento**

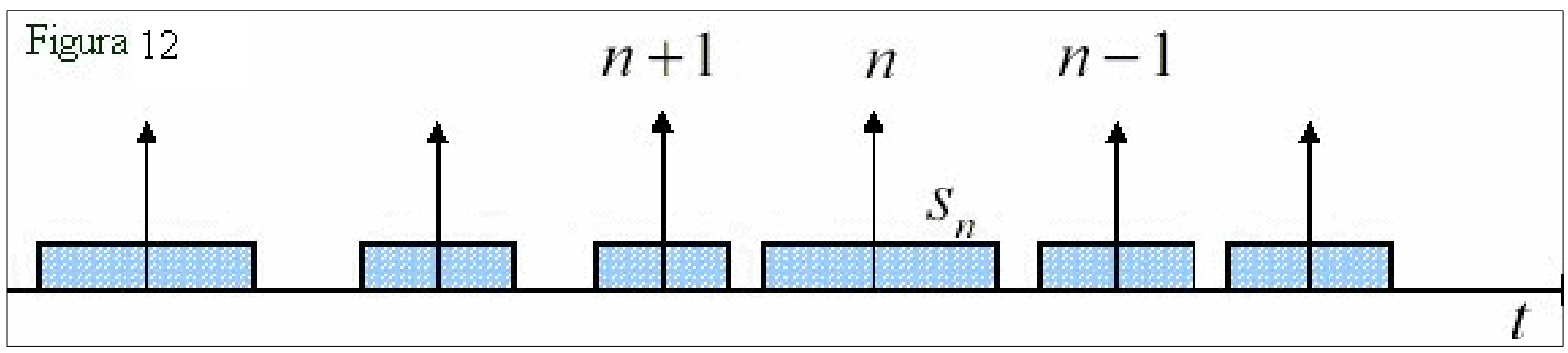


Fig 12: Formas de Atendimento

A figura 12 mostra:

S_n = Tempo de atendimento ao cliente *n* no servidor S

(S_n , n ≥ 1) é um processo estocástico

Os tempos de atendimento têm distribuições idênticas e têm o mesmo valor esperado $E(S_n) = E(S) = 1/\mu$

μ é conhecido como a razão de atendimento. (Cf.Korilis, 2001, P.84)

2.6.4.1 **Nomenclatura das Formas de Atendimento**

S = O número de canais de atendimento. Todos são supostamente idênticos.

L_s = O valor esperado do número de clientes em atendimento num sistema estável.

W_s = O valor esperado do tempo de atendimento num sistema estável.

μ_n = O valor esperado da razão de atendimento no sistema quando n clientes estão presentes.

μ = O valor esperado do número de clientes atendidos por unidade de tempo (razão de atendimento) quando o estado do sistema não afeta a razão de atendimento. O valor esperado do tempo em completar um atendimento é $\frac{1}{\mu}$.

ρ = Intensidade do tráfego . É o valor entre a razão de chegada que os clientes tentam implantar no sistema e a máxima razão de atendimento dos atendentes no sistema.

E = Eficiência ou utilização. A razão entre o valor esperado do número de clientes em atendimento e o número de atendentes. (Cf. Jensen, 2006, P.84).

2.7 O PROCESSO DE NASCIMENTO E MORTE

O processo de nascimento e morte é um processo estocástico de tempo contínuo no qual o estado do sistema em um tempo qualquer é um inteiro positivo. Se o processo de nascimento e morte está num estado j num tempo t, o movimento do processo está governado pelas seguintes regras:

1. A probabilidade de uma chegada (nascimento) é $P_{j, j+1}(\Delta t) = \lambda_j \Delta t + o(\Delta t)$
2. A probabilidade de um atendimento (morte) é $P_{j, j-1}(\Delta t) = \mu_j \Delta t + o(\Delta t)$
3. Nascimentos (chegadas) e mortes (atendimentos completos) são independentes.

As regras 1-3 podem ser usadas para mostrar que a probabilidade de mais de um evento (nascimento ou morte) acontecer entre t e t+Δt é dada por $o(\Delta t)$. Onde $o(\Delta t)$ é uma quantidade qualquer que satisfaz o seguinte: $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$

Isto permite concluir que a probabilidade de acontecer mais de um evento num tempo (t + Δt) é zero.

2.7.1 M/M/1/FCFS/∞/∞ como um processo de nascimento e morte

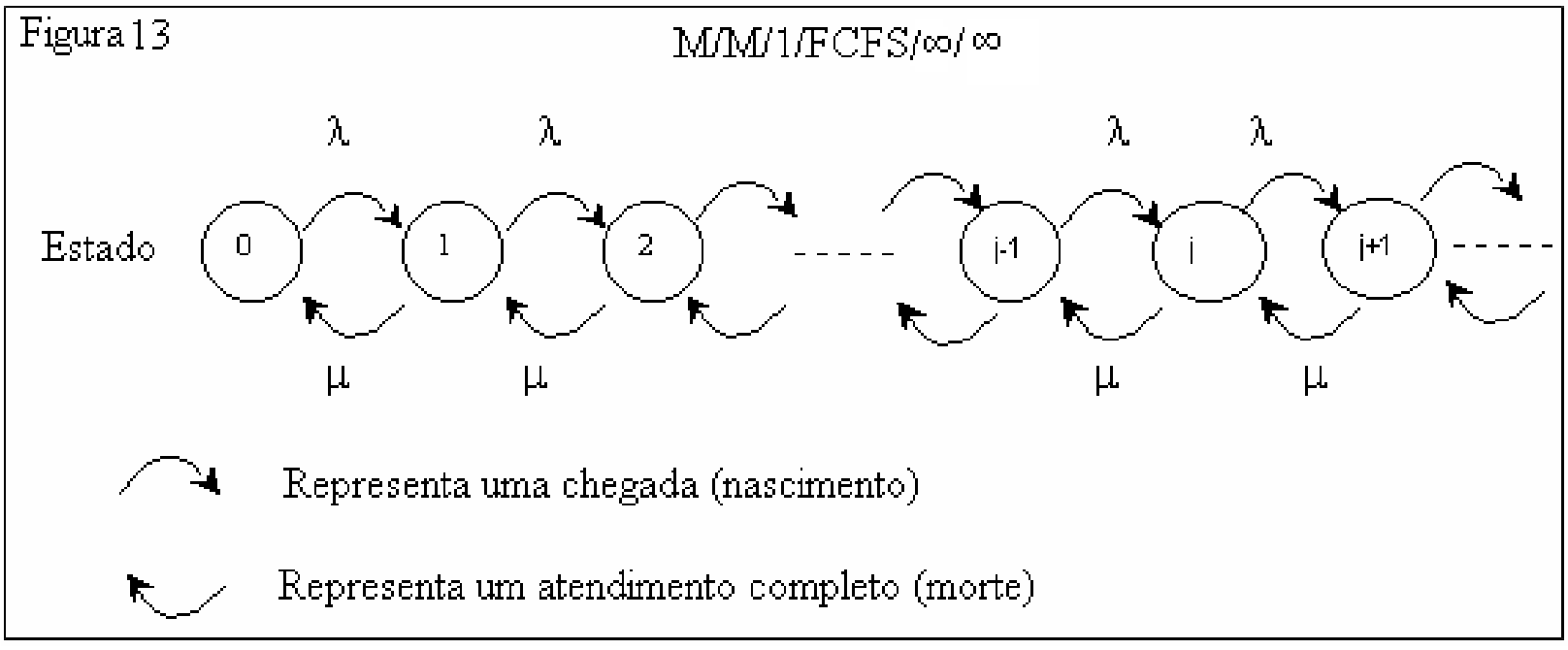


Fig 13: Rede de Transição de Estados do Modelo M/M/1

A figura 13 pode ser modelada como um processo de nascimento e morte. Como o tempo de chegada e atendimento é descrito pela distribuição exponencial, então é aplicável a propriedade da não-memória, ou seja, os intervalos $(t + \Delta t)$ e Δt são similares no sentido de que a $P(t<x<\Delta t) = P(0<x<\Delta t)$. Então:

- A probabilidade de uma chegada no intervalo $(t + \Delta t)$ é:

$$P_{j,\,j+1}(\Delta t) = \int_0^{\Delta t} \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda \Delta t} = \lambda \Delta t + o(\Delta t).$$

Demonstração

$\int e^{f(x)} f'(x)\, dx = e^{f(x)}$, logo

$$\int_0^{\Delta t} \lambda e^{-\lambda t}\, dt = -\int_0^{\Delta t} -\lambda e^{-\lambda t}\, dt = -e^{-\lambda t}\Big|_0^{\Delta t} = -e^{-\lambda \Delta t} + 1$$

também as series de Taylor dizem: $e^{-\lambda \Delta t} = 1 - \lambda \Delta t + o(\Delta t)$, logo

$$\int_0^{\Delta t} \lambda e^{-\lambda t}\, dt = 1-[1-\lambda \Delta t + o(\Delta t)] = 1-1+ \lambda \Delta t + o(\Delta t) = \lambda \Delta t + o(\Delta t)$$

Por tanto, $P_{j,\,j+1}(\Delta t) = \lambda \Delta t + o(\Delta t)$.

A probabilidade de um atendimento é: $P_{j,\,j-1}(\Delta t) = \int_0^{\Delta t} \mu e^{-\mu t}\, dt = 1 - e^{\mu \Delta t} = \mu \Delta t + o(\Delta t)$

Demonstração.

Proceder da forma anterior, somente mudar λ por μ. Portanto

$P_{j,\,j-1}(\Delta t) = \mu \Delta t + o(\Delta t)$.

2.7.2 **Probabilidades em estado estável do processo nascimento-morte**

Probabilidades em estado estável (π_j) ou a fração de tempo que o sistema fica num estado (j) no processo de nascimento-morte é dada pelas seguintes possibilidades descritas na seguinte tabela.

Estado no Tempo t	Estado no Tempo (t + Δt)	Probabilidade da Seqüência dos Eventos
j –1	j	$(P_{i,j-1}(t))\ (\lambda_{j-1}\ \Delta t + o\ (\Delta t))$ = I
j+1	j	$(P_{i,j+1}(t))\ (\mu_{j+1}\ \Delta t + o\ (\Delta t))$ = II
J	j	$(P_{i,j}(t))\ (1 - \mu_j\ \Delta t - \lambda_j\ \Delta t - 2\ o\ (\Delta t))$ = III
Qualquer outro estado	j	$o\ (\Delta t)$ = IV

Quadro 1: Probabilidades do Processo Nascimento-Morte

A figura 14 mostra a primeira possibilidade das quatro possíveis.

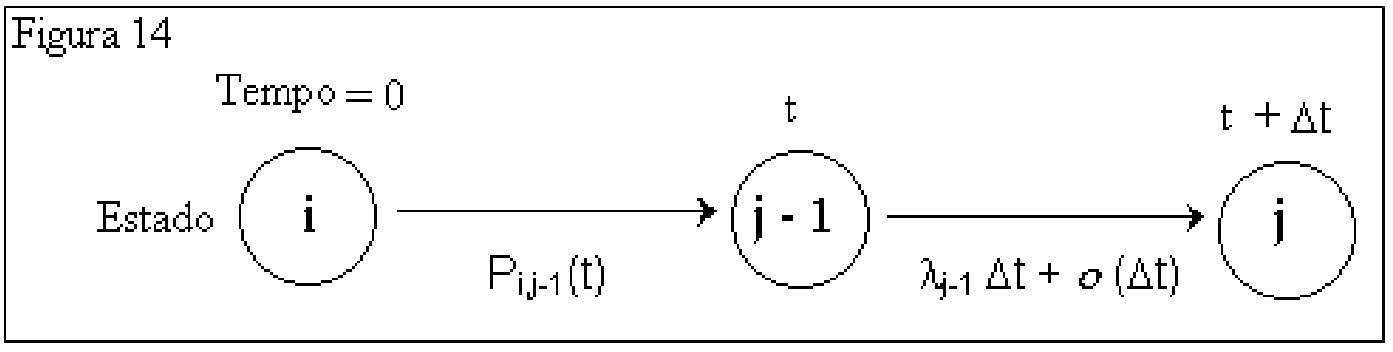


Fig 14: Primeira Possibilidade de Seqüência dos Eventos

Observar que $P_{i,j}(t + \Delta t)$ = I + II + III + IV

Substituindo,

$$P_{i,j}(t + \Delta t) = P_{i,j}(t) + \Delta t\ [\lambda_{j-1}\ P_{i,j-1}(t) + \mu_{j+1}\ P_{i,j+1}(t) - P_{i,j}(t)\ \mu_j\ - P_{i,j}(t)\ \lambda_j\ ] +$$

$$\underline{o\ (\Delta t)\ [P_{i,j-1}(t)\ + P_{i,j+1}(t) + 1 - 2\ P_{i,j}(t)\ ]}$$

Logo, o que está sublinhado pode ser escrito somente como o (Δt) porque o fator o (Δt) é zero. Reagrupando os termos:

$$P_{i,j}(t + \Delta t) - P_{i,j}(t) = \Delta t\ [\lambda_{j-1}\ P_{i,j-1}(t) + \mu_{j+1}\ P_{i,j+1}(t) - P_{i,j}(t)\ \mu_j\ - P_{i,j}(t)\ \lambda_j\ ] + o\ (\Delta t)$$

Para todo estado i e j $\geq$ 1, dividir ambos os lados por (Δt), ficando:

$$\frac{P_{i,j}(t+\Delta t)-\mathrm{Pi,\,j(t)}}{\Delta t}=\lambda_{j-1}P_{i,j-1}(t)+\mu_{j+1}P_{i,j+1}(t)-P_{i,j}(t)\mu_j-P_{i,j}(t)\lambda_j+\frac{o\Delta t}{\Delta t}.$$

Logo aproximar (Δt) a zero fica:

$$\lim_{\Delta t\to 0}\left(\frac{P_{i,j}(t+\Delta t)-\mathrm{Pi,\,j(t)}}{\Delta t}\right)=\lim_{\Delta t\to 0}\left(\lambda_{j-1}P_{i,j-1}(t)+\mu_{j+1}P_{i,j+1}(t)-P_{i,j}(t)\mu_j-P_{i,j}(t)\lambda_j+\frac{o\Delta t}{\Delta t}\right)$$

$$\text{Por definição de derivada}: \mathrm{P}'_{i,j}(t)=\lim_{\Delta t\to 0}\left(\frac{P_{i,j}(t+\Delta t)-\mathrm{Pi,\,j(t)}}{\Delta t}\right)$$

e $\lim_{\Delta t\to 0}\frac{o\Delta t}{\Delta t}=0$. Substituindo: $P'_{i,j}(t) = \lambda_{j-1}P_{i,j-1}(t) + \mu_{j+1}P_{i,j+1}(t) - P_{i,j}(t)\ \mu_j - P_{i,j}(t)\ \lambda_j$

Quando o sistema alcança um estado estável o $\lim_{T\to\infty}P_{i,j}(t)=\pi_j$. Em estado estacionário, π_j é uma constante e o estado inicial é irrelevante. A derivada de uma constante é zero. Então a equação fica assim:

$$0=\lambda_{j-1}\pi_{j-1}+\mu_{j+1}\pi_{j+1}-\pi_j\mu_j-\pi_j\lambda_j$$

$$\pi_j\left(\mu_j+\lambda_j\right)=\lambda_{j-1}\pi_{j-1}+\mu_{j+1}\pi_{j+1}\qquad\left(j=1,2,3,\cdots\cdots\right)\qquad(1)$$

Para j = 0: $\mu_1\pi_1=\pi_0\lambda_0\qquad\left(j=0\right)$, Não existem estados negativos. $\mu_0=0$

A equação (1) para j=1, 2, 3, ... significa que em qualquer instante (t>0) no qual se observa o modelo do processo de nascimento e morte, acontece que para cada estado j: o número de entradas ao estado (j) e o número de saídas do estado (j) diferem no máximo em um. Por exemplo, ao entrar em média cinco vezes ao estado j, logo o número de vezes que saímos do estado j em média será quatro ou cinco ou seis vezes.

Logo para um longo período de tempo t, com j = 1, 2, 3, ...(e para qualquer condição inicial), será verdade que:

$$\frac{\text{o número esperado de saídas do estado j}}{\text{unidade de tempo}}=\frac{\text{o número esperado de entradas ao estado j}}{\text{unidade de tempo}}\qquad(2)$$

Assumir que o sistema atinge um estado estável, e que o sistema espera uma fração π_j de tempo no estado j. Agora usar a equação (2) para determinar as probabilidades π_j em estado estável. Para j $\geq$ 1, ao sair do estado j somente se chega aos estados: j +1 ou j –1. Então, para j $\geq$ 1:

$$\pi_j\left(\mu_j+\lambda_j\right)=\lambda_{j-1}\pi_{j-1}+\mu_{j+1}\pi_{j+1}\qquad\left(j=1,2,3,\cdots\cdots\right)$$

Esta equação é chamada de equação de movimento de equilíbrio para um processo de nascimento e morte.

$$\text{Para } j=0, \pi_1 = \frac{\pi_0 \lambda_0}{\mu_1}; \quad \text{Para } j=1, \pi_0\lambda_0 + \pi_2\mu_2 = \pi_1(\mu_1 + \lambda_1). \text{ Logo :}$$

$$\pi_0\lambda_0 + \pi_2\mu_2 = \frac{\pi_0\lambda_0}{\mu_1}(\mu_1 + \lambda_1) \rightarrow \pi_2 = \frac{\pi_0(\lambda_0\lambda_1)}{\mu_1\mu_2}; \text{ Define - se } C_j = \frac{\lambda_0\lambda_1\lambda_2\lambda_{.3}.....\lambda_{j-1}}{\mu_1\mu_2\mu_3.....\mu_j},$$

$$\pi_j = \pi_0 c_j \qquad \text{Onde } \sum_{j=0}^{j=\infty}\pi_j = 1, \text{então } \sum_{j=0}^{j=\infty}\pi_0 c_j = 1, \text{ ou seja : } \pi_0 C_0 + \sum_{j=1}^{j=\infty}\pi_0 c_j = 1$$

$$\pi_0 = \frac{1}{C_0 + \sum_{j=1}^{j=\infty} C_j}, \text{logo, } \pi_0 = \pi_0 C_0, \text{então, finalmente, } \pi_0 = \frac{1}{1 + \sum_{j=1}^{j=\infty} C_j}$$

3. DESEMPENHO DOS MODELOS

Existem os seguintes modelos: Markovianos, Não-Markovianos e em Série. Os modelos em Série são: Modelos de Rede Aberta e Modelos de Rede Fechada

3.1 O teorema de Little

A fórmula de Little é uma analogia com a fórmula de Física Fundamental. Isto é distância = velocidade * tempo. Ou seja, os princípios físicos são aplicados no sistema de filas de espera.

Número médio no sistema de filas = razão de chegada * tempo médio no sistema

$L = \lambda * W$

Número médio na fila de espera = razão de chegada * tempo médio na fila

$L_q = \lambda * W_q$

Número médio de clientes em atendimento = razão de chegada * tempo médio de atendimento

$L_s = \lambda * W_s$

3.2 Modelos Markovianos

Quando os tempos de chegada e atendimento têm probabilidade com distribuição exponencial, o sistema de filas é um sistema de filas Markoviano. A distribuição exponencial satisfaz a hipótese de Markovian na qual ela não tem memória. Ou seja, ao esperar a ocorrência de uma chegada com o tempo distribuído exponencialmente, o valor esperado do tempo de chegada não depende

de quanto tempo já transcorreu, esperando essa chegada. Esta situação mostra-se peculiar, mas existem situações em que esta hipótese é valida. E a suposição da distribuição exponencial é necessária para obter soluções aproximadas para variáveis estatísticas. Com a hipótese de distribuições exponenciais, os processos de chegada e atendimento são ***processos de Poisson***.

Todos os modelos markovianos podem ser analisados como um processo de nascimento e morte.

3.2.1 O Modelo M/M/1/GD/∞/∞

A intensidade de Tráfego é dada por: $\rho = \frac{\lambda}{\mu}$, $0 \leq \rho < 1$ para que o sistema seja estável

Notar que: $\lambda_j = \lambda$ (j = 0, 1, 2,.....); $\mu_0 = 0$; $\mu_j = \mu$ (j = 1,2,.....)

$\pi_1 = \frac{\pi_0 \lambda_0}{\mu_1}$, $\pi_2 = \frac{\pi_0 (\lambda_0 \lambda_1)}{\mu_1 \mu_2}$,....., $\pi_j = \pi_0 \frac{\lambda^j}{\mu^j}$, $\pi_j = \pi_0 \rho^j$ (3)

Pela lei da probabilidade:

$\pi_0 + \pi_1 + \pi_2 + \ldots\ldots\ldots = 1$ → Substituindo: $\pi_0 + \pi_0 \frac{\lambda}{\mu} + \pi_0 \frac{\lambda^2}{\mu^2} + \ldots\ldots = 1$

$\pi_0 \ (1 + \rho + \rho^2 + \ldots.) = 1$

Definindo $s = 1 + \rho + \rho^2 + \ldots.$ logo $\rho s = \rho + \rho^2 + \ldots.$

$s - \rho s = 1 \rightarrow s = \frac{1}{1 - \rho}$

$\pi_0 s = 1 \rightarrow \pi_0 = 1 - \rho$ para $0 \leq \rho < 1$, logo na equação 3, substituindo, fica:

$\pi_j = (1 - \rho)\, \rho^j$ para $0 \leq \rho < 1$ (4)

3.2.1.1 Média do número de clientes no sistema (L)

Assumindo que um sistema alcança o estado estável, $0 \leq \rho < 1$. O número médio de clientes presentes num sistema de filas (L) é dado por $L = \sum_{j=0}^{j=\infty} j\pi_j$. Logo

$L = \sum_{j=0}^{j=\infty} j(1-\rho)\rho^j \Rightarrow (1-\rho)\sum_{j=0}^{j=\infty} j\rho^j$

Definindo S' = $\sum_{j=0}^{j=\infty} j\rho^j$ = ρ + 2 ρ^2+3 ρ^3+ ....., logo

ρ S'= ρ^2 + 2 ρ^3 + 3 ρ^4 +.....

s'- ρs'= ρ + ρ^2+ ρ^3+.... = ρs → s'(1- ρ) = ρs → s'(1- ρ) = ρ $\frac{1}{1-\rho}$

s'= $\frac{\rho}{(1-\rho)^2}$, então L = (1- ρ) S' → L = $\frac{\rho}{1-\rho}$

L = $\frac{\lambda}{\mu-\lambda}$ (5)

3.2.1.2 **Média do número de clientes na fila de espera (L_q)**

O número médio de clientes na fila de espera é:

$L_q = \sum_{j=1}^{j=\infty}(j-1)\pi_j = \sum_{j=1}^{j=\infty} j\pi_j - \sum_{j=1}^{j=\infty}\pi_j$ = L – (1-π_0)

L_q = L – ρ = $\frac{\rho}{1-\rho}$ - p = $\frac{\rho^2}{1-\rho}$

L_q = $\frac{\lambda^2}{\mu(\mu-\lambda)}$ (6)

3.2.1.3 **Média do número de clientes em atendimento (L_s)**

O número médio de clientes em atendimento é: $\sum_{j=0}^{j=\infty}\pi_j = 1$. Logo

$L_s = \sum_{j=0,i=0}^{j=\infty,i=1} i\pi_j$ = 0 π_0 + 1 (π_1 + π_2 + π_3 +.....)=0+1(1- π_0) = 1 – (1-ρ) = ρ (7)

3.2.1.4 **Média do tempo total no sistema (W)**

O tempo médio de um cliente no sistema de filas é:

W = $\frac{L}{\lambda} = \frac{\rho}{(1-\rho)\lambda} = \frac{1}{\mu-\lambda}$ (8)

3.2.1.5 **Média do tempo na fila de espera(W_q)**

O tempo médio de um cliente na fila de espera é:

$W_q = \frac{Lq}{\lambda} = \frac{\lambda^2}{\mu(\mu-\lambda)\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$ (9)

3.2.1.6 **Média do tempo de atendimento (W_s)**

$W_s = \frac{1}{\mu}$. (Cf. Winston, 2004, p.85)

3.2.1.7 **Probabilidade de existir no mínimo N clientes no sistema**

$P(n \geq N) = \rho^N$

3.2.1.8 **Tempo total dos períodos de atendimento**

$$E(t) = \frac{\rho t}{\lambda t(1-\rho)} \quad E(t) = \frac{1}{\mu - \lambda}$$

O numerador ρt da primeira fração é a duração total de todos os intervalos não ociosos. Pois $\rho = 1-P_0$ indica os atendentes ocupados que multiplicando por (t) se teria o tempo total de atendimento. O denominador $\lambda t\ (1-\rho)$ resulta da relação $\frac{t(1-\rho)}{\frac{1}{\lambda}}$, onde $t(1-\rho) = (t)(P_0)$ é o tempo total sem atendimento. O valor $\frac{1}{\lambda}$ é o valor esperado do tempo de chegadas descrita pela distribuição exponencial. Neste caso $\frac{1}{\lambda}$ é também a média de tempo dos intervalos sem atendimento (o tempo de chegada e o tempo sem atendimento é o mesmo porque o tempo sem atendimento começa quando uma chegada foi atendida e termina quando existe uma nova chegada). Logo $\lambda t(1-\rho)$ indica o número separado de intervalos sem atendimento, mas como períodos ociosos e ocupados se alternam, então o número de períodos ocupados é o mesmo que o de períodos ociosos. Por tanto, o denominador $\lambda t(1-\rho)$ da primeira fração representa o número de períodos separados ocupados durante o tempo t.

3.2.1.9 **Número de atendimentos em todos os períodos ocupados**

$$E(n) = \mu * \frac{1}{\mu - \lambda} \qquad E(n) = \frac{\mu}{\mu - \lambda}$$

As fórmulas E(t) e E(n) são válidas também para o modelo M/G/1. (Cf. Harvey, 1969, P.84)

Exemplo:

Dez passageiros por minuto que chegam ao aeroporto. Para checar os passageiros com armas, o aeroporto deve ter um ponto de checagem que consiste num detector de metais e uma máquina de raios X. Quando um ponto de checagem está ativo, dois empregados são requeridos. Um ponto pode checar 12 passageiros por minuto (o tempo de checagem do passageiro é exponencial). Assumindo que o aeroporto tem um ponto ativo de checagem, calcular: a) Qual é a probabilidade de que um passageiro deverá esperar antes de ser atendido? b) Em média quantos passageiros estão em fila de espera antes de entrar no ponto de checagem? c) Em média quanto tempo um passageiro demora no ponto de checagem? d) Calcular o tempo total dos períodos de atendimento e) Encontrar o número total de atendimentos nos períodos de atendimento. Logo: $\rho = 10 / 12 = 5/6$. Logo

a)$P(n \geq 1)=\rho^1 = 10/12 = 5/6$. b)$L_q=\rho^2 / (1-\rho) = (5/6)^2 / [1-(5/6)] = 25/36\ (6) = 25/6 \cong 4$ passageiros na fila. c)$W_s = 1/\mu = (1/12)\ 60 = 5$ segundos.

d) $E(t) = \frac{1}{\mu-\lambda} = \frac{1}{12-10} = \frac{1}{2}$ = 30 segundos. e) $\frac{\mu}{\mu-\lambda} = \frac{12}{12-10} = 6$ atendimentos

3.2.2 O Modelo M/M/1/GD/c/∞

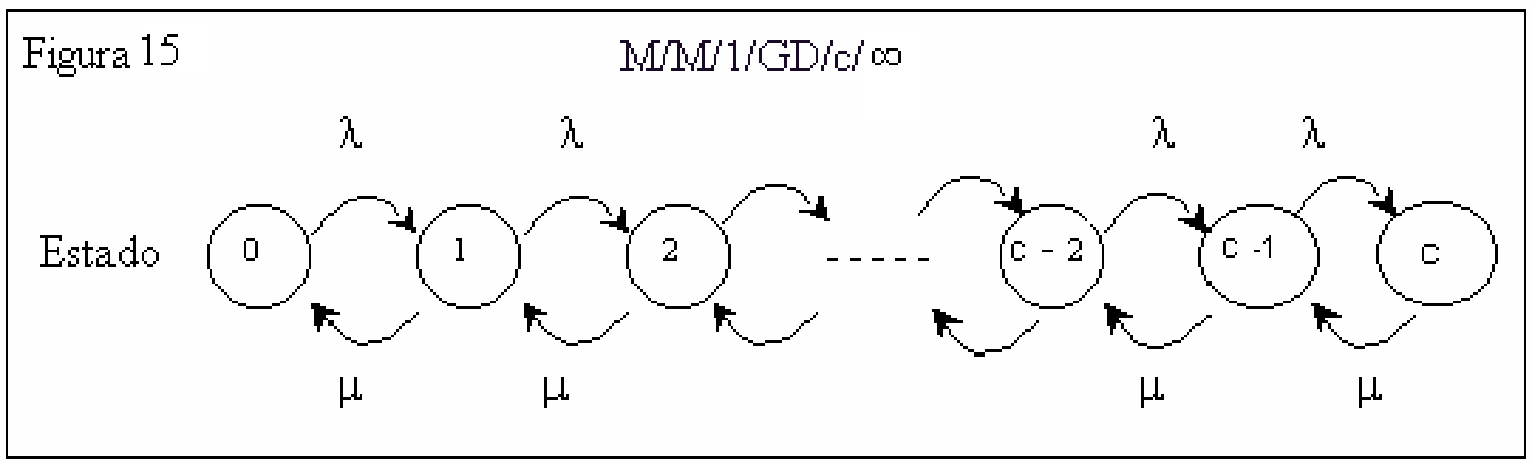


Fig 15: Rede de Transição de Estados do Modelo M/M/1/GD/c/∞

$\lambda_j = \lambda$ para (j = 0, 1, 2,.....,c-1)

$\lambda_c = 0$

$\mu_0 = 0$

$\mu_j = \mu$ para (j =1, 2,....., c)

$$\pi_0 = \frac{1-\rho}{1-\rho^{c+1}} \qquad \pi_j = \rho^j \pi_0 \qquad \text{para (j = 1,2,.....,c)}$$

$\pi_j = 0$ para (j = c+1, c+2,....)

$$L = \sum_{j=0}^{j=c} j\pi_j = \frac{[1-(c+1)\rho^c + c\rho^{c+1}]\rho}{(1-\rho^{c+1})(1-\rho)}$$

$W = \frac{L}{\lambda(1-\pi_c)}$ $W_s = 1/\mu$ $W_q = \frac{L_q}{\lambda(1-\pi_c)}$

Para $\lambda = \mu$ → $\pi_j = \frac{1}{c+1}$ para (j = 0, 1, 2, ...,c)

$L = \frac{c}{2}$ $L_s = 1 - \pi_0$ $L_q = L - L_s$

O estado do sistema continua estável mesmo que $\lambda \geq \mu$, porque a capacidade do sistema é finita, ou seja, quando o estado do sistema alcança seu limite, as próximas chegadas são perdidas para sempre.

Exemplo:

As informações chegam a um roteador de *Internet* numa razão de 125 pacotes por minuto em média. As chegadas acontecem independentemente. O tempo de processamento dos pacotes é de 0.002 segundos por pacote em média. O roteador é desenhado a ter uma capacidade limitada para armazenar mensagens em espera. Qualquer mensagem que chega quando o *buffer* está cheio é perdida. Assumindo que as chegadas e o atendimento têm tempos com distribuição exponencial. a) Qual será o tamanho da memória do roteador (*buffer*) para garantir uma perda máxima de uma mensagem em um milhão de mensagens que chegam ao roteador? b) Encontrar as probabilidades do estado estacionário do sistema.

Como o modelo deste sistema tem capacidade limitada, o problema pede para encontrar o tamanho máximo do sistema, com perda de 0.0001% de mensagens. A probabilidade de alcançar a capacidade máxima do sistema deve ser menor ou igual $1(10)^{-6}$. Ou seja: $\pi_c \leq 1(10)^{-6}$. Em que c é o número máximo de pacotes no sistema.

Logo $\lambda = 2{,}08$ pacotes/sg e $\mu = 500$ pacotes/sg, onde $\rho = \frac{2{,}08}{500} \cong 0{,}0041$

$\pi_j = \rho^j \pi_0$ e $\pi_0 = \frac{1-\rho}{1-\rho^{C+1}}$

Logo, os resultados da seguinte planilha, calculados num *add-in* de filas de espera instalado em Excel, mostram as probabilidades de π_j até que $\pi_j \leq 1(10)^{-6}$. Notar que, $\pi_j \leq 1(10)^{-6}$ quando j ≥ 3. Então, com j = 3: $\pi_3 \leq 1(10)^{-6}$.

Quando a capacidade do roteador aumenta, a probabilidade de existir *overflow* diminui. Quando j=1, existe a probabilidade de perder mais de uma

mensagem num milhão de mensagens. Quando j=2, ainda a perda de mensagens é maior que a razão um em um milhão mas menor que se for j=1. Quando j=3, em média o roteador perde uma mensagem num milhão. Portanto, a capacidade do *buffer* do roteador deve ser de duas mensagens, para que exista uma perda de uma mensagem num milhão. Então, a capacidade no sistema será de c = 3. Ou seja, um pacote está sendo processado e os dois restantes em espera.

	A	B			
1	**Sistema**	Que5	12	Número Médio Atendimento	0.004166
2	Razão de Chegada	2.083333333	13	Tempo Médio Atendimento	0.002
3	Razão de Atendimento	500	14	Razão Média Chegada	2.083333
4	Atendentes	1	15	Eficiência	0.004166
5	N Máximo no Sistema	3	16	Prob Atendentes Ociosos	0.995833
6	População	***	17	Prob Atendentes Ocupados	0.004166
7	Modelo	M/M/1/3	18	Prob Sistema Cheio	7.20E-08
8	Número Médio Sistema	0.004184099	19	Tempo Crítico Espera	1
9	Tempo Médio Sistema	0.002008368	20	P(Espera ≥ Espera Crítica)	0
10	Número Médio na Fila	1.74328E-05	21	P(0)	0.993278
11	Tempo Médio na Fila	8.36777E-06	22	P(1)	0.009305
			23	P(2)	1.72E-05
			24	P(3)	7.20E-08

Quadro 2: Exemplo do Modelo M/M/1/c/∞

A fração π_3 de todas as chegadas encontrarão o sistema cheio, ou seja, em média $\lambda\pi_3$ das mensagens não entrarão no sistema ou serão perdidas. Então, o valor esperado das mensagens que entram no roteador será $\lambda - \lambda\pi_3$.

b) As probabilidades em estado estacionário são mostradas nas filas 21, 22, 23 e 24 respectivamente. Logo, P ($j > 3$) = 0.

Nota: Para gerar esta planilha em Excel, com os resultados já mostrados, deve existir um *add-in* de filas de espera. Este *add-in* é um programa criado na linguagem de programação *Visual Basic* para gerar respostas dos diferentes modelos de um sistema de filas e deve ser instalado em Excel.

3.2.3 O Modelo M/M/S/GD/∞/∞

Este sistema pode ser modelado como um processo de nascimento e morte com parâmetros:

$\lambda_j = \lambda$ (j = 0, 1, 2,.....)

$\mu_j = j\mu$ (j =0,1, 2,....., s)

$\mu_j = s\mu$ (j = S +1, S+2,.....)

Onde S é o número de servidores em atendimento. Definir

$\rho = \frac{\lambda}{\mu * S}$ para $\rho < 1$

$\pi_0 = \frac{1}{\left[\sum_{i=0}^{i=(s-1)} \frac{(s\rho)^i}{i!} \oplus \frac{(s\rho)^s}{S!(1-\rho)}\right]}$, onde π_0 é a probabilidade de zero clientes no sistema

$\pi_j = \frac{(s\rho)^j \pi_0}{j!}$ (j = 1, 2, ...S) $\qquad \pi_j = \frac{(s\rho)^j \pi_0}{s!\, s^{j-s}}$ (j = S, S+1, S+2,.....)

P [j ≥ s] = $\pi_0 \frac{(s\rho)^S}{S!(1-\rho)}$, onde P[j≥s] significa todos os servidores em atendimento.

Se S = 1, logo P [j ≥ 1] = λ/μ como no Modelo M/M/1

$L_q = P[j \geq s] \cdot \frac{\rho}{1-\rho}$, $L = L_q + L_s$ → $L = L_q + \frac{\lambda}{\mu}$, $W_q = \frac{L_q}{\lambda} = \frac{p(j \geq s)}{s\mu - \lambda}$

$W = W_q + W_s$ $\quad W = \frac{p(j \geq s)}{s\mu - \lambda} \oplus \frac{1}{\mu}$

P(tempo de espera W_q > espera crítica t) = $P[j \geq s] * e^{[-s\mu(1-\rho)t]}$

Gráfico

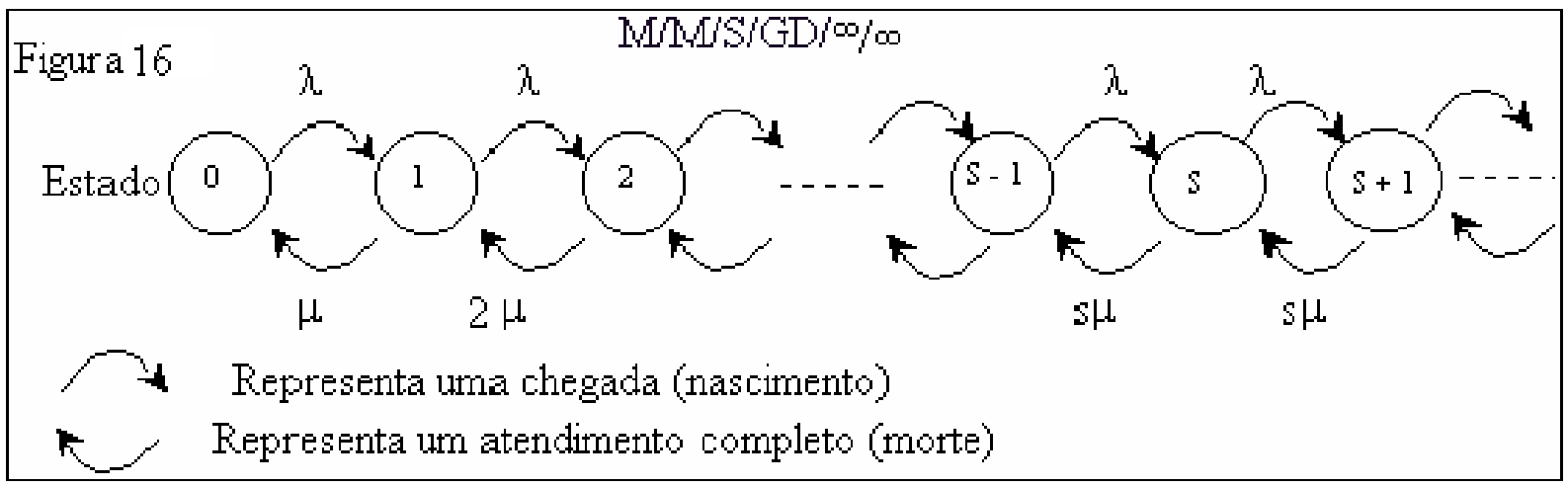


Fig 16: Rede de Transição de Estados do Modelo M/M/S/GD/∞/∞

Exemplo:

Uma estação de conserto de uma fábrica de computadores tem três empregados que consertam computadores com defeitos. O valor esperado do tempo de conserto é de 30 minutos, ou seu equivalente, a razão de conserto é de dois computadores por hora. Devido à ampla variação do número de defeitos, a distribuição do tempo de reparo é aproximada à distribuição exponencial. As

máquinas chegam à estação numa razão de cinco por hora. As chegadas acontecem independentemente, portanto, se justifica que o tempo entre chegadas tem uma distribuição exponencial com valor esperado de 12 minutos. Quantas máquinas em média estão em espera para ser reparadas? Por quanto tempo uma máquina estará em uma oficina de reparo? Com que freqüência os empregados ficam ociosos?

	A	B	16	Probabilidade de Atendentes Ociosos	0.0449
1	**Queue Station**		17	Probabilidade de Atendentes Ocupados	0.7022
2	Razão de Chegada	5	18	Probabilidade do Sistema Cheio	0
3	Razão de Atendimento/Canal	2	19	Tempo de Espera Crítico	1
4	Número de Atendentes	3	20	P(Espera >= Espera Crítica)	0.2583
5	Número no Sistema	***	21	P(0)	0.0449
6	Número na População	***	22	P(1)	0.1121
7	Modelo	M/M/3	23	P(2)	0.1404
8	Número médio no Sistema	6.011235	24	P(3)	0.1170
9	Tempo médio no Sistema	1.202247	25	P(4)	0.0975
10	Número médio na Fila	3.511235	26	P(5)	0.0812
11	Tempo médio na Fila	0.702247	27	P(6)	0.0677
12	Número médio em Atendimento	2.5	28	P(7)	0.0564
13	Tempo médio em Atendimento	0.5	29	P(8)	0.0470
14	Média da Razão de Chegada	5	30	P(9)	0.0391
15	Eficiência	0.833333	31	P(10)	0.0326

Quadro 3: Exemplo do Modelo M/M/S

Então a planilha mostra as respostas nas filas 10, 11 e 16 respectivamente. Ou seja:

a) L_q= 3,5 máquinas em espera

b) W_q = 1,2 horas por máquina no sistema

c) P(0) = 4,5 %. Probabilidade que todos os empregados estejam ociosos

3.2.4 O Modelo M/M/R/GD/K/K de reparo de máquinas

Consiste em K máquinas e R pessoas de reparo. Em todo instante as máquinas estão com defeito ou sem defeito. O intervalo de tempo em que uma máquina está em boa condição segue uma distribuição exponencial com razão λ. Quando uma máquina é defeituosa ou começa a falhar, ela é enviada ao centro de reparos onde R pessoas de reparo estão disponíveis. O tempo que se toma em reparar uma máquina é assumido que é exponencial com razão μ. Uma vez que uma máquina é reparada, ela novamente volta a funcionar corretamente mais outra vez fica suscetível a falhar no futuro.

O modelo de reparo de máquinas pode ser definido como um processo de nascimento e morte, em que o estado (j) em qualquer instante é o número de máquinas que estão com defeito. Exemplo:

M/M/R/GD/K/K com k = 5 e R = 2

Interpretação de cada estado para o modelo de reparo de máquina

Estado do Sistema	Número de Máquinas sem defeito	Máquinas Defeituosas em Espera	Número de Atendentes Ocupados
0	ND ND ND ND ND	-	0
1	ND ND ND ND	-	1
2	ND ND ND	-	2
3	ND ND	D	2
4	ND	DD	2
5	-	DDD	2

Quadro 4: Modelo de Reparo de Máquinas K = 5 e R = 2

Gráfico

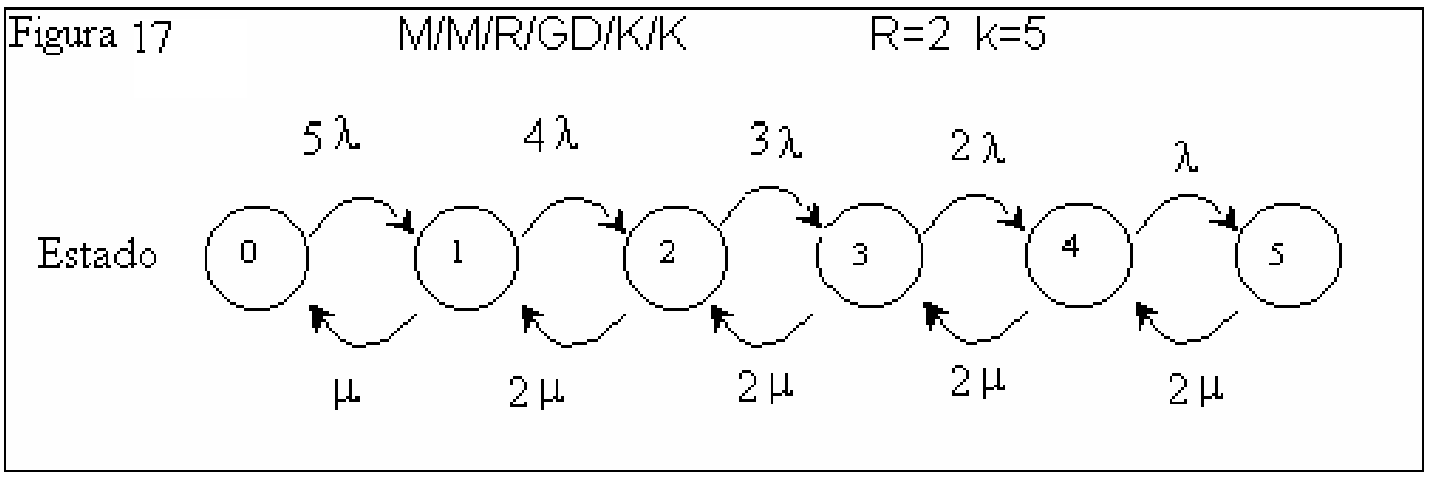


Fig 17: Rede de Transição de Estados do Modelo M/M/R/GD/K/K

$$\pi_j = \binom{k}{j} \rho^j \pi_0 \qquad (j = 0, 1, 2, \ldots\ldots, R)$$

$$\pi_j = \frac{\binom{k}{j} \rho^j \, j! \pi_0}{R! \, R^{j-R}} \qquad (j = R+1, R+2, \ldots\ldots, k)$$

$$L = \sum_{j=0}^{j=k} j\pi_j \qquad L_q = \sum_{j=R}^{j=k} (j-R)\pi_j$$

$$\bar{\lambda} = \sum_{j=0}^{j=k} \lambda_j \pi_j = \sum_{j=0}^{j=k} (k-j)\lambda\pi_j = \lambda\left(K\sum_{j=0}^{j=k} \pi_j - \sum_{j=0}^{j=k} j\pi_j \right) = \lambda(K*1-L) = \lambda(K-L)$$

$$W = \frac{L}{\bar{\lambda}} \qquad Wq = \frac{Lq}{\bar{\lambda}}$$

Exemplo:

O Departamento de Polícia tem cinco carros. Um carro quebra por mês. O departamento tem dois mecânicos, cada um leva em média três dias para consertar um carro. Os tempos de quebra e de reparo são exponenciais.

1. Determinar a média de carros em boas condições
2. Encontrar a média do tempo em que um carro fica parado
3. Encontrar a fração de tempo em que um mecânico fica ocioso

Este é um modelo M/M/R/GD/k/k. Primeiro encontrar o tráfego de intensidade $\rho = \lambda/\mu$

$\rho = \frac{1/30}{1/3} = 0.1$ Para a primeira pergunta é necessário encontrar K – L. Então:

$\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = 1$

$$\pi_0 + \binom{5}{1}\left(\frac{1}{10}\right)\pi_0 + \binom{5}{2}\left(\frac{1}{10}\right)^2\pi_0 + \binom{5}{3}\left(\frac{1}{10}\right)^3\frac{3!}{2!2}\pi_0 + \binom{5}{4}\left(\frac{1}{10}\right)^4\frac{4!}{2!2^2}\pi_0 + \binom{5}{5}\left(\frac{1}{10}\right)^5\frac{5!}{2!2^3}\pi_0 = 1$$

$$\pi_0(1+0.309296+0.061859+0.009279+0.000928+4.64\text{E-}05) = 1 \Rightarrow \pi_0 = 1-0.381408$$

$$\pi_0 = 0.618592$$

Logo $L = \sum_{j=0}^{j=5} j\pi_j$ = 1(0,3092)+2(0,0618)+3(0,0092)+4(0,00092)+5(0,000046)=0,465

K - L = 5 – 0,465 = 4,535 carros em boas condições

A segunda pergunta pede encontrar W.

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

$\bar{\lambda}$ = λ (K–L)

$\bar{\lambda}$ = (1/30)*4,535 = 0,515 , logo W = 0,465/0,515 = 3,08 dias

A terceira pergunta pede para encontrar em média a fração do tempo em que um mecânico fica ocioso, logo simplesmente essa fração de tempo é uma probabilidade. Ela é: (π_0 *1 + π_1 * 0,5) = 0,154 * 0,618 = 0,77 ou 77%

3.2.5 **Os Modelos M/G/∞/GD/∞/∞ e GI/G/∞/GD/∞/∞**

À continuação, são mostrados alguns exemplos de sistemas com servidores infinitos ou auto-atendimento.

Situação	Chegadas	Tempo de Atendimento	Estado do Sistema
Programa Acadêmico	Estudantes ingressam no programa	Tempo que o estudante permanece no programa	Número de estudantes no programa
Indústria	Empresa ingressa à indústria	Tempo até que a empresa sai da indústria	Número de empresas na indústria

Quadro 5: Exemplos de Filas de Espera com Auto-atendimento

$L = L_s = \frac{\lambda}{\mu}$ $W = W_s = \frac{1}{\mu}$ $W_q = L_q = 0$

$\pi_j = \frac{(\frac{\lambda}{\mu})^j e^{-\frac{\lambda}{\mu}}}{j!}$ com parâmetro $\rho = \lambda/\mu$

Exemplo:

Existem 40 empresas para captação de energia solar atualmente no estado de Indiana, Estados Unidos. Em média 20 destas empresas são abertas a cada ano no estado. A média de sobrevivência das firmas é de dez anos. Se continuar essa tendência, a) qual será o número de empresas de energia solar que encontra no estado americano? Se o tempo entre as empresas que surgem nesta atividade obedece à distribuição exponencial, b) qual será a probabilidade de existirem (em estado estável) mais de 300 empresas de energia solar nesta indústria? (Ajuda: para valores grandes de λ, a distribuição de Poisson pode ser aproximada pela distribuição Normal).

O modelo adequado para esse problema é o M/G/∞, no qual o surgimento de empresas não tem limite e os tempos de chegada são exponenciais. As informações dizem que E(x) = λ = 20 empresas por ano e a média de sobrevivência é W = $W_s = \frac{1}{\mu}$ = dez anos por indústria. Logo o valor esperado em 10 anos será:

a) L = $\lambda \frac{1}{\mu}$ = 20 (10) = 200 empresas. A probabilidade em estado estável será:

b) $\pi(x > 300) = \sum_{j=300}^{\infty} \frac{\rho^j}{j!} e^{-\rho}$. Distribuição de Poisson com parâmetro ρ.

Aproximando a distribuição de Poisson como uma distribuição normal:

Usa-se a fórmula da distribuição normal acumulada (F(x)) para encontrar a área. Pois a área representa a probabilidade de uma variável randômica contínua ficar no intervalo que limita essa área. Utilizando o programa Excel, usa-se a fórmula: NormDist (x, λ, σ, 1), onde x é a variável randômica, λ é o valor esperado, σ é o desvio padrão e 1 significa que o valor retornado é a distribuição normal acumulada (área da curva).

Poisson tem sua variância igual a sua média, assim, na média de dez anos:

var(x)= 20(10); O desvio padrão é: $\sigma = 200^{1/2}$

$\pi(x > 300) = 1 - \pi(x \leq 300) = 1 -$ NormDist(300,200,$200^{1/2}$,1) = 1-1 = 0

3.3 **Modelos Não Markovianos**

Resultados analíticos disponíveis somente existem em poucas situações em que não é necessária a hipótese de Markov nos processos de chegada e/ou atendimento. Eles são importantes porque, em situações práticas, as distribuições de tempo de chegada e atendimento não poderão ser razoavelmente aproximadas por distribuições exponenciais. Para sistemas de filas com modelo M/G/1, o resultado das fórmulas são exatos. Para os demais modelos de sistemas de filas, os resultados das fórmulas são aproximados. Não existem fórmulas disponíveis para filas de espera finitas ou populações finitas num sistema de filas Não-Markovianos.

Os sistemas Não-Markovianos precisam da especificação dos coeficientes de variação (COV) do processo de chegada e atendimento. O COV de chegadas é o resultado do desvio padrão de tempo entre chegadas dividido pelo o valor esperado do tempo entre chegadas. O COV de uma distribuição exponencial é 1. Distribuições com menor variabilidade que a distribuição exponencial têm o COV < 1, enquanto que distribuições com maior variabilidade que a distribuição exponencial têm o COV >1. O COV do processo de atendimento é o desvio padrão do tempo de atendimento dividido pelo valor esperado do tempo de atendimento.

3.3.1 **O Modelo M/G/1/GD/∞/∞**

Este sistema não pode ser modelado como um sistema do processo de nascimento-morte porque o tempo de atendimento não é exponencial. Portanto o sistema não tem a propriedade da não-memória. Ou seja, a probabilidade de concluir um atendimento entre o tempo t e t+Δt quando o estado do sistema no tempo t é j depende da duração do tempo a partir do último atendimento concluído.

Por conseguinte a probabilidade de um atendimento concluído entre o tempo t e t+Δt não é da forma $\mu\Delta$t, e o processo de nascimento-morte não é apropriado. Como conseqüência, a determinação de probabilidades em estado estável para o sistema M/G/1/GD/∞/∞ é difícil de determinar. A teoria das cadeias de Markov é usada para determinar π_i'.

π_i' é a probabilidade de um sistema que esteja em operação num longo período de tempo, com i clientes que estão presentes no instante imediato após que um atendimento for concluído.

$\pi_i' = \pi_i$, onde π_i é a fração de tempo após o sistema alcançar um estado estacionário, em que i clientes estão presentes. (ver Kleinrock 1975). Por exemplo, π_0' é a probabilidade de ter zero clientes no sistema mas π_0 também é igual à fração de tempo onde o atendimento fica ocioso.

Afortunadamente, utilizando os resultados de **Pollaczek** e **Khintchine**, se determina L, L_q, L_s, W, W_q, e W_s.

Pollaczek e **Khintchine** demonstraram que para um sistema M/G/1/GD/∞/∞ de filas de espera:

$L_q = \dfrac{\lambda^2 V \oplus \rho^2}{2(1-\rho)}$, onde V é a variância do atendimento e $\rho=\lambda/\mu$

Definindo o valor esperado da distribuição do tempo de atendimento: E(S) = $\dfrac{1}{\mu}$ e sua variância: var S = σ^2:

$L = L_s + L_q = \rho \oplus \dfrac{\lambda^2 V \oplus \rho^2}{2(1-\rho)}$, onde V é a variância de atendimento

Desde que $W_s = \dfrac{1}{\mu}$, Logo $L_s = \lambda W_s \rightarrow L_s = \dfrac{\lambda}{\mu}$

$W_q = \frac{L_q}{\lambda}$ $W = W_s + W_q \rightarrow W = \frac{1}{\mu} + \frac{L_q}{\lambda}$

Também pode ser demonstrado que π_0 (a fração de tempo em que o atendimento fica ocioso) é igual a 1 - ρ. Este resultado é o mesmo do modelo M/M/1/GD/∞/∞.

Exemplo:

Um processo de ordem de pedidos num depósito recebe chamadas para atendimento com valor esperado de 8,5 por hora. O valor esperado de tempo para preencher um pedido é 0,1 horas. Por motivos de análises assumir que os dois tempos são distribuídos exponencialmente. Analisando o sistema como M/M/1, o valor esperado do tempo na fila de espera (fila 12) é de 0,5667 horas. Uma oportunidade existe para reduzir a variabilidade do processo de preencher os pedidos. O gerente de inventários quer saber se essa mudança justifica seu custo.

	A	B	C	D
1	**Sistema**	Que2	Que3	Que4
2	Razão de Chegada	8.5	8.5	8.5
3	Razão de Atendimento/Canal	10	10	10
4	Número de Atendentes	1	1	1
5	COV de Tempos de Chegada	1	1	1
6	COV de Tempos de Atendimento	1	0.5	0
7	COV de Saídas	1	0.676849	0.526783
8	Modelo	M/M/1	M/G/1	M/G/1
9	Número médio no Sistema	5.666668	3.860417	3.258334
10	Tempo médio no Sistema	0.666667	0.454167	0.383333
11	Número médio na Fila	4.816668	3.010417	2.408334
12	Tempo médio na Fila	0.566667	0.354167	0.283333
13	Número médio em Atendimento	0.85	0.85	0.85
14	Tempo médio em Atendimento	0.1	0.1	0.1
15	Eficiência	0.85	0.85	0.85

Quadro 6: Exemplo Modelos MM1 e MG1

Nos resultados na planilha de Excel observar que, reduzindo a variabilidade (fila 6 mostra 1, ½, e 0 como valores da variabilidade respectivamente) no processo do atendimento, causa um decréscimo do tempo na fila de espera (fila 12). O tempo médio de atendimento (fila 14) não muda porque é fixado pelos dados iniciais e não é influenciado pela variabilidade. A eficiência (fila 15) somente depende da quantia de trabalho disponível comparado com o número de atendentes, portanto não muda com a mudança da variabilidade. A mudança compensa para o gerente de inventários, porque o cliente fica menos tempo na fila de espera (ver fila 12).

3.3.2 **O Modelo M/G/S/GD/S/∞**

Se os clientes que chegam encontram os atendentes ocupados, eles saem do sistema sem ser atendidos. Este sistema se chama “saída de clientes não atendidos” (**blocked customers cleared**) ou o sistema BCC. Ou seja, nesse modelo o estado máximo do sistema (S) é igual ao número de atendentes (S). Como uma fila de espera não existe, $L_q = W_q = 0$, e $W = W_s = \frac{1}{\mu}$.

Na maioria dos sistemas BCC, o objetivo principal é focado na fração de todas as chegadas que não entram no sistema. Por tanto a fração de chegadas que não entram ocorre quando S clientes estão presentes no sistema, esta fração é π_s. Ou seja, o produto ($\lambda^*\pi_s$) das chegadas será perdido pelo sistema, de tal forma que o produto $\lambda(1 - \pi_s)$ de chegadas por unidade de tempo entrará no sistema. Conclui-se que $L = L_s = \frac{\lambda(1-\pi_s)}{\mu}$. Para este modelo pode ser mostrado que π_s (probabilidade de que existem S clientes no sistema) depende somente da média do tempo de atendimento $\frac{1}{\mu}$ e da média do número de chegadas ao sistema λ. Este fato é conhecido como a fórmula de clientes não atendidos de Erlang (**Erlang’ loss formula**). Portanto, qualquer modelo M/G/S/GD/S/∞ com razão de chegada λ e tempo médio de atendimento $\frac{1}{\mu}$ terá a mesma π_s.

Exemplo:

O Departamento do Corpo de Bombeiros recebe uma média de 24 solicitações de carros por hora. Cada pedido gera uma média de 20 minutos de ter carros não disponíveis. Para ter no máximo 1% de possibilidades de não responder a uma solicitação, quantos carros o corpo de bombeiros deveria ter? Assumir que os tempos entre os pedidos dos carros ao Corpo de Bombeiros são exponenciais.

Cada pedido que não seja atendido imediatamente é perdido; ou seja, não existe fila de espera. Portanto, o modelo certo deste problema de congestionamento é o M/G/S/GD/S/∞. Uma vez que o número de clientes seja igual a S, os demais clientes não podem ser atendidos e eles simplesmente são considerados como perdidos para sempre.

Os pedidos são perdidos quando o estado do sistema alcança S. Então a probabilidade de que existam S clientes no sistema (π_s) deverá ser no máximo igual ou menor que 1%. Ou seja: $\pi_s = \pi_0 * C_s \leq 0{,}01$

Logo: $C_j = \frac{\lambda_0 \lambda_1 \lambda_2 \lambda_3 \ldots\ldots\ldots \lambda_{j-1}}{\mu_1 \mu_2 \mu_3 \ldots\ldots\ldots \mu_j}$ e $\pi_0 = \frac{1}{1 + \sum_{j=1}^{j=\infty} C_j}$

Logo, as probabilidades dos estados do sistema começam desde S = 0. O cálculo continua até que o produto de $\pi_0 * C_s$ não ultrapasse o valor 0,01.

Para S = 0 $\pi_0 * C_0 \leq 0{,}01$

Para S = 1 $\pi_0 * C_1 \leq 0{,}01$

Para S = 2 $\pi_0 * C_2 \leq 0{,}01$

A seguinte planilha mostra os cálculos em Excel. Na planilha criada em Excel observar que na fila 21 e na coluna E (que representa o valor de π_j) apresenta o valor de 0.01706 que ultrapassa o valor 0,01; portanto, na fila 22, o valor 0,009 é menor que 0,01. Então se conclui que o número mínimo de carros disponíveis que o corpo de bombeiros deve ter é de 15 carros mostrados na fila 22 e na coluna A.

	A	B	C	D	E	F	G
1	MODELO	λ	μ	s			
2	M/G/S/GD/S/∞	24	3	15			
3		L ou LS	π15				
4		7.927192889	0.009100889				
5						# EM	
6	ESTADO j	λ	μ	Cj	πj	ESPERA	j*πj
7	0	24	0	1	0.000338247	0	0
8	1	24	3	8	0.002705974	0	0.002705974
9	2	24	6	32	0.010823896	0	0.021647791
10	3	24	9	85.33333333	0.028863722	0	0.086591166
11	4	24	12	170.6666667	0.057727444	0	0.230909775
12	5	24	15	273.0666667	0.09236391	0	0.46181955
13	6	24	18	364.0888889	0.12315188	0	0.73891128
14	7	24	21	416.1015873	0.140745006	0	0.985215039
15	8	24	24	416.1015873	0.140745006	0	1.125960045
16	9	24	27	369.8680776	0.125106672	0	1.125960045
17	10	24	30	295.8944621	0.100085337	0	1.000853373
18	11	24	33	215.1959724	0.072789336	0	0.800682699
19	12	24	36	143.4639816	0.048526224	0	0.58231469
20	13	24	39	88.28552715	0.029862292	0	0.388209793
21	14	24	42	50.44887266	0.017064167	0	0.238898334
22	15	0	45	26.90606542	0.009100889	0	0.136513334
23	16	0	45	0	0	1	0

Quadro 7: Exemplo do Modelo M/G/S/GD/S/∞

Com 15 carros, o máximo percentual de solicitações perdidas é de 1%, então o Corpo de Bombeiros atende 99% dos pedidos do público sem perda de solicitações.

3.3.3 O Modelo G/G/M

Na maioria das situações, os tempos de chegada seguem uma variável aleatória exponencial. Mas, freqüentemente os tempos de atendimentos não seguem uma distribuição exponencial. Quando os tempos de chegada e os tempos de atendimento, cada um, seguem uma distribuição não exponencial, chamar o sistema como G/G/m. A primeira G indica que os tempos de chegada sempre seguem a mesma variável aleatória (não necessariamente exponencial), e a segunda G indica que o tempo de atendimento segue sempre uma mesma variável aleatória (mas não necessariamente exponencial). Por esta razão, os padrões de modelos anteriores já discutidos não são válidos para este modelo. Afortunadamente, as aproximações de Allen-Cunneen são geralmente boas para encontrar L, W, L_q, e W_q para o modelo G/G/M.

$$\rho = \frac{\lambda}{s\mu},\ W_s = 1/\mu,\ R(s,\mu) = 1 - \left(\frac{Poisson(s, \frac{\lambda}{\mu}, 0)}{Poisson(s, \frac{\lambda}{\mu}, 1)} \right)$$

onde $R(s,\mu)$ utiliza a função de Probabilidade de Poisson com variável aleatória S (número de atendentes), razão $\rho=\lambda/\mu$ e o valor lógico um para retornar a função acumulada ou zero para retornar o valor da função de Poisson

$$W_q = \left[\frac{E_c(s,\mu)}{s(1-\rho)} \times W_s \left(\frac{COV^2\ \text{Chegada} \oplus COV^2 \text{Atendimento}}{2} \right) \right], E_c(s,\mu) = \frac{1 - R(s,\mu)}{1 - R(s,\mu) \times \rho}$$

Exemplo:

Existe uma média de 230 clientes por hora que chegam à área de atendimento de passageiros onde oito agentes estão em atendimento. Cada agente pode atender 30 clientes por hora. Os quadrados dos coeficientes de variação de chegada e atendimento são 1,5 e dois respectivamente. a) Em média quantos clientes estarão presentes na área de atendimento? b) Quanto tempo em média um cliente deve esperar antes de ser atendido?.

$\lambda = 230$ clientes por hora, $\mu = 30$ clientes por hora.

S = 8, $\rho = \frac{230}{8*30} = \frac{23}{24} = 0{,}9583$

O cálculo mais longo é encontrar W_q. Uma vez encontrado o tempo de espera, encontrar W e usando o teorema de Little encontrar L.

Inserindo os dados no *Add-in* de Excel, ele mostra o seguinte. Analisando estes resultados observar que na fila 9, L em média é igual a 42,65 clientes, ou seja, 42 clientes aproximadamente e o tempo de espera na fila 12 é em média 0,152 horas ou nove minutos e sete segundos aproximadamente.

Aproximação de Allen-Cunneen

	A	B
1	**Sistema**	Que1
2	Razão de Chegada	230
3	Razão de Atendimento	30
4	Número de Atendentes	8
5	COV de tempos de chegada	1.224745
6	COV de tempos de atendimento	1.414214
7	COV de Saídas	1.399715
8	Modelo	G/G/8
9	Número Médio no Sistema	42.65348
10	Tempo Médio no Sistema	0.18545
11	Número Médio na Fila de Espera	34.98682
12	Tempo Médio na Fila	0.152117
13	Número Médio em Atendimento	7.666667
14	Tempo Médio em Atendimento	0.033333
15	Eficiência	0.958333

Quadro 8: Exemplo do Modelo G/G/S

3.4 Modelos em Série

Em muitas situações, uma unidade de chegada não somente passa por uma fila de espera, mas por uma série de filas de espera. Um exemplo é matricular-se em um colégio. Estudantes visitam um número de departamentos requeridos para a aprovação do programa acadêmico. Cada departamento tem um ou mais atendentes e uma fila de espera para os estudantes. Outro exemplo é em uma fábrica onde os equipamentos são agrupados de acordo com a sua função em diferentes estações. Um grupo de equipamentos similares pode ser modelado como uma estação de um sistema de filas. A unidade de chegada passa por uma ou mais estações antes de sair da fábrica. A análise deste tipo de redes pode ser sumamente difícil a menos que eles tenham uma estrutura especial.

Existem resultados analíticos disponíveis quando todas as chegadas externas são Poisson, cada sistema de filas dentro da rede tem capacidade ilimitada em que os tempos de atendimento são exponenciais; e as transferências de um sistema de filas a outro são feitas randomicamente de acordo a probabilidades já determinadas. Este tipo de sistema é nomeado **rede de Jackson** (Jackson Network), e alguns de seus estados com comportamentos estáveis podem ser analisados usando as mesmas fórmulas desenvolvidas para um só sistema de filas de uma estação Markoviana. Estas fórmulas são simples de usar e os resultados das fórmulas são críticos para o desenho de um sistema de filas.

Quando os sistemas de filas individuais não são Markovianos os resultados são aproximados.

Exemplo:

Um conjunto de programas processados chega consecutivamente a um centro de informática e passa por três estações: Um processador de *input*, o processador central, e uma impressora. Os programas chegam ao centro de informática aleatoriamente com valor esperado de dez por minuto. Para conseguir atender todos os programas, o centro de informática poderá ter vários processadores dos três tipos, operando em paralelo. Os tempos das três estações têm distribuições exponenciais com os seguintes valores esperados: processador de *input*, dez segundos; processador central, três segundos; e impressora 70 segundos. Quando os processadores não estão imediatamente disponíveis para atendimento, os programas devem esperar numa fila. A fila de espera que exista em cada processador é infinita. O trabalho é encontrar o mínimo número de cada tipo de processadores e calcular o valor esperado do tempo requerido por um programa em sair do sistema com sucesso.

A seguinte planilha de Excel mostra os resultados. A primeira pergunta é contestada na fila 4 com 2, 1, e 12 tipos de processadores respectivamente. Estes números são o mínimo para manter o sistema estável. A segunda pergunta tem a resposta na fila 7 com uma média de cinco minutos aproximadamente que um programa permanece no sistema.

	A	B	C	D	E	F
1	**Sistema**	Input	Processador	Impressora	**Rede Aberta**	QueNet1
2	Razão de Chegada	10	10	10	Razão de Chegada	10
3	Razão de Atendimento/Canal	6	20	0.857		
4	Número de Atendentes	2	1	12		Em
5	Modelo	M/M/2	M/M/1	M/M/12	Total	Serie
6	Número médio no Sistema	5.455	1	43.071083		49.5256
7	Tempo médio no Sistema	0.545	0.1	4.3071079		4.95256
8	Número médio na Fila	3.788	0.5	31.40247		35.6903
9	Tempo médio na Fila	0.379	0.05	3.140247		3.56903
10	Número médio em Atendimento	1.667	0.5	11.668612		13.8353
11	Tempo médio em Atendimento	0.167	0.05	1.1668612		1.38353
12	Eficiência	0.833	0.5	0.9723843		

Quadro 9: Exemplo de Filas em Serie

3.4.1 Rede de Fila de Espera Aberta

A rede de fila de espera aberta é uma generalização do modelo de filas de espera em série.

Exemplo

	A	B	C	D	E	F	G
1	**Queue Station**	Input	Processador	Impressora	**Rede Aberta**	QueNet	
2	Razão de Chegada Independente	10	0	0	Razão de Chegada	10	
3	Razão de Chegada	10	8	3.2	Total		
4	Razão de Atendimento/Canal	6	20	0.857			
5	Número de Atendentes	2	1	4	No Sistema		
6	Modelo	M/M/2	M/M/1	M/M/4	Modelo	General	
7	Número Médio na Estação	5.4545	0.66666669	15.756813		21.88	
8	Tempo Médio na Estação	0.5455	0.08333334	4.9240036		2.188	
9	Número Médio na Fila	3.7879	0.26666669	12.022857		16.08	
10	Tempo Médio na Fila	0.3788	0.03333334	3.7571427		4.169	
11	Número Médio em Atendimento	1.6667	0.40000001	3.7339559		5.801	
12	Tempo Médio em Atendimento	0.1667	0.05	1.1668612		1.384	
13	Eficiência	0.8333	0.40000001	0.933489			
14	Tempo de Espera Crítico	1	1	1			
15	P(Espera >= Espera Crítica)	0.1025	2.4577E-06	0.6819823			
16	**Matriz de Transição**	Input	Processador	Impressora	**Matriz Inversa**		
17	Input	0	0.8	0	1	0.8	0.32
18	Processador	0	0	0.4	0	1	0.4
19	Impressora	0	0	0	0	0	1
20	**Matriz Augmented**	Input	Processador	Impressora			
21	Input	1	-0.8	0			
22	Processador	0	1	-0.4			
23	Impressora	0	0	1			

Quadro 10: Exemplo de Fila Aberta

Continuando com o exemplo anterior, tem-se condições adicionais: Todos os trabalhos precisam passar pelo processador de *input*. Mas, por motivos

de erros, somente 84% dos programas passam pelo processador central. Finalmente somente os 40% dos programas que passam pelo processador central vão para a impressora. Igualmente, o no quadro 10, mostra-se os resultados com as condições requeridas.

A razão de chegada mostrada na linha 3 é calculada com base na equação que envolve a inversa da matriz *Augmented*. A matriz *Augmented* é o resultado da matriz de Transição subtraída da matriz identidade. As entradas na matriz de Transição são inteiramente gerais desde que a matriz Inversa exista. Uma entrada na diagonal representa uma reciclagem na estação. Em muitos casos, os números na matriz representam proporções de fluxo de transição. Um requisito da rede de Jackson é ter caminhos probabilísticos. Ou seja, uma entidade particular saindo de uma estação será transferida à próxima com uma distribuição de probabilidade determinada.

Os números na coluna F representam o valor esperado das variáveis de todo o sistema, respectivamente.

Alternativamente as razões de chegada podem ser encontradas manualmente, resolvendo um sistema de equações da seguinte maneira: Cada estação tem uma equação. Por exemplo: duas estações teriam duas equações com duas incógnitas, três estações teriam três equações com três incógnitas etc. Neste caso existem três estações; portanto, se tem três equações com três incógnitas.

O sistema de equações geral é o seguinte: $\lambda_j = r_j + \sum_{j=1}^{j=k} \lambda_i \pi_{i,j}$.

j representa uma estação qualquer.

r_j = Razão de chegada externa à estação j

λ_i = Razão de chegada à estação i

$\pi_{i,j}$ = Probabilidade de transição da estação i à estação j

Logo: λ_1= 10; λ_2= (0,8)x λ_1 = 8; λ_3= (0,4) x λ_2 = (0,4)x(8) = 3,2

$W = \frac{L}{\lambda}$ Onde $\lambda = r_1 + r_2 + r_3 + r_4 + \ldots\ldots + r_k$

W representa o tempo médio que um cliente qualquer fica no sistema.

3.4.2 **Rede de Fila de Espera Fechada**

Em fábricas, indústrias e sistemas de computação, onde existem um constante número de diversas atividades acontecendo; estes podem ser modelados com uma rede de fila de espera fechada. Quando um trabalho (cliente) termina de ser atendido (sai do sistema), ele é substituído por outro que entra no sistema, assim, sempre mantendo constante o número de clientes no sistema. Lembrando que nas redes em fila de espera aberta, o número de atividades em cada atendimento eram variáveis aleatórias independentes; aqui, como o número de trabalhos (clientes) no sistema é sempre constante, a distribuição de atividades (clientes) nos diferentes servidores (atendentes) não pode ser independente. O algoritmo de **Buzen** pode ser usado para determinar as probabilidades em estado estável para redes de fila de espera fechada.

Para determinar λ_j usar a mesma equação da Rede de Fila de Espera Aberta, ou seja: $\lambda_j = r_j + \sum_{j=1}^{j=s} \lambda_i \pi_{i,j}$. Mas neste caso r_j= 0 porque não existem chegadas externas, uma vez que o sistema alcança um estado estacionário. S representa o número de servidores (atendentes). N é o número total de trabalhos no sistema. Para determinar as probabilidades em estado estável, se mostra a fórmula seguinte: $\pi_N(n) = \frac{\rho^{n1}{}_1 \rho^{n2}{}_2 \cdots \rho_n^{ns}}{G(N)}$ Onde n_1, n_2,....n_s representam o número de trabalhos em cada estação respectivamente. A somatória de cada n_i para i= 1, 2,....s deve ser igual a N. G(N) é determinado pelo algoritmo de Buzen e ρ_n é a intensidade do tráfego em cada estação.

Exemplo:

As chegadas de diversas atividades chegam a um servidor de arquivos que tem um CPU, um disco rígido I e um disco rígido II. Com uma probabilidade de 65%, um trabalho vai do CPU até o disco rígido 1 e com uma probabilidade de 30%, um trabalho vai do CPU até o disco rígido 2. Com probabilidade de 5%, um trabalho termina após o processamento no CPU e é substituído imediatamente por outro trabalho. Existem sempre três tipos de atividade no sistema. O tempo médio em completar uma operação no CPU é de 0,039 segundos. O tempo médio para completar uma operação no disco rígido I é de 0,18 segundos e o tempo médio em completar uma operação no disco rígido II é de 0,26 segundos.

a) Determinar a probabilidade de estados estáveis do número de trabalhos de cada parte do sistema?

b) Qual é o valor esperado do número de trabalhos no CPU? No disco rígido I? No disco rígido II?

c) Qual é a probabilidade de que o CPU esteja ocupado? O disco rígido I? O disco rígido II?

d) Qual é o valor esperado do número de trabalhos completados por segundo no CPU? No disco 1? No disco 2?

As perguntas a, b, c, e d são resolvidas usando o algoritmo de Buzen na seguinte planilha de Excel. Em a) é perguntada a probabilidade dos estados estáveis (0 até 3) das partes do sistema respectivamente (CPU, Disco 1 e o Disco 2). Estas probabilidades são mostradas no oval destacada na planilha com a letra a. Assim mesmo, as perguntas b, c, e d são respondidas na planilha que mostra os ovais respectivos com as letras b, c e d.

	A	B	C	D	E	F
1	Lambda	1	0.65	0.3		
2	Miu	25.64102564	5.555555556	3.846153846	Partes	Prob
3	phoi	0.039	0.117	0.078	CPU	
4		CPU	Disco 1	Disco 2	0	0.722222
5	0	1	1	1	1	0.211111
6	1	0.039	0.156	0.234	2	0.055556
7	2	0.001521	0.019773	0.038025	G(N) 3	0.011111
8	3	0.000059319	0.00237276	0.0053387	Partes	
9	Partes	Partes	Partes		Disco 1	
10	no CPU	no Disco 1	no Disco 2	Probabilidade	0	0.166667
11	0	0	3	0.088888889	1	0.233333
12	0	1	2	0.133333333	$\Pi_N(n)$ 2	0.3
13	0	2	1	0.2	3	0.3
14	0	3	0	0.3	Partes	
15	1	0	2	0.044444444	Disco 2	
16	1	1	1	0.066666667	0	0.444444
17	1	2	0	0.1	1	0.288889
18	2	0	1	0.022222222	2	0.177778
19	2	1	0	0.033333333	3	0.088889
20	3	0	0	0.011111111		
21		Número Médio de Trabalhos	Prob Ocupado	Trabalhos Completos por segundo		
22	CPU	0.355555556 b	0.277777778 c	7.122507123 d		
23	Disco 1	1.733333333	0.833333333	4.62962963		a
24	Disco 2	0.911111111	0.555555556	2.136752137		

Quadro 11: Exemplo de Fila Fechada

O algoritmo de Buzen permite determinar de uma maneira eficiente em Excel o valor de G(N). Uma vez que encontradas as probabilidades em estado estável, usando G(N), se pode determinar outras medidas de efetividade como o número médio de trabalhos na fila de espera de cada estação e o valor esperado do tempo que um trabalho toma em cada estação (atendimento). Além de outras medidas como a fração de tempo que um atendente está ocupado e o número de

trabalhos processados de cada atendente por unidade de tempo.(Cf. Winston, 2004, p.85)

3.5 **Custo Mínimo de um Sistema de Filas de Espera M/M/1**

- CT: Custo médio total do sistema
- CE: Custo de espera médio no sistema por unidade de tempo
- CA: Custo de atendimento médio no sistema por unidade de tempo
- CE_{unit} : Custo de espera unitário (por cliente) por unidade de tempo
- CA_{unit}: Custo de atendimento unitário (por cliente) por unidade de tempo

Dessa forma se pode escrever a relação: CT = CE + CA

$CE = CE_{unit} * L$, onde L = valor esperado do número de clientes no sistema

$CA = CA_{unit} * \mu$, onde μ é o valor esperado do número de clientes já atendidos por unidade de tempo no sistema. Substituindo:

$$CT = (CE_{unit} * \frac{\lambda}{\mu - \lambda}) + (CA_{unit} * \mu)$$

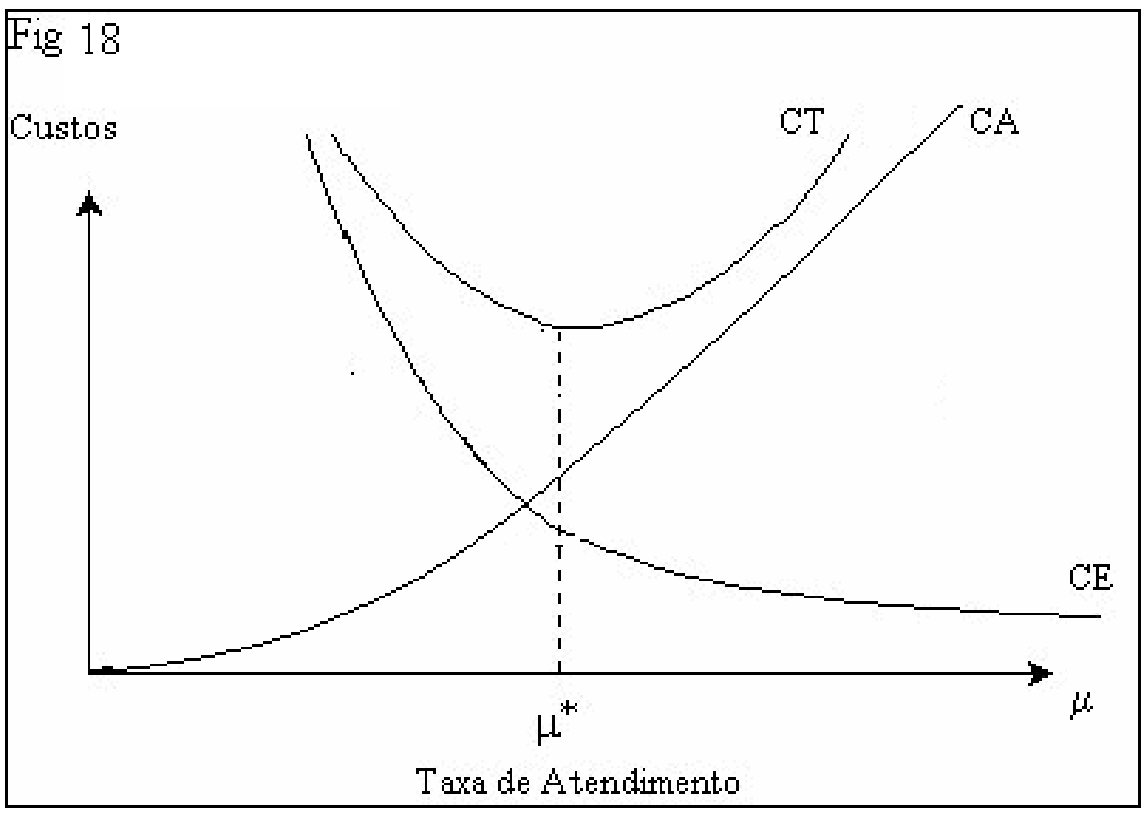


Fig 18: Custo Mínimo do Modelo M/M/1

Na figura 18, mostra a relação entre o custo($) e o valor da média de atendimentos por unidade de tempo (μ), onde o custo representa a linha central vertical Y e (μ) representa a linha horizontal X. Nesta figura existem três curvas: o custo total (CT), o custo de espera (CE), e o custo de atendimento (CA).

Observando a curva do custo total (CT), o custo mínimo fica no ponto mais baixo da curva. Logo Y teria a mínima altura, portanto, o custo mínimo. O valor desta altura (custo mínimo) pode ser encontrado usando μ. Ou seja, observando a figura 18, com o valor de μ encontrar o valor de Y (custo mínimo).

Assim mesmo, a derivada da curva do custo total no ponto mais baixo é zero, porque este ponto pertence a uma reta horizontal que é a tangente da curva.

Então, na equação do (CT), existem 2 variáveis: o custo total e a taxa de atendimento μ, ficando como constantes as demais quantidades. Logo, deriva-se com relação à μ; ou seja, $\frac{d(CT)}{d\mu} = 0$ no ponto mínimo da curva. Assim, substituindo o valor de (CT) e efetuando a derivação resulta em:

$$\frac{d(CT)}{d\mu} = - CE_{unit.} \frac{\lambda}{(\mu - \lambda)^2} + CA_{unit} = 0$$

$(\mu-\lambda)^2 . CA_{unit} = \lambda . CE_{unit}$, onde $\mu = \lambda + \sqrt{\frac{\lambda * CE_{unit}}{CA_{unit}}}$

μ é a razão de atendimento que resulta no menor custo total no modelo M/M/1

Exemplo:

Uma oficina de reparos de eletrodomésticos recebe por dia uma média de dois pedidos de consertos, segundo uma distribuição de Poisson. O eletricista consegue reparar uma média 2,5 aparelhos por dia, também segundo a distribuição de Poisson. A oficina estima que cada dia de espera de um aparelho custa $84 em termos de seguros e deterioração da imagem da firma. Por outro lado, de mão de obra, cada conserto custa em média $84 por dia.

- Determinar o custo total de operação da firma por dia.
- Determinar a eficiência do eletricista que resultaria no menor custo total.

Dados:

λ = 2 pedidos de conserto por dia

μ = 2,5 consertos por dia

CE_{unit} = $84 por aparelho/dia

CA_{unit} = $84 por aparelho/dia

Custo de espera:

CE = CE_{unit} * $\frac{\lambda}{\mu - \lambda}$ = 84 * $\frac{2}{2,5 - 2}$ = $ 320 por dia.

Custo de reparos:

CA = CA_{unit} * μ = 84 * 2,5 = $ 200 por dia.

Custo total:

CT = 320 + 200 = $ 520 por dia.

Eficiência que resulta no custo total mínimo:

μ^* = 2 + $\sqrt{\frac{2x80}{80}}$ = 3,41 aparelhos por dia.

Assim, conclui-se que, para obter o menor custo total de operação, o eletricista deve ser capaz de consertar uma média de 3,41 equipamentos por dia.

A eficiência é a probabilidade de o atendente estar ocupado, ou seja: P(n≥1) = 1 – P_0 = 1- (1-ρ) = ρ. Neste caso $\rho = 2 / 3,41 = 0,585 = 58.5\%$. O seja, neste caso o atendente está ocupado 58,5%. Quando μ = 2,5; a eficiência é ρ= 4/5 = 84%.

Nota: A fórmula CT = CA + CE é válida para obter o custo de qualquer modelo de filas de espera. Onde os valores de CA e CE são obtidos tomando em conta as características da cada problema. (Cf. Andrade, 1998, P.84)

3.6 COMPORTAMENTO TRANSITÓRIO DE UM SISTEMA DE FILAS

"Por tudo dito até agora, a razão de chegada, a razão de atendimento e o número de atendentes são constantes através do tempo. Isto permite falar da existência de um estado estável. Mas em muitas situações as razões de chegada e atendimento, assim como o número de atendentes, podem variar. Quando os parâmetros que definem o sistema de uma fila de espera variam através do tempo, chamar esse sistema de não estacionário (**nonstationary**).

Probabilidades Transitórias

Considere um exemplo de um restaurante de comida rápida (fast-food) que abre às 10 da manhã e fecha às 18 horas. A distribuição de probabilidade do número de clientes presentes a todo o momento entre as 10 e as 18 são chamadas probabilidades transitórias. Por exemplo, para determinar a probabilidade de que

pelo menos dez clientes estejam presentes, esta probabilidade certamente será maior às 12:30 PM que às 15 horas.

Processo Não Homogêneo de Poisson

Neste processo acontecem três fatos:

1. Num instante qualquer t, a probabilidade de uma chegada num sistema de filas de espera é $\lambda(t) \times \Delta t$.

2. No instante t, a probabilidade de chegada de mais de um cliente é $o(\Delta t)$.

3. Chegadas durante intervalos diferentes são independentes (a razão de chegada pode variar em cada intervalo)" (Winston, 2004, p.85)

4. APLICAÇÕES

Existem muitas aplicações importantes na teoria de filas de espera, incluindo o tráfego de: veículos, aviões, pessoas, comunicações; no ordenamento de: pacientes de hospitais, programas de computador, tarefas em máquinas; desenho de agências como: bancos, parques de diversão, restaurantes de comida rápida, correios. Hoje se encontra uma vasta quantidade de filas espera no dia-a-dia, e a teoria das filas pode ajudar a lidar com esses problemas.

Outra das aplicações modernas de filas de espera ocorre numa Rede de computadores. É importante notar que nesta aplicação o congestionamento tem uma característica singular chamada *fractal*. Conceito que será explicado no subitem 4.1.1.

As filas de espera formadas numa rede (Queueing Networks) são geralmente usadas para modelar uma rede de comunicação de informação. Os modelos de um sistema de filas de espera numa rede permitem determinar o atraso típico ou perda dos pacotes de informação que são transmitidos entre os nodos que formam uma rede e também permitem desenhá-la. Deve-se assumir que a razão de chegada dos pacotes de dados a um nodo é constante. (Cf. Winston, 2004, p.85)

4.1 O CASO DE UM SERVIDOR NA UECE

A continuação se usa a teoria das filas para otimizar o servidor da UECE. Analisando os dados do mês de Janeiro.

Assumir que o tempo entre chegadas de mensagens é descrita pela distribuição exponencial com uma razão de chegada constante. De igual forma

assumir o tempo de processamento de mensagens descritas pela distribuição exponencial com uma razão de atendimento constante.

Primeiro a razão de chegada (λ) é 2,93 mensagens por minuto. Pois em Janeiro o total de mensagens que chegam ao servidor durante todo o mês é 130929. Calcula-se (λ) fazendo conversão de dias a minutos. Logo se pode calcular o seguinte: Qual será o tempo de processamento das mensagens no servidor para, por exemplo, perder no máximo uma mensagem em mil mensagens que chegam.

O modelo adequado é o M/M/1/GD/c/∞. Onde c indica a capacidade do sistema.

Usando a fórmula $P(\geq N) = \left(\frac{\lambda}{\mu}\right)^N$ que indica a probabilidade do número mínimo (N) de mensagens no sistema; logo: $\left(\frac{2,93}{\mu}\right)^N \leq \frac{1}{1000}$. Existem 2 variáveis: μ (razão de atendimento) e N (número máximo de mensagens no sistema). É necessário ter N como dado. Supondo N = 4. Se calcula:

$$\left(\frac{2,93}{\mu}\right)^4 \leq \frac{1}{1000} \Rightarrow \frac{1}{\mu} = 0,060 \text{ minutos} \times \frac{60\,\text{sg}}{\text{m}} = 3,6\,\text{sg}$$

Então o servidor deve processar uma mensagem em menos de 3,6 segundos para perder no máximo uma mensagem de cada mil que chegam. Com 3 mensagens na fila e uma mensagem sendo processado.

Por outra parte, calculando o processamento de mensagens por minuto (μ=0,08315) usando como referência o número de mensagens enviados do servidor (3712 mensagens no mês de Janeiro), o servidor processa uma mensagem em aproximadamente 12 minutos!! $\left(\frac{1}{\mu} = 12.025\right)$.

Análise

Obviamente o tempo de processamento das mensagens é muito grande e não representa a realidade. Logo se conclui o seguinte:

- A razão de atendimento das mensagens pode ser calculada a partir da fórmula do número mínimo de mensagens no sistema e não das mensagens enviados pelo servidor.
- A porcentagem das mensagens perdidos depende das razões de chegada e atendimento.

o O tempo que fica uma mensagem no servidor (espera no *buffer* mais processamento) depende somente da razão de atendimento. Pois a média de tempo de processamento de uma mensagem é:

$E(Ws) = \int_0^t tf(t)dt = \frac{1}{\mu}$. Onde f(t) representa uma distribuição qualquer que descreve o tempo de atendimento e μ a razão de atendimento. É necessário ressaltar que f(t) não necessariamente representa uma distribuição exponencial.

A análise para os outros meses é o mesmo. Somente muda a razão de chegada. Por tanto se considera somente Janeiro, para evitar redundância de resultados. Para conhecer outras variáveis, inserir os dados no formulário do *Add-in* de filas de espera em Excel e os resultados são mostrados na seguinte planilha. Os dados em minutos como unidade de tempo são: λ = 2,93 , μ = 16,67 (calculada previamente), o número máximo de mensagens N = 4 (supondo que a capacidade do sistema seja c = 4), e número de servidores = 1.

	B	C
1	**Sistema**	Janeiro
2	Razão de Chegada	2.932997312
3	Razão de Atendimento	16.66666667
4	Servidores	1
5	N Máximo no Sistema	4
6	População	***
7	Modelo	M/M/1/4
8	Número Médio Sistema	0.212718502
9	Tempo Médio Sistema	0.072583355
10	Número Médio na Fila	0.036877771
11	Tempo Médio na Fila	0.012583355
12	Número Médio Atendimento	0.175840735
13	Tempo Médio Atendimento	0.060000002
14	Razão Média Chegada	2.930678844
15	Eficiência	0.175840735
16	Prob Atendentes Ociosos	0.824159265
17	Prob Atendentes Ocupados	0.175840745
18	Prob Sistema Cheio	0.000790429
19	Tempo Crítico Espera (1 min)	1
20	P(Espera ≥ Espera Crítica)	7.51199E-08
21	P(0)	0.824159265
22	P(1)	0.145035416
23	P(2)	0.025523309
24	P(3)	0.004491588
25	P(4)	0.000790429

Quadro 12: Resultados do Servidor da UECE

4.1.1 A Teoria das Filas não se aplica ao Tráfego no Servidor

A teoria das filas não se aplica ao congestionamento no servidor porque as distribuições de Poisson e Exponencial não são apropriadas para descrever seu comportamento. A razão principal é que as características do tráfego numa rede de computadores são descritas com o termo *burstiness* (explosões). *Burstiness* é um conceito qualitativo, mas pode ser descrito no sentido analítico como auto-semelhança (self-similiar) em diferentes intervalos de tempo. Auto-semelhança significa que um objeto ou figura é exatamente igual ou aproximadamente igual a uma parte do mesmo objeto ou figura. O conceito matemático chamado *fractal* é a mais apropriada ferramenta matemática para descrever alguns aspectos do comportamento do tráfego numa rede de computadores. Fractais estão relacionados com distribuições que têm cauda longa (long tail) especificamente a distribuição de Pareto e outras como a Lognormal e Weibull. Mas estas distribuições limitam o uso dos modelos da teoria das filas porque, por exemplo, a distribuição de Pareto não tem definido o valor esperado e a variância nas chegadas e no atendimento, que são medidas fundamentais para usar a teoria das filas. Para superar estas limitações, a comunidade científica está pesquisando vários métodos com o objetivo de analisar, de uma forma mais eficiente e realista, o congestionamento num servidor. Os pesquisadores usam vários métodos como por exemplo:

- Eles tentaram acomodar as distribuições de cauda longa com distribuições tipo fase (distribuição de Erlang) que já têm modelos consolidados na teoria das filas. Este método conseguiu certo sucesso, mas é limitado porque a distribuição de acomodação se torna complicada.
- Outro caminho é criando métodos para encontrar de forma aproximada a transformada de Laplace das distribuições de cauda longa. Logo se aplica os resultados padronizados da teoria das filas. Para visualizar melhor, a figura seguinte mostra as diferenças entre estes métodos e o método clássico markoviano. A figura 19 mostra as diferenças no tempo médio de espera no tráfego num servidor quando se usa as velhas regras (modelos de Erlang) em comparação com o método de aproximação da transformação (TAM). Tráfego que por natureza mostra um comportamento de auto-semelhança (*fractal*), característica que sobressai no tráfego de rede de computadores. A figura mostra que o tempo médio de espera é substancialmente maior usando as novas regras. É importante mencionar que o

método TAM é base para outro método que é o TRM, método TAM de recursão. O TRM tem versão em primeira ordem e versão em segunda ordem. Outros métodos numéricos de reversão incluem o método de Fourier.

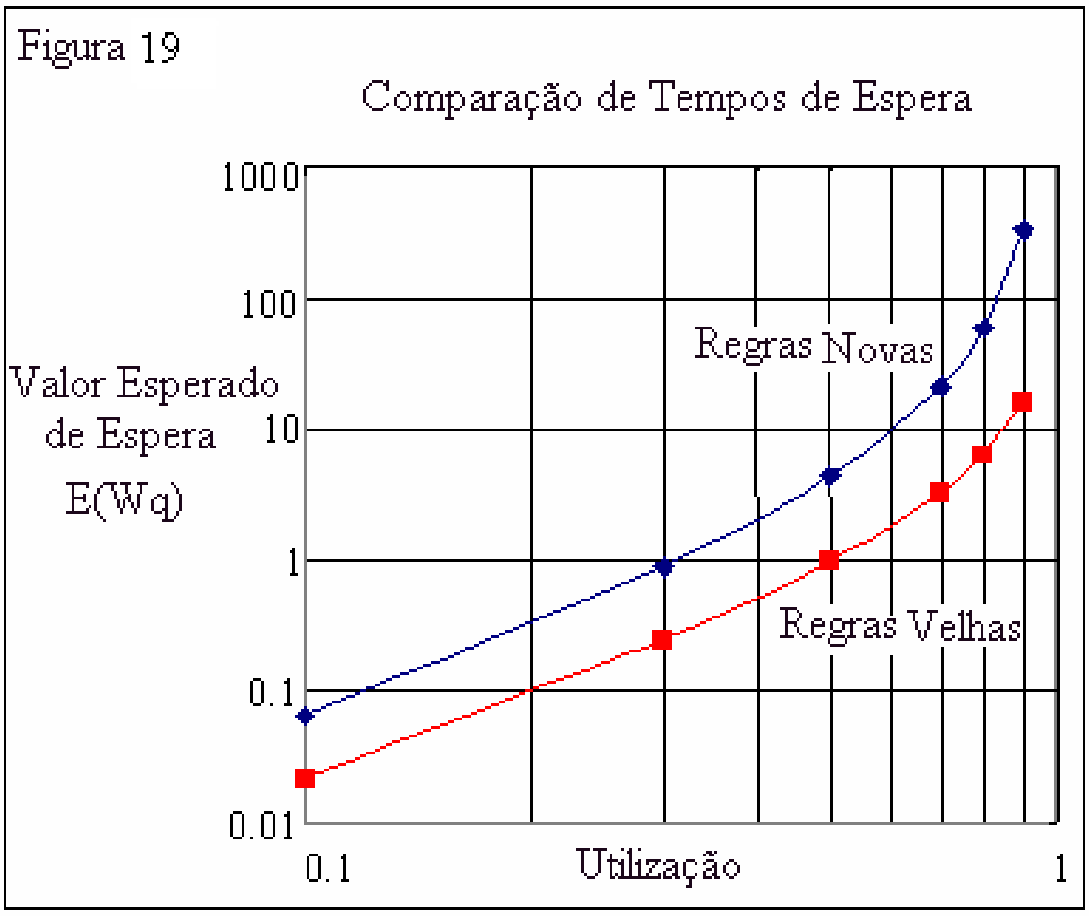


Fig 19: Tempos de Espera do Tráfego num Servidor

- Finalmente o outro método encontra-se na área da Simulação com programas como o Arena e GPSS/H. Análises iniciais mostraram sua potencial, mas precisa de longos tempos de execução num computador.

A figura 20 mostra o tráfego num servidor com a característica clássica *fractal* (auto-semelhança) na coluna direita, comparada com o tráfego de voz (Poisson) na coluna esquerda.O tráfego num servidor é medido na camada de ligação de dados (*link layer*). A linha vertical mostra a medida em pacotes por unidade de tempo. A linha horizontal tem como medida os intervalos de tempo. A figura mostra quatro pares de gráficos onde as medidas verticais e horizontais são incrementadas por um fator de dez em cada par de gráficos, desde o topo até a base da figura. Ou seja, as médias dos intervalos de tempo crescem.

A coluna direita mostra a característica *fractal* (auto-semelhança) do tráfego num servidor comparado com o tráfego de voz (Poisson) mostrado na coluna esquerda da figura. Ou seja, as características mostradas de um tráfego de voz no telefone, que descrevem as chegadas de voz com uma distribuição de Poisson, são diferentes das características do tráfego num Servidor, que não têm uma distribuição de Poisson.

Figura 20 Medida do tráfego digital mostra auto semelhança comparado com o tráfego de voz.

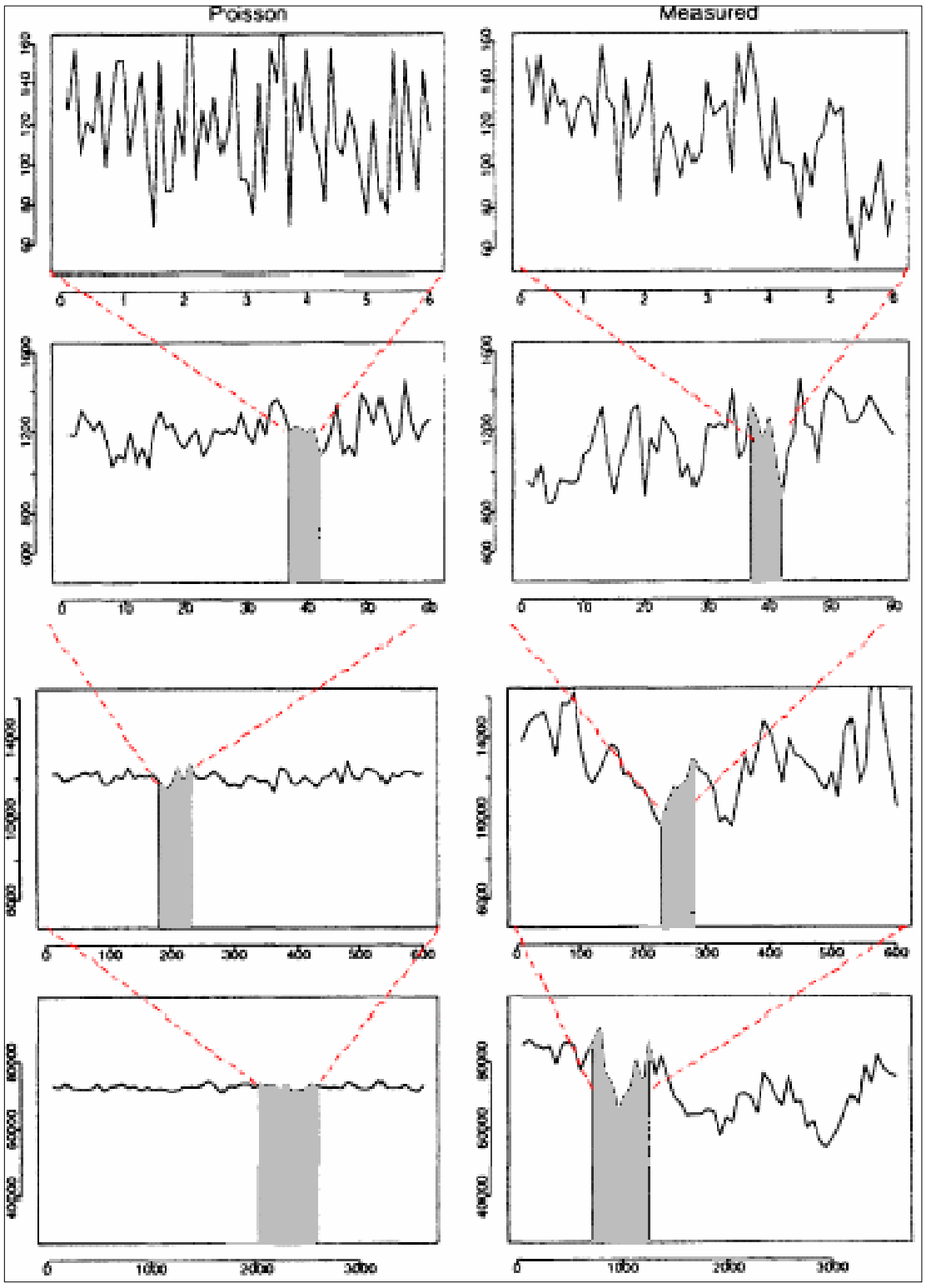


Fig 20: Tráfego de Voz Versus Tráfego num Servidor

Observa-se na figura que o fenômeno de *burstiness* (explosões) no caso do tráfego de voz (Poisson), tem tendência a desaparecer (consegue atingir em média um valor constante) quando os intervalos de tempo são incrementados. Em comparação com o tráfego num Servidor, o fenômeno de *burstiness* (explosões) não muda quando se incrementa os intervalos de tempo. Isto significa que, no caso do tráfego de voz, os picos de tráfego são limitados a ocorrer com freqüência, incluindo sua acentuação, porque a informação do tráfego de voz pode ser induzida a reduzir os efeitos ruins dos picos dentro de um nível menos severo. Esse não é o caso do tráfego num Servidor, porque este não pode ser induzido a reduzir os efeitos ruins dos picos de tráfego, como, por exemplo, armazenar temporariamente a informação (*buffering*). Portanto, o cálculo da probabilidade de perda de pacotes

de informação no tráfego no servidor não pode ser avaliado como no caso do tráfego de voz (telefone) em que sim se pode avaliar a probabilidade de perda de dados.

Finalmente se reforça a idéia de que os gráficos na parte direita da figura mostram ter igual aparência, independentemente do incremento do intervalo de tempo, insinuando ou sinalizando o comportamento de tipo *fractal*. (Cf. Fischer, 2001, P.84)

5. CONCLUSÕES

A teoria das filas é um processo estocástico, onde as atividades de chegada e atendimento estão governadas por uma distribuição de probabilidade. A culminação de uma atividade termina num evento. Um evento tem o poder de mudar o estado (número de unidades no sistema) de um sistema de filas. O tipo de distribuição de probabilidade tem um papel determinante na análise de uma fila de espera.

A distribuição exponencial desempenha um papel fundamental na teoria das filas, na descrição das distribuições de tempo de chegada e atendimento, porque esta suposição permite representar um sistema de filas como um processo Markoviano. Pela mesma razão, distribuições de tipo fase, tais como a distribuição de Erlang (em que o tempo total é dividido em fases individuais com distribuições exponenciais), são muito convenientes. Existem resultados analíticos exatos somente para poucos modelos de sistemas de filas de espera, usando outras suposições, ou seja, processos não Markovianos, como por exemplo o modelo M/G/1.

A teoria das filas gera modelos matemáticos para predizer o comportamento de um determinado congestionamento. Isto permite desenhar sistemas de filas eficientes para obter um ótimo equilíbrio entre o custo de ficar em uma fila de espera, aguardando atendimento, e o custo em providenciar atendimento. (Por exemplo, um empregado esperando receber suas ferramentas de trabalho numa fila gera tempo perdido. Este tempo perdido não pode ser recuperado e isto gera perda de produtividade. O retardamento no atendimento também eleva o custo.) A teoria das filas é uma ferramenta de otimização para ser aplicada a diversos tipos de congestionamento. Mas é fundamental previamente determinar as classes de distribuições de probabilidade que descrevem um

determinado congestionamento. É necessário conhecer também alguns dados como por exemplo: a razão de chegada (λ), razão de atendimento (μ) etc.

Usando o modelo markoviano se analisou o congestionamento do servidor e se obteve resultados que foram estudados objetivando otimizar o desempenho do servidor. O desempenho do servidor depende do valor da razão de chegada (λ), razão de atendimento (μ), e capacidade do *buffer* (N-1). Não se pode obter a razão de atendimento a partir dos dados do servidor. Mas se poderia determinar a razão de atendimento partindo do conhecimento de λ, capacidade do *buffer* e a percentagem de mensagens perdidas. Tomando o mês de Janeiro como referencia se determinou a necessidade que o servidor processe uma mensagem em menos de 3,6 segundos para ter uma perda máxima de 1 em mil mensagens, tendo como dado de entrada a razão de chegada λ = 2,08 mensagens por minuto e a capacidade do *buffer* (N -1 = 3 mensagens). Os valores de perda de mensagens e capacidade do *buffer* podem ser mudados, pois seus valores são assumidos. O único valor real do servidor é a razão de chegada. Por exemplo, supondo perder no máximo uma mensagem num milhão, o tempo de processamento de mensagens teria que ser menor que 3,6 segundos (aproximadamente 2 segundos). Assim, olhando os resultados, se determina que a perda de mensagens no servidor depende do tempo de processamento das mensagens.

Também outras variáveis que estão envolvidas neste tipo de congestionamento são calculadas de forma aproximada num *add-in* (programa escrito em Visual Basic instalado em Excel). E com estes resultados se pode observar a área ou áreas que o servidor precisa melhorar para incrementar seu desempenho ou eficiência.

Por outra parte, nos artigos científicos se estudou o comportamento dos pacotes de informação que chegam a um servidor utilizando o modelo de Interconexão de Sistemas Abertos. É na camada de ligação de dados onde se evidencia o comportamento do tráfego de pacotes que chegam a um servidor. O congestionamento de mensagens num servidor tem comportamento *fractal*. Como resultado se chegou a concluir que um congestionamento gerado pelo excesso de informação que chega a um servidor (tráfego tecnologia *packet switch*) tem diferente comportamento que um congestionamento gerado por exemplo pelo tráfego de voz humana no telefone (tecnologia *circuit switch*). Por tanto, foram

identificadas as distribuições que se adaptam melhor a este tipo de congestionamento. Estas distribuições são as chamadas distribuições de cauda longa. Estas são a distribuição de Pareto, LogNormal, Weibull etc. Estas distribuições de cauda longa não têm fórmulas da transformada de Laplace que sejam úteis para os propósitos de otimizar o congestionamento gerado num servidor. Isto levou aos cientistas a criar métodos para encontrar de forma aproximada os momentos das distribuições de cauda longa. Estes métodos são: o método Comparativo de Transformação (Transformation Matching Method, TMM), o método de Aproximação da Transformação (Transformation Aproximation Method, TAM), o método TAM de Recursão (Transformation Recursive Method, TRM), o método Fourier etc. Estes métodos não foram analisados aqui.

5.1 **Comentários Finais**

Como comentário final é importante mencionar que esta pesquisa tem varias limitações:

- Não foram utilizadas as distribuições de cauda longa no congestionamento do servidor da UECE por razões matemáticas. Por tanto, os métodos aproximados para encontrar a transformada de Laplace das distribuições de cauda longa não são analisados analiticamente porque requerem o conhecimento de matemática avançada para analisar os momentos de cauda longa.

- Também é conveniente dar um enfoque mais profundo no modelo de Interconexão de Sistemas Abertos (*Open Systems Interconnection*), especialmente na camada de ligação de dados (*link layer*), onde o comportamento *fractal* do congestionamento se torna evidente.

- Mesmo que se obtiveram os resultados das variáveis formadas no congestionamento usando a teoria das filas, não se pode dizer que os resultados representam a realidade do congestionamento no servidor. Primeiro porque os cálculos são baseados em suposições. Suposições que são necessárias para simplificar os cálculos. Pois o único dado real do servidor foi a razão de chegada. Segundo porque não foram usadas outras técnicas como por exemplo programas de simulação, os quais aumentam a possibilidade de analisar o desempenho do servidor. Mas simulação não está no escopo deste estudo. Finalmente este tipo de congestionamento atualmente é motivo de novas pesquisas sobre o assunto.

6. BIBLIOGRAFIA

ANDERSON, David e outros. (2005). Cisco Systems. http://www.cisco.com/en/US/products/hw/univgate/ps501/products_implementation_design_guide_chapter09186a00840eaafd.htm. Acesso em: dez 2006

ANDRADE, Eduardo Leopoldino. *Introdução à Pesquisa Operacional*: Métodos e Modelos para a Análise de Decisão. 2 ed. Rio de Janeiro: Livros Técnicos e Científicos, 1998.

BERTSEKAS, Dimitri P.; TSITSIKLIS, John N. *Introduction to Probability.* s.e. Massachusetts: Athena Scientific, 2002.

CHELST, Kenneth e outros. (2006). http://www.hsor.org/index.cfm. Acesso em: jun. 2006

FISCHER, Martin. J., FOWLER, Thomas.B. (2001). *Fractals, Heavy-Tails and the Internet*. http://www.mitretek.org/SigmaSummer2001_2.pdf. Acesso em: jun. 2006

HARVEY, Wagner. M. *Principle of Operations Research*: Applications to Managerial Decisions. Englewood Cliffs, New Jersey: Prentice-Hall Inc, 1969.

HILLIER, Frederick S., LIEBERMAN, Gerald J. *Introduction to Operations Research.* 4 ed. California: Holden-Day Inc, 1986

JENSEN, Paul A. (2006). Texas University; *Operations Research Models and Methods*. http://www.me.utexas.edu/~jensen/ORMM/index.html. Acesso em: mai. 2007

KALINSKY, David. (2005). Embedded Systems: *How to Size Messages Queues*. http://www.embedded.com/showArticle.jhtml?articleID=9901103. Acesso em: jul. 2006

KORILIS, Yannis.(2006). Pennsylvania University; *Networking and Queuing Course*. http://www.seas.upenn.edu/~tcom501/. Acesso em: fev. 2005

KRETCHMAR, James. (2004). *Open Source Network Administration*. http://searchopensource.techtarget.com/tip/1,289483,sid39_gci1130326,00.htm. Acesso em: jun. 2006.

McNICKLE, Donald.(1998). *Queuing for toilets*. http://www.orsoc.org.uk/about/topic/insight/article_orinsight_toilets.htm. Acesso em: fev. 2004.

MEYER, Paul.L. *Probabilidade*: Aplicações à Estatística. 2 ed. Rio de Janeiro: Livros Técnicos e Científicos, 1965

SALIBY, Eduardo. *Repensando a Simulação:* A Amostragem Descritiva. São Paulo: Atlas, 1989

SMOOT, Janet, JACOBS, Bill, FADELY, Michelle. (1997).Virginia Department of Motor Vehicles Queueing Management System. Richmond. http://www.hsor.org/modules.cfm?name=Arm-and-a-Leg. Acesso em: jun. 2006.

WINSTON, Wayne.L. *Operations Research:* Applications and Algorithms. 4 ed. California: BrooksCole Thomson Learning, 2004