"ANÁLISE DE DESEMPENHO DE SUB-SISTEMAS DE COMUNICAÇÃO DE DADOS: APLICAÇÃO DE MODELOS DE MULTIFILAS"

EDUARDO DE VASCONCELLOS

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS (M.Sc.).

APROVADA POR:

Pierre Jean Lavelle
(Presidente)

Daniel A. Menascé

José Manoel Carvalho de Mello

Gerhard Schwarz

RIO DE JANEIRO, RJ - BRASIL

NOVEMBRO DE 1982

VASCONCELLOS, EDUARDO DE

Análise de desempenho de sub-sistemas de comunicação de dados: aplicação de modelos de multifilas (Rio de Janeiro) 1982.

ix, 209 p. 29,7 cm (COPPE-UFRJ, M.Sc., Engenharia de Sistemas e Computação, 1982).

Tese - Univ. Fed. Rio de Janeiro, Fac. de Engenharia.

1. Comunicação de dados. 2. Multifilas. 3. Modelagem. 4. Análise de desempenho. I.COPPE/UFRJ II.Título (série).

CURRICULUM VITAE DO AUTOR

Eduardo de Vasconcellos nasceu a 6 de dezembro de 1950 no Rio de Janèiro, RJ. Em 1972 graduou-se em Engenharia Operacional de Telecomunicações no Instituto Nacional de Telecomunicações, Santa Rita do Sapucai, MG, e em 1978 em Engenharia Eletrônica na Universidade Gama Filho, Rio de Janeiro, RJ. Em 1979 cursou na COPPE as matérias relativas ao mestrado em Engenharia de Sistemas e Computação. No período 1972-1980 trabalhou na Companhia de Pesquisas e Recursos Minerais - CPRM, Siemens S/A, Ecodata Comércio e Indústria Ltda. e Empresa Brasileira de Telecomunicações - EMBRATEL. Trabalha desde 1981 na Itaú Tecnologia S/A - ITAUTEC, onde desenvolve atividades relacionadas com arquitetura, projeto e avaliação de desempenho de redes de computadores, no contexto de sistemas on-line distribuídos.

AGRADECIMENTOS

A Regina Coeli, pelas expressões de apoio, incentivo e inspiração constantes e incondicionais, determinantes nos processos de decisão pelo início e desenvolvimento do curso de mestrado e de realização e consecução desta tese.

Ao Prof. Pierre Jean Lavelle, pela orientação, em particular na fase de identificação de uma área de pesquisa relacionada com problemas práticos,no sentido de obtenção de resultados concretos e utilizáveis a nível profissional.

Ao Prof. Gerhard Schwarz, pela co-orientação,sob a forma de um acompanhamento crítico-construtivo estreito deste meu trabalho, desde o início da pesquisa até a fase de elaboração final do mesmo.

A José Martins de Castro, Erich Jantsch, Kant, Rapoport, Paulo Vilhena de Moraes, Horowitz, Rostropovich,Schumann, Prokofiev, Chopin, Bach, Thomaz Collet, Ravel e Kristina Stobaeus, entre outros, pela criação do ambiente de consciência e serenidade de espírito dentro do qual me foi possível desenvolver, de forma intensa e gratificante, este trabalho.

RESUMO

Nesta tese são estudados aspectos relativos à análise de desempenho de sub-sistemas de comunicação de dados "half-duplex" sob disciplina de "polling". A abordagem adotada consiste em um tratamento quantitativo aproximado baseado na utilização de modelos analíticos de teoria de filas, em particular, modelos de multifilas analisados/disponíveis na literatura especializada.

Em seguimento a etapas de identificação e conceituação da classe de modelos de multifilas, revisão/investigação de tópicos de matemática aplicada relevantes e análise detalhada de alguns modelos de maior potencial de aplicação, são apresentados resultados relativos à utilização e validação de um modelo desenvolvido por J. Sykes, dos Laboratórios Bell, em 1970, na avaliação de desempenho de sub-sistemas de comunicação de dados reais, em operação, do tipo mencionado acima.

Adicionalmente, são apresentados resultados e indicadas extensões que, na opinião do autor, podem vir a constituir temas de interesse para pesquisas adicionais, na direção de um melhor entendimento deste e de uma classe mais geral e abrangente de problemas na área de análise de desempenho de sistemas de comunicação de dados.

ABSTRACT

In this thesis I study topics related to the performance analysis of half-duplex data communication sub-systems under polling discipline. The approach is that of an approximate quantitative treatment, based on the utilization of queueing-theoretic analytical models, particularly multiqueue models, available in the specialized literature.

Following steps on the identification and description of the class of multiqueue models, review/investigation of relevant topics of applied mathematics and detailed analysis of some models with higher potential of application, I present results related to the utilization and validation of a model developed in 1970 by J. Sykes of the Bell Laboratories, in the performance evaluation of actual operating data communications sub-systems of the above mentioned type.

Additionally, results are presented and extensions indicated that, in my opinion, may constitute subjects of interest for further research, towards a better insight into this and a more general and broader class of problems in the field of performance analysis of data communication systems.

INDICE

CAPÍTULO I

INTRODUÇÃO

Problemas relacionados com a análise de desempenho de sistemas de comunicação de dados, em contextos de complexidade diversa, de ambientes de acesso remoto de terminais a sofisticados sistemas on-line de computadores distribuídos, vêm merecendo crescente interesse por parte dos segmentos científico e industrial envolvidos com a teleinformática, segundo abordagens as mais variadas.

O surto de desenvolvimento internacional apresentado na última década pelas redes de comunicação de dados, em particular pela tecnologia de comutação de pacotes, tem motivado esforços substanciais da comunidade acadêmico-científica, inclusive no Brasil, no sentido de encontrar soluções eficientes para problemas de projeto e otimização de sistemas desta natureza. Nesta perspectiva pode ser colocado também, mais recentemente, o assunto redes locais.

Coexistem, entretanto, com o formalismo, a elegância analítico-matemática e a eficiência computacional de soluções propostas para estes problemas, relativamente defasados do cotidiano e da realidade nacionais nesta área, perguntas ainda sem respostas, nem sequer equacionadas, de dirigentes, engenhei

ros e analistas, em diversos níveis de decisão empresarial, relativas a aspectos simples, objetivos e relevantes de sistemas reais, em fase de projeto/implantação ou operação/expansão.

Exemplos de tais questões aparecem principalmente na área de distribuição local, ou seja, em sub-sistemas que interligam terminais a computadores "host", tipicamente através de concentradores, multiplexadores ou configurações multiponto (a denominação área de distribuição local não deve ser confundida com rede local, uma vez que compreende meios convencionais de comunicação remotos; é devida a autores americanos, particularmente àqueles ligados aos Laboratórios Bell da AT&T e a outras concessionárias de serviços de comunicações):

a) qual o impacto causado no tempo de resposta pelo acréscimo de n terminais em uma ou mais linha(s) ou concentrador(es)?

b) qual o desempenho estimado atingível a custos mínimos em termos de velocidades de linhas? qual o efeito de operação a 2 fios x 4 fios?

c) qual o desempenho estimado, e respectiva sensibilidade, face a intensidades e perfis de tráfego esperados (mensagens/tempo, comprimentos de mensagens, protocolos utilizados, etc.)?

d) quantos concentradores/terminais podem ser ligados a uma linha multiponto, a um desempenho acei

tável estabelecido?

e) qual a precisão/confiabilidade das estimativas de desempenho, margens de segurança, regiões críticas de operação e fatores adicionais a serem considerados na tomada de decisões pertinentes?

A lista de questões identificadas mostra-se hoje já bastante extensa, e amplia-se consideravelmente na medida da diversificação de aplicações, configurações, meios/equipamentos de comunicação de dados disponíveis e respectivos custos e tarifas, e da rigidez/forma de especificação dos requisitos de desempenho estabelecidos.

A literatura especializada disponível acusa diversos títulos/autores que sugerem, ou mesmo afirmam, a existência de soluções/respostas, para tais problemas, considerados, por muitos, triviais, uma vez que estes vêm sendo, há algum tempo , objeto de estudos e pesquisas em todo o mundo, por constituirem situações práticas naturais e bastante comuns no âmbito das redes de distribuição local, estágio no qual nos encontramos no Brasil, com algumas exceções no campo de comunicação de computadores (a maioria destas, entretanto, a nível de emulação de terminais, utilizando tecnologia e meios característicos da área de distribuição local).

A pesquisa de literatura inicial realizada indica , entretanto, a predominância de trabalhos relacionados com problemas que o autor denomina, neste contexto, de "idealizados" .

Sem confundir com o exercício de análise de modelos simplificados ou em situações ideais de operação (equilíbrio, ausência de erros, situações limites, moderação de rigor e detalhes), o termo "idealizado" indica graus de distorção, ou ausência total , de isomorfismo com os problemas objetivos, tais que comprometem a utilização dos resultados obtidos, por eventuais usuários interessados nos assuntos-título. Trabalhos de um maior realismo funcional carecem totalmente de suporte de validação prática.

Consequência desta situação, é uma prática generalizada de tomadas de decisões em bases totalmente empíricas ou ad hoc, sem qualquer garantia quanto à qualidade/eficácia das mesmas, principalmente no que toca a aspectos de utilização de recursos, e, consequentemente, de custos, dada uma tendência natural para soluções super-dimensionadas. Exceções constituem alguns ambientes operacionais restritos, de grandes fornecedores, para os quais existem sistemas de simulação, orientados para planejamento a médio e longo prazos.

Dentro deste quadro delineado nos parágrafos anteriores, o posicionamento desta tese consiste em articular e direcionar elementos de teoria disponíveis no campo de modelos analíticos de sistemas de filas, no sentido de obter resultados explícitos, confiáveis e de fácil utilização, aplicáveis a questões práticas do tipo colocadas nos itens a-e anteriores. Filosoficamente, sugere-se um maior comprometimento da pesquisa científica com a realidade de um determinado contexto espaço-tempo, no caso, o nacional/1982.

A amplitude do assunto e a diversidade de métodos/a

bordagens de tratamento admissíveis, conduziram o autor a um direcionamento da pesquisa, segundo motivações e objetivos próprios.

Nas Seções I.1 a I.4 deste Capítulo I, são apresentados, respectivamente, a descrição do problema objeto investigado, as motivações e objetivos correspondentes, a abordagem adotada e trabalhos relacionados, e o esquema de organização desta tese.

SEÇÃO I.1

DESCRIÇÃO DO PROBLEMA INVESTIGADO

Esta tese tem seu desenvolvimento voltado para o tratamento quantitativo de problemas de análise de desempenho de sub-sistemas de comunicação de dados "half-duplex" sob disciplina de "polling".

Particularmente, são considerados sub-sistemas que envolvem uma unidade de controle de terminais ("cluster") IBM do tipo 327X, uma linha de comunicação de dados a 2 ou 4 fios, com velocidade no conjunto {2400, 4800, 9600 bps} (linha privada urbana ou interurbana, especializada ou não-especializada), e uma unidade de controle de comunicações IBM do tipo 3705, operando com disciplina de "polling" segundo protocolo BSC 3 (no início do Capítulo V são indicadas referências relativas a características/detalhes deste ambiente).

A variável de desempenho estudada é a soma dos atrasos e tempos médios de serviço (transmissão) de entrada e saída relacionados com uma transação.

A parcela experimental deste trabalho, que trata da aplicação e validação dos resultados propostos sobre sub-sistemas reais (Capítulo V), toma como base configurações típicas equivalentes encontradas em operação no contexto do Sistema de Agências On-Line, desenvolvido e implantado pela Itaú Tecnologia S/A(*), onde as unidades de controle de terminais são substituídas por concentradores compatíveis, com capacidade atual de 16 terminais.

Por condicionante natural do laboratório de trabalho disponível, é dada ênfase nesta parte experimental a sistemas on-line interativos de alto desempenho, orientados para transações.

Neste trabalho não é considerada a comunicação terminal-concentrador (protocolo particular de uso restrito), constituindo então o concentrador e uma terminação da unidade de controle de comunicações 3705, as fontes de tráfego para a linha de comunicação. Em ambientes estritamente IBM ("cluster"+ termi

---

(*) As referências feitas neste trabalho a equipamentos e/ou a sistemas construídos pela Itaú Tecnologia S/A, não implicam em quaisquer compromissos industriais ou comerciais desta empresa, com relação a especificações técnicas de produtos ou a dados funcionais de desempenho de sistema.

nais IBM), os tempos envolvidos na comunicação "cluster"-terminal são pràticamente desprezíveis, se comparados a outras componentes.

SEÇÃO I.2

MOTIVAÇÕES E OBJETIVOS DA PESQUISA

As motivações iniciais para a pesquisa realizada tiveram suas origens a partir de inclinações acadêmico-profissionais do autor, com relação à teoria de filas e problemas de tráfego/desempenho em redes de dados.

Ainda na fase dos cursos de mestrado (Teleprocessamento I e II) desenvolveram-se discussões, juntamente com a orientação acadêmica, das quais resultou a definição de um tratamento quantitativo, baseado em modelos de filas, do problema de "polling" em redes de distribuição local, tomando-se como ponto de referência a grande quantidade de sistemas deste tipo em operação,no Brasil e em diversos outros países, e a inexistência prática de métodos, técnicas ou modelos que viessem a fornecer respostas rápidas e confiáveis sobre aspectos relevantes do problema (a respeito, ver artigo recente de Data Communications Junho/82, /SUBRN82/).

Um aspecto importante a ressaltar é o fato de que a totalidade dos sistemas monitores de teleprocessamento atuais (mais genericamente, conjuntos de software DC/DB - "data communication/data base") não implementa a medição dos componentes

de tempo de comunicação em seus monitores de desempenho/tempo de resposta, computando apenas os intervalos decorridos entre a recepção de mensagens de consulta, ou transações, e o envio das respostas correspondentes, ou seja, o tempo de processamento da consulta dentro do computador "host", ou seja, no segmento computacional (ver / GELLK82/).

A constatação de tal fato gera a necessidade real de se estimar os tempos/atrasos relativos ao segmento de comunicações, em função de parâmetros como intensidades de tráfego, comprimentos de mensagens de entrada e saída, velocidades das linhas, tempos de "turnaround" em linhas a 2 fios, frequências de "polling", "overhead" das sequências de controle do protocolo, entre outros.

O envolvimento do autor com atividades de análise de desempenho e dimensionamento de sub-sistemas deste tipo, **profissionalmente**, durante os anos 1981-1982, ampliou as motivações e os objetivos iniciais da pesquisa, tornando-os mais abrangentes e concretos, na medida que integrados ao cotidiano profissional.

Adicionalmente, facilidades disponíveis no software de suporte desenvolvido pela Itaú Tecnologia S/A para o gerenciamento do sistema mencionado, permitem a avaliação dos tempos/atrasos no segmento de comunicações (protocolos especiais entre inteligências de terminal/concentrador/monitor), dados estes que, coletados e processados por um sistema de análise estatística acoplado, permitiram a realização de medições, levanta-

tamentos de estatísticas de tráfego e parâmetros relevantes, e de exercícios de validação do modelo aplicado.

Do ponto de vista dos usuários considerados (pesquisadores/engenheiros/analistas), os objetivos consistem em fornecer uma base conceitual sólida e abrangente, que viabilize incursões adicionais na área, e em apresentar, como seguimento natural, resultados de utilização simples e eficiente, e confiáveis, aplicáveis aos sub-sistemas introduzidos na Seção I.1 anterior (ver também Seções I.3 e II.3 quanto a este objetivo).

Adicionalmente, pretende-se que este trabalho constitua contribuições efetivas para os segmentos acadêmico e profissional envolvidos com problemas desta classe, através da metodologia refletida na organização/desenvolvimento do mesmo, do esforço de compatibilização/homogeneização de linguagens e dos resultados práticos fornecidos.

Neste contexto situam-se também os resultados e extensões indicados no Capítulo VI, que, na opinião do autor, podem vir a constituir temas de interesse para pesquisas adicionais, na direção de um melhor entendimento deste e de outros problemas desta natureza.

SEÇÃO I.3

ABORDAGEM E LITERATURA RELACIONADA

Em consonância com colocações e posicionamentos do autor, apresentandos no início deste Capítulo I e também na Se-

ção I.2, a abordagem adotada neste trabalho é a de assimilar e utilizar, dentro do maior grau possível de isomorfismo, modelos de teoria de filas, na solução do problema descrito na Seção I.1 anterior.

É óbvio que tal atitude implica em conhecimentos bastante detalhados de ambos modelo(s) e sistema real, para que se proceda à fase de mapeamento/utilização. A experiência profissional do autor mostra, entretanto, que o funcionamento de sistemas reais deste tipo é relativamente bem conhecido, cabendo então a este trabalho a tarefa de identificar modelos ou classes de modelos, de alto potencial de aplicação, conceituá-los e estudá-los, provendo os subsídios teóricos necessários ao seu entendimento e explorá-los detalhadamente até o nível de mapeamento "ponto-a-ponto" com o problema objeto.

A crítica feita anteriormente aos trabalhos afetos ao assunto "polling", quando então aparece a denominação "idealizados", pode ser ampliada, se se incorporam as seguintes características adicionais, emergentes de uma extensa e cuidadosa pesquisa bibliográfica realizada: hermetismo (de conceitos, terminologia, notação, etc.), esforço computacional de solução (alguns resultados tomam tão somente a forma de equacionamentos, e são, muitas vezes reconhecidamente, de solução extremamente difícil e/ou custosa) e ausência de suporte de validação prática (atinge a totalidade dos trabalhos; alguns comparam resultados teóricos com resultados de simulação, ambos referindo-se ao mesmo sistema idealizado, incorporando aos modelos de simulação alguns aspectos de maior realismo (detalhes) não refletidos pelos

modelos analíticos).

Nesta tese propõe-se, no Capítulo V, a utilização de um modelo analítico de multifilas estudado por J. Sykes, dos Laboratórios Bell, em 1970 /SYKEJ70/. Com relação às deficiências identificadas nos demais modelos, constantes da crítica do parágrafo anterior, tem-se:

a) o modelo de Sykes, através de pequenos ajustes/manipulações de mapeamento (explicitadas pelo autor nas Seções V.1 e V.2), atinge um alto grau de isomorfismo com o problema objeto;

b) o hermetismo do modelo, e de maneira geral, de toda a classe de modelos de multifilas, é quebrado, através da apresentação de material pertinente, nos Capitulos II, III e IV desta tese;

c) o modelo em questão, bem como outros selecionados/indicados para estudos adicionais, apresenta resultados sob forma de expressões analíticas fechadas, de extrema facilidade computacional;

d) o modelo de Sykes é alimentado com dados reais de operação, os resultados fornecidos comparados com valores medidos, e a validação/precisão dos resultados discutida na Seção V.4 desta tese.

Não se pretende aqui elaborar, de forma exaustiva, sobre a escolha da abordagem e de um modelo particular. O inte-

resse e aplicabilidade que vêm encontrando os modelos de teoria de filas, em análise de desempenho de sistemas computacionais e de comunicações, são hoje fatos patentes e indiscutíveis, dispensando, neste ponto, quaisquer referências específicas. A escolha do modelo de Sykes sobre os demais modelos da classe de multifilas, constitui uma das tarefas essenciais deste trabalho, e é conduzida, de forma progressiva, desde o Capítulo II até o Capítulo V desta tese. No Capítulo VI é mencionado o fato da consideração de simulação nesta fase inicial de orientação de abordagem, e os respectivos escôpo e resultados.

Entretanto, delineia-se resumidamente a seguir o processo de pesquisa de literatura sobre o assunto "polling", aqui relatado baseado em algumas referências clássicas disponíveis , alguns comentários, e indicações que conduziram o autor a concentrar suas investigações na classe de modelos de multifilas.

Em /MARTJ72/, Martin (na época, da IBM) escreve aproximadamente 100 páginas sobre o assunto configurações multiponto e o efeito de "polling" nestas. Os modelos propostos são baseados em filas M/G/1 independentes, e resultados numéricos são comparados com resultados de simulações. A incorporação de tempos de "polling" é baseada em estimativas, desenvolvidas por Martin, das probabilidades de "polling" positivo e negativo em sistemas simétricos. A interação de tráfegos bidirecionais, sujeitos a "polling" (entrada) e a "addressing" (saída) não é modelada. Martin não indica quaisquer referências adicionais sobre o assunto. O autor desconhece, a nível de literatura, qualquer experiência publicada de utilização do material apresen

tado em /MARTJ72/.

Em /EVERW72/, Everling (IBM) dedica um capítulo (aproximadamente 20 páginas) de suas notas de aula ao assunto "polling" de terminais, considerando, inicialmente, um esquema do tipo multifila, com tráfego unidirecional: "M terminais transmitindo para uma CPU. Calcular o atraso de entrada". Tal exemplo é o protótipo do que o autor denomina de idealizado, sendo esta a grande desvantagem, inclusive de grande parte dos sistemas de multifilas. Em seguida Everling sugere o tratamento de sistemas de terminais conversacionais utilizando-se o modelo de interferência de máquinas, sistema fechado proposto por Mack , Murphy e Webb /MACKC57a e MACKC57b/ .O autor não considera surpresa que Everling não tenha conseguido resolver o problema,uma vez que esta foi uma das primeiras alternativas consideradas na pesquisa, e a quantidade de dependências estatísticas e de operação não-conservativa a ser introduzida tornaria o trabalho de complexidade equivalente a desenvolver um novo modelo dedicado, e nada trivial, comparado com o estado atual de desenvolvimento de teoria de filas. Nestas notas de aula Everling já menciona autores/trabalhos da área de multifilas, como Leibowitz /LEIBM61/ , Cooper /COOPR69/ e um trabalho na época ainda não publicado de Konheim /KONHA72/ (ver Capítulo II).

Kleinrock apresenta em /KLEIL76/ uma breve seção (4.14) do Capítulo 4, sobre acesso remoto de terminais a computadores, que consiste fundamentalmente em um resumo dos trabalhos realizados por Gaver, baseados em premissas de alto-tráfego, aproximações de difusão e uma série de configurações com ter

minais exclusivos de entrada e saída. O mapeamento com o sistema de interesse nesta tese não é encontrado, e, mais importante, na quase totalidade dos casos, sistemas reais operando com "polling" apresentam valores de tráfego (utilização) bastante baixos (< 30%), invalidando assim a aplicação da aproximação de difusão (ver referências de Gaver em /KLEIL76/), não justificando, na opinião do autor, um aprofundamento nesta direção.

Adicionalmente, Kleinrock comenta ligeiramente sobre os trabalhos relativos a "loops", de Pierce, Farmer e Newhall, Hayes e Sherman, Spragins e Konheim e Meister, incluindo também os trabalhos de Chu sobre multiplexadores estatísticos.

Cabe ressaltar que, embora não diretamente ligado ao assunto objeto desta dissertação, durante todo o processo de pesquisa bibliográfica, realizou-se também um acompanhamento detalhado de toda a literatura relativa a "loops", atualizada até 1981.

Mischa Schwartz, com o Capítulo 12 do livro /SCHWM77/, forneceu os indicadores decisivos para a incursão em um vasto campo da pesquisa operacional, particularmente, a área de modelos de multifilas, em trabalhos teóricos/genéricos ou orientados para o assunto de transmissão de dados e "polling". Tratando do assunto "polling", após comentários sobre "loops" e indicação de referências clássicas (ver Kleinrock acima), Schwartz, baseado no sistema PARS da IBM, explica o funcionamento qualitativo de tais sistemas, e passa a um item sobre análise quantativa, onde, de maneira não muito clara, utiliza simultâneamente (transcreve) resultados de Kaye e Richardson /KAYEA73/, Hayes e

Sherman /HAYEJ72/ e Konheim e Meister /KONHA74/.

Schwartz declara explìcitamente que os resultados a presentados referem-se apenas a tráfego unidirecional e, consequentemente, a atrasos de entrada (cada uma das referências indicadas foi minuciosamente analisada, bem como as respectivas referências secundárias, confirmando o fato), mas, no entendimento do autor, Schwartz super-simplifica a matéria, quando diz que, por razões de simplicidade, considera apenas o atraso de entrada, e que algumas de suas referências incorporam também atrasos de saída, o que pode ser fàcilmente realizado somando-se a cada mensagem de entrada, o comprimento médio da mensagem de saída correspondente.

Foi, entretanto, a partir das referências apontadas em /SCHWM77/, particularmente /HAYEJ72/ e /KONHA74/ (respectivamente, Bell e IBM), que o autor, de posse de outros nomes como Leibowitz, Cooper e Murray e Kuehn, iniciou uma segunda etapa de de pesquisa bibliográfica, mais direcionada, na área dos modelos de multifilas, pesquisando quase uma centena de artigos sobre o assunto, até que as referências mostraram-se recorrentes, e tornaram-se evidentes os trabalhos mais básicos e de maior potencial para exploração. Os resultados desta pesquisa são apresentados no Capítulo II desta dissertação.

SEÇÃO I.4

ORGANIZAÇÃO DA TESE

Complementando a visão da tese fornecida pelo índi-

ce, podem ser colocados, adicionalmente, os seguintes esclarecimentos:

a) a organização e o desenvolvimento desta tese refletem exatamente a metodologia adotada pelo autor, na elaboração da mesma;

b) sintéticamente, admite-se o seguinte mapeamento:

Cap. I: posicionamento, definições, direcionamento.

Cap. II, III e IV: conceituação e aprofundamento seletivos.

Cap. V: contato com a realidade prática, exercícios.

Cap. VI: síntese, extensões (hipóteses e teses!)

O autor se exime de maiores comentários específicos a respeito dos diversos capítulos/seções, uma vez que os títulos no índice pretendem ser auto-explicativos, suficientemente detalhados, e que explicações adicionais de conteúdo são apresentadas ao início de cada capítulo/seção desta tese.

CAPÍTULO II

MODELOS DE SISTEMAS DE MULTIFILAS

Modelos de sistemas de multifilas constituem uma classe de modelos matemáticos em teoria de filas, cujo tratamento não se encontra disponível, até o momento, em livros sobre a matéria, considerando-se aqui, também, os diversos livros que tratam assuntos como avaliação de desempenho, projeto e análise de sistemas de computação e de redes de comunicação de computadores.

Embora os primeiros trabalhos sobre multifilas, identificados pelo autor, datem do início da década de 1950, este assunto vem sendo exclusivamente tratado através de artigos publicados em periódicos especializados, resultados de pesquisas realizadas, principalmente, por pesquisadores de grandes instituições industriais (Laboratórios Bell-AT&T, IBM e NTT) e de algumas universidades americanas e alemãs, neste caso, geralmente , sob a forma de teses de doutorado.

Neste Capítulo II é apresentada uma introdução aos sistemas de multifilas, suficiente para o entendimento da mecânica e da utilização dos modelos disponíveis, bem como para um eventual aprofundamento nesta área.

A Seção II.1 consiste na formulação de um modelo ge

ral de multifilas, incluindo conceituação, explicação e especificação dos principais processos estocásticos envolvidos e medidas de desempenho.

Na Seção II.2 é realizada uma revisão orientada de literatura, que, sem a pretensão de ser exaustiva, pretende ser bastante abrangente e objetiva, no que toca ao relacionamento dos trabalhos/autores considerados com o ambiente de tráfego de dados, objeto deste estudo. Assim, na descrição de alguns processos, são utilizados/introduzidos conceitos de mensagens, tráfego, velocidades de canais de comunicação, comprimento de mensagens, entre outros.

A Seção II.3 conclui com a identificação de classes de modelos de interesse, a serem abordados especificamente no contexto deste estudo, segundo critérios de grau de isomorfismo com o problema-objeto, simplicidade operacional de solução e utilização e potencial de aplicação posterior a problemas correlatos ou a variantes dos mesmos.

Os modelos selecionados serão, então, objeto de estudos detalhados no Capítulo IV, e um modelo específico, o tema de exemplo de aplicação real de multifilas em avaliação de desempenho em comunicação de dados, no Capítulo V.

SEÇÃO II.1

FORMULAÇÃO DE MODELO GERAL DE MULTIFILAS

II.1.1 CONCEITUAÇÃO

O modelo geral de multifilas apresentado nesta seção pode ser conceituado como um super-modelo com a estrutura básica de um sistema de multifilas, qual seja, um conjunto de filas que compartilha numa única unidade de serviço, ao qual são incorporadas generalizações das características de diversos modelos, de forma a permitir a geração, através de reduções, restrições e particularizações, de todos os modelos referenciados na Seção II.2.

Este modelo é representado esquematicamente na Figura II.1, com base na qual são fornecidas descrições dos processos de chegada e de serviço, bem como da disciplina de atendimento, incluindo explicações relativas a terminologia, notações e convenções básicas utilizadas.

Embora sejam encontrados na literatura alguns tratamentos de modelos de multifilas com múltiplas unidades de serviço ou com filas finitas, como, por exemplo, /WHITB75/ e /ARTHE79/, tais modelos não são aqui considerados, por razões de aplicabilidade e de limitação da complexidade. O modelo geral desta seção sintetiza a classe de sistemas abertos de multifilas infinitas mono-atendidos.

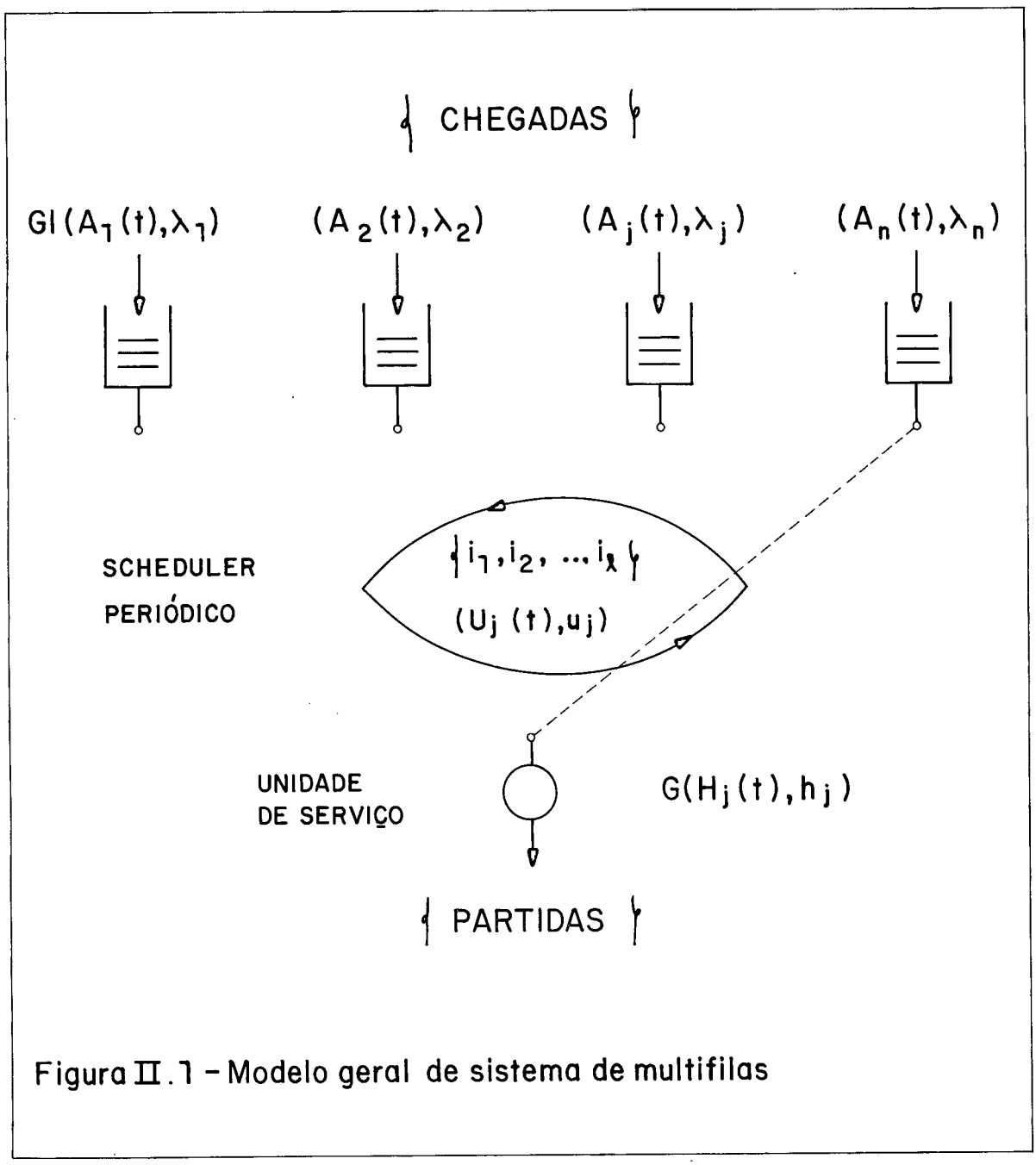


Figura II.1 - Modelo geral de sistema de multifilas

II.1.2 PROCESSOS DE CHEGADA

Itens pertencentes à população da fila j (população infinita, sistema aberto, taxas exógenas) realizam um processo geral e independente (GI) de chegadas, com f.d.a. (função de distribuição acumulada) $A_j(t)=Pr[T_{A_j} \leq t]$, onde T_{A_j} denota a variável aleatória tempo entre chegadas sucessivas à fila j e $\lambda_j=(E[T_{A_j}])^{-1}$ define a taxa de chegadas de itens a esta fila. A

sequência $\{T_{A_j}\}$ descreve, em geral, um processo de renovação standard (ver III.2.3), sendo o processo de Poisson (M) a particularização dominante ($\{T_{A_j}\}$ exponencialmente distribuidos). Outros casos particulares considerados são processos descritos por intervalos entre chegadas independentes e identicamente distribuidos (i.i.d.), segundo distribuições gamma (Γ), Erlang-k (E_k), determinística (D) ou hiper-exponencial-2 (H_2).

São admitidas também chegadas em lotes (grupos), quando então o processo de chegadas é especificado pelo intervalo entre chegadas de lotes, T_{B_j}, e o tamanho aleatório do lote, K_j, $j=1,2,\ldots,n$, dado por sua distribuição de probabilidades $q_{jk}=Pr[K_j=k]$, $k=0,1,2,\ldots$.

A taxa total de chegadas à fila j, λ_j, relaciona-se com a taxa de chegadas de lotes, λ_{B_j}, através de $\lambda_j=\lambda_{B_j} . E[K_j]$. Neste caso são caracterizados processos de renovação compostos (ver III.2.3), dentre os quais destacam-se, novamente, os processos de Poisson compostos ($M^{[X]}$), pela facilidade relativa de tratamento matemático e larga faixa de aplicabilidade.

Em sistemas de manipulação de dados ou tráfego digital, os itens, ou clientes, de uma fila são representados pelas unidades de dados correspondentes, como, por exemplo, bits, bytes, caracteres, blocos, pacotes, mensagens, etc. Neste trabalho é adotada, uniformemente, a unidade mensagem, como uma seqüência, de comprimento variável, de caracteres de comprimento fixo, de forma a obter compatibilização com os modelos estudados, que, conforme apresentado na Seção II.2, consideram, em sua maioria, che

gadas isoladas.

Exceção constituem os modelos da linha de análise operacional, que consideram uma unidade de dados determinística (para fins deste estudo, o caratter),implicando em que a chegada de uma mensagem, ou qualquer outro agregado de caracteres,de comprimento variável, resulte em um processo de chegadas em lote.

Todas as filas possuem espaço de armazenamento ilimitado (filas infinitas), de forma que não há ocorrência de perda "overflow" nos processos de chegada.

Os processos de chegadas considerados são, geralmente, a estado e tempo contínuos, embora os modelos da linha de análise operacional considerem, sistematicamente, processos discretos.

II.1.3 PROCESSOS DE SERVIÇO

As unidades de dados, ou itens, residentes em uma fila j recebem serviço sob a forma de transmissão ou processamento destas pela unidade de serviço, que representa geralmente um meio de transmissão com capacidade de canal finita determinada; em aplicações relacionadas com processamento de dados, sistemas de controle e supervisão, etc., são atribuídas unidades/capacidades análogas.

A capacidade de canal é expressa em bps (bits por segundo), o que determina unívocamente o tempo de transmissão de um caracter, t_c, dado o comprimento, em bits, deste caracter. Assim, a variável aleatória tempo de serviço é descrita pelo comprimento de mensagens em caracteres, que, a menos de um fator de escala (tempo de transmissão de um caracter), equivale à descrição dos tempos de transmissão correspondentes.

Itens da fila j recebem, então, um tempo de serviço aleatório T_{H_j}, com a f.d.a. $H_j(t)=Pr[T_{H_j} \leq t]$ e média $h_j=E[T_{H_j}]$, $j=1,2,\ldots,n$. Em geral, não há restrições quanto à distribuição de T_{H_j}, que pode ser arbitrária (G).

A adoção do caracter como unidade de dados, com tempo de transmissão determinado, para uma dada capacidade de canal, resulta em uma distribuição constante, ou determinística (D), para T_{H_j} (f.d.a = degrau unitário com origem em t_c, onde t_c = tempo de transmissão de um caracter).

Os itens de uma determinada fila j são sempre despachados em ordem de chegada (FIFO ou FCFS), a cada estágio de atendimento a esta fila, conforme a disciplina de atendimento do sistema.

II.1.4 DISCIPLINA DE ATENDIMENTO

A operação periódica da unidade de serviço, ou ci-

clo, é descrita por uma seqüência $\{i_1, i_2, \ldots, i_\ell\}$, onde $i_k \in (1,2,\ldots,n)$ denota o índice da fila que recebe serviço em k-ésima posição no ciclo de comprimento ℓ; se a i_k-ésima fila está vazia, a unidade de serviço transfere-se para a i_{k+1}-ésima fila (módulo ℓ).

O serviço dedicado a uma determinada fila pode ser exaustivo (E) ("exhaustive") quando se despacham todos os itens, inclusive aqueles que chegam durante o período de atendimento, ou limitado (L) ("gated"), quando apenas são despachados aqueles itens presentes na fila no instante do início de atendimento. Em termos de serviço não-exaustivo, considera-se também o serviço limitado-* (L*), onde * indica um número máximo constante de itens a serem despachados em cada estágio de atendimento a uma fila.

Ciclos de atendimento com diferentes freqüências de visitas a determinadas filas em um ciclo, geralmente para compensar desbalanceamentos de carga, caracterizam um atendimento cíclico com prioridades (por exemplo, $n = 4$ e $\ell = 6$, com $\{1,2,1,3,1,4\}$), enquanto seqüências do tipo $\{1,2,\ldots,n\}$, com $\ell=n$ caracterizam um atendimento cíclico ordinário, ou estrito. Estas duas disciplinas são também denominadas, respectivamente, rotação com prioridades e rotação estrita.

Com relação ao movimento, transição, ou transferência da unidade de serviço ao longo do ciclo de atendimento conceituam-se:

a) Tempos de transição nulos - a transferência da

unidade de serviço de uma fila j para a fila (j+1) no ciclo de atendimento realiza-se em tempo zero,e a mesma permanece em estado de repouso,ou estacionária,enquanto todo o sistema estiver vazio,até que uma das filas acuse a chegada de um item.Tal mecânica torna,em realidade,um pouco difusa a noção de atendimento cíclico,ou periódico,embora a abstração de um número infinito de ciclos de duração nula seja matematicamente admitida em casos limites.

Observar que neste tipo de sistema,pressupõe-se que a unidade de serviço receba informações instantâneas sobre as chegadas às diversas filas,o que,fisicamente,sugere um esquema local de interrupção ou sinalização.O autor cogita ainda que este modo de operação,e os modelos correspondentes,possam ser utilizados na obtenção de estimativas (limites inferiores) de atrasos de acesso em sistemas onde os tempos de transição sejam negligíveis quando comparados com os tempos de serviço,principalmente em condições de alto tráfego.

b) Tempos de transição finitos - também denominados tempos de transferência,trânsito,estabelecimento,orientação ou aprendizado,geralmente referidos como "overhead" ou "changeover", correspondem aos tempos de transição da unidade de serviço de uma fila j para a fila (j+1),ou,de maneira mais geral,entre duas quaisquer filas,ao ser completado o estágio de atendimento desta fila j.

O tempo de transição é modelado por uma variável aleatória T_{U_j},com f.d.a. $U_j(t)=Pr[T_{U_j} \leq t]$ e média $u_j=E[T_{U_j}]$, satisfazendo a condição $0 < u_j < \infty$, $j=1,2,\ldots,n$. Estes tempos de

transição são geralmente considerados variáveis aleatórias i.i.d. e, em alguns casos, constantes, para fins de simplificação de análise/expressões.

Uma característica fundamental de sistemas com tempos de transição finitos é a regra prevista, pela disciplina de atendimento, para a posição da unidade de serviço quando todas as filas estão vazias:

- permanecer estacionária, até que uma fila acuse a chegada de um item ao sistema, ou
- continuamente transicionar entre as filas, conforme o ciclo estabelecido, independentemente do estado do sistema e das diversas filas, atendendo as filas não-vazias segundo serviço (E), (L) ou (L*).

Observar, neste ponto, que o segundo modo de operação descrito corresponde, basicamente, ao conceito de operação de sistemas de comunicação/processamento de dados e de controle de dispositivos, que trabalham com controle de acesso regido por disciplinas de "polling" ou "scanning". A formulação genérica deste tempo de transferência, modelando períodos de não-utilização da unidade de serviço com despacho de tráfego disponível no sistema (sistema não-conservativo), permite a associação destes modelos a métodos de controle de acesso centralizados, do tipo "roll-call polling" ou "scanning", ou distribuídos, do tipo "bus-polling" (ver /SCHWM77,pp.265-267/).

II.1.5 MEDIDAS DE DESEMPENHO

O objetivo básico dos estudos de análise e simulação realizados sobre modelos de multifilas consiste na determinação de indicadores de desempenho para os sistemas em estudo, sendo consideradas, universalmente, as variáveis aleatórias que descrevem:

- tempo de espera ("waiting time") ou atraso de espera ("waiting delay"), em cada fila;
- tempo no sistema ("system time") ou atraso no sistema/fila ("system/queueing delay"), em cada fila;
- comprimento/tamanho da fila ("queue length/size"), para cada fila.

São ainda considerados o tempo de ciclo ("cycle time") e a utilização da unidade de serviço ("server utilization"). Identifica-se também interesse no estudo do processo de saída, ou de partidas ("output/departure process"), com vistas à análise de sistemas mais complexos (redes), que envolvam sub-sistemas de multifilas.

Idealmente, uma caracterização completa, e desejável, das variáveis aleatórias mencionadas, seria fornecida por uma função de distribuição, o que se mostra, geralmente, de obtenção bastante dificil. A alternativa dominante consiste, então, na obtenção de soluções em domínios transformados complexos (Laplace ou Laplace-Stieltjes, funções geratrizes, transfor

madas Z), as quais, ainda que satisfeitas as condições de existência e unicidade da transformada inversa, apresentam, na maioria dos casos, dificuldades consideráveis, ou mesmo impossibilidade prática, no processo de inversão, inclusive numérico. Buscam-se, então, caracterizações parciais, através de momentos de primeira a terceira ordem, fornecidos por soluções transformadas. Em alguns casos são obtidos apenas valores médios de determinadas variáveis, através de raciocínios/manipulações simples e de propriedades do operador de valor esperado.

SEÇÃO II.2

REVISÃO DE LITERATURA E CLASSIFICAÇÃO

Com base na pesquisa bibliográfica realizada, que abrange diversos periódicos especializados nas áreas de comunicações, computação, pesquisa operacional e matemática aplicada, além de livros, relatórios e teses de doutorado publicados por diversas universidades, é apresentada a seguir uma revisão de literatura sobre a classe de modelos estudada, onde são mostradas as relações de grupos de trabalhos de vários autores, em diferentes épocas , com o modelo geral formulado na seção anterior.

Pretende-se que esta revisão seja bastante completa, no sentido de fornecer um quadro dos esforços e resultados relevantes que caracterizam um período de estudo do assunto de aproximadamente 30 anos, e de colocar à disposição de interessados subsídios para estudos mais específicos.

J.C.Tanner, em 1953, propõe e analisa um modelo com duas filas para aplicação em problemas de tráfego em interseções e trechos estreitos de rodovias sujeitos a tráfego bidirecional /TANNJ53/. Este trabalho aparece na literatura como uma das primeiras formulações conhecidas de um modelo de multifilas. O modelo consiste em duas filas do tipo M/G/* (o asterisco,como número de unidades de serviço, indica a existência de uma única, compartilhada pelas n filas do modelo) e Tanner sugere três variações na disciplina de atendimento, de tratamento bastante difícil,chegando a resultados reconhecidamente insatisfatórios e concluindo com situações limite fisicamente irrealizáveis, como por exemplo,tempos de serviço nulos.Interessante notar que Tanner já emprega em sua análise o conceito de cadeias de Markov imersas (CMI) (ver III.3.1),a que denomina pontos de regeneração,referenciando o, na época, recente trabalho de Kendall /KENDD51/. Não são fornecidos aqui maiores detalhes quanto a este modelo, uma vez que esta menção pretende ser apenas histórico-referencial.

Durante um período de aproximadamente 10 anos são publicados diversos trabalhos na linha de Tanner,tratando modelos de duas filas,com aplicações específicas em controle de tráfego.

Em 1964, encontra-se Darroch, Newell e Morris analisando um sistema de duas filas M/G/*, com tempos de transição finitos e uma constante adicional de temporização ("headway"),para aplicação em um sistema de sinais de trânsito atuados por veículos, em uma interseção /DARRJ64/.

Referências sobre estes modelos/aplicações específicos podem ser encontradas, tanto entre referências secundárias de trabalhos como /DARRJ64/, como entre as referências indicadas por muitos dos trabalhos a serem ainda comentados nesta seção.

No que toca à associação de modelos de multifilas a problemas específicos de comunicação de dados, encontra-se, em 1961, M.A.Leibowitz, da IBM, motivado pelo problema de um sistema de comunicações multiterminal com "polling", formulando um modelo de multifilas simétrico (estatísticas de tráfego idênticas em todas as filas), com N filas M/G/*, serviço limitado (L), atendimento cíclico ordinário e tempos de transição finitos /LEIBM61/.

O sistema não é resolvido, em termos das medidas de desempenho universais, uma vez que a preocupação explícita do autor é com a formulação de uma premissa de independência entre as filas, que tem o objetivo de evitar a descrição de estado completa do sistema, qual seja, a especificação de probabilidades conjuntas de um vetor (N+1)-dimensional de variáveis aleatórias (estado de cada uma das N filas e posição corrente da unidade de serviço), e o respectivo tratamento, que geralmente implica na solução de um número significativo de sistemas de equações .

A proposta de Leibowitz baseia-se em aproximações utilizadas em física atômica, em particular, no conceito de distribuição de probabilidade auto-consistente, para definir as probabilidades de estado estacionárias $\{p_n\}$ de cada fila, nos instantes de chegada da unidade de serviço (ver Seção IV.1). Será

visto que muitos dos tratamentos aproximados fundamentam-se nesta premissa de independência de Leibowitz.

Em um trabalho complementar /LEIBM62/, reforça alguns conceitos, estabelecendo relações e diferenças entre estes e alguns elementos de análise disponíveis em teoria de filas e probabilidades, e introduz os conceitos de tempo de varredura , ou ciclo de "polling", e de eficiência, ou utilização,fornecendo expressões para o valor médio do ciclo, a utilização,bem como uma expressão aproximada para o valor médio do tempo de espera na fila. Em /LEIBM68/, Leibowitz, em artigo não-técnico, explica, de modo bastante acessível, idéias correntes em teoria de filas, bem como o modelo de multifilas e o princípio da independência.

Na década de 60 ocorre um estímulo ao desenvolvimento de modelos de multifilas, na área de pesquisas em teoria de filas, em conexão com o interesse no estudo de classes e prioridades, motivado por aplicações industriais e de engenharia de produção, e também por recentes aplicações identificadas em sistemas de computação.

Neste período, Avi-Itzhak, Maxwell e Miller /AVIIB65/, Takács /TAKAL68/ e outros (ver referências dos trabalhos citados), analisam modelos assimétricos com duas filas M/G/*,serviço exaustivo e tempos de transição nulos, com o objetivo de comparar, em desempenho, esquemas convencionais de prioridades (tipo FIFO e HOL) com uma disciplina de atendimento baseada em serviço exaustivo denominada, por Avi-Itzhak, de prioridades al-

ternantes, com a seguinte mecânica: itens de classe i, (i=1,2), têm prioridade sobre itens de classe j, (j=1,2; j≠i), enquanto existir um item de classe i em serviço; quando a unidade de serviço está em repouso (sistema vazio), o primeiro item que chega recebe serviço e adquire o direito de prioridade para a sua respectiva classe. Os itens desta classe i são então despachados em ordem de chegada, até o término do serviço exaustivo,quando então é transferida a prioridade para a classe j, caso exista algum item desta classe no sistema.

O método de análise utilizado em /AVIIB65/ baseia-se em uma abordagem proposta por Cobham /COBHA54/, e já anteriormente empregada por Avi-Itzhak /AVIIB63/, e que consiste , essencialmente, em argumentações e manipulações com valores médios. Em /AVIIB65/ este método é extendido de forma a fornecer a f.g.m. (função geratriz de momentos) do tempo de espera, em regime de equilíbrio.

Em /TAKAL68/, Takács analisa o mesmo modelo de /AVIIB65/, citando explícitamente o autor, a quem atribui o uso de métodos intuitivos, em substituição aos quais, propõe uma solução mais simples e rigorosa, baseada no tratamento de uma cadeia de Markov imersa definida nos instantes de partida.

Os resultados obtidos por Takács e Avi-Itzhak para o tempo médio de espera são idênticos,entretanto, Takács fornece adicionalmente o segundo momento do tempo de espera. É importante destacar, entre estes dois trabalhos, a seguinte diferença de nomenclatura: "queueing time" (tempo no sistema, inclu

indo serviço), até então utilizado também para designar o tempo de espera na fila, a critério de alguns autores, é claramente diferenciado de "waiting time" (tempo de espera na fila, não incluindo serviço). Esta distinção é fundamental na comparação dos resultados destes dois trabalhos.

No período de 1969 a 1972, verifica-se a publicação de uma série de trabalhos sobre modelos de multifilas, orientados para aplicações em comunicação de dados, tráfego digital e sistemas de comutação, resultantes de atividades de pesquisa aplicada realizadas por grandes organizações do setor de telecomunicações, como os Laboratórios Bell, da AT&T americana, e os laboratórios de pesquisa da NTT japonesa.

Cooper e Murray /COOPR69, COOPR70/ analisam modelos assimétricos com N filas M/G/*, serviços exaustivo e limitado (L) e tempos de transição nulos, na procura de soluções para problemas relacionados com o Sistema de Comutação Eletrônico No. 1 ESS da Bell, acesso a computadores centrais em tempo compartilhado e aplicações afins. Observar que a nomenclatura de filas cíclicas, gerada pelo conceito de serviço cíclico, utilizada nestes e em trabalhos subseqüentes, não deve ser confundida com a noção correntemente encontrada em teoria de filas e, particularmente,em avaliação de desempenho, de sistemas de filas cíclicas, que descreve sistemas de filas simples conectadas em série, com realimentação da saída da última fila para a entrada da primeira (ver, por exemplo, /KOBAH78/.

O método utilizado baseia-se também no conceito de

cadeias de Markov imersas, e os resultados obtidos, em sua maioria relativos ao serviço exaustivo, incluem transformadas de Laplace-Stieltjes do tempo de ciclo e do tempo de espera, bem como valores médios destas variáveis e do número de itens em uma fila nos instantes de início de serviço a esta fila. O cerne do trabalho consiste na derivação de um conjunto de equações funcionais para as f.g.p!s (função geratriz de probabilidades) indexadas do estado do sistema, que, resolvidas por um método iterativo, podem ser utilizadas para obter expressões explícitas para valores médios de algumas variáveis de interesse. As transformadas obtidas são expressas em forma adequada para inversão numérica computacional, e o cálculo do valor médio do tempo de espera em uma dada fila requer a solução numérica de N(N+1) equações lineares. A formulação para serviço limitado (L) é apenas iniciada, e indicado desenvolvimento análogo ao caso de serviço exaustivo.

Sykes, também dos Laboratórios Bell , estudou em 1969/1970 um modelo assimétrico com duas filas M/G/*, serviço exaustivo, tempos de transição finitos e transições contínuas entre as filas, em caso de sistem vazio. Em /SYKEJ69a/ o objetivo era a obtenção de um modelo analítico para interligação "half-duplex" de dois computadores, onde os tempos de transição finitos representam os tempos de reversão na direção de transmissão. O método utilizado baseia-se na premissa de independência de Leibowitz /LEIBM61/, e fornece resultados aproximados para o atraso médio de uma mensagem aguardando transmissão em fila, e o espaço médio de memória necessário para armazenamento de mensagens em cada computador.

Em /SYKEJ69b/, Sykes utiliza os resultados acima,juntamente com um modelo para conversão de tráfego de voz em volume de tráfego de dados e, a partir daí, em números de terminais e canais de comunicação, na análise de desempenho de um sistema de consultas, do tipo utilizado por companhias aéreas.

O artigo /SYKEJ70/ foi em realidade, submetido para publicação em maio de 1968 e corresponde à primeira análise realizada por Sykes do modelo apresentado em /SYKEJ69a/,que já consiste em uma aplicação dos resultados obtidos neste primeiro trabalho, elaborado por volta do início de 1968.

No período 1970-1972, Hashida, da NTT, orientado para o problema de sistemas multiterminais controlados por disciplina de "polling", e motivado pelo modelo aproximado proposto por Leibowitz e por tentativas de análise aproximada realizadas pelo próprio Hashida nos laboratórios da NTT, publicou,entre outros trabalhos, /HASHO70/, /HASHO72a/ e /HASHO72b/, de cujas referências constam outros trabalhos correlatos, de menor relevância em termos de resultados.

Em /HASHO70/ é realizada uma análise exata de um modelo assimétrico, com N filas M/G/*, serviço limitado (L), disciplina de atendimento estritamente cíclica, tempos de transição finitos com distribuição arbitrária e movimento contínuo da unidade de serviço em condição de sistema vazio. Através de definições de cadeias de Markov imersas nos instantes de chegada e de partida da unidade de serviço a uma determinada fila, Hashida obtém uma expressão funcional recursiva para a função geratriz

indexada da distribuição do comprimento da fila em estado de equilíbrio, cuja forma explícita não pode ser derivada. É também obtido um conjunto de equações simultâneas de dimensão N^3, cuja solução permite, em princípio, a computação do tempo médio de espera na fila, o que se mostra, entretanto, bastante trabalhoso.

Hashida ilustra então a utilização de seus resultados através de particularizações para: a) sistemas simétricos e b) sistemas com duas filas. As expressões obtidas são ainda,com raras exceções, relativamente complicadas para cálculo,uma vez que dependem, em sua maioria, de cálculos iterativos de funções auxiliares e intermediárias definidas. São também feitas comparações com resultados aproximados de análise baseada no princípio de independência de Leibowitz.

Em /HASHO72a/ são analisados ambos os serviços exaustivo e limitado (L), numa extensão, e mesmo duplicação de conteúdo de /HASHO70/, no que toca a modelos com serviço limitado. As características dos modelos, os métodos de análise e a natureza dos resultados (baixa aplicabilidade !) são idênticos aos apresentados no trabalho anterior. Hashida mostra-se,em ambos os trabalhos, bastante rigoroso no tratamento matemático,incluindo em apêndices provas de teoremas, condições de existência de estado de equilíbrio, etc.

Em /HASHO72b/ é apresentada a aplicação de um modelo de multifilas a problemas relacionados com capacidades de linhas de unidades de controle de comunicação (CCU).As principais

diferenças com relação aos modelos previamente analisados por Hashida são: filas finitas ("buffers" com duas posições) e serviço limitado (L^1). A unidade de dados adotada é o caracter, e o atendimento cíclico corresponde ao processo de varredura dos "buffers" das diversas linhas conectadas à CCU; o serviço consiste na transferência destes caracteres para a unidade de memória, através de um canal multiplexador (MXC). Hashida estuda o processo de overflow, o número de linhas e velocidades comportadas, o efeito de retransmissões e o atraso aproximado introduzido no canal, em caso de conexão de múltiplas CCU's. Para evitar o que seria o tratamento de um sistema "tandem" (CCU+MXC), é utilizado um modelo de fontes finitas para o MXC, onde cada CCU constitui uma fonte. O tempo de serviço da CCU inclui então esta componente de atraso estimado. A análise do modelo com serviço limitado (L^1) é efetuada de forma aproximada, a partir do princípio de independência de Leibowitz.

Ainda neste período de início dos anos 70, Eisenberg, do Sistema Bell, contribuiu para o estudo dos sistemas de multifilas com /EISEM71/ e /EISEM72/, ambos baseados em tese de doutorado do autor, apresentada em setembro de 1967 ao M.I.T. /EISEM67/ e publicada externamente em abril de 1968 pelo Centro de Pesquisa Operacional daquela universidade como Relatório Técnico Nº 35, sob financiamento de instituições militares americanas /EISEM68/.

Em /EISEM67, (EISEM68)/, Eisenberg trata, com graus de detalhamento e preocupação didática excelentes, o modelo assimétrico com duas filas M/G/* , serviço exaustivo, com tempos

de transição finitos, prioridades alternantes e estacionariedade da unidade de serviço em condições de sistema vazio. São também mostradas comparações com atendimento segundo esquema de prioridades estritas (análogo a uma fila simples com duas classes de itens) e discutidos aspectos de seleção de uma disciplina ótima. Em um apêndice desta sua tese, Eisenberg delineia algumas proposições aproximadas para um sistema com N filas e tempos de serviço nulos (!), de reduzido valor prático, expressando,entretanto, intenções explícitas quanto ao desenvolvimento e aplicação da técnica de análise a problemas de natureza mais geral,por exemplo, sistemas com tempos de serviço arbitrariamente distribuídos.

/EISEM71/ constitui, segundo menção do próprio Eisenberg, um extrato de sua tese de 1967, onde procura resumir , em artigo de 16 páginas, os principais conceitos e resultados relativos ao sistema de duas filas M/G/* mencionado acima,particularmente no que concerne aos esquemas de prioridades.

Em /EISEM72/, Eisenberg analisa um modelo assimétrico com N filas M/G/*, serviços exaustivo e limitado (L) (denominados, interessantemente, "come right in" e "please wait"), disciplina de atendimento periódico arbitrário, tempos de transição não-nulos e movimento contínuo da unidade de serviço com o sistema vazio. Este último item diferencia o modelo em questão dos demais analisados/publicados por Eisenberg, fato este que, aparentemente, não vem sendo notado e devidamente indicado, por uma série de autores,em suas referências, comparações e atribuições de resultados.

Eisenberg parte da definição de quatro cadeias de Markov imersas e obtém relações em forma funcional, entre as f.g.p. associadas, considerando, inicialmente, serviço exaustivo. A obtenção de momentos do tempo de espera e do tempo entre visitas requer diferenciações de 1ª e 2ª ordens destas expressões, resultando em sistemas de (N-1) equações simultâneas no caso de valores médios, e em sistemas de N(N-1)/2 equações no caso de momentos de segunda ordem.

O trabalho é concluído com resultados particulares para o caso N=2, com serviço cíclico ordinário, e observações quanto a uma aproximação que poderia ser computacionalmente mais eficiente, que corresponde, em essência, ao princípio de independência de Leibowitz. É ainda indicado um método para análise do mesmo modelo, com serviço limitado (L), que consiste basicamente, em uma expansão do vetor de estado que possibilita, para a fila em serviço, a contagem dos itens que serão despachados no ciclo corrente, bem como daqueles que deverão esperar pelo próximo ciclo. A partir desta nova descrição de estado, Eisenberg indica o desenvolvimento, de maneira análoga ao realizado para o serviço exaustivo.

Posteriormente, em 1979, Eisenberg formula um modelo de duas filas com serviço alternante, baseado ainda am sistemas de multifilas, em uma variação que demonstra uma analogia de conceituação com o modelo de prioridades alternantes /EISEM79/. Trata-se de um sistema com duas filas M/G/* , serviço limitado (L^1), disciplina de atendimento cíclica e tempos de transição nulos. A técnica de análise utilizada baseia-se ainda em cadei-

as de Markov imersas e nas f.g.p.'s correspondentes,até a obtenção de uma equação funcional que, mediante transformação em equação integral em um domínio complexo, permite a obtenção de uma solução para o número médio de itens em cada fila,a qual conduz, por relações simples, a expressões para os tempos médios de espera em cada fila. Este método de análise é,aparentemente, inédito nesta área e mostra relativas dificuldades matemáticas, uma vez que envolve considerável quantidade de conceitos em funções de variáveis complexas, como mapeamentos conformes, analiticidade, integrais de contorno e "kernels", entre outros.

Na primeira metade dos anos 70, encontram-se diversos trabalhos realizados por Konheim e Meister, da IBM, orientados para a avaliação de desempenho de sistemas multiterminais com topologias em anel e multiponto, desenvolvidos segundo métodos de análise operacional.

/KONHA74/ representa a contribuição fundamental destes dois autores para a área de análise de sistemas de multifilas, consistindo de um tratamento matemático bastante rigoroso do problema de linhas multiponto com disciplina de "polling",ainda que bastante idealizado, conforme colocado no Capítulo I. É analisado um modelo simétrico com N filas, processos de chegada gerais, do tipo renovação (geralmente compostos, uma vez que a unidade de dados adotada é o caracter, nos casos de mensagens de comprimento variável) serviço exaustivo, tempos de transição não-nulos e atendimento cíclico, com movimento contínuo da unidade de serviço, independentemente do estado global do sistema.

Os resultados obtidos são expressões fechadas para o valor esperado e a variância do tempo de ciclo e valores esperados do comprimento de fila e atraso de espera, em qualquer fila, uma vez que o sistema é simétrico. Observa-se também que parte deste trabalho (definições iniciais, resultados parciais e desenvolvimentos resumidos) já havia sido publicada em 1971 /KONHA71/.

Em 1975, Halfin, dos Laboratórios Bell, motivado por problemas associados com o Sistema de Comutação Crossbar No.5 , formulou e analisou uma classe especial de modelos de multifilas, com ênfase em aspectos de eficiência computacional, propondo e validando por simulação um método aproximado de análise,baseado em uma redução da dimensão do vetor de estado, geralmente (N+1), o que já ocasionava custos computacionais (tempo e espaço) na época não aceitáveis, na solução de sistemas onde $N \gtrsim 15$. Devido a uma série de especializações do modelo de Halfin, tais como serviço em ordem aleatória, retorno de itens despachados à fila, etc., não serão aqui apresentados sua descrição detalhada ou resultados. A menção a este trabalho pretende, essencialmente, chamar atenção para preocupações desta natureza e recomendar o estudo deste método /HALFS75/, que gera resultados bastante próximos dos esperados, segundo comparações, realizadas por Halfin, com resultados de simulações, em uma larga faixa de parâmetros e condições de carga dos sistemas estudados.

Em 1979, P.J.Kuehn, da Universidade de Siegen, Alemanha, publicou resultados de análise aproximada de sistemas assimétricos com N filas $M^{[X]}/G/*$ (chegadas em lotes), atendimento cíclico (com e sem prioridades), serviço não-exaustivo, par-

ticularmente o caso limite (L^1), tempos de transição finitos e movimento contínuo da unidade de serviço /KUEHP79b/. Resultados analíticos são validados por simulações exaustivas, abrangendo um número de estudos sobre o desempenho do sistema, em relação às propriedades estatísticas do tráfego oferecido. Referências secundárias indicam estudos anteriores de Kuehn, orientados para sistemas em tempo real e sistemas de comutação computadorizados.

O método de análise utilizado por Kuehn neste trabalho baseia-se em uma extensão de conceitos introduzidos por Hashida (ver /KUEHP80a/), e fornece resultados no domínio transformado sobre as probabilidades de estado, a distribuição de atrasos e os tempos médios de espera, em cada fila. É também formulado um critério de estabilidade para o caso geral de sistemas de multifilas do tipo GI/G/* , em regime de atendimento cíclico com prioridades.

Em continuação ao método de decomposição por ele desenvolvido e apresentado em /KUEHP79a/, Kuehn vem trabalhando em uma linha de aplicação de métodos de decomposição ao problema de análise de sistemas de multifilas mais complexos, como se encontra em /KUEHP79c/, onde são tratadas estruturas hierárquicas de multifilas com controle centralizado, pertinentes a sistemas de comutação de voz e dados digitais. Ainda nesta linha, em /KUEHP80b/ são combinados conceitos e resultados de análises anteriores, na avaliação de desempenho de protocolos do tipo ARQ em sistemas hierárquicos controlados por disciplina de "polling", também idealizados.

Considera-se que esta orientação possa ser ainda bastante explorada, no sentido de se obter modelos e ferramentas para avaliação de sistemas multiponto multiníveis e de redes complexas de filas que incluam sub-sistemas de multifilas mono-ou multiníveis.

Em 1980, G.B.Swartz, do Monmouth College (New Jersey, USA) publica um resumo /SWARG80/ de sua tese de doutorado na Universidade de New York, em 1977 /SWARG77/. O trabalho de Swartz consiste em uma generalização da análise de Konheim e Meister /KONHA74/, extendida para modelos assimétricos, e contou com orientação acadêmica do próprio Konheim. A linha de análise adotada é a análise operacional, acompanhada de graus de rigor e formalismo matemáticos que tornam o trabalho bastante hermético, não só quanto ao seu desenvolvimento, como quanto à utilização dos resultados obtidos.

O objetivo é fornecer relações entre a carga oferecida ao sistema, a seqüência de "polling" e o desempenho,em termos de atrasos e comprimentos de filas. Resultados finais são fornecidos sob a forma de valores esperados e variâncias das variáveis de desempenho, entretanto, o cálculo destes valores requer a avaliação computacional de uma série de relações recursivas obtidas no desenvolvimento do trabalho, relacionadas com variância de uma distribuição invariante auxiliar, definida por Konheim para processos de chegada idênticos. Swartz obtém expressões para as variâncias indexadas desta distribuição invariante para cada fila, em função de sua posição no ciclo de "polling", e apresenta também considerações sobre seqüências ótimas de "polling",em

termos de minimização dos efeitos de segunda ordem no desempenho.

Em 1981, L.F.M. de Moraes, brasileiro em doutoramento na UCLA, apresenta sua tese de doutorado /MORAL81/, onde são abordados, entre outros, tópicos de multifilas assimétricas, a tempo discreto, no contexto de diversas disciplinas de controle de acesso a canais de comunicação de dados.

Os resultados principais apresentados em /MORAL81/, estão relacionados com "polling" em redes de comunicações, incluindo as disciplinas de serviço exaustiva e limitadas, bem como um esquema de "polling" como prioridades posicionais. São ainda obtidos resultados relativos à distribuição de tempos de espera de mensagens em esquemas TDMA, e proposto um esquema TPAC ("tree probing random access") destinado a suprir certas deficiências notadas em disciplinas de "polling" em ambientes de muito baixo tráfego.

Em sua análise dos esquemas de "polling" exaustivo e limitado, Moraes considera um canal de comunicação síncrono, atendendo a um conjunto de M terminais assimétricos, com processos de chegada gerais e independentes, a tempo discreto.

Entretanto, os conjuntos de equações obtidos para a determinação de tempos médios de espera de mensagens, nos dois casos, não são passíveis de solução em forma fechada, implicando assim na solução numérica de um conjunto de M^3 equações lineares.

Nos casos de se considerar simetria, e adicionalmente processos de chegada do tipo Poisson, resultados são disponíveis sob a forma de expressões analíticas fechadas.

SEÇÃO II.3

IDENTIFICAÇÃO DE CLASSES DE MODELOS DE INTERESSE

O objetivo de obtenção de um modelo para análise de desempenho de sub-sistemas de comunicação de dados,conforme colocado no Capítulo I, e o material apresentado nas Seções II.1 e II.2 anteriores, descrevendo os modelos de multifilas disponíveis e os respectivos resultados, conduzem à escolha de alguns modelos de maior potencial para concentração dos estudos.

Nesta seção é delineado o desenvolvimento do processo de identificação destes modelos, através de considerações e justificativas conceituais e práticas.

A quase totalidade dos modelos disponíveis considera filas do tipo M/G/*, o que corresponde ao estado atual do desenvolvimento de teoria de filas, em termos de resultados utilizáveis. Processos de chegada do tipo Poisson mostram-se bastante adequados no estudo de tráfego de dados, enquanto serviço geral, sem conhecimento mesmo das distribuições envolvidas, constitui uma situação bastante confortável para o engenheiro/analista de sistemas, que geralmente dispõe de dados sobre média e variância, ou desvio padrão, de comprimentos de mensagens,base pa

ra a caracterização de processos de serviço.

A existência de unidades envolvidas no processo de comunicação, situadas remotamente, não permite a estas unidades, ou a uma eventual unidade central (ou primária) de controle, o conhecimento instantâneo do estado do sistema, acarretando em utilização de alguma capacidade do canal de comunicações para a veiculação destas informações. Tal capacidade pode ser tomada em freqüência ou em tempo, sendo este último o caso dos sistemas controlados por disciplina de "polling", o que sugere a utilização de modelos que incorporem tempos de transição finitos (não-nulos), para a representação de componentes de tempo associadas com este controle.

Nestes sistemas, esta amostragem do estado do sistema (por exemplo, a pergunta a uma estação secundária sobra a existência de mensagens a transmitir) é realizada de maneira periódica e regular, admitindo inclusive prioridades, o que implica em atividades de transição contínuas entre as filas, mesmo com o sistema em estado zero (operação "keep-switching").

Os resultados disponíveis são, na maioria dos casos, expressões, exatas ou aproximadas, para os valores médios dos atrasos nas filas do sistema.

Para uma primeira análise do problema central deste estudo (comunicação estação-primária estação-secundária com "polling") podem ser utilizados modelos com 2 filas, geralmente assimétricos, com tempos de transição finitos e operação "keep-

-switching". A extensão para diversas estações secundárias em multiponto exigiria a utilização de modelos assimétricos com N filas.

Considerando ainda o objetivo de colocar à disposição de profissionais de engenharia e análise interessados, um modelo de utilização simples e eficiente, procura-se propor a aplicação de modelos para os quais existam soluções em forma fechada, ou seja, expressões analíticas relativamente simples, que não envolvam cálculos computacionais intermediários (por exemplo, expressões recursivas ou iterativas, avaliação de séries ou somatórios infinitos, algoritmos numéricos e similares) ou soluções de sistemas de equações simultâneas.

Resulta desta seqüência de considerações a seleção dos modelos analisados por Sykes /SYKEJ70/, Eisenberg /EISEM72/ e Konheim e Meister /KONHA74/, para investigação mais detalhada e posterior utilização, na análise da classe dos problemas de interesse. No Capítulo IV desta dissertação são estudados com maior atenção estes modelos, e no Capítulo V é apresentada e discutida a utilização do modelo de Sykes na análise de desempenho de um sub-sistema de comunicação de dados "half-duplex" sob disciplina de "polling".

A utilização do modelo de Swartz /SWARG77/ é proposta como extensão deste trabalho, dada a complexidade de manipulação e solução.

Ainda nesta seção deve ser ressaltada a aplicabili-

dade de outras classes de modelos no contexto de análise de desempenho de sistemas voltados para comunicação/processamento de dados.

Modelos de multifilas com tempos de transição nulos, ou finitos com estacionariedade da unidade de serviço, podem ser utilizados onde se verificam transições em tempos praticamente nulos entre as filas (por exemplo, unidades locais amostradas por varredura cíclica ou por interrupção), ou quando estes tempos de transição se tornam bastante pequenos, quando comparados ao tráfego efetivo de mensagens (por exemplo, em um canal com utilização alta, em que a probabilidade de uma estação ter mensagens para transmissão é grande, as diversas estações presentes são penalizadas por estas transmissões em si, e não pelo "overhead" associado com "polling" das estações).

Por outro lado, em linhas com muito baixa utilização o fator dominante consiste no "overhead" de controle.Tal fato levou, por exemplo, Hayes a pesquisar esquemas alternativas de "polling" adaptativo, com o objetivo de melhorar o desempenho de tais sistemas /HAYEJ78/.

Ambientes possíveis para utilização dos modelos de multifilas mencionados acima, seriam sistemas digitais voltados para comunicação (por exemplo, centrais de comutação) ou controle de dispositivos em tempo real, controlados por microprocessadores gerenciando diversos dispositivos (fontes de tráfego), ou ainda, redes locais ou sistemas de microprocessadores distribuídos, onde estes, então, constituem fontes de tráfego para um

meio de capacidade finita, por exemplo, um barramento ou anel compartilhado.

Particularmente com relação a redes locais, W. Bux, da IBM, apresenta em /BUXWE81/, uma avaliação comparativa do desempenho de sub-redes locais, em barramento e em anel, baseada na utilização de modelos de multifilas, incluindo algumas variantes introduzidas por Bux.

São considerados quatro métodos de acesso ao meio de transmissão: "token ring", "slotted ring", CSMA/CD e "multilevel multiple-access scheme" (MLMA), descritos brevemente no artigo em questão.

Os sistemas analisados são sempre simétricos, e os resultados de modelos de multifilas utilizados, qualitativa e quantitativamente, são aqueles obtidos por Kuehn, Konheim e Meister e Sykes, respectivamente, em /KUEHP79b/, /KONHA74/ e /SYKEJ70/.

Tomando os atrasos médios de acesso fornecidos por estes modelos, juntamente com os tempos de transmissão e atrasos de propagação, como medida de desempenho, Bux analisa/compara os efeitos, nos diferentes esquemas, de parâmetros de sistema como taxa de transmissão, comprimento do cabo, comprimentos de pacotes e "overhead" de controle.

CAPÍTULO III

TÓPICOS SELECIONADOS DE MATEMÁTICA APLICADA

Este Capítulo III compreende a apresentação de um conjunto de ferramentas matemáticas fundamentais para o desenvolvimento, melhor entendimento e utilização dos modelos de multifilas disponíveis, em particular, aqueles introduzidos nos Capítulos II e IV deste estudo.

A orientação adotada nas Seções III.1 a III.3 pressupõe conhecimentos básicos em cálculo diferencial e integral, funções de variáveis complexas, teoria de probabilidades, processos estocásticos e teoria de filas simples, visando uma revisão orientada e, em alguns casos, fornecendo indicações ou derivações explícitas de resultados relevantes no tratamento da classe de modelos estudada.

Na medida do realizável, são ilustradas aplicações de conceitos, de forma a mostrar o relacionamento destes com o ambiente de comunicação de dados e análise de desempenho, principalmente no que diz respeito à caracterização de entradas de modelos (variáveis aleatórias, processos estocásticos).

Tópicos relevantes em teoria de filas M/G/1 são incluídos, com base na experiência do autor, de que o domínio de tópicos desta natureza, contribui expressivamente para o enten

dimento, utilização, discernimento de alternativas e aquisição de sensibilidade sobre o assunto global, incluíndo os aspectos metodológicos.

Na Seção III.1 são revistos os conceitos de funções características e geratrizes, e de transformadas, Z e de Laplace, de variáveis aleatórias, e suas propriedades. Seguem-se exemplos de aplicações destes operadores a distribuições de probabilidade mais utilizadas na caracterização de tráfego de dados (/JACKP69/,/FUCHE70/,/DUDIA71/,/ONEIP80/ e /PAWLP81/), na obtenção de representações equivalentes nos domínios complexos (Z e s), ou na caracterização parcial, através de momentos.

A Seção III.2 introduz definições e descrições de processos estocásticos, e apresenta fundamentos e resultados centrais em cadeias de Markov e processos de renovação, incluindo a obtenção de processos de renovação compostos, baseada em teoria da renovação e em transformadas das distribuições envolvidas. Processos de Poisson, bastante comuns no ambiente em estudo, são tratados como casos particulares de processos de renovação.

Na Seção III.3 são abordados tópicos especiais em teoria de filas M/G/1, como cadeias de Markov imersas, períodos de ocupação e filas com períodos de ausência de serviço, uma vez que, conforme pode ser observado no Capítulo II, a grande maioria das análises de modelos de multifilas disponíveis considera filas do tipo M/G/*, e utiliza frequentemente estes conceitos na análise.

Adicionalmente, pretende-se que o material apresentado neste Capítulo III, venha a contribuir para uma ampliação da capacidade geral de realização de análises quantitativas de engenheiros e analistas interessados, na medida em que procura cobrir alguns "gaps" conceituais e mesmo de comunicação (linguagem, notação, siglas, etc.).

Neste contexto, o autor recomenda fortemente o contato (e estudo) com trabalhos introdutórios e/ou de levantamento, na(s) área(s) de interesse, como, por exemplo, /REISM82/, /KOBAH77/, e /CHUWW72/, entre outros. Tais trabalhos, contém extensas listas de bibliografia, com um total global aproximado de 300 referências, eliminadas as redundâncias.

SEÇÃO III.1

TEORIA DE PROBABILIDADES - FUNÇÕES GERATRIZES E TRANSFORMADAS

Esta Seção III.1 consiste de uma revisão de teoria de probabilidades, com ênfase em funções geratrizes e transformadas, apresentando, no Item III.1.1, os principais conceitos e propriedades e, no Item III.1.2, aplicações destes a distribuições mais utilizadas na caracterização de processos de tráfego de dados. Referências básicas para esta Seção III.1 são /KLEIL75, Ap.II/, /GIFFW75/, /PATEJ76/ e IBM/GF20-0007-01/.

III.1.1 Definições e Propriedades

Dado um espaço ou sistema de probabilidades (Ω, ε, P),

onde Ω representa um espaço amostral, ε uma família de eventos e P uma medida de probabilidade, define-se uma variável aleatória contínua (discreta) como uma função $X(w): w \in \Omega \rightarrow Y \subseteq \mathbb{R}$, que mapeia o espaço amostral em um conjunto Y, subconjunto infinito (finito ou enumerável) do conjunto dos números reais $\mathbb{R}$.

A uma variável aleatória (v.a.) X são geralmente associadas duas funções: a função de distribuição acumulada $F_X(x)$, denotada por f.d.a, e a função de densidade de probabilidade $f_X(x)$, denotada por f.d.p., onde

$$F_X(x) \triangleq Pr[X \leq x] = \{w : X(w) \leq x\} \qquad e$$

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}, \quad tal\ que$$

$$\int_{-\infty}^{x} f_X(y)dy = F_X(x) \qquad e \qquad \int_{-\infty}^{\infty} f_X(x)dx = 1.$$

As notações comumente encontradas na literatura em inglês para f.d.a. e f.d.p. são, respectivamente PDF e pdf.

Com a finalidade de garantir a existência da derivada $f_X(x)$ nos casos em que a f.d.a. contenha descontinuidades (variável aleatória discreta, v.a.d.), deve-se permitir a introdução de impulsos, ou pontos de acumulação, na f.d.p., obtendo-se, desta forma, uma formulação unificada para variáveis aleatórias contínuas, v.a.c., e discretas, v.a.d. .

Entretanto, encontram-se frequentemente as denominações de função de probabilidades, função de massa de probabili

dade, ou <u>função densidade discreta</u>, g_k, na caracterização de v.a.d.'s:

$$g_k \triangleq \Pr\,[X=k].$$

Os <u>momentos ordinários e centrais</u> de uma v.a. podem ser expressos em notação integral de Riemann, se são incorporados impulsos na f.d.p., e em caso contrário, em notação integral de Stieltjes, uma vez que ambas são utilizadas/encontradas na literatura corrente, respectivamente:

$$E[X^n] \triangleq \overline{X^n} = \int_{-\infty}^{\infty} x^n\, f_X(x)dx \;, \quad \text{ou} \quad \int_{-\infty}^{\infty} x^n\, dF_X(x), \quad \text{e}$$

$$E\left[(X-E[X])^n\right] \triangleq \overline{(X-\bar{X})^n} \triangleq \int_{-\infty}^{\infty} (x-\bar{X})^n\, f_X(x)dx \;, \text{ ou } \int_{-\infty}^{\infty} (x-\bar{X})^n\, dF_X(x).$$

Existe a seguinte relação entre os momentos centrais de ordem n de uma v.a. e os n momentos ordinários associados:

$$\overline{(X-\bar{X})^n} = \sum_{k=0}^{n} \binom{n}{k}\, \overline{X^k}\, .\, (-\bar{X})^{n-k}.$$

Da mesma forma como para a média (valor esperado), $E[X] \triangleq \bar{X} \triangleq \int_{-\infty}^{\infty} x\, f_X(x)dx$, notações especiais são ainda adotadas para o segundo momento central (variância) e para o coeficiente de variação de uma v.a., quais sejam;

$$\sigma_X^2 \triangleq \overline{(X-\bar{X})^2} \triangleq \overline{X^2} - (\bar{X})^2 \quad \text{e}$$

$$C_X \triangleq \frac{\sigma_X}{\bar{X}} \triangleq (\overline{X^2}/(\bar{X})^2 - 1)^{1/2} \quad .$$

Bastante utilizado é também o desvio padrão, σ_X, que corresponde à raiz quadrada da variância σ_X^2.

A <u>função característica</u> (f.c.) de uma v.a. X, denotada por $\phi_X(n)$, é definida como

$$\phi_X(n) \triangleq E\left[e^{jnX}\right] \triangleq \int_{-\infty}^{\infty} e^{jnx} f_X(x)dx \quad , \quad \text{onde}$$

$$j = \sqrt{-1} \qquad e \qquad n \in \mathbb{R} \quad .$$

Observe-se que, exceto pelo sinal do expoente, a função característica corresponde à transformada de Fourier da f.d.p. de X.

Uma <u>propriedade</u> importante da f.c. permite o cálculo de todos os <u>momentos de X</u>, conforme:

$$\left.\frac{d^n\phi_X(u)}{du^n}\right|_{u=0} = j^n \overline{X^n} \quad , \quad \text{ou, em notação mais compacta}$$

$$\phi_X^{(n)}(0) = j^n \overline{X^n} \quad , \quad \text{onde } \phi_X^{(n)}(x_o) \triangleq \left.\frac{d^n\phi_X(x)}{dx^n}\right|_{x=x_o} \quad ,$$

sendo $\phi_X(x)$ uma função arbitrária continuamente diferenciável.

A função geratriz de momentos (f.g.m.) de uma v.a.X,

denotada por $M_X(v)$, é definida como

$$M_X(v) \triangleq E\left[e^{vX}\right] \triangleq \int_{-\infty}^{\infty} e^{vx} f_X(x)dx \ , \ \text{com } v\in\mathbb{R} \ ,$$

e gera todos os momentos da v.a.X, através da relação

$$M_X^{(n)}(0) = \overline{X^n} \ .$$

As propriedades de <u>geração de momentos</u> da f.c. e da f.g.m. podem ser explicitadas, expandindo-se a exponencial do integrando em série de potências, e integrando-se termo a termo, conforme

$$\emptyset_X(u) = \int_{-\infty}^{\infty} f_X(x) \left[1+jux + \frac{(jux)^2}{2!} + \ldots\right] dx$$

$$= 1 + ju\bar{X} + \frac{(ju)^2}{2!} \overline{X^2} + \ldots \ , \ \text{e}$$

$$M_X(v) = \int_{-\infty}^{\infty} f_X(x) \left[1+vx + \frac{(vx)^2}{2!} + \ldots\right] dx$$

$$= 1 + v\bar{X} + \frac{v^2}{2!} \overline{X^2} + \ldots \ .$$

Uma função também bastante utilizada é a <u>transformada de Laplace</u> (t.L.) da f.d.p. de uma v.a.X. Seja F(x) a f.d.a. da v.a.X, tal que $F(x) = Pr[X\leq x]$, e f(x) a f.d.p. correspondente. A transformada de Laplace de f(x), denotada por $F^*(s)$ é definida por

$$F^*(s) = E\left[e^{-sX}\right] \triangleq \int_{0^-}^{\infty} e^{-sx} f(x)dx \ , \ \text{onde } x\in\mathbb{C} \ ,$$

e $\mathbb{C}$ representa o conjunto dos números complexos.

Notar que o limite inferior de integração é definido como 0^-, uma vez que as variáveis aleatórias em questão são, em sua grande maioria, não-negativas, e que o limite pela esquerda garante a consideração de quaisquer eventuais impulsos na origem da f.d.p..

De maneira análoga à f.c. e à f.g.m. , a t.L. da f.d.p. de uma v.a.X também gera momentos ordinários, através de

$F^{*(n)}(0) = (-1)^n \overline{X^n}$, e estas três funções são relacionadas por

$\emptyset_X(sj) = M_X(-s) = F^*(s)$, de forma que o n-ésimo momento de X pode ser calculado opcionalmente como

$\overline{X^n} = j^{-n} \emptyset_X^{(n)}(0) = M_X^{(n)}(0) = (-1)^n F^{*(n)}(0)$, valendo ainda a relação $\emptyset_X(0) = M_X(0) = F^*(0) = 1$, que pode ser interpretada como o momento de ordem zero, que é unitário.

No caso de uma v.a.d., descrita por $g_k = Pr[X=k]$, define-se ainda a função geratriz de probabilidades (f.g.p.), denotada por G(z), como

$G(z) \triangleq E[z^X] \triangleq \sum_k z^k g_k$, com $z \in \mathbb{C}$, e a seguinte expansão em série de potências

$$G(z) = \sum_{k=0}^{\infty} g_k z^k = g_0 + g_1 z + g_2 z^2 + g_3 z^3 + \ldots .$$

Observe-se que, a menos do sinal do expoente de z, a f.g.p. equivale à transformada Z da sequência discreta $\{g_k\}$, sendo assim denominada quase que exclusivamente na literatura. Entretanto, conforme pode ser visto em /GIFFW75,Cap.3/, a f.g.p. é, em realidade, uma transformada geométrica, que é a correspondente discreta da transformada exponencial.

Pela expansão de G(z) em série de potências, torna-se aparente que

$$G(1) = 1$$

$$G^{(1)}(1) = \bar{X}$$

$$G^{(2)}(1) = \overline{X^2} - \bar{X} \text{ , e assim sucessivamente, o que explicita a propriedade de geração de momentos da f.g.p.}$$

A soma de uma coleção de v.a.'s independentes, $\{X_i\}$, com f.d.p.'s $\{f_{X_i}(x)\}$, pode ser avaliada através de convoluções.

Seja $Y = \sum_{i=1}^{n} X_i$ e $f_Y(y)$ a f.d.p. associada a Y. Demonstra-se que

$$f_Y(y) = f_{X_1}(y) \circledast f_{X_2}(y) \circledast \ldots \circledast f_{X_n}(y) \text{ , onde o operador } \circledast \text{ indica a convolução, que possui propriedades associativas e comutativas, e, que no caso n=2 é expresso por}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(y-x_2)\ f_{X_2}(x_2)dx_2.$$

Entretanto, a variável Y pode ser mais facilmente avaliada com o uso de funções geratrizes ou transformadas. Por exemplo, sejam $\phi_Y(u)$ e $\{\phi_{X_i}(u)\}$ as f.c.'s correspondentes a Y e a $\{X_i\}$. Demonstra-se, apenas pela definição, que

$$\phi_Y(u) = \prod_{i=1}^{n} \phi_{X_i}(u), \quad \text{que, no caso de } X_i\text{'s idênticamente distribuídas, se reduz a}$$

$$\phi_Y(u) = [\phi_X(u)]^n .$$

A soma S_n de uma coleção de v.a.'s independentes e idênticamente distribuídas (i.i.d.), com f.d.a. F(x) e t.L.F*(s), na qual o número de parcelas, N, é uma v.a.d. , independente de $\{X_i\}$ e descrita por $g_n = Pr[N=n]$, com f.g.p. G(z), pode ser avaliada através de sua t.L. H*(s), como a função composta

$$H^*(s) = G[F^*(s)].$$

Derivando-se esta expressão e tomando-se s=0, podem ser determinados os momentos de S_n, em termos dos momentos de N e X. As duas primeiras derivadas de H*(s) são

$$H^{*(1)}(s) = G^{(1)}[F^*(s)].F^{*(1)}(s) \quad \text{e}$$

$$H^{*(2)}(s) = G^{(2)}[F^*(s)].[F^{*(1)}(s)]^2 + G^{(1)}[F^*(s)].F^{*(2)}(s).$$

No ponto s=0 tem-se

$$H^{*(1)}(0) = G^{(1)}(1).F^{*(1)}(0) \qquad \text{e}$$

$$H^{*(2)}(0) = G^{(2)}(1).[F^{*(1)}(0)]^2 + G^{(1)}(1).F^{*(2)}(0).$$

Relacionando-se os dois primeiros momentos de N e X com as respectivas f.g.p. e t.L., e resolvendo-se para a média e variância de S_n, tem-se, finalmente, que

$$\bar{S}_n = \bar{N}.\bar{X}$$

$$\sigma^2_{S_n} = \bar{N}\sigma^2_X + (\bar{X})^2\sigma^2_N.$$

Neste caso, a f.d.a. associada à variável aleatória resultante, S_n, é denominada distribuição composta, conceito este que será utilizado na caracterização de certos tipos de processos de chegada na Seção III.2 deste Capítulo III.

Como extensão desta revisão são recomendados o Capítulo 1 de /PATEJ76/ (momentos, cumulantes, funções geratrizes e características, para distribuições discretas e contínuas) e os Capítulos 1,2,3 e 4 de /GIFFW75/ (técnicas de transformadas aplicadas a modelos de probabilidades, incluindo relações com funções características e geratrizes, e funções simples de variáveis aleatórias).

III.1.2 Exemplos de Aplicações

Em problemas práticos de rotina dispõe-se, geralmen

te, de estimadores empíricos para a média e a variância, ou desvio padrão, de variáveis aleatórias observadas. Em alguns casos são disponíveis formas gráficas aproximadas das funções f.d.a. e/ou f.d.p., a partir das quais são formuladas e verificadas hipóteses sobre a forma analítica da distribuição.

Com o objetivo de familiarização/exercícios com as expressões analíticas das distribuições de maior interesse (taxas e comprimentos de mensagens), incluindo as respectivas médias e variâncias, são apresentados, a seguir, exemplos de obtenção destes dois momentos, através da utilização das funções transformadas definidas no Item III.1.1 anterior.

Algumas das expressões obtidas serão utilizadas diretamente na caracterização de processos de chegada do tipo renovação compostos, na Seção III.2 seguinte. As expressões analíticas iniciais apresentadas são encontradas em /PATEJ76/.

1. Distribuição POISSON

(aplicação, quase exclusiva, em processos de contagem de chegadas no tempo)

Seja uma v.a.d. X com distribuição Poisson de parâmetro λ:

$$g_k \triangleq Pr[X=k] = \frac{e^{-\lambda}\lambda^k}{k!} \qquad (3.1.1)$$

com $\lambda>0$ e $k=0,1,2,\ldots$.

A f.g.p. associada é $G(z) = e^{\lambda(z-1)}$, (3.1.2)

então:

$$G^{(1)}(z) = \lambda e^{\lambda(z-1)} \Longrightarrow G^{(1)}(1) \triangleq \bar{X} = \lambda \quad (3.1.3)$$

$$G^{(2)}(z) = \lambda^2 e^{\lambda(z-1)} \Longrightarrow G^{(2)}(1) \triangleq \overline{X^2} - \bar{X} \triangleq \sigma_X^2 = \lambda . \quad (3.1.4)$$

2. Distribuição EXPONENCIAL

(aplicação, vinculada ao processo Poisson, na descrição de intervalos de tempo entre chegadas)

Seja uma v.a.c. X com distribuição exponencial de parâmetro $\lambda > 0$:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} , \quad (3.1.5)$$

donde

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} . \quad (3.1.6)$$

A t.L. associada é $F_X^*(s) = \dfrac{\lambda}{\lambda+s}$, (3.1.7)

então:

$$\bar{X} \triangleq -F_X^{*(1)}(0) = \left. \frac{\lambda}{(\lambda+s)^2} \right|_{s=0} = 1/\lambda \quad (3.1.8)$$

$$\overline{X^2} \triangleq F_X^{*(2)}(0) = \left. \frac{2\lambda(\lambda+s)}{(\lambda+s)^4} \right|_{s=0} = 2/\lambda^2 \text{ , donde}$$

$$\sigma_X^2 \triangleq \overline{X^2} - (\bar{X})^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2 . \quad (3.1.9)$$

Ainda, $C_X \triangleq \frac{\sigma_X}{\bar{X}} = 1$. (3.1.10)

3. Distribuição GEOMÉTRICA

(aplicação na descrição de comprimentos de mensagens ou distribuição de serviço discreto)

Seja uma v.a.d. X com distribuição geométrica de parâmetro p=1-q, onde 0<p<1:

$$g_k \triangleq Pr[X=k] = pq^{(k-1)} \quad , \qquad (3.1.11)$$

onde k = 1,2,3 ... .

A f.g.p. associada é $G(z) = \frac{(1-q)z}{1-qz}$, (3.1.12)

então:

$$G^{(1)}(z) = \frac{1-q}{(1-qz)^2} \Longrightarrow G^{(1)}(1) \triangleq \bar{X} = \frac{1}{1-q} = \frac{1}{p} \qquad (3.1.13)$$

$$G^{(2)}(z) = \frac{2q(1-q)}{(1-qz)^3} \quad , \quad \text{donde}$$

$$G^{(2)}(1) \triangleq \overline{X^2} - \bar{X} = \frac{2q}{(1-q)^2} \Rightarrow \sigma_X^2 = \frac{q}{(1-q)^2} \; . \qquad (3.1.14)$$

4. Distribuição GAMMA

(largo espectro de aplicações, incluindo descrições de processos de chegada e serviço, e variáveis de desempenho)

Seja uma v.a.c.X com distribuição gamma de parâme-

tros α e θ ; $\alpha, \theta > 0$:

$$f_X(x) = \Gamma(x;\alpha,\theta) = \begin{cases} 0 & x \leq 0 \\ \frac{\theta^\alpha}{\Gamma(\alpha)} .x^{\alpha-1} e^{-\theta x} & x>0 \end{cases}, \quad (3.1.15)$$

onde $\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1} e^{-x} dx.$

A f.g.m. associada é $M_X(v) = \left(\frac{\theta}{\theta-v} \right)^\alpha$, (3.1.16)

então:

$$M_X^{(1)}(v) = \frac{\alpha\theta^\alpha}{(\theta-v)^{\alpha+1}} \Rightarrow M_X^{(1)}(0) \triangleq \bar{X} = \alpha/\theta \quad (3.1.17)$$

$$M_X^{(2)}(v) = \frac{\alpha(\alpha+1)\theta^\alpha}{(\theta-v)^{\alpha+2}} \Rightarrow M_X^{(2)}(0) \triangleq \overline{X^2} = \frac{\alpha(\alpha+1)}{\theta^2}$$

logo, $\sigma_X^2 \triangleq \overline{X^2} - (\bar{X})^2 = \alpha/\theta^2$. (3.1.18)

5. Distribuição CONSTANTE

(aplicação em aproximações para processos determinísticos, e na composição de funções empíricas, constantes por trechos)

Seja uma v.a. degenerada X, com distribuição constante em L (função δ-Dirac, ou impulso, no ponto x=L):

$$f_X(x) = \begin{cases} 1 & x=L \\ 0 & x \neq L \end{cases}, \quad \text{ou} \quad f_X(x) = \delta(x-L). \quad (3.1.19)$$

A t.L. associada é $F_X^*(s) \triangleq \int_{0^-}^\infty e^{-sx} \delta(x-L)dx = e^{-sL}$, (3.1.20)

então:

$$F_X^{*(1)}(s) = -Le^{-sL} \Rightarrow \bar{X} \triangleq -F_X^{*(1)}(0) = L \qquad (3.1.21)$$

$$F_X^{*(2)}(s) = L^2 e^{-sL} \Rightarrow \overline{X^2} \triangleq F_X^{*(2)}(0) = L^2 \ ,$$

$$\text{logo, } \sigma_X^2 \triangleq \overline{X^2} - (\bar{X})^2 = L^2 - L^2 = 0 \ . \qquad (3.1.22)$$

Considerando X como uma v.a.d., com

$$g_k = \begin{cases} 1 & k=L \\ & \\ 0 & k\neq L \end{cases} \ , \qquad (3.1.23)$$

tem-se a f.g.p. $G(z) = z^L$, e então:

$$G^{(1)}(z) = Lz^{L-1} \Rightarrow \bar{X} \triangleq G^{(1)}(1) = L \qquad (3.1.24)$$

$$G^{(2)}(z) = L(L-1)z^{L-2} \Rightarrow \overline{X^2} - \bar{X} = L^2 - L \ , \text{ e}$$

$$\text{logo, } \sigma_X^2 \triangleq \overline{X^2} - (\bar{X})^2 = L^2 - L^2 = 0. \qquad (3.1.25)$$

Este exemplo conclui esta Seção III.1, juntamente com a recomendação para exercícios adicionais de manipulação de expressões analíticas relacionadas com distribuições relevantes , particularmente no que diz respeito a obtenção e inversão de funções transformadas.

SEÇÃO III.2

TÓPICOS EM TEORIA DE PROCESSOS ESTOCÁSTICOS

Processos de chegada e de serviço em sistemas de filas constituem exemplos bastante elucidativos, e de fácil assimilação, de processos estocásticos, uma vez que consistem em sequências, no tempo, de variáveis aleatórias (por exemplo, a sequência dos intervalos entre chegadas consecutivas de mensagens a um sistema, ou a sequência dos tempos de transmissão destas mensagens através de um canal de comunicações), podendo exibir as relações mais diversas entre os termos destas sequências.

O objetivo básico da teoria de processos estocásticos consiste no estudo do comportamento destas sequências e dos relacionamentos probabilísticos existentes (quando algum) entre os termos das mesmas.

Esta Seção III.2 compreende, então, um resumo de definições básicas importantes em processos estocásticos, apresentado no Item III.2.1, ao qual se seguem os Itens III.2.2 e III.2.3, introduzindo duas classes de processos relevantes neste estudo, quais sejam, processos de Markov e de renovação, incluindo processos compostos. O Item III.2.4 apresenta exemplos de aplicações dos conceitos introduzidos, relacionados com o ambiente de tráfego de dados.

Referências básicas para esta Seção III.2 são /BHATU72/, /ROSSS72, Cap.7/, /PARZE62, Cap.5/, /GROSD74, Ap.4/, /KLEIL75, Cap.2/, /CHUWW72/ e /KOBAH77/.

A referência /BHATU72/ é bastante recomendada pelo autor, no caso de interesse maior, de caráter aplicado, no assunto processos estocásticos.

III.2.1 Processos Estocásticos - Definição e Descrição

Processos estocásticos, ou processos aleatórios, são abstrações matemáticas de processos empíricos, cuja evolução é regida por leis probabilísticas /BHATU72, Cap.2/.

Do ponto de vista da teoria matemática de probabilidades, um processo estocástico pode ser definido como uma família, ou coleção, de variáveis aleatórias, $\{X(t),\ t \in T\}$, indexadas pelo parâmetro t, sobre um conjunto de índices, ou espaço de parâmetro, T.

O processo $\{X(t)\}$ é denominado processo estocástico, os valores assumidos pelo processo são denominados estados, e o conjunto de valores possíveis constitui o espaço de estado, que pode ser discreto (conjunto enumerável de pontos) ou contínuo (intervalo com infinitos pontos).

O conjunto de valores possíveis para o parâmetro-índice é denominado espaço de parâmetro, que também pode ser discreto ou contínuo. Em caso de parâmetro-índice discreto utiliza-se a notação $\{X_n,\ n=0,1,2,\ldots\}$.

A partir da natureza dos espaços de estado e de parâmetro, podem então ser conceituados quatro classes de proces-

sos estocásticos, correspondendo às quatro combinações possíveis entre os tipos de espaços de estado e de parâmetro. Assim, tem-se, processos de parâmetro discreto e espaço de estado discreto, processos de parâmetro contínuo e espaço de estado discreto, etc.

Para um dado valor do parâmetro t, o processo estocástico $\{X(t)\}$ é uma variável aleatória simples, e sua distribuição de probabilidade pode ser obtida, da forma como para qualquer outra variável. Entretanto, quando t varia em um espaço T, uma distribuição de probabilidade, para um dado t, não fornece informação sobre o processo $\{X(t)\}$. Para uma informação completa sobre o processo deve-se especificar a distribuição conjunta das variáveis aleatórias básicas da família $\{X(t),\ t \in T\}$.

Quando T é contínuo, a obtenção de tal distribuição torna-se impossível, uma vez que o número de membros desta família é infinito. Nestas circunstâncias, supõe-se que o comportamento do processo possa ser estudado sobre um conjunto discreto adequado de pontos (análogamente ao conceito de amostragem), definindo-se uma função de distribuição conjunta em instantes $(t_1, t_2,\ldots,t_n)$, tal que $t_1<t_2<\ldots<t_n \in T$, da forma

$$\Pr[X(t_1) \leq x_1;\ X(t_2) \leq x_2;\ldots;\ X(t_n) \leq x_n]. \quad (3.2.1)$$

Esta distribuição assume sua forma mais simples quando as variáveis aleatórias são independentes, quando equivale, então, ao produto das distribuições individuais. Situações práticas implicam, entretanto, quase que invariavelmente, na exis-

tência de alguma forma de dependência entre estas variáveis.

A título de ilustração/exercício, o autor sugere que sejam imaginadas as diversas situações de dependência que ocorrem, ou podem ocorrer, em ambientes de comunicação/processamento de dados, nos contextos de processos de chegada e de serviço.

Embora sejam necessárias distribuições conjuntas, do tipo definido na eq. (3.2.1), para uma descrição completa do processo, grande parte da informação necessária na prática pode ser fornecida por funções de distribuição de transição, que são funções de distribuição de probabilidades condicionais, disponíveis para valores específicos do parâmetro t.

Sejam, então, t_o e $t_1 \in T$, $t_o \leq t_1$. Define-se a função de distribuição de transição condicional como

$$F(x_o, x_1; t_o, t_1) = Pr[X(t_1) \leq x_1 \mid X(t_o) = x_o]. \qquad (3.2.2)$$

Quando o processo estocástico possui espaços de estado e de parâmetro discretos, definem-se as probabilidades de transição como

$$P_{ij}^{(m,n)} = Pr[X_n = j \mid X_m = i]. \qquad (3.2.3)$$

Um processo estocástico $\{X(t), t \in T\}$, ou $\{X_n, n = 0,1,2,\ldots\}$, é considerado homogêneo, ou invariante no tempo (a translações), se a função de distribuição de transição, conforme eq. (3.2.2) ou (3.2.3), depende apenas da diferença $(t_1 - t_o)$,

ou (n-m), e não dos valores particulares de t_o e t_1, ou n e m. Tem-se, então

$$F(x_o,x;t_o,t_o+t) = F(x_o,x;0,t) = F(x_o,x;t) \qquad (3.2.4)$$

e

$$P_{ij}^{(n_o+n,n_o)} = P_{ij}^{(n,0)} = P_{ij}^{(n)} \quad , \qquad (3.2.5)$$

para quaisquer t_o, $n_o \in T$.

Em /KLEIL75, Cap.2/ é apresentada uma classificação (e conceituação) bastante abrangente e clara dos diversos tipos de processos estocásticos. Os processos homogêneos são destacados neste Item III.2.1 por constituírem a classe de processos mais comumente utilizada/encontrada nas análises disponíveis, e corresponderem a situações reais de análise de desempenho em períodos de pico, nos quais as taxas de tráfego são consideradas constantes, o que equivale a dizer, também, que os processos associados são homogêneos.

Nos Itens III.2.2 e III.2.3 seguintes são caracterizados, respectivamente, os processos de Markov e de renovação, a partir das definições e equações básicas apresentadas neste Item III.2.1.

III.2.2 Processos e Cadeias de Markov

Processos estocásticos resultantes de observações de situações reais são tais que, conforme mencionado no Item III.2.1

anterior, para um conjunto discreto de parâmetros $t_1, t_2, \ldots, t_n \in T$, as variáveis aleatórias $X(t_1), X(t_2), \ldots, X(t_n)$ exibem algum tipo de dependência, tornando mais complicada a análise do processo, à medida que a estrutura de dependência se torna complexa /BHATU72, Cap. 2/.

O tipo mais simples de dependência é a dependência de primeira-ordem, ou de Markov, definida conforme a seguir.

Seja um conjunto finito, ou enumerável, de pontos $(t_o, t_1, \ldots, t_n, t)$, $t_o < t_1 < t_2 < \ldots < t_n < t$, com $t, t_r \in R(r=0,1,2,\ldots, n)$, onde T é o espaço de parâmetro do processo $\{X(t)\}$.

$\{X(t), t \in T\}$ é um processo de Markov se a distribuição condicional de $X(t)$, para dados valores de $X(t_o), X(t_1), \ldots, X(t_n)$, depende apenas de $X(t_n)$, que é o valor conhecido mais recente do processo, ou seja

$$\Pr[X(t) \leq x | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots, X(t_o) = x_o] =$$

$$= \Pr[X(t) \leq x | X(t_n) = x_n] \qquad (3.2.6)$$

$$= F(x_n, x; t_n, t) \; . \qquad (3.2.7)$$

Observa-se, então, que o conhecimento do estado de um processo de Markov, em um instante específico do tempo, contém informação suficiente para a predição do comportamento do processo a partir daquele ponto.

Como consequência da propriedade expressa nas eq.

(3.2.6) e (3.2.7), tem-se a seguinte relação

$$F(x_o,x;t_o,t)= \int_{y\in S} F(y,x;\tau,t)dF(x_o,y;t_o,\tau) \ , \qquad (3.2.8)$$

onde $t_o<\tau<t$, e S representa o espaço de estado do processo $\{X(t)\}$.

Quando o processo estocástico possui espaços de estado e de parâmetro discretos, as eq. (3.2.6) e (3.2.8) tomam as seguintes formas, para $n>n_1>n_2>\ldots>n_k \in T$:

$$Pr[X_n = j \,|\, X_{n_1} = i_1, X_{n_2} = i_2, \ldots, X_{n_k} = i_k] =$$

$$= Pr[X_n = j \,|\, X_{n_1} = i_1] \qquad (3.2.9)$$

$$= P_{i_1 j}^{(n_1,n)} \qquad (3.2.10)$$

Aplicando-se a mesma propriedade para $m<r<n$, tem-se

$$P_{ij}^{(m,n)} = Pr[X_n = j \,|\, X_m=i]$$

$$= \sum_{k\in S} Pr[X_n=j \,|\, X_r=k].Pr[X_r=k \,|\, X_m=i] \qquad (3.2.11)$$

$$= \sum_{k\in S} P_{ik}^{(m,r)} . P_{kj}^{(r,n)} . \qquad (3.2.12)$$

As expressões (3.2.8) e (3.2.12) são denominadas <u>equações de Chapmam-Kolmogorov</u> (CK) dos processos, e constituem as equações básicas no estudo de processos de Markov.

Quanto à classificação, processos de Markov com espaços de estado discretos são denominados cadeias de Markov, a despeito da natureza do espaço de parâmetro, enquanto que, processos de Markov com espaços de estado contínuos não recebem qualquer denominação especial. Uma cadeia de Markov é finita se o espaço de estado é finito, sendo infinita ou enumerável, em caso contrário.

Uma vez que, como será visto na Seção III.3 seguin (Item III.3.1), o processo estocástico associado à variável de estado de uma fila simples (número de itens no sistema) evolui em um espaço de estado discreto, e que estes processos, ainda que não Markovianos, permitem, em alguns casos, a identificação de sub-processos de interesse com propriedades Markovianas (cadeias de Markov imersas, ver Item III.3.1), apresenta-se em seguida um resumo dos principais conceitos e resultados da teoria de cadeias de Markov.

O autor recomenda, entretanto, /BHATU72/,/KEMEJ76a/ e /KEMEJ76b/ para maior aprofundamento no assunto cadeias de Markov.

Se, em uma cadeia de Markov, a variável aleatória X_n assume o valor j, o sistema é considerado no estado j após n passos, ou transições. As probabilidades condicionais $Pr[X_n=j|X_{n-1}=i]$ são denominadas probabilidades de transição de um passo, ou, apenas, probabilidades de transição.

Se estas probabilidades são independentes de n, a

cadeia é homogênea e as probabilidades $Pr[X_n=j|X_{n-1}=i]$ podem ser denotadas por p_{ij}. A matriz formada pela atribuição de p_{ij} à entrada (i,j) é conhecida como matriz de transição ou matriz da cadeia.

Em cadeias homogêneas as probabilidades de transição de m-passos

$$Pr[X_{n+m} = j|X_n=i] = p_{ij}^{(m)} \quad ,$$

também são independentes de n, e a probabilidade incondicional do estado j, na n-ésima transição é expressa por

$Pr[X_n=j] = p_j^{(n)}$, tal que a distribuição inicial é dada por $p_j^{(0)}$.

A seguir são enunciadas diversas definições pertinentes a cadeias de Markov.

Dois estados i e j são comunicantes ($i \leftrightarrow j$) se i é acessível a partir de j ($j \rightarrow i$) e vice-versa ($i \rightarrow j$). Uma cadeia é denominada irredutível se todos os seus estados se comunicam, ou seja, se existe um n tal que $p_{ij}^{(n)}>0$, para todos os pares (i, j).

O período de retorno a um estado k é definido como o máximo divisor comum (MDC) do conjunto de inteiros {n} para o qual $p_{kk}^{(n)}>0$. Um estado é aperiódico se este MDC é 1, ou seja, possui período unitário. Uma cadeia é denominada aperiódica se

todos os seus estados são aperiódicos.

Seja $f_{jj}^{(n)}$ a probabilidade de que uma cadeia retorne, pela primeira vez, ao estado j, em n transições, partindo do estado j. Então, a probabilidade de retorno a j (em algum tempo) é dada por

$$f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)} . \qquad (3.2.13)$$

Se $f_{jj} = 1$, j é um estado recorrente; se $f_{jj}<1$, j é um estado transitório. Quando $f_{jj}=1$, define-se

$$m_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)} , \qquad (3.2.14)$$

como o tempo médio de recorrência. Se $m_{jj}<\infty$, j é um estado recorrente positivo, enquanto que, se $m_{jj}=\infty$, j é um estado recorrente nulo.

Define-se $f_{ij}^{(n)}$, $i\neq j$, como a probabilidade de ocorrência da primeira passagem do estado i para o j, em exatamente n passos. Então, a probabilidade de que o estado j seja (em algum tempo) atingido a partir de i, é dada por

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)} . \qquad (3.2.15)$$

O valor esperado da sequência $\{f_{ij}^{(n)}, n=1,2,\ldots\}$ de probabilidades de primeira passagem para um dado par (i,j), $i\neq j$, é denotado por m_{ij}, e denominado tempo médio de primeira passa-

gem, ou seja

$$m_{ij} = \sum_{n=1}^{\infty} n f_{ij}^{(n)} \quad , \; i \neq j;$$ quando $i=j$, m_{ij} torna-se o tempo médio de recorrência do estado i.

Uma cadeia aperiódica, irredutível e recorrente positiva é dita ergódica.

Uma distribuição de probabilidades $\{\Pi_j, \; j \in C\}$ é uma distribuição estacionária se $\Pi_j = \sum_{i=0}^{\infty} \Pi_i \, p_{ij}$, $j \in C$, onde C denota uma cadeia de Markov.

Uma cadeia de Markov possui uma distribuição limite se existe uma distribuição de probabilidades $\{\Pi_j, \; j \in C\}$ com a propriedade

$$\lim_{n \to \infty} p_{ij}^{(n)} = \Pi_j \qquad (\forall i,j) \; . \qquad (3.2.16)$$

Com base nestas definições, são então enunciados (sem demonstração, ver /GROSD74, Ap.4/ e respectivas referências) alguns teoremas relevantes, relativos a cadeias de Markov.

T1: Seja C uma cadeia irredutível; então C é recorrente ou transitória, isto é, todos os seus estados são, ou recorrentes, ou transitórios.

T2: Se C é uma cadeia irredutível recorrente, então todos os seus estados são, ou positivos, ou nulos.

T3: Em uma cadeia irredutível e aperiódica as probabilidades limite

$$\lim_{n\to\infty} \Pr[X_n=j]=\Pi_j \quad , \quad (\forall j), \tag{3.2.17}$$

sempre existem e são independentes da distribuição dos estados iniciais. Se todos os estados são transitórios ou recorrentes nulos, então $\Pi_j=0$, $\forall j$, e não existe distribuição estacionária. Se, entretanto, todos os estados são recorrentes positivos (ergódicos), então $\Pi_j>0$, $\forall j$, e $\{\Pi_j\}$ é uma distribuição de probabilidade, onde $\Pi_j = 1/m_{jj}$. Esta distribuição limite é solução única do sistema de equações estacionárias

$$\Pi_j = \sum_{i=0}^{\infty} \Pi_i p_{ij} \quad , \text{ com } \quad \sum_{i=0}^{\infty} \Pi_i=1 \ . \tag{3.2.18}$$

T4: Uma cadeia irredutível e aperiódica é ergódica se existe uma solução não-negativa para o sistema

$$\sum_{j=0}^{\infty} p_{ij}x_j \le x_i-1 \quad , \quad (i\neq 0), \text{ tal que}$$

$$\sum_{j=0}^{\infty} p_{oj}x_j<\infty \ . \tag{3.2.19}$$

T5: Um sistema aperiódico e irredutível é ergódico se, e somente se, existe uma solução não-nula pa

ra as equações

$$\sum_{j=0}^{\infty} x_j p_{ij} = x_i \quad , \text{ tal que}$$

$$\sum_{j=0}^{\infty} |x_j| < \infty \quad . \tag{3.2.20}$$

Este teorema T5 conclui esta apresentação de conceitos, definições e resultados em cadeias de Markov (Item III.2.2), ressaltando-se que, além da conceituação básica, os itens relativos à classificação dos estados e das cadeias em si, bem como aqueles que tratam de aspectos de ergodicidade e existência e solução de uma distribuição estacionária, mostram-se fundamentais no entendimento da análise e solução de modelos de filas, mesmo M/G/1, conforme será visto na Seção III.3 seguinte, particularmente, Item III.3.1.

III.2.3 Processos de Renovação

Sejam consideradas as seguintes situações:

1) Mensagens chegam, individualmente, a um sistema de processamento de dados. Seja N(t) o número de chegadas durante (0,t].

2) Componentes de equipamentos eletrônicos são substituídos, assim que detetadas falhas. Seja N(t) o número de substituições realizadas no intervalo (0,t].

Nas situações acima, há interesse na observação da

ocorrência de um certo evento, e N(t) é o número destes eventos que ocorrem no tempo (0,t]. Trata-se de um processo estocástico não-decrescente, a estado discreto, que pode ser denominado um processo de contagem. Se os intervalos de tempo entre épocas consecutivas de ocorrência do evento são independentes e idênticamente distribuídos, então N(t) é um processo (de contagem) de renovação, e o evento de interesse é dito evento de renovação.

Formalmente, sejam {N(t), $t \geq 0$} um processo de contagem, e X_n o intervalo de tempo entre o (n-1)-ésimo e o n-ésimo evento deste processo, $n \geq 1$.

Se os termos da sequência de variáveis aleatórias não-negativas $\{X_1, X_2, \ldots\}$ são independentes e idênticamente distribuídos (i.i.d.), então {N(t), $t \geq 0$} é um processo de renovação.

Entre diversos tópicos e caracterizações de interesse, a teoria da renovação /SMITW58/ visa o estudo da distribuição de N(t) e respectivo valor esperado de renovações no tempo t, E[N(t)]; quando o parâmetro tempo é contínuo, E[N(t)] é conhecida como função de renovação .

A seguir, são apresentados, a partir de /BHATU72 , Cap.8/ formulação e resultados de teoremas relativos a N(t) e a E[N(t)].

1. Processo de Renovação a Tempo Discreto

Seja uma sequência de tentativas repetidas com saí

das possíveis E_i, e um evento de interesse particular E_*. Uma renovação ocorre na n-ésima tentativa se, e somente se, E_* ocorre na n-ésima tentativa, e se as saídas posteriores a esta n-ésima tentativa ocorrem independentemente das tentativas anteriores. O intervalo entre duas épocas consecutivas de ocorrência de E_* (número de tentativas) é denominado período de renovação do processo (ou tempo de espera para a primeira ocorrência de um evento de renovação).

Sejam $p^{(n)}$ a probabilidade de que E_* ocorra após a n-ésima tentativa, e $f^{(n)}$ a probabilidade de que E_* ocorra pela primeira vez na n-ésima tentativa.

Sejam ainda

$$p^{(0)} = 1 \qquad e \qquad f^{(0)} = 0 \tag{3.2.21}$$

e

$$f^* = \sum_{n=1}^{\infty} f^{(n)} \quad , \quad \text{onde} \tag{3.2.22}$$

f^* é a probabilidade de eventual ocorrência de E_*.

Quando $f^*=1$ (evento recorrente), seja Z_r uma variável aleatória própria que representa o comprimento do r-ésimo período de renovação, e $\{f^{(n)}\}$ a sua distribuição de probabilidade, onde

$$E(Z_r) = \sum_{n=1}^{\infty} nf^{(n)} = \mu \quad , \tag{3.2.23}$$

$r = 1,2,\ldots$.

Seja $f_{(r)}^{(n)}$ a probabilidade de que E_* ocorra, pela r-ésima vez, na n-ésima tentativa. Claramente, para r=2,

$$f_{(2)}^{(n)} = f^{1}f^{(n-1)} + f^{(2)}f^{(n-2)}+\ldots+f^{(n-1)}f^{1}. \quad (3.2.24)$$

Generalizando este resultado, para o tempo de espera até a r-ésima ocorrência, um teorema da teoria da renovação fornece:

T6: Seja S_r o tempo de espera até a r-ésima ocorrência de um evento de renovação E_*, e Z_i o seu período de renovação. Ainda,

$$f^{(n)} = Pr(Z_i=n) , \quad i=1,2,... \quad (3.2.25)$$

$$f_{(r)}^{(n)} = Pr(S_r=n). \quad (3.2.26)$$

Então, $S_r=Z_1+Z_2+\ldots+Z_r$ (3.2.27)

e $$f_{(r)}^{(n)} = \{f^{(n)}\}^{r\circledast} , \quad (3.2.28)$$

onde $r^{\circledast}$ indica convolução de ordem r.

A partir de

$$F(z) = \sum_{n=1}^{\infty} f^{(n)}z^n , \ |z|<1, \quad (3.2.29)$$

usando-se propriedades da f.g.p. de somas de v.a.'s independentes, tem-se

$$F_{(r)}(z) = \sum_{n=1}^{\infty} f_{(r)}^{(n)} z^n = [F(z)]^r. \quad (3.2.30)$$

Notar que, na eq.(3.2.21), considera-se $p^{(0)}=1$, ou seja, o tempo inicial é uma época de renovação, o que não necessariamente é uma premissa realista em muitas situações práticas. De forma a remover esta restrição, seja a distribuição do período de renovação inicial

$$b^{(n)}(k) = \Pr(Z_1=n \mid \text{última renovação ocorreu k tentativas antes de 0}),\ n=1,2,\ldots . \quad (3.2.31)$$

Relacionando com $\{p^{(n)}\}$ e $\{f^{(n)}\}$ tem-se

$$p^{(n)} = b^{(n)} + \sum_{r=1}^{n} p^{(n-r)}\, f^{(r)} \ , \ n\geq 1 \ , \quad (3.2.32)$$

que é denominada equação discreta de renovação.

A partir da eq. (3.2.32), definidas as f.g.p.'s

$$P(z) = \sum_{n=0}^{\infty} p^{(n)}\, z^n \ , \ |z|<1 \quad (3.2.33)$$

e

$$B(z) = \sum_{n=0}^{\infty} b^{(n)}\, z^n \ , \ |z|<1$$

tem-se $P(z) = B(z) + F(z)\, P(z)$, donde

$$P(z) = \frac{B(z)}{1-F(z)} \ . \quad (3.2.34)$$

No caso de uma primeira ocorrência de E_* em n=0, tem-se

$$P(z) = \frac{1}{1-F(z)} \ . \quad (3.2.35)$$

2. Processo de Renovação a Tempo Contínuo

Seja um evento E_* que ocorre nos instantes $t_1, t_2, t_3, \ldots$, e sejam $Z_r = t_r - t_{r-1}$ $(r=2,3,\ldots)$ variáveis aleatórias i.i.d. com

$$Pr(Z_r \leq x) = F(x) \tag{3.2.36}$$

e

$$E(Z_r) = \mu \ , \tag{3.2.37}$$

onde $t_o = 0$ e $Z_1 = t_1 - t_o$ é distribuída segundo

$$Pr(Z_1 \leq x) = F_1(x) \ . \tag{3.2.38}$$

Seja $N(t)$ o número de vezes que E_* ocorre no tempo $(0,t]$. O processo $N(t)$ é um processo de renovação a tempo contínuo e $Z_r (r=1,2,\ldots)$ são os períodos de renovação.

Como $F_1(x)$ não necessáriamente coincide com $F(x)$, a renovação inicial não coincide obrigatóriamente com o instante da observação inicial. Seja

$$S_r = Z_1 + Z_2 + \ldots + Z_r \qquad (r=1,2,\ldots) \tag{3.2.39}$$

o tempo necessário para a ocorrência de r renovações. A distribuição de S_r pode ser expressa por

$$Pr(S_r \leq x) = F_1(x) \circledast F_{r-1}(x) \ , \tag{3.2.40}$$

onde $\circledast$ denota convolução, e $F_{r-1}(x)$ representa a convolução de

ordem (r-1) de F(x) consigo própria, $F_r^{(0)} = 0$ para r>0 e $F_r^{(0)}=1$ para r=0.

Definem-se as transformadas de Laplace-Stieltjes

$$\phi(\theta) = \int_0^\infty e^{-\theta x}\, dF(x) \quad , \quad Re[\theta]>0 \qquad (3.2.41)$$

e

$$\phi_1(\theta) = \int_0^\infty e^{-\theta x}\, dF_1(x) \quad , \quad Re[\theta]>0 \quad , \qquad (3.2.42)$$

e, a partir da eq. (3.2.40), tem-se

$$\int_0^\infty e^{-\theta x}\, d_x Pr(S_r \leq x) = \phi_1(\theta).[\phi(\theta)]^{r-1} \; . \qquad (3.2.43)$$

Suponha-se que um processo de renovação seja observado no instante t. Surgem as seguintes questões: (a) qual a probabilidade de ocorrência de E_* no intervalo infinitesimal $(t,t+\Delta t)$? (b) qual a distribuição do tempo decorrido desde a última renovação? (c) qual a distribuição do tempo até o próximo ponto de renovação?

Estes dois últimos tempos são denominados <u>tempo de ocorrência para trás</u> ("backward") e tempo <u>de ocorrência para frente</u> ("forward"), respectivamente.

No caso geral, não são deriváveis expressões explícitas para estes tempos (estes serão mostrados para o processo de Poisson no Item III.2.4 seguinte, a título de exemplo), sen-

do, entretanto, disponíveis equações integrais e formas limites, alguns destes resultados mostrados a seguir.

T7: Para um dado valor de t, N(t) é uma variável aleatória própria, e sua distribuição é dada por

$$P_n(t) \triangleq Pr[N(t)=n] = F_n(t) - F_{n+1}(t), \quad n=0,1,2... \qquad (3.2.44)$$

e

$$U(t) \triangleq E[N(t)] = \sum_{n=1}^{\infty} F_n(t), \qquad (3.2.45)$$

onde, por convenção, /BHATU72, Cap.8/, $F_n(t) = F_1(t) \circledast F_{n-1}(t)$.

A função U(t) é denominada, na literatura, função de renovação, e constitui uma das características mais importantes de um processo contínuo. Sua primeira derivada é denominada densidade de renovação, e expressa a densidade de probabilidade de uma renovação no tempo t.

Através de técnicas transformadas e convoluções obtem-se, respectivamente, as seguintes expressões para estas funções:

$$U(t) = F_1(t) + \int_0^t U(t-\tau)dF(\tau) \qquad (3.2.46)$$

e

$$u(t) = f_1(t) + \int_0^t u(t-\tau)f(\tau)d\tau. \qquad (3.2.47)$$

Em termos de comportamento limite da densidade de

renovação, mostra-se que

$$U(t+\Delta) - U(t) \to \frac{\Delta}{\mu} \quad , \; t \to \infty \tag{3.2.48}$$

e, logo,

$$u(t) \to \frac{1}{\mu} \quad , \; t \to \infty \; . \tag{3.2.49}$$

Quanto ao processo em regime estacionário (em operação por período suficientemente longo), mostra-se que

$$u(t) = \frac{1}{\mu} \quad , \tag{3.2.50}$$

quando o período de renovação inicial assume a forma do comportamento limite do tempo de recorrência para frente, tornando-se, assim, independente do parâmetro tempo. Ainda

$$E[N(t)] \triangleq U(t) = \frac{t}{\mu} \tag{3.2.51}$$

e

$$E[N(s+t) - N(s)] = \frac{t}{\mu} \; . \tag{3.2.52}$$

Ocorre que $f_1(x) = [1-F(x)]/\mu$ constitui condição necessária para a estacionariedade do processo de renovação.

III.2.4 Exemplos de Aplicações

Identificados processos de renovação, aplicações de grande interesse prático de análise, consistem em, dada a distribuição de N(t) obter probabilidades e tempos relevantes, e no sentido inverso, obter a distribuição, ou momentos, de N(t),

a partir de, por exemplo, conhecimento dos tempos entre ocorrências de renovação.

Resolvidos, ou equacionados, estes problemas pode-se chegar à caracterização de processos de renovação compostos, onde o evento E_* é também uma variável aleatória, com distribuição conhecida, por exemplo, épocas Poisson de chegadas de mensagens, cujo comprimento segue uma distribuição geométrica, gamma, ou constante.

A seguir são apresentados exemplos referentes a estas situações descritas acima.

1. Processo Poisson

Um dos mais utilizados processos de contagem de renovação é o processo de Poisson, de parâmetro λ, para o qual são obtidos:

$$Pr[N(t)=n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \quad n=0,1,2,... \qquad (3.2.53)$$

e

$$E[N(t)] = \lambda t \, . \qquad (3.2.54)$$

A probabilidade Pr[uma renovação ocorra entre $(t,t+\Delta t)] = \lambda\Delta t + o(\Delta t)$, logo

$$F(x) = F_1(x) = 1-e^{-\lambda x} \quad , \quad x \geq 0 \qquad (3.2.55)$$

$$Pr(S_r \leq x) = \int_0^{\infty} e^{-\lambda y} \frac{\lambda^r y^{r-1}}{(r-1)!} \, dy \, . \qquad (3.2.56)$$

Então os tempos de ocorrência para trás, S(t), e para frente, R(t), são dados por

$$\Pr\,[R(t) \leq x] = 1 - e^{-\lambda x} \qquad (3.2.57)$$

$$\Pr\,[S(t) = t] = e^{-\lambda t}$$

$$\Pr\,[S(t) \leq x] = 1 - e^{-\lambda x} \quad , \quad 0 \leq x < t \; . \qquad (3.2.58)$$

Uma observação importante é, então, que um processo de renovação do tipo Poisson é composto por intervalos entre ocorrências de renovação exponencialmente distribuídos. A mesma observação é válida para R(t) e S(t).

2. Distribuição Gamma de Tempos entre Ocorrências

A distribuição gamma é uma família de distribuições a dois parâmetros, que pode ser usada para aproximar quase qualquer distribuição geral de tempos entre chegadas. Consequentemente, processos de contagem de renovação gerados por tempos entre chegadas com distribuição gamma, são frequentemente de interesse, especialmente, uma vez que as probabilidades de tais processos podem ser facilmente computadas.

Seja $\{N(t), t \geq 0\}$ um processo de renovação correspondente a tempos entre chegadas i.i.d. segundo uma distribuição gamma com parâmetros $\lambda > 0$ e $k = 1, 2, \ldots$. Assim

$$f(t) = \frac{\lambda}{(k-1)!} (\lambda t)^{k-1} e^{-\lambda t} \quad , \quad t > 0$$

$$= 0 \quad , \quad t \leq 0 \; .$$

Mostra-se que /PARZE62, Cap.5/ a f.g.p. de N(t) é expressa por

$$\Psi(z,t)=1+\left(\frac{z-1}{z}\right)\frac{1}{k}\sum_{r=0}^{k-1}\frac{z^{1/k}\varepsilon^{r}}{1-z^{1/k}\varepsilon^{r}}\{1-\exp[-\lambda t(1-z^{1/k}\varepsilon^{r})]\}, \quad (3.2.59)$$

onde $\varepsilon = \exp\left(\frac{2\pi i}{k}\right)$, $\varepsilon^{o}=1$, e $\varepsilon^{r}=\exp\left(\frac{2\pi i r}{k}\right)$. (3.2.60)

Como ilustração, sejam os casos k=1 e k=2; o caso k =1 corresponde a tempos entre chegadas exponencialmente distribuídos (processo Poisson). Da eq.(3.2.59) obtém-se

$$\Psi(z,t) = 1+\left(\frac{z-1}{z}\right)\left(\frac{z}{1-z}\right)\{1-\exp[-\lambda t(1-z)]\}$$

$$= \exp[\lambda t(z-1)] \text{ , que é a f.g.p. de uma distribuição Poisson (ver eq.(3.1.2), Item III.1.2, Cap.III).}$$

No caso k=2, mostra-se que /PARZE62, Cap.5/ a eq. (3.2.59) se reduz a

$$\Psi(z,t) = e^{-\lambda t}\{\cosh(\lambda t\sqrt{z}) + \frac{1}{\sqrt{z}}\sinh(\lambda t\sqrt{z})\}, \quad (3.2.61)$$

e que o processo de renovação associado, $\{N(t)\ t\geq 0\}$, possui tempos entre chegadas i.i.d., com f.d.p.

$$f(x) = \lambda^{2} x e^{-\lambda x} \quad , \quad x>0 \quad , \quad (3.2.62)$$

e média e variância, respectivamente,

$$E[N(t)] = \frac{\lambda}{2}\, t - \frac{1}{4} - \frac{1}{4}\, e^{-2\lambda t} \qquad (3.2.63)$$

$$Var[N(t)] = \frac{\lambda}{4}\, t - \frac{\lambda}{2}\, t\, e^{-2\lambda t} + \frac{1}{4}\, e^{-\lambda t} \sinh \lambda t$$

$$- \frac{1}{4}\, e^{-2\lambda t}\, \sinh^2 \lambda t. \qquad (3.2.64)$$

Caracterizados processos de renovação, por meio de 1º e 2º momentos, pode-se realizar a composição de processos , conforme a propriedade da soma de uma coleção de v.a'.s i.i.d.'s, onde o número de parcelas é uma v.a.d., conforme

$$H^*(s) = G[F^*(s)]. \qquad (3.2.65)$$

(ver Seção III.1, Cap. III).

3. Processo de Renovação Composto - Poisson + Geométrica

Pelos resultados obtidos na Seção III.1, Item III.1.2, tem-se que a média e a variância de um processo de renovação composto, deste tipo, podem ser obtidas conforme a seguir:

$$\text{POISSON:}\quad \bar{X}_p = \lambda \quad , \quad \sigma_p^2 = \lambda \quad \text{e} \quad g_p(z) = e^{\lambda(z-1)}$$

$$\text{GEOMÉTRICA:}\quad \bar{X}_\lambda = \frac{1}{1-q} \quad , \quad \sigma_g^2 = \frac{q}{p^2} \quad \text{e} \quad g_g(z) = \frac{(1-q)z}{1-qz}$$

A média, $\bar{X}_s$, e a variância, σ_s^2 , do processo com-

posto resultante, podem ser calculados através da eq.(3.2.65) , ou, diretamente, por meio da relação entre os momentos das variáveis envolvidas.

Utilizando-se a relação transformada tem-se:

$$g_s(z) = e^{\lambda\left[\frac{(1-q)z}{1-qz} - 1\right]} ,$$

$$g_s'(1) = \lambda\{1 + \frac{q}{1-q}\} = \frac{\lambda}{1-q} = \bar{X}_s$$

$$g_s''(1) = \frac{\lambda^2+2q\lambda}{(1-q)^2} = \overline{X_s^2} - \bar{X}_s \ , \quad \text{logo,}$$

$$\sigma_s^2 = \frac{\lambda^2+2q\lambda}{(1-q)^2} + \frac{\lambda}{1-q} - \frac{\lambda^2}{(1-q)^2} = \frac{\lambda(1+q)}{(1-q)^2} .$$

Através da relação explícita entre os momentos (ver Seção III.1, Item III.1.1), obtêm-se

$$\bar{X}_s = \bar{X}_p \ . \ \bar{X}_g = \lambda \ . \ \frac{1}{1-q} = \frac{\lambda}{1-q}$$

e

$$\sigma_s^2 = \bar{X}_p . \sigma_g^2 + (\bar{X}_g)^2 . \sigma_p^2 = \frac{\lambda q}{p^2} + \frac{\lambda}{(1-q)^2} = \frac{\lambda(1+q)}{(1-q)^2} ,$$

que são, naturalmente, idênticos aos obtidos por meio da relação transformada, eq. (3.2.65).

Recomenda-se a leitura de /CHUWW72/ e /KOBAH77/ para contato com outras aplicações desta natureza.

SEÇÃO III.3

TÓPICOS EM TEORIA DE FILAS M/G/1

Nesta Seção III.3 são apresentados tópicos relevantes na análise de filas M/G/1, em particular, aqueles presentes na formulação de análises de modelos de multifilas,visando subsidiar o entendimento e eventuais envolvimentos posteriores com modelos correlatos.

III.3.1 - FILAS M/G/1 E CADEIAS DE MARKOV IMERSAS

Dado que a variável de estado associada a uma fila simples é, geralmente, o número de itens em fila, incluindo aquele que, no instante de observação, é atendido pela unidade de serviço, ou seja, o número de itens no sistema, N(t), e que o relaxamento da restrição de serviço exponencialmente distribuído implica na especificação do tempo de serviço decorrido,ou do residual, para o item em serviço, verifica-se que, para modelos M/G/1, não é possível um tratamento, relativamente simples, através de análise baseada em equações de Chapman-Kolmogorov , conforme realizado para filas markovianas /GROSD74, Cap. 3 e 4/ e /KLEIL75, Cap. 2 e 3/.

Dentre diversos métodos para a solução deste modelo, resumidos em /KLEIL75 Cap. 5/, destaca-se o método das cadeias de Markov imersas (CMI), desenvolvido por Palm /PALMC43/ e Kendall /KENDD51/, que permite, então, o emprego de elementos da

teoria de cadeias de Markov, e a conservação da variável de estado discreta unidimensional N(t).

O sistema é então observado em uma sequência de pontos imediatamente após os instantes de completação de serviço, ou partidas de itens da fila, uma vez que, nestes instantes, garante-se que o tempo de serviço decorrido de um eventual item em serviço, é implicitamente nulo.

Sejam B(t) e b(t), respectivamente, as f.d.a. e f.d.p. da variável aleatória que descreve os tempos de serviço, e $1/\mu$ o seu valor médio. O processo de chegada é Poisson, com parâmetro λ. O processo estocático imerso $X(t_i)$, onde X denota o estado do sistema e $t_1, t_2, t_3, \ldots$, são instantes de partida sucessivos, pode ser demonstrado markoviano, segundo o seguinte raciocínio: como t_i é o instante de partida do i-ésimo item, então, $X(t_i)$ representa o número de itens deixados em fila por este i-ésimo item. Como o espaço de estado é discreto, simplifica-se a notação $X(t_i) \equiv X_i$ e pode-se escrever, para todo $n > 0$ que

$$X_{n+1} = \begin{cases} X_n - 1 + A_{n+1} & (X_n \geq 1) \\ A_{n+1} & (X_n = 0) \end{cases}, \qquad (3.3.1)$$

o que equivale a

$$X_{n+1} = X_n - U(X_n) + A_{n+1} \qquad (X_n \geq 0), \qquad (3.3.2)$$

onde

$$U(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$ representa a função unitária

de Heaviside, X_n o número no sistema no n-ésimo instante de partida, e A_{n+1} o número de chegadas ocorridas durante o tempo de serviço $S^{(n+1)}$, do (n+1)-ésimo item. Como a variável $S^{(n+1)}$ é, por suposição, independente de tempos de serviço prévios e do comprimento da fila, a mesma pode ser denotada por S, e ainda, como as chegadas são Poisson, a variável aleatória A_{n+1} depende apenas de S, recendo a notação de A. Segue-se então que

$$Pr[A=a] = \int_0^\infty Pr[A=a \mid S = t\,]b(t)\,dt \quad \text{e}$$

$$Pr[A=a|S=t] = \frac{e^{-\lambda t}(\lambda t)^a}{a!}, \text{ de forma que}$$

$$Pr[X_{n+1}=j|X_n=i] = Pr[A=j-i+1] \text{ , o que resulta em}$$

$$Pr[X_{n+1}=j|X_n=i] = \begin{cases} \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^{j-i+1}}{(j-i+1)!}\, b(t)dt & (j \geq i-1,\ i \geq 1) \\ 0 & (j < i-1,\ i \geq 1). \end{cases} \tag{3.3.3}$$

Se um cliente em partida deixar o sistema vazio, o estado permanece nulo até uma próxima chegada, de forma que as probabilidades de transição para o caso i=0 são idênticas áquelas para i=1. Pode-se então notar que o processo imerso é markoviano, uma vez que apenas os indices (i,j) estão envolvidos na eq. (3.3.3); ainda, como a variável de estado é discreta, trata-se de uma cadeia de Markov imersa (CMI).

Seja,então, Π_n a probabilidade estacionária do estado n em um ponto da CMI, e p_n a probabilidade estacionária do estado n em um instante arbitrário de tempo (estas probabilidades estacionárias, ou de equilíbrio, existem a partir do instante no qual o sistema atinge o estado estacionário, o que é, neste momento, pressuposto para continuação da análise; a existência deste equilíbrio é discutida no final deste Item III.3.1 . Em geral, as probabilidades $\{\Pi_n\}$ e $\{p_n\}$ são diferentes. Entretanto, o interesse no estudo da CMI que gera as probabilidades $\{\Pi_n\}$, fundamenta-se no fato de se poder demonstrar que,em sistemas M/G/1, as probabilidades $\{\Pi_n\}$ e $\{p_n\}$ são idênticas /GROSD 74, Cap. 5/5.1,3/. Adicionalmente, as probabilidades $\{q_n\}$, que descrevem o sistema em regime estacionário, nos instantes de ocorrência de chegadas, também são idênticas às $\{p_n\}$ /KLEIL75 , Cap. 5/5.3/.

A matriz de probabilidades de transição (MPT) da CMI identificada acima é denotada por

$$\mathbb{P} = [p_{ij}] \quad , \text{ onde } \quad p_{ij} = Pr[X_{n+1}=j \mid X_n=i] \quad .$$

Da eq. (3.3.3) resulta que

$$p_{ij} = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^{j-i+1}}{(j-i+1)!}\, b(t)\, dt \quad , (j \geq i-1, i \geq 1) \, , \quad \text{e}$$

definindo-se $k_n \triangleq$ Pr [n chegadas durante um serviço S=t] ,

$$k_n \triangleq \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, b(t)\, dt \, , \quad \text{de forma que } p_{ij} \triangleq k_{j-i+1}, \quad \text{ob-}$$

tem-se a MPT

$$\mathbb{P} = [p_{ij}] = \begin{bmatrix} k_o & k_1 & k_2 & \cdots \\ k_o & k_1 & k_2 & \cdots \\ 0 & k_o & k_1 & \cdots \\ 0 & 0 & k_o & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} .$$

Pressuposta a existência de estado estacionário, o vetor de probabilidades estacionárias $\boldsymbol{\Pi} = \{\Pi_n\}$ pode ser obtido como solução do sistema de equações estacionárias (ver Item III.2.2, T3), resultando em

$$\Pi_i = \Pi_o k_i + \sum_{j=1}^{i+1} \Pi_j \, k_{i-j+1} \qquad (i=0,1,2,\ldots) . \quad (3.3.4)$$

As probabilidades $\{\Pi_i\}$ podem então ser utilizadas na derivação da f.g.p. do número de itens no sistema, definida como

$$Q(z) = \sum_{i=0}^{\infty} \Pi_i \, z^i \quad , \qquad (|z| \leq 1) \quad , \qquad (3.3.5)$$

a partir da qual podem ser obtidos momentos da distribuição desta variável de estado.

Neste ponto, são indicados dois métodos, não totalmente disjuntos, para a obtenção de Q(z):

a) Gross e Harris /GROSD74, Cap. 5/ definem, sobre os coeficientes da MPT, a seguinte f.g.p.:

$K(z) \triangleq \sum_{i=0}^{\infty} k_i z^i$, e utilizam a expressão dos $\{\Pi_i\}$ na eq. (3.3.4), para substituição direta na eq. (3.3.5), obtendo

$$Q(z) = \frac{\Pi_o (1-z) K(z)}{K(z) - z}$$. Definindo $\rho = \lambda/\mu$, e utilizando as propriedades $Q(1)=1$, $K(1)=1$ e $K'(1)=\rho$, chega-se a $\Pi_o = 1-\rho$, o que resulta em

$$Q(z) = \frac{(1-\rho)(1-z)K(z)}{K(z) - z} . \qquad (3.3.6)$$

Como $Q'(1)$ representa o valor esperado do estado do sistema ou comprimento da fila, denotado por L, pode-se,pela definição calcular

$$L = \rho + \frac{\rho^2+\lambda^2 \sigma_s^2}{2(1-\rho)} , \qquad (3.3.7)$$

onde σ_s^2 é a variância do tempo de serviço. Esta expressão é conhecida como a equação Pollaczek-Khintchine (P-K) do valor médio (esta atribuição é relativamente recente, e autores como Cox e Smith /COXDR61/ referem-se apenas a Pollaczek), a partir da qual podem ser obtidas outras grandezas de interesse usando-se a fórmula de Little, tais como o valor médio do tempo de espera, o valor médio do tempo total no sistema (espera + serviço), etc.

b) Kleinrock /KLEIL75, Cap. 5/ parte diretamente da equação de estado, eq. (3.3.2) e, inicialmente, elevando-a ao

quadrado, aplicando o operador $E[\cdot]$ a ambos os membros, e tomando o limite quando $n \to \infty$ (segundo um método sugerido por Kendall), e chega a uma expressão intermediária para a equação P-K do valor médio, que, por sua forma, induz a definição de uma f.g.p. $V(z)$ (em realidade análoga a $K(z)$ em a)), cuja relação com a t.L. da f.d.p. $b(t)$, $B^*(s)$, é expressa por

$$V(z) = B^*(\lambda - \lambda z). \qquad (3.3.8)$$

Utilizando as propriedades de geração de momentos apresentadas por $V(z)$ e $B^*(s)$, e a relação eq. (3.3.8), obtém-se uma forma da equação P-K do valor médio, em termos do coeficiente de variação quadrado do tempo de serviço, C_b^2, onde $\bar{q}$ denota o valor médio do comprimento da fila

$$\bar{q} = \rho + \rho^2 \frac{(1+C_b^2)}{2(1-\rho)} \quad . \qquad (3.3.9)$$

Desenvolvimento análogo ao descrito acima, baseado na definição da f.g.p. $Q(z)$, e utilizando ambos os membros da eq. (3.2.2) como expoentes para a variável complexa $z(a=b \Rightarrow z^a=z^b)$, resulta em:

$$Q(z) = B^*(\lambda - \lambda z) \frac{(1-\rho)(1-z)}{B^*(\lambda - \lambda z) - z} \quad , \qquad (3.3.10)$$

denominada de equação transformada de P-K, e que equivale à equação (3.3.6), tomando-se $K(z)=V(z)=B^*(\lambda - \lambda z)$ (novamente, esta denominação é proposta por Kleinrock, não sendo, entretanto, adotada universalmente, nem encontrada uniformemente em autores clássicos anteriores).

Além das abordagens a) e b) aqui resumidas, recomenda-se um estudo completo das mesmas e de suas referências secundárias. Particularmente simples e elucidativa mostra-se a monografia /COXDR61/, indicada para estudo prévio a trabalhos mais especializados.

As condições para existência de estado estacionário, ou de equilíbrio, em filas M/G/1, ou seja, condições que garantam a ergodicidade do processo estocástico descrito pela variável de estado, até o momento pressupostas, e o respectivo processo de derivação/demonstração, podem ser encontradas em /GROSD74, Cap. 5/5.14/. Determina-se que $\rho < 1$ é condição necessária e suficiente para existência de solução de equilíbrio; a suficiência desta condição é demonstrada com base em teoremas relativos a cadeias de Markov, e a necessidade de $\rho < 1$ advém diretamente da existência da f.g.p. $Q(z)$ sobre o intervalo $|z| \leq 1$.

III.3.2 - PERÍODOS DE OCUPAÇÃO EM FILAS M/G/1

A evolução do estado de uma fila M/G/1, quando observada, apresenta ciclos alternantes de períodos de ocupação e repouso. O objetivo da análise destes períodos é a determinação das distribuições das durações dos mesmos, tarefa bastante trabalhosa, quando não impossível, o que leva, então à procura de obtenção de transformadas, as quais fornecem, pelo menos, uma caracterização parcial das distribuições, através de momentos.

Kleinrock /KLEIL75, Cap. 5/5.3-5.11/ apresenta uma con-

ceituação e um método de análise bastante detalhados e didáticos, baseado em uma mecânica de reordenação dos itens em fila , proposta por Takács /TAKAL62/. Os resultados obtidos são uma equação funcional da transformada de Laplace da distribuição do período de ocupação, geralmente não inversível, uma expressão explícita para a f.d.a. do período de ocupação, possível de avaliação numérica aproximada (envolve um somatório infinito!), e os quatro primeiros momentos do período de ocupação. São analisados também o número de itens despachados em um período de ocupação, e estabelecidas relações entre estes períodos e os tempos de espera.

A análise efetuada por Gross e Harris /GROSD74,Cap. 5/5.1.8/ mostra-se menos extensa e detalhada, utilizando, implicitamente, em sua argumentação a noção de sub-períodos de ocupação, de Takács, resultando na obtenção da mesma equação funcional, da qual são derivados os momentos do período de ocupação , fornecendo um quadro simples e sintético do conceito, considerada, em escôpo e propósitos, adequada para apresentação neste Item III.3.2.

Seja G(x) a f.d.a. do período de ocupação, X,de uma fila M/G/1 com f.d.a. de serviço B(t). A variável X é então condicionada à duração do primeiro tempo de serviço iniciando o período de ocupação, e, como cada chegada durante este tempo de serviço vai contribuir para o período de ocupação com novas chegadas, durante seus respectivos tempos de serviço, cada uma das chegadas que ocorre durante o primeiro serviço de um período de ocupação pode ser considerada, essencialmente, como a geratriz

de seu próprio período de ocupação. Pode-se então escrever

$$G(x) = \int_0^x \Pr[\text{ dado primeiro serviço } t, \text{ período de ocupação gerado por todas as chegadas durante } t \leq x - t\ ]\, dB(t)$$

$$= \int_0^x \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, G^{(n)}(x-t)\, dB(t)\ , \qquad (3.3.11)$$

onde $G^{(n)}(x)$ é a n-ésima convolução de $G(x)$. Seja então

$$G^*(s) \triangleq \int_0^{\infty} e^{-sx}\, dG(x)\ ,\ \text{a t.L.S. de } G(x)\ ,\ \text{e } B^*(s) \quad \text{a}$$

t.L.S. de $B(t)$. Levando-se ambos os membros da eq. (3.3.11) ao domínio transformado, tem-se

$$G^*(s) = \int_0^{\infty}\int_0^x \sum_{n=0}^{\infty} e^{-xs}\ \frac{e^{-\lambda t}(\lambda t)^n}{n!}\ G^{(n)}(x-t)\, dB(t)dx, \qquad (3.3.12)$$

que, após rearranjamento da ordem de integração, resulta em

$$G^*(s) = \int_0^{\infty} \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, dB(t) \int_t^{\infty} e^{-xs}\, G^{(n)}(x-t)\, dx. \qquad (3.3.13)$$

Aplicando-se a propriedade da convolução,

$$G^*(s) = \int_0^{\infty} \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, e^{-st}[G^*(s)]^n\, dBt \qquad (3.3.14)$$

$$= \int_0^{\infty} e^{-\lambda t}\, e^{\lambda t G^*(s)}\, e^{-st}\, dB(t), \text{ que equivale a}$$

$$G^*(s) = B^*\,[s + \lambda - \lambda G^*(s)] \; . \qquad (3.3.15)$$

A equação (3.3.15) é exatamente a obtida por Kleinrock, a partir da qual são obtidos os primeiros momentos do período de ocupação. Sejam $\bar{x} = 1/\mu$ e $\overline{x^2}$ os dois primeiros momentos do tempo de serviço. Diferenciando-se a eq. (3.3.15) em s=0, e resolvendo para o valor esperado, tem-se

$$E[X] = \frac{\bar{x}}{1-\rho} \; , \quad \text{onde} \quad \rho = \lambda\bar{x} = \lambda/\mu \; . \qquad (3.3.16)$$

O segundo momento é obtido através da segunda derivada em s=0, o que resulta em

$E[X^2] = \dfrac{\overline{x^2}}{(1-\rho)^3}$, de forma que a variância do período de ocupação, σ_X^2 , pode, então, ser calculada como

$$\sigma_X^2 \triangleq E[X^2] - (E[X])^2$$

$$= \frac{\overline{x^2}}{(1-\rho)^3} - \frac{(\bar{x})^2}{(1-\rho)^2} = \frac{\sigma_b^2 + \rho(\bar{x})^2}{(1-\rho)^3} \; , \qquad (3.3.17)$$

onde σ_b^2 representa a variância da distribuição do tempo de serviço.

No que diz respeito a este tópico, recomenda-se, além da bibliografia básica já citada, e respectivas referências secundárias, o tratamento dado por Cox e Smith /COXDR61, Cap.5/ 5.6/, incluindo duas abordagens/métodos distintos, de conceitua

ção bastante elucidativa, apresentando, por um lado, um grau de similaridade com os métodos discutidos em /GROSD74/ e /KLEIL75/, e, por outro lado, uma certa complexidade matemática, principalmente no segundo método (integrais de contorno, aproximações assintóticas, integral inversa de Laplace, aproximações por método de pontos, etc.).

III.3.3 FILAS M/G/1 COM INTERRUPÇÃO DE SERVIÇO

Na análise de modelos de filas considera-se geralmente que o número de unidades de serviço, ou taxa de serviço de uma única unidade, é constante, ou seja, mantém-se no tempo, segundo as especificações do(s) processo(s) estocástico(s) correspondente(s).

Entretanto, em muitas situações práticas, a capacidade global de realização de serviço de um sistema mostra-se variável, sendo esta variabilidade produzida por interrupções, geralmente aleatórias, do funcionamento da(s) unidade(s) de serviço, a(s) qual(is) se mantém inoperante(s) períodos também aleatórios.

Conforme sugerido por Avi-Itzhak em /AVIIB65/, o compartilhamento de uma única unidade de serviço por diversas filas em um sistema de multifilas pode ser posto em analogia com uma situação de interrupção de serviço, do ponto de vista de cada uma das filas componentes.

Em /AVIIB63/, Avi-Itzhak e Naor obtêm resultados para diversas situações em que interrupções de serviço são verificadas, sendo estes resultados utilizados por Avi-Itzhak em sua análise de multifilas (prioridades alternantes) apresentada em /AVIIB65/.

A seguir apresenta-se o desenvolvimento resumido do modelo E (Avi-Itzhak analisa 5 modelos diferentes),conforme /AVIIB 63/, com o objetivo de fornecer subsídios para melhor entendimento (eventualmente, posterior análise detalhada) do trabalho de Avi-Itzhak, considerado pelo autor desta tese, de formulação bastante interessante e original, além de constituir base potencial para estudos futuros, ainda no campo de interesse deste estudo, no sentido de refletir situações de quebra/reparo de dispositivos de serviço.

As premissas básicas para a análise dos modelos de interrupção de serviço /AVIIB63/ são:

1. Chegadas representadas por processos Poisson estacionários, de parâmetro ξ .
2. Processos de serviço arbitrários, com f.d.p. $f(t_s)$ e segundo momento finito.
3. O processo de quebra da unidade de serviço é Poisson, no sentido de que a duração de períodos de disponibilidade ininterrupta é uma v.a. exponencialmente distribuída.

4. O tempo de reparo é uma v.a. com f.d.p. $f(t_r)$ e segundo momento finito.

5. Os sistemas considerados operam em condições de não-saturação, garantindo assim o atingimento de estacionariedade.

Seja i (i=0,1,2,...) o número de itens no sistema, incluindo aquele que esteja (possívelmente) sendo servido. Seja v (v=0,1) o número de unidades de serviço que estão fora de operação (neste trabalho, Avi-Itzhak trata o caso de filas simples).

E_{vi} descreve um estado no qual v unidades estão inoperantes e i itens esperam pelo término de serviço, sendo p_{vi} a probabilidade deste estado. Assim a probabilidade da única unidade de serviço estar inoperante (estado $E_{1.}$) é dada por

$$p_{1.} = \sum_{i=0}^{i=\infty} p_{1i} \qquad (3.3.18)$$

e a probabilidade complementar (estado $E_{o.}$, disponível)

$$p_{o.} = 1 - p_{1.} = \sum_{i=o}^{i=\infty} p_{oi} \ . \qquad (3.3.19)$$

Ao longo do tempo, a unidade de serviço quebrará, enquanto atendendo a um item (será visto que, no caso do modelo E, a unidade só quebra quando desocupada). Se se considerar que este serviço passado e incompleto não é perdido, a fração de tempo durante a qual a unidade está ocupada, servindo itens, é igual a

$$b = \xi \ . \ E(t_s) \ . \qquad (3.3.20)$$

Como a fração de tempo durante a qual a unidade de serviço está disponível é $p_{o.}$, a condição de não-saturação é satisfeita por $b < p_{o.}$, ou

$$1 - p_{1.} - b = p_{oo} > 0. \qquad (3.3.21)$$

Seja a quantidade q_v o valor esperado do número de itens no sistema, dado que v unidades estão inoperantes. Então

$$q_v \triangleq E(i|v) = (1/p_{v.}) \sum_{i=o}^{i=\infty} i\, p_{vi} , \qquad (3.3.22)$$

e q, o valor esperado incondicional desta variável,

$$q = E(i) = \sum_{i=o}^{i=\infty} \sum_{v=o}^{v=1} i\, p_{vi} = q_o\, p_{o.} + q_1\, p_{1.} \qquad (3.3.23)$$

O modelo E considera que a unidade de serviço só pode quebrar, e requerer um tempo de reparo, em momentos em que o sistema esteja vazio (serviço exaustivo !).

O número médio de interrupções no tempo é dado por $\lambda(p_{o.}-b)$. Multiplicando-se por $E(t_r)$ obtém-se a probabilidade $p_{1.}$, da unidade estar inoperante. Assim

$$p_{1.} = \lambda(p_{o.}-b).E(t_r) = \lambda(1-p_{1.}-b).E(t_r)$$

$$= \lambda(1-b).E(t_r)/[1+ \lambda E(t_r)] , \qquad (3.3.24)$$

e

$$p_{o.} = [1+\lambda b E(t_r)]/[1+\lambda E(t_r)] . \qquad (3.3.25)$$

O valor médio do tempo de espera, θ_w, é expresso por

$$\theta_w = E(t_s) + p_{1.} E(T_r) + bE(T_s), \qquad (3.3.26)$$

onde: $w = \xi\theta$ (pela fórmula de Little);

$E(T_r)$ = valor esperado do atraso para frente ("forward delay", ver /AVIIB63, pp. 306/) de t_r, ou seja, o período esperado de tempo que decorre entre a chegada de um item a uma unidade inoperante e o término do processo de reparo;

$E(T_s)$ = valor esperado do atraso para frente ("forward delay", ver /AVIIB63, pp.306/) de t_s, ou seja, o período de tempo esperado que decorre entre a chegada de um item a uma unidade ocupada e o término de serviço do item corrente.

Por definição, w representa o comprimento médio da fila. A substituição das quantidades básicas, derivadas no Apêndice de /AVIIB63/, conduz a

$$w = [\xi p_{1.} E(T_r) + \xi b E(T_s)]/(1-b), \qquad (3.3.27)$$

e

$$w = \{\xi\lambda(1-b)E^2(t_r)(\gamma_r^2+1)/[1+\lambda E(t_r)] + b^2(\gamma_s^2+1)\}/2(1-b). \qquad (3.3.28)$$

A equação de Pollaczek-Khintchine pode ser obtida como um caso particular da eq. (3.3.28), se $\lambda = 0$, e o valor esperado do número de itens no sistema (incluindo fila+serviço) é expresso por

$$q = b + w . \qquad (3.3.29)$$

O autor recomenda a interessados o acompanhamento do trabalho completo de Avi-Itzhak e Naor /AVIIB63/, bem como de /AVIIB65/, onde são estabelecidas as relações entre atendimento de multifilas e interrupções de serviço em sistemas de filas simples.

Dentro da classe de modelos com variabilidade do número de unidades de serviço, o autor sugere ainda a investigação do trabalho de K. Singh /SINGK78/, onde uma unidade de serviço adicional é ativada, a um determinado custo, ao ser atingido o comprimento máximo de uma fila finita, e desativada quando o comprimento da fila é reduzida a um limite fixo tolerável.

Este item III.3.3 conclui os tópicos em teoria de filas, objeto da Seção III.3, e com isto, o Capítulo III desta dissertação.

CAPÍTULO IV

ANÁLISE DE MODELOS DE MULTIFILAS - DESENVOLVIMENTO E RESULTADOS

Na Seção II.3 foram identificados, segundo critérios definidos, alguns modelos para constituirem objetos de estudos detalhados neste trabalho, no sentido de um aprofundamento em aspectos de mecânica de funcionamento, definição/caracterização de entradas e tipos e formas de resultados disponíveis.

Este contato mais estreito com os processos de formulação dos modelos e de solução/derivação de resultados, contribui efetivamente para a assimilação das notações utilizadas (os autores referenciados neste Capítulo IV adotam notações bastante distintas), facilitando assim a utilização prática dos modelos/resultados.

Adicionalmente, são verificados e reforçados conceitos e métodos relativos a tópicos de probabilidade, processos estocásticos e teoria de filas simples, conforme aqueles apresentados no Capítulo III.

A associação da conceituação clara e precisa ao domínio prático de um modelo, ou conjunto de modelos, mostra-se fundamental na utilização correta e oportuna, bem como, no desenvolvimento de variantes ou extensões dos mesmos.

Com estes objetivos, são apresentados neste Capítulo IV resumos das análises e soluções dos modelos estudados por Leibowitz, Eisenberg, Sykes e Konheim e Meister, respectivamente, em /LEIBM61/, /EISEM72/, /SYKEJ70/ e /KONHA74/.

O princípio da independência de Leibowitz, já mencionado frequentemente no Capítulo II, é o tema da Seção IV.1 , uma vez que constitui um marco expressivo, além de fundamentação teórica de grande valor, no desenvolvimento de análises de modelos de multifilas, como pode ser verificado na revisão de literatura do Capítulo II.

Na Seção IV.2 são resumidas as análises de Eisenberg e Sykes, que se baseiam em análise estocástica, correspondendo à utilização de técnicas, métodos e premissas convencionalmente empregados no campo de teoria de filas.

A Seção IV.3 consiste no resumo da análise realizada por Konheim e Meister, conforme referência, que se caracteriza pelo emprego de técnicas de análise operacional.

Em /DENNP78/, Denning e Buzen discutem a abordagem operacional a modelos de redes de filas, em trabalho de carater introdutório/conceitual, orientado para avaliação de desempenho, onde podem ser encontradas as bases e princípios da análise operacional, bem como comparações/analogias com elementos de análise estocástica.

Sobre o modelo de Sykes será mapeado, no Capítulo V

seguinte, um problema real de análise de desempenho. O modelo de Eisenberg é praticamente equivalente a este, para o caso de 2 filas, apresentando, entretanto, grande potencial de utilização posterior (N filas simétricas), e consequente maior abrangência de desenvolvimento, que justificam sua inclusão neste Capítulo IV.

Adicionalmente à utilização direta do modelo de Konheim e ao interesse conceitual de sua análise, pretende-se que um estudo detalhado deste, atue como base na direção do modelo assimétrico de Swartz (/SWARG77/), cuja aplicação na solução de desempenho de sub-sistemas em configuração multiponto é considerada pelo autor desta tese, conforme extensão proposta no Capítulo VI.

SEÇÃO IV.1

O PRINCÍPIO DE INDEPENDÊNCIA DE LEIBOWITZ/LEIBM61/

Em /LEIBM61/, Leibowitz formula um princípio de independência aplicável a sistemas de multifilas, que resulta em um método aproximado para solução eficiente de modelos desta classe. Esta Seção IV.1 apresenta um resumo da formulação deste princípio, de importância conceitual e prática fundamentais no estudo de multifilas, na medida em que, conforme pode ser observado na revisão de literatura, um grande número de trabalhos relevantes no assunto baseia-se neste princípio de independência.

No tratamento de sistemas de multifilas, uma descri

ção completa do estado do sistema implica na especificação de probabilidades conjuntas, dependendo de um grande número de índices, cada índice correspondendo, em geral, a uma fila. A determinação destas probabilidades demanda geralmente a solução de uma grande quantidade de sistemas de equações. Entretanto, uma descrição completa faz-se raramente necessária, uma vez que o interesse concentra-se geralmente, por exemplo, na distribuição de probabilidades do comprimento de uma determinada fila ou do tempo de espera, ou ainda em alguma outra distribuição ou parâmetro que forneça uma caracterização mais simples do sistema.

É entretanto praticamente impossível escrever equações que envolvam apenas estas grandezas e que sejam satisfeitas de maneira exata. Em outras palavras, uma caracterização parcial exata do sistema, só pode ser obtida a partir de uma descrição completa e exata do sistema.

É considerado por Leibowitz um sistema simétrico com N filas infinitas, processos de chegada independentes, do tipo Poisson, com tempo médio entre chegadas $1/\lambda$ serviço limitado (L) cíclico $\{1,2,\ldots,N\}$ e tempo médio de serviço $1/\mu$. O tempo de transição ("walking time", em /LEIBM61/) é definido como o intervalo entre o término de serviço em uma fila e o início de serviço na próxima, e tem f.d.p. $w(t)$ e valor médio $\bar{w}$. A f.d.p. do tempo de serviço é $s(t)$, arbitrária.

A descrição de estado completa, no instante de chegada da unidade de serviço a qualquer fila, por exemplo, à fila 1, requer a especificação do vetor de números de itens nas di

versas filas, $(n_1,n_2,\ldots,n_N)$. Por esta razão, as probabilidades de estado $\{p_n\}$ não constituem uma distribuição estacionária de uma cadeia de Markov, como no caso de uma única fila. Deve ser considerada a distribuição de probabilidades conjuntas $\{p_{n_1,\ldots,n_N}\}$ do vetor de variáveis $(n_1,\ldots,n_N)$. Ainda que seja possível encontrar uma matriz de transição e um sistema de equações lineares para $p_{n_1,\ldots,n_N}$, o cálculo dos elementos desta matriz torna-se bastante complicado com N crescente. Assim, embora p_n possa ser obtido pela somatória de $p_{n_1,\ldots,n_N}$ sobre $n_2,\ldots,n_N$, tal método não se mostra viável se N é grande.

Seja o seguinte método heurístico para determinação do valor médio

$$\bar{n} = \Sigma\, np_n \ ,$$

e seja T o tempo médio de um ciclo completo no sistema, composto de N transições (com duração média total $N\bar{w}$) e da somatória dos períodos de serviço em cada uma das N filas. Para determinar a duração média destes períodos de serviço, suponha-se que:

(A) EM CADA FILA A UNIDADE DE SERVIÇO ENCONTRE $\bar{n}$ ITENS, o que resulta em

$$T = N(\bar{n}/\mu) + N\bar{w} \ .$$

Durante este tempo T, em média chegam $\bar{n}=T\lambda$ ítens à fila 1. Então

$$\bar{n}/\lambda = N(\bar{n}/\mu) + N\bar{w} \text{ , ou}$$

$$\bar{n} = \frac{N\lambda\bar{w}}{1-N\lambda/\mu} \text{ .} \qquad (4.1.1)$$

A condição de não-saturação do sistema ($\bar{n}$ finito) é expressa então por $N\lambda/\mu<1$.

Suponha-se adicionalmente que:

(B) EM CADA FILA A UNIDADE DE SERVIÇO ENCONTRE A MESMA DISTRIBUIÇÃO DE PROBABILIDADES $\{p_n\}$, INDEPENDENTEMENTE DAS DEMAIS FILAS.

Com base nesta premissa pode ser articulada a seguinte argumentação: se a unidade de serviço encontra a mesma distribuição de probabilidades de número de itens em cada fila em um ciclo do sistema, começando pela fila 1, então a mesma distribuição será encontrada quando esta retornar à fila 1. $\{p_n\}$ é denominada de "distribuição auto-consistente".

O desenvolvimento da função geratriz para $\{p_n\}$, segundo a premissa (B), fornece expressões para os momentos de 1a e 2a ordem do número de itens na fila, tais que, a expressão para $\bar{n}$ é idêntica à eq. (4.1.1) obtida acima. Através de uma expansão de p_n em série de potências, Leibowitz conjectura que as aproximações resultantes são corretas até termos da ordem de λ^3, se λ/μ é pequeno. Desta forma, é natural esperar-se boas aproximações para pequenos valores de λ, uma vez que, como a entrada de itens no sistema é pequena, o tempo de transição é o fa

tor dominante, e a dependência entre as filas é relativamente fraca. As mesmas expectativas mostram-se válidas quando o número de filas, N, é muito grande, devido a um cancelamento cruzado de efeitos entre diversas filas onde a premissa (B) é violada, de tal forma que esta premissa ainda prevalece, de alguma maneira, em termos médios.

Em realidade, Leibowitz mostra que seu tratamento fornece valores exatos para o caso N=1. Para o caso N=2, é demonstrado que a aproximação fornece valores corretos até a ordem de λ^3, para pequenos λ's. Assim, Leibowitz considera razoável esperar que o mesmo seja válido para qualquer N>2, uma vez que, mantida a condição de baixo tráfego no sistema, a tendência é de um enfraquecimento crescente da dependência entre as filas, com o aumento de N.

Recomenda-se um estudo completo dos artigos /LEIBM61/ /LEIBM62/ e /LEIBM68/, aos interessados no desenvolvimento integral das idéias e resultados que constituem o trabalho de Leibowitz na área de análise de sistemas de multifilas.

SEÇÃO IV.2

ANÁLISE ESTOCÁSTICA - EISENBERG E SYKES

IV.2.1 EISENBERG/EISEM72/

Em /EISEM72/ é analisado um modelo assimétrico com M filas M/G/*, serviço exaustivo ("come-right-in discipline", em

/EISEM72/), sequência de atendimento periódica arbitrária, tempos de transição ("changeover times", em /EISEM72/) finitos com distribuição arbitrária e movimento contínuo da unidade de serviço com o sistema vazio. Um estágio é definido como o período de tempo durante o qual a unidade de serviço atende continuamente a uma única fila, e o tempo entre visitas é o tempo decorrido entre a partida e o retorno subsequente da unidade de serviço a uma mesma fila. Devido ao movimento contínuo da unidade de serviço, o tempo total de transição em um ciclo deve ser não-nulo, o que não permite comparações dos resultados de /EISEM72/ com outros trabalhos mencionados na revisão de literatura, que consideram tempos de transição nulos, fazendo-se o mesmo tender a zero e tomando resultados limites. A análise de Eisenberg fornece resultados exatos e apresenta expressões fechadas para um sistema de duas filas.

A taxa média de chegadas à fila m é λ_m, e o tempo de serviço nesta fila possui f.d.a. $F_{s_m}(t)$ e valor médio $1/\mu_m$, m=1,2,...,M. A transformada de Laplace-Stieltjes (TLS) desta distribuição é denotada por $S_m(s)$, onde:

$$S_m(s) \triangleq \int_o^\infty e^{-st}\, dF_{s_m}(t) \quad , \quad (m=1,2,\ldots,M). \qquad (4.2.1)$$

A sequência de serviço é definida por um conjunto ordenado de I inteiros $(m_1,m_2,\ldots,m_I)$, onde m_i indica a fila atendida durante o estágio i. Entre estágios sucessivos de um ciclo ocorre sempre uma transição entre filas, onde $F_{c_i}(t)$ denota a f.d.a. do tempo de transição entre estágios (i-1) e i , com

TLS $C^i(s)$ e valor médio c^i, onde:

$$C^i(s) \triangleq \int_0^\infty e^{-st}\, dF_{c^i}(t) \quad , \quad (i=1,2,\ldots,I) . \qquad (4.2.2)$$

São definidas cadeias de Markov imersas nos instantes de início e término de serviço e de início e término de estágio, com as respectivas probabilidades de estado em equilíbrio:

$w^i_{\boldsymbol{n}} = w^i_{n_1,\ldots,n_M}$ = probabilidade de estado do processo imerso nos instantes de início de serviço

$\Pi^i_{\boldsymbol{n}} = \Pi^i_{n_1,\ldots,n_M}$ = idem instantes de término de serviço

$\alpha^i_{\boldsymbol{n}} = \alpha^i_{n_1,\ldots,n_M}$ = idem instantes de início de estágio

$\beta^i_{\boldsymbol{n}} = \beta^i_{n_1,\ldots,n_M}$ = idem instantes de término de estágio.

As funções geratrizes associadas são da forma

$$w^i(z) = w^i(z_1,\ldots,z_M) = \Sigma_{n_1=0}^{n_1=\infty} \ldots \Sigma_{n_M=0}^{n_M=\infty} w^i_{n_1,\ldots,n_M} z_1^{n_1} \ldots z_M^{n_M},$$

definidas similarmente para $\Pi^i(z)$, $\alpha^i(z)$ e $\beta^i(z)$. Observar que $\beta^i(z)$ é independente de z_{m_i}, uma vez que, ao final de um estágio i, a fila m_i deve estar vazia.

No intervalo de tempo (0,t) define-se:

$w^i(t;\boldsymbol{n}) = w^i(t;n_1,\ldots,n_M)$ = número de inícios de serviço no estágio i com estado $(n_1,\ldots,n_M)$ que ocorrem em (0,t).

$w(t)$ = número total de inícios de serviço em $(0,t)$.

De maneira análoga são definidas $\Pi^i(t;n)$ e $\Pi(t)$, $\alpha^i(t;n)$ e $\alpha^i(t)$ e $\beta^i(t;n)$ e $\beta^i(t)$. Se $\lambda_m<\mu_m$, $m=1,2,\ldots,M$ (condição para não-saturação), tem-se a seguinte relação:

$$\alpha^i(t;n) + \Pi^i(t;n) = w^i(t;n) + \beta^i(t;n), \qquad (4.2.3)$$

que expressa o fato de que para cada início de estágio ou término de serviço, ocorre um início de serviço ou término de estágio. Estabelecendo-se algumas relações limites entre estas funções, e utilizando-se propriedades/teoremas de ergodicidade, obtém-se:

$$\gamma\alpha^i(z) + \Pi^i(z) = w^i(z) + \gamma\beta^i(z), \qquad (4.2.4)$$

onde $\gamma = \gamma^i_{\forall i} \triangleq \lim_{t\to\infty} [\beta^i(t)/\Pi(t)]$, e o limite existe com probabilidade 1.

Manipulações adicionais conduzem a uma relação entre as probabilidades de estado nos instantes de término de serviço e as probabilidades de estado nos instantes de término de estágio:

$$\Pi^i(z) = \{\gamma\tilde{S}_{m_i}(z)/[z_{m_i}-\tilde{S}_{m_i}(z)]\}[\beta^{i-1}(z)\tilde{C}^i(z)-\beta^i(z)]. \qquad (4.2.5)$$

Nesta expressão $\tilde{S}_m(z) = S_m(\lambda_1-\lambda_1 z_1+\ldots+\lambda_M-\lambda_M z_M)$ e

$\tilde{C}^i(z) = C^i(\lambda_i-\lambda_1 z_1+\ldots+\lambda_M-\lambda_M z_M)$ representam, respectivamente, as funções geratrizes das distribuições do número

de itens de cada tipo que chegam durante o serviço de um item da fila m, e do número de itens que chegam durante uma transição de estágio (i-1) para i.

O modelo é resolvido para $\beta^i(z)$ e normalizado, pela determinação de γ, resultando em

$$\gamma = 1/\lambda C , \tag{4.2.6}$$

$$\text{onde} \quad C = (c^1+c^2+\ldots+c^I)/(1-\rho_1-\rho_2-\ldots-\rho_M) \tag{4.2.7}$$

representa o tempo médio de ciclo, $\sum_{i=1}^{i=I} c^i$ corresponde ao tempo total de transição em um ciclo e $\rho_i = \lambda_i/\mu_i$ representa o tráfego em cada fila.

A condição de não-saturação, obtida por Eisenberg de maneira rigorosa, consiste em

$$1-\rho_1-\rho_2-\ldots-\rho_M > 0 \tag{4.2.8}$$

e mostra-se independente dos tempos de transição. A explicação é que, em níveis de carga perto da saturação, a proporção de tempo gasta pela unidade de serviço em transições torna-se desprezivelmente pequena.

Através da variável aleatória auxiliar tempo entre visitas, definida no início deste Item IV.2.1, e dos resultados anteriormente obtidos, a partir da eq. (4.2.5) obtém-se a relação entre o tempo médio de espera e os dois primeiros momentos

do tempo entre visitas:

$$w^i = [E(V^{i\,2})/2v^i] + [\lambda_{m_i} E(S^2_{m_i})/2(1-\rho_{m_i})]. \qquad (4.2.9)$$

Nesta expressão, w^i é o tempo médio de espera na fila atendida no estágio i, m_i, $E(V^{i\,2})$ é o segundo momento do tempo entre visitas a esta fila e $E(S^2_{m_i})$ é o segundo momento do tempo de serviço na fila m_i. Este é o resultado final, genérico para M filas, fornecido por Eisenberg neste trabalho. Entretanto a avaliação dos dois primeiros momentos do tempo entre visitas é bastante trabalhosa, implicando em diferenciações de 1ª e 2ª ordens de uma expressão funcional para $\beta^i(z)$, que envolvem operadores especiais de composição de funções ("ℓ-fold nesting", conforme /EISEM72/, eq. 26), e a solução de dois sistemas de equações simultâneas, um com (M-1) e outro com M(M-1)/2 equações.

O trabalho de Eisenberg é então concluído com um caso particular de duas filas e serviço cíclico ordinário. São fornecidos resultados para o tempo médio de espera na fila 1, em função dos dois primeiros momentos de tempo entre visitas à fila 1. Resultados análogos podem ser obtidos para a fila 2, por substituição de índices. Assim, M=2, I=2, m_1=1, m_2=2 e

$$w^1 = \{E[V_1^2]/2v^1\} + \{\lambda_1 E[S_1^2]/2(1-\rho_1)\}\ , \qquad (4.2.10)$$

onde

$$v^1 = (1-\rho_1)(c^1+c^2)/(1-\rho_1-\rho_2) \qquad (4.2.11)$$

$$\text{e} \quad E(V_1^2) = \frac{2\rho_1\rho_2[\rho_2 c^1 + (1-\rho_1)c^2]^2 + (c^1+c^2)[\rho_2^2\lambda_1 E(S_1^2) + (1-\rho_1)^2\lambda_2 E(S_2^2)]}{(1-\rho_1-\rho_2)^2(1-\rho_1-\rho_2+2\rho_1\rho_2)}$$

$$+ \frac{2c^1c^2(1-\rho_1)\rho_2 + \rho_2^2 E(C_1^2) + (1-\rho_1)^2 E(C_2^2)}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2+2\rho_1\rho_2)}$$

$$+ \frac{2c^1[\rho_2 c^1 + (1-\rho_1)c^2]}{1-\rho_1-\rho_2} + E(C_1^2) \cdot \qquad (4.2.12)$$

Ao final desta análise, Eisenberg descreve brevemente as alterações necessárias para o caso do serviço limitado, denominado pelo mesmo de "please-wait discipline".

Básicamente, é necessária uma expansão do vetor de estado, de forma a manter, para aquela fila que está sendo atendida, contagens separadas, de quantos itens serão despachados nesta visita corrente e de quantos deverão esperar pela próxima visita. Com o novo estado expandido, são definidas funções geratrizes análogas às do caso exaustivo, e indicado o desenvolvimento, que é, entretanto, bem mais complexo, e não resolvido por Eisenberg, neste trabalho.

IV.2.2 SYKES/SYKEJ70/

Em /SYKEJ70/, Sykes apresenta uma análise simplificada de um sistema de duas filas com prioridades alternantes e tempos de transição ("switching times", em /SYKEJ70/) finitos. O modelo analisado considera duas filas M/G/* que compartilham uma

única unidade de serviço, com serviço exaustivo e movimento contínuo da unidade de serviço com o sistema vazio (denominado, pelo autor, "keep-switching"). Uma situação equivalente consiste em uma única fila com duas classes de clientes, e a disciplina de prioridades alternantes caracteriza-se pela quantidade mínima de transições realizadas entre as duas filas.

Motivação para esta análise constituiu a obtenção de um modelo analítico para interligação"half-duplex"de dois computadores, que resultasse em estimativas de duas grandezas relevantes no desempenho de sistemas de comunicação de dados: o atraso médio de mensagens, aguardando transmissão e a quantidade média de memória utilizada para enfileiramento de mensagens em cada estação. O artigo /SYKEJ69a/ apresenta alguns resultados resumidos e gráficos, relativos a esta aplicação do modelo.

A descrição do modelo considera processos de chegada independentes, entre si e do estado da unidade de serviço,do tipo Poisson, com taxas de chegadas médias λ_i (i=1,2). Nenhuma restrição é colocada quanto à ordem de serviço em uma fila. A notação $s_i(t)$, (i=1,2) representa a f.d.p. arbitrária do tempo de serviço de itens da fila (ou classe) i, cujo valor médio é denotado por $\bar{t}_i$. Os tempos de serviço são ainda considerados variáveis aleatórias não-negativas, independentes entre si e dos intervalos entre chegadas. A f.d.p. do tempo de transição, ou comutação, da fila i para a fila j é denotada por $h_{ij}(t)$, (ij=12,21), e possui valor médio $\bar{r}_{ij}$.

A Figura IV.1 ilustra um ciclo típico de operação do

sistema, onde chegadas acumulam na fila 2 durante um período de acumulação A_2', que corresponde à soma de r_{21}', T_1' e r_{12}. Ao final deste período A_2' os itens acumulados na fila 2 são despachados, durante o tempo T_2, que representa o tempo necessário para esvaziar a fila 2, e é seguido de um tempo de transição r_{21}. Inicia-se então o tempo T_1, que representa o tempo necessário para esvaziar a fila 1, o qual é seguido de um tempo de transição r_{12}, reiniciando-se então o ciclo, com o início de novo T_2. Notar que, durante um ciclo, ambos T_1 e/ou T_2 podem ter duração nula, uma vez que as filas 1 e 2 podem ser encontradas vazias.

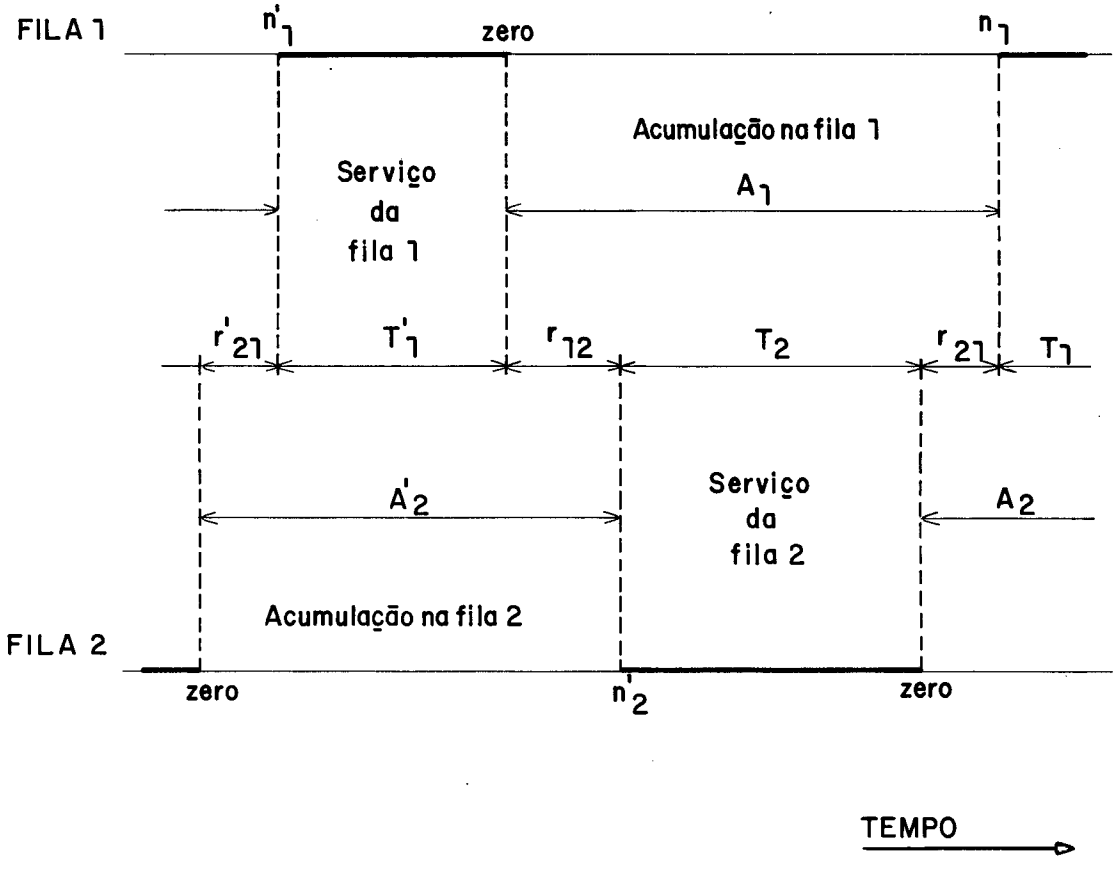


Figura IV.1 - Ciclo típico de operação do modelo de prioridades alternantes

Sejam $\rho_1=\lambda_1\bar{t}_1$ e $\rho_2=\lambda_2\bar{t}_2$ as intensidades de tráfego das filas 1 e 2. A fração total do tempo, na qual a unidade de serviço é utilizada é $\rho=\rho_1+\rho_2$, onde $\rho<1$ para existência de um estado de equilíbrio do sistema. Logo, $(1-\rho)$ é a fração do tempo na qual a unidade de serviço não realiza trabalho, fração esta consumida pelos tempos r_{12} e r_{21}. Assim, a duração média de um ciclo pode ser dada por

$$\bar{C} = (\bar{r}_{12} + \bar{r}_{21})/(1-\rho) \qquad (4.2.13)$$

ou, alternativamente, por

$$\bar{C} = \bar{T}_1' + \bar{r}_{12} + \bar{T}_2 + \bar{r}_{21} \qquad (4.2.14)$$

que resulta por simples inspeção visual da Figura IV.1.

Uma vez que os processamentos das filas 1 e 2 são equivalentes, são derivadas expressões para a fila 1, que podem ser aplicadas para a fila 2 por substituição de índices.

Neste modelo, todos os itens que chegam à fila 1 (ou 2) sofrem atrasos, dependentemente da ocorrência da chegada no período de acumulação ou de serviço desta fila, bem como em função de n_1, o número de itens acumulados no período de acumulação anterior, A_1. Ciclos de operação iniciados com n_1 itens na fila 1, $n_1=0,1,2,\ldots$, são denotados por $C(n_1)$.

O atraso médio $\bar{D}_1$ de itens na fila 1, até o início dos respectivos serviços pode ser determinado a partir de

$$\bar{D}_1 = \bar{W}_1/\bar{N}_1 = E[\bar{W}_1(n_1)]/E[\bar{N}_1(n_1)]$$

$$= [\Sigma_{n_1=0}^{n_1=\infty} P(n_1)\bar{W}_1(n_1)]/[\Sigma_{n_1=0}^{n_1=\infty} P(n_1)\bar{N}_1(n_1)], \quad (4.2.15)$$

onde: $\bar{W}_1$ = valor médio da somatória dos atrasos dos itens da fila 1, em um ciclo de operação

$\bar{N}_1$ = número médio de itens da fila 1 atendidos durante um ciclo

$\bar{W}_1(n_1)$ = valor médio no tempo, de $W_1(n_1)$, a soma dos atrasos de todos os itens atendidos durante um ciclo $C(n_1)$

$\bar{N}_1(n_1)$ = número médio de itens da fila 1 atendidos durante um ciclo $C(n_1)$

$P(n_1)$ = probabilidade de um ciclo $C(n_1)$.

Na determinação das quantidades que compõem a expressão (4.2.15), Sykes utiliza o conceito de <u>períodos de ocupação</u> (ver III.3.2) e respectivos resultados, sob a forma dos dois primeiros momentos de cada um dos períodos de ocupação -n_1, gerados pelo início de um ciclo com n_1 itens na fila 1, denotados, respectivamente, por $\bar{B}_1$ e $B_2^{(2)}$, onde

$$\bar{B}_1 = \bar{t}_1/(1-\rho_1) \quad (4.2.16)$$

$$B_1^{(2)} = t_1^{(2)}/(1-\rho_1)^3, \quad (4.2.17)$$

chegando a

$$\bar{D}_1 = \rho_1 E(A_1^2)/2\bar{n}_1\bar{B}_1 + [n_1^{(2)}/\bar{n}_1 - 1]\bar{t}_1/2 + \lambda_1 t_1^{(2)}/2(1-\rho_1), \quad (4.2.18)$$

onde

$$E(A_1^2) = E[(r_{12} + T_2 + r_{21})^2] = r_{12}^{(2)} + r_{21}^{(2)} + 2\bar{r}_{12}\bar{r}_{21}$$

$$+ \bar{t}_2^2[n_2^{(2)} - \bar{n}_2]/(1-\rho_2)^2 + \bar{n}_2 t_2^{(2)}/(1-\rho_2)^3$$

$$+ 2\rho_2[(\bar{r}_{12}+\bar{r}_{21})^2/(1-\rho) + (r_{12}^{(2)} - \bar{r}_{12}^2)/(1-\rho_2)] \quad (4.2.18)$$

e expressões para $\bar{n}_1$, $\bar{n}_2$, $n_1^{(2)}$ e $n_2^{(2)}$ devem ser ainda obtidas (Sykes mostra a derivação da eq. (4.2.18) no Apêndice de /SYKEJ70/).

As expressões correspondentes à fila 1 são avaliadas por meio de uma função geratriz $G_{12}(x,y)$ para a probabilidade conjunta $P(n_1,m_2)$ dos seguintes eventos: em condições de equilíbrio, no instante imediato do início do período T_1, n_1 itens, estão acumulados na fila 1, e m_2 itens na fila 2.

Simbolicamente,

$$G_{12}(x,y) = \sum_{n_1=0}^{n_1=\infty} \sum_{m_2=0}^{m_2=\infty} P(n_1,m_2) x^{n_1} y^{m_2} = E[x^{n_1} y^{m_2}], \text{ onde}$$

n_1 e m_2 são variáveis aleatórias descritas por distribuições Poisson independentes, com parâmetros λ_1 e λ_2, respectivamente.

O desenvolvimento desta função geratriz $G_{12}(x,y)$ conduz à obtenção de expressões fechadas para n_1 e $n_1^{(2)}$ (bem como n_2 e $n_2^{(2)}$), em função de variáveis conhecidas, resultando na seguinte expressão para o atraso médio

$$\bar{D}_1 = \frac{\lambda_2 t_2^{(2)}(1-\rho_1)^2 + \lambda_1 t_1^{(2)} \rho_2^2}{2(1-\rho_1)(1-\rho)(1-\rho+2\rho_1\rho_2)} + \frac{\lambda_2 t_1^{(2)}}{2(1-\rho_1)}$$

$$+ \frac{\bar{r}_{21}^2}{2(1-\rho_1)(\bar{r}_{12}+\bar{r}_{21})(1-\rho)} [\xi + \rho_2^2 \eta]$$

$$+ \frac{\bar{r}_{21}^2 \theta}{2(1-\rho_1)(\bar{r}_{12}+\bar{r}_{21})} + \frac{(1-\rho_1)\ \bar{r}_{12}\bar{r}_{21}}{(1-\rho)(\bar{r}_{12}+\bar{r}_{21})} \quad , \quad (4.2.19)$$

onde $\xi = [(1-\rho_1)^2 - \rho_2^2 + (1-\rho)^2\ C^2(r_{21})]$,

$\eta = [(1-\rho)C^2(r_{21}) + 1-\rho +2\rho_1\rho_2]/[1-\rho + 2\rho_1\rho_2]$ e

$\theta = [(1-\rho)C^2(r_{12}) + 1-\rho +2\rho_1\rho_2]/[1-\rho + 2\rho_1\rho_2]$.

Nestas expressões, $C^2(r_{ij}) = r_{ij}^{(2)}/\bar{r}_{ij}^2 - 1$, (ij=12,21), representam os coeficientes de variação quadrados das variáveis aleatórias que descrevem os tempos de transição entre as filas. Se r_{12} e r_{21} são constantes, implica em $C^2(r_{12}) \equiv C^2(r_{21}) \equiv 0$, e a expressão (4.2.19) se reduz a

$$\bar{D}_1 = \frac{\lambda_2 t_2^{(2)}(1-\rho_1)^2 + \lambda_1 t_1^{(2)} \rho_2^2}{2(1-\rho_1)(1-\rho)(1-\rho+2\rho_1\rho_2)} + \frac{\lambda_1 t_1^{(2)}}{2(1-\rho_1)} + \frac{(1-\rho_1)[\bar{r}_{12}+\bar{r}_{21}]}{2(1-\rho)} .$$

(4.2.20)

O comprimento médio da fila 1, $\bar{L}_1$, incluindo o item que está em serviço, pode ser obtido através da aplicação da fórmula de Little:

$$\bar{L}_1 = \lambda_1 [\bar{D}_1 + \bar{t}_1] = \lambda_1 \bar{D}_1 + \rho_1 . \quad (4.2.21)$$

Chama-se a atenção para o fato de que Sykes não men

ciona explícitamente em /SYKEJ70/ a utilização do princípio da independência de Leibowitz, apresentado na Seção IV.1. Sua lista de referências nem mesmo o inclui. Entretanto, em /SYKEJ69a/, Sykes faz menção explícita à sua utilização deste princípio de independência, citando Leibowitz, no contexto da avaliação das probabilidades conjuntas necessárias ao desenvolvimento da função geratriz $G_{12}(x,y)$ (ver /SYKEJ69a/, pag. 237).

Observar, entretanto, que o trabalho /SYKEJ70/ é auto-contido, apresentando o desenvolvimento completo da análise e derivações de Sykes, sendo este a base para este Item IV.2.2, conforme explicado no início deste Capítulo IV. A referência a /SYKEJ69a/ tem, neste contexto, objetivo exclusivo de chamar a atenção e esclarecer um fato relevante e, em certo aspecto, curioso, segundo opinião do autor desta tese.

Resultados deste modelo serão objeto de aplicação na análise de desempenho de sub-sistemas de comunicação de dados, no Capítulo V seguinte.

SEÇÃO IV.3

ANÁLISE OPERACIONAL - KONHEIM E MEISTER/KONHA74/

Em /KONHA74/, Konheim e Meister apresentam uma análise exata e bastante rigorosa de sistemas de multifilas simétricos com N filas, processos de chegada independentes e idênticamente distribuídos, processos de serviço determinísticos, serviço exaustivo, atendimento cíclico estrito e tempos de transição fi

nitos, arbitrariamente distribuídos (Konheim denomina os tempos de transição de "reply interval").

A linha de análise adotada é a análise operacional, e o modelo físico em que se baseiam os autores corresponde a um sistema de comunicações em que terminais dotados de "buffers" infinitos transmitem dados para uma estação central, por exemplo, uma CPU, compartilhando um canal de comunicações multiponto, sob disciplina de "polling". Do ponto de vista de topologia, a análise pode ser igualmente aplicada a configurações estrela ou anel. O objetivo é fornecer relações entre o tráfego oferecido ao sistema e o desempenho, avaliado em termos de comprimentos de filas e atrasos de espera nas filas.

Seja um canal de comunicações multiplexado no tempo. O período $0 \leq t < \infty$ é dividido em intervalos contíguos, ou janelas ("slots"), $s_j : (j-1)\Delta \leq t < j\Delta$, $(1 \leq j < \infty)$, cada uma das quais acomoda uma única unidade de dados (por exemplo, "bytes", caracteres , grupos de caracteres, etc.). A operação de serviço é sempre iniciada em tempos de forma $j\Delta$, $(0 \leq j < \infty)$, de maneira que a entrada de dados no sistema pode ser especificada completamente pelo número de itens que chegam em um intervalo $(j-1)\Delta \leq t < j\Delta$.

Nesta análise, pressupõe-se que as chegadas de dados aos vários terminais são descritas por processos estocásticos

$$X^{(i)} = \{X_j^{(i)} : 1 \leq j < \infty\} \quad , \quad (1 \leq i \leq N) \qquad (4.3.1)$$

que satisfazem as seguintes condições: a) os processos $\{X^{(i)}: 1 \leq i \leq N\}$ são independentes e b) para cada i, $1 \leq i \leq N$, as variáveis aleatórias $\{X_j^{(i)}: 1 \leq j < \infty\}$ assumem valores inteiros não-negativos, independentes e idênticamente distribuídos, segundo distribuições conhecidas

$$p^{(i)}(k) = Pr\{X_j^{(i)} = k\} \quad , \quad (0 \leq k < \infty) \quad , \qquad (4.3.2)$$

onde $X_j^{(i)}$ é o número de unidades de dados que chegam ao i-ésimo terminal durante a j-ésima janela, $(j-1)\Delta \leq t < j\Delta$.

Os terminais recebem "polling" sequencialmente, $T^{(1)}, T^{(2)}, \ldots, T^{(N)}$, e adquirem direito para uso do canal alternadamente, removendo dados de seus "buffers" à taxa de uma unidade de dados por janela. A chegada de dados continua durante o processo de transmissão, de forma que o "buffer" é simultâneamente esvaziado e ocupado com novos dados, até que seu conteúdo atinja zero, quando então o terminal perde controle do canal. A partir deste instante, o canal torna-se não disponível para todos os terminais, por um intervalo de tempo aleatório, que é utilizado para "overheads" de sistema relativos ao controle de acesso ("polling","addressing", confirmações, etc.), e/ou para transmissão de dados da estação central para os terminais ("reply interval"). Uma varredura completa de todos os N terminais é denominada ciclo do sistema.

O estado do sistema é descrito pelas variáveis: $W_j^{(i)}$, número de unidades de dados no "buffer" do i-ésimo terminal $T^{(i)}$ no tempo $j\Delta - 0$ $(0 \leq j < \infty)$ e U_j, estado do canal na (j+1)-ésima jane

la, onde U_j=i se o canal está disponível para $T^{(i)}$ e U_j=0 em caso contrário.

O sistema evolui segundo a seguinte equação de estado

$$W_j^{(i)} = (W_{j-1}^{(i)} - \chi(U_{j-1}=i))^+ + X_j^{(i)} \quad , \quad (1\leq j<\infty \ , 1\leq i\leq N) \ ; \qquad (4.3.3)$$

onde $a^+=\max(a,0)$ e χ_A é a função característica do evento A:1 se A ocorre, e 0 em caso contrário. O processo aleatório

$$\{W_j = (W_j^{(1)}, W_j^{(2)}, \ldots, W_j^{(N)}) : 0\leq j<\infty\} \qquad (4.3.4)$$

não é um processo Markoviano, mas contém imerso um sub-processo Markoviano, denominado pelos autores de "natural", onde naturais são considerados os tempos do processo acima nos quais o canal se torna disponível para cada um dos terminais (instantes de início de serviço de uma fila). Estes instantes são denotados por $\{\tau_{i,j} : 1\leq i<\infty, \ 1\leq j\leq N\}$, $0=\tau_{1,1}<\tau_{1,2}<\ldots<\tau_{1,N}<\tau_{2,1}<\ldots$ $<\tau_{i,j-1}<\tau_{i,j}<\tau_{i,j+1}<\ldots$. O instante aleatório em que o canal se torna disponível para j-ésimo terminal, no i-ésimo ciclo é denotado por $\tau_{i,j}$. O sub-processo imerso mencionado acima, consistindo de observações do sistema nos instantes $\{\tau_{i,j}\}$, $W = \{W_{\tau_{i,j}} : 1\leq i<\infty, \ 1\leq j\leq N\}$ é Markoviano, com as probabilidades de transição

$$p_j(x/y) = \Pr\{W_{\tau_{i,j+1}} = x \mid W_{\tau_{i,j}} = y\} \ , \ (1\leq j\leq N) \qquad (4.3.5)$$

estacionárias e independentes de i. Como os processos de chega

da $\{\boldsymbol{X}^{(i)}: 1 \leq j \leq N\}$ são regidos por uma lei comum, então as probabilidades $\{p_j: 1 \leq j \leq N\}$ somente diferem por uma permutação cíclica. No desenvolvimento do trabalho, os autores demonstram que $\boldsymbol{W}$ é recorrente positivo se são satisfeitas as seguintes condições : a) os tempos de transição tem duração esperada finita e b) a taxa total de chegadas de unidades de dados ao sistema é menor que 1 (taxa de serviço efetiva). Nestas condições $\boldsymbol{W}$ possui uma distribuição invariante, que, determinada, fornece uma descrição adequada do sistema, na medida em que possibilita a obtenção de distribuições de comprimentos de filas e de atrasos de espera nas filas. A premissa de processos de chegadas idênticamente distribuídos permanece, e possibilita substancial simplificação da análise, através de certas simetrias resultantes.

O tipo de processo estocástico utilizado para modelar as chegadas ao sistema é

$\boldsymbol{X} = \{X_j: 1 \leq j < \infty\}$, que consiste em variáveis aleatórias inteiras não-negativas i.i.d., com função geratriz

$$P(z) = E\{z^{X_j}\} = \Sigma_{k=0}^{\infty} p(k) z^k, \text{ onde } p(k) = Pr\{X_j = k\}, \ (0 \leq k \leq \infty).$$

X_j corresponde então ao número de unidades de dados que chegam a um terminal durante a j-ésima janela.

Considerando um "buffer" que em t=0 contém W_o unidades de dados, processo de chegada de dados do tipo $\boldsymbol{X}$ descrito , e remoção de dados à taxa de uma unidade por janela, tem-se que o número total de itens no "buffer", no início da (j+1)-ésima ja

nela, W_j, satisfaz à equação

$$W_j = (W_{j-1} - 1)^+ + X_j \quad , \quad (1 \leq j < \infty). \qquad (4.3.6)$$

O propósito desta equação é estabelecer uma analogia entre o problema da "ruína do jogador" ("gamblers ruin") , desenvolvido na Seção 3 de /KONHA74/ e o comportamento do "buffer" de um terminal no problema em análise. A ruína corresponde ao esvaziamento do "buffer", ou seja, o evento $\{W_j>0,\ 1\leq j<\infty\}$ é um evento nulo, ou ainda, $Pr\{W_j>0,\ 1\leq j<\infty\} = 0$, e o tempo para esta ruína, denotado por $T=\min\{j:W_j=0\}$, é finito com probabilidade 1, e corresponde ao tempo de retenção do canal por uma determinada fila. T possui f.g.p. $E\{w^T\}=H(\theta(w))$, onde $H(z)=E\{z^{W_o}\}$ e $\theta(w)$ é uma função que satisfaz $\theta(w)-wP(\theta(w))=0$, $(|w|<1,\ |\theta(w)|<1).P(z)$ é a f.g.p. do processo de chegadas a um terminal.

A duração dos tempos de transição é determinada por um processo aleatório auxiliar $\mathbf{R} = \{R_{i,j}: 1\leq j\leq N\ ,\ 1\leq i<\infty\}$, independente do processo de chegada $\{\mathbf{X}^{(i)}: 1\leq i\leq N\}$. As variáveis aleatórias $\{R_{i,j}\}$ assumem valores inteiros positivos, independentes e idênticamente distribuídos, com função geratriz comum $R(z) = E\{z^{R_{i,j}}\}$.

O estado do sistema no instante $\tau_{i,j}$ é descrito pelo vetor de estado $W_{\tau_{i,j}} = (W^{(1)}_{\tau_{i,j}}, W^{(2)}_{\tau_{i,j}}, \ldots, W^{(N)}_{\tau_{i,j}})$, com função geratriz

$$F_{i,j}(z_1,z_2,\ldots,z_N) = E\{z_1^{W_{\tau_{i,j}}}, \ldots, z_N^{W^{(N)}_{\tau_{i,j}}}\} ,$$

que é analítica no polidisco $\{(z_1,z_2,\ldots,z_N):|z_1|,|z_2|,\ldots,|z_N|<1\}$.

A <u>evolução do processo</u> $\{W_k:0\leq k<\infty\}$ é dada por

$$W_k^{(j)} = \begin{cases} W_{\tau_{i,j}}^{(j)} + \sum_{v=1+\tau_{i,j}}^{k} X_v^{(j)} - (k-\tau_{i,j}) & \text{se } \tau_{i,j}\leq k<\bar{\tau}_{i,j}\ , \\ 0 & \text{se } k=\bar{\tau}_{i,j}\ , \\ \sum_{v=1+\bar{\tau}_{i,j}}^{k} X_v^{(j)} & \text{se } \bar{\tau}_{i,j}<k<\tau_{i+1,j}. \end{cases} \tag{4.3.7}$$

Durante o primeiro intervalo $(\tau_{i,j}\leq k<\bar{\tau}_{i,j})$ dados chegam ao terminal e são removidos, à taxa de uma unidade por janela. No instante $\bar{\tau}_{i,j}$ o "buffer" do terminal $T^{(j)}$ está vazio, e a partir do tempo $\bar{\tau}_{i,j}+0$, dados acumulam no "buffer", até a próxima alocação do canal ao terminal $T^{(j)}$.

Baseados nesta formulação matemática inicial do modelo, os autores desenvolvem na Seção 4 de /KONHA74/, os resultados centrais da análise, que consistem em uma caracterização unívoca da <u>distribuição limite invariante do conteúdo do "buffer"</u> no terminal $T^{(1)}$, no <u>início de um ciclo</u> do sistema, $\bar{F}^*(z) = F^*(z,1,1,\ldots,1)$, cujos <u>valor esperado</u> e <u>variância</u> são apresentados a seguir:

$$E(\bar{F}^*) = Nr\mu(1-\mu)/(1-N\mu) \tag{4.3.8}$$

e

$$Var(\bar{F}^*)=\delta^2\mu^2N(1-\mu)/(1-N\mu)+[Nr\sigma^2/(1-N\mu)^2].[1-(N+1)\mu+(2N-1)\mu^2] \ , \tag{4.3.9}$$

onde

$$\mu = E\{X_j^{(i)}\} \ , \ \sigma^2 = Var\{X_j^{(i)}\}, \ r=E\{R_{i,j}\} \text{ e } \delta^2=Var\{R_{i,j}\}.$$

No desenvolvimento destes resultados podem ser observados, entre outros, os seguintes aspectos:

a) A simetria implicada pela hipótese de processos de chegada $\{\boldsymbol{X}^{(i)}:1\leq j\leq N\}$ idênticamente distribuídos, ou seja, a indistinguibilidade estatística dos terminais, que permite que não seja mantido controle do índice do terminal que inicia o ciclo, e a simplificação resultante na análise.

b) As condições de existência e unicidade da distribuição invariante: $N\mu<1$ e $r<\infty$, equivalentes às condições de equilíbrio ou ergodicidade na análise estocástica de sistemas similares.

c) O fato da distribuição limite do conteúdo do "buffer" no início de um ciclo ser independente da carga inicial do mesmo.

Na Seção 5 de /KONHA74/, é realizada a conexão entre o processo imerso $\{W_{\tau_{i,j}}:1\leq i<\infty, \ 1\leq j\leq N\}$ até então estudado e o processo original $\{W_k:0\leq k<\infty\}$, uma vez que se busca a determinação de comprimentos de filas e atrasos para todos os valores de k, ou seja, ao longo de um determinado período de observação do sistema. O processo utilizado consiste em se diluir em média

os observações $f(W_j)$ do estado no tempo $j\Delta$ (f é uma função qualquer mensurável), ao longo da evolução do sistema, segundo a relação

$$\tau_{m,1}^{-1} \sum_{j=0}^{\tau_{m,1}-1} f(W_j) \qquad (4.3.10)$$

que expressa uma média sobre os primeiros (m-1) ciclos. O valor limite desta expressão, quando $m\to\infty$ (se existe) fornece a diluição média temporal desejada.

Obtém-se então o comprimento médio do ciclo, e as distribuições estacionárias do comprimento de fila e do atraso na fila, válidas para qualquer fila no sistema. São obtidas funções geratrizes das distribuições mencionadas, a partir das quais são derivadas, na Seção 6 de /KONHA74/, expressões para os valores médios e as variâncias correspondentes. Os resultados finais (utilizáveis) são apresentados a seguir.

O <u>comprimento médio estacionário do ciclo, $\bar{C}$</u> é dado por

$$\bar{C} = \lim_{m\to\infty} (\tau_{m,1}/(m-1)) = N\mu/(1-N\mu) \qquad (4.3.11)$$

satisfeitas as condições $r<\infty$ e $N\mu < 1$ (nestas condições, o limite $\lim_{m\to\infty} (\tau_{m+1,1}/m)$ existe e é constante com probabilidade 1) e a <u>variância estacionária do comprimento do ciclo, Var{C}</u> é expressa por

$$Var\{C\} = \delta^2 N/(1-\mu)(1-N\mu) + \sigma^2 r N^2/(1-\mu)(1-N\mu)^2 . \qquad (4.3.12)$$

Se $r<\infty$ e $N\mu < 1$, o comprimento médio estacionário da fila pode ser calculado por

$$E\{W^{(1)*}\} = \frac{1}{2}\delta^2\mu/r + \frac{1}{2}\sigma^2/(1-N\mu) + \frac{1}{2}Nr\mu(1-\mu)/(1-N\mu) \quad (4.3.13)$$

e o atraso médio estacionário por

$$E\{D^{(1)*}\} = \delta^2/2r + \frac{1}{2}N\sigma^2/(1-N\mu) + \frac{1}{2}(1-\mu) + Nr(1-\mu)/2(1-N\mu). \quad (4.3.14)$$

Para as expressões (4.3.11) a (4.3.14) valem $r = E\{R_{i,j}\}$, $\delta^2 = Var\{R_{i,j}\}$, $\mu = E\{X_j^{(i)}\}$ e $\sigma^2 = Var\{X_j^{(i)}\}$, e os valores correspondentes calculados são expressos em "número de janelas", e devem ser convertidos para as unidades de tempo adequadas, através da velocidade do canal e da unidade de dados utilizada em uma aplicação particular.

Neste ponto, recomenda-se, para maior esclarecimento, uma inspeção detalhada da Seção 7 de /KONHA74/ (resultados numéricos), onde são ilustradas algumas utilizações do modelo, a forma dos parâmetros de entrada, bem como diversos conjuntos de curvas sobre o desempenho do sistema considerado.

CAPÍTULO V

ANÁLISE DE DESEMPENHO DE SUB-SISTEMAS DE COMUNICAÇÃO DE DADOS "HALF-DUPLEX" SOB DISCIPLINA DE "POLLING"

Em seguimento a etapas de conceituação e definições, apresentação de tópicos de teoria matemática relevantes e inspeção de algumas análises detalhadas de modelos de multifilas, conforme realizado, respectivamente, nos Capítulos II, III e IV desta tese, este Capítulo V é, objetivamente a resolução desta fase de preparação conceitual e didática orientada, apresentando resultados de caráter prático, sob a forma de exercícios de aplicação de um modelo específico /SYKEJ70/ ao tratamento do problema objeto colocado no Capítulo I.

É analisado o desempenho de sub-sistemas de comunicação de dados "half-duplex" (HDX) sob disciplina de "polling". Em particular, são considerados sub-sistemas típicos de comunicação entre uma unidade de controle de comunicações 3705-IBM (ou equivalente) e uma unidade de controle de terminais do tipo 327X-IBM (ou equivalente), com protocolo BSC e "polling" genérico (ver /HOUST79,Cap. 4/, /STUTB72/, /BJØRD70/, IBM/GA27-3004-02/ e IBM/GA27-2749-10/, para detalhes deste ambiente). A análise é realizada em regime estacionário (processos homogêneos) e sem a presença de erros (não ocorrem retransmissões de mensagens).

Estes exercícios de análise refletem experiências realizadas pelo autor, no contexto de sub-sistemas de comunicação de dados em operação, as quais ilustram e validam a utilização e a qualidade dos resultados obtidos através do modelo considerado.

Na Seção V.1 é mostrado o mapeamento "sistema físico x modelo" proposto, a partir de descrição geral da mecânica de funcionamento do sistema e de um conjunto modelo esquemático -diagrama de transição referente à classe de modelos considerada.

Na Seção V.2 realiza-se um mapeamento detalhado do funcionamento e variáveis do sistema sobre o modelo (parâmetros e resultados) de Sykes, cujo desenvolvimento é apresentado na Seção IV.2 deste trabalho.

Em ambas as Seções V.1 e V.2 são evidenciadas as aproximações ou "distorções de modelagem" assumidas pelo autor . O mapeamento proposto não tem pretensões de unicidade ou de maior fidelidade, mostrando-se, entretanto, bastante natural e conseqüente, na opinião do autor, além de bastante robusto aos testes de validação realizados.

Na Seção V.3 são apresentadas e discutidas diversas curvas de desempenho, parametrizadas em função de variáveis de maior interesse prático, em cuja discussão se procura chamar a atenção para aspectos de tipo, oportunidade, escôpo e espaço de validade dos resultados/conclusões obtidos.

Aspectos relativos à validação dos resultados fornecidos pelo modelo, junto ao sistema em operação disponível , bem como considerações adicionais aplicáveis, são objeto da Seção V.4, que conclui este Capítulo V.

SEÇÃO V.1

MAPEAMENTO SISTEMA FÍSICO x MODELO

Os sistemas físicos geradores de motivação para a pesquisa realizada correspondem à classe dos sub-sistemas de comunicação de dados "half-duplex" (HDX) entre uma estação primária e uma ou mais estações secundárias, regidos por disciplina de "polling", conforme aqueles encontrados em sistemas de teleprocessamento que utilizam componentes da família 3270 de terminais/controladores IBM ou equipamentos compatíveis de outros fornecedores.

Em particular, é proposta a utilização de um modelo de multifilas com 2 filas, conforme o analisado por SyKes (ver Seção IV.2), na análise de desempenho da comunicação entre uma estação primária, por exemplo, uma unidade de controle de comunicações de um computador, e uma estação secundária,por exemplo, uma unidade de controle de terminais, utilizando um protocolo de comunicações do tipo BSC com "polling" genérico (ver IBM/GA27-3004-02/).

No Capítulo VI deste trabalho é indicada uma exten

são deste estudo, já em realização pelo autor, para o tratamento de diversas estações secundárias em ligação multiponto com uma estação primária.

A pesquisa dos elementos funcionais e experimentais apresentados neste Capítulo V foi realizada através de estudos, observações e medições pelo autor, em sub-sistemas de comunicação do Sistema de Agências On-Line, desenvolvido pela Itaú Tecnologia S/A. Nestes sistemas-objeto a estação primária corresponde a uma terminação de uma unidade de controle de comunicações 3705-IBM e a estação secundária a um concentrador de terminais (por ocasião deste estudo, um modelo com capacidade de 16 terminais), desenvolvido pela Itaútec, que emula uma unidade de controle de terminais IBM do tipo 3272 (ver IBM/GA27-2749-10/).

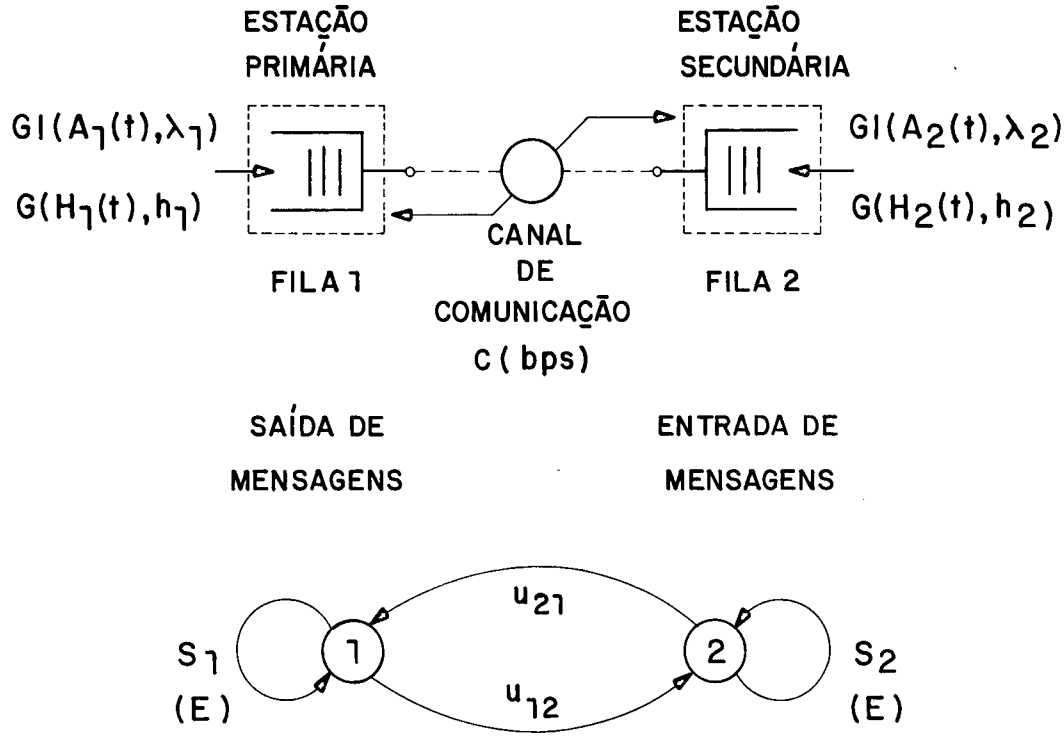
A operação de um sistema desta natureza pode ser visualizada, de modo simplificado, através do modelo esquemático e do diagrama de transição apresentados na Figura V.1, observando-se a seguinte atribuição de índices:

FILA1 = estação primária = fila de saída da unidade de controle de comunicações.

FILA2 = estação secundária = fila de saída da unidade de controle de terminais.

Para efeito da análise em regime estacionário realizada, são consideradas independentes e infinitas as filas de entrada e saída de cada estação, ou seja, a unidade de controle de

terminais pode sempre receber uma mensagem, quando endereçada, independentemente do estado de sua fila de entrada, enquanto na unidade de controle de comunicações não se verifica operação em "slow-down mode" (falta de "buffers" para recepção de mensagens da estação secundária, o que implica em alteração da política de "polling", no sentido de liberar "buffers" de sua fila de saída). Tais premissas são geralmente verificadas em condições normais de operação, a menos de erros de dimensionamento de recursos no sistema ou condições de sobrecarga não previstas.



u_{ij} = tempo de transição fila i - fila j

$S_i(E)$ = serviço exaustivo na fila i

Figura V.1 - Modelo esquemático e diagrama de transição do sistema-objeto

A operação do sistema pode ser descrita pela seguinte seqüência de eventos, considerando que o controle de acesso à linha é exercido pela estação primária:

A estação primária convida a secundária a transmitir mensagens eventualmente existentes em sua fila de saída (no caso de "polling" genérico, sem endereçar terminais específicos, mas sim, a estação secundária como uma única fonte de tráfego). Caso a estação secundária não apresente qualquer mensagem em seus "buffers" por ocasião deste convite ("poll"), a mesma responde com uma seqüência de controle apropriada, significando nada a transmitir (para detalhes do protocolo de comunicações ver IBM /GA 27-3004-02). Este tempo de convite/resposta é atribuido ao tempo de transição u_{12}. Se existem mensagens a serem transmitidas, a estação secundária ganha controle da linha e transmite mensagens até que sua fila atinja o estado zero.

Três observações: a) nesta situação ocorre uma diferença no tempo u_{12}, que só corresponde ao tempo de convite ("poll"), a qual é, entretanto, bastante diluída ao longo da operação, geralmente em faixas de baixa utilização de linha (<30%); b) o que consideramos tempo de serviço de uma mensagem na direção 2→1 deve incluir todos os elementos de tempo associados com a confirmação da recepção (seqüência de "ack"), bem como tempos de reversão de linha ("turnaround") e tempos de reação de equipamentos terminais envolvidos (por exemplo, tempo de cálculo de um "bcc"); c) após a transmissão da última mensagem ocorre o envio de uma seqüência de controle de fim de transmissão, não representado pelo modelo.

Após atendimento exaustivo da fila 2,a estação primária verifica a existência de mensagens para transmissão, em sua própria fila de saída, em um tempo de transição u_{21}, geralmente bastante pequeno, a menos de condições de sobrecarga, uma vez que se trata de um tempo interno de máquina, não envolvendo transmissões remotas. Caso não exista tráfego de saída, a estação primária procede ao "polling" da fila 2 (observar que esta situação corresponde à operação "keep-switching" estudada por Sykes e Eisenberg).

Se existem mensagens destinadas à estação remota (mais exatamente, a qualquer terminal conectado àquela estação secundária), a estação primária procede ao endereçamento da unidade de controle/terminal, recepção de uma seqüência de "ack" , envio da mensagem e recepção da confirmação associada. Todos estes componentes de tempo, adicionados dos devidos tempos de reversão e reação compõem o tempo de retenção da linha para serviço de uma mensagem na direção 1→2. A fila de saída da estação primária é atendida exaustivamente até o estado zero, quando então se reinicia o ciclo de operação descrito, conforme ilustrado na Figura V.2.

A despeito de ligeiras diferenças encontradas entre a dinâmica do modelo considerado e a operação do sistema real em estudo, o quadro delineado mostrou-se bastante encorajador , indicando a continuação do trabalho de aplicação do modelo ao problema de análise de desempenho do sistema físico.

A atitude adotada neste ponto da pesquisa caracte-

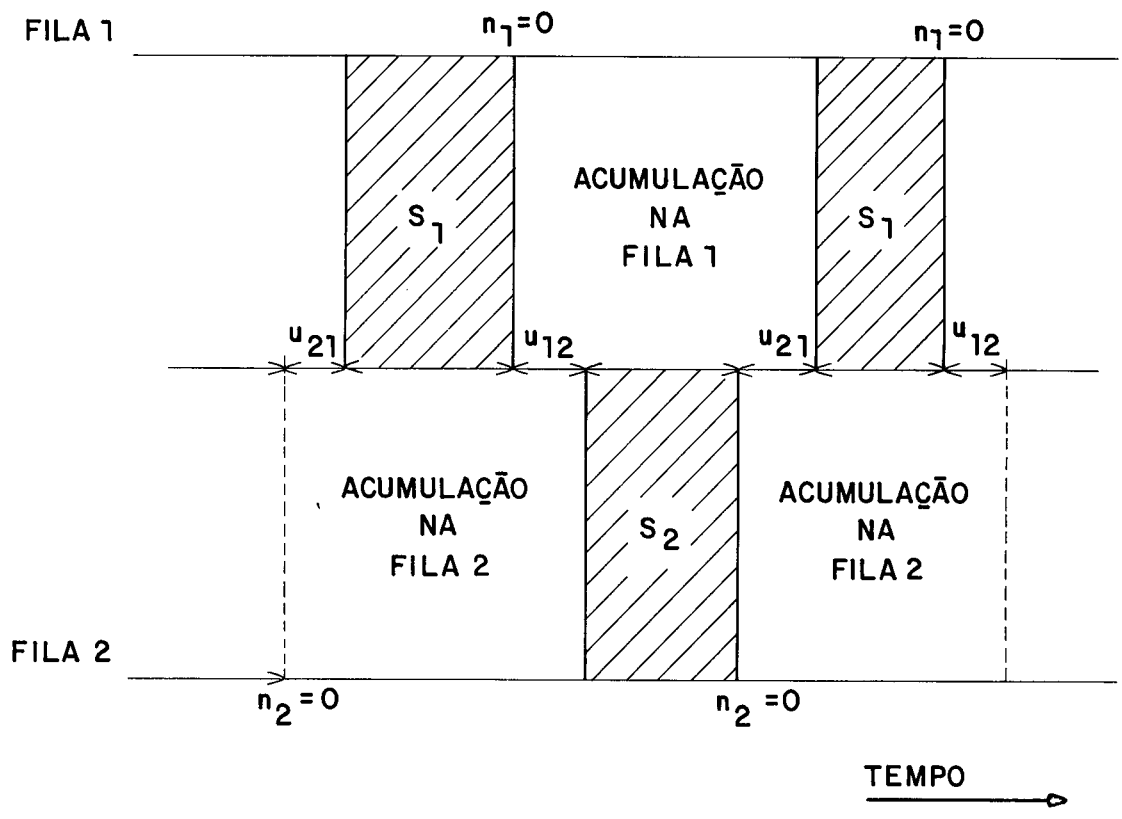


Figura V.2 - Ciclo de operação do sistema

rizou-se por uma forte intuição e mesmo, otimismo com relação à utilidade prática de resultados de um modelo analítico , ainda que aproximado (ver Seção V.4 quanto às aproximações obtidas) , sem preocupações excessivas com um rigor absoluto na representação.

Na Seção V.2 seguinte é apresentada com detalhes a utilização dos resultados de Sykes para o modelo de multifilas com duas filas M/G/*, serviço exaustivo, atendimento cíclico ordinário, tempos de transição finitos arbitrariamente distribuídos e transição contínua da unidade de serviço com sistema em estado zero, conforme /SYKEJ70/.

A premissa de tráfego Poisson, embora não subsidia-

da por identificações rigorosas das distribuições de tempos entre chegadas, pode ser verificada aproximadamente em diversos pontos (fontes de tráfego) no sistema, através de inspeções visuais de distribuições empíricas e de coeficientes de variação bem próximos de 1, que caracterizam distribuições exponenciais.

A premissa de serviço geral, caracterizado por média e 2º momento é bastante confortável para o(s) usuário(s) do modelo e não requer qualquer validação/testes, apenas a coleta dos dados.

SEÇÃO V.2

UTILIZAÇÃO DO MODELO DE SYKES

Nesta seção é apresentada com detalhes, a utilização do modelo de multifilas analisado por Sykes em /SYKEJ70/ , cujo desenvolvimento e resultados estão descritos de maneira resumida na Seção IV.2, na análise de desempenho dos sistemas objeto da Seção V.1 anterior.

A medida de desempenho básica considerada é a soma dos atrasos médios de entrada e saída de mensagens, conforme fornecidos pela avaliação da expressão 4.2.20 ($\bar{D}_i$) para as filas 1 (saída) e 2 (entrada).

Para efeito da avaliação dos componentes de tempo relacionados com a transmissão no segmento de comunicações, adota-se neste trabalho a medida do tempo ou atraso no sistema (es

pera + serviço), em ambas as direções de transmissão.

A razão para estudo da soma dos atrasos de entrada e de saída, consiste no objetivo de se avaliar, particulamente, a contribuição do segmento de comunicações para o tempo de resposta total de transações que envolvem uma mensagem de entrada e uma de saída correspondente.

Ressalva-se entretanto que o modelo em questão não impõe qualquer restrição quanto ao ambiente/tipo de aplicação considerados, no que toca à relação entre os tráfegos de entrada e de saída.

Conforme descrito na Seção IV.2 ,o modelo de Sykes considera os seguintes parâmetros:

λ_i = taxa média de chegadas Poisson à fila i, i=1,2.

$\bar{t}_i$ = tempo médio de serviço de itens da fila i, i=1,2.

$t_i^{(2)}$ = segundo momento do tempo de serviço de itens da fila i, i=1,2.

$\bar{r}_{ij}$ = tempo médio de transição fila i-fila j, i,j=1,2; i≠j.

$r_{ij}^{(2)}$ = segundo momento do tempo de transição fila i-fila j, i,j=1,2; i≠j.

Ainda, por definição:

$\rho_i = \lambda_i \bar{t}_i$ = fator de utilização da fila i, i=1,2.

$C^{(2)}(r_{ij}) = r_{ij}^{(2)} / \bar{r}_{ij}^2 - 1$ = coeficiente de variação quadrático do tempo de transição r_{ij}, i,j=1,2; i≠j.

$\rho=\rho_1+\rho_2$ = utilização total da unidade de serviço.

Para os sistemas em estudo, descritos na Seção V.1, estes parâmetros podem ser obtidos, diretamente ou através de manipulações simples, a partir de estimadores estatísticos de primeira e segunda ordem (média e variância ou desvio padrão) geralmente disponíveis, ou facilmente observáveis, em sistemas de coleta e análise de dados estatísticos associados a sistemas de teleprocessamento. A unidade de tempo adotada no desenvolvimento que se segue é, uniformemente, o segundo.

a) λ_i = taxa média de chegadas Poisson à fila i , i=1,2;

A partir da verificação da hipótese de chegadas Poisson, seja conforme descrito na Seção V.1 anterior, ou por meio de métodos mais rigorosos, ou ainda por suposição a priori (este tipo de processo é sempre utilizado na falta de maiores informações, vide base adotada internacionalmente para todos os cálculos relativos a tráfego telefônico, telex e dados), resulta imediatamente a taxa média de chegadas λ_i, ou o tempo médio entre chegadas $1/\lambda_i$.

Este parâmetro é tipicamente expresso em mensagens/segundo ou transações/segundo. Em sistemas de transações on-line, ou sistemas de consulta/resposta ocorre freqüentemente que $\lambda_1 \cong \lambda_2$, não havendo, entretanto, quaisquer restrições quanto à relação entre estes parâmetros.

b) $\overline{t}_i$ = tempo médio de serviço de itens da fila i , i=1,2;

b.1) i = 2

O tempo de serviço médio para itens da fila 2 (mensagens no sentido unidade de controle de terminais → unidade de controle de comunicações) pode ser sintetizado pelas seguintes componentes:

.$\overline{t}_{21}$ = tempo médio para transmissão da mensagem (texto), com tamanho médio de $\overline{nt}_2$ caracteres, e caracteres de controle associados, em número $\overline{nc}_2$, geralmente constante; se a velocidade da linha é c bps tem-se

$$\overline{t}_{21} = \frac{(\overline{nt}_2+\overline{nc}_2)x8}{c}$$, considerando 8 bits por caracter.

.$\overline{t}_{22}$ = tempo médio para transmissão de uma mensagem de confirmação da recepção correta ("ack"), pela estação primária, com tamanho médio (geralmente constante) de $\overline{na}_1$ caracteres, na linha de c bps

$$\overline{t}_{22} = \frac{\overline{na}_1 x8}{c}$$

.$\overline{t}_{2R}$ = tempos médios de reação e reversão envolvidos na seqüência mensagem-confirmação,

como:

- tempo de reação da estação primária para resposta à recepção da mensagem, incluindo atrasos de atendimento internos e tempo para cálculo do "bcc", até início do envio do "ack";
- tempo de reversão da direção de transmissão "turnaround" em linhas a 2 fios, que no caso de linhas a 4 fios pode ser desconsiderado, ou condicionado apenas pela estação primária.

Por razões de simplicidade e limitação da notação, não se atribuem nomes ou se listam exaustivamente estes fatores componentes de $\overline{t}_{2R}$, que podem inclusive variar com o ambiente. Na utilização do modelo a parcela $\overline{t}_{2R}$ deve conter todos os eventuais tempos desta natureza, ou seja, aqueles não decorrentes de transmissão líquida de caracteres na linha.

Desta forma, o tempo médio de serviço para mensagens da fila 2 pode ser expresso por

$$\overline{t}_2 = \frac{8x(\overline{nt}_2+\overline{nc}_2+\overline{na}_1)}{c} + \overline{t}_{2R}. \qquad (5.2.1)$$

b.2) i = 1

O tempo de serviço médio para itens da fila 1 (mensagens no sentido estação primária → estação secundária) pode ser sintetizado pelas seguintes componentes:

. $\overline{t}_{11}$ = tempo para transmissão da mensagem (texto), com tamanho médio de $\overline{nt}_1$ caracteres, e caracteres de controle associados, em número $\overline{nc}_1$, geralmente constante; se a velocidade da linha é c bps tem-se

$$\overline{t}_{11} = \frac{(\overline{nt}_1+\overline{nc}_2)x8}{c}$$, considerando 8 bits por caracter.

. $\overline{t}_{12}$ = tempo para transmissão de uma mensagem de confirmação da recepção correta ("ack"), pela estação secundária,com tamanho médio (geralmente constante) de $\overline{na}_2$ caracteres, na linha de c bps

$$t_{12} = \frac{\overline{na}_2x8}{c}$$

. $\overline{t}_{13}$ = tempo para transmissão de uma seqüência de endereçamento pela estação primária , composta de $\overline{ne}_1$ caracteres, geralmente constante, e respectiva resposta positiva pela estação secundária (seqüência "ack"), com comprimento de $\overline{na}_2$ caracteres, na linha de c bps

$$\overline{t}_{13} = \frac{(\overline{ne}_1+\overline{na}_2)x8}{c}$$

. t_{1R} = tempos médios de reação e reversão envol

vidos na seqüência endereçamento-confirmação-mensagem-confirmação, de maneira análoga ao descrito para a parcela $\overline{t}_{2R}$, permanecendo válidas as observações relativas ao detalhamento da composição desta parcela.

Desta forma, o tempo médio de serviço para mensagens da fila 1 pode ser expresso por

$$\overline{t}_1 = \frac{(\overline{nt}_1+\overline{nc}_1+\overline{ne}_1+2x\overline{na}_2)x8}{c} + \overline{t}_{1R} \ . \qquad (5.2.2)$$

c) $t_i^{(2)}$ = segundo momento do tempo de serviço de itens da fila i, i=1,2:

O fator determinante da dispersão estatística do tempo de serviço destes sistemas é a dispersão (variância) do comprimento das mensagens a serem transmitidas a partir das filas 1 e 2. Conforme vem sendo sistematicamente indicado no decorrer desta Seção V.2, os comprimentos das seqüências de controle envolvidas são geralmente fixas em um determinado ambiente de operação, contribuindo com variâncias nulas para o tempo de serviço.

Exceção pode, em alguns casos, ser feita a componentes de reação/processamento interno de equipamentos terminais, que podem apresentar variações em função da carga a que estão submetidos instantâneamente.

Entretanto, o levantamento estatístico e a inclusão deste grau de precisão em análises de valor médio,em condições estacionárias, e geralmente de tráfego de pico, não nos parece justificável ou compatível com a utilização prática pretendida e proposta para este modelo.

Sugere-se a utilização de valores típicos medidos/estimados/especificados, considerados com variância nula , bem como, em casos aplicáveis, de limites superiores destes tempos, na obtenção de padrões de desempenho em condições de pior caso. Afirma-se, entretanto, que a consideração de variâncias destes parâmetros é matéria trivial.

c.1) i=2

Seja $var(nt_2)$ a variância do comprimento de mensagem da estação secundária. Como tempos de serviço (transmissões na linha) estão relacionados com comprimentos de mensagens através da constante 8/c, onde c é a velocidade da linha em bps, a variância da variável aleatória correspondente pode ser obtida através da constante $(8/c)^2$. Uma vez que a variância do tempo de serviço para a fila 2 corresponde à $var(nt_2)$ tem-se

$$t_2^{(2)} = var(nt_2) \times (\frac{8}{c})^2 + (\bar{t}_2)^2 . \qquad (5.2.3)$$

c.2) i=1

De maneira análoga, tem-se para a estação primária

$$t_1^{(2)} = \text{var}\ (nt_1)x(\frac{8}{\bar{c}})^2 + (\bar{t}_1)^2\ , \qquad (5.2.4)$$

onde $\text{var}(nt_1)$ representa a variância do comprimento de mensagens transmitidas pela estação primária.

d) $\bar{r}_{ij}$ = tempo médio de transição fila i - fila j, $i=1,2;\ i \neq j$.

d.1) $\bar{r}_{21}$

Conforme pode ser observado na Figura V.2 da Seção V.1 anterior, o tempo de transição r_{21}, na representação genérica denotado por u_{21}, corresponde ao intervalo decorrido entre o atingimento do estado zero na fila 2 (não há mais mensagens para transmissão neste ciclo) e um início de serviço na fila 1, ou, como se trata de operação "keep-switching" , um início de transição r_{12}, caso a fila 1 esteja vazia.

No caso da classe dos sistemas reais em estudo, tal tempo de transição corresponde apenas a um pequeno tempo interno que a unidade de controle de comunicações gasta para verificar o estado de sua fila de saída (existência de mensagens para início de um ciclo de serviço) após atendimento exaustivo da fila 2.

Este tempo pode ser obtido por estimativas ou medições nos(s) sistema(s) em questão, e dada a não-trivialidade de sua determinação exata, é considerado, para fins de uti

lização do modelo, como constante (variância nula), embora ainda levado a valores mínimos típicos ou máximos em exercícios de análise de sensibilidade.

d.2) $\bar{r}_{12}$

Conforme pode ser observado na Figura V.2 da Seção V.1, o tempo de transição r_{12} na representação genérica denotado por u_{12}, corresponde ao intervalo decorrido entre o atingimento do estado zero na fila 1 (não há mais mensagens para transmissões neste ciclo) e um início de serviço na fila 2 , ou um início de transição r_{21}, caso a fila 2 esteja vazia.

Este tempo corresponde ao tempo gasto para a transmissão de uma seqüência de "polling" e respectiva resposta. A resposta negativa (não há tráfego a transmitir) é uma seqüência de controle ("eot") e a resposta positiva é o início da própria transmissão de mensagens.

Uma vez que a incidência de respostas negativas é dominante (o "polling" é contínuo, com uma determinada freqüência, em função da carga total da estação primária,e a utilização destas linhas é tipicamente menor que 30%), sugere-se incluir em $\bar{r}_{12}$ os tempos médios para um "poll" e respectiva resposta negativa. Em caso de existência de tráfego nas condições citadas, o acréscimo devido à resposta (que nesta situação não ocorre) é diluído ao longo dos atrasos sofridos por todas as mensagens transmitidas naquele ciclo.

Este tempo pode ser obtido por estimativas

ou medições, direta ou indiretamente inclusive através de recursos de monitoração do próprio sistema central, que pode fornecer dados como freqüência de "polling" (número de "polls"/segundo), resultados destas operações (positivo/negativo), entre outros. Da mesma maneira que para $\bar{r}_{21}$, tempo $\bar{r}_{12}$ é considerado constante.

e) $r_{ij}^{(2)}$ = segundo momento do tempo de transição fila i - fila j, i=1,2; i≠j:

Por razões já mencionadas ao longo desta Seção V.2 (itens c e d), não são investigados neste trabalho, estimadores de segunda ordem dos tempos de transição r_{ij}, ou seja, os coeficientes de variação quadráticos $C^{(2)}(r_{ij})$ são identicamente nulos.

Tal consideração reduz a expressão original obtida por Sykes, para o atraso médio $\bar{D}_i$, a uma forma simplificada, apresentada como eq. 4.2.20 no Capítulo IV desta dissertação, aqui transcrita:

$$\bar{D}_i = \frac{\lambda_j t_j^{(2)} (1-\rho_i)^2 + \lambda_i t_i^{(2)} \rho_j^2}{2(1-\rho_i)(1-\rho)(1-\rho+2\rho_i\rho_j)} + \frac{\lambda_i t_i^{(2)}}{2(1-\rho_i)} + \frac{(1-\rho_i)[\bar{r}_{12}+\bar{r}_{21}]}{2(1-\rho)}$$

$i,j = 1,2 \; ; \quad i \neq j, \quad$ onde

$\rho_i = \lambda_i . t_i \quad$ e $\quad \rho = \sum_{i=1}^{2} \rho_i \; .$

É também apresentada uma expressão para o comprimento médio da fila i (na realidade, do sistema i, já que in

clui o item em serviço no instante), conforme eq. 4.2.21 do Capítulo IV:

$$\bar{L}_i = \lambda_i[\bar{D}_i+\bar{t}_i] = \lambda_i\bar{D}_i + \rho_i \cdot$$

Aplicando, então, as fórmulas de atrasos médios nas duas direções da comunicação,acrescidas dos respectivos tempos de serviço, obtém-se a parcela total de tempo gasta por uma transação, ou mensagem de consulta /resposta,no sub-sistema de comunicações em estudo. Desta forma, o que denominamos de tempo médio no sub-sistema de comunicações, $\bar{t}_{sc}$, pode ser expresso por

$$\bar{t}_{sc} = \bar{D}_1 + \bar{t}_1 + \bar{D}_2 + \bar{t}_2 \ , \quad \text{onde}$$

$\bar{t}_1$ e $\bar{t}_2$ estão calculados no item b desta Seção V.2, eqs. 5.2.2 e 5.2.1, e $\bar{D}_1$ e $\bar{D}_2$ são fornecidos pela expressão (4.2.20),tomando-se, respectivamente i=1 e j=2 e i=2 e j=1.

SEÇÃO V.3

PARAMETRIZAÇÃO DO COMPORTAMENTO DO SISTEMA

A Seção V.2 anterior é concluída com uma expressão para o tempo médio no sub-sistema de comunicações, $\bar{t}_{sc}$, cujas componentes são tempos de serviço e atrasos, nas filas de entrada e saída de sistemas do tipo descrito na Seção V.1.

No desenvolvimento da Seção V.2 pode ser identificada a existência de dois conjuntos de dados de entrada, ou parâ-

metros de operação: o conjunto de parâmetros "primitivos" do sistema, pertencentes ao cotidiano de engenheiros e/ou analistas, utilizados para descrição completa das condições de operação , e um conjunto de parâmetros derivados, "enxergados" pelo modelo, cujos elementos geralmente sintetizam diversos elementos do conjunto primitivo.

Tal característica é resultante do modo como foi realizado o mapeamento sistema físico x modelo, e do grau de detalhe adotado para a descrição do sistema.

Nesta Seção V.3 será estudado o comportamento da variável $\bar{t}_{sc}$, em relação aos parâmetros primitivos definidos, o que corresponde a necessidades práticas reais. Assim,estuda-se, por exemplo, a sensibilidade da variável $\bar{t}_{sc}$ com relação ao comprimento de mensagens ou ao tempo de reversão em uma linha a dois fios, e, não, em relação à média ou segundo momento do tempo de serviço ou ao tempo de transição $\bar{r}_{12}$, embora implícitos no processo.

A seguir são apresentados estes dois conjuntos de parâmetros (para facilidade de trabalho nesta Seção V.3),bem como as relações existentes entre os mesmos :

a) PARÂMETROS PRIMITIVOS:

.λ_1,λ_2 : taxas médias de mensagens, ou transações, por unidade de tempo,geradas nas estações primária e secundária .

.nt_1, nt_2 : números de caracteres de informação que compõem as mensagens das estações primária e secundária; os valores médios e variâncias associados são representados respectivamente por $\overline{nt}_1$, $var(nt_1)$ e $\overline{nt}_2$,$var(nt_2)$.

.$\overline{nc}_1$, $\overline{nc}_2$: números médios de caracteres de controle associados às mensagens das estações primária e secundária.

.c : velocidade (taxa de sinalização de dados) da linha de comunicação de dados, em bps; o circuito é considerado simétrico.

.$\overline{na}_1$, $\overline{na}_2$: números médios de caracteres de controle que compõem os blocos de confirmação ("ack") emitidos pelas estações primária e secundária.

.$\overline{ne}_1$: número médio de caracteres em blocos de endereçamento emitidos pela estação primária.

.$\overline{t}_{1R}$,$\overline{t}_{2R}$: tempos médios de reação e reversão envolvidos nas comunicações estação-primária e secundária-primária (ver Seção V.2, itens b e c).

.componentes de tempo que compõem os tempos de transição 1→2 e 2→1 do modelo (ver Seção V.2 , item d).

b) PARÂMETROS DERIVADOS:

.λ_1, λ_2 : correspondem às taxas médias de chegadas Poisson às filas 1 e 2 (são os próprios λ's primitivos).

.t_1, t_2 : tempos de serviço para itens das filas 1 e 2; os valores médios e segundos momentos associados são denotados respectivamente, por $\overline{t}_1$, $t_1^{(2)}$ e $\overline{t}_2$, $t_2^{(2)}$.

.r_{12}, r_{21} : tempos de transição fila 1→2, fila 2→1, após atendimento exaustivo; os valores médios associados são denotados, respectivamente, por $\overline{r}_{12}$ e $\overline{r}_{21}$.

.definições: $\rho_i = \lambda_i t_i$ = fator de utilização do recurso (linha de comunicação) pela fila i, i=1,2; $\rho=\rho_1+\rho_2$ = fator de utilização total do recurso.

Entre estes parâmetros são estabelecidas as seguintes relações:

$$\bar{t}_1 = 8 \times (\overline{nt}_1+\overline{nc}_1+\overline{ne}_1+ 2 \times \overline{na}_2)/c + \bar{t}_{1R} \quad (5.3.1)$$

$$t_2 = 8 \times (\overline{nt}_2+\overline{nc}_2+\overline{na}_1)/c + \bar{t}_{2R} \quad (5.3.2)$$

$$t_1^{(2)} = \text{var}(nt_1) \times (\frac{8}{c})^2 + (\bar{t}_1)^2 \quad (5.3.3)$$

$$t_2^{(2)} = \text{var}(nt_2) \times (\frac{8}{c}) + (\bar{t}_1)^2 \ . \quad (5.3.4)$$

Conforme mencionado na Seção V.2 anterior, não se detalha, intencional e sistematicamente neste estudo, a composição dos elementos de tempo associados com reações de estações , reversões de linha e transições entre filas, os quais são indicados para levantamento nos contextos específicos de ambiente/aplicação.

Uma análise de sensibilidade, do comportamento do sistema com relação a todos estes parâmetros,que corresponderia a um objetivo ideal e exaustivo da utilização deste modelo, implicaria na realização de um conjunto bastante extenso de experiências fatoriais (suponha-se, por exemplo, os 14 parâmetros primitivos com 6 níveis, ou seja, cada um assumindo 6 valores distintos !).

Em situações práticas o que ocorre, entretanto, são interesses localizados em alguns parâmetros, tais como tráfego oferecido (taxa de mensagens/tempo x comprimento de mensagens), velocidade das linhas, e tempos de reação/reversão (linhas 2 fios x 4 fios), uma vez que, em ambientes típicos, pouca variação sofrem, ou pequeno controle se tem sobre os demais parâmetros.

Nesta Seção V.3 é, então, analisado o comportamento do sistema, em termos da medida $\bar{t}_{SC}$, com relação ao tráfego (através das taxas λ's, que podem resultar de uma relação aproximada com o número de terminais da estação secundária, para um número não muito pequeno de terminais), comprimentos de mensagens, velocidades de linha e tempos de reação/reversão (através da consideração de tempos de "turnaround" típicos em linhas a 2 fios). Considera-se operação em ambiente livre de erros, ou seja, não ocorrem retransmissões de mensagens.

Os resultados são apresentados sob a forma de gráficos discretos interpolados, acompanhados de indicações a respeito dos valores adotados para os parâmetros relevantes.

Neste estudo, são considerados os seguintes valores, ou intervalos de valores, para os parâmetros em questão, conforme Figura V.3.

Os valores apresentados são bastante plausíveis em sistemas orientados para transações, e fundamentam-se, basicamente, tanto em observações, como em necessidades de predição de desempenho, relativas a diversos sub-sistemas de comunicação no Sistema de Agências On-Line Itautec.

Em /CHANJ75/, Chang, da IBM, em um tratamento analítico aproximado de sistemas que envolvem sub-sistemas do tipo considerado neste estudo, indica tabelas de parâmetros de operação relativos ao controle da linha (tempos e comprimentos de blocos de controle do protocolo BSC) e a estatística de tráfego de terminais, onde são encontrados valores bastante próximos/coe-

PARÂMETROS	INTERVALOS DE VALORES	VALORES TÍPICOS
$\lambda_1 = \lambda_2$	ε (0, 1, 2) mensagens/seg	0,5
$\overline{nt}_1$	ε (100, 2000) caracteres/mensagem	200
$var(nt_1)$	ε (1000,50000) (caracteres/mensagem)^2	30000
$\overline{nt}_2$	ε (50, 250) caracteres/mensagem	80
$var(nt_2)$	ε (50,1000) (caracteres/mensagem)^2	200
c	ε {2400,4800,9600} bits/seg	*
$\overline{nc}_1=\overline{nc}_2$	= 6 caracteres	-
$\overline{na}_1=\overline{na}_2$	= 4 caracteres	-
$\overline{ne}_1$	= 12 caracteres	-
$\overline{t}_{1R}$	ε (40, 300) ms, 4 fios	60
	ε (300,1000) ms, 2 fios	600
$\overline{t}_{2R}$	ε (10, 200) ms, 4 fios	20
	ε (50, 400) ms, 2 fios	200
$\overline{r}_{12}$	ε (50, 500) ms, 4 fios	200
	ε (150,1500) ms, 2 fios	600
$\overline{r}_{21}$	ε (10, 100) ms	20

FIGURA V.3 - INTERVALOS E VALORES TÍPICOS DE PARÂMETROS UTILIZADOS NO ESTUDO DO COMPORTAMENTO DO SISTEMA.

* = conforme indicado no(s) gráficos a seguir.

rentes com os indicados na Figura V.3.

Diferenças fundamentais consistem na especificação de tráfego por terminal (e não por unidade de controle de ter-

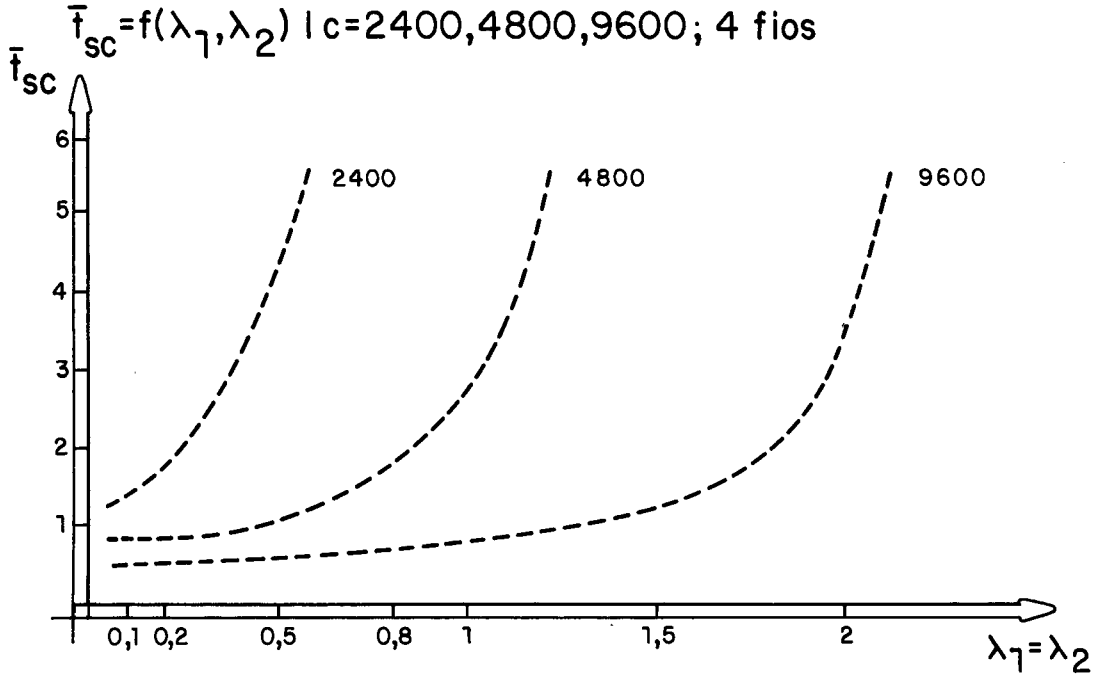
minais) e em um detalhamento maior de alguns parâmetros do NCP (programa de controle de rede da unidade de controle de comunicações 3705), uma vez que o autor não utiliza um modelo de multifilas, e procura caracterizar isoladamente cada um dos componentes de tempo/atraso resultantes da dinâmica de interação dos tráfegos de entrada e saída e da política de "polling"/"addressing" do NCP (ver Tabelas 3 e 4, pag. 280, de /CHANJ75/).

Ainda no mesmo trabalho, Chang apresenta resultados teóricos estimados relativos a tempos de resposta de terminais, incluindo processamento no computador "host". Estes resultados, entretanto,não são objeto de comparação ou validação com relação a valores reais observados em sistemas em operação.

Em /PAWLP81/, que sintetiza recentes resultados relativos a medições de tráfego de dados em sistemas interativos, podem também ser encontradas estatísticas de tráfego que corroboram diversos intervalos de valores adotados para diversos parâmetros constantes da Figura V.3.

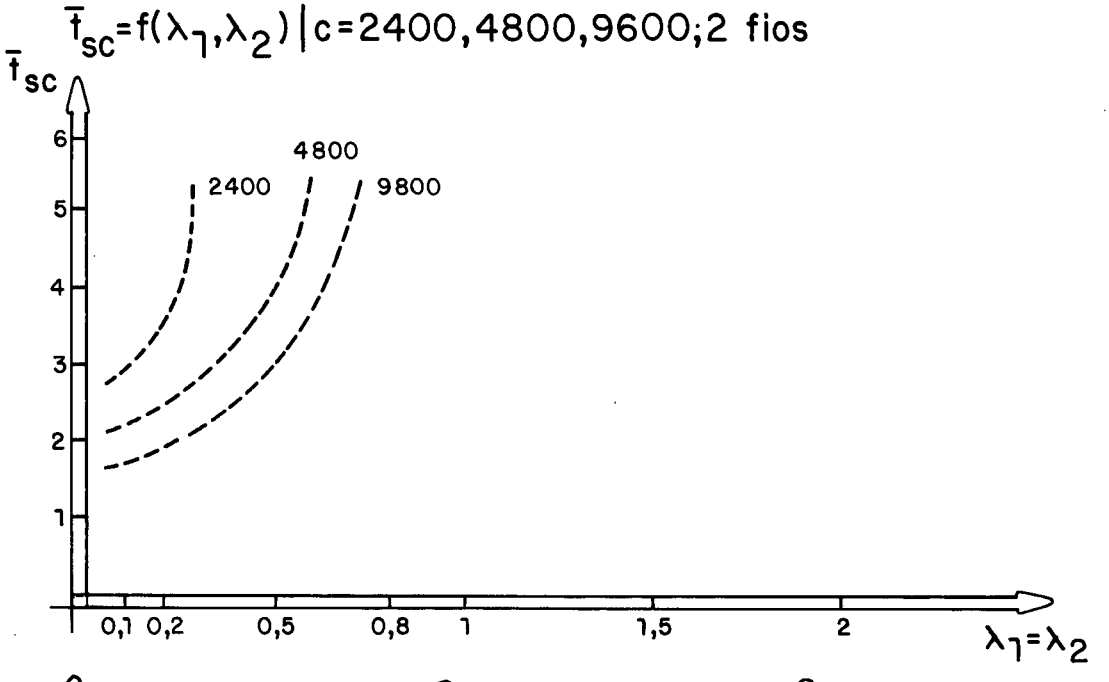
A seguir são apresentados alguns gráficos que descrevem o comportamento da variável $\bar{t}_{sc}$ em diferentes condições de operação. Parâmetros não mencionados explicitamente assumem os valores típicos indicados na Figura V.3.

GRÁFICO 5.1



$\rho_{2400} \in (0.115, 0.573); \rho_{4800} \in (0.061, 0.613); \rho_{9600} \in (0.035, 0.693)$

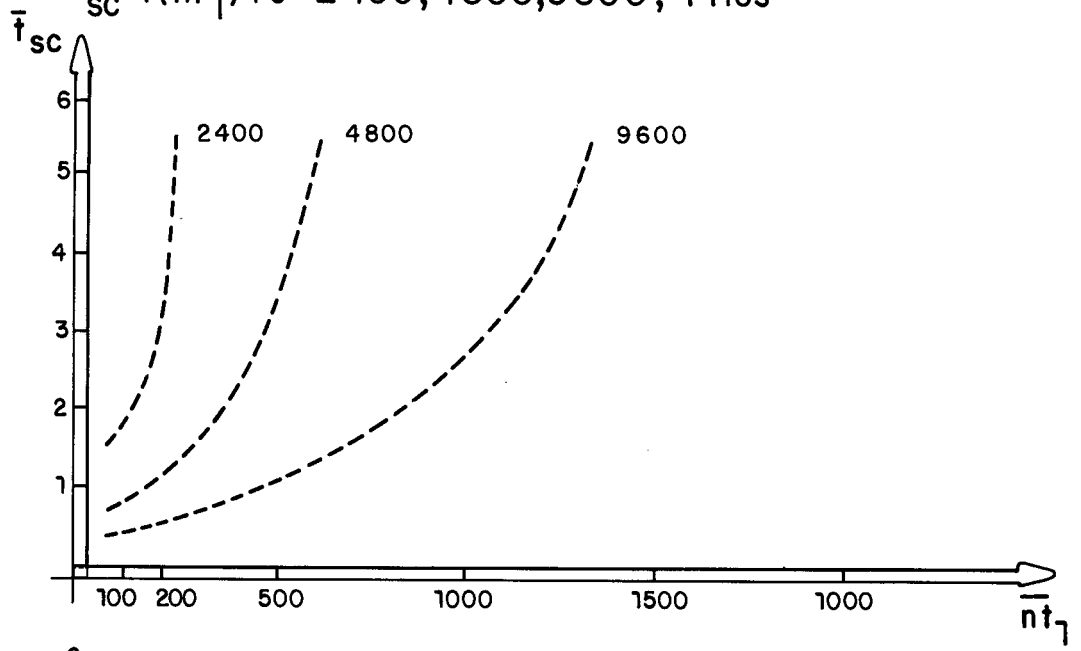
GRÁFICO 5.2



$\rho_{2400} \in (0.187, 0.373); \rho_{4800} \in (0.133, 0.667); \rho_{9600} \in (0.107, 0.533)$

GRÁFICO 5.3

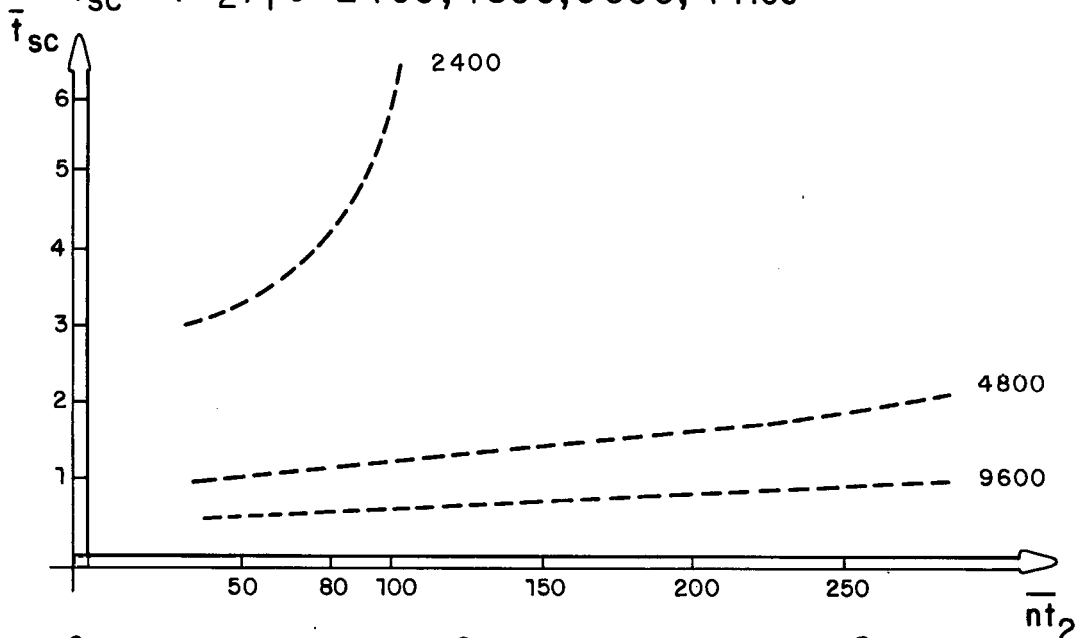
$\bar{t}_{sc} = f(\bar{nt}_1) \mid c = 2400, 4800, 9600; 4 \text{ fios}$



$\rho_{2400} \in (0.407, 0.573); \rho_{4800} \in (0.223, 0.56); \rho_{9600} \in (0.132, 0.715)$

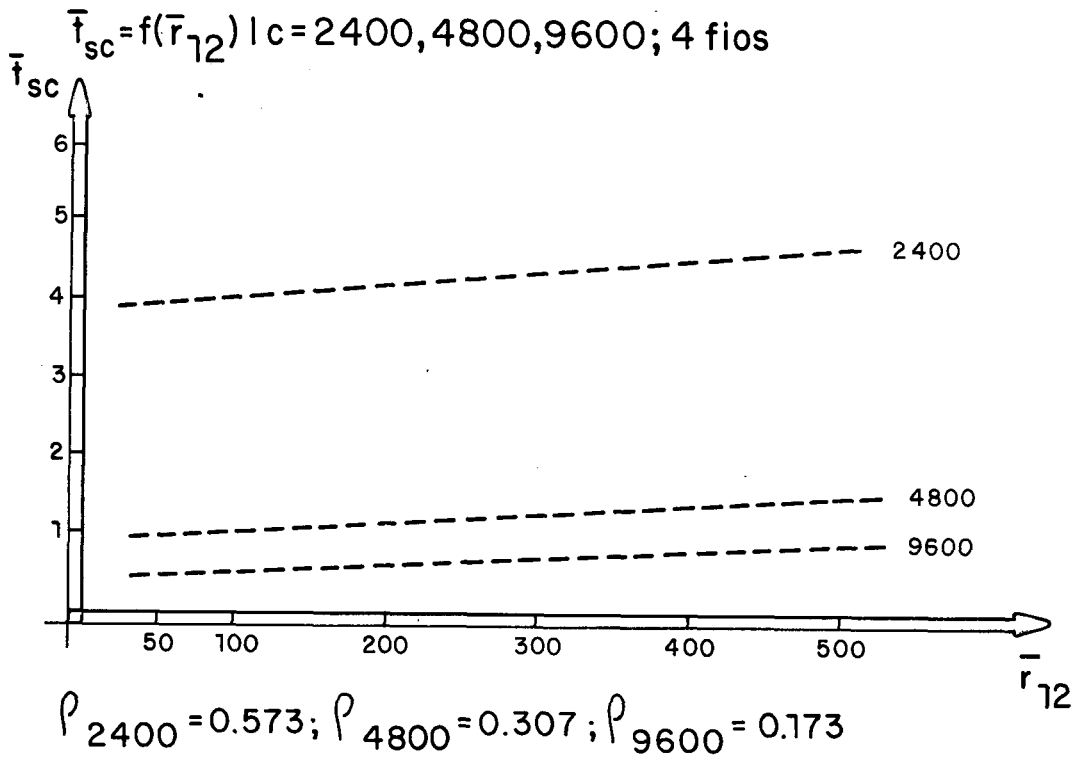
GRÁFICO 5.4

$\bar{t}_{sc} = f(\bar{nt}_2) \mid c = 2400, 4800, 9600; 4 \text{ fios}$



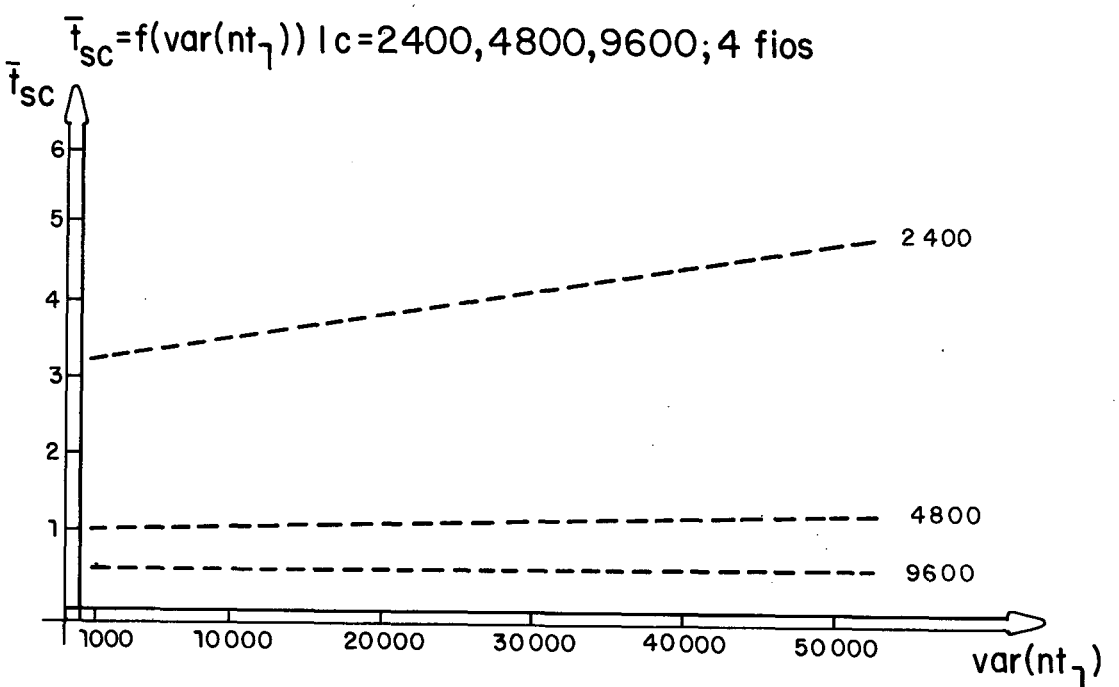
$\rho_{2400} \in (0.523, 0.69); \rho_{4800} \in (0.282, 0.45); \rho_{9600} \in (0.161, 0.244)$

GRÁFICO 5.5



$\rho_{2400}=0.573$; $\rho_{4800}=0.307$; $\rho_{9600}=0.173$

GRÁFICO 5.6



$\rho_{2400}=0.573$; $\rho_{4800}=0.307$; $\rho_{9600}=0.173$

A observação dos GRÁFICOS 5.1-5.6, juntamente com a Figura V.3, pode conduzir aos seguintes comentários/conclusões (não exaustivas), aplicáveis a sistemas de consulta/resposta on-line, de alta performance, onde se desejam tempos médios de respostas da ordem de 3 a 4 segundos (considerando as parcelas de tempo de processamento no "host", atrasos no método de acesso e outros eventuais componentes de atraso até os terminais, atingindo um total inferior a 1 segundo):

1 - O GRÁFICO 5.1 mostra que taxas de mensagens até 0,2 mensagens/segundo são perfeitamente suportadas por linhas a 4 fios com as 3 velocidades consideradas, indicando $\bar{t}_{sc}<2$ segundos em todos os casos. Notar, entretanto, que a 2400 bps começa-se a atingir uma região de instabilidade de operação, ou sensibilidade acentuada ao tráfego, na medida em que a 0.4 mensagens/segundo, $\bar{t}_{sc}$ eleva-se a aproximadamente 3 segundos, atingindo a região de saturação em torno de 0,5 mensagens/segundo.

As curvas a 4800 e 9600 bps mostram-se bastante mais estáveis, respectivamente, nas faixas até 0,5 e 1,5 mensagens/segundo.

Devem ser considerados, entretanto, na escolha de velocidades de linhas para uma operação segura e, ao mesmo tempo, eficiente (em termos de utilização do recurso), determinados limites de tráfego atingíveis no contexto parti

cular onde se realizam análises desta natureza.

Por exemplo, no caso de um concentrador com 16 terminais, onde cada operador de terminal trabalhasse com máxima atividade, digamos, realizando uma transação a cada 30 segundos, o tráfego máximo de mensagens gerado por este concentrador, para a linha de comunicação, seria de aproximadamente 0,5 mensagens/segundo, o que indicaria operação segura a 4800 bps. Nesta situação, linhas de 9600 bps seriam recursos sub-utilizados que não acarretariam melhoria substancial no desempenho, ou segurança adicional quanto a flutuações de tráfego.

Ainda neste caso, supondo-se atingíveis taxas de até 0,25 mensagens/segundos (mais razoáveis em se tratando de operadores humanos em ambientes de atendimento, correspondendo a uma transação a aproximadamente cada minuto), poderia ser também considerada operação relativamente segura a 2400 bps, por exemplo, por razões de custo de uma linha dedicada especializada interurbana ou de limitação por distância de velocidade de operação em banda-base.

Outras situações poderiam ser criadas, por outro lado, se se considera a comunicação entre unidades inteligentes, segundo este mesmo mecanismo, onde o tráfego não seria gerado por operadores e, sim, automáticamente,

ou resultante de um processo de concentração multinível por exemplo, em uma comunicação entre processadores/concentradores regionais, ou mesmo entre computadores ou quaisquer outros equipamentos que emulam unidades de controle do tipo 327X, onde o tráfego poderia atingir, digamos, 1,5 mensagens/segundo em uma linha de 9600 bps (notar que são mantidas as estatísticas de comprimentos de mensagens da Figura V.3, o que não é necessáriamente verdade, ou típico, para uma comunicação entre computadores; pretende-se aqui, exatamente, demonstrar e estimular o exercício da utilização de um modelo e a análise dos resultados por ele fornecidos, no contexto de ambientes reais e/ou plausíveis).

2 - O GRÁFICO 5.2 mostra, por exemplo, que operação em linhas a 2 fios com tempos de reversão típicos (conforme Figura V.3), apenas seria viável a 9600 bps, no limite de 0,5 mensagens/segundos comentado no item anterior. Linhas de 4800 bps poderiam ser utilizadas com relativa segurança e estabilidade na presença de taxas de até 0,2 mensagens/segundo. Operação a 2400 bps não seria recomendável neste ambiente, uma vez que $\bar{t}_{sc} \cong 3$ segundos, mesmo a baixas taxas de tráfego.

3 - O comportamento da variável de desempenho $\bar{t}_{sc}$ em função do comprimento médio das mensagens de saída (respostas),

$\overline{nt}_1$ é mostrado no GRÁFICO 5.3, em linhas a 4 fios. Observa-se que, mantidas as demais condições de operação, um aumento no comprimento médio das mensagens de saída mostra-se praticamente insuportável a 2400 bps, uma vez que com $\lambda_1 = \lambda_2 = 0,5$ e $\overline{nt}_1 = 200$ (ver GRÁFICO 5.1) a operação a 2400 bps atinge regiões já não admissíveis. Mensagens de saída menores que 100 caracteres poderiam, entretanto, viabilizar operação a 2400 bps, embora o aspecto da curva indique a proximidade de uma região de saturação, o que pode ser eventualmente atingido por uma sensibilidade acentuada a um outro parâmetro, em torno do ponto $\overline{nt}_1 = 100$ caracteres.

Enfatiza-se aqui, mais uma vez, a necessidade do exercício de análises mais abrangentes e completas, uma vez definido(s) ponto(s) desejado(s) ou de interesse para operação do sistema. A existência destas condições objetivas vem então justificar maior aprofundamento e eventualmente, recomendar experiências fatoriais exaustivas, com relação aos parâmetros mais determinantes do desempenho.

Aumentos de $\overline{nt}_1$ até valores em torno de 400 e 1000 caracteres são suportados, com desempenho satisfatório, respectivamente, por 4800 e 9600 bps.

Deve-se considerar também, que os requisitos de tempo de

resposta, bem como as taxas de mensagens atingíveis em aplicações onde $\overline{nt}_1$ pertence ao intervalo de 500 a 1.500 caracteres (grandes áreas de tela de terminais de video, por exemplo), não necessáriamente correspondem aos padrões aqui colocados.

Por exemplo, a taxa de mensagens será sensívelmente mais baixa, devido ao "processamento" e tomada de decisão do operador diante de tal quantidade de informações, o qual, certamente, admitirá do sistema um maior tempo de resposta.

Tal situação pode ser analisada, por exemplo, através de um gráfico do tipo do GRÁFICO 5.3, levantado com valores adequados de λ e interpretado segundo outros critérios para a variável $\bar{t}_{sc}$.

4 - O GRÁFICO 5.4 apresenta a influência do comprimento médio de mensagens de entrada, $\overline{nt}_2$, no desempenho do sistema, considerando operação a 4 fios.

Observa-se que operação a 2400 bps é praticamente inviável neste ambiente descrito pela Figura V.3, e que a sensibilidade da variável $\bar{t}_{sc}$ com relação a $\overline{nt}_2$, a 4800 e 9600 bps é relativamente baixa, apresentando comportamento bastante similar em ambas as velocidades, em uma faixa onde $\overline{nt}_2$ atinge 250 caracteres.

É interessante notar um impacto inicial que a comparação dos GRÁFICOS 5.3 e 5.4 pode provocar, quanto à influência das mensagens de saída e entrada no desempenho. Deve-se atentar, entretanto, para o fato de que a faixa de variação de $\overline{nt}_1$ considerada é bastante mais extensa que a de $\overline{nt}_2$, além de nt_1 e nt_2 apresentarem valores típicos já diferentes. Tais fatos, considerados em conjunto com a forma geral da curva de resposta de sistemas de filas, podem dissipar esta impressão aparentemente estranha.

Adicionalmente, há que se considerar a maneira "assimétrica" como é realizado o atendimento das filas de entrada e saída, comparando-se os tempos envolvidos em "polling" e "addressing", no caso de 4 fios (ver Figura V.3 , componentes $\overline{t}_{1R}$, $\overline{t}_{2R}$, $\overline{r}_{12}$ e $\overline{r}_{21}$).

5 - Com relação ao GRÁFICO 5.5, pode-se dizer que, com exceção da operação a 2400 bps (praticamente excluída em termos de valores absolutos para a variável $\overline{t}_{sc}$), o desempenho de um sistema nas condições descritas pela Figura V.3, a 4 fios, é relativamente "resistente" ao parâmetro $\overline{r}_{12}$, ou seja, em termos práticos, à freqüência de "polling" (incluindo resposta negativa ao mesmo, conforme Seção V.2).

O valor limite considerado neste gráfico foi determinado a partir de medições e de alguma extrapolação sobre o que seria um comportamento admissível para uma unidade de controle de comunicações, implicando no valor $\overline{r}_{12} \cong 500$ ms, o que corresponde a 2 "polling" por segundo.

Qualquer valor abaixo deste pode, entretanto, ser considerado.

Pode-se imaginar (e, mesmo, verificar, traçando-se o gráfico adequado) o efeito da operação a 2 fios, com tempos de reversão incluidos em $\bar{r}_{12}$, quando este parâmetro poderia atingir valores de 1,5 segundos.

6 - Embora a medida de desempenho adotada neste estudo, seja uma estatística de 1ª ordem ($\bar{t}_{sc}$), é interessante analisar a influência de momentos de ordem superior (no caso, 2ª ordem) de parâmetros de operação, neste desempenho.

O modelo adotado incorpora características estatísticas de 2ª ordem das variáveis nt_1 e nt_2, através dos momentos de segunda ordem dos tempos de serviço nas filas 1 e 2 (ver Seção V.2).

No GRÁFICO 5.6 é mostrado o comportamento da variável $\bar{t}_{sc}$ em função da variância do comprimento das mensagens de saída, var(nt_1), onde se observa que a 4800 e 9600 bps a sensibilidade de $\bar{t}_{sc}$ a var(nt_1) é bastante reduzida, em termos práticos, quase nula, em uma faixa bastante extensa.

A 2400 bps, a despeito dos valores absolutos não admissíveis de $\bar{t}_{sc}$, verifica-se uma variação aproximadamente linear no intervalo considerado, com um coeficiente de variação já possível de consideração, caso se pretendesse operar em faixa de desempenho ligeiramente aci-

ma da especificada, por exemplo, 4 a 5 segundos de tempo de resposta nos terminais. Nesta situação, valores de $var(nt_1)$ acima de 20000 (bastante comuns em ambientes onde coexistem respostas curtas e longas) poderiam comprometer o desempenho desejado para o sistema.

As observações 1 a 6 apresentadas fornecem algumas indicações quantitativas a respeito da influência dos diversos parâmetros no desempenho dos sistemas em estudo, bem como elementos que conduzem à construção de uma base qualitativa de conhecimentos e sensibilidade relativos ao assunto objeto desta dissertação. Pretende-se também, e, principalmente, que o exercício realizado nesta Seção V.3 possa transmitir, a interessados no assunto, a metodologia adotada (e proposta como contribuição deste trabalho) na sua realização.

SEÇÃO V.4

VALIDAÇÃO DO MODELO

Conforme já indicado em diversos pontos ao longo deste trabalho, o laboratório de estudos que o viabilizou completamente, e ampliou a motivação inicial do autor no sentido da obtenção de respostas (entendimento, análises de sensibilidade, predições qualitativas e quantitativas) sobre um sistema real em operação, consiste no Sistema de Agências On-Line, desenvolvido pela Itaú Tecnologia S/A.

Em particular, são objeto de estudos diversas linhas de comunicação urbanas e interurbanas, que interligam concentradores de terminais a uma unidade de controle de comunicações 3705-IBM.

O relacionamento deste estudo com este ambiente pode ser caracterizado pelos seguintes aspectos:

a) o autor é funcionário de empresa nacional (por ocasião da realização de parte expressiva deste trabalho) diretamente envolvida com a análise, predição e melhoria do desempenho do referido sistema.

b) o levantamento inicial sobre o funcionamento de tais sub-sistemas, protocolo BSC, mecânica de acesso à linha, etc., realizado por ocasião do início deste trabalho, foi bastante aprimorado e enriquecido, na direção de maior fidelidade ao funcionamento real, a partir de observações e análises feitas "ao vivo" e/ou através de documentação detalhada específica, gerada na Itautec, ou de fornecedores (IBM).

c) a coleta de dados a respeito dos parâmetros considerados (intervalos, valores típicos, estimadores de 1ª e 2ª ordens), foi realizada através da utilização de diversos sistemas de suporte à avaliação disponíveis, tais como, monitores de medição do sistema operacional e programas de "tracing", módulos de geração de dados estatísticos do software de suporte a redes e a dados, bem como de sistemas de análise estatística, os quais fornecem os dados já tratados/manipulados

estatisticamente, como por exemplo, número de transações/tempo, comprimentos médios e variâncias de mensagens de entrada e saída, etc., permitindo ainda a avaliação dos tempos e/ou atrasos de interesse (p. ex., $\bar{t}_{sc}$), sobre intervalos de tempo considerados, para fins de comparação com resultados teóricos obtidos a partir de parâmetros de operação vigentes nestes intervalos.

d) comparações de resultados medidos e processados pelo(s) sistemas(s) de estatística com resultados calculados para a variável $\bar{t}_{sc}$, a partir dos parâmetros correspondentes, indicaram um desvio máximo de aproximadamente 15 a 20% com valores teóricos inferiores aos valores medidos; os desvios típicos constatados, em grande parte das experiências, situaram-se em torno de 5 a 10% para baixo, em faixas de utilização inferiores a 30%.

Considerando-se o estado atual do desenvolvimento de teoria de filas, a nível internacional, e a situação de empresários, engenheiros, analistas (brasileiros, em particular; talvez também no exterior), na tomada de decisões relativas ao binômio custo-desempenho de sistemas desta natureza, julgamos bastante oportunos os resultados obtidos, ou seja, um modelo analítico, de simples utilização e confiabilidade relativamente alta.

A manipulação deste modelo pode ser tornada extremamente eficiente, o que foi realizado durante este estudo, através da implementação de um programa computacional que:

a) realize um pré-processamento no conjunto dos parâmetros primitivos (eventualmente, se não disponível programa especializado, mesmo o cálculo de estimadores estatísticos, a partir de valores medidos), de forma a gerar os parâmetros de entrada para o modelo;

b) alimente o modelo com os dados obtidos na fase anterior e forneça o resultado final ($\bar{t}_{sc}$), bem como resultados intermediários de interesse (tais como, $\bar{D}_i$, $\bar{t}_i$, ρ_i, i=1,2), juntamente com uma listagem dos parâmetros primitivos correspondentes, para facilitar trabalhos de análise e levantamento de gráficos; emissão automática de resultados sob forma gráfica pode também ser considerada (não o foi neste trabalho).

O programa utilizado foi implementado em linguagem APL, mostrando-se de utilização bastante simples, principalmente no que diz respeito à execução com vetores de parâmetros, para obtenção de pontos de um gráfico, dadas as características convenientes desta linguagem em termos de manipulação de arranjos (vetores, matrizes, etc.) e avaliação de expressões correspondentes.

Com relação às aproximações obtidas, mencionadas no item d desta Seção V.4, cabem ainda as seguintes observações:

a) os valores teóricos aproximados fornecidos pelo

modelo de Sykes, devem, por definição (fundamentação no princípio da independência de Leibowitz) situar-se abaixo daqueles fornecidos por um tratamento exato do mesmo modelo;

b) comparação de valores teóricos em alguns casos calculados, entre resultados de Sykes (aproximado, tempo contínuo), Eisenberg (exato, tempo contínuo) e Konheim e Meister (exato, tempo discreto) indicam a ordenação crescente esperada entre estes valores, no caso de duas filas simétricas;

c) a premissa de sistema aberto, em ambas as filas, verifica-se, na prática com maior ou menor aproximação, em função do número de terminais conectados à unidade de controle de terminais ; se este número for pequeno, e a taxa λ=constante não se verificar, independentemente do estado da(s) fila(s), neste caso, decrescendo, os valores teóricos (qualquer que seja o modelo aberto) devem ser superiores aos observados;

d) entende-se que uma série de outros pequenos fatores, da natureza dos descritos acima, podem violar determinadas premissas do modelo, implicando em resultados teóricos tanto acima como a baixo dos observados; alie-se a tal fato as imperfeições de mapeamento sistema físico x mode-

lo, conforme discutidas nas Seções V.1 e V.2.

Concluindo, pode-se conceber diversos artifícios que, em conjunto com um conhecimento bem mais detalhado da dinâmica e de certas propriedades de ambos modelo e sistema físico, conduzam a correções e/ou compensações que resultem em um melhor ajuste (empírico) do modelo, medido por uma maior aderência entre valores teóricos e observados.

Entende-se, entretanto, que os seguintes fatores devem ser levados, neste ponto, em consideração:

- em muitos dos casos, trata-se de aproximações da ordem de, digamos, 10%, ou mesmo 20%; são encontradas na literatura, mesmo, desvios maiores que 20%, e um fato bastante importante é que estes desvios sejam conhecidos e considerados, em qualquer análise.

- o exposto acima, alia-se ao fato de que decisões relevantes no projeto (ou alterações) de sistemas desta natureza, por exemplo, visando redução de custos, são tomadas, não com base em alguns décimos de segundos (ou frações do desempenho desejado), mas, sim, com maiores margens, procurando , na maioria dos casos, atingir operação segura em faixas de desempenho de baixa sensibilidade aos parâmetros de tráfego e controle de linha,com relativa distância de regiões críticas de saturação.

Estas considerações reforçam e concluem os comentários anteriormente feitos ao longo desta Seção V.4, quanto à validação, confiabilidade e oportunidade e forma de utilização deste, ou de modelos desta natureza.

Dentro deste contexto pode ser perfeitamente assimilado e utilizado um modelo otimista conforme o proposto, embora a disponibilidade de modelos pessimistas possa ser, em princípio, mais tranquilizadora, na presença de imprecisões.

Quando, como e quão taxativo ser com relação a resultados de um modelo de aproximação conhecida, depende em muito, da situação e da sensibilidade e experiência profissionais do engenheiro/analista que detém o conhecimento dos princípios da utilização, e da interpretação do modelo.

Deve-se manter também sempre em mente que soluções técnicas isoladas não constituem soluções de grande parte dos problemas de empresas e indústrias,mas, sim, subsídio para, ou componentes de uma decisão global vetorial, resultante de componentes de natureza política, econômica, social, organizacional, mercadológica e técnica, entre outras.

CAPÍTULO VI

RESULTADOS E EXTENSÕES

Sintetizando diversos comentários, observações e conclusões apresentados ao longo do conteúdo desta dissertação, e ao mesmo tempo, procurando analisá-la em perspectiva, no espaço acadêmico-profissional em que teve seu desenvolvimento, são apresentados, neste Capítulo VI:

a) Os resultados atingidos, em termos de contribuição efetiva aos segmentos acadêmico e profissional, bem como ao enriquecimento de conhecimentos e experiência globais do autor e

b) Extensões possíveis sugeridas pelo autor, que possam vir a acrescentar contribuições à matéria, tanto a nível acadêmico (direções e temas para estudos e pesquisas posteriores, na área de proposição e solução de modelos de filas correlatos) como a nível profissional (utilização de modelos similares, realização de levantamentos/medições, considerações de modelos e condições de operação menos ideais e exercício sistemático desta metodologia de análise de desempenho).

SEÇÃO VI.1

RESULTADOS

O Capítulo V desta dissertação fornece resultados

quantitativos e qualitativos sobre o desempenho de uma classe de sub-sistemas de comunicações de dados (comunicação de dados HDX com "polling"/BSC), encontrada com bastante frequência, e em grande número, em redes de teleprocessamento, ou de comunicação de computadores, de diversos portes, em operação no Brasil e no exterior.

Conforme introduzido no Capítulo I, um esforço bastante expressivo de pesquisa bibliográfica, subsidiado por contatos pessoais, indica a inexistência de um modelo analítico fechado, de utilização simples e rápida e confiabilidade relativamente alta, conforme o proposto neste trabalho, para solução do problema de análise de desempenho destes sub-sistemas, baseado na utilização de um modelo de multifilas analisado por Sykes, dos Laboratórios Bell, no início da década de 1970 /SYKEJ70/.

A utilização deste modelo é conceituada, ilustrada e validada através de experiências e medições realizadas, pelo autor, em diversas linhas de comunicação de um sistema on-line em operação . Ao longo do trabalho é, entretanto, recomendada a utilização do modelo em ambientes diversos de operação, como exercícios adicionais de validação do modelo e da metodologia proposta.

É também fornecida uma orientação, conforme discernimento do autor, quanto à utilização deste modelo, ou de modelos similares, em ambientes profissionais, no que toca ao papel de resultados teóricos estimadas no processo de tomada de decisões.

Os Capítulos II e III apresentam, respectivamente, uma conceituação clara e abrangente dos modelos de multifilas (com respectiva revisão de literatura e referências adicionais) e uma revisão orientada de tópicos de matemática aplicada, relevantes no entendimento, utilização e desenvolvimento de modelos de multifilas.

O Capítulo IV revê os desenvolvimentos e resultados mais relevantes em multifilas, do ponto de vista deste trabalho.

Pretende-se que este material atue em uma região "nebulosa", que separa as áreas acadêmica e profissional, ou como querem alguns, teórica e prática, como elemento catalisador e de ligação, no sentido de levar a pesquisadores interessados em análise matemática de multifilas, algumas motivações/orientações objetivas (tais como alguns problemas específicos citados ao longo dêstes capítulos anteriores), quando não, o assunto multifilas em si, não muito explorado e bastante disperso na literatura, e ao mesmo tempo, de transmitir, através de uma linguagem relativamente acessível, a profissionais interessados da área de computadores e comunicações, a potência e os benefícios da utilização de tais técnicas/modelos em seu cotidiano.

Resumindo, pretende-se estar atingindo resultados que promovam uma interação tal entre os "mundos" acadêmico e profissional, que conduza a uma minimização, sabidamente lenta e gradativa, do dipolo, grave distorção, existente entre estes dois "mundos": pesquisadores, professores e alunos dedicando es

esforços e capacidades à solução de problemas desvinculados de qualquer compromisso com a realidade, particularmente a nacional, sem qualquer chance de conversão em resultados concretos, mesmo a longo prazo, e, de outro lado, profissionais atuando na indústria, tomando decisões ou estimando situações de maneira totalmente empírica ou ad hoc, de eficácia geralmente limitada, em ambientes bastante complexos, típicos de processamento/comunicação de dados, sem lançar mão de quaisquer recursos formais ou genéricos, colocados à sua disposição por alguns setores da comunidade acadêmica.

A modificação desta situação certamente requer esforços decisivos de ambas as partes.

A realização desta dissertação desenvolveu-se a partir da motivação para solução de um problema real, cursou uma fase de orientação predominantemente didática/acadêmica, caracterizada por estudos de análise x simulação, concluindo com uma fase de estudos análise x sistema físico, na medida em que o autor teve à sua disposição, um "laboratório de trabalho", no contexto de um sistema real em operação.

Embora apenas parcialmente relatado neste estudo, afirma-se ter sido bastante amplo (quantitativa e qualitativamente) o material pesquisado (e assimilado) pelo autor neste processo, particularmente em áreas não explícitas neste estudo, como, modelagem, simulação (implementação de sistemas de multifilas, calibração e validação, análise estatística de resultados), análise operacional, tópicos avançados de processos estocásticos

e descrição de protocolos de comunicação, além de um levantamento sobre redes em anel ("loops").

Conta-se com um total de aproximadamente 200 referências adicionais, relativas aos tópicos acima, as quais são colocadas à disposição pelo autor, para cópias, anotações ou mesmo, discussões pessoais a respeito.

SEÇÃO VI.2

EXTENSÕES

Como extensões deste trabalho, podem ser sugeridos, ainda em termos de aplicação da classe de modelos de multifilas considerada, os seguintes estudos/pesquisas:

a) A utilização de um modelo de multifilas com 2 filas, do tipo analisado por Sykes ou Eisenberg, na análise de desempenho de uma comunicação HDX com disciplina de "contention" (ver/DOLLD77/), nos moldes da proposta realizada neste trabalho para "polling"; o mesmo pode ser investigado com relação ao protocolo SDLC, em ambiente 3270 IBM, uma vez que a comunicação, em termos de mensagens de texto, é realizada em modo HDX.

b) A utilização do modelo de N filas assimétrico, a tempo discreto, analisado por Swartz em sua tese de doutorado /SWARG77/, na análise de desempenho de linhas multiponto, onde ter-se-ia uma fila de saída da unidade de controle de

comunicações e (N-1) filas de entrada de unidades de controle de terminais (em função da disciplina de serviço da unidade de controle de comunicações, ter-se-iam, eventualmente, N/2 filas de saída e N/2 filas de entrada).

c) Aplicações de modelos de multifilas na avaliação de desempenho de sistemas de microprocessadores com arquitetura multimicro (barramento, anel, etc.)

d) Aplicações de modelos de multifilas na avaliação e comparação de desempenho de esquemas de controle de acesso em redes locais, do tipo "token ring", "slotted ring", CSMA/CD, "ordered-bus MLMA" e outros, conforme recente trabalho de W.Bux , da IBM, /BUXW81/, utilizando o modelo de Konheim e Meister, descrito em /KONHA74/.

e) Aplicações de modelos de multifilas do tipo, por exemplo, analisado em /KONHA74/, /COOPR69,70/ e /SWARG77/, em sistemas de controle em tempo real, que envolvam processos de "scanning" ou amostragem de dispositivos de entrada/saída de dados; considerar também sistemas por interrupção.

f) Aplicações da sub-classe de modelos de multifilas estudada especialmente por Kühn, da Universidade de Siegen, Alemanha, que se dedica, há alguns anos ao estudo de modelos de multifilas com serviço limitado (L^1), no contexto de sistemas de comutação de dados, utilizando-se de métodos de decomposição por ele desenvolvidos para análise de estruturas hierárquicas nestes sistemas (ver /KUEHP79a,b,c/ e /KUEHP80a,b/).

g) Pesquisa relativa à aplicação de modelos hierárquicos com serviço parcialmente compartilhado, em redes multiponto multiníveis, modelos estes estudados e resolvidos por Foster e Perros /FOSTF79/ no contexto de análise de desempenho de sistemas hierárquicos de I/O.

h) Aplicações do modelo proposto por Chang, da IBM, em /CHANJ 75/, visando a validação do mesmo em ambientes reais de operação; Chang considera comunicação BSC e SDLC, atrasos na TCU (3705) e também o desempenho do computador "host".

i) Exploração do tema simulação discreta, com ênfase em sistemas de filas, nos aspectos estatísticos de entrada, dinâmica de modelo e saída, e na construção, calibração e validação de modelos de simulação computacionais; bastante úteis no caso de sistemas mais complexos, para os quais não existam soluções analíticas, ou, mesmo, em sistemas simples, onde se exijam resultados e respostas mais detalhados que aqueles fornecidos por modelos em regime estacionário, livres de erros, geralmente sob a forma de apenas valores médios (referências iniciais: /FISHG78/, /JACOS80/ e /KLEIJ74, 75/; ver referências secundárias e também /KLEIJ79/).

j) Estudos e pesquisas generalizados em aplicações de modelos analíticos de filas, incluíndo redes de filas, na análise de desempenho de sub-sistemas de comunicações e de computação; esta recomendação é baseada em resultados positivos adicionais que vem sendo obtidos, pelo autor, em atividades de avaliação de desempenho (particularmente, sistemas de

computação) relacionadas com seu ambiente profissional na Itaú Tecnologia S/A (referências iniciais: /KLEIL76/ , /REISM82/, /KOBAH78/ e /SAUEC81/; ver referências secundárias).

No que toca ao desenvolvimento de extensões na área de análise matemática, podem ser mencionados como desejáveis, modelos do tipo multifilas, ou mesmo outros, que:

a) incorporem resultados de variáveis aleatórias de desempenho sob a forma de distribuições, que permitam uma caracterização mais completa, e a obtenção de percentiles das mesmas; reconhecida a dificuldade de obtenção de distribuições, considerar como alternativa momentos de 2ª e 3ª ordem, que permitam, através de ajustes empíricos aproximados, a obtenção dos percentiles acima mencionados.

b) reflitam o efeito de erros nas linhas (probabilidade de retransmissão de uma mensagem, eventualmente, uma função do comprimento das mensagens!); em /KUEHP80b/, Kühn analisa um esquema de retransmissão de pacotes de tamanho fixo, com probabilidade p, em multifilas com serviço limitado (L^1) ; faz-se necessária a consideração de retransmissões em multifilas com serviço exaustivo, por exemplo, uma adaptação do próprio modelo de Sykes, ou outro modelo com 2 filas (ver /KONHA80/ e /TOWSD79/ como base).

c) permitam uma abordagem de síntese de sistemas, no sentido de encontrar parâmetros ou políticas de atendimento, prio-

ridades, etc, que satisfaçam a requisitos de desempenho pré-definidos para o mesmo, conforme colocado em /COFFE80/, onde Coffman e Mitrani enfatizam e resolvem um problema relacionado com síntese a partir de requisitos realizáveis por sistemas de filas simples.

d) considerem filas finitas, sua ocupação, "overflow", etc, fato de grande interesse prático, uma vez que em sistemas de computação / comunicação reais não existem filas infinitas; ver estudos e propostas iniciais de Arthurs e Stuck, da Bell, em /ARTHE79/.

e) considerem processos de chegada e de serviço heterogêneos, no sentido de viabilizar análises ao longo de quaisquer intervalos, e não apenas como típico, em intervalos de pico, em regime estacionário, onde as taxas de chegada, por exemplo, são consideradas homogêneas no tempo; nesta área cita-se, por exemplo, como referência inicial, /NIUSC80/, que analisa sistemas de filas simples com processos de chegada heterogêneos, do tipo Poisson não-estacionário, com uma função de intensidade cuja evolução é regida por uma cadeia de Markov com dois estados a tempo contínuo.

f) incorporem mecanismos e dinâmicas bastante comuns no ambiente de tráfego de dados, tais como, temporizações ("time-outs") aplicadas sobre atrasos limites (ver /HAUGR80/ e /WALLB76/) e entradas correlacionadas (não independentes , não em lotes), sob a forma de sequências estocásticas finitas de chegadas, conforme formuladas e analisadas por Gopi

nath e Morrison, dos Laboratórios Bell, em /GOPIB77/; em /FRASA78/, é estudada por Fraser, também da Bell, uma aplicação dos resultados de Gopinath e Morrison ao problema de análise de "buffers" de um concentrador de terminais lentos, onde as mensagens são compostas de pacotes correlacionados, separados por intervalos de tempo fixos.

g) permitam o tratamento de processos de entrada do tipo irregular, ou em "rajadas" ("bursty"), conforme estudo de caracterização do tráfego de dados, de Lam /LAMSS78/; neste sentido, Gopinath, Mitra e Sondhi, da Bell, apresentam em /GOPIB73/ fórmulas para filas com este tipo de processo de entrada.

h) sugere-se ainda como tema de crescente interesse na comunidade ligada à teoria de filas, a utilização de "Petri-Nets" numéricas, na descrição e definição de sistemas de filas ; se tal técnica pode ser utilizada, com benefícios, na classe de multifilas constitui ainda uma pergunta (ver/SYMOF80/).

Conforme apresentado acima, durante o processo de pesquisa realizado para o desenvolvimento e elaboração desta dissertação, foi despertado, no autor, o interesse por alguns temas possíveis de investigação futura, apresentados nesta Seção VI.2 como extensões do estudo realizado.

Alguns destes estudos apresentam relações mais estreitas e imediatas com a classe de problemas objeto desta dissertação, como, por exemplo, as extensões para "contention", SDLC,

multiponto e sistemas multimicro, bem como a investigação detalhada e eventual validação do modelo de Chang, já considerados pelo autor para abordagem imediata.

A disponibilidade de resultados sob a forma de distribuições, ou quaisquer outras formas que possibilitem a obtenção de percentiles é altamente desejável, assim como também a introdução de erros de comunicação (retransmissões) em modelos de multifilas exaustivos com 2 filas e tempos de transição finitos.

Uma intenção adicional do autor consiste em estudar a introdução, em problemas de determinação de topologia ótima de redes de distribuição local (localização de terminais, concentradores e/ou multiplexadores), de restrições de atraso (ou função-objetivo) conforme aquelas obtidas nesta tese, quando aplicáveis.

Os outros temas colocados devem ser avaliados com bastante critério, para o que são recomendados, de qualquer forma, estudos/pesquisas exploratórios.

Acrescenta-se ainda que diversos temas para estudos e investigações futuras, relativos àquelas áreas de pesquisa mencionadas na final da Seção VI.1 podem ser identificados na literatura correspondente disponível, ou sugeridas, pessoalmente, pelo próprio autor deste trabalho, com base nas pesquisas e experiências realizadas e nas dificuldades encontradas.

Concluindo, espera-se que as extensões aqui sugeri

das, sirvam ao propósito de reunir esforços, habilidades, experiências e dedicação de elementos das comunidades acadêmicas e científico-industrial nacionais, na obtenção de soluções simples, adequadas e eficientes para problemas de interesse da comunidade de teleinformática.

REFERÊNCIAS BIBLIOGRÁFICAS

ARTHE79 ARTHURS, E., B.W.Stuck
Analysis of a single-buffer loss-delay system
Operations Research, 27, 1979, pp.65-79

AVIIB63 AVI-ITZHAK, B., P.Naor
Some queueing problems with the service station subject to breakdown
Operations Research, 11, 1963, pp.303-320

AVIIB65 AVI-ITZHAK, B., W.L. Maxwell, L.W. Miller
Queueing with alternating priorities
Operations Research, 13, 1965, pp.306-318

BHATU72 BHAT, U.N.
Elements of applied stochastic processes
John Wiley & Sons, Inc., New York, 1972

BJØRD70 BJØRNER, D.
Finite state automation - definition of data communi cation line control procedures
Fall Joint Comput. Conf., 1970, pp.477-491

BUXWE81 BUX, W.
Local-area subnetworks: a performance comparison
IEEE Trans. Comm., Vol. COM-29, Nº.10, October 1981 pp.1465-1473

CHANJ75 CHANG, J.H.
Terminal response times in data communications systems
IBM J. Res. Develop., Vol.19, Nº3, 1975, pp.272-282

CHUWW72 CHU, W.W., A.G. Konheim
On the analysis and modeling of a class of computer communication systems
IEEE Trans. Comm., Vol.COM-20, Nº3, June 1972, pp.645-660

COBHA54 COBHAM, A.
Priority assignment in waiting line problems
Operations Research, 8, 1954, pp.70-76

COFFE80 COFFMAN Jr., E.G., I. Mitrani
A characterization of waiting time performance realizable by single-server queues
Operations Research, 28, 1980, pp.810-821

COOPR69 COOPER, R.B., G. Murray
Queues served in cyclic order
Bell Syst. Tech. J., Vol.48, Nº3, March 1969, pp.675-689

COOPR70 COOPER, R.B.
Queues served in cyclic order: waiting times
Bell Syst. Tech. J., Vol.49, Nº3, March 1970, pp.399-413

COXDR61 COX, D.R., W.L. Smith
Queues
Chapman and Hall, London, 1961

DARRJ64 DARROCH, J.N., G.F. Newell, R.W.J. Morris
Queues for a vehicle-actuated traffic light
Operations Research, 12, 1964, pp.882-895

DENNP78 DENNING, P.J., J.P. Buzen
The operational analysis of queueing network models
ACM Comput. Surveys, Vol.10, Nº3, September 1978,
pp.225-261

DOLLD77 DOLL, D.
Data communications - facilities, networks and
systems design
John Wiley & Sons, Inc., New York, 1977

DUDIA71 DUDICK, A.L., E. Fuchs, P.E. Jackson
Data traffic measurements for inquiry-response
computer communication systems
Proc. IFIP - Information Processing 71, Ljubljana,
August 1971, North-Holland Publ. Co., pp.634-641

EISEM67 EISENBERG, M.
Multi-queues with changeover times
MIT-Department of Electrical Engineering, Ph.D.,1967

EISEM68 EISENBERG, M.
Multi-queues with changeover times
Tech. Report N°35, Operations Research Center, MIT, 1968

EISEM71 EISENBERG, M.
Two queues with changeover times
Operations Research, 19, 1971, pp.386-401

EISEM72 EISENBERG, M.
Queues with periodic service and changeover time
Operations Research, 20, 1972, pp.440-451

EISEM79 EISENBERG, M.
Two queues with alternating service
SIAM J. Appl. Math., Vol.36, N°2, April 1979, pp.287-303

EVERW72 EVERLING, W.
Exercises in computer systems analysis
Springer-Verlag, Berlin-Heidelberg, 1972

FISHG78 FISHMAN, G.S.
Principles of discrete event simulation
John Wiley & Sons, Inc., New York, 1978

FOSTF79 FOSTER, F.G., H.G. Perros
Hierarchical queue networks with partially shared service
J. Opl. Res. Soc., Vol.30, N°2, 1979, pp.157-166

FRASA78 FRASER, A.G., B. Gopinath, J.A. Morrison
Buffering of slow terminals
Bell Syst. Tech. J., Vol.57, N°8, October 1978,
pp.2865-2885

FUCHE70 FUCHS, E., P.E. Jackson
Estimates of distributions of random variables for
certain computer-communication traffic models
Comm. ACM, Vol.13, N°12, December 1970, pp.752-757

GELLK82 GELL, K.
Improving SNA response time
Computerworld, Vol.XVI, N°21, May 1982, In depth 52-53

GIFFW75 GIFFIN, W.C.
Transform techniques for probability modeling
Academic Press, Inc., New York, 1975

GOPIB73 GOPINATH; B., D. Mitra, M.M. Sondhi
Formulas on queues in burst processes-I
Bell Syst. Tech. J., Vol.52, N°1, January 1973, pp.9-33

GOPIB77 GOPINATH, B., J.A. Morrison
Discrete-time single server queues with
correlated inputs
Bell Syst. Tech. J.. Vol.56, N°9, November 1977,
pp.1743-1768

GROSD74 GROSS, D., C.M. Harris
Fundamentals of queueing theory
John Wiley & Sons, Inc., New York, 1974

HALFS75 HALFIN, S.
An approximate method for calculating delays for a family of cyclic-type queues
Bell Syst. Tech. J., Vol.54, Nº10, December 1975, pp.1733-1753

HASHO70 HASHIDA, O.
Gating multiqueues served in cyclic order
Syst. Comput. Contr., Vol.1, Nº1, 1970, pp.1-8

HASHO72a HASHIDA, O.
Analysis of multiqueue
Rev. Electr. Comm. Lab. NTT, Vol.20, Nº3-4, March--April 1972, pp.189-199

HASHO72b HASHIDA, O., K. Ohara
Line accommodation capacity of a communication control unit
Rev. Electr. Comm. Lab. NTT, Vol.20, Nº3-4, March--April 1972, pp.231-239

HAUGR80 HAUGEN, R.B., E. Skogan
Queueing systems with stochastic time out
IEEE Trans. Comm., Vol. COM-28, Nº12, December 1980, pp.1984-1989

HAYEJ72 HAYES, J.F., D.N.Sherman
A study of data multiplexing techniques and delay performance
Bell Syst. Tech. J., Vol.51, Nº9, November 1972, pp.1983-2011

HAYEJ78 HAYES, J.F.
An adaptive technique for local distribution
IEEE Trans. Comm., Vol.COM-26, Nº8, August 1978, pp.1178-1186

HOUST79 HOUSLEY, T.
Data communications and teleprocessing systems
Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1979

IBM/GA27-2749-10/
IBM 3270 information display system-component description
Publicação técnica, distribuída pela IBM

IBM/GA27-3004-02/
General information-binary synchronous communications
Publicação técnica, distribuída pela IBM

IBM/GF20-0007-01/
Analysis of some queueing models in real-time systems
Publicação técnica, distribuída pela IBM

JACKP69 JACKSON, P.E., C.D. Stubbs
A study of multiaccess computer communications
Spring Joint Comput. Conf., 1969, pp.491-504

JACOS80 JACOBY, S.L.S., J.S. Kowalik
Mathematical modeling with computers
Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980

KAYEA73 KAYE, A.R., T.G. Richardson
A performance criterion and traffic analysis for polling systems
INFOR, Vol.11, N°2, June 1973, pp.93-112

KEMEJ76a KEMENY, J.G., J.L. Snell
Finite Markov chains
Springer-Verlag, New York, 1976

KEMEJ76b KEMENY, J.G., J.L. Snell, A.W. Knapp
Denumerable Markov chains
Springer-Verlag, New York, 1976

KENDD51 KENDALL, D.G.
Some problems in the theory of queues
J. Royal Stat. Soc., Ser.B, 13, 1951, pp.151-185

KLEIJ74,75 KLEIJNEN, J.P.C.
Statistical techniques in simulation-PartI/PartII
Marcel Dekker, Inc., New York, 1974/1975

KLEIJ79 KLEIJNEN, J.P.C.,A.J. van der Burg, R.Th. van der Ham
Generalization of simulation results - practicality of statistical methods
Europ. J. Opl. Res., Vol.3, Nº1, January 1979, pp.50-64

KLEIL75 KLEINROCK, L.
Queueing systems - volume I
John Wiley & Sons, Inc., New York, 1975

KLEIL76 KLEINROCK, L.
Queueing systems - volume II
John Wiley & Sons, Inc., New York, 1976

KOBAH77 KOBAYASHI, H., A.G. Konheim
Queueing models for computer communications system analysis
IEEE Trans. Comm., Vol.COM-25, Nº1, January 1977, pp.2-28

KOBAH78 KOBAYASHI, H.
Modeling and analysis: an introduction to system performance evaluation methodology
Addison Wesley Publ. Co., Inc., California, 1978

KONHA71 KONHEIM, A.G., B. Meister
Polling in a multidrop communication system: waiting line analysis
2nd ACM/IEEE Symp.Probl. Optim. Data Comm. Syst., Palo Alto, October 1971, pp.124-129

KONHA72 KONHEIM, A.G., B. Meister
Service in a loop system
Journal ACM, Vol.19, Nº1, January 1972, pp.92-108

KONHA74 KONHEIM, A.G., B.Meister
Waiting lines and times in a system with polling
Journal ACM, Vol.21, Nº3, July 1974, pp.470-490

KONHA80 KONHEIM, A.G.
A queueing analysis of two ARQ protocols
IEEE Trans. Comm., Vol.COM-28, Nº7, July 1980,
pp.1004-1014

KUEHP79a KUEHN, P.J.
Approximate analysis of general queueing networks
by decomposition
IEEE Trans. Comm., Vol.COM-27, Nº1, January 1979,
pp.113-126

KUEHP79b KUEHN, P.J.
Multiqueue systems with nonexhaustive cyclic service
Bell Syst. Tech. J., Vol.58, Nº3, March 1979, pp. 671-698

KUEHP79c KUEHN, P.J.
Analysis of switching system control structures
by decomposition
9. Int. Teletraffic Congress, Torremolinos, October
1979, Congr. Paper#514

KUEHP80a KUEHN, P.J.
Comunicação pessoal
Siegen-Rio de Janeiro, Agosto 1980

KUEHP80b KUEHN, P.J.
Performance of ARQ-protocols for HDX-transmission in hierarchical polling systems
5. Int. Conf. Comput. Comm. ICCC-80, Atlanta, October 1980

LAMSS78 LAM, S.S.
A new measure for characterizing data traffic
IEEE Trans. Comm., Vol.COM-26, Nº1, January 1978, pp.137-140

LEIBM61 LEIBOWITZ, M.A.
An approximate method for treating a class of multiqueue problems
IBM J. Res. Develop., Vol.5, Nº3, 1961, pp.204-209

LEIBM62 LEIBOWITZ, M.A.
A note on some fundamental parameters of multiqueue systems
IBM J. Res. Develop., Vol.6, Nº4, 1962, pp.470-471

LEIBM68 LEIBOWITZ, M.A.
Queues
Scientific American, 219, August 1968, pp.96-103

MACKC57a MACK, C., T. Murphy, N.L. Webb
The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants
J. Royal Stat. Soc., Ser. B, 19, 1957, pp.166-172

MACKC57b MACK, C.
The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable
J. Royal Stat. Soc., Ser. B, 19, 1957, pp.173-178

MARTJ72 MARTIN, J.
Systems analysis for data transmission
Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1972

MORAL81 MORAES, L.F.M. de
Message queueing delays in polling schemes with applications to data communication networks
UCLA - Department of System Science, Ph.D., 1981
(ref. UCLA-ENG-8106)

NIUSC80 NIU, S.C.
A single server queueing loss model with heterogeneous arrival and service
Operations Research, 28, 1980, pp.584-593

ONEIP80 O'NEILL, P., A. O'Neill
Performance statistics of a time sharing network at a small university
Comm. ACM, Vol.23, N°1, January 1980, pp.10-13

PALMC43 PALM, C.
Intensitätschwankungen im Fernsprechverkehr
Ericsson Technics, N°6, 1943, pp.1-189

PARZE62 PARZEN, E.
Stochastic processes
Holden-Day, Inc., San Francisco, 1962

PATEJ76 PATEL, J.K., C.H. Kapadia, D.B. Owen
Handbook of statistical distributions
Marcel Dekker, Inc., New York, 1976

PAWLP81 PAWLITA, P.F.
Traffic measurements in data networks, recent measurement results, and some implications
IEEE Trans. Comm., Vol.COM-29, N°4, April 1981, pp.525-535

REISM82 REISER, M.
Performance evaluation of data communication systems
Proc. IEEE, Vol. 70, N°2, February 1982, pp.171-196

ROSSS72 ROSS, S.M.
Introduction to probability models
Academic Press, Inc., New York, 1972

SAUEC81 SAUER, C.H., K.M.Chandy
Computer systems performance modeling
Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1981

SCHWM77 SCHWARTZ, M.
Computer - communication network design and analysis
Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1977

SINGK78 SINGH, K.B.
A queueing system with random additional service facility
R.A.I.R.O., Vol.12, N°3, Août 1978, pp.311-318

SMITW58 SMITH, W.L.
Renewal theory and its ramifications
J.Royal. Stat. Soc., Vol.XX, N°2, 1958, pp.243-284

STUTB72 STUTZMAN, B.W.
Data communication control procedures
ACM Comput. Surveys, Vol.4, N°4, December 1972,
pp.197-220

SUBRN82 SUBRAMANIAN, N., D.A. Rubin
Polled networks: modeling helps the user
evaluate their performance
Data Communications, June 1982, pp.77-80

SWARG77 SWARTZ, G.B.
Polling with unequal arrival rates
New York University, Ph.D., Mathematics, 1977
(available from University Microfilms International, Ann Arbor, Michigan, USA)

SWARG80 SWARTZ, G.B.
Polling in a loop system
Journal ACM, Vol.27, N°1, January 1980, pp.42-59

SYKEJ69a SYKES, J.S.
Analytical model of half-duplex interconnections of computers
IEEE Trans. Comm. Tech., Vol.COM-17, N°2, April 1969 pp.235-238

SYKEJ69b SYKES, J.S.
Analysis of the communication aspects of an inquiry--response system
Fall Joint Comput. Conf., 1969, pp.655-667

SYKEJ70 SYKES, J.S.
Simplified analysis of an alternating-priority queueing model with setup times
Operations Research, 18, 1970, pp.1182-1192

SYMOF80 SYMONS, F.J.W.
The description and definition of queueing systems by numerical Petri-nets
Austral. Telecomm. Res., Vol.13, N°2, 1980, pp.20-31

TAKAL62 TAKÁCS, L.
Introduction to the theory of queues
Oxford University Press, New York , 1962

TAKAL68 TAKÁCS, L.
Two queues attended by a single server
Operations Research, 16, 1968, pp.639-650

TANNJ53 TANNER, J.C.
A problem of interference between two queues
Biometrika, 40, 1953, pp.58-69

TOWSD79 TOWSLEY, D., J.K. Wolf
On the statistical analysis of queue lengths and quiting time for statistical multiplexers with ARQ retransmission schemes
IEEE Trans. Comm., Vol.COM-27, Nº4, April 1979, pp.693-702

WALLB76 WALLSTRÖM, B.
A queueing system with time-outs and random departures
Ericsson Technics, Nº2, 1977, pp.154-174, também 8. Int. Teletraffic Congress, Melbourne, November 1976

WHITB75 WHITAKER, B.A.
Analysis and optimal design of a multiserver multiqueue system with finite waiting space in each queue
Bell Syst. Tech. J., Vol.54, Nº3, March 1975,pp.595-623