Este exemplar corresponde à redação final da tese defendida por ANTONIO CARLOS LAVELHA e aprovada pela Comissão Julgadora em 31 / 10 / 1991

Orientador

ANÁLISE DE SISTEMAS MULTIFILAS COM MÚLTIPLOS SERVIDORES CÍCLICOS

Antonio Carlos Lavelha

Orientadores: Prof. Dr. Jorge Moreira de Souza
Prof. Dr. João Bosco Ribeiro do Val

Campinas, 31 de outubro de 1991

Dissertação apresentada na Faculdade de Engenharia Elétrica
da Universidade Estadual de Campinas — UNICAMP —,
como requisito parcial para a obtenção
do título de Doutor em Engenharia Elétrica

UNICAMP
BIBLIOTECA CENTRAL

COMISSÃO JULGADORA

Jorge Moreira de Souza, CPqD-TELEBRÁS

João Bosco R. do Val, FEE-UNICAMP

Luis Felipe M. de Moraes, IPRJ

Ivanil S. Bonatti, FEE-UNICAMP

Michel D. Yacoub, FEE-UNICAMP

Shusaburo Motoyama, FEE-UNICAMP

Rafael dos Santos Mendes, FEE-UNICAMP

AGRADECIMENTOS

O autor agradece sinceramente a todos que contribuiram — ainda que involuntariamente — para a elaboração deste trabalho; em particular, ao CPqD-TELEBRÁS, à UNICAMP, aos orientadores e demais membros da comissão julgadora e à equipe do projeto TRÓPICO.

Sumário

Consideramos modelos de sistemas com múltiplos nós servidos ciclicamente por vários servidores idênticos. Em cada nó há uma fila de transmissão e uma fila de recepção. Usuários chegam nas filas de transmissão dos nós de acordo com um processo Poissoniano. Após o atendimento em um nó de origem, um usuário deve ser encaminhado à fila de recepção de um nó de destino. Um nó pode utilizar no máximo um servidor em um dado instante. Os processos de caminhada dos servidores pelos nós e os processos de serviço são supostos gerais. Esses modelos são apropriados para a avaliação de desempenho de uma ampla classe de redes de interligação de processadores. Desenvolvemos um modelo analítico aproximado e um modelo de simulação para a obtenção do tempo médio dos usuários em uma fila de transmissão. O modelo analítico trata sistemas simétricos ou assimétricos com filas de capacidade infinita, servidores operando no modo repetição no caso de bloqueio do nó de destino e com serviço 1-limitado. O enfoque utilizado é o da agregação dos servidores em um único servidor equivalente. O modelo é uma extensão ao caso multi-servidor do modelo de Hashida e Ohara relativo a servidor em férias e serviço não exaustivo. O tempo de ciclo do servidor equivalente é calculado através de dois métodos distintos. No primeiro método, nós utilizamos uma equivalência entre as taxas de serviço e caminhada do servidor equivalente e dos servidores originais. No segundo método, o tempo de ciclo é a superposição dos tempos de ciclo condicionais dos servidores originais, análogo ao método proposto por Kuehn, estendendo-se aqui ao caso multi-servidor. É desenvolvida uma expressão fechada para a transformada de Stieltjes-Laplace da distribuição do tempo de espera dos usuários em fila. O modelo de simulação é a eventos discretos e trata sistemas multi-servidores simétricos ou assimétricos, com serviço exaustivo, limitado, com barreira ou não exaustivo, e filas com capacidade finita ou infinita; os servidores operam com escalonamento do tipo repetição ou espera no caso de bloqueio do nó de destino. Ele é utilizado para propósitos de validação do modelo analítico. Extensões dos modelos e aplicações à avaliação de desempenho de redes de processadores, incluindo a rede do sistema de comutação brasileiro TRÓPICO, são consideradas.

Abstract

We consider models of systems with multiple nodes served cyclically by a number of identical servers. At each node there is one transmiting queue and one receiving queue. Customer arrival processes in the transmiting queues are Poissonian. After the service at a transmiting queue of a node is completed, the customer must be directed to a receiving queue of another node. A node cannot use more than one server at the same time. The walking and service times are general. These models are appropriate for the performance evaluation of a wide class of networks of processors. We developed one approximate analytical model and one simulation model for the evaluation of the mean waiting time at each transmiting queue. The analytical model deals with symmetric or asymmetric systems with infinite capacity queues. Blocking at the receiveing queues is considered, with servers working in the repeated mode and 1-limited service. The approach is to aggregate the servers in one equivalent server. The model is an extension to the multiserver case of the model of Hashida e Ohara. We utilize two methods for the evaluation of the cycle time of the equivalent server. In the first method, we suppose an equivalence between the service and walk rates of the equivalent server and the correspondent of the original servers. In the second method, the analysis uses the conditional cycle times, that is analogous to the method proposed by Kuehn, extended here for the multiserver case. We present a closed expression for the Laplace-Stieltjes transform of the delay distribution at each queue. The simulation model is an event discrete type and it deals with multiserver symmetric or asymmetric systems with exhaustive, limiting, gating or nonexhaustive service, and finite or infinite queue capacity. The servers work at a repeat or wait mode. It is used to validate the analytical model. Some extensions of the models and applications in the performance evaluation of networks of processors, such as the network of the brazilian switching system TRÓPICO, are considered.

Conteúdo

Capítulo 1

INTRODUÇÃO

Os sistemas computacionais distribuídos vêm atualmente ocupando cada vez mais o espaço de aplicações antes exclusivo dos sistemas centralizados e descentralizados convencionais. As principais razões para isso são ligadas a custo, modularidade, degradação suave em presença de falhas e capacidade potencial para o atendimento de novos serviços. Nesses sistemas, as unidades de processamento acopladas se comunicam através da passagem de mensagens por uma *estrutura (rede) de interligação*, a qual deve operar com um alto desempenho, para ser capaz de tratar milhares de mensagens por segundo e satisfazer a requisitos estritos relativos a tempos de transferência de mensagens entre as unidades. Cada mensagem é constituída por um conjunto de *bytes* com estrutura bem determinada e denotada por sinal *software* [23]; daí, a estrutura de interligação também é chamada *estrutura de sinalização*.

As redes de interligação que tiveram até hoje uma maior aceitação são aquelas estruturadas topologicamente em barramento ou em anel. Quanto às redes em barramento, nós consideramos neste trabalho aquelas com esquemas de acesso ao canal com passagem de permissão (*token-bus*) e com alocador centralizado [33],[34],[42]. Por sua vez, as redes em anel aqui consideradas são baseadas nos três diferentes esquemas de acesso ao canal: com passagem de permissão (*token*), anel segmentado (*slotted*) e com inserção de registro [22],[33],[34].

Redes estruturadas em barramento e em anel têm sido utilizadas tanto em sistemas comerciais quanto em protótipos de pesquisa. Elas apresentam várias características desejáveis, tais como alta taxa de utilização máxima do canal e tempo médio de transferência de mensagens entre unidades de processamento limitado, sob uma disciplina de serviço do tipo limitado. Entretanto, em redes em anel ou barramento único, os tempos de transferência podem se tornar elevados, sob condições de carga de tráfego altas e, até mesmo, moderadas. Conseqüentemente, elas não são adequadas para aplicações que requerem uma alta taxa de comunicação interprocessadores, supondo que a capacidade de transmissão de um único anel ou barramento está no seu limite máximo. Nesse caso, torna-se fundamen-

tal prover uma ampla largura de faixa através de barramentos ou anéis múltiplos [18], [22], o que for mais apropriado para a aplicação particular em questão. Um exemplo típico é a estrutura de sinalização interprocessadores do sistema de comutação TRÓPICO[1], cujo estudo foi a motivação prática do nosso trabalho. Nós denotamos genericamente por *plano de sinalização* um anel ou barramento particular. Em redes com múltiplos planos de sinalização, surgem importantes questões relativas à analise da degradação de desempenho.

A estrutura de interligação básica analisada detalhadamente em nossa abordagem é uma estrutura em barramentos múltiplos, com as seguintes características particulares:

- cada unidade de processamento é composta por um processador com sua própria memória local; por simplicidade, às vezes nós denotamos uma unidade por processador;
- cada unidade acessa todos os planos de sinalização através de uma única porta individual;
- cada porta tem um *buffer* de transmissão e um *buffer* de recepção, com capacidades de armazenamento supostamente infinitas para propósitos práticos, operando em modo semiduplex;
- a alocação de um plano a cada unidade particular é organizada por meio de um escalonamento cíclico;
- em um dado instante de alocação, no máximo uma mensagem é transmitida;
- em caso de bloqueio da porta de recepção, o escalonador passa a tratar o *buffer* de transmissão seguinte e a mensagem aguarda no seu *buffer* de transmissão por um novo atendimento (modo repetição);
- os planos operam em paralelo e independentemente; portanto, pode ocorrer interferência entre planos.

Nessa estrutura, as unidades não são capazes de utilizar completamente a capacidade de transferência de mensagens. A degradação depende principalmente dos seguintes efeitos de bloqueio:

- o *buffer* de transmissão pode ser servido somente por um plano de cada vez;
- o *buffer* de recepção só pode receber uma mensagem, se a porta correspondente não estiver transmitindo nem recebendo uma outra mensagem.

Nós também investigamos os efeitos causados no desempenho através da alteração das seguintes características:

[1] Marca registrada da Telecomunicações Brasileiras S/A - TELEBRÁS

- os *buffers* de transmissão e recepção operando em modo *full duplex*;
- cada unidade de processamento com uma porta associada a cada plano;
- em caso de bloqueio da porta de recepção, o escalonador permanece ocupado e a porta da unidade de transmissão espera pela liberação do *buffer* de recepção (modo espera);

e consideramos estruturas em barramentos múltiplos com passagem de permissão, anéis múltiplos com passagem de permissão, segmentados ou com inserção de registro operando em modo repetição ou espera, em caso de bloqueio da unidade de recepção de uma mensagem, através de extensões da estrutura básica.

O principal objetivo deste trabalho é a investigação do desempenho de estruturas de interligação de processadores com essas características. Os principais índices de desempenho que utilizamos são:

- o tempo médio de permanência das mensagens no *buffer* de transmissão;
- o tempo médio de transferência de mensagens entre duas unidades;
- a taxa global de mensagens interprocessadores, sujeita a um tempo médio de permanência no *buffer* de transmissão predeterminado.

Modelos (sistemas) multifilas com múltiplos servidores cíclicos surgem naturalmente como modelos para a análise de desempenho de redes estruturadas em barramento ou anel. Algumas vezes esses modelos são referenciados como modelos de *polling* [7],[8],[11], [19],[20]. Os usuários fazem o papel das mensagens a serem transmitidas entre as unidades (nós), enquanto que os servidores cíclicos representam os barramentos ou anéis, controlados pelos alocadores, permissões ou quadros, conforme for o caso. O tempo de caminhada de um servidor entre dois nós sucessivos em um ciclo é associado ao atraso de propagação, ao tempo de varredura ou ao adicional envolvido no armazenamento de dados, assim como deve incluir o tempo de comutação do controle de um nó para o próximo.

Nós usamos os termos modelos multifilas e sistemas multifilas indistintamente durante este texto.

Os sistemas multifilas são classificados de acordo com a disciplina de serviço, a distribuição de tráfego e a capacidade das filas.

Com respeito ao número de usuários atendidos por visita de um servidor, os esquemas de serviço usualmente encontrados na literatura são [9],[12],[15]: exaustivo, com barreira (*gated*) e limitado. Quanto à ordem de atendimento, as disciplinas mais usuais são as do tipo FIFO, LIFO e ordem aleatória. Nós consideramos sistemas onde os usuários, após o atendimento em um nó de origem, devem ser encaminhados a um nó de destino. Em caso de bloqueio do nó de destino, os servidores podem operar no modo espera ou no modo repetição [8],[18].

A distribuição de tráfego está ligada aos processos de chegada de usuários em cada nó, aos processos de caminhada e de serviço e ao encaminhamento dos

usuários entre os nós. Um sistema simétrico é aquele no qual todas as filas são idênticas, os servidores se comportam da mesma maneira e os usuários são encaminhados eqüitativamente entre os nós [12],[15].

Quanto à capacidade de armazenamento nos nós, as filas podem ser consideradas finitas (limitadas) ou infinitas (ilimitadas).

Um modelo de filas cíclicas é dito resolvido, se o tempo médio de espera em fila exato é conhecido [7],[12].

A literatura relativa a filas cíclicas, particularmente aquelas com servidor único, é muito extensa, o que pode ser verificado na seção de referências deste trabalho, que constitui uma amostra pequena — porém bastante ilustrativa — sobre a matéria. Os trabalhos encontrados na literatura são dedicados principalmente a sistemas com servidor único [1]–[15]. Vide, por exemplo, a referência [21], onde é citado que Takagi relacionou cerca de três centenas de trabalhos sobre modelos dessa classe. Devido a características de desempenho e confiabilidade, os sistemas com múltiplos servidores têm recebido recentemente uma grande atenção [16]–[25].

Neste trabalho, nós:

- apresentamos um modelo analítico da estrutura de interligação de processadores básica;
- estendemos esse modelo para torná-lo apropriado para a avaliação do desempenho de uma ampla classe de redes de interligação de processadores;
- apresentamos um modelo de simulação apropriado para a validação do modelo analítico e suas correspondentes extensões, e para analisar sistemas de filas cíclicas com múltiplos servidores ainda não considerados na literatura;
- utilizamos os modelos no dimensionamento e verificação do desempenho de redes de interligação de sistemas reais.

O modelo analítico é aproximado e trata sistemas multi-servidores simétricos ou assimétricos, com serviço limitado, disciplina FIFO e filas com capacidade infinita. Um nó pode utilizar no máximo um servidor em um dado instante, ou para transmissão ou para recepção de um usuário. Em caso de bloqueio do nó que contém a fila de recepção, os servidores operam no modo repetição. O processo de chegada dos usuários em cada nó é Poissoniano e os processos de caminhada e serviço são gerais. São tratados em uma extensão do modelo analítico os seguintes casos:

- servidores operando em modo espera;
- filas com capacidade limitada;
- nós podendo utilizar um servidor para transmissão e um servidor para recepção em um dado instante;

- acessibilidade total dos servidores a um nó.

O enfoque utilizado na solução é o da agregação dos servidores em um único servidor, o *servidor equivalente*, que tende a se comportar como o conjunto dos servidores originais, visto por um nó particular [20]. A equivalência é refletida em dois conceitos: *probabilidade de visita bem sucedida a um nó* e *tempo de ciclo do servidor equivalente*. A probabilidade de visita bem sucedida é estimada em função da proporção do tempo em regime estacionário no qual um nó não está utilizando um servidor [16],[18],[19],[20]. Para calcular o tempo de ciclo equivalente, nós utilizamos dois métodos distintos. No primeiro, nós assumimos uma certa equivalência entre as taxas de serviço e de caminhada do servidor equivalente e dos servidores originais [20]. No segundo, nós supomos que o tempo de ciclo equivalente corresponde à superposição dos tempos de ciclo condicionais dos servidores originais, análogo ao método proposto por Kuehn [4], estendendo-se aqui ao caso multi-servidor. Na análise do modelo, nós observamos uma fila marcada j, supomos que as outras filas estão em equilíbrio e seguimos o enfoque de cadeias de Markov imersas, de maneira análoga ao modelo de Hashida e Ohara [3] relativo a servidor único em férias e serviço limitado. Nós desenvolvemos uma expressão fechada para a transformada de Stieltjes-Laplace da distribuição do tempo de espera dos usuários em fila. A correspondente expressão para o tempo médio de espera em fila inclui explicitamente a probabilidade de visita bem sucedida e os dois primeiros momentos do tempo de ciclo do servidor equivalente. Assim, o modelo pode ser generalizado de maneira direta para tratar uma ampla classe de sistemas multifilas com múltiplos servidores cíclicos, através de uma caracterização conveniente desses parâmetros.

O modelo de simulação é a eventos discretos e trata sistemas multi-servidores simétricos ou assimétricos, com serviço exaustivo, limitado ou com barreira, disciplina FIFO e filas com capacidade finita ou infinita. O processo de chegada dos usuários nos nós é Poissoniano e os processos de caminhada e serviço são do tipo Erlang-k, $k = 1, 2, \ldots$ (o caso constante corresponde a $k = \infty$). Isso é suficiente para tratar os sistemas que motivaram nossas análises. No entanto, outros tipos de processos de chegada, caminhada e serviço são considerados em extensões do modelo. Em caso de bloqueio de um nó que contém a fila de recepção para onde deve ser encaminhado um usuário, os servidores podem operar no modo repetição ou no modo espera. Ele permite a obtenção de alguns importantes parâmetros de desempenho para propósitos de comparação com os resultados do modelo analítico. As comparações servem para verificar a região de validade das aproximações, o que é muito importante, pois não existe solução exata para o problema aqui focalizado. Alguns autores chegam até mesmo a duvidar que uma solução exata possa vir a existir um dia [4],[22]. O modelo de simulação assume, conseqüentemente, um papel relevante, constituindo uma parte integrante da modelagem.

No Capítulo 2, nós descrevemos sumariamente as principais estruturas de interligação de processadores às quais se aplicam os modelos propostos. No Capítulo 3, nós comentamos trabalhos anteriores existentes na literatura, introduzimos de-

finições e hipóteses, desenvolvemos o modelo analítico aproximado e discutimos possíveis extensões. No Capítulo 4, nós descrevemos o modelo de simulação e suas extensões, com o auxílio dos diagramas de transição de estados de suas entidades, e mostramos como as entidades se relacionam a fim de reproduzir no modelo as mesmas características dos sistemas de filas cíclicas em consideração. As análises numéricas relativas à validação do modelo analítico através da comparação com resultados de simulação são mostradas no Capítulo 5. Aplicações numéricas dos modelos à avaliação de desempenho de estruturas de interligação de processadores são apresentadas no Capítulo 6. Nesse capítulo, nós analisamos o desempenho das estruturas de interligação do sistema TRÓPICO, de um ponto de transferência de sinalização por canal comum e de um sistema de comutação de mensagens de alto desempenho. Constatações importantes relativas ao tipo de acesso dos processadores a um plano e ao tipo de transmissão são apresentadas. No Capítulo 7, nós concluímos e apontamos direções para pesquisas futuras, no sentido de melhorar a precisão do modelo analítico e estender as suas aplicações à análise do desempenho de outras estruturas de interligação de processadores. Uma descrição simplificada da estrutura de interligação que motivou as nossas investigações é apresentada no Apêndice A. Finalmente, nos Apêndices B e C, respectivamente, nós apresentamos os manuais de utilização dos correspondentes programas de computador do modelo analítico e do modelo de simulação. Os programas foram escritos na linguagem S-Port SIMULA[2] [30] do sistema VAX/VMS[3].

[2]Marca Registrada do *Norwegian Computing Center*

[3]Marca Registrada da *Digital Equipment Corporation*

Capítulo 2

TIPOS DE ESTRUTURAS DE INTERLIGAÇÃO DE PROCESSADORES

Nós consideramos a classe de sistemas distribuídos que têm as seguintes características:

1. As funções e a carga de tráfego são distribuídas em um grande número de unidades de processamento.

2. As funções são particionadas em fases elementares. O tratamento de uma fase é feito pela concatenação de tarefas apropriadas. Cada unidade é autônoma e o processamento é controlado por fluxos de dados transportados por sinais *software* chamados mensagens.

3. O controle é feito da maneira que segue. A primeira tarefa da primeira fase é ativada por um evento externo ao sistema. Após a execução de uma tarefa qualquer, é gerada uma mensagem que ativa a próxima tarefa da fase. Cada tarefa é realizada por uma unidade de processamento. Ao ser ativada, a tarefa entra em uma fila onde espera ser executada pela unidade.

4. Quando duas tarefas consecutivas em uma fase são executadas por uma mesma unidade, a segunda é ativada imediatamente após o término da execução da primeira. Caso contrário, há que ser utilizada a estrutura de interligação e, então, um certo tempo é requerido para a transmissão da mensagem. A estrutura de interligação também é chamada *estrutura de sinalização*, pois a comunicação entre as unidades é feita por meio de sinais (mensagens).

5. Os programas que executam as tarefas são distribuídos pelas unidades de processamento. Um programa é armazenado em uma única unidade ou é repetido em várias unidades.

6. Não há fortes correlações entre as unidades que cooperam para a execução das fases.

7. Não há pares de unidades que cooperam para a execução das fases, tal que uma trabalha durante uma alta porcentagem de tempo em relação à outra.

8. A escolha de uma unidade para a execução de cada tarefa é independente do estado do sistema, o que facilita o cálculo do número médio de tarefas executadas em uma unidade de processamento por unidade de tempo.

Nós descrevemos sumariamente nas Seções 2.1 e 2.2 alguns tipos de estruturas de interligação apropriadas aos sistemas distribuídos em consideração, que podem ser modeladas como sistemas de filas cíclicas com múltiplos servidores. As características importantes para a modelagem são ressaltadas na Seção 2.3. Finalmente, na Seção 2.4, nós caracterizamos a necessidade da avaliação de desempenho dessas estruturas.

2.1 Redes Tipo Barramento

Uma rede em barramento é estruturada em $N \geq 2$ unidades de processamento: $P_1, P_2, \ldots, P_N$, ligadas a $B \geq 1$ barramentos através de portas padronizadas (Figura 2.1).

No caso em que $B \geq 2$, cada unidade pode acessar os barramentos através de uma única porta (Figura 2.2) ou através de B portas, cada uma associada a um barramento particular (Figura 2.3). Cada porta, por sua vez, tem um *buffer* de transmissão (TX) e um *buffer* de recepção (RX).

Em redes em barramento baseadas em esquemas de acesso do tipo *Carrier Sense Multiple Access/Collision Detection* ($CSMA/CD$), o tempo para entregar uma mensagem pode se tornar ilimitado, devido ao grande número de colisões e retransmissões[22]. Assim, a tendência atual tem sido desenvolver redes em barramentos múltiplos com passagem de permissão (*token-bus*) e com alocador centralizado [42].

Nas descrições apresentadas a seguir, nós supomos que cada unidade acessa os barramentos através de uma única porta.

2.1.1 Barramento Com Passagem de Permissão

Em uma rede estruturada em barramentos múltiplos e com esquema de acesso do tipo passagem de permissão, a regra de operação é normalmente como segue.

As B fichas nos B barramentos são independentes e operam assincronamente em cada barramento. Se uma unidade de processamento não tem mensagem a transmitir, a porta dessa unidade simplesmente passa as fichas que chegam à

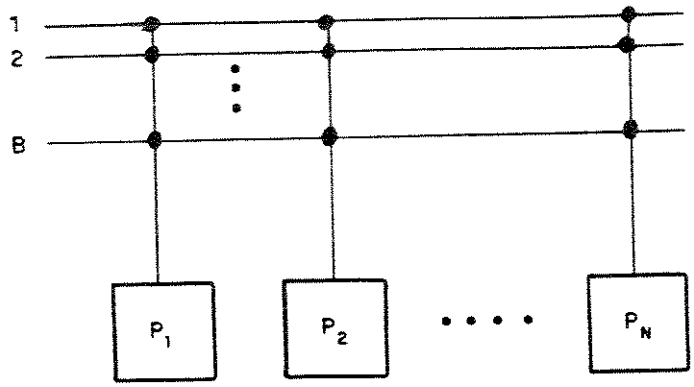


Figura 2.1:

Rede Em Barramento Múltiplo

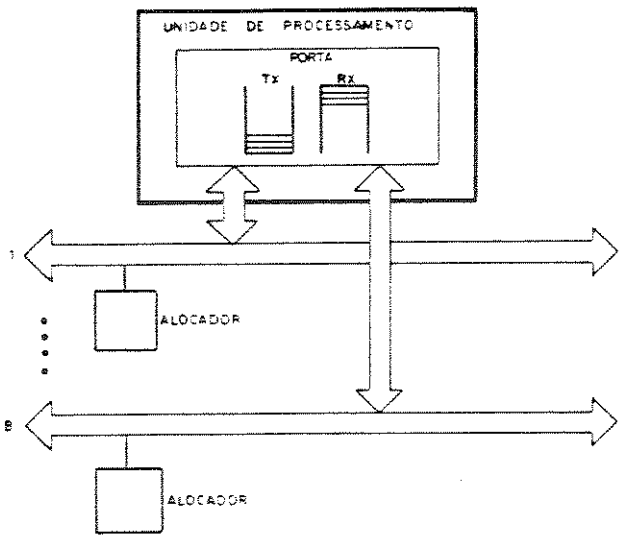


Figura 2.2:

Acesso das Unidades à Rede Através de Porta Única

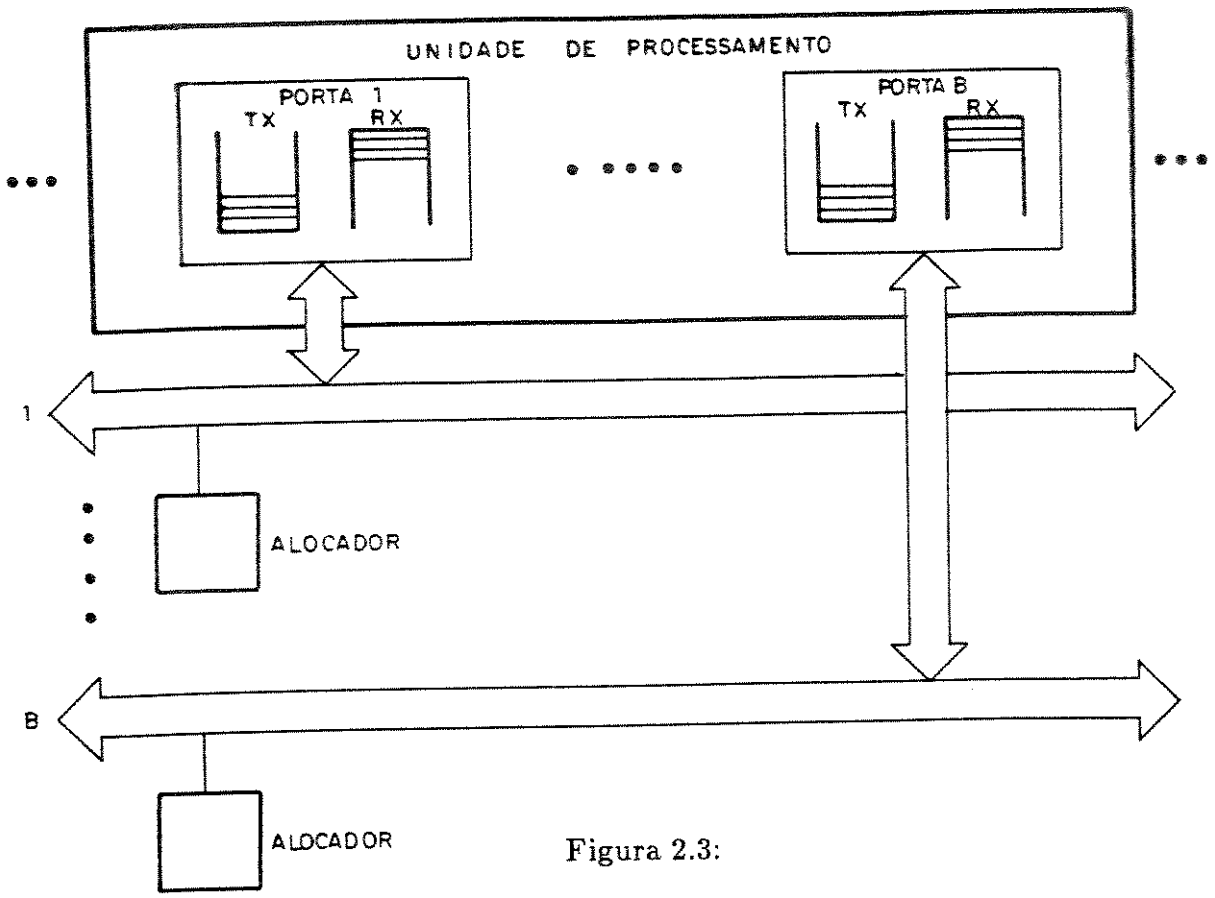


Figura 2.3:

Acesso das Unidades à Rede Através de Portas Múltiplas

próxima unidade. Caso contrário, essa porta monitora simultaneamente todos os barramentos e captura a primeira ficha que chega em qualquer um dos B barramentos. A porta, então, converte a ficha em um conector e o segue juntamente com o pacote da mensagem. O conector, como a ficha, é uma seqüência curta e única de *bits*, distinta de qualquer seqüência possível do campo de um pacote. Enquanto uma unidade transmite uma mensagem em um barramento particular, ela não pode transmitir outra mensagem por qualquer outro barramento. As fichas que chegam à unidade durante a transmissão são passadas para a próxima unidade do barramento. Cada unidade "escuta" os barramentos, para receber as mensagens que lhe são destinadas. A capacidade de armazenamento do *buffer* de recepção pode ser considerada infinita para propósitos práticos. A disciplina de transmissão de mensagem pode ser exaustiva, limitada ou com barreira.

2.1.2 Barramento Com Alocador Centralizado

Em uma rede em barramento com alocador centralizado, cada barramento é gerenciado por um alocador, e é constituído por enlaces de controle e enlaces de comunicação. Um enlace de controle serve à troca de informações entre o alocador e as unidades a ele associadas, a fim de possibilitar o acesso à rede de interligação. Um enlace de comunicação é o meio físico por onde se propagam as mensagens entre os processadores. A alocação de um barramento a cada unidade particular

é organizada por meio de um escalonamento cíclico.

Consideremos uma tentativa de envio de mensagem de uma unidade de transmissão qualquer para outra unidade de recepção qualquer. Em caso de bloqueio da unidade de recepção, duas alternativas são possíveis [18]:

- o alocador passa a tratar a unidade seguinte e a mensagem espera no seu *buffer* de transmissão até que a sua porta seja atendida novamente (modo barramento com repetição);
- o barramento permanece ocupado e a porta da unidade de transmissão espera pela liberação do *buffer* de recepção, quando então a unidade de recepção torna-se apta a aceitar a mensagem (modo barramento com espera).

Normalmente, os *buffers* de transmissão são servidos de maneira limitada, ou seja, por instante de varredura no máximo uma única mensagem é transmitida. O bloqueio da unidade de recepção pode ser devido:

1. à porta de recepção já estar sendo ocupada por outro barramento. No caso de operação em modo *full duplex*, considera-se que a porta de recepção está ocupada quando ela está recebendo uma outra mensagem. Entretanto, se o modo de operação for semiduplex a porta de recepção está ocupada quando ela está transmitindo ou recebendo uma outra mensagem.
2. à limitação da capacidade do *buffer* de recepção dessa unidade.

No Apêndice A, é apresentada uma descrição simplificada da estrutura de interligação em barramentos múltiplos com alocador centralizado e modo repetição de um sistema que está atualmente em operação comercial.

2.2 Redes Tipo Anel

As redes em anel pertencem à classe das redes assíncronas com controle descentralizado, que apresentam grandes vantagens com respeito a problemas de distribuição de relógio, à estabilidade e à utilização eficiente da largura de faixa[22]. As disciplinas de transmissão de mensagens nessas redes normalmente são do tipo exaustivo, limitado ou com barreira.

Uma rede em anel múltiplo é ilustrada na Figura 2.4. Ela é estruturada fisicamente em $N \geq 2$ unidades de processamento: $P_1, P_2, \ldots, P_N$, conectadas a $B \geq 1$ anéis através de portas padronizadas.

A seguir, nós descrevemos os esquemas de acesso ao canal que até o momento tiveram maior aceitação: anel com passagem de permissão (*token*), segmentado (*slotted*) e com inserção de registro.

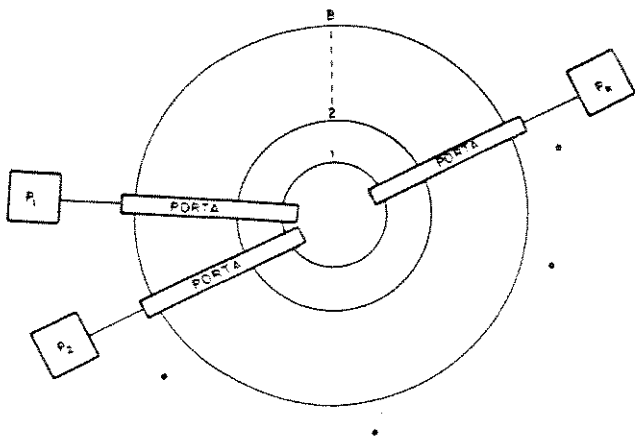


Figura 2.4:

Rede Em Anel Múltiplo

2.2.1 Anel Com Passagem de Permissão

Em uma rede em anéis múltiplos com passagem de permissão, a regra de operação é normalmente como segue, supondo que cada unidade acessa todos os anéis através de uma única porta.

As B fichas nos B anéis são independentes e operam assincronamente em cada anel. Se uma unidade de processamento não tem mensagem a transmitir, a porta simplesmente passa as fichas que chegam à próxima unidade. Caso contrário, a porta monitora simultaneamente todos os anéis e captura a primeira ficha que chega em qualquer um dos B anéis. A porta, então, converte a ficha em um conector e o segue com o pacote da mensagem. A unidade remove a mensagem transmitida quando ela retorna pelo anel. Enquanto uma unidade transmite uma mensagem em um anel particular, ela não pode transmitir outra mensagem em qualquer outro anel. As fichas que chegam à unidade durante a transmissão são passadas para a próxima unidade do anel. Cada unidade tem um *buffer*, que pode ser considerado de capacidade infinita para propósitos práticos. Assim, ela pode receber mais que uma mensagem simultaneamente, de modo a evitar qualquer perda de pacote. Na disciplina de transmissão não exaustiva, uma unidade só pode transmitir a próxima mensagem por um dado anel após a ficha correspondente retornar à mesma.

2.2.2 Anel Segmentado

Em redes em anel múltiplo segmentado, cada um dos B anéis é formatado por um número constante de quadros de tamanho fixo que circulam continuamente em torno do anel. Um indicador (*flag*) dentro do cabeçalho de cada quadro é usado para indicar se o quadro está vazio ou cheio. Uma unidade de processamento que tem uma mensagem a transmitir espera pela chegada de um quadro vazio em qualquer anel, marca-o como cheio e começa a transmitir. Normalmente, é permitida somente uma transmissão ativa em qualquer instante. Entretanto, uma unidade de destino pode receber até B pacotes nos B anéis simultaneamente (*buffer* de capacidade infinita em cada unidade). Na disciplina não exaustiva, não é permitido a uma estação de origem reusar o quadro do qual ela tenha removido uma mensagem. Após o quadro ter sido marcado como vazio, é sempre passado à próxima unidade. O número total de quadros em qualquer anel é conhecido durante a operação de iniciação. Assim, a unidade de origem pode reconhecer sua mensagem transmitida simplesmente contando o número de quadros que passam.

Nota: Embora os quadros dos diferentes anéis sejam assíncronos e seus cabeçalhos distribuídos uniformemente sobre os comprimentos dos anéis, Bhuyan *et alii* [22] assumiram que a circulação dos quadros nos B anéis é sincronizada, isto é, uma unidade pode observar B indicadores nos B quadros dos diferentes anéis ao mesmo tempo. Segundo eles conjecturaram, em termos do tempo de espera de uma mensagem em uma unidade, a diferença entre as duas situações normalmente é pequena.

2.2.3 Anel Com Inserção de Registro

Normalmente, em redes em anel múltiplo com inserção de registro, uma unidade de processamento tem um registro de recepção (RRX), um registro de inserção (RIN) e uma chave de recepção (CRX) para cada um dos B anéis e um único registro de transmissão (RTX) e uma única chave de transmissão (CTX) para todos os anéis [22] (Figura 2.5).

Uma CRX pode conectar uma linha de saída a uma linha de entrada, ao RRX ou ao RIN. O RTX é utilizado para transmitir uma mensagem em um dos B anéis; portanto, uma unidade não pode transmitir mais do que uma mensagem por vez. No entanto, uma unidade pode ter mais que um RRX inserido nos diferentes anéis, desde que eles sejam inseridos em instantes diferentes. Quando uma unidade não tem uma mensagem a transmitir, as linhas de entrada são conectadas às linhas de saída em todos os anéis. Os dados das linhas de entrada são também transferidos para os RRX's correspondentes. Quando uma mensagem destinada à unidade está inteiramente contida em um RRX, ela é removida e enviada à unidade. Quando uma unidade deseja transmitir uma mensagem, a porta monitora

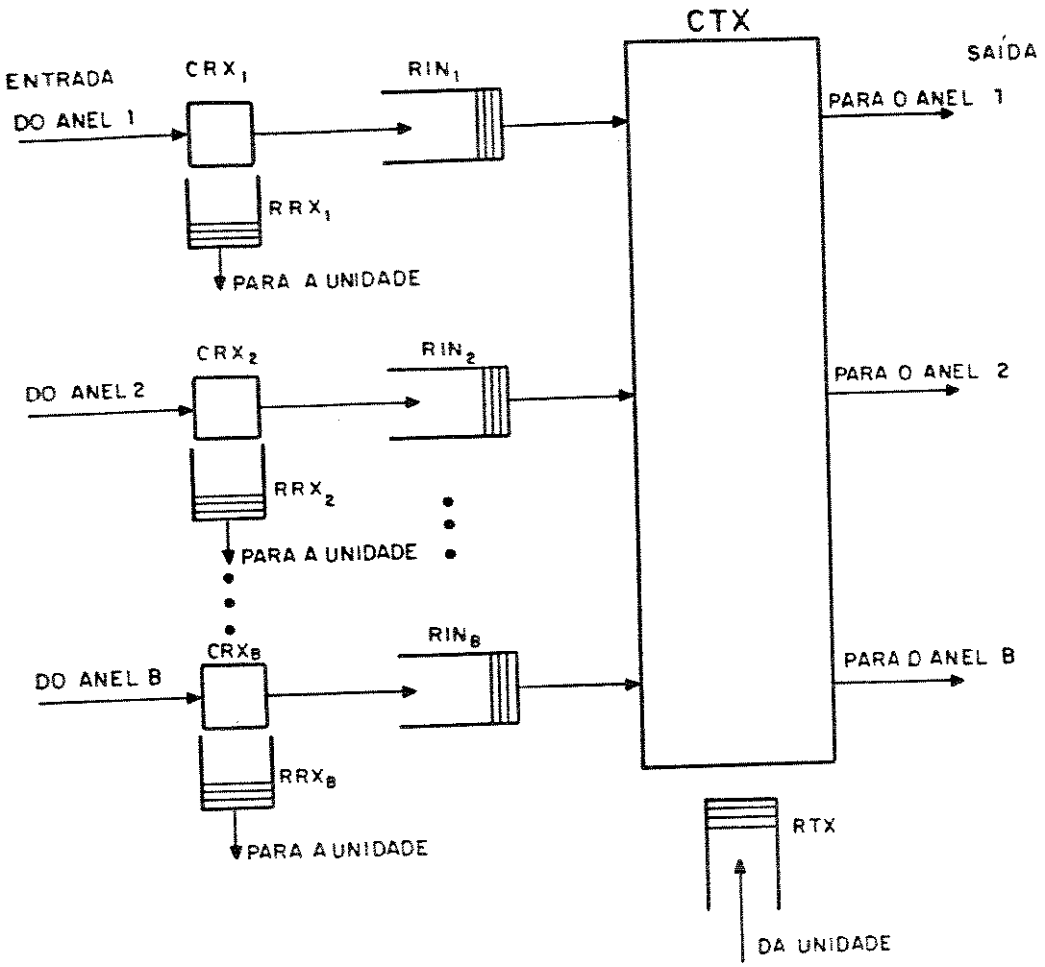


Figura 2.5:

Porta do Anel Múltiplo Com Inserção de Registro

os anéis. Se houver um anel cujos dados por ele passando não são parte de uma mensagem, isto é, são *bits* vazios, um registro contendo os *bits* da mensagem a transmitir é inserido na linha imediatamente. Se pacotes de mensagens estão passando por todos os anéis, o registro é inserido no primeiro anel onde ocorrer que o último *bit* da mensagem tenha por ele passado. A inserção da mensagem no anel é obtida pela conexão do RTX à linha de saída do anel em questão, e pelo armazenamento dos dados na linha de entrada do RRX correspondente durante a transmissão. O RIN contém a mensagem completa formatada no anel. Quando uma unidade reconhece que uma mensagem transmitida retornou e está completamente contida no RIN, o registro é removido do anel, removendo-se, conseqüentemente, a mensagem transmitida. Na disciplina de transmissão não exaustiva, caso uma unidade já tenha uma mensagem no anel, ela pode transmitir a próxima mensagem somente após a mensagem pendente ter sido removida do anel.

2.3 Similaridade Entre as Estruturas

Sob o ponto de vista de modelagem para a avaliação de desempenho, as estruturas de interligação apresentadas em 2.1 e 2.2 são similares. Elas podem ser vistas como sistemas de múltiplas filas com múltiplos servidores cíclicos.

Os usuários são as mensagens a serem transmitidas entre as estações de serviço.

Os servidores representam os barramentos ou os anéis, conforme for o caso. Os alocadores, quadros ou permissões determinam os instantes de disponibilidade dos servidores.

Nas estruturas em anel ou barramento com passagem de permissão, as fichas circulam continuamente em torno do anel ou barramento. Analogamente, nas estruturas em anel segmentado ou inserção de registro, os quadros ou os *bits* vazios circulam em torno do anel. Já nas estruturas em barramento com alocador centralizado, os alocadores varrem os canais do enlace de controle em busca de solicitação de transmissão de mensagem. Isso dá o caráter cíclico aos servidores.

Enquanto uma estação transmite uma mensagem em um anel ou barramento particular, em geral ela não pode transmitir outra mensagem em qualquer outro anel ou barramento. As fichas, os quadros ou *bits* vazios que chegam nesse ínterim são passados para a próxima estação. No caso da estrutura com alocador centralizado, a estação que foi atendida por um alocador retira a solicitação de transmissão de mensagem dos enlaces de controle de todos os outros alocadores. Isso caracteriza a possibilidade de haver interferência entre os servidores.

Assim, nós temos que considerar um modelo em que os servidores visitam as estações, de acordo com um escalonamento cíclico, em busca de usuários a serem servidos. Quando um servidor visita uma estação onde há pelo menos um usuário em espera, ele atende um usuário, desde que não haja interferência com outros

servidores (sobreposição) e a estação para onde deve ser encaminhado esse usuário esteja apta para recebê-lo. Após o serviço, o servidor visita a próxima estação no seu ciclo após um tempo chamado tempo de caminhada. Caso a estação para onde deve ser encaminhado o usuário não esteja apta para recebê-lo, o comportamento do servidor depende do modo de escalonamento:

modo repetição: o servidor visita a próxima estação no seu ciclo após o tempo de caminhada;

modo espera: o servidor aguarda a estação de destino se tornar apta para receber o usuário.

Se não houver usuários em espera em um instante de visita, o servidor também visita a próxima estação no seu ciclo após o tempo de caminhada, que normalmente não pode ser considerado nulo. Ele é associado ao atraso de propagação, tempo de varredura ou o adicional envolvido no armazenamento de dados e tempo de comutação de uma estação para a próxima.

Um modelo deve, portanto, refletir a interferência entre os servidores e entre as estações, de modo a estimar um dado parâmetro de desempenho, como, por exemplo, o tempo médio de espera dos usuários em fila ou o tempo médio de transferência dos usuários entre duas estações.

O modelo analítico e o modelo de simulação apresentados nos capítulos 3 e 4, respectivamente, são especialmente adequados para redes de interligação com múltiplos planos e esquemas de acesso ao canal do tipo alocação centralizada ou passagem de permissão. Deve-se ressaltar, entretanto, que alguns autores modelam também redes com esquemas do tipo anel segmentado ou com inserção de registro por sistemas com múltiplas filas e múltiplos servidores cíclicos [19],[22]. Assim, nossos modelos podem, em princípio, ser adaptados para analisar redes em anéis múltiplos com esses esquemas de acesso ao canal.

Nota: Nas estruturas de múltiplos processadores que estamos considerando, é muito importante a rápida detecção de um plano de sinalização em falha. Nesse sentido, nós propomos em [23] um algoritmo dinâmico de supervisão de falhas, que é baseado em uma generalização do *problema das caixas de fósforos de Banach* [35].

2.4 Necessidade da Avaliação de Desempenho

Os requisitos de desempenho de tráfego se reportam aos tempos de execução das fases das funções consideradas importantes pelo usuário do sistema. Na área de Telefonia, por exemplo, normalmente são adotados os requisitos de qualidade de serviço recomendados pelo CCITT[1][40], que são especificados a dois níveis

[1] Comitê Consultivo Internacional de Telegrafia e Telefonia

de tráfego: nível de carga normal, referenciado como carga nominal, e nível de carga acima do normal, referenciado como sobrecarga. Uma sobrecarga pode ser caracterizada por um acréscimo da ordem de, por exemplo, 20% ou 40% em relação à carga nominal. Os padrões numéricos de qualidade, também chamados graus de serviço, não são baseados em nenhuma fórmula teórica de otimização entre a satisfação do usuário e a economia do sistema. Eles têm, na verdade, uma conotação puramente prática e têm sido justificados no decorrer de muitos anos de experiência operacional. Os graus de serviço visam dotar o sistema de uma margem considerável de proteção contra aumentos imprevistos de tráfego que venham a degradar o serviço a níveis alarmantes de congestionamento e perceptíveis aos usuários.

Para não perturbarem o bom funcionamento do sistema, os usuários não podem perceber nenhum atraso significativo na execução das fases, ou seja, devem ter uma ilusão de instantaneidade de serviço. Ao perceberem um atraso anormal, eles podem ter uma reação de impaciência, que pode, eventualmente, causar uma sobrecarga catastrófica no sistema. Qualquer que seja essa reação, o efeito sobre o sistema é sempre traduzido por um tráfego espúrio, ou seja, uma utilização ineficaz de recursos, onde tarefas prioritárias, porém inúteis, tomam tempo e ocupam memória dos processadores. Um exemplo bem conhecido é o que ocorre em uma central telefônica, onde o usuário é conhecido como assinante e a primeira fase da função processamento de chamada originada é denominada pré-seleção. Ela consiste na identificação da retirada do fone do gancho feita pelo assinante e do envio a este assinante de um sinal audível, o tom de discar, que representa o convite à numeração. Se, após retirar o fone do gancho, o assinante não receber imediatamente esse tom, ele pode ter comumente duas reações distintas de impaciência:

numeração intempestiva : o assinante começa a numeração sem verificar a presença do tom; nesse caso, os primeiros dígitos são perdidos ou alterados e a central rejeita a chamada, após um certo tempo;

desligamento prematuro : o assinante repõe o fone no gancho, mesmo que sua demanda já esteja sendo tratada.

A ilusão de instantaneidade é conseguida estipulando-se que o tempo médio de execução de uma fase seja bem pequeno, ou, alternativamente, que a probabilidade que o tempo de execução ultrapasse um dado valor especificado seja bem pequena. Como decorrência, a rede de interligação deve ter uma disponibilidade suficiente, tal que qualquer efeito de bloqueio possa ser considerado desprezível. Normalmente, as variáveis de interesse que servem para caracterizar o desempenho da rede de interligação são

- o tempo médio de espera das mensagens no *buffer*

e

- o quantil da distribuição do tempo de espera.

Os modelos matemáticos analíticos e os modelos de simulação são utilizados nas fases de especificação, planejamento e projeto e durante o desenvolvimento de um sistema. Eles constituem uma importante ferramenta de apoio técnico na tomada de decisões e previsão de desempenho, uma vez que possibilitam a realização de experimentos que seriam impossíveis ou impraticáveis em sistemas reais.

O modelo analítico desenvolvido no Capítulo 3, as suas correspondentes extensões e as simulações têm como objetivo caracterizar o tempo médio de espera e a estabilidade de uma fila em um sistema de múltiplas filas com múltiplos servidores cíclicos.

Capítulo 3

A MODELAGEM ANALÍTICA E DEFINIÇÕES

Neste capítulo, nós usamos os termos genéricos usuários, tempo de serviço e tempo de caminhada para caracterizar, respectivamente, as mensagens, o tempo de transferência de mensagem e o tempo de varredura, e, com relação a uma variável aleatória genérica Y, nós denotamos por:

- P: a probabilidade;
- E: a esperança matemática;
- $F_Y(t) = P[Y \leq t]$: a sua função de distribuição de probabilidade. $f_Y(t) = \frac{dF_Y}{dt}$, caso exista, é a correspondente função densidade de probabilidade.
- $y = E[Y] = \int_0^\infty t dF_Y(t)$ e $y^{(2)} = E[Y^2] = \int_0^\infty t^2 dF_Y(t)$: o primeiro e segundo momentos, respectivamente. Supomos, neste trabalho, que todas as distribuições possuem os dois primeiros momentos finitos.
- $\phi_Y(\xi) = \int_{0^-}^\infty e^{-\xi t} dF_Y(t)$: a transformada de Stieltjes-Laplace (T.S.L.) de F_Y, sendo que $y = -\phi'(0)$ e $y^{(2)} = \phi''(0)$. ξ é um parâmetro complexo e a integral converge para uma parte real de ξ maior que um certo τ_0 (abcissa de convergência).
- $VAR[Y] = y^{(2)} - y^2$: a variância da distribuição.

Na Seção 3.1, nós introduzimos as definições e hipóteses que caracterizam o nosso modelo analítico. Na Seção 3.2, nós comentamos brevemente trabalhos anteriores existentes na literatura sobre sistemas de múltiplas filas com múltiplos servidores cíclicos. O desenvolvimento do modelo analítico aproximado é apresentado na Seção 3.3. Finalmente, possíveis extensões do modelo são discutidas na Seção 3.4.

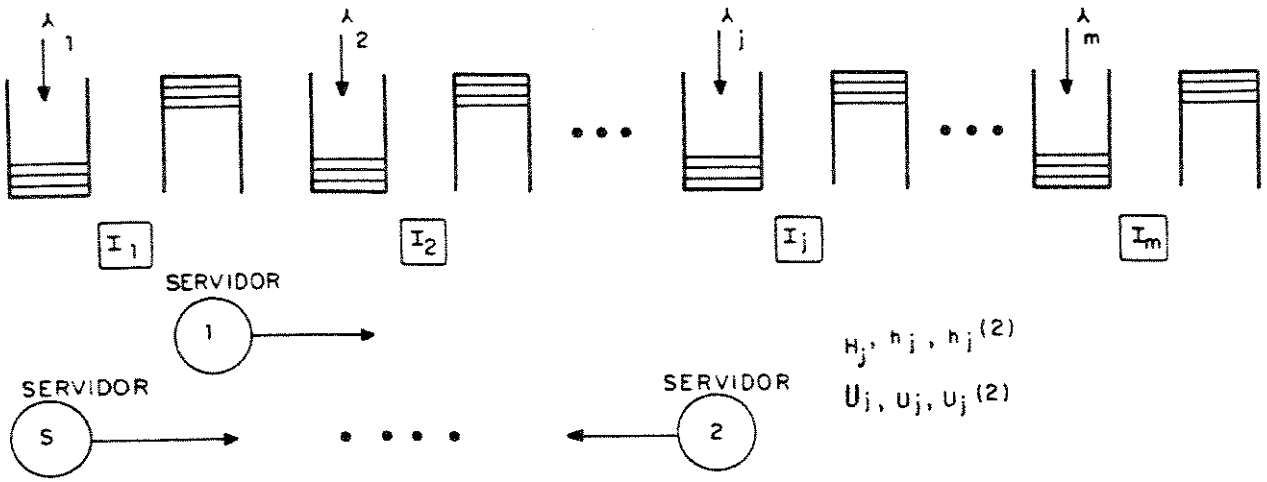


Figura 3.1:

Sistema Multifilas Com Múltiplos Servidores Cíclicos

3.1 Caracterização do Problema

Nós consideramos o sistema de filas cíclicas ilustrado na Figura 3.1. Ele consiste de um certo número $m \geq 2$ de nós de serviço, cada um contendo uma fila de transmissão e uma fila de recepção (*buffers*). As filas de transmissão recebem usuários do mundo externo de acordo com taxas médias denotadas por λ_j, $j = 1, 2, \ldots, m$. O sistema é servido por s servidores, tal que $1 \leq s < m$, que visitam os nós de acordo com um escalonamento cíclico a fim de, após o serviço, encaminhar os usuários da fila de transmissão de um nó para a fila de recepção de outro nó. O tempo requerido por um servidor para se mover de um nó ao próximo em um ciclo é chamado tempo de caminhada.

Em geral, um sistema multifilas como esse é classificado de acordo com o número de servidores, a disciplina de serviço, a distribuição de tráfego e a capacidade das filas. Em relação ao número de servidores, temos sistemas com servidor único e sistemas com múltiplos servidores cíclicos. Os esquemas de serviço com respeito ao número de usuários atendidos por visita de um servidor, usualmente encontrados na literatura, são [9],[12],[15]:

exaustivo: todos os usuários são servidos até que a fila se torne vazia;

com barreira *(gated)*: todos os usuários que estão na fila são servidos, mas nenhum que eventualmente chegar após a visita do servidor é atendido;

ℓ–limitado: um limite, digamos ℓ, é colocado como o número máximo de usuários que podem ser servidos. Este pode operar ainda em uma maneira exaustiva ou com barreira. O caso $\ell = 1$ para todas as filas é chamado serviço *um-por-vez*.

Quanto à ordem de atendimento dos usuários, as disciplinas mais usuais são as do tipo FIFO, LIFO e ordem aleatória. Em caso de bloqueio do nó que contém a fila de recepção para onde deve ser encaminhado um usuário, há dois modos de escalonamento possíveis [8],[18]:

modo servidor com espera: o servidor permanece ocupado e o usuário do nó de origem espera pela liberação do nó de destino;

modo servidor com repetição: o servidor passa a tratar a fila do próximo nó de origem no ciclo de caminhada e o usuário espera por um novo atendimento de um servidor.

Quanto à distribuição de tráfego, o sistema é classificado como simétrico ou assimétrico [12],[15], de acordo com os processos de chegada de usuários em cada nó, com os processos de caminhada e de serviço e com o encaminhamento dos usuários entre os nós. Um sistema simétrico é aquele no qual os processos de chegada, caminhada e serviço possuem as mesmas distribuições para todos os nós e os usuários são encaminhados eqüitativamente entre os nós. Quanto à capacidade, as filas podem ser finitas (limitadas) ou infinitas (ilimitadas).

É muito difícil definir uma notação compacta para descrever os diferentes sistemas com filas múltiplas e múltiplos servidores cíclicos. Em [21], é feita uma tentativa nesse sentido, utilizando uma notação derivada da notação de Kendall.

Um modelo de filas cíclicas é dito resolvido, se o tempo médio de espera em fila exato é conhecido [7],[12]. De acordo com esta definição, a maioria dos modelos de filas cíclicas com servidor único e serviço exaustivo ou com barreira e o modelo não exaustivo simétrico já foi resolvida. Com relação aos modelos com múltiplos servidores, somente o modelo Markoviano com filas de capacidade finita e sem a consideração do efeito de bloqueio do nó de destino já foi resolvido [21]. Devido à complexidade numérica ou à impossibilidade matemática de obtenção de resultados, é comum a utilização de aproximações em algum estágio da análise desses modelos.

Outras características importantes de um sistema de filas cíclicas são:

1. a distribuição dos intervalos de tempo entre chegadas dos usuários: a hipótese de que os processos de chegada são Poissonianos é bastante razoável, pois a distribuição exponencial negativa se ajusta estatisticamente bem ao tempo entre chegadas nos sistemas reais que estamos considerando e, além disso, favorece a tratabilidade matemática;

2. a distribuição do tempo de serviço: obviamente, o tempo de serviço pode ser considerado constante para mensagens de tamanho fixo, mas há que ser adotada uma distribuição mais realista quando isto não ocorre (exponencial, hiperexponencial, etc.);

3. a distribuição do tempo de caminhada: normalmente, o tempo de caminhada não pode ser considerado nulo e tem uma distribuição geral;

4. a capacidade das filas de transmissão e recepção: a hipótese de que há um número infinito de *buffers* é bastante razoável nos sistemas atuais, devidos aos avanços tecnológicos na implementação de memórias com grande capacidade;

5. a estabilidade: uma fila é estável, se para taxas de entrada finitas o seu comprimento médio é limitado. O sistema de filas cíclicas é estável, se todas as filas são estáveis.

Assumimos no nosso modelo as seguintes hipóteses operacionais:

H_1: As filas de transmissão e recepção dos m nós são de capacidade infinita.

H_2: Os s servidores são idênticos. Os usuários de um nó podem ser atendidos por qualquer servidor. No entanto, um nó pode utilizar no máximo um servidor em um dado instante, seja para transmissão, seja para recepção. Portanto, dizemos que um nó está livre quando ele não está usando nenhum servidor. A cada nó j, $j = 1, 2, \ldots, m$, está associado um indicador de ocupação $I_j(t)$, onde

$$I_j(t) = \begin{cases} 0, & \text{se o nó está livre no instante } t, \\ 1, & \text{caso contrário.} \end{cases}$$

Quando o sistema é estacionário,

$$P[I_j(t) = 0] = p_j$$

e

$$P[I_j(t) = 1] = 1 - p_j \ ,$$

para um dado p_j tal que $0 < p_j \leq 1$. Assim, um usuário de uma fila de transmissão está apto para ser servido quando o correspondente nó e o nó que contém a fila de recepção para onde ele deve ser transportado estão ambos livres.

H_3: Os processos de chegadas de usuários nas filas de transmissão são Poissonianos e independentes. Se o processo de chegada é simétrico então $\lambda_j = \lambda$, $\forall j$.

H_4: Os usuários de uma fila de transmissão de um nó de origem j devem ser encaminhados à fila de recepção de um nó de destino k com probabilidade constante α_{jk}, onde $0 \leq \alpha_{jk} \leq 1$ $\forall j, k$ e $\alpha_{jj} = 0$, $\forall j$. Se $\lambda_j > 0$ então

$\sum_{k=1}^{m} \alpha_{jk} = 1, \forall j$. Se o processo de encaminhamento é simétrico então $\alpha_{jk} = \frac{1}{m-1}, \forall j,k$. Quando um servidor visita um nó j que está ocupado $(I_j = 1)$ ou que está livre $(I_j = 0)$, mas a sua fila de transmissão está vazia, ele o ignora e prossegue com a caminhada. Analogamente, se o nó j está livre $(I_j = 0)$ e há um usuário na sua fila de transmissão, mas o nó de destino k está ocupado $(I_k = 1)$, o servidor ignora o nó j e prossegue com a caminhada. O primeiro usuário da fila de transmissão do nó j espera por uma nova visita de um servidor, ou seja, a disciplina de atendimento é *FIFO* e o modo de escalonamento é do tipo ***repetição.***

H_5: Os usuários de uma fila de transmissão j recebem serviço por um tempo aleatório independente do servidor, denotado por H_j. Se o processo de serviço é simétrico então $H_j = H, \forall j$.

H_6: A disciplina de transmissão é do tipo *um-por-vez*, ou seja, no máximo um usuário da fila de transmissão é atendido por visita de um servidor ao nó j (serviço 1-limitado).

H_7: Após ter completado o serviço em uma particular fila de transmissão j ou após uma visita mal sucedida a esta fila, o servidor passa à fila $(j \bmod m) + 1$ após um tempo de caminhada aleatório U_j. Se o processo de caminhada é simétrico então $U_j = U, \forall j$.

Nota: Um sistema é simétrico, se os processos de chegada, encaminhamento, serviço e caminhada são simétricos.

H_8: Os processos de chegadas e de serviço dos usuários em cada nó são independentes uns dos outros e também são independentes daqueles dos outros nós. Para cada servidor, os tempos de caminhada entre os nós são independentes daqueles dos outros servidores e também independentes dos processos de chegadas e de serviço em cada nó.

H_9: Quando um usuário é transmitido com sucesso à fila de recepção, ele deixa o sistema.

Sem perda de generalidade, as hipóteses H_1-H_9 caracterizam o modelo analítico de uma estrutura de interligação de processadores do tipo multibarramento com alocador centralizado, onde:

- cada unidade de processamento acessa os barramentos através de uma única porta;
- os *buffers* de transmissão e recepção das portas, supostos com capacidades de armazenamento infinitas para propósitos práticos, operam no modo semiduplex;

- a disciplina de transmissão de mensagem é do tipo 1-limitado e na ordem de chegada;

- o modo de escalonamento em caso de bloqueio da unidade de destino de uma mensagem é do tipo barramento com repetição.

Entretanto, esse modelo pode, eventualmente, ser estendido para tratar as outras estruturas de interligação descritas no Capítulo 2.

Nós definimos o tempo de espera de um usuário na fila de transmissão de um nó j, W_j, $j = 1, 2, \ldots, m$, como o intervalo de tempo decorrido desde o instante de chegada desse usuário até o instante em que o serviço começa. O nosso problema consiste em *determinar o tempo médio de espera em fila*, w_j, para o modelo caracterizado pelas hipóteses H_1-H_9.

No modelo analítico apresentado na Seção 3.3, nós desenvolvemos uma expressão fechada aproximada para a T.S.L. da distribuição do tempo de espera em fila. Assim, além de w_j, pode também ser obtido o segundo momento do tempo de espera, $w_j^{(2)}$, ou outro momento qualquer de ordem superior.

3.2 Trabalhos Anteriores

Desde que o nosso principal interesse é tratar o caso de filas cíclicas com múltiplos servidores ($s > 1$), nós comentamos brevemente a seguir as referências [16]–[24].

Na referência [16], é apresentado um estudo sobre a estrutura de sinalização interprocessadores da central telefônica de pequeno porte TRÓPICO-R, que atualmente está em operação comercial. A estrutura é do tipo *multibarramento com alocador centralizado operando em modo semiduplex e com repetição.* As solicitações de barramento para transmissão de mensagem são feitas a cada 4 ms, sob a escalação do relógio de tempo real (RTR). Em um instante de solicitação, a unidade de processamento mantém o pedido durante um certo intervalo de tempo (t_{sol}). Caso não seja atendida por um alocador nesse intervalo, passa a realizar outras tarefas e volta a tentar na próxima escalação do RTR. Uma vez conseguido o barramento, se a unidade de destino está livre, a mensagem é transmitida. Caso contrário, é feita uma nova solicitação de barramento na próxima escalação do RTR. São apresentados um modelo analítico e um modelo de simulação, que estimam a probabilidade de perda de mensagem em função de dois parâmetros: $p_1 \equiv$probabilidade que o tempo de espera pelo atendimento de um alocador supere t_{sol} e $p_2 \equiv$probabilidade de bloqueio da unidade de destino. O primeiro parâmetro é estimado aproximadamente no modelo analítico como a probabilidade que o tempo de espera em uma fila M/D/s supere t_{sol} (s denota o número de servidores). O segundo parâmetro é estimado aproximadamente como a proporção do tempo em regime estacionário no qual uma unidade esteja transmitindo ou recebendo

mensagens. Fixados os valores limites de tentativas de tomada de barramento e tentativas de transmissão, a probabilidade de perda é obtida através de um passeio aleatório em uma cadeia de Markov com barreiras absorventes. É apresentado um caso de estudo para um sistema assimétrico com 42 unidades de processamento, 2 servidores, chegadas Poissonianas e tempos de serviço e de caminhada constantes. O modelo analítico não se aplica ao nosso caso, pois a modelagem é orientada para a determinação da probabilidade de perda de mensagem. Desse modelo, nós aproveitamos apenas os conceitos de p_1 e p_2, a partir dos quais nós introduzimos no nosso atual modelo analítico o conceito de probabilidade de visita bem sucedida de um servidor a um nó, o qual será esclarecido na próxima seção.

Na referência [23], é apresentado um modelo de simulação para a avaliação de desempenho da estrutura de sinalização interprocessadores da central telefônica de grande porte TRÓPICO-RA. A estrutura é uma evolução da estrutura do TRÓPICO-R. Cada alocador controla um certo número de barramentos. Uma unidade de processamento acessa os alocadores através de circuitos dedicados de sinalização, não havendo mais a escalação periódica do *RTR*. As repetições são devidas apenas ao bloqueio da unidade de destino. O modelo de simulação estima os dois primeiros momentos do tempo de transferência de mensagens entre duas unidades. A probabilidade do tempo de transferência ultrapassar um dado valor fixado é obtida pelo ajuste de uma função gama incompleta. É apresentado um caso de estudo para um sistema assimétrico com 768 unidades de processamento, chegadas Poissonianas, tempo de serviço geral e tempo de caminhada constante. Desde que é feita uma abordagem apenas por simulação, esse modelo não se aplica ao nosso caso.

Morris e Wang [17] consideram filas cíclicas com disciplinas *exaustiva* e *limitada* e a possibilidade de *mais que um servidor estar sendo usado pelo mesmo nó em um dado instante*. Após o serviço ter terminado, *os usários deixam juntos o sistema*. O *tempo médio de ciclo de um servidor particular*, o *tempo médio intervisita de um servidor genérico* e a *condição de estabilidade* são deduzidos exatamente por meio de argumentos simples de conservação. O tempo médio de passagem dos usuários no sistema é estimado aproximadamente com base no cálculo do segundo momento do tempo intervisita, no caso de serviço com barreira, e na distribuição do tempo intervisita, no caso de serviço limitado. Neste último caso, o tempo intervisita é tomado como a superposição dos tempos de ciclo e o modelo é uma variação da fila de Bailey com serviço *bulk* [28],[29]. São apresentados resultados numéricos para sistemas simétricos e assimétricos com 7 filas e 3 servidores ou 7 servidores e 3 filas, no caso de tempo de serviço exponencial e tempo de caminhada constante. As aproximações são validadas por simulação, e são inaplicáveis, se o tempo de caminhada é muito pequeno ou se o sistema está próximo da saturação. Os autores ressaltam que a principal fonte de imprecisão vem do fato que os servidores tendem a se mover juntos de fila a fila. Em vista disso, eles propõem escalonamentos dispersivos, que tendem a melhorar a aproximação. Do modelo de Morris e Wang, nós utilizamos apenas o conceito de tempo médio de ciclo de

um servidor particular. As disciplinas de serviço e o enfoque de modelagem que utilizamos são diferentes.

O artigo de Raith [18] apresenta um modelo aproximado de um sistema de unidades de processamento que se *comunicam através de um número finito de buffers* de transmissão e recepção conectados a uma rede multibarramento. A disciplina de serviço é do tipo *limitado*. Cada unidade pode *acessar somente um barramento em um dado instante* para uma transmissão. As unidades podem *transmitir e receber simultaneamente*, mas *não podem receber de dois barramentos ao mesmo tempo*. O processo de *esvaziamento* de um *buffer* de recepção é *Poissoniano*. Quando ele está lotado, as mensagens não podem ser aceitas. Neste caso, o *modo de escalonamento* dos barramentos é do tipo *repetição* ou *espera*. A análise aproximada é baseada na solução numérica de uma cadeia de Markov imersa. São apresentados resultados para o tempo médio de espera, para a probabilidade de bloqueio devido à falta de espaço no *buffer* de transmissão e para a média e variância do tempo intervisita. São considerados sistemas simétricos com 8 filas, 2 servidores e 10 posições no *buffer* de transmissão, distribuição do tempo de serviço exponencial e hiperexponencial e tempo de caminhada constante. As aproximações são testadas por meio de comparações com resultados de simulação. A precisão do modelo melhora à medida que o tempo de caminhada aumenta e o coeficiente de variação do tempo de serviço diminui. O modelo de Raith não é adequado ao nosso caso, desde que ele considera *buffers* com capacidade finita. Isto dificulta a solução das equações de estado e não é representativo da situação que motivou nossas investigações, onde os sistemas dispõem de memórias de grande capacidade e as filas podem ser consideradas na prática com capacidades de armazenamento ilimitadas.

Kamal e Hamacher [19] apresentam um modelo aproximado para a estimação do tempo médio de espera de uma rede de área local em anel segmentado e com inserção de registro. Eles consideram *sistemas simétricos, servidores idênticos, serviço limitado e tempo de caminhada entre nós não nulo*. A análise se baseia na dedução de expressões aproximadas, relacionando três tipos de ciclos: *o ciclo do servidor*, definido como o tempo entre a partida de um servidor de um dado nó e o retorno do mesmo servidor a esse nó; *o ciclo do nó*, definido como o tempo entre a saída de um servidor de um nó e a chegada do próximo servidor a esse nó e o *ciclo nó-servidor*, definido como o tempo entre a saída de um servidor que visitou a fila quando ela não estava sendo servida e a chegada do mesmo servidor à mesma fila que não está novamente sendo servida por outro servidor. Um procedimento iterativo é usado para estimar os dois primeiros momentos do ciclo do nó. Os valores obtidos são então introduzidos na expressão do tempo médio de espera em uma fila M/G/1 com férias do servidor [32]. A precisão das aproximações é testada para os casos de 15 e 40 nós no anel, 3, 4, 6, 10, 12 e 16 quadros e tempos de transmissão e caminhada constantes. As estimativas obtidas pelo modelo analítico são interiores aos correspondentes intervalos de confiança obtidos por simulação, para cargas baixas e médias. A precisão das aproximações piora para cargas altas e melhora com o aumento do número de servidores. O modelo de Kamal

e Harnacher não é aplicável ao nosso caso, desde que ele trata somente sistemas simétricos, não considera o efeito do eventual bloqueio do nó de destino em uma transmissão de mensagem e utiliza um método de estimação dos primeiros dois momentos do tempo de ciclo muito particular ao caso de redes em anel segmentado e com inserção de registro, sendo, portanto, de difícil generalização.

Marsan *et alii* [20] apresentam um modelo aproximado de um sistema multi-filas *simétrico não exaustivo*, onde no *máximo um servidor* ou, alternativamente, *qualquer número de servidores* pode atender simultaneamente uma fila. Sob a hipótese de capacidade infinita de armazenamento dos *buffers* em cada nó, são obtidas várias equações relacionadas com a análise de ciclo dos servidores e a estabilidade das filas. A análise se baseia no conceito de *chegada utilizável* de servidor a um nó, definida como a chegada de um servidor que pode prover serviço ao usuário em espera, e na dedução de expressões aproximadas relacionando três tipos de ciclos: o *ciclo nó-servidor*, o *tempo de férias nó-servidor* e o *tempo de ciclo do nó*. O primeiro é definido como o tempo decorrido entre duas chegadas utilizáveis de um servidor particular ao nó. O segundo é definido como o tempo decorrido entre a saída de um servidor particular, após uma chegada utilizável, e a próxima chegada utilizável do mesmo servidor ao nó. Finalmente, o terceiro é definido como o intervalo de tempo entre duas chegadas utilizáveis consecutivas de qualquer servidor ao nó, não importando se o nó recebeu ou não serviço. Um procedimento aproximado é usado para estimar os dois primeiros momentos dos ciclos condicionais do nó (com e sem serviço). Os valores obtidos são então introduzidos na expressão do tempo médio de espera obtida por Kuehn [4] para o caso de servidor único. A precisão é testada para os casos de 2 servidores e o número de estações variando de 2 a 100, e 6 servidores e o número de estações entre 2 e 5. São usados tempos de serviço e de caminhada exponenciais e constantes. Comparando-se os resultados do modelo analítico com os correspondentes obtidos via simulação, os autores concluem que a precisão é aceitável na maioria dos casos. Este modelo não é suficiente para os nossos propósitos, desde que, por exemplo, ele trata somente sistemas simétricos e não leva em conta as possíveis retentativas de transmissão de um usuário devido ao bloqueio do nó de destino.

Marsan *et alii* [21] apresentam um modelo *exato* de um sistema de filas com capacidade finita servidas por s servidores cíclicos. A *distribuição do intervalo de tempo entre chegadas dos usuários, do tempo de serviço e de caminhada dos servidores é a exponencial negativa*, isto é, o sistema é Markoviano. Os usuários podem ser servidos por qualquer servidor. O *número máximo de servidores* usados simultaneamente por um nó se situa *entre 1 e s*. A disciplina de serviço considerada por instante de visita de um servidor a um nó é a *limitada*. O enfoque utilizado é o da aplicação de *redes de Petri estocásticas generalizadas* para descrever a operação do sistema sob diferentes conjuntos de hipóteses. Através dessa descrição, é deduzida numericamente a distribuição de probabilidade exata sobre o espaço de estados e, daí, são obtidos os parâmetros de desempenho de interesse, tais como os tempos médios de espera e de passagem dos usuários. O preço a

pagar para a resolução numérica, com complexidade em termos de tempo e espaço aceitável, são as restrições relativas a distribuições exponenciais e filas com capacidade finita e a dimensão do modelo. Essas restrições tornam o modelo inaplicável ao nosso problema. A conveniência dessa técnica de análise é ilustrada por meio da resolução numérica de alguns sistemas multifilas de pequeno porte.

Bhuyan *et alii* [22] apresentam um modelo analítico aproximado para a avaliação de tempos médios de espera e de passagem em *redes em anel único e múltiplo*, baseadas em protocolos do tipo passagem de permissão, segmentado e inserção de registro. Eles consideram *sistemas simétricos, filas infinitas e tempos de serviço e de caminhada dos servidores entre os nós constantes.* A disciplina de serviço é do tipo *limitado.* Por considerar somente sistemas simétricos e tempos de serviço e de caminhada constantes, o modelo é insuficiente para o nosso problema. As aproximações envolvidas na análise são devidas a certas hipóteses de independência que são adotadas e à hipótese de que os intervalos de tempo entre chegadas dos servidores às filas são uniformemente distribuídos. A precisão das aproximações é verificada por meio de comparação com resultados de simulação. Eles consideram um sistema com 50 filas e 1, 2 ou 4 servidores. Os erros das estimativas do tempo médio de transferência relativos às estimativas obtidas por simulação são menores que 10%. Com respeito à simulação, são previstas esperas maiores, sob condições de carga baixas, e esperas menores, perto da saturação.

Marsan *et alii* [24] apresentam um estudo sobre *sistemas multifilas simétricos, chegadas Poissonianas, tempos de serviço exponenciais e tempos de caminhada nulos.* Para o caso em que qualquer número de servidores pode atender simultaneamente uma mesma fila, são apresentados resultados exatos para o tempo médio de espera em fila. Para o caso em que apenas um único servidor pode atender uma fila em um dado instante, são apresentados limites inferiores e superiores e uma aproximação bem precisa para o valor médio do tempo de espera em fila. As restrições relativas a sistemas simétricos, tempos de serviço exponenciais e tempos de caminhada nulos tornam o modelo inaplicável à nossa situação.

Todos esses modelos possuem características que os tornam inaplicáveis ou insuficientes para o nosso caso. Na próxima seção, nós apresentamos o nosso modelo analítico aproximado [25]. Ele trata sistemas multifilas simétricos ou assimétricos com múltiplos servidores cíclicos, que visitam as filas de transmissão dos nós a fim de transportar usuários para as filas de recepção. As chegadas dos usuários são Poissonianas e os tempos de serviço e de caminhada são gerais. A assimetria dos sistemas pode ser caracterizada pelo processo de chegada, de caminhada ou de serviço ou, ainda, pela atratividade de tráfego entre as filas de transmissão e recepção. A disciplina de serviço é FIFO e não exaustiva e o modo de escalonamento é do tipo repetição. A transmissão é do tipo unidirecional, isto é, no máximo um servidor pode estar servindo um nó, seja para a transmissão ou para a recepção de um usuário. Os servidores originais são substituídos por um único servidor equivalente. Nós desenvolvemos uma expressão fechada para a T.S.L. da distribuição do tempo de espera em uma fila de transmissão. A correspondente expressão para o

tempo médio de espera em fila inclui explicitamente a probabilidade de visita bem sucedida de um servidor a um nó e os dois primeiros momentos do tempo de ciclo do servidor equivalente. Assim, o modelo pode ser generalizado de maneira direta para tratar uma ampla classe de sistemas multifilas com múltiplos servidores cíclicos, através de uma caracterização conveniente desses parâmetros.

3.3 Desenvolvimento do Modelo

Para uma análise exata dos sistemas de filas cíclicas que estamos considerando, a variável de estado em um instante t deveria ser representada por um vetor que descrevesse o número de usuários em espera em cada fila de transmissão, a localização dos servidores no ciclo e a duração do tempo de serviço ou de caminhada residual de cada servidor. Um enfoque exato parece ser atualmente inviável, desde que o processo de evolução de estado não é Markoviano e o espaço de estados não é enumerável. Por essas razões, desenvolvemos um modelo analítico aproximado, em cuja análise nós observamos uma fila marcada j, supomos que as outras filas estão em equilíbrio e seguimos o enfoque de cadeias de Markov imersas. O *estado do sistema* é caracterizado pelo *número de usuários* presentes nessa fila. A idéia básica consiste em agregar todos os servidores em um único *servidor equivalente* [20] e estender o modelo de Hashida e Ohara [3] relativo a servidor único em férias e serviço 1-limitado ao caso multi-servidor. Após servir um usuário, o servidor equivalente deixa a fila j por um período de tempo aleatório (férias). Para calcular este tempo, nós utilizamos dois métodos distintos. No primeiro, nós supomos que o servidor equivalente é um servidor com uma taxa de serviço e uma taxa de caminhada igual a s vezes as correspondentes taxas individuais dos servidores originais [20]. No segundo, nós supomos que o efeito das outras filas no tempo de espera da fila marcada é concentrado nos tempos de ciclo condicionais, de maneira análoga ao enfoque proposto por Kuehn [4].

A Disciplina de Operação

Consideremos uma fila de transmissão particular j. Após servir um usuário (serviço um-por-vez), o servidor equivalente deixa a fila j por um período de tempo aleatório chamado *tempo de ciclo equivalente* (férias) e denotado por C_{e_j}. Ao final deste período de tempo, ele visita novamente a fila j. Se a fila estiver vazia, ele retorna ao ciclo. Caso contrário, com uma certa probabilidade ρ_j, tal que $0 < \rho_j \leq 1$, ele serve o primeiro usuário da fila e, com probabilidade $1 - \rho_j$, ele retorna ao ciclo. A probabilidade ρ_j, denominada *probabilidade de visita bem sucedida ao nó j*, deve refletir a interferência entre os servidores originais.

O tempo de serviço e o tempo de caminhada do servidor equivalente em relação a uma fila i são denotados, respectivamente, por H_{e_i} e U_{e_i}, $i = 1, 2, \ldots, m$.

Tudo se passa, então, como se o servidor equivalente fosse um servidor com uma taxa de serviço e uma taxa de caminhada equivalentes àquelas dos s servidores originais, porém, ao visitar uma fila não vazia, às vezes ele não enxerga nenhum usuário nessa fila e, conseqüentemente, passa em branco.

As Equações de Transição de Estados

Sejam t_q o instante da $q-\acute{e}sima$ visita do servidor equivalente à fila j e $M_j^{(q)}(t_q)$ o número de usuários nessa fila no instante t_q. As probabilidades de estado da cadeia de Markov imersa são definidas como

$$P_{n,j}^{(q)} = P[M_j^{(q)} = n],\ n = 0, 1, 2, \ldots$$

No estado estacionário, nós temos que

$$P_{n,j} = P[M_j(t) = n] = \lim_{q \to \infty} P_{n,j}^{(q)},\ n = 0, 1, 2, \ldots$$

Agora, por facilidade de leitura, suprimiremos o índice j associado à fila de transmissão observada todas as vezes que estiver claro relativamente ao contexto.

Devido à propriedade de falta de memória do processo de chegada, o estado da fila considerada forma uma cadeia de Markov imersa no conjunto discreto dos instantes de visita (pontos de renovação). Se a fila não está vazia em um instante de visita, um usuário é servido com probabilidade ρ e não é servido com probabilidade $1 - \rho$. Denotando por P_{in} a probabilidade de transição do estado i para o estado n entre dois instantes de visita sucessivos do servidor equivalente a essa fila e por $\star$ o operador convolução, nós temos que para $i = 0$

$$P_{0n} = \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} dF_{C_e}(t)$$

e, para $i > 0$

$$P_{in} = \begin{cases} \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} e^{-\lambda t} dF_{C_e}(t) \star F_{H_e}(t), & \text{com probabilidade } \rho, \\ \int_0^\infty \frac{(\lambda t)^{n-i}}{(n-i)!} e^{-\lambda t} dF_{C_e}(t), & \text{com probabilidade } 1 - \rho. \end{cases}$$

Assim, as equações de transição de estados podem ser escritas como

$$P_n^{(q+1)} = P_0^{(q)} \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} dF_{C_e}(t) + \rho \sum_{i=1}^{n+1} P_i^{(q)} \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} e^{-\lambda t} dF_{C_e}(t) \star F_{H_e}(t)$$

$$+ (1 - \rho) \sum_{i=1}^{n} P_i^{(q)} \int_0^\infty \frac{(\lambda t)^{n-i}}{(n-i)!} e^{-\lambda t} dF_{C_e}(t) \cdot \qquad (3.1)$$

No estado estacionário, nós temos que

$$P_n = P_0 \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} dF_{C_e}(t) + \rho \sum_{i=1}^{n+1} P_i \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} e^{-\lambda t} dF_{C_e}(t) \star F_{H_e}(t)$$

$$+ (1-\rho) \sum_{i=1}^{n} P_i \int_0^\infty \frac{(\lambda t)^{n-i}}{(n-i)!} e^{-\lambda t} dF_{C_e}(t), \qquad (3.2)$$

onde $\sum_{n=0}^{\infty} P_n = 1$. Introduzindo a função geradora de probabilidade da distribuição de estado $\{P_n\}$, $G(x) = \sum_{n=0}^{\infty} P_n x^n$, nós obtemos

$$G(x) = \sum_{n=0}^{\infty} \left[P_0 \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF_{C_e}(t) \right] x^n + \rho \sum_{n=0}^{\infty} \left[\sum_{i=1}^{n+1} P_i \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} dF_{C_e}(t) \star F_{H_e}(t) \right] x^n$$

$$+ (1-\rho) \sum_{n=0}^{\infty} \left[\sum_{i=1}^{n} P_i \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n-i}}{(n-i)!} dF_{C_e}(t) \right] x^n.$$

Sejam, agora, $z = \lambda(1-x)$ e $G(x) = G_1(x) + G_2(x) + G_3(x)$, onde

a)

$$G_1(x) = \sum_{n=0}^{\infty} \left[P_0 \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF_{C_e}(t) \right] x^n$$

$$\Rightarrow \; G_1(x) = \sum_{n=0}^{\infty} P_0 \int_0^\infty e^{-\lambda t} \frac{(\lambda x t)^n}{n!} dF_{C_e}(t)$$

$$\Rightarrow \; G_1(x) = P_0 \int_0^\infty e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda x t)^n}{n!} dF_{C_e}(t)$$

$$\Rightarrow \; G_1(x) = P_0 \int_0^\infty e^{-\lambda(1-x)t} dF_{C_e}(t)$$

$$\Rightarrow \; G_1(x) = P_0 \phi_{C_e}[\lambda(1-x)]$$

$$\Rightarrow \; G_1(x) = P_0 \phi_{C_e}(z)$$

b)

$$G_2(x) = \rho \sum_{n=0}^{\infty} \left[\sum_{i=1}^{n+1} P_i \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} dF_{C_e}(t) \star F_{H_e}(t) \right] x^n$$

$$\Rightarrow \; G_2(x) = \rho \sum_{n=0}^{\infty} [\sum_{i=0}^{n+1} P_i \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} dF_{C_e}(t) \star F_{H_e}(t)$$

$$- P_0 \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n+1}}{(n+1)!} dF_{C_e}(t) \star F_{H_e}(t)] x^n.$$

Rearranjando a ordem das somatórias

$$\Rightarrow G_2(x) = \rho\{\sum_{i=0}^{\infty} P_i x^i \int_0^{\infty} \sum_{n-i+1=0}^{\infty} e^{-\lambda t}\frac{(\lambda x t)^{n-i+1}}{(n-i+1)!}\frac{1}{x} dF_{C_e}(t) \star F_{H_e}(t)$$

$$-P_0\left[\int_0^{\infty}\sum_{n+1=0}^{\infty} e^{-\lambda t}\frac{(\lambda x t)^{n+1}}{(n+1)!}\frac{1}{x} dF_{C_e}(t)\star F_{H_e}(t)\right]\}.$$

Fazendo as mudanças de variáveis: k=n-i+1, em relação à primeira somatória, e l=n+1, em relação à segunda somatória, resulta que

$$G_2(x) = \rho\{\frac{1}{x}G(x)\int_0^{\infty} e^{-\lambda t}\sum_{k=0}^{\infty}\frac{(\lambda x t)^k}{k!} dF_{C_e}(t)\star F_{H_e}(t)$$

$$-P_0\frac{1}{x}\int_0^{\infty} e^{-\lambda t}\sum_{l=0}^{\infty}\frac{(\lambda x t)^l}{l!} dF_{C_e}(t)\star F_{H_e}(t)\}$$

$$\Rightarrow G_2(x) = \rho\left\{\frac{1}{x}G(x)\int_0^{\infty} e^{-\lambda(1-x)t} dF_{C_e}(t)\star F_{H_e}(t) - \frac{1}{x}P_0\int_0^{\infty} e^{-\lambda(1-x)t} dF_{C_e}(t)\star F_{H_e}(t)\right\}$$

$$\Rightarrow G_2(x) = \rho\left[\frac{1}{x}G(x)\phi_{C_e}(z)\phi_{H_e}(z) - \frac{1}{x}P_0\phi_{C_e}(z)\phi_{H_e}(z)\right]$$

$$\Rightarrow G_2(x) = \frac{\rho}{x}\phi_{C_e}(z)\phi_{H_e}(z)\left[G(x) - P_0\right]$$

c)

$$G_3(x) = (1-\rho)\sum_{n=0}^{\infty}\left[\sum_{i=1}^{n} P_i\int_0^{\infty} e^{-\lambda t}\frac{(\lambda t)^{n-i}}{(n-i)!} dF_{C_e}(t)\right]x^n$$

$$\Rightarrow G_3(x) = (1-\rho)\sum_{n=0}^{\infty}[\sum_{i=0}^{n} P_i\int_0^{\infty} e^{-\lambda t}\frac{(\lambda t)^{n-i}}{(n-i)!} dF_{C_e}(t)$$

$$-P_0\int_0^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!} dF_{C_e}(t)]x^n$$

$$\Rightarrow G_3(x) = (1-\rho)[\sum_{i=0}^{\infty} P_i x^i\int_0^{\infty}\sum_{n-i=0}^{\infty} e^{-\lambda t}\frac{(\lambda x t)^{n-i}}{(n-i)!} dF_{C_e}(t)$$

$$-P_0\int_0^{\infty}\sum_{n=0}^{\infty} e^{-\lambda t}\frac{(\lambda x t)^n}{n!} dF_{C_e}(t)]$$

$$\Rightarrow G_3(x) = (1-\rho)\left[G(x)\phi_{C_e}(z) - P_0\phi_{C_e}(z)\right]$$

$$\Rightarrow G_3(x) = (1-\rho)\phi_{C_e}(z)\left[G(x) - P_0\right].$$

Portanto,

$$G(x) = P_0\phi_{C_e}(z) + \frac{\rho}{x}\phi_{C_e}(z)\phi_{H_e}(z)\left[G(x) - P_0\right] + (1-\rho)\phi_{C_e}(z)\left[G(x) - P_0\right]$$

e, após algumas manipulações algébricas, resulta que

$$G(x) = P_0\rho\frac{\phi_{C_e}(z)\left[x - \phi_{H_e}(z)\right]}{x\left[1-(1-\rho)\phi_{C_e}(z)\right] - \rho\phi_{C_e}(z)\phi_{H_e}(z)}. \qquad (3.3)$$

A Probabilidade de Fila Vazia e Estabilidade

O importante parâmetro P_0 é obtido da condição de normalização $\lim_{x\to 1} G(x) = 1$. Mas, desde que se $x \to 1 \Rightarrow z \to 0$ e $\phi_{C_e}(0) = \phi_{H_e}(0) = 1$, resulta que $\lim_{x\to 1} G(x)$ é da forma $\frac{0}{0}$. Aplicando, então, a regra de L'Hospital, nós temos que

$$\lim_{x\to 1} G(x) = P_0\rho \lim_{x\to 1} \frac{\frac{d}{dx}\left\{\phi_{C_e}(z)\left[x - \phi_{H_e}(z)\right]\right\}}{\frac{d}{dx}\left\{x\left[1-(1-\rho)\phi_{C_e}(z)\right] - \rho\phi_{C_e}(z)\phi_{H_e}(z)\right\}}$$

$$\Rightarrow P_0\rho \frac{\phi_{C_e}(z)\left[1-\phi'_{H_e}(z)\right] + \left[x-\phi_{H_e}(z)\right]\phi'_{C_e}(z)}{-x\left[(1-\rho)\phi'_{C_e}(z)\right] + 1 - (1-\rho)\phi_{C_e}(z) - \rho\phi_{C_e}(z)\phi'_{H_e}(z) - \rho\phi'_{C_e}(z)\phi_{H_e}(z)} = 1.$$

Agora, desde que $\phi'_{C_e}(0) = -c_e$, $\lim_{x\to 1}\phi'_{C_e}(z) = \lambda c_e$, $\phi'_{H_e}(0) = -h_e$ e $\lim_{x\to 1}\phi'_{H_e}(z) = \lambda h_e$ resulta, após algumas manipulações algébricas, que

$$P_0 = \frac{\rho(1-\lambda h_e) - \lambda c_e}{\rho(1-\lambda h_e)}. \tag{3.4}$$

Portanto, a condição de estabilidade de uma fila de transmissão é

$$0 < \frac{\lambda c_e}{1-\lambda h_e} < \rho \leq 1. \tag{3.5}$$

A partir de (3.4), pode ser verificado que a probabilidade do servidor equivalente encontrar uma fila não vazia é maior no caso em que $\rho < 1$ do que no caso em que $\rho = 1$, de acordo com a intuição.

Distribuição do Número de Usuários Deixados Para Trás Por Um Usuário de Partida

Seja Q_n a probabilidade que, em um ponto de conclusão de serviço, um usuário de partida da fila j deixa para trás n usuários dessa fila. Então, Q_n é dada por

$$Q_n = \sum_{i=1}^{n+1} \left[P_i|_{i>0}\right] \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} e^{-\lambda t} dF_{H_e}(t),$$

onde $P_i|_{i>0}$ denota a probabilidade condicional de haver i usuários na fila, dado que ela não está vazia. Mas $P_i|_{i>0} = \frac{P_i}{1-P_0}$, donde

$$Q_n = \sum_{i=1}^{n+1} \frac{P_i}{1-P_0} \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} dF_{H_e}(t).$$

Introduzindo a função geradora de probabilidade da distribuição $\{Q_n\}$, $R(x) = \sum_{n=0}^{\infty} Q_n x^n$, resulta que

$$R(x) = \sum_{n=0}^{\infty} \left[\sum_{i=1}^{n+1} \frac{P_i}{1-P_0} \int_0^{\infty} \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} dF_{H_e}(t) \right] x^n$$

$$\Rightarrow R(x) = \frac{1}{1-P_0} \sum_{n=0}^{\infty} [\sum_{i=0}^{n+1} P_i \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} dF_{H_e}(t)$$

$$-P_0 \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^{n+1}}{(n+1)!} dF_{H_e}(t)] x^n$$

$$\Rightarrow R(x) = \frac{1}{1-P_0} [\sum_{i=0}^{\infty} P_i x^i \int_0^{\infty} e^{-\lambda t} \sum_{n-i+1=0}^{\infty} \frac{(\lambda x t)^{n-i+1}}{(n-i+1)!} \frac{1}{x} dF_{H_e}(t)$$

$$-P_0 \int_0^{\infty} e^{-\lambda t} \sum_{n+1=0}^{\infty} \frac{(\lambda x t)^{n+1}}{(n+1)!} \frac{1}{x} dF_{H_e}(t)]$$

$$\Rightarrow R(x) = \frac{1}{1-P_0} \left[\frac{G(x)\phi_{H_e}(z)}{x} - P_0 \frac{1}{x} \phi_{H_e}(z) \right]$$

$$\Rightarrow R(x) = \frac{1}{1-P_0} \frac{G(x) - P_0}{x} \phi_{H_e}(z) \cdot \tag{3.6}$$

O Tempo Médio de Espera em Fila

Se $F_W(t)$ denota a distribuição do tempo de espera para a fila j, nós temos que [3],[32]

$$Q_n = \int_0^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} dF_W(t) \star F_{H_e}(t) \cdot$$

Como $R(x) = \sum_{n=0}^{\infty} Q_n x^n$, então

$$R(x) = \sum_{n=0}^{\infty} \left[\int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF_W(t) \star F_{H_e}(t) \right] x^n$$

$$\Rightarrow R(x) = \int_0^{\infty} \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda x t)^n}{n!} dF_W(t) \star F_{H_e}(t)$$

$$\Rightarrow R(x) = \int_0^{\infty} e^{-\lambda(1-x)t} dF_W(t) \star F_{H_e}(t)$$

$$\Rightarrow R(x) = \phi_W[\lambda(1-x)] \phi_{H_e}[\lambda(1-x)]$$

$$\Rightarrow R(x) = \phi_W(z) \phi_{H_e}(z). \tag{3.7}$$

Assim, de (3.6) e (3.7), resulta que

$$\phi_W(z) = \frac{1}{1-P_0}\frac{G(x)-P_0}{x}.$$

Substituindo $G(x)$ e P_0 dados, respectivamente, por (3.3) e (3.4), nesta expressão e usando a relação $\lambda x = \lambda - z$ nós obtemos, após algumas manipulações algébricas,

$$\phi_W(z) = \frac{\rho[1-\lambda h_e]-\lambda c_e}{c_e}\frac{1-\phi_{C_e}(z)}{z[1-(1-\rho)\phi_{C_e}(z)]-\lambda[1-(1-\rho)\phi_{C_e}(z)-\rho\phi_{C_e}(z)\phi_{H_e}(z)]}. \tag{3.8}$$

O tempo médio de espera em fila é $w = -\phi'_W(0)$. Para determinar $\lim_{z\to 0}\frac{d}{dz}\phi_W(z)$, nós temos de aplicar sucessivamente duas vezes a regra de L'Hospital, resultando, após algumas manipulações algébricas, que

$$w = \frac{c_e^{(2)}}{2c_e} + \frac{2(1-\rho)c_e + \lambda\left[c_e^{(2)} + \rho(h_e^{(2)} + 2c_e h_e)\right]}{2\left[\rho(1-\lambda h_e) - \lambda c_e\right]}. \tag{3.9}$$

Repare que para $\rho = 1$ e $s = 1$, nós temos que $\phi_{C_e}(z) = \phi_C(z)$, $c_e = c$, $c_e^{(2)} = c^{(2)}$, $\phi_{H_e}(z) = \phi_H(z)$, $h_e = h$, $h_e^{(2)} = h^{(2)}$ e, então,

$$G(x) = P_0\frac{\phi_C(z)[x-\phi_H(z)]}{x-\phi_C(z)\phi_H(z)},$$

$$P_0 = 1 - \frac{\lambda c}{1-\lambda h},$$

$$\phi_W(z) = \frac{1-\lambda(c+h)}{c}\frac{1-\phi_C(z)}{z-\lambda[1-\phi_C(z)\phi_H(z)]}$$

e

$$w = \frac{c^{(2)}}{2c} + \frac{\lambda(c^{(2)}+2ch+h^{(2)})}{2[1-\lambda(c+h)]},$$

ou seja, nós obtemos as mesmas expressões relativas ao caso de *servidor único em férias e serviço do tipo um-por-vez* [3],[19].

Para calcular w temos, portanto, que determinar os parâmetros ρ, h_e, $h_e^{(2)}$, c_e e $c_e^{(2)}$, de acordo com a expressão em (3.9).

A Probabilidade de Visita Bem Sucedida

A fim de calcular ρ, nós consideramos, inicialmente, que o tráfego oferecido por uma fila de transmissão j a um servidor marcado é igual a $\frac{\lambda_j h_j}{s}$, uma vez que o tráfego é equidistribuído entre os servidores [19]. Assim, a proporção do tempo em regime estacionário em que o correspondente nó está sendo usado em transmissão por outros $s-1$ servidores é igual a $\frac{s-1}{s}\lambda_j h_j$. Por outro lado, o tráfego oferecido

pelas outras $m-1$ filas de transmissão dos outros nós à fila de recepção do nó j através do servidor marcado é igual a $\frac{1}{s}\sum_{i=1,i\neq j}^{m}\alpha_{ij}\lambda_i h_i$. Assim, a proporção do tempo em regime estacionário que a fila de recepção do nó j está recebendo usuários das outras $m-1$ filas de transmissão através dos outros $s-1$ servidores é igual a $\frac{s-1}{s}\sum_{i=1,i\neq j}^{m}\alpha_{ij}\lambda_i hi$. A probabilidade do servidor encontrar o nó j livre, p_j, é tal que

H_{10}: p_j pode ser estimado, aproximadamente, pela proporção do tempo em regime estacionário em que o nó j não está usando outros $s-1$ servidores.

Assim, nós temos que [18]

$$p_j = 1 - \frac{s-1}{s}\lambda_j h_j - s\frac{s-1}{s}\sum_{i=1,i\neq j}^{m}\alpha_{ij}\lambda_i h_i$$

$$\Rightarrow\ p_j = 1-(s-1)\left[\frac{\lambda_j h_j}{s}+\sum_{i=1,i\neq j}^{m}\alpha_{ij}\lambda_i h_i\right] \tag{3.10}$$

e

$$\rho_j = p_j\sum_{k=1,k\neq j}^{m}\alpha_{jk}p_k. \tag{3.11}$$

Donde

$$\rho_j = \left\{1-(s-1)[\frac{\lambda_j h_j}{s}+\sum_{i=1,i\neq j}^{m}\alpha_{ij}\lambda_i h_i]\right\}\cdot\left\{\sum_{k=1,k\neq j}^{m}\alpha_{jk}\left\{1-(s-1)[\frac{\lambda_k h_k}{s}\sum_{i=1,i\neq k}^{m}\alpha_{ik}\lambda_k h_k]\right\}\right\}. \tag{3.12}$$

Em um sistema simétrico, nós temos que

$$p_j = p = 1-\frac{s^2-1}{s}\lambda h$$

e

$$\rho_j = \rho = \left[1-\frac{s^2-1}{s}\lambda h\right]^2.$$

Análise do Tempo de Serviço e do Tempo de Ciclo do Servidor Equivalente

Nós descrevemos a seguir os métodos de obtenção dos dois primeiros momentos da distribuição do tempo de serviço e do tempo de caminhada do servidor equivalente, h_e, $h_e^{(2)}$, c_e e $c_e^{(2)}$, respectivamente.

a) Método da Equivalência Entre Taxas de Serviço e Caminhada

Considerando o tempo de serviço e o tempo de caminhada do servidor equivalente em relação a uma fila i, $i = 1, 2, \ldots, m$, nós supomos que

$$\mathrm{H}_{11}: \quad H_{e_i} = \tfrac{1}{s} H_i$$

e

$$\mathrm{H}_{12}: \quad U_{e_i} = \tfrac{1}{s} U_i \ ,$$

donde

$$h_{e_i} = \frac{1}{s} h_i \ , \tag{3.13}$$

$$h_{e_i}^{(2)} = \frac{1}{s^2} h_i^{(2)} \ , \tag{3.14}$$

$$u_{e_i} = \frac{1}{s} u_i$$

e

$$u_{e_i}^{(2)} = \frac{1}{s^2} u_i^{(2)}.$$

A fim de determinar os dois primeiros momentos da distribuição do tempo de ciclo equivalente, c_e e $c_e^{(2)}$, respectivamente, definimos a variável aleatória E_i como o intervalo de tempo decorrido entre os instantes de visita a uma fila i e à fila $(i \bmod m) + 1$ pelo servidor equivalente, isto é, o segmento de tempo de ciclo do servidor equivalente associado à fila i. A T.S.L. da distribuição de E_i é dada por

$$\phi_{E_i}(\xi) = \phi_{U_{e_i}}(\xi) \left[(1 - P_{0_i}) \rho_i \phi_{H_{e_i}}(\xi) + P_{0_i} + (1 - P_{0_i})(1 - \rho_i) \right].$$

Nós supomos que

$$\mathrm{H}_{13}: \quad E_i \ i = 1, 2, \ldots, m, \text{ são independentes.}$$

Assim, a T.S.L. da distribuição do tempo de ciclo equivalente em relação à fila particular j pode ser escrita como

$$\phi_{C_{e_j}}(\xi) = \phi_{U_{e_j}}(\xi) \prod_{i=1, i \neq j}^{m} \phi_{E_i}(\xi).$$

Conseqüentemente,

$$\phi_{C_{e_j}}(\xi) = \prod_{i=1}^{m} \phi_{U_{e_i}}(\xi) \prod_{i=1, i \neq j}^{m} \left[P_{0_i} + (1 - P_{0_i})(1 - \rho_i) + (1 - P_{0_i}) \rho_i \phi_{H_{e_i}}(\xi) \right] \tag{3.15}$$

e, então, nós temos que

$$c_{e_j} = \frac{1}{s} \left[\sum_{i=1}^{m} u_i + \sum_{i=1, i \neq j}^{m} (1 - P_{0_i}) \rho_i h_i \right] \tag{3.16}$$

e

$$c_{e_j}^{(2)} = c_{e_j}^2 + \frac{1}{s^2}\left\{\sum_{i=1}^{m} VAR[U_i] + \sum_{i=1, i\neq j}^{m}\left[(1-P_{0_i})\rho_i h_i^{(2)} - (1-P_{0_i})^2\rho_i^2 h_i^2\right]\right\}. \quad (3.17)$$

De (3.4) e (3.16), nós obtemos o sistema

$$\rho_j(s-\lambda_j h_j)P_{0_j} - \lambda_j \sum_{i=1, i\neq j}^{m} P_{0_i}\rho_i h_i = s\rho_j - \lambda_j\left[\sum_{i=1}^{m}(u_i+\rho_i h_i)\right], \quad (3.18)$$

$$j = 1, 2, \ldots, m,$$

cuja solução é

$$P_{0_j} = \frac{\rho_j(s-\sum_{i=1}^{m}\lambda_i h_i) - \lambda_j\sum_{i=1}^{m}u_i}{\rho_j(s-\sum_{i=1}^{m}\lambda_i h_i)}, \quad j = 1, 2, \ldots, m, \quad (3.19)$$

o que pode ser verificado por substituição em (3.18). Em um sistema simétrico, nós temos que para $j = 1, 2, \ldots, m$,

$$P_{0_j} = P_0 = \frac{\rho(s-m\lambda h) - m\lambda u}{\rho(s-m\lambda h)}.$$

De posse de P_0, calculamos c_e e $c_e^{(2)}$ a partir de (3.16) e (3.17), respectivamente. Resulta, então, que

$$c_e = \frac{1}{s}\left[\sum_{i=1}^{m} u_i + \sum_{i=1, i\neq j}^{m}\frac{\lambda_i h_i\sum_{k=1}^{m}u_k}{s-\sum_{k=1}^{m}\lambda_k h_k}\right] \quad (3.20)$$

e

$$c_e^{(2)} = c_e^2 + \frac{1}{s^2}\left\{\sum_{i=1}^{m} VAR[U_i] + \sum_{i=1, i\neq j}^{m}\left[\frac{\lambda_i\sum_{k=1}^{m}u_k\left[h_i^{(2)}(s-\sum_{k=1}^{m}\lambda_k h_k) + \lambda_i h_i^2\sum_{k=1}^{m}u_k\right]}{\left(s-\sum_{k=1}^{m}\lambda_k h_k\right)^2}\right]\right\}. \quad (3.21)$$

Portanto, para calcular o tempo médio de espera em fila, nós calculamos ρ, P_0, h_e, $h_e^{(2)}$, c_e e $c_e^{(2)}$ através de (3.12), (3.19),(3.13),(3.14),(3.20) e (3.21), respectivamente, e, então, determinamos w por (3.9), para cada fila j, $j = 1, 2, \ldots, m$.

Nota: Análise do Tempo Entre Duas Visitas Bem Sucedidas

Consideremos a variável aleatória V_j^b, que denota o intervalo de tempo entre duas visitas bem sucedidas do servidor equivalente à fila j. Com respeito ao comportamento do servidor equivalente, temos um processo semi-Markoviano com dois estados: 1⇒serviço equivalente e 2⇒ciclo equivalente. Ao visitar o estado 1, o servidor equivalente nele permanece por um tempo H_e. Com probabilidade 1,

vai ao estado 2. Nesse estado, permanece por C_e unidades de tempo e então vai ao estado 1, com probabilidade $(1 - P_0)\rho$, ou permanece no mesmo estado, com probabilidade $1-(1-P_0)\rho$, por mais C_e unidades de tempo. V^b é o tempo decorrido entre duas visitas sucessivas ao estado 1.

A T.S.L. da distribuição de V^b é

$$\phi_{V^b}(\xi) = \phi_{H_e}(\xi)\{(1 - P_0)\rho\phi_{C_e}(\xi) + [1 - (1 - P_0)\rho](1 - P_0)\rho\phi^2_{C_e}(\xi)$$

$$+[1 - (1 - P_0)\rho]^2(1 - P_0)\rho\phi^3_{C_e}(\xi) + \ldots\},$$

ou seja,

$$\phi_{V^b}(\xi) = (1 - P_0)\rho\frac{\phi_{H_e}(\xi)\phi_{C_e}(\xi)}{1 - [1 - (1 - P_0)\rho]\,\phi_{C_e}(\xi)} \tag{3.22}$$

e, portanto, os dois primeiros momentos de V^b são dados por

$$v^b = h_e + \frac{c_e}{(1 - P_0)\rho} \tag{3.23}$$

e

$$v^{b(2)} = \frac{[(1 - P_0)\rho]^2 h_e^{(2)} + (1 - P_0)\rho(c_e^{(2)} + 2c_e h_e - 2c_e^2) + 2c_e^2}{[(1 - P_0)\rho]^2}\,. \tag{3.24}$$

Repare que se calcularmos v^b diretamente do diagrama de transição de estado, nós temos que

$$v^b = h_e + (1 - P_0)\rho c_e + 2c_e[1 - (1 - P_0)\rho](1 - P_0)\rho + 3c_e[1 - (1 - P_0)\rho]^2(1 - P_0)\rho + \ldots$$

$$\Rightarrow v^b = h_e + \frac{c_e}{(1 - P_0)\rho}\,,$$

ou seja, obtemos o mesmo resultado, conforme era esperado.

b) Método da Superposição dos Tempos de Ciclo Condicionais

Consideremos uma fila de transmissão particular j. Nós definimos o *tempo de ciclo de um servidor marcado*, C_j, e o *tempo intervisita de um servidor genérico*, V_j, como o intervalo de tempo decorrido entre duas visitas sucessivas à fila j, respectivamente, pelo servidor marcado e por servidores quaisquer. Nós definimos, também, as seguintes variáveis aleatórias:

$C'_j(V'_j)$: *tempo de ciclo (intervisita)* condicionado a uma visita mal sucedida no instante da visita anterior à fila j;

$C''_j(V''_j)$: *tempo de ciclo (intervisita)* condicionado a uma visita bem sucedida no instante da visita anterior à fila j.

Temos, então, a seguinte correspondência:

$$V_j^{'} = C_{e_j}$$

e

$$V_j^{''} = C_{e_j} + H_{e_j},$$

onde C_{e_j} e H_{e_j} denotam, respectivamente, o tempo de ciclo e o tempo de serviço do servidor equivalente utilizados na abordagem *a*).

Nós supomos que

$$\mathrm{H}_{14}: \quad C_{e_j} \text{ e } H_{e_j} \text{ são independentes.}$$

Donde $\phi_{V'}(z) = \phi_{C_e}(z)$ e $\phi_{V''}(z) = \phi_{C_e}(z)\phi_{H_e}(z)$. Resulta, então, que os dois primeiros momentos dos tempos intervisitas condicionais dos tipos (') e (") de um servidor genérico são dados por

$$v^{'} = c_e \ ,$$
$$v^{'(2)} = c_e^{(2)} \ ,$$
$$v^{''} = c_e + h_e$$

e

$$v^{''(2)} = c_e^{(2)} + 2c_e h_e + h_e^{(2)}.$$

Portanto, a partir de (3.3),(3.4),(3.5),(3.8) e (3.9), nós obtemos os seguintes resultados:

1. a função geradora de probabilidade da distribuição de estado $\{P_n,\ n = 0,1,2,\ldots\}$ fica sendo

$$G(x) = P_0\rho\frac{x\phi_{V'}(z) - \phi_{V''}(z)}{x\left[1-(1-\rho)\phi_{V'}(z)\right] - \rho\phi_{V''}(z)}\ . \tag{3.25}$$

2. a probabilidade de fila vazia passa a ser dada por

$$P_0 = \frac{\rho\left[1-\lambda(v^{''}-v^{'})\right] - \lambda v^{'}}{\rho\left[1-\lambda(v^{''}-v^{'})\right]} \tag{3.26}$$

e, conseqüentemente, a condição de estabilidade de uma fila de transmissão fica sendo

$$0 < \frac{\lambda v^{'}}{1-\lambda(v^{''}-v^{'})} < \rho \leq 1. \tag{3.27}$$

O número esperado de usuários na fila de transmissão no instante de visita segue de $E[M] = \lim_{x\to 1}\frac{dG(x)}{dx}$, o que resulta em

$$E[M] = P_0\rho\lambda\frac{\lambda v^{'(2)}(1-\lambda v^{''}) + v^{'}(\lambda^2 v^{''(2)} + 2 - 2\lambda v^{''})}{2\left[\rho(1-\lambda v^{''}) - \lambda v^{'}(1-\rho)\right]^2}\ . \tag{3.28}$$

3. a T.S.L. da distribuição do tempo de espera na fila j passa a ser dada por

$$\phi_W(z) = \frac{\left\{\rho\left[1-\lambda(v''-v')\right]-\lambda v'\right\}\left[1-\phi_{V'}(z)\right]}{v'\left\{z\left[1-(1-\rho)\phi_{V'}(z)\right]-\lambda\left[1-(1-\rho)\phi_{V'}(z)-\rho\phi_{V''}(z)\right]\right\}}\,. \tag{3.29}$$

4. o tempo médio de espera em fila passa a ser dado por

$$w = \frac{v'^{(2)}}{2v'} + \frac{2(1-\rho)v' + \lambda\left[v'^{(2)} + \rho(v''^{(2)} - v'^{(2)})\right]}{2\left\{\rho\left[1-\lambda(v''-v')\right]-\lambda v'\right\}}\,. \tag{3.30}$$

Para calcular w temos, então, que determinar os parâmetros ρ, v', $v'^{(2)}$, v'' e $v''^{(2)}$, de acordo com a expressão em (3.30). O parâmetro ρ é estimado através de (3.12) e o procedimento de cálculo dos demais parâmetros será descrito posteriormente.

É interessante notar que, para $\rho = 1$ e $s = 1$, nós temos que $v' = c'$, $v'' = c''$, $v'^{(2)} = c'^{(2)}$, $v''^{(2)} = c''^{(2)}$ e, então,

$$G(x) = P_0\frac{x\phi_{C'}(z) - \phi_{C''}(z)}{x - \phi_{C''}(z)}\,,$$

$$P_0 = \frac{1-\lambda c''}{1-\lambda(c''-c')}\,,$$

$$E[M] = P_0\lambda\frac{\lambda c'^{(2)}(1-\lambda c'') + c'(\lambda^2 c''^{(2)} + 2 - 2\lambda c'')}{2(1-\lambda c'')^2}$$

e

$$w = \frac{c'^{(2)}}{2c'} + \frac{\lambda c''^{(2)}}{2(1-\lambda c'')},$$

ou seja, obtemos as mesmas expressões para $G(x)$, P_0, $E[M]$ e w que foram obtidas por Kuehn [4] para o caso de sistemas com múltiplas filas e servidor único.

Notas:

1. As equações (3.25),...,(3.30) podem também ser obtidas considerando que as probabilidades de transição do estado i ao estado n da fila, entre dois instantes de visita sucessivos de servidores quaisquer, são dadas por

$$P_{0n} = \int_0^\infty \frac{(\lambda t)^n}{n!}e^{-\lambda t}dF_{V'}(t) \quad \text{e}$$

$$P_{in} = \begin{cases} \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!}e^{-\lambda t}dF_{V''}(t), & \text{com probabilidade } \rho, \\ \int_0^\infty \frac{(\lambda t)^{n-i}}{(n-i)!}e^{-\lambda t}dF_{V'}(t), & \text{com probabilidade } 1-\rho, \\ i > 0 \end{cases}$$

e, no estado estacionário, as equações de transição de estados podem ser escritas como

$$P_n = P_0 \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} dF_{V'}(t) + \rho \sum_{i=1}^{n+1} P_i \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} e^{-\lambda t} dF_{V''}(t)$$

$$+(1-\rho) \sum_{i=1}^{n} P_i \int_0^\infty \frac{(\lambda t)^{n-i}}{(n-i)!} e^{-\lambda t} dF_{V'}(t).$$

2. Alternativamente, a fim de obter o tempo médio de espera de um usuário na fila de transmissão, w, nós podemos calcular a distribuição estacionária do número M^* de usuários na fila considerada, em um instante de observação arbitrário, $\{P_n^*,\ n = 0, 1, 2, \ldots\}$. Sejam π_s', π_c' e π'' as probabilidades de um observador encontrar um ciclo do tipo (') com fila de transmissão vazia, encontrar um ciclo do tipo (') com fila de transmissão não vazia e encontrar um ciclo do tipo ("), respectivamente. Denotando o tempo médio intervisita incondicional, v, por

$$v = [P_0 + (1 - P_0)(1 - \rho)]\, v' + (1 - P_0)\rho v'' \tag{3.31}$$

pode ser verificado que

$$\pi_s' = \frac{P_0 v'}{v},$$

$$\pi_c' = \frac{(1 - P_0)(1 - \rho) v'}{v}$$

e

$$\pi'' = \frac{(1 - P_0)\rho v''}{v}.$$

Note que a identidade

$$\rho(1 - P_0) = \lambda v$$

pode ser mostrada utilizando-se (3.26) e (3.31), ou seja, *o número médio de usuários servidos é igual ao número médio de usuários que chegam em um ciclo*, em um sistema em equilíbrio estacionário. As probabilidades de chegada de n usuários durante os tempos de recorrência para trás, V_r' e V_r'', definidos como o intervalo de tempo decorrido desde o último instante de visita até o ponto de observação arbitrário nos casos (') e ("), respectivamente, são dadas por

$$\int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_{V_r'}(t) dt$$

e

$$\int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_{V_r''}(t) dt, \quad n = 0, 1, 2, \ldots$$

onde

$$f_{V_r'}(t) = \frac{1 - F_{V'}(t)}{v'}$$

e

$$f_{V_r''}(t) = \frac{1 - F_{V''}(t)}{v''}.$$

Considerando os dois tipos de tempos de ciclo condicionais, as probabilidades de estado avaliadas em um instante arbitrário podem ser escritas como

$$P_n^* = \pi_s' \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_{V_r'}(t)dt + \pi_c' \sum_{i=1}^{n} P_i^* \int_0^\infty \frac{(\lambda t)^{n-i}}{(n-i)!} e^{-\lambda t} f_{V_r'}(t)dt$$

$$+\pi'' \sum_{i=1}^{n+1} P_i^* \int_0^\infty \frac{(\lambda t)^{n-i+1}}{(n-i+1)!} e^{-\lambda t} f_{V_r''}(t)dt \qquad n = 0, 1, 2, \ldots$$

A função geradora de $\{P_n^*,\ n = 0, 1, 2, \ldots\}$ é $G^*(x) = \sum_{n=0}^{\infty} P_n^* x^n$. Usando o fato que $P_i^* = \frac{P_i}{1-P_0}$, nós obtemos, após algumas manipulações algébricas,

$$G^*(x) = \frac{x\left[1 - \phi_{V'}(z)\right]\left[(1-\rho)P_0 G(x) + \rho P_0\right] + \rho\left[1 - \phi_{V''}(z)\right]\left[G(x) - P_0\right]}{vxz}.$$

O número esperado de usuários na fila de transmissão em um instante arbitrário resulta de $E[M^*] = \lim_{x\to 1} \frac{dG^*(x)}{dx}$ e, então,

$$E[M^*] = \frac{2v'\{E[M]\left[\rho + (1-\rho)P_0\right] + \rho(P_0 - 1)\} + \lambda\left[v'^{(2)} + \rho(1-P_0)v''^{(2)}\right]}{2v}.$$

Finalmente, pela lei de Little, o tempo médio de espera em fila é dado por $w = \frac{E[M^*]}{\lambda}$, donde

$$w = \frac{2v'\left\{E[M]\left[\rho + (1-\rho)P_0\right] + \rho(P_0 - 1)\right\} + \lambda\left[v'^{(2)} + \rho(1-P_0)v''^{(2)}\right]}{2\lambda v}.$$

Substituindo $E[M]$ e P_0 dados, respectivamente, por (3.28) e (3.26), nesta expressão, nós podemos obter o mesmo resultado dado por (3.30). Note que este método de cálculo de w é mais complicado que o anterior. Portanto, ele deve ser utilizado somente quando se estiver interessado na obtenção de $E[M^*]$.

Análise dos Tempos de Ciclo e Intervisitas Condicionais

A fim de determinar os parâmetros v', $v'^{(2)}$, v'' e $v''^{(2)}$, nós definimos a variável aleatória E_i como o intervalo de tempo decorrido entre os instantes de visita a uma fila i e à fila $(i \bmod m) + 1$ por um servidor marcado, isto é, o segmento de tempo de ciclo de um servidor particular associado à fila i. A T.S.L. da distribuição de E_i é dada por

$$\phi_{E_i}(\xi) = \phi_{U_i}(\xi)\left[(1 - P_{0_i})\rho_i \phi_{H_i}(\xi) + P_{0_i} + (1 - P_{0_i})(1 - \rho_i)\right].$$

Nós supomos que

$$H_{15}: \quad E_i,\ i = 1, 2, \ldots, m, \text{ são independentes.}$$

Assim, as T.S.L.'s das distribuições dos tempos de ciclo condicionais para um servidor marcado em relação à fila particular j podem ser escritas como

$$\phi_{C_j'}(\xi) = \phi_{U_j}(\xi) \prod_{i=1, i\neq j}^{m} \phi_{E_i}(\xi)$$

e

$$\phi_{C_j''}(\xi) = \phi_{U_j}(\xi)\phi_{H_j}(\xi) \prod_{i=1, i\neq j}^{m} \phi_{E_i}(\xi).$$

Conseqüentemente,

$$\phi_{C_j'}(\xi) = \prod_{i=1}^{m} \phi_{U_i}(\xi) \prod_{i=1, i\neq j}^{m} \left[(1 - P_{0_i}')\rho_i\phi_{H_i}(\xi) + P_{0_i}' + (1 - P_{0_i}')(1 - \rho_i)\right] \tag{3.32}$$

e

$$\phi_{C_j''}(\xi) = \prod_{i=1}^{m} \phi_{U_i}(\xi) \prod_{i=1, i\neq j}^{m} \left[(1 - P_{0_i}'')\rho_i\phi_{H_i}(\xi) + P_{0_i}'' + (1 - P_{0_i}'')(1 - \rho_i)\right] \phi_{H_j}(\xi). \tag{3.33}$$

P_{0_i}' e P_{0_i}'' são obtidos de (3.26), resultando em

$$P_{0_i}' = \frac{\rho_i \left[1 - \lambda_i(v_i'' - v_i')\right] - \lambda_i v_i'}{\rho_i \left[1 - \lambda_i(v_i'' - v_i')\right]}$$

e, supondo que

$$H_{16}: \quad v_i' = v_i'' = v_j' ,$$

temos que $P_{0_i}' = \frac{\rho_i - \lambda_i v_j'}{\rho_i}$. Analogamente,

$$P_{0_i}'' = \frac{\rho_i \left[1 - \lambda_i(v_i'' - v_i')\right] - \lambda_i v_i'}{\rho_i \left[1 - \lambda_i(v_i'' - v_i')\right]}$$

e supomos, neste caso, que

$$H_{17}: \quad v_i' = v_i'' = v_j'' ,$$

o que resulta em $P_{0_i}'' = \frac{\rho_i - \lambda_i v_j''}{\rho_i}$. De (3.32) e (3.33), nós obtemos

$$c_j' = \sum_{i=1}^{m} u_i + \sum_{i=1, i\neq j}^{m} \lambda_i v_j' h_i, \tag{3.34}$$

$$c_j^{'(2)} = \sum_{i=1}^{m} VAR[U_i] + \sum_{i=1, i\neq j}^{m} \lambda_i v_j^{'} \left\{ VAR[H_i] + h_i^2(1 - \lambda_i v_j^{'}) \right\} + c_j^{'2}, \tag{3.35}$$

$$c_j^{''} = \sum_{i=1}^{m} u_i + h_j + \sum_{i=1, i\neq j}^{m} \lambda_i v_j^{''} h_i \tag{3.36}$$

e

$$c_j^{''(2)} = \sum_{i=1}^{m} VAR[U_i] + VAR[H_j] + \sum_{i=1, i\neq j}^{m} \lambda_i v_j^{''} \left\{ VAR[H_i] + h_i^2(1 - \lambda_i v_j^{''}) \right\} + c_j^{''2}. \tag{3.37}$$

A fim de determinar $v^{'}$ e $v^{'(2)}$, nós supomos que

H_{18}: o tempo intervisita condicional $V^{'}$ corresponde à superposição de $s > 1$ tempos de ciclo condicionais $C^{'}$, recorrentes e independentes.

Por outro lado, a fim de determinar $v^{''}$ e $v^{''(2)}$, nós definimos o tempo de ciclo condicional efetivo $C_r^{''}$ como

$$C_r^{''} = C^{''} + \sum_{i=0}^{\kappa} C_i^{'} ,$$

onde κ é uma variável aleatória inteira associada ao número de visitas mal sucedidas do servidor marcado ao nó da fila de transmissão j [19]. A caracterização da distribuição desta variável é muito difícil, devido à forte dependência com o estado de outros servidores e outros nós. Nós supomos, aproximadamente, que:

H_{19}: κ é distribuída geometricamente com média igual a ρ^{-1}

e

H_{20}: as variáveis aleatórias $C^{''}$ e $\sum_{i=0}^{\kappa} C_i^{'}$ são independentes.

Conseqüentemente, a T.S.L. de $F_{C_r^{''}}$ é dada por

$$\phi_{C_r^{''}}(\xi) = \rho \frac{\phi_{C^{''}}(\xi)}{1 - (1 - \rho)\phi_{C^{'}}(\xi)}$$

e os dois primeiros momentos são dados, respectivamente, por

$$c_r^{''} = \frac{1 - \rho}{\rho} c^{'} + c^{''} \tag{3.38}$$

e

$$c_r^{''(2)} = \frac{\rho^2 c^{''(2)} + \rho(1 - \rho) c^{'(2)} + 2\rho(1 - \rho) c^{'} c^{''} + 2(1 - \rho)^2 c^{'2}}{\rho^2}. \tag{3.39}$$

Note que se $\rho = 1$, nós temos que $c_r^{''} = c^{''}$ e $c_r^{''(2)} = c^{''(2)}$.

Finalmente, nós supomos que

$\mathbf{H_{21}}$: V'' corresponde à superposição de $s > 1$ tempos de ciclo condicionais efetivos recorrentes e independentes C_r''.

A superposição pode ser realizada utilizando-se um dos dois procedimentos que descrevemos a seguir.

b_1): Procedimento de Superposição de Kuehn

A superposição pode ser realizada recursivamente, como proposto por Kuehn [26], em $s-1$ passos, considerando dois processos de cada vez e, então, utilizando os correspondentes processos substitutos hipo ou hiperexponenciais, se o coeficiente de variação é menor que 1 ou maior ou igual a 1, respectivamente. A seguir, nós apresentamos apenas uma visão geral desse procedimento. Para uma compreensão mais ampla, há que ser necessariamente consultada a referência [26], na qual são mostrados todos os cálculos detalhadamente.

Dados dois processos recorrentes estacionários genéricos com intervalos de tempo interchegadas aleatórios Y_1 e Y_2 e funções de distribuição $F_{Y_1}(t)$ e $F_{Y_2}(t)$, respectivamente, se ambos forem superpostos em um ponto, a função de distribuição do processo resultante é dada por $F_Y(t) = P[Y \leq t]$, onde Y é o intervalo de tempo interchegada do processo superposto. Nós temos que:

- as funções de distribuição complementares de Y_i são dadas por

$$F_{Y_i}^c(t) = 1 - F_{Y_i}(t) \ ,$$

- $\lambda_i = \frac{1}{y_i}$: taxa média de chegada e $cv_i = \sqrt{\frac{y_i^{(2)}}{y_i^2} - 1}$: coeficiente de variação do processo i, $i = 1, 2$ e

- $\lambda = \frac{1}{y}$: taxa média de chegada e $cv = \sqrt{\frac{y^{(2)}}{y^2} - 1}$: coeficiente de variação do processo superposto.

Pode ser mostrado que

$$F_Y(t) = 1 - \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \left\{ F_{Y_1}^c(t) \int_{\xi=t}^{\infty} F_{Y_2}^c(\xi) d\xi + F_{Y_2}^c(t) \int_{\xi=t}^{\infty} F_{Y_1}^c(\xi) d\xi \right\}.$$

Infelizmente, não estamos aptos para calcular os momentos $E[Y^k]$, exceto no caso $k = 1$, onde obtemos o resultado intuitivamente plausível $\lambda = \lambda_1 + \lambda_2$. O segundo momento $E[Y^2]$ e seu coeficiente de variação podem ser calculados usando os processos substitutos hipo e hiperexponenciais, de acordo com as equações

$$F_{Y_i}(t) = \begin{cases} 0 & 0 \leq t \leq t_{i1} \\ 1 - exp\{-\epsilon_{i2}(t - t_{i1})\} & t \geq t_{i1} \\ & 0 \leq cv_i \leq 1 \\ 1 - p_{i1} exp(-\epsilon_{i1} t) - p_{i2} exp(-\epsilon_{i2} t) & \\ & cv_i \geq 1 \end{cases}$$

onde

$$E[Y_{i\nu}] = t_{i\nu} = \frac{1}{\epsilon_{i\nu}} \quad \nu = 1,2$$

$$0 \leq cv_i \leq 1 \ : \ \epsilon_{i1} = \frac{\lambda_i}{1 - cv_i}, \ \epsilon_{i2} = \frac{\lambda_i}{cv_i}$$

$$cv_i \geq 1 \ : \ \epsilon_{i1,2} = \lambda_i \left\{ 1 \stackrel{+}{-} \sqrt{\frac{cv_i^2 - 1}{cv_i^2 + 1}} \right\}$$

$$p_{i1,2} = \frac{\epsilon_{i1,2}}{2\lambda_i}$$

$$(p_{i1}t_{i1} = p_{i2}t_{i2}).$$

Aplicando o procedimento no nosso caso, para obter V' e V'', respectivamente, nós superpomos s processos C' e s processos C''_r, resultando que $v' = \frac{c'}{s}$,$v'' = \frac{c''_r}{s}$ e $v'^{(2)}$ e $v''^{(2)}$ são calculados numericamente em função do coeficiente de variação dos processos superpostos nos $s - 1$ passos.

Nós chamamos este procedimento de *procedimento de superposição de Kuehn.*

b_2): Procedimento de Superposição de Bhuyan

Alternativamente, a superposição pode ser realizada generalizando o enfoque dado em [22]. Assim, nós supomos que

H_{22}: os $s - 1$ servidores remanescentes estão distribuídos uniformemente entre os outros $m - 1$ nós, no instante em que um servidor visita uma fila de transmissão particular.

Denotemos, sem perda de generalidade, por s o número do servidor marcado. Assim, em um ciclo do tipo ('), nós temos que V' — o tempo após o qual a fila j recebe a visita do próximo servidor — é dado por

$$V' = \min \left\{ C'_1, C'_2, \ldots, C'_{s-1} \right\} ,$$

onde C'_i é o tempo de ciclo do tipo (') do servidor i em relação à fila j, $i = 1, 2, \ldots, s - 1$. Donde

$$P[V' > t] = P[C'_1 > t, C'_2 > t, \ldots, C'_{s-1} > t].$$

Assumindo que

H_{23}: as variáveis aleatórias C'_i são independentes e identicamente distribuídas,

resulta que

$$P[V' > t] = \prod_{i=1}^{s-1} P[C_i' > t] = \left\{ \int_t^{\infty} f_{C'}(\tau) d\tau \right\}^{s-1}$$

$$\Rightarrow \; F_{V'}(t) = 1 - \left\{ 1 - \int_0^t f_{C'}(\tau) d\tau \right\}^{s-1} .$$

Em termos de unidades de tempo, o valor médio do maior valor C_i' se aproxima de c'. Daí, assumimos simplificadamente que

$\mathbf{H_{24}}$: C_i' é distribuída uniformemente entre 0 e c'

e, então,

$$F_{C_i'}(t) = \frac{t}{c'} \; , \; 0 \leq t < c' \quad i = 1, 2, \ldots, s-1.$$

Donde

$$F_{V'}(t) = 1 - (1 - \frac{t}{c'})^{s-1} \; , \quad 0 \leq t < c'. \tag{3.40}$$

Seguindo raciocínio análogo, ou seja, assumindo que:

$\mathbf{H_{25}}$: as variáveis aleatórias $C_{r,i}''$ são independentes e identicamente distribuídas

e

$\mathbf{H_{26}}$: $C_{r,i}''$ é distribuída uniformemente entre 0 e c_r'' ,

nós temos que

$$F_{C_{r,i}''}(t) = \frac{t}{c_r''} \; , \; 0 \leq t < c_r'' \quad i = 1, 2, \ldots, s-1$$

e, neste caso, podemos mostrar que

$$F_{V''}(t) = 1 - (1 - \frac{t}{c_r''})^{s-1} \; , \quad 0 \leq t < c_r''. \tag{3.41}$$

A partir de (3.40) e (3.41), nós obtemos, então,

$$v' = \frac{c'}{s} \; , \tag{3.42}$$

$$v'^{(2)} = \frac{2c'^2}{s(s+1)} \; , \tag{3.43}$$

$$v'' = \frac{c_r''}{s} \tag{3.44}$$

e

$$v''^{(2)} = \frac{2c_r''^2}{s(s+1)} \; . \tag{3.45}$$

Nós chamamos este procedimento de *procedimento de superposição de Bhuyan.*

UNICAMP
BIBLIOTECA CENTRAL

Cálculo dos Dois Primeiros Momentos dos Tempos Intervisitas Condicionais

Finalmente, a fim de calcular os primeiros dois momentos dos tempos intervisitas condicionais, nós procedemos como segue. De (3.42) e (3.34), nós obtemos

$$c^{'} = \frac{\sum_{i=1}^{m} u_i}{1 - \frac{1}{s}\sum_{i=1, i\neq j}^{m} \lambda_i h_i} \tag{3.46}$$

e

$$v^{'} = \frac{\sum_{i=1}^{m} u_i}{s - \sum_{i=1, i\neq j}^{m} \lambda_i h_i} . \tag{3.47}$$

De (3.46),(3.47) e (3.35), nós obtemos $c^{'(2)}$. Desde que $v^{''} = \frac{c_r^{''}}{s}$, isto resulta em

$$v^{''} = \frac{(1-\rho)c^{'} + s\rho c^{''}}{s\rho} . \tag{3.48}$$

De (3.48) e (3.36), nós obtemos

$$c^{''} = \frac{\sum_{i=1}^{m} u_i + h + \frac{(1-\rho)c^{'}}{s\rho}\sum_{i=1, i\neq j}^{m} \lambda_i h_i}{1 - \frac{1}{s}\sum_{i=1, i\neq j}^{m} \lambda_i h_i} . \tag{3.49}$$

De (3.48),(3.49) e (3.37), nós obtemos $c^{''(2)}$. Com os valores de $c^{'}$, $c^{'(2)}$, $c_r^{''}$ e $c_r^{''(2)}$ disponíveis, nós obtemos $v^{'(2)}$ e $v^{''(2)}$ utilizando o procedimento de superposição de Kuehn ou, alternativamente, o procedimento de Bhuyan. Neste último caso, eles são diretamente obtidos de (3.43) e (3.45).

O tempo de ciclo incondicional de um servidor marcado, c, é dado exatamente por [17]

$$c = \frac{\sum_{i=1}^{m} u_i}{1 - \frac{1}{s}\sum_{i=1}^{m} \lambda_i h_i} . \tag{3.50}$$

Desde que $[P_0 + (1-P_0)(1-\rho)]c^{'} + (1-P_0)\rho c^{''} = c$, nós temos, então, os seguintes passos :

1. Calcular $c^{'}$ de (3.46).
2. Calcular $v^{'}$, assumindo que $v^{'} = \frac{c^{'}}{s}$.
3. Calcular $c^{'(2)}$, $v^{''}$, $c^{''}$ e $c^{''(2)}$ a partir de (3.35),(3.48),(3.49) e (3.37), respectivamente.
4. Recalcular $c^{'}$, usando que

$$c^{'} = \frac{c - (1-P_0)\rho c^{''}}{P_0 + (1-P_0)(1-\rho)} . \tag{3.51}$$

5. Se o valor previamente calculado de c' e aquele obtido do passo 4 diferem por uma quantidade maior que um valor especificado ϵ, onde ϵ é um valor positivo bem pequeno que determina a precisão do procedimento, voltar ao passo 2. Caso contrário, parar e calcular $v'^{(2)}$ e $v''^{(2)}$. Nós usamos $\epsilon = 10^{-4}$ em nossas experiências e observamos que este procedimento iterativo é convergente.

Com λ, ρ, v', $v'^{(2)}$, v'' e $v''^{(2)}$ disponíveis, nós obtemos P_0, $E[M]$ e o tempo médio de espera na fila w de (3.26),(3.28) e (3.30), respectivamente.

No Apêndice B, nós apresentamos um manual de utilização do programa de computador do modelo analítico. Esse programa inclui algumas extensões do modelo, que são discutidas na próxima seção.

Nós apresentamos na Tabela 3.1, onde temos que $\overline{\rho} = 1 - \rho$, um quadro resumo dos principais resultados fornecidos pelo modelo analítico. Note que o método da superposição dos ciclos condicionais é obtido a partir do método da equivalência entre taxas de serviço e caminhada, considerando que:

- o tempo intervisita de um servidor genérico condicionado a uma visita mal sucedida é equivalente ao tempo de ciclo do servidor equivalente

e

- o tempo intervisita condicionado a uma visita bem sucedida corresponde à soma do tempo de ciclo com o tempo de serviço do servidor equivalente.

3.4 Extensões

As extensões do modelo analítico aproximado aqui discutidas se reportam à alteração de algumas hipóteses operacionais H_1-H_9 da Seção 3.1.

1. Se a acessibilidade dos servidores a um nó é total, ou seja, se o número máximo de servidores que podem ser simultaneamente usados por um nó é igual a s, nós temos que $\rho_j = 1$, $j = 1, 2, \ldots, m$. O modelo pode, então, ser usado diretamente e ele representa, por exemplo:

   - uma rede em barramento múltiplo com alocador centralizado ou com passagem de permissão, onde cada unidade de processamento acessa os planos de sinalização através de uma porta associada a cada barramento;
   - uma rede em anel múltiplo com passagem de permissão ou segmentado, onde cada unidade acessa os planos através de uma porta associada a cada anel;
   - uma rede em anel múltiplo com inserção de registro, onde há um RTX associado a cada anel

Tabela 3.1:

Principais Resultados do Modelo Analítico

Resultado	**Método**	
	Equivalência Entre Taxas	*Superposição dos Ciclos Condicionais*
$G(z)$	$P_0\rho\frac{\phi_{C_e}(z)\left[z-\phi_{H_e}(z)\right]}{z\left[1-\overline{\rho}\phi_{C_e}(z)\right]-\rho\phi_{C_e}(z)\phi_{H_e}(z)}$	$P_0\rho\frac{z\phi_{V'}(z)-\phi_{V''}(z)}{z\left[1-\overline{\rho}\phi_{V'}(z)\right]-\rho\phi_{V''}(z)}$
P_0	$\frac{\rho(1-\lambda h_e)-\lambda c_e}{\rho(1-\lambda h_e)}$	$\frac{\rho\left[1-\lambda(v''-v')\right]-\lambda v'}{\rho\left[1-\lambda(v''-v')\right]}$
Estabilidade	$0<\frac{\lambda c_e}{1-\lambda h_e}<\rho\leq 1$	$0<\frac{\lambda v'}{1-\lambda\left(v''-v'\right)}<\rho\leq 1$
$\phi_W(z)$	$\frac{\{\rho[1-\lambda h_e]-\lambda c_e\}\left[1-\phi_{C_e}(z)\right]}{c_e\left\{z\left[1-\overline{\rho}\phi_{C_e}(z)\right]-\lambda\left[1-\overline{\rho}\phi_{C_e}(z)-\rho\phi_{C_e}(z)\phi_{H_e}(z)\right]\right\}}$	$\frac{\left\{\rho\left[1-\lambda(v''-v')\right]-\lambda v'\right\}\left[1-\phi_{V'}(z)\right]}{v'\left\{z\left[1-\overline{\rho}\phi_{V'}(z)\right]-\lambda\left[1-\overline{\rho}\phi_{V'}(z)-\rho\phi_{V''}(z)\right]\right\}}$
w	$\frac{c_e^{(2)}}{2c_e}+\frac{2\overline{\rho}c_e+\lambda\left[c_e^{(2)}+\rho\left(h_e^{(2)}+2c_eh_e\right)\right]}{2[\rho(1-\lambda h_e)-\lambda c_e]}$	$\frac{v'^{(2)}}{2v'}+\frac{2\overline{\rho}v'+\lambda\left[v'^{(2)}+\rho\left(v''^{(2)}-v'^{(2)}\right)\right]}{2\left\{\rho\left[1-\lambda(v''-v')\right]-\lambda v'\right\}}$

e, em cada porta, independentemente do tipo de estrutura particular em questão, há um *buffer* de transmissão e um *buffer* de recepção com capacidades de armazenamento consideradas ilimitadas para propósitos práticos.

2. Para uma rede em barramento múltiplo operando em modo *full duplex*, ou seja, uma rede onde uma porta pode ter seu *buffer* de transmissão usando um barramento e seu *buffer* de recepção usando outro barramento simultaneamente, nós temos que

$$p_j = 1 - \frac{s-1}{s}\lambda_j h_j.$$

Basta, então, calcular o parâmetro ρ_j em função deste novo p_j e utilizar o modelo, desde que as demais expressões não se alteram. Nós temos, então, que se o escalonamento é do tipo repetição

$$\rho_j = p_j \sum_{k=1,k\neq j}^{m} \alpha_{jk}\left[1 - \frac{s-1}{s}\sum_{i=1,i\neq k}^{m}\alpha_{ik}\lambda_k h_k\right]$$

e, em um sistema simétrico,

$$\rho_j = \rho = \left[1 - \frac{s-1}{s}\lambda h\right]^2.$$

Essa situação modela, também, uma rede em anel múltiplo onde é considerado o bloqueio da unidade de recepção. Nesse caso, quando a mensagem retorna pelo anel, deve haver uma indicação de que ela não foi recebida pela unidade de recepção, o anel é liberado e a porta tenta a transmissão da mensagem novamente por qualquer um dos anéis.

3. Em uma rede em anel múltiplo onde não há o efeito de bloqueio da unidade de destino de uma mensagem, que ocorre na rede em barramento múltiplo com alocador centralizado, nós temos que

$$\rho_j = p_j = 1 - \frac{s-1}{s}\lambda_j h_j$$

e, no caso simétrico,

$$\rho_j = \rho = 1 - \frac{s-1}{s}\lambda h.$$

Nesse caso, o modelo se aplica diretamente, se forem mantidas as demais hipóteses operacionais e, se a rede é do tipo passagem de permissão ou segmentado, há uma única porta associada aos anéis e, se a rede é do tipo inserção de registro, há um único RTX associado aos anéis.

Nota: Nós temos, então, que o tipo de acessibilidade dos servidores a um nó caracteriza mais um fator de assimetria de um sistema de filas cíclicas. Por exemplo, um sistema onde a acessibilidade dos servidores a alguns nós é parcial e a outros é total é assimétrico, mesmo que os processos de chegada, caminhada, serviço e atratividade sejam simétricos.

4. O modelo pode ser generalizado de maneira direta para tratar o caso de uma rede em barramento múltiplo com alocador centralizado e filas com capacidade finita. Consideremos, por exemplo, o modelo analítico com a aplicação do método da superposição dos ciclos condicionais. Denotando por tx_j e rx_j os limites de capacidade das filas de transmissão e recepção, respectivamente, $j = 1, 2, \ldots, m$, as equações de estado em regime estacionário nos instantes de visita passam a ser dadas por

$$P_n = P_0 b'_n + (1-\rho)\sum_{i=1}^{n} P_i b'_{n-i} + \rho \sum_{i=1}^{n+1} P_i b''_{n-i+1},\ n = 0, 1, \ldots, tx-1$$

$$P_{tx} = P_0 \sum_{i=tx}^{\infty} b'_i + (1-\rho)\sum_{i=1}^{tx} P_i \sum_{f=tx-i}^{\infty} b'_f + \rho \sum_{i=1}^{tx} P_i \sum_{f=tx-i+1}^{\infty} b''_f$$

onde

$$b'_l = \int_0^{\infty} \frac{(\lambda t)^l}{l!} e^{-\lambda t} dF_{V'}(t)$$

e

$$b''_l = \int_0^{\infty} \frac{(\lambda t)^l}{l!} e^{-\lambda t} dF_{V''}(t)\ ,\quad l = 0, 1, 2, \ldots$$

Estas equações têm que ser resolvidas numericamente, desde que não é mais aplicável o método das transformações de domínio. Por seu turno, o parâmetro ρ deve incorporar as probabilidades de bloqueio da fila de transmissão e de recepção associadas, respectivamente, aos limites tx_j e rx_j. A primeira é a própria P_{tx_j}, enquanto que a segunda pode ser aproximada, por exemplo, pela probabilidade de bloqueio de uma fila $GI/M/1-rx_j$, ou seja, uma fila com servidor único, onde os processos de chegada são gerais e independentes, os processos de serviço são Poissonianos e a capacidade da fila é limitada em rx_j. Nós temos, então, que

$$p_j = 1-(s-1)\left[\frac{\lambda_j h_j}{s}(1-P_{tx_j}) + \sum_{i=1,i\neq j}^{m} \alpha_{ij}\lambda_i h_i (1-P_{rx_i})\right],$$

no caso semiduplex, e

$$p_j = 1-\frac{s-1}{s}\lambda_j h_j (1-P_{tx_j})(1-P_{rx_j}),$$

no caso *full duplex*. O parâmetro ρ_j é calculado em função deste novo p_j. As extensões desse modelo para tratar outras estruturas de interligação são feitas de maneira análoga ao caso em que as filas têm capacidades infinitas.

5. Consideremos uma rede em barramento múltiplo com alocador centralizado e modo de escalonamento do tipo espera. Nesse caso, $G(x)$, P_0, $E[M]$, $G^*(x)$ e $E[M^*]$ são dadas pelas mesmas expressões obtidas no modo repetição, desde que as equações de estado são as mesmas. Entretanto, a T.S.L. do segmento de tempo de ciclo de um servidor marcado correspondente à fila j fica sendo

$$\phi_E(\xi) = \phi_U(\xi)\left\{(1-P_0)\left[\sum_{k=1,k\neq j}^{m}\alpha_{jk}\left[(1-\rho_k)\phi^r_{H,k}(\xi)+\rho_k\right]\phi_H(\xi)\right] + P_0\right\},$$

onde $\phi^r_{H,k}$ é a T.S.L. da distribuição do tempo de espera pela liberação da parte recepção do nó k, $k=1,2,\ldots,m$, $k\neq j$. Sob a hipótese aproximada de independência entre os segmentos de tempo de ciclo, temos, então, que

$$\phi_{C'}(\xi) = \prod_{i=1}^{m}\phi_{U_i}(\xi)\prod_{i=1,i\neq j}^{m}\left\{(1-P_{0_i})[\sum_{k=1,k\neq i}^{m}\alpha_{ik}[(1-\rho_k)\phi^r_{H_k}(\xi)+\rho_k]\phi_{H_i}(\xi)+P_{0_i}\right\}$$

e

$$\phi_{C''}(\xi) = \prod_{i=1}^{m}\phi_{U_i}(\xi)[\sum_{k=1,k\neq j}^{m}\alpha_{jk}[(1-\rho_k)\phi^r_{H_k}(\xi)+\rho_k]\prod_{i=1,i\neq j}^{m}\{(1-P_{0_i})$$

$$[\sum_{k=1,k\neq i}^{m}\alpha_{ik}[(1-\rho_k)\phi^r_{H_k}(\xi)+\rho_k]\phi_{H_i}(\xi)]+P_{0_i}\}.$$

Raith [18] considerou que o tempo de espera pela liberação da parte recepção é igual ao tempo de recorrência à frente do tempo de serviço, o que é verdade no caso em que $s = 2$ servidores. Para $s > 2$, há a possibilidade de formação de uma fila de espera pela liberação da parte recepção, cujo comprimento máximo é igual a $s - 1$. A determinação da distribuição do tempo de espera não é trivial e depende da disciplina de serviço. Nós temos aqui uma questão em aberto. Raith e Tran-Gia [8] obtiveram aproximadamente o tempo de espera pela liberação da parte recepção de um sistema com servidor único e fila de recepção limitada. Temos também que

$$p_j = 1 - \frac{s-1}{s}(\lambda_j h_j + \sum_{i=1, i\neq j}^{m} \alpha_{ij}\lambda_i h_i)$$

e

$$\rho_j = p_j \sum_{k=1, k\neq j}^{m} \alpha_{jk} p_k \ ,$$

no modo semiduplex. Assim, se for aplicado o método da superposição dos ciclos condicionais, após a obtenção de v', $v'^{(2)}$, v'' e $v''^{(2)}$, nós podemos, então, aplicar o modelo analítico ou a sua extensão para o caso de filas limitadas, para obter os resultados de interesse no sistema onde os servidores operam no modo espera.

Repare que esse modelo analítico com múltiplos servidores operando em modo espera se aplica também a uma rede em anel múltiplo na qual é considerado o efeito de bloqueio da unidade de recepção. Nesse caso, quando a mensagem retorna pelo anel e a porta da unidade de transmissão verifica que ela não foi recebida pela unidade de recepção, o anel é mantido ocupado e a porta tenta a transmissão da mensagem novamente pelo mesmo anel.

As extensões do modelo para representar outros tipos de estruturas que eventualmente possam ser modeladas por sistemas multifilas com múltiplos servidores cíclicos procedem de maneira análoga aos casos anteriores.

Capítulo 4

O MODELO DE SIMULAÇÃO

A fim de adquirir uma visão crítica mais apurada sobre o modelo analítico aproximado e analisar sistemas de filas cíclicas ainda não considerados na literatura, nós desenvolvemos um modelo de simulação estocástico, dinâmico e a eventos discretos [37]. Ele permite a obtenção dos seguintes parâmetros de uma fila particular j: probabilidade de fila vazia, probabilidade de visita bem sucedida, primeiro e segundo momentos do tempo de ciclo de um servidor marcado, do tempo intervisita de servidores quaisquer e do tempo de espera.

A simulação modela um sistema não Markoviano em tempo contínuo. O correspondente programa de computador, cujo manual de utilização é apresentado no Apêndice C, foi escrito em linguagem SIMULA [30]. Essa linguagem foi escolhida por satisfazer requisitos ligados à portabilidade, generalidade, simplicidade, compacidade e estruturação do modelo. Através de uma caracterização conveniente dos dados de entrada desse programa, podem ser simulados todos os sistemas de filas cíclicas considerados na nossa abordagem, assim como outros sistemas ainda não tratados analiticamente.

O modelo de simulação é caracterizado pelas mesmas hipóteses operacionais H_1-H_9 da Seção 3.1, com algumas modificações descritas a seguir:

- as filas de transmissão e recepção dos m nós podem ser de capacidade finita ou infinita;
- o modo de escalonamento de um servidor pode ser do tipo repetição ou do tipo espera;
- após o serviço, um usuário entra na fila de recepção, onde ele aguarda ser atendido por um elemento esvaziador. Após esse atendimento, o usuário deixa o sistema simulado;
- os processos de caminhada, de serviço e de esvaziamento das filas de recepção são do tipo Erlang-r, r inteiro positivo. Os casos particulares $r = 1$ e

$r = \infty$ correspondem, respectivamente, ao processo Poissoniano e ao processo constante;

- a disciplina de atendimento de um usuário da fila de transmissão e a disciplina de esvaziamento da fila de recepção podem ser do tipo exaustivo, limitado ou com barreira;

- os usuários podem ser transmitidos das filas de transmissão às filas de recepção dos nós no modo semiduplex ou no modo *full duplex*.

Sem perda de generalidade, essas hipóteses caracterizam o modelo de simulação de uma estrutura de interligação de processadores com barramentos ou anéis múltiplos, onde:

- cada unidade acessa os barramentos ou anéis, conforme for o caso, através de uma única porta;

- os *buffers* de transmissão e recepção das portas, que podem ser considerados com capacidades de armazenamento finitas ou infinitas, operam no modo semiduplex ou no modo *full duplex*;

- a disciplina de transmissão de mensagem e a disciplina de esvaziamento da fila de recepção são na ordem de chegada e a transmissão e o esvaziamento podem ser do tipo exaustivo, limitado ou com barreira;

- o modo de escalonamento em caso de bloqueio da unidade de recepção de uma mensagem pode ser do tipo espera ou repetição.

Entretanto, esse modelo pode, eventualmente, ser estendido para tratar outros tipos de estruturas de interligação de processadores ou os mesmos tipos considerados na nossa abordagem, mas que tenham esquemas de acesso ao canal diferentes.

As variáveis aleatórias são geradas a partir de sementes fornecidas como dados de entrada. Os elementos esvaziadores permitem refletir no modelo a capacidade de processamento das unidades relativa aos *buffers* de recepção de mensagens.

Neste capítulo, nós utilizamos a mesma notação que foi introduzida no capítulo anterior. Na Seção 4.1, nós apresentamos as entidades do modelo de simulação. Na Seção 4.2, nós mostramos, com o auxílio dos diagramas de transição de estados, como as entidades se relacionam para reproduzir no modelo as mesmas características dos sistemas de filas cíclicas em consideração. Na Seção 4.3, nós mostramos como são tratados estatisticamente os resultados de uma rodada de simulação, e descrevemos o procedimento que deve ser adotado para refletir no modelo de simulação as mesmas características do modelo analítico. Finalmente, possíveis extensões do modelo são discutidas na Seção 4.4.

4.1 Entidades do Modelo

Nós identificamos no modelo de simulação as seguintes entidades: *sistema simulado, nó, usuário, gerador de usuários, servidor cíclico, temporizador de liberação de caminho e esvaziador da fila de recepção.* Ele representa um sistema com $m \geq 2$ nós, cada nó com uma fila de transmissão e uma fila de recepção, servidos por s servidores cíclicos idênticos, tal que $1 \leq s < m$. Cada servidor controla um certo número $l_s \geq 1$ de caminhos de comunicação, que ligam as filas de transmissão com as filas de recepção dos nós. O caso em que $l_s > 1$ corresponde a uma estrutura em barramento com alocador centralizado, onde cada alocador controla, eventualmente, mais do que um enlace de comunicação, sendo que cada enlace pode ser composto, por exemplo, por duas vias de comunicação multiplexadas no tempo.

Denominamos genericamente *operador* o analista e usuário do programa de simulação. Neste sentido, o sistema simulado faz a interface do modelo com o operador, sob o ponto de vista de obtenção dos dados de entrada, manutenção da base de dados, controle das entidades e apresentação dos resultados durante uma rodada de simulação. O início de uma rodada é desencadeado por um pedido do operador. O fim de uma rodada ocorre após ter sido simulado um número predeterminado de usuários ou, alternativamente, por um pedido de supressão da simulação, feito pelo operador em qualquer instante da rodada. Um usuário é considerado simulado quando sai da fila de recepção.

Os resultados de uma rodada de simulação são relativos a um nó e um servidor particulares, denominados *marcados para estatística*, conforme estipulado pelo operador. Eles são apresentados em relatórios periódicos, de modo a permitir o acompanhamento da evolução no tempo das respostas do modelo. Através da inspeção desses relatórios, o operador pode, inclusive, verificar a estabilidade das filas.

Após receber um pedido de início de uma rodada, o sistema simulado obtém os dados de entrada (vide Apêndice C), que são subdivididos em:

dados de configuração: caracterizam o sistema sob o ponto de vista de identificação e dimensionamento das entidades;

dados de tráfego: caracterizam o sistema sob o ponto de vista dos parâmetros numéricos relativos a taxas, valores médios, variâncias e probabilidades de encaminhamento;

dados de controle: caracterizam o sistema sob o ponto de vista de determinação das disciplinas de serviço, tipos dos processos e término de uma rodada.

Se os dados de entrada forem infactíveis, a rodada não é iniciada. Caso contrário, o sistema simulado constrói a base de dados, cria e ativa as entidades que estão sob seu controle direto e permanece passivamente aguardando os

instantes de elaboração dos relatórios de resultados. Esses instantes ocorrem de acordo com um certo número predeterminado de usuários a serem simulados por relatório periódico. Quando uma entidade é ativada, ela se torna apta a iniciar as suas ações.

As entidades controladas diretamente pelo sistema simulado são os m nós, os m esvaziadores das filas de recepção, os s servidores cíclicos e o gerador de usuários. Elas interagem no decorrer de uma rodada, evoluindo nos seus respectivos conjuntos de estados discretos, de modo a refletir no modelo o procedimento que permite o transporte dos usuários das filas de transmissão às filas de recepção.

Cada nó do modelo de simulação representa um nó do sistema, e controla, durante uma rodada, uma fila de transmissão e uma fila de recepção. O nó marcado para estatística calcula em uma rodada os valores dos primeiros dois momentos do tempo de ciclo do servidor marcado para estatística e do tempo intervisita de servidores quaisquer.

O gerador simula o processo de aparição dos usuários no modelo. Os usuários chegam na fila de transmissão do $j-$ *ésimo* nó segundo um processo Poissoniano com taxa média λ_j e se destinam à fila de recepção do nó k com probabilidade α_{jk}, $j,k = 1,2,\ldots,m$. Seja $\lambda_T = \sum_{j=1}^{m} \lambda_j$ a taxa média global de chegada de usuários ($\lambda_T > 0$). O gerador obtém intervalos de tempo entre geração de usuários, G, de acordo com a função densidade de probabilidade

$$f_G(t) = \begin{cases} \lambda_T e^{-\lambda_T t}, & t > 0, \\ 0, & \text{caso contrário.} \end{cases}$$

Em um instante de geração de usuário, ele escolhe o número da fila de transmissão, J, o número da fila de recepção, K, e o tempo de ocupação de caminho (serviço), H, segundo as distribuições

$$P[J = j] = \frac{\lambda_j}{\lambda_T},$$

$$P[K = k/J] = \alpha_{Jk}$$

e

$$f_H(t) = \begin{cases} \frac{h^{-r_h}}{\Gamma(r_h)} t^{r_h - 1} e^{\frac{-t}{h}}, & t > 0, \\ 0, & \text{caso contrário,} \end{cases}$$

onde r_h denota o parâmetro de escala da distribuição Erlang. É óbvio que $P[K = J/J] = \alpha_{JJ} = 0$. Em seguida, cria e ativa um usuário com os atributos J, K e H. Após G unidades de tempo, ele repete o procedimento de escolha.

Ao ser ativado pelo gerador, o usuário entra na fila de transmissão do nó J, se ela é de capacidade infinita ou é de capacidade finita, mas não está lotada. Se a fila é de capacidade finita e está lotada, ele incrementa em uma unidade os contadores relativos a usuários da fila J e usuários totais perdidos por fila lotada e deixa o sistema simulado. Os quocientes entre esses contadores e os respectivos contadores relativos a usuários da fila J e totais gerados são os estimadores de máxima verosimilhança das probabilidades de bloqueio das filas de transmissão

[36]. Na fila de transmissão, o usuário aguarda o atendimento de um servidor. Ao receber de um dos servidores cíclicos o aviso de que o nó K está livre, ele sai da fila J, ocupa o caminho por H unidades de tempo e entra na fila de recepção K, onde permanece até que seja avisado pelo esvaziador correspondente. Se ela é de capacidade infinita, o usuário é avisado no mesmo instante de entrada. Ao sair dessa fila, ele avisa o sistema simulado que deve ser elaborado um relatório, se for o caso, e então deixa o sistema. Cada usuário simulado obtém, em relação a um nó marcado para estatística, os tempos de permanência na fila de transmissão e na fila de recepção.

Os servidores cíclicos visitam as filas de transmissão dos nós de acordo com o mesmo escalonamento descrito no modelo analítico. Quanto à ordem de atendimento, a disciplina é do tipo FIFO. Quanto ao número de usuários atendidos por visita, o modo pode ser exaustivo, com barreira, limitado exaustivo ou limitado com barreira, se $l_s = 1$. Se $l_s > 1$, a operação é apenas no modo 1-limitado. Ao visitar um nó onde há um usuário na fila de transmissão apto a ser servido, o servidor obtém do usuário o número do nó de destino, K, e o tempo de ocupação de caminho, H. Após um tempo de comunicação usuário-servidor, C, suposto constante e que pode ser nulo, inclusive, o servidor verifica se o nó de destino está apto para recepção. Em caso positivo, informa ao usuário que ele pode sair da fila de transmissão, avisa o nó de destino que ele vai receber um usuário, ocupa um de seus l_s caminhos e cria e ativa um temporizador de liberação de caminho, que mantém o caminho ocupado durante H unidades de tempo. Se todos os seus caminhos foram ocupados, o servidor aguarda a liberação de um caminho. Caso contrário, ou seja, se há pelo menos um caminho livre, ele prossegue na caminhada. Cada caminho corresponde a uma estrutura de dados, onde é indicado o seu tipo: unidirecional (semiduplex) ou bidirecional (*full duplex*) e o seu estado: livre ou ocupado. O tempo de caminhada da fila j à fila $(j \bmod m) + 1$ é escolhido segundo a distribuição

$$f_U(t) = \begin{cases} \frac{u^{-r_u}}{\Gamma(r_u)} t^{r_u - 1} e^{\frac{-t}{u}}, & t > 0, \\ 0, & \text{caso contrário,} \end{cases}$$

onde r_u denota o parâmetro de escala da distribuição Erlang. Caso o nó de destino esteja ocupado, ele pode operar em modo repetição ou em modo espera. Neste último caso, eles obtêm os primeiros dois momentos do tempo de espera pela liberação da nó de destino marcado para estatística. Os servidores obtêm, ainda, o valor do parâmetro ρ relativo ao nó marcado para estatística.

Cada fila de recepção tem um esvaziador a ela associado. Se a fila é de capacidade ilimitada, o esvaziador informa ao usuário que acabou de entrar que ele deve sair imediatamente e deixar o sistema simulado. Se a fila é de capacidade limitada, o tempo de retirada de um usuário da fila de recepção, ESV, é escolhido de acordo com a distribuição

$$f_{ESV}(t) = \begin{cases} \frac{esv^{-r_{esv}}}{\Gamma(r_{esv})} t^{r_{esv} - 1} e^{\frac{-t}{esv}}, & t > 0, \\ 0, & \text{caso contrário,} \end{cases}$$

onde esv denota o tempo médio de esvaziamento e r_{esv} o parâmetro de escala da distribuição Erlang. A operação de esvaziamento pode ser do tipo exaustivo, com barreira, limitado exaustivo ou limitado com barreira.

4.2 Diagramas de Transição de Estados

Como um auxílio à memória nos procedimentos de verificação e validação do modelo e do programa de simulação, nós introduzimos os diagramas de transição de estados das entidades.

Nós supomos que as entidades realizam uma série de ações, se comunicam através de mensagens e transicionam em um conjunto discreto e finito de estados. Um estado é uma condição na qual as ações da entidade são suspensas à espera de uma mensagem. Ao receber uma mensagem, a entidade realiza uma seqüência de ações e, então, permanece no mesmo estado ou muda para outro estado, de acordo com o resultado das ações. Denomina-se transição essa seqüência de ações. A mensagem que provoca uma transição pode ser interna à própria entidade ou oriunda de outra entidade (externa). Nós adotamos as seguintes convenções :

-  denota um estado. Cada estado tem um nome que é escolhido arbitrariamente, com exceção do estado inicial, que é denominado *Repouso*. O nome pode ser em forma de mnemônico. Neste caso, nós usamos um mnemônico com 6 caracteres.
- $\longrightarrow$ denota uma transição. Junto deste arco dirigido deve vir o nome da mensagem que provocou a transição e, se a mensagem é externa, deve ser colocado entre parêntesis o nome da entidade que a originou.
- —| denota um ponto de divergência.
- ---[denota um comentário.
- Seja X um estado qualquer de uma entidade diferente do estado *Repouso*. Uma criação e ativação dessa entidade corresponde a uma transição do tipo $Repouso \rightarrow X$. Analogamente, uma desativação corresponde a uma transição do tipo $X \rightarrow Repouso$.

Sistema Simulado

Na Figura 4.1, nós apresentamos o diagrama de transição de estados do sistema simulado. Repare que o sistema simulado sai do estado *Repouso*, após receber

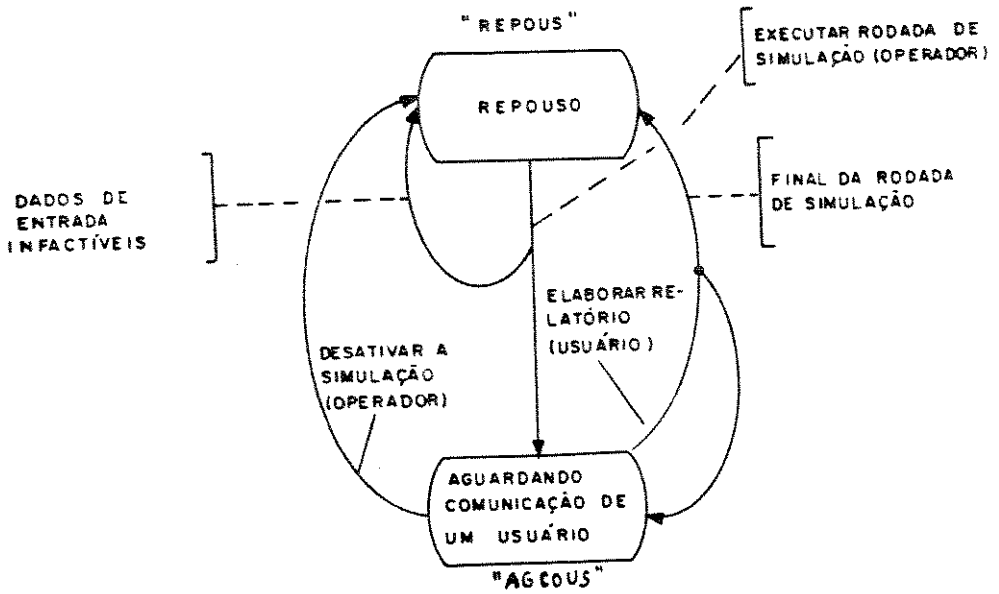


Figura 4.1:

Diagrama de Transição de Estados do Sistema Simulado

do operador a mensagem *Executar Rodada de Simulação*, e vai para o estado *Aguardando Comunicação de um Usuário*, se os dados de entrada são factíveis. Caso contrário, permanece no estado *Repouso*, ou seja, a rodada de simulação não é iniciada. Ao receber de um usuário simulado a mensagem *Elaborar Relatório*, ele elabora um relatório de resultados e permanece no mesmo estado, se o relatório não é o último a ser elaborado, ou vai ao estado *Repouso*, caso contrário, assim como se receber do operador a mensagem *Desativar a Simulação*. Antes de ir a esse estado, ele avisa todas as outras entidades sob seu controle que é o fim de uma rodada de simulação.

Nó

O diagrama de um nó é mostrado na Figura 4.2. Um nó vai para o estado *Apto Para Transmissão e Recepção* ao receber do sistema simulado a mensagem *Ativação*. Esse estado caracteriza um nó livre. Os outros estados são função do tipo de transmissão e da capacidade da fila de recepção. O estado *Em Transmissão* e o estado *Em Recepção* sempre podem existir. O estado *Em Transmissão e Recepção* só pode existir se o modo de operação dos caminhos dos servidores for do tipo *full duplex* (bidirecional). O estado *Aguardando Abrir Vaga na Fila de Recepção* só pode existir no caso de fila de recepção com capacidade limitada. Quando um nó recebe a mensagem *Visita* de um servidor, ele não muda de estado;

apenas realiza as estatísticas relativas aos tempos de ciclo e intervisitas, se for um nó marcado para estatística. Ele é desativado pelo sistema simulado.

Usuário

O diagrama de transição de um usuário é mostrado na Figura 4.3. Estando no estado *Repouso*, um usuário vai ao estado *Aguardando Atendimento de um Servidor*, se há lugar na fila de transmissão do nó de origem, ou retorna ao estado *Repouso*, caso contrário, ao receber do gerador a mensagem *Ativação*. A soma dos tempos em que um usuário fica nos estados *Aguardando Atendimento de um Servidor* e *Aguardando Fim de Comunicação Com Servidor* é o seu tempo de espera na fila de transmissão. O estado *Aguardando Liberação do Destino* só pode existir no modo servidor com espera. O tempo em que ele fica no estado *Aguardando Fim de Transmissão* é o seu tempo de serviço. Se as filas de recepção forem de capacidade ilimitada, ele fica 0 unidades de tempo no estado *Aguardando Atendimento de um Esvaziador*, ou seja, é atendido instantaneamente. Ele é avisado do fim de uma rodada pelo gerador, através da mensagem *Desativação* ou, no caso de filas de recepção limitadas, pode ser desativado pelo esvaziador da fila de recepção do nó de destino.

Gerador de Usuários

O diagrama do gerador de usuários é mostrado na Figura 4.4. O gerador só recebe mensagens do sistema simulado, ou seja, é criado, ativado e desativado pelo sistema simulado. A mensagem *Temporização de Geração de Usuário* é uma mensagem interna; ao recebê-la, ele cria e ativa um usuário. Os intervalos de tempo entre gerações dessas mensagens têm distribuição exponencial negativa, conforme anteriormente descrito.

Servidor Cíclico

O diagrama de transição de estados de um servidor cíclico é mostrado na Figura 4.5. Ao receber do sistema simulado a mensagem *Ativação*, um servidor vai ao estado *Aguardando Instante de Visitação*. Sai desse estado quando recebe a mensagem interna *Temporização de Caminhada* e verifica que a visita que fez a uma dada fila j, no seu ciclo de caminhada, foi bem sucedida. Nesse caso, vai ao estado *Aguardando Fim de Comunicação Com Usuário*, onde permanece até receber a mensagem interna *Temporização de Comunicação*. Quando sai desse estado, vai a um dos estados: *Aguardando Instante de Visitação, Aguardando Retorno à Caminhada ou Aguardando Liberação do Destino*, de acordo com o seu

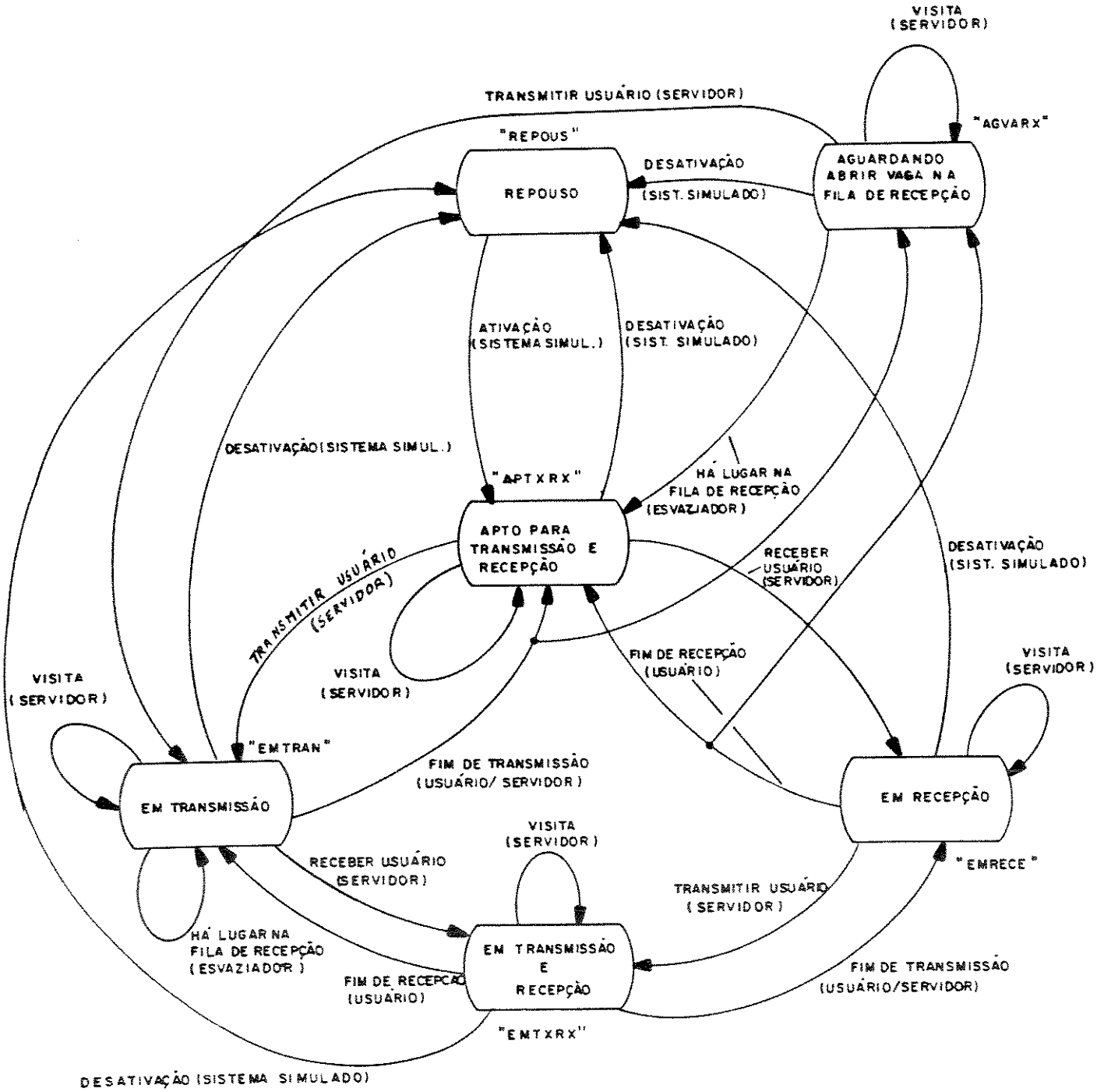


Figura 4.2:

Diagrama de Transição de Estados de Um Nó

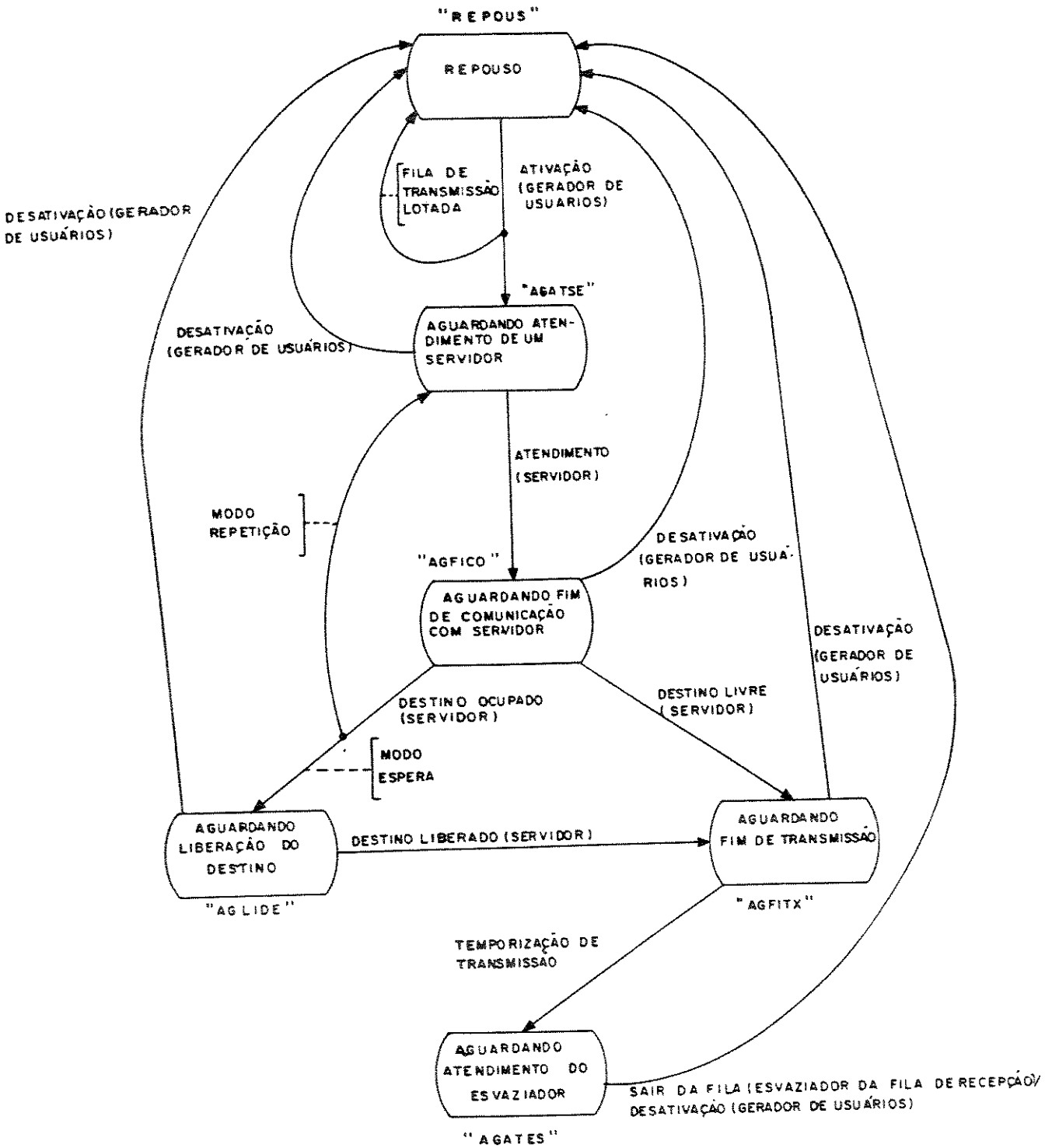


Figura 4.3:

Diagrama de Transição de Estados de Um Usuário

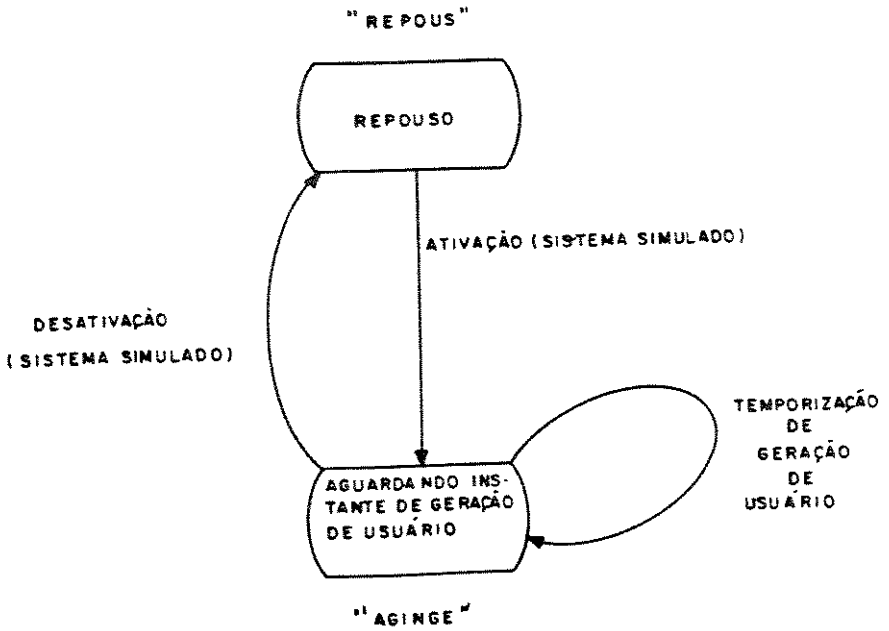


Figura 4.4:

Diagrama de Transição de Estados do Gerador de Usuários

modo de escalonamento e com a quantidade de caminhos que controla. Ele é desativado pelo sistema simulado.

Temporizador de Liberação de Caminho

Na Figura 4.6 é mostrado o diagrama do temporizador de liberação de caminho. Ao ser ativado pelo servidor, o temporizador vai ao estado *Aguardando Instante de Liberação de Caminho*, onde permanece pelo tempo correspondente ao tempo de serviço do usuário que foi atendido pelo servidor. Sai desse estado ao receber a mensagem interna *Temporização de Liberação de Caminho*, quando então, se todos os caminhos do servidor que o criou estavam ocupados, ele o avisa que um caminho foi liberado.

Esvaziador da Fila de Recepção

Finalmente, na Figura 4.7, é apresentado o diagrama do esvaziador da fila de recepção. Ao receber a mensagem *Ativação* do sistema simulado, ele vai ao estado *Aguardando Instante de Visitação*, onde permanece até que receba um aviso de chegada de usuário, através da mensagem *Esvaziar Fila*, quando então passa ao estado *Retirando Usuário da Fila de Recepção*. No caso de fila de recepção

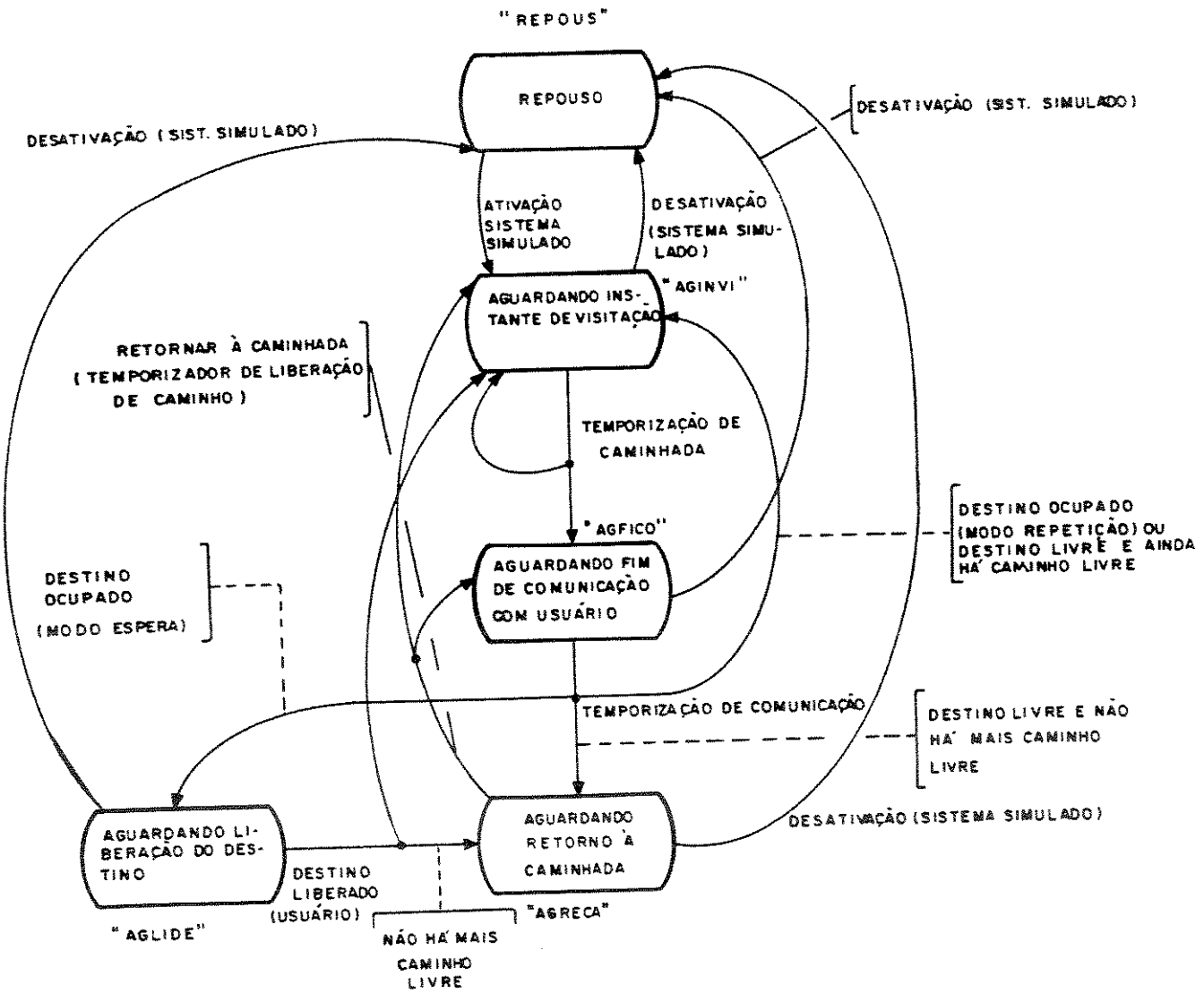


Figura 4.5:

Diagrama de Transição de Estados de Um Servidor Cíclico

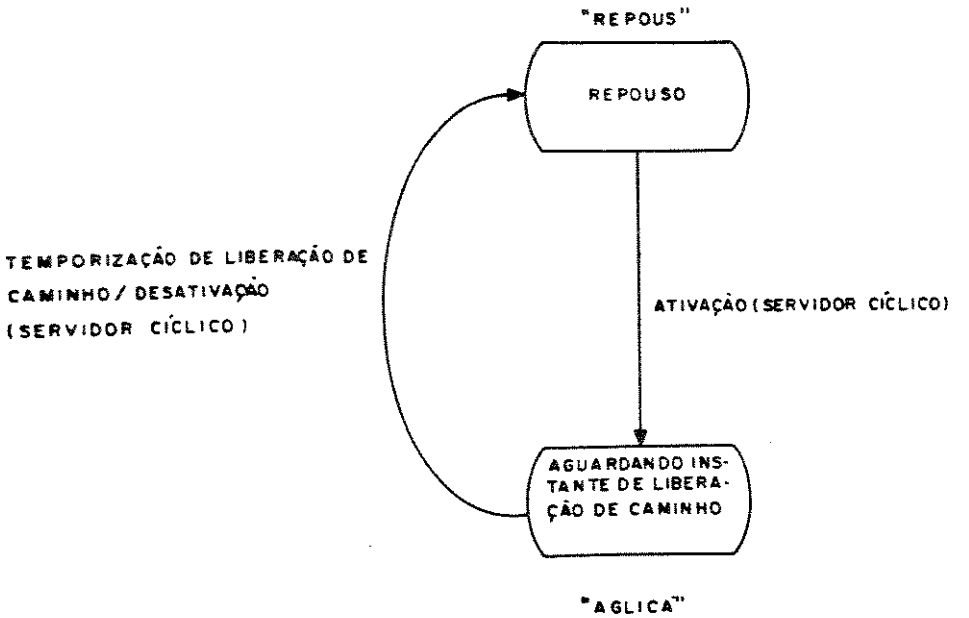


Figura 4.6:

Diagrama de Transição de Estados de Um Temporizador de Liberação de Caminho

limitada, ele gera mensagens internas *Temporização de Esvaziamento* segundo o procesto Erlang-r_e, conforme anteriormente descrito. Ao receber essa mensagem, informa o usuário em questão que ele deve sair da fila.

4.3 Resultados de Simulação

Os resultados de uma rodada de simulação são obtidos em relação ao nó e ao servidor marcados para estatística.

Consideremos uma rodada de simulação particular. O índice $\star$ indica que o estimador é obtido por simulação. Nesta seção, nós utilizamos os mnemônicos dos estados, conforme estipulado nos diagramas de transição de estados das entidades.

Estatísticas

Denotemos por $n > 1$ o número de saídas do estado *Repous* dos usuários originados no nó marcado em uma rodada. Fazendo:
$x(i) = 1$, se a transição de um usuário i é *Repous*$\rightarrow$*Repous*,
$x(i) = 0$, caso contrário, $i = 1, 2, \ldots, n$, o estimador da probabilidade de

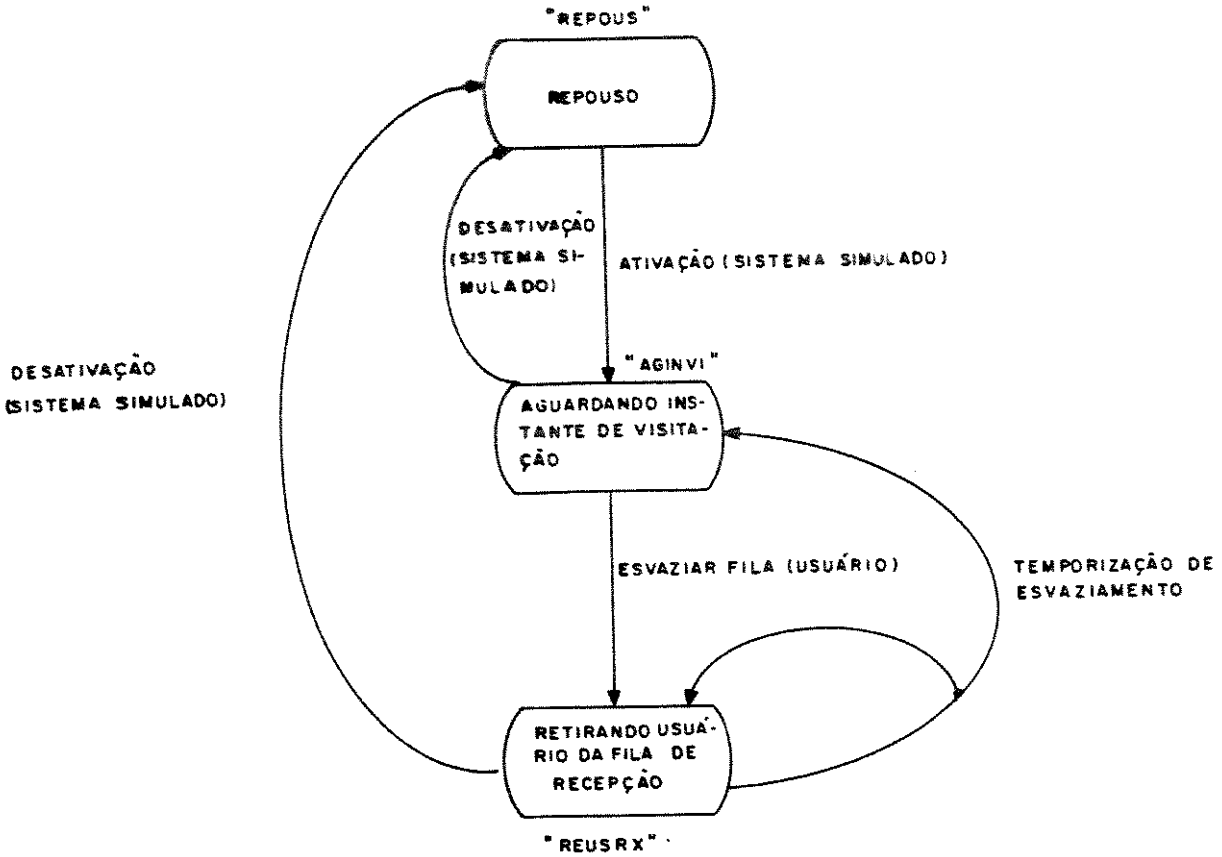


Figura 4.7:

Diagrama de Transição de Estados de Um Esvaziador de Fila de Recepção

usuários perdidos por fila lotada é dado por

$$p^*_{tx} = \frac{\sum_{i=1}^{n} x(i)}{n}.$$

Seja, agora, $n > 1$ o número de saídas do estado *Aginvi* do servidor marcado em relação ao nó marcado. Fazendo
$y(i) = 1$, se a transição é *Aginvi*→*Agfico*,
$y(i) = 0$, caso contrário
e
$z(i) = 1$, se a transição é *Aginvi*→*Aginvi*, devido a fila de transmissão vazia,
$z(i) = 0$, caso contrário, $i = 1, 2, \ldots, n$, então os estimadores de ρ e P_0 são dados por

$$\rho^* = \frac{\sum_{i=1}^{n} y(i)}{n}$$

e

$$P_0^* = \frac{\sum_{i=1}^{n} z(i)}{n}.$$

Os estimadores dos dois primeiros momentos do tempo de ciclo do servidor marcado e do tempo intervisita de servidores quaisquer, dos tipos (') e ("), são obtidos pelo nó marcado para estatística. Ele registra, no decorrer de uma rodada, os instantes de recebimento de mensagens dos servidores. Se a mensagem é *Visita*, é registrado um instante de visita do tipo ('). Se a mensagem é *Transmitir Usuário*,

é registrado um instante de visita do tipo (''). A diferença entre o instante atual e o instante anterior corresponde ao tempo de ciclo ou intervisita de um dos dois tipos, conforme for o caso. Ao obter um valor desse tempo, o nó atualiza os estimadores dos dois primeiros momentos dos tempos de ciclo e intervisitas condicionais.

Cada usuário que realiza a transição *Agfico*→*Agfitx* obtém seu tempo de espera na fila de transmissão. Esse tempo é definido como a diferença entre o instante de saída do estado *Agfico* para o estado *Agfitx* e o instante de entrada no estado *Agatse*. A partir desse valor, ele atualiza os estimadores $w^{\star}$ e $w^{(2)\star}$. Analogamente, cada usuário que realiza a transição*Agates*→ *Repous* obtém seu tempo de espera na fila de recepção, W_{rx}. Esse tempo é definido como a diferença entre o instante de saída e o instante de entrada no estado *Agates*. A partir desse valor, ele atualiza os estimadores dos dois primeiros momentos do tempo de espera na fila de recepção marcada, $w^{\star}_{rx}$ e $w^{(2)\star}_{rx}$.

No caso de operação em modo espera, os servidores cíclicos obtêm o tempo de espera pela liberação do nó de recepção marcado, W_{prx}. Esse tempo é definido como a diferença entre o instante de saída e o instante de entrada no estado *Aglide*. Cada servidor que realiza a transição*Aglide*→*Agreca* atualiza os estimadores $w^{\star}_{prx}$ e $w^{(2)\star}_{prx}$.

Método de Obtenção dos Estimadores dos Momentos

A atualização dos estimadores dos dois primeiros momentos de uma variável aleatória genérica Y, $y^{\star}$ e $y^{(2)\star}$, é realizada a partir de um valor atual da amostra, Y_{i+1}, e da média e variância da amostra de tamanho i, $i = 1, 2, \ldots$, ou seja,

$$y_i^{\star} = \frac{1}{i}\sum_{j=1}^{i} Y_j,$$

$$y_{i+1}^{\star} = \frac{i}{i+1}y_i^{\star} + \frac{1}{i+1}Y_{i+1},$$

$$VAR[Y_{i+1}]^{\star} = \frac{i}{i+1}VAR[Y_i]^{\star} + \frac{i}{(i+1)^2}(Y_{i+1} - y_i^{\star})^2,$$

onde $VAR[Y_1]^{\star} = 0$ e

$$y_{i+1}^{(2)\star} = VAR[Y_{i+1}]^{\star} + y_{i+1}^{\star 2}.$$

Nós utilizamos o *método das replicações independentes com supressão dos dados iniciais* [31] para estimar um parâmetro qualquer do modelo de simulação, que pode ser, por exemplo, o primeiro ou o segundo momento de uma variável aleatória. Cada replicação é obtida escolhendo-se um conjunto diferente de sementes das variáveis aleatórias geradas pelas entidades do modelo. Sejam, por exemplo, as $n \geq 2$ replicações independentes, cada uma com tamanho igual a $l > 1$ observações

e $l_0 \geq 0$ o número de observações iniciais suprimidas em cada rodada, de um estimador genérico x^*: $\{x^*_{ij}\ ,\ i = 1,2,\ldots,n$ e $j = 1,2,\ldots,l\}$. Um procedimento para determinar l_0 é dado em [31]. A média amostral da $i-$*ésima* replicação é dada por

$$\overline{x^*_i}(l) = \frac{1}{l - l_0} \sum_{j=l_0}^{l} x^*_{ij}.$$

A média amostral das n replicações independentes é

$$\overline{x^*}(n) = \frac{1}{n} \sum_{i=1}^{n} \overline{x^*_i}(l)\ ,$$

a variância amostral é dada por

$$\sigma^{*^2}(n) = \frac{1}{n-1} \sum_{i=1}^{n} \left[\overline{x^*_i}(l) - \overline{x^*}(n)\right]^2$$

e o intervalo aproximado de $100(1-\gamma)\%$ de confiança é

$$\overline{x^*}(n) \pm t_{n-1,1-\frac{\gamma}{2}} \times \frac{\sigma^*(n)}{\sqrt{n}}\ ,$$

cuja precisão relativa é igual a

$$\frac{t_{n-1,1-\frac{\gamma}{2}} \sigma^*(n)}{\overline{x^*}(n)\sqrt{n}} \times 100\%\ ,$$

onde $t_{n-1,1-\frac{\gamma}{2}}$ é o valor crítico superior $1-\frac{\gamma}{2}$ da distribuição t de Student com $n-1$ graus de liberdade.

Reprodução das Características do Modelo Analítico

A fim de reproduzir no modelo de simulação as mesmas condições do modelo analítico introduzido no Capítulo anterior, o operador deve considerar:

- filas de transmissão e recepção com capacidade ilimitada;
- os servidores operando no modo repetição;
- $l_s = 1$ para todos os servidores, ou seja, cada servidor com apenas um único caminho de comunicação;
- tempo de comunicação usuário-servidor nulo, ou seja, o servidor verifica instantaneamente o estado (livre ou ocupado) do nó de destino de um usuário;
- os caminhos utilizados em modo semiduplex.

As características das extensões do modelo analítico também podem ser reproduzidas facilmente no modelo de simulação. Por exemplo, podem ser considerados sistemas com filas de capacidade limitada, operação dos caminhos em modo *full duplex* e servidores operando em modo espera.

4.4 Extensões

O modelo de simulação pode ser estendido de maneira relativamente fácil, em termos de complexidade de espaço e tempo, para tratar sistemas de filas cíclicas com outras características. A seguir, nós relatamos possíveis extensões.

1. Quando o processo de chegada de usuários não pode ser considerado Poissoniano, o gerador de usuários deve obter os intervalos de tempo entre chegadas, G, e o número do nó da fila de transmissão, J, de acordo com distribuições apropriadas em cada caso, o que pode ser feito, inclusive, a partir de dados experimentais.

2. Podem ser adotadas distribuições de probabilidade diferentes da distribuição Erlang-r para o tempos de serviço, de caminhada ou de esvaziamento.

3. Os servidores podem ter tempos de serviço ou de caminhada diferentes (alguns servidores podem ser mais "lentos" que outros). Nesse caso, os servidores devem gerar H ou U com distribuições ou médias diferentes. O escalonamento cíclico de um servidor pode ser também determinado de acordo com uma tabela prefixada. Alguns nós podem ser visitados com mais freqüência que outros.

4. Quanto à ordem de atendimento dos usuários em uma visita, os servidores podem operar de acordo com uma disciplina diferente da FIFO.

5. Quanto ao número de usuários atendidos por visita, uma outra disciplina pode ser adotada. Por exemplo, o serviço decremental [15]: quando há pelo menos um usuário apto a ser servido em uma visita, o serviço continua até que o número de usuários em fila decresça a um menos que aquele encontrado no instante de visita.

6. O número máximo de servidores que podem ser simultaneamente usados por um nó pode ser qualquer valor entre 1 e s. Nesse caso, o nó visitado deve informar ao servidor o número de servidores que ele já está usando.

7. Um usuário pode ser servido somente por um subconjunto particular de servidores. Essa situação modela, por exemplo, uma rede em anel segmentado com quadros de tamanho variável: quando um quadro vazio chega à estação, deve ser indicado também o seu tamanho, de modo que ela possa verificar se o quadro é adequado para transmitir a mensagem em questão.

8. Após sair da fila de recepção, um usuário pode, com uma certa probabilidade, retornar ao sistema.

Capítulo 5

VALIDAÇÃO DO MODELO ANALÍTICO

A fim de explorar a sensibilidade do modelo analítico com respeito à variação de certos parâmetros de entrada, nós apresentamos aplicações dos modelos ao estudo de alguns sistemas de filas cíclicas. A sensibilidade é medida através da comparação com resultados do modelo de simulação, o que mostra objetivamente a importância deste modelo. Em nossos experimentos, nós empregamos um nível de confiança igual a 95% relativo às simulações.

Nós utilizamos a mesma notação introduzida nos dois capítulos anteriores. Denotamos também:

- o tráfego oferecido a um servidor por a, onde

$$a = \frac{1}{s}\sum_{i=1}^{m}\lambda_i h_i;$$

- o tráfego escoado por um nó j, $j = 1, 2, \ldots, m$ por a_{e_j}, onde

$$a_{e_j} = \lambda_j h_j + \sum_{i=1, i\neq j}^{m} \alpha_{ij}\lambda_i h_i.$$

  Em um sistema simétrico, $a_{e_j} = a_e = 2\lambda h$. O tráfego é medido na unidade Erlang (erl).

- a taxa média de chegada global por

$$\lambda_T = \sum_{i=1}^{m}\lambda_i;$$

- o tempo médio de ciclo absoluto por

$$c = [P_0 + (1 - P_0)(1 - \rho)]\, c^{'} + (1 - P_0)\rho c^{''};$$

- o segundo momento do tempo de ciclo absoluto por

$$c^{(2)} = [P_0 + (1 - P_0)(1 - \rho)]\, c'^{(2)} + (1 - P_0)\rho c''^{(2)};$$

- o tempo médio intervisita absoluto por

$$v = [P_0 + (1 - P_0)(1 - \rho)]\, v' + (1 - P_0)\rho v'';$$

- o segundo momento do tempo intervisita absoluto por

$$v^{(2)} = [P_0 + (1 - P_0)(1 - \rho)]\, v'^{(2)} + (1 - P_0)\rho v''^{(2)}.$$

Quando aplicamos o método da superposição dos ciclos condicionais, o procedimento de superposição que fornece resultados melhores é sempre citado em qualquer exemplo particular. A fim de evitar ambigüidade com a variável que denota o número de servidores, nós abreviamos a unidade de tempo segundo por *seg*. Entretanto, os múltiplos de segundo são abreviados de acordo com as normas internacionais.

Na Seção 5.1, nós descrevemos os sistemas de filas cíclicas de referência utilizados para os propósitos de validação do modelo analítico. Os resultados numéricos obtidos através da aplicação do método da equivalência entre as taxas de serviço e caminhada e do método da superposição dos ciclos condicionais são apresentados na Seção 5.2. Finalmente, na Seção 5.3, nós fazemos comentários sobre a precisão dos resultados aproximados e descrevemos o método da verificação da influência da probabilidade de visita bem sucedida e dos dois primeiros momentos do tempo de ciclo na precisão.

5.1 Sistemas de Referência

Um sistema de filas cíclicas é dito *fracamente acoplado*, se o tráfego escoado por qualquer nó pode ser considerado baixo, ou seja, se

$$a_{e_j} < a_{e_0},\ j = 1, 2, \ldots, m,$$

onde a_{e_0} é um tráfego escoado limite, cujo valor numérico depende da configuração particular do sistema considerado.

Os sistemas que têm motivado as nossas investigações são fracamente acoplados, têm um grande número de filas e servidores e o tempo de caminhada é pequeno em relação ao tempo de serviço. Para efetuar os estudos de validação do modelo analítico com respeito aos dois métodos de obtenção dos momentos do tempo de ciclo equivalente, nós escolhemos quatro sistemas de filas cíclicas simétricos com as seguintes características:

sistema I: $m = 10$ nós, $s = 1, 2$ ou 3 servidores, tempo de serviço exponencial com $h = 1$ ms e tempo de caminhada constante $u = 0,1$ ms

sistema II: $m = 50$ nós, $s = 4$ servidores, tempo de serviço $h = 1,500$ *bit time* e tempo de caminhada $u = 0,060$ *bit time*, ambos constantes

sistema III: $m = 100$ nós, $s = 1, 2, 3, 4$ ou 5 servidores, tempo de serviço exponencial com $h = 1$ ms e tempo de caminhada constante $u = 0,1$ ms

sistema IV: $m = 1024$ nós, $s = 4, 12, 18$ ou 24 servidores, tempo de serviço $h = 2$ ms e tempo de caminhada $u = 8$ μs, ambos constantes.

Nós efetuamos os cálculos para $a = 0,2$; 0,4; 0,6 e 0,8 erl. Nesse campo de valores de tráfego, o sistema I é fortemente acoplado enquanto que os outros três sistemas são fracamente acoplados. O sistema II corresponde a uma rede em anel apresentada em [22] e é chamado anel de Bhuyan. Ele foi escolhido para propósitos de comparação com resultados de um modelo analítico existente na literatura. O sistema IV é um exemplo significativo da classe de sistemas que motivaram nossos estudos.

Com relação ao modelo de simulação, por simplicidade de notação nós temos que para $s = 1$, c e $c^{(2)}$, e, para $s > 1$, $v = c_e$ e $v^{(2)} = c_e^{(2)}$, respectivamente, denotam os dois primeiros momentos do tempo de ciclo do servidor equivalente.

5.2 Resultados Numéricos

a) Método da Equivalência Entre as Taxas de Serviço e Caminhada

Sistema I

Os resultados são apresentados na Tabela 5.1, juntamente com os valores médios obtidos por simulação. Para $s = 1$, vemos que os valores de P_0 obtidos analiticamente concordam bem com os valores médios das simulações, para todos os valores de a. No entanto, os valores de c, $c^{(2)}$ e, conseqüentemente, w obtidos analiticamente são otimistas em relação aos correspondentes valores das simulações. Para $s = 2$ e $s = 3$, nós vemos que os valores de c_e, $c_e^{(2)}$ e P_0 obtidos analiticamente se afastam dos correspondentes valores médios obtidos por simulação, à medida que a aumenta. Com isso, o mesmo ocorre com relação a w e o modelo analítico às vezes é otimista e às vezes é pessimista se comparado ao modelo de simulação. Ele poderia ser utilizado eventualmente como uma aproximação razoável para tráfegos moderados, digamos algo no entorno de 0,4 ou 0,6 erl.

Para adquirir uma orientação mais objetiva sobre a validade das aproximações, nós consideramos o cálculo de P_0 e w, dados por (3.4) e (3.9), respectivamente, em função dos valores médios dos parâmetros ρ, v (c_e) e $v^{(2)}$ $(c_e^{(2)})$, obtidos das várias rodadas de simulação, e de λ, h_e e $h_e^{(2)}$ dados, ou seja, em função da probabilidade de visita bem sucedida e dos dois primeiros momentos do tempo intervisita obtidos por simulação e de taxas de chegada e dois primeiros momentos do tempo de serviço conhecidos. Os resultados comparativos são mostrados na Tabela 5.2, onde os resultados de simulação são apresentados juntamente com os seus respectivos intervalos de 95% de confiança. Assim, com respeito a ambos os parâmetros P_0 e w, temos intervalos de confiança simultâneos de 90%, devido à desigualdade de Bonferroni [31]. Vemos que para $s = 1$, 2 ou 3 servidores, os valores de P_0 obtidos analiticamente se situam no interior dos correspondentes intervalos de 95% de confiança para todos os valores de a, exceto no caso em que $s = 3$. Com respeito ao parâmetro w, não são obtidos bons resultados.

Sistema II

Os resultados serão apresentados posteriormente, juntamente com aqueles obtidos através da aplicação do método da superposição dos ciclos condicionais.

Sistema III

Os resultados comparativos são mostrados na Tabela 5.3. Para $s = 1$, vemos que os valores de P_0, c, $c^{(2)}$ e w obtidos analiticamente concordam bem com os valores médios das simulações, para todos os valores de a. No entanto, para $s = 2$, 3, 4 ou 5, vemos que, embora os valores de P_0 e c_e concordem relativamente bem com os correspondentes valores médios das simulações, os valores de $c_e^{(2)}$ e, conseqüentemente, w obtidos analiticamente se afastam bastante dos correspondentes valores médios obtidos via simulação. A influência do segundo momento $c_e^{(2)}$ é marcante e o modelo analítico é otimista em relação ao modelo de simulação.

Seguindo a mesma orientação do exemplo anterior, nós apresentamos na Tabela 5.4 os valores de P_0 e w calculados analiticamente em função dos valores médios de ρ, v (c_e) e $v^{(2)}$ $(c_e^{(2)})$, obtidos nas rodadas de simulação correspondentes, e λ,h_e e $h_e^{(2)}$ dados. Por legibilidade, omitimos os resultados obtidos para o caso $s = 1$. Como podemos ver, uma boa parte dos valores de P_0 e w obtidos analiticamente se situam no interior dos respectivos intervalos de 95% de confiança.

Sistema IV

Os resultados comparativos em relação ao tempo médio de espera são mostrados na tabela 5.5. Para $s = 4$, vemos que não são obtidos bons resultados para nenhum valor de a. Já para $s = 4$, 18 ou 24, são obtidos resultados razoáveis apenas para $a = 0,6$ erl.

b) Método da Superposição dos Ciclos Condicionais

Sistema I

Os resultados são apresentados na Tabela 5.6. Para o caso de um único servidor vemos que os valores de P_0 e c obtidos analiticamente concordam bem com os valores obtidos por simulação, para todos os valores de a. Entretanto, os valores de $c^{(2)}$ obtidos analiticamente tendem a se afastar dos correspondentes valores obtidos por simulação, à medida que a aumenta. Com isso, ocorre o mesmo fenômeno com relação ao parâmetro w. Para $s = 2$ e $s = 3$ servidores, vemos que os valores de P_0 e v obtidos analiticamente concordam relativamente bem com os valores obtidos por simulação, para todos os valores de a. No entanto, os valores de $v^{(2)}$ obtidos analiticamente tendem a se afastar daqueles obtidos por simulação, à medida que a aumenta. Assim, ocorre o mesmo fenômeno com relação a w. Para

adquirir mais evidências sobre a validade das aproximações, nós consideramos o cálculo dos parâmetros P_0 e w, dados por (3.26) e (3.30), respectivamente, em função dos valores médios dos parâmetros ρ, v', $v'^{(2)}$, v'' e $v''^{(2)}$, obtidos das várias rodadas de simulação, e um dado valor de λ. Os resultados comparativos são mostrados na Tabela 5.7, onde os resultados de simulação são apresentados juntamente com seus respectivos intervalos de 95% de confiança. Convém lembrar novamente que, com relação a ambos os parâmetros P_0 e w, temos intervalos de confiança simultâneos de 90%. Vemos que para $s = 1$, os valores de P_0 e os valores de w, exceto, neste caso, para $a = 0,8$ erl, obtidos analiticamente, se situam no interior dos correspondentes intervalos de 95% de confiança. Agora, para $s = 2$ e $s = 3$, a maioria dos valores de P_0 e todos os valores de w se situam no exterior dos correspondentes intervalos de 95% de confiança. Assim, o modelo analítico não se aplica bem nesses casos, desde que não obtemos bons resultados mesmo fornecendo ao modelo os dados médios obtidos via simulação.

Sistema II

Usamos o procedimento de superposição de Bhuyan e alguns resultados são mostrados na Figura 5.1. Pode ser visto que há uma boa concordância entre os resultados analíticos, os de simulação e os resultados obtidos em [22] para uma ampla faixa de valores de a. Quando o tráfego aumenta, o modelo analítico se torna um pouco pessimista em relação aos resultados de simulação. Na Tabela 5.8, nós apresentamos os valores de w obtidos através da aplicação dos dois métodos, para alguns valores de a. Os resultados médios de simulação estão acompanhados dos seus respectivos intervalos de 95% de confiança. Repare que, com relação aos resultados de simulação, o método da superposição dos ciclos condicionais é mais preciso do que o método da equivalência entre taxas, neste exemplo particular.

Sistema III

Os resultados são apresentados na Figura 5.2, onde, por legibilidade, os resultados de simulação foram suprimidos. Eles são mostrados na Tabela 5.9. Como anteriormente, consideramos o cálculo de P_0 e w em função dos valores médios de ρ e dos primeiros dois momentos dos tempos intervisitas condicionais, obtidos nas rodadas de simulação, e de um dado λ. Os resultados obtidos para $s = 1$ foram omitidos. Examinando os resultados comparativos dessa tabela, notamos que os valores de P_0 e w obtidos analiticamente estão muito próximos dos correspondentes valores obtidos nas simulações, o que é uma indicação segura de que o método da superposição dos ciclos condicionais funciona bem para uma ampla gama de valores do parâmetro a neste exemplo particular. Esse comportamento também se verificou nos outros exemplos deste trabalho e em outros não aqui apresentados,

que podem ser caracterizados como sistemas de filas cíclicas fracamente acoplados. Na Tabela 5.10, apresentamos valores de w obtidos através da aplicação dos dois métodos. Neste exemplo, o método da superposição dos ciclos condicionais forneceu resultados mais precisos do que o método da equivalência entre as taxas de serviço e caminhada, com exceção do caso em que $s = 5$ e $a = 0,8$ erl.

Sistema IV

Nós consideramos inicialmente $s = 4$ servidores. A dependência do tempo médio de espera na fila de transmissão com respeito ao tráfego oferecido por servidor é mostrada na Figura 5.3.a, onde nós usamos o procedimento de superposição de Kuehn. Supomos que o tempo médio de transmissão de mensagem entre dois nós, sob condições operacionais nominais, deve ser menor ou igual a 10 ms. Desde que o tempo de transmissão é composto pelo tempo de espera na fila e o tempo de serviço, w deve ser menor ou igual a 8 ms, donde $a = 0,60$ erl aproximadamente. Segue dos resultados na Figura 5.3.a que isto corresponde a uma taxa global igual a 1200 mensagens/seg. Seja, agora, $s = 12$ servidores. A dependência de w com respeito a a é mostrada na Figura 5.3.b, onde nós usamos o procedimento de superposição de Bhuyan. Para w da ordem de 8 ms, devemos ter pelo menos que $a = 0,90$ erl. Mas, devido ao fato que alguns requisitos devem também ser satisfeitos sob condições de sobrecarga, que corresponde a um acréscimo de 20% em a, devemos ter, na verdade, que pelo menos $a = 0,80$ erl. Então, da Figura 5.3.b, nós temos que w deve ser da ordem de 4 ms, correspondendo a uma taxa global igual a 4800 mensagens/seg. Consideremos nos dois casos que seguem, respectivamente, um sistema com $s = 18$ e outro sistema com $s = 24$ servidores. A dependência de w com respeito a a para os dois casos é mostrada nas Figuras 5.3.c e 5.3.d, onde nós usamos o procedimento de superposição de Bhuyan. Pelas mesmas razões mencionadas anteriormente, devemos ter que $a \approx 0,80$ erl, que corresponde, no caso $s = 18$, a uma taxa global igual a 7200 mensagens/seg e, no caso $s = 24$, a uma taxa global igual a 9600 mensagens/seg aproximadamente.

Nos quatro casos aqui apresentados, nós observamos que os resultados analíticos concordam suficientemente bem com os os resultados de simulação, no intervalo de 0,20 erl a 0,95 erl para a, o que corresponde a um campo normal de operação em aplicações desse tipo. O método da superposição dos ciclos condicionais tem uma aplicação significativa a essa situação.

5.3 Precisão das Aproximações

Com relação ao método da equivalência entre as taxas de serviço e caminhada, os resultados numéricos obtidos mostram que:

1. o modelo funciona bem em algumas situações particulares, independente de se ter sistemas fracamente ou fortemente acoplados; a principal razão que justifica a validade da aproximação é que, nessas situações, a variância do tempo de ciclo do servidor equivalente é um bom estimador da variância do tempo intervisita de um servidor genérico;

2. o modelo analítico não tem nenhuma tendência em ser otimista e nem tampouco pessimista em relação a qualquer parâmetro do modelo de simulação;

3. é possível melhorar a precisão do modelo analítico, desde que aprimoremos o método de obtenção de P_0, do primeiro e, principalmente, do segundo momento do tempo de ciclo do servidor equivalente.

Nesse sentido, nós seguimos uma sugestão dada em [20] e, então, supomos que a taxa de caminhada do servidor equivalente é igual a s vezes a taxa de caminhada de um servidor original apenas para valores pequenos de tráfego, e não é alterada à medida que o tráfego oferecido por servidor tende à unidade. A nova equivalência entre os tempos de caminhada passa a ser dada por

$$U_{e_j} = \frac{U_j}{(1-a)s + a} \quad j = 1, 2, \ldots, m. \tag{5.1}$$

Repare que se $a \to 0 \Rightarrow U_{e_j} \to \frac{U_j}{s}$ e se $a \to 1 \Rightarrow U_{e_j} \to U_j$. Denotamos este novo método por *método da equivalência modificada.*

Os resultados comparativos obtidos aplicando-se o método da equivalência modificada aos sistemas I, II, III e IV são mostrados nas tabelas 5.11, 5.12, 5.13 e 5.14, respectivamente. Comparando-se esses resultados com os correspondentes obtidos pela aplicação do outro método da equivalência, podemos notar que:

- sistema I: o método da equivalência modificada forneceu resultados iguais ou piores;

- sistema II: o método da equivalência modificada forneceu resultados melhores para todos os valores de a, exceto para $a = 0,6$ erl;

- sistema III: o método da equivalência modificada forneceu resultados iguais ou melhores para $a = 0,2$, 0,4 e 0,6 erl; para $a = 0,8$ erl, ele forneceu resultados excessivamente pessimistas (instabilidade), mesmo se comparados com as simulações;

- sistema IV: o método da equivalência modificada forneceu resultados melhores para $s = 4$; para $s = 12$, 18 e 24, ele forneceu resultados melhores apenas para $a = 0,2$ e $a = 0,4$.

O método da superposição dos ciclos condicionais funciona bem, em geral, no caso de sistemas de filas cíclicas fracamente acoplados e geralmente não funciona

bem para sistemas de filas cíclicas fortemente acoplados com $s > 1$ servidores, principalmente para tráfegos elevados. Isto significa que, mesmo calculando-se P_0 e w em função de um λ dado e dos valores médios dos parâmetros ρ, v', $v'^{(2)}$, v'' e $v''^{(2)}$ obtidos por simulação, em geral, não são obtidos bons resultados. Um exemplo típico foi apresentado na seção anterior. A principal razão da não validade das aproximações é o fato que o efeito das outras filas sobre a fila marcada é limitado, no modelo, aos tempos intervisitas condicionais, o que não parece ser suficiente em sistemas cíclicos fortemente acoplados, onde os servidores tendem a caminhar juntos [17]. Para esses sistemas, torna-se imprescindível o emprego do modelo de simulação, especialmente para um número pequeno de filas, quando então o tempo de execução de uma rodada é geralmente pequeno.

Uma outra tendência verificável em um sistema simétrico é que a precisão das aproximações tende a melhorar, à medida que a razão entre o tempo de caminhada total e o tempo médio de serviço aumenta, de acordo com a conjectura de Tran-Gia e Raith [11].

Quando o método da superposição dos ciclos condicionais funciona bem, nós constatamos que a influência dos parâmetros v', $v'^{(2)}$, v'' e $v''^{(2)}$ no cálculo de P_0 e w é preponderante em relação à influência do parâmetro ρ. A verificação, nesse caso, é feita de acordo com o procedimento:

1. calcula-se P_0 e w através de (3.26) e (3.30), respectivamente, em função do parâmetro ρ obtido analiticamente, dos valores médios de v', $v'^{(2)}$, v'' e $v''^{(2)}$ obtidos nas rodadas de simulação e um λ dado;

2. calcula-se P_0 e w através de (3.26) e (3.30), respectivamente, em função do valor médio de ρ obtido das rodadas de simulação, dos valores de v', $v'^{(2)}$, v'' e $v''^{(2)}$ obtidos analiticamente e um λ dado;

3. compara-se os valores de P_0 e w obtidos nos passos 1 e 2 com os correspondentes valores médios e respectivos intervalos de 95% de confiança obtidos nas simulações.

Nós constatamos também que o método de obtenção de v' e $v'^{(2)}$ é mais preciso que o método de obtenção de v'' e $v''^{(2)}$, de acordo com a intuição. A razão para isso é a hipótese de independência entre os tempos de ciclo condicionais.

É muito difícil prever a priori qual dos dois procedimentos de superposição [22],[26] funciona melhor. Nos nossos experimentos numéricos eles deram, em geral, resultados alternados com respeito à precisão das aproximações. Em algumas situações, entretanto, especialmente para um grande número de servidores, nós constatamos que o procedimento de Kuehn é divergente, conforme já fora observado por Whitt [27]. É claro que, devido à simplicidade dos cálculos, o procedimento de Bhuyan é preferível ao procedimento de Kuehn, quando ambos forem suficientemente precisos.

Com base nos experimentos aqui relatados e em outros que realizamos e não relatamos, algumas conjecturas podem ser feitas:

- para sistemas fortemente acoplados, não deve ser utilizado o método da superposição; recomenda-se, nesse caso, o método da equivalência modificada para tráfegos oferecidos por servidor baixos e moderados, e o método da equivalência para tráfegos altos, digamos, iguais ou superiores a 0,7 erl;

- para sistemas fracamente acoplados, recomenda-se utilizar o método da equivalência para tráfegos baixos, o método da equivalência modificada para tráfegos moderados e o método da superposição para tráfegos altos.

Deve-se ressaltar todavia que, devido ao caráter aproximado do modelo analítico, torna-se indispensável em qualquer situação a utilização do modelo de simulação, ao menos para alguns valores de tráfego convenientemente selecionados. Quando a precisão fornecida pelos métodos é comparável, é óbvio que os métodos da equivalência e da equivalência modificada são preferíveis ao método da superposição, por serem explícitos e mais simples sob o ponto de vista computacional.

No próximo capítulo, nós consideramos somente sistemas fracamente acoplados.

Tabela 5.1:

Resultados Obtidos Para o Sistema I
Método: Equivalência Entre Taxas de Serviço e Caminhada

a	P_0	c	$c^{(2)}$	c_e	$c_e^{(2)}$	w	*Modelo*	s
0,2	0,975	1,225	1,945			0,861	Analítico	1
0,4	0,933	1,600	3,720			1,362		
0,6	0,850	2,350	8,020			2,259		
0,8	0,600	4,600	26,920			5,688		
0,2	0,975	1,249	2,138			0,896	Simulação	
0,4	0,935	1,675	4,697			1,692		
0,6	0,854	2,483	11,095			3,014		
0,8	0,598	4,946	39,421			10,373		
0,2	0,972	1,225	1,945	0,613	0,486	0,517	Analítico	2
0,4	0,914	1,600	3,720	0,800	0,930	0,961		
0,6	0,777	2,350	8,020	1,175	2,005	2,020		
0,8	0,307	4,600	26,920	2,300	6,730	11,496		
0,2	0,972	1,248	2,088	0,624	0,562	0,637	Simulação	
0,4	0,918	1,679	4,467	0,836	1,130	1,180		
0,6	0,795	2,488	9,898	1,255	2,720	2,679		
0,8	0,419	4,790	32,325	2,414	9,195	10,908		
0,2	0,965	1,225	1,945	0,408	0,216	0,470	Analítico	3
0,4	0,856	1,600	3,720	0,533	0,413	1,250		
0,6	0,445	2,350	8,020	0,783	0,891	6,487		
0,8	Sistema Instável							
0,2	0,963	1,248	2,077	0,418	0,278	0,649	Simulação	
0,4	0,861	1,664	4,266	0,560	0,502	1,465		
0,6	0,616	2,536	9,943	0,838	1,138	4,225		
0,8	0,149	4,180	23,606	1,400	3,020	28,658		

Tabela 5.2:

Probabilidade de Fila Vazia (P_0) e Tempo Médio de Espera em Fila (w) Obtidos Para o Sistema I Em Função dos Valores Médios de Simulação
Método: Equivalência Entre Taxas de Serviço e Caminhada

a) Valores de P_0

a (erl)	s					
	1		2		3	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,975	0,975 [0,974; 0,976]	0,973	0,972 [0,968; 0,976]	0,971	0,963 [0,958; 0,968]
0,4	0,930	0,935 [0,928; 0,942]	0,921	0,918 [0,908; 0,928]	0,907	0,861 [0,846; 0,876]
0,6	0,842	0,854 [0,841; 0,867]	0,805	0,795 [0,783; 0,807]	0,747	0,616 [0,595; 0,637]
0,8	0,570	0,598 [0,567; 0,629]	0,447	0,419 [0,391; 0,447]	0,330	0,149 [0,123; 0,175]

b) Valores de w

a (erl)	s					
	1		2		3	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,890	0,896 [0,781; 1,011]	0,487	0,637 [0,591; 0,683]	0,372	0,649 [0,588; 0,710]
0,4	1,511	1,692 [1,450; 1,934]	0,784	1,180 [1,080; 1,280]	0,564	1,465 [1,252; 1,678]
0,6	2,547	3,014 [2,633; 3,395]	1,379	2,679 [2,317; 3,041]	0,985	4,225 [3,612; 4,838]
0,8	5,406	10,373 [8,830; 11,916]	3,062	10,908 [8,989; 12,827]	1,989	28,658 [21,917; 35,399]

Tabela 5.3:

Resultados Obtidos Para o Sistema III
Método: Equivalência Entre Taxas de Serviço e Caminhada

a	P_0	c	$c^{(2)}$	c_e	$c_e^{(2)}$	w	*Modelo*	s
0,2	0,975	12,475	160,514			6,626	Analítico	1
0,4	0,933	16,600	288,320			9,380		
0,6	0,850	24,850	644,995			15,452		
0,8	0,600	49,600	2523,521			43,078		
0,2	0,974	12,510	162,678			6,697	Simulação	
0,4	0,936	16,657	297,646			9,651		
0,6	0,852	24,909	678,009			16,202		
0,8	0,626	49,409	2640,557			45,097		
0,2	0,975	12,475	160,514	6,238	40,128	3,392	Analítico	2
0,4	0,932	16,600	288,320	8,300	72,080	4,917		
0,6	0,844	24,850	644,995	12,425	161,249	8,324		
0,8	0,580	49,600	2523,521	24,800	630,880	24,424		
0,2	0,974	12,501	160,991	6,250	64,692	5,530	Simulação	
0,4	0,930	16,597	289,225	8,296	121,831	7,424		
0,6	0,837	24,990	661,186	12,533	292,503	15,937		
0,8	0,577	48,736	2531,157	24,476	1156,451	45,054		
0,2	0,974	12,475	160,514	4,158	17,835	2,351	Analítico	3
0,4	0,929	16,600	288,320	5,533	32,036	3,544		
0,6	0,834	24,850	644,995	8,283	71,666	6,279		
0,8	0,543	49,600	2523,521	16,533	280,391	20,192		
0,2	0,975	12,457	157,950	4,162	37,179	4,745	Simulação	
0,4	0,929	16,656	289,754	5,540	70,578	7,118		
0,6	0,825	25,165	665,545	8,432	174,729	13,283		
0,8	0,518	48,318	2474,001	16,323	683,071	39,248		
0,2	0,973	12,475	160,514	3,119	10,032	1,861	Analítico	4
0,4	0,925	16,600	288,320	4,150	18,020	2,961		
0,6	0,819	24,850	644,995	6,213	40,312	5,594		
0,8	0,483	49,600	2523,521	12,400	157,720	20,897		
0,2	0,973	12,494	160,930	3,118	25,129	4,180	Simulação	
0,4	0,921	16,668	290,886	4,189	47,285	6,224		
0,6	0,805	24,918	656,569	6,354	115,065	12,416		
0,8	0,482	47,264	2376,659	12,045	412,824	32,130		
0,2	0,972	12,475	160,514	2,495	6,421	1,595	Analítico	5
0,4	0,918	16,600	288,320	3,320	11,533	2,718		
0,6	0,795	24,850	644,995	4,970	25,800	5,596		
0,8	0,387	49,600	2523,521	9,920	100,941	27,075		
0,2	0,971	12,494	161,040	2,505	18,581	4,074	Simulação	
0,4	0,919	16,661	292,704	3,358	34,088	5,664		
0,6	0,808	24,531	637,736	5,039	80,296	10,271		
0,8	0,439	47,242	2415,011	9,610	294,366	33,024		

Tabela 5.4:

Probabilidade de Fila Vazia (P_0) e Tempo Médio de Espera em Fila (w) Obtidos Para o Sistema III Em Função dos Valores Médios de Simulação
Método: Equivalência Entre Taxas de Serviço e Caminhada

a) Valores de P_0

a (erl)	s							
	2		3		4		5	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,975	0,974 [0,970;0,978]	0,974	0,975 [0,973;0,977]	0,973	0,973 [0,969;0,977]	0,974	0,971 [0,966;0,976]
0,4	0,931	0,930 [0,921;0,939]	0,929	0,929 [0,919;0,939]	0,927	0,921 [0,911;0,931]	0,926	0,919 [0,900;0,938]
0,6	0,839	0,837 [0,820;0,854]	0,829	0,825 [0,793;0,857]	0,823	0,805 [0,786;0,824]	0,815	0,808 [0,783;0,833]
0,8	0,565	0,577 [0,543;0,611]	0,517	0,518 [0,476;0,560]	0,513	0,482 [0,436;0,528]	0,492	0,439 [0,393;0,485]

b) Valores de w

a (erl)	s							
	2		3		4		5	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	5,292	5,530 [4,746;6,314]	4,579	4,745 [4,217;5,273]	4,062	4,180 [3,646;4,714]	3,821	4,074 [3,473;4,675]
0,4	7,761	7,424 [6,929;7,919]	6,776	7,118 [6,795;7,441]	6,033	6,224 [5,321;7,127]	5,425	5,664 [4,893;6,435]
0,6	13,088	15,937 [13,754;18,120]	11,748	13,283 [11,605;14,961]	10,314	12,416 [10,732;14,100]	9,220	10,271 [8,813;11,729]
0,8	31,058	45,054 [38,128;51,980]	28,155	39,248 [33,795;44,701]	23,015	32,130 [27,731;36,529]	20,674	33,024 [28,090;37,958]

Tabela 5.5:

Tempo Médio de Espera em Fila (w) Para o Sistema IV
Método: Equivalência Entre Taxas de Serviço e Caminhada

a (erl)	s							
	4		12		18		24	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	1,347	4,416 [4,370; 4,462]	0,494	0,957 [0,912; 1,002]	0,375	0,916 [0,890; 0,942]	0,337	0,839 [0,779; 0,899]
0,4	1,852	5,554 [5,418; 5,690]	0,743	1,013 [1,003; 1,023]	0,639	0,973 [0,939; 1,007]	0,682	0,856 [0,827; 0,885]
0,6	2,819	9,118 [8,905; 9,331]	1,239	1,163 [1,005; 1,321]	1,207	1,070 [1,026; 1,114]	1,561	0,983 [0,920; 1,046]
0,8	5,658	21,699 [20,378; 23,020]	2,742	2,855 [2,505; 3,203]	3,065	2,363 [2,095; 2,639]	5,074	3,384 [2,915; 3,853]

Tabela 5.6:

Resultados Obtidos Para o Sistema I
Método: Superposição dos Ciclos Condicionais
Procedimento de Superposição: Bhuyan

a	P_0	c	$c^{(2)}$	v	$v^{(2)}$	w	Modelo	s
0,2	0,975	1,250	2,068			0.870	Analítico	1
0,4	0,933	1,666	4,157			1,424		
0,6	0,850	2,500	9,481			2,565		
0,8	0,600	5,000	34,045			7,944		
0,2	0,975	1,249	2,138			0,896	Simulação	
0,4	0,935	1,675	4,697			1,692		
0,6	0,854	2,483	11,095			3,014		
0,8	0,598	4,946	39,421			10,373		
0,2	0,972	1,250	2,070	0,625	0,534	0,542	Analítico	2
0,4	0,914	1,666	4,156	0,833	0,977	1,033		
0,6	0,777	2,500	9,482	1,250	2,284	3,084		
0,8	Sistema Instável							
0,2	0,972	1,248	2,088	0,624	0,562	0,637	Simulação	
0,4	0,918	1,679	4,467	0,836	1,130	1,180		
0,6	0,795	2,488	9,898	1,255	2,720	2,679		
0,8	0,419	4,790	32,325	2,414	9,195	10,908		
0,2	0,965	1,250	2,069	0,417	0,267	0,540	Analítico	3
0,4	0,856	1,667	4,154	0,556	0,449	1,729		
0,6	0,445	2,500	9,477	0,834	1,142	163,699		
0,8	Sistema Instável							
0,2	0,963	1,248	2,077	0,418	0,278	0,649	Simulação	
0,4	0,861	1,664	4,266	0,560	0,502	1,465		
0,6	0,616	2,536	9,943	0,838	1,138	4,225		
0,8	0,149	4,180	23,606	1,400	3,020	28,658		

Tabela 5.7:

Probabilidade de Fila Vazia (P_0) e Tempo Médio de Espera em Fila (w) Obtidos Para o Sistema I Em Função dos Valores Médios de Simulação
Método: Superposição dos Ciclos Condicionais

a) Valores de P_0

a	s					
	1		2		3	
(erl)	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,975	0,975 [0,974; 0,976]	0,974	0,972 [0,968; 0,976]	0,973	0,963 [0,958; 0,968]
0,4	0,933	0,935 [0,928; 0,942]	0,928	0,918 [0,908; 0,928]	0,920	0,861 [0,846; 0,876]
0,6	0,849	0,854 [0,841; 0,867]	0,830	0,795 [0,783; 0,807]	0,803	0,616 [0,595; 0,637]
0,8	0,593	0,598 [0,567; 0,629]	0,554	0,419 [0,391; 0,447]	0,540	0,149 [0,123; 0,175]

b) Valores de w

a	s					
	1		2		3	
(erl)	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,897	0,896 [0,781; 1,011]	0,492	0,637 [0,591; 0,683]	0,383	0,649 [0,588; 0,710]
0,4	1,584	1,692 [1,450; 1,934]	0,810	1,180 [1,080; 1,280]	0,618	1,465 [1,252; 1,678]
0,6	3,077	3,014 [2,633; 3,395]	1,634	2,679 [2,317; 3,041]	1,246	4,225 [3,612; 4,838]
0,8	9,119	10,373 [8,830; 11,916]	5,113	10,908 [8,989; 12,827]	3,349	28,658 [21,917; 35,399]

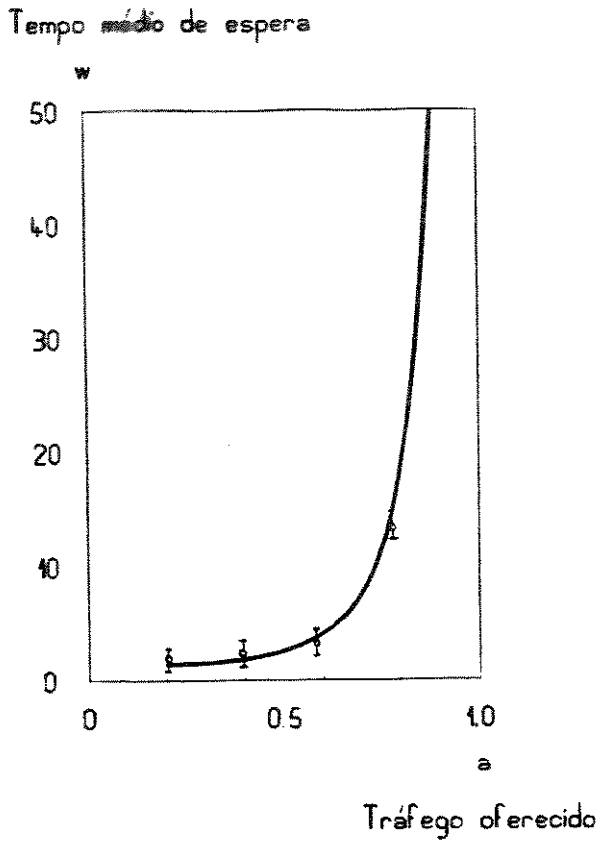


Figura 5.1:

Desempenho do Sistema II (Anel Simétrico de Bhuyan)
Método: Superposição dos Ciclos Condicionais
$\overline{\Phi}$: Resultado de Simulação Com 95% de Confiança

Tabela 5.8:

Valores do Tempo Médio de Espera em Fila (w) Obtidos Para o Sistema II Através da Aplicação dos Dois Métodos

	w		
	Método		
a	Equivalência	Superposição	Simulação
0,2	0,726	0,893	0,938 [0,818; 1,058]
0,4	1,329	1,478	1,618 [1,415; 1,821]
0,6	2,699	3,195	2,897 [2,518; 3,276]
0,8	8,988	22,246	19,859 [17,092; 22,626]

Tabela 5.9:

Probabilidade de Fila Vazia (P_0) e Tempo Médio de Espera em Fila (w) Obtidos Para o Sistema Sistema III Em Função dos Valores Médios de Simulação
Método: Superposição dos Ciclos Condicionais

a) Valores de P_0

a (erl)	s							
	2		3		4		5	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,975	0,974 [0,970;0,978]	0,974	0,975 [0,973;0,977]	0,974	0,973 [0,969;0,977]	0,974	0,971 [0,966;0,976]
0,4	0,932	0,930 [0,921;0,939]	0,930	0,929 [0,919;0,939]	0,929	0,921 [0,911;0,931]	0,928	0,919 [0,900;0,938]
0,6	0,842	0,837 [0,820;0,854]	0,834	0,825 [0,793;0,857]	0,833	0,805 [0,786;0,824]	0,825	0,808 [0,783;0,833]
0,8	0,568	0,577 [0,543;0,611]	0,536	0,518 [0,476;0,560]	0,568	0,482 [0,436;0,528]	0,584	0,439 [0,393;0,485]

b) Valores de w

a (erl)	s							
	2		3		4		5	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	5,509	5,530 [4,746;6,314]	4,775	4,745 [4,217;5,273]	4,335	4,180 [3,646;4,714]	3,988	4,074 [3,473;4,675]
0,4	8,574	7,424 [6,929;7,919]	7,504	7,118 [6,795;7,441]	6,667	6,224 [5,321;7,127]	5,971	5,664 [4,893;6,435]
0,6	15,953	15,937 [13,754;18,120]	14,271	13,283 [11,605;14,961]	12,400	12,416 [10,732;14,100]	11,137	10,271 [8,813;11,729]
0,8	48,758	45,054 [38,128;51,980]	43,576	39,248 [33,795;44,701]	32,883	32,130 [27,731;36,529]	27,303	33,024 [28,090;37,958]

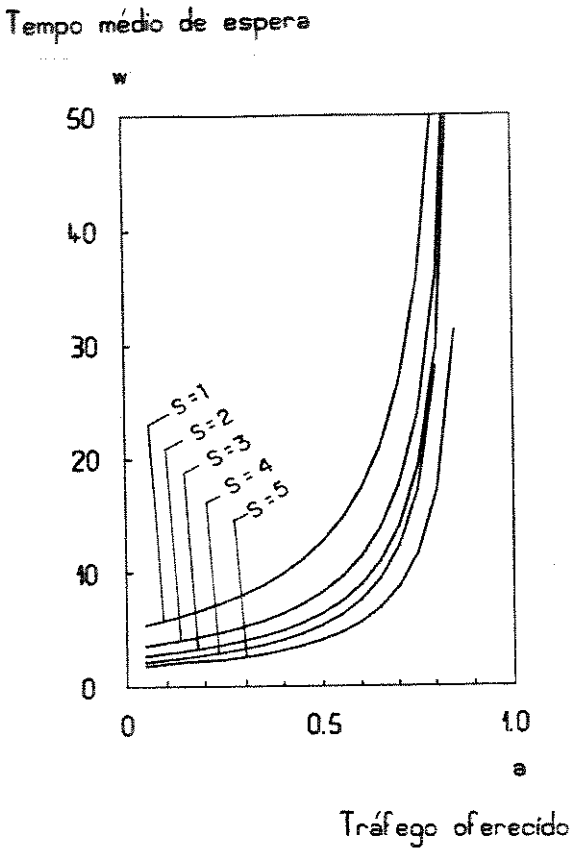


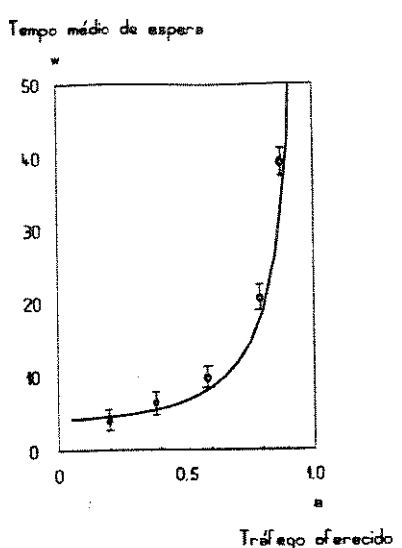
Figura 5.2:

Desempenho do Sistema III Para Diferentes Valores de Número de Servidores
Método: Superposição dos Ciclos Condicionais
Procedimento de Superposição: Bhuyan

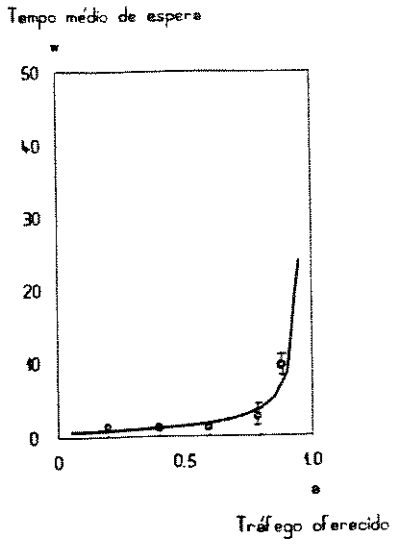
Tabela 5.10:

Valores do Tempo Médio de Espera em Fila (w) Obtidos Para o Sistema III Através dos Dois Métodos
Procedimento de Superposição: Bhuyan
E: Equivalência Entre Taxas de Serviço e Caminhada
S: Superposição dos Ciclos Condicionais

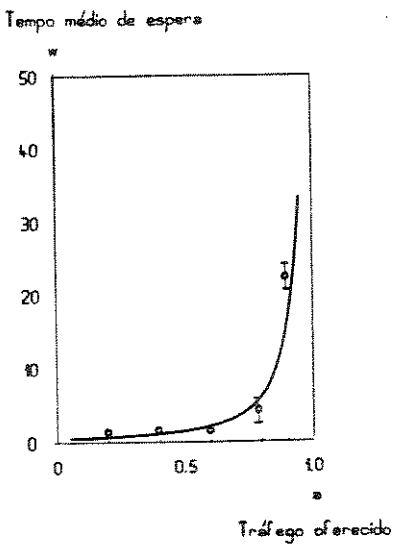
a (erl)	s									
	1		2		3		4		5	
	E	S	E	S	E	S	E	S	E	S
0,2	6,626	6,630	3,392	4,438	2,351	3,364	1,861	2,788	1,595	2,930
0,4	9,380	9,409	4,917	6,256	3,544	4,989	2,968	4,342	2,718	4,043
0,6	15,452	15,605	8,324	10,836	6,279	9,267	5,594	8,991	5,596	9,848
0,8	43,078	44,534	24,424	39,160	20,192	49,080	20,897	83,396	27,075	Instabilidade



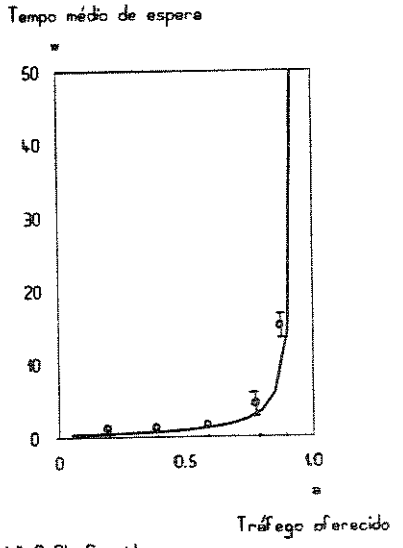
(a) S=4 Servidores



(b) S=12 Servidores



(c) S=18 Servidores



(d) S=24 Servidores

Figura 5.3:

Desempenho do Sistema Simétrico IV
$\pm$: Resultado de Simulação Com 95% de Confiança

Tabela 5.11:

Tempo Médio de Espera em Fila (w) Para o Sistema I
Método: Equivalência Modificada

a (erl)	s					
	1		2		3	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	0,861	0,896 [0,781; 1,011]	0,517	0,637 [0,591; 0,683]	0,470	0,649 [0,588; 0,710]
0,4	1,362	1,692 [1,450; 1,934]	0,961	1,180 [1,080; 1,280]	1,958	1,465 [1,252; 1,678]
0,6	2,259	3,014 [2,633; 3,395]	5,090	2,679 [2,317; 3,041]	25,173	4,225 [3,612; 4,838]
0,8	5,688	10,373 [8,830; 11,916]	Instável	10,908 [8,989; 12,827]	Instável	28,658 [21,917; 35,399]

Tabela 5.12:

Tempo Médio de Espera em Fila (w) Para o Sistema II
Métodos: Equivalência Modificada e Superposição dos Ciclos Condicionais

a	w		
	Método		
	Eq. Modificada	Superposição	Simulação
0,2	0,842	0,893	0,938 [0,818; 1,058]
0,4	1,466	1,478	1,618 [1,415; 1,821]
0,6	4,656	3,195	2,897 [2,518; 3,276]
0,8	21,076	22,246	19,859 [17,092; 22,626]

Tabela 5.13:

Tempo Médio de Espera em Fila (w) Para o Sistema III
Método: Equivalência Modificada

a (erl)	s							
	2		3		4		5	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	3,392	5,530 [4,746;6,314]	2,351	4,745 [4,217;5,273]	2,487	4,180 [3,646;4,714]	1,997	4,074 [3,473;4,675]
0,4	4,917	7,424 [6,929;7,919]	5,453	7,118 [6,795;7,441]	4,021	6,224 [5,321;7,127]	4,754	5,664 [4,893;6,435]
0,6	19,969	15,937 [13,754;18,120]	10,321	13,283 [11,605;14,961]	14,130	12,416 [10,732;14,100]	11,137	10,271 [8,813;11,729]
0,8	174,393	45,054 [38,128;51,980]	Instável	39,248 [33,795;44,701]	Instável	32,130 [27,731;36,529]	Instável	33,024 [28,090;37,958]

Tabela 5.14:

Tempo Médio de Espera em Fila (w) Para o Sistema IV
Método: Equivalência Modificada

a (erl)	s							
	4		12		18		24	
	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação	Analítico	Simulação
0,2	1,779	4,416 [4,370;4,462]	0,589	0,957 [0,912;1,002]	0,448	0,916 [0,890;0,942]	0,424	0,839 [0,779;0,899]
0,4	2,439	5,554 [5,418;5,690]	1,100	1,013 [1,003;1,023]	1,033	0,973 [0,939;1,007]	1,083	0,856 [0,827;0,885]
0,6	5,521	9,118 [8,905;9,331]	2,932	1,163 [1,005;1,321]	2,704	1,070 [1,026;1,114]	3,783	0,983 [0,920;1,046]
0,8	11,301	21,699 [20,378;23,020]	11,476	2,855 [2,507;3,203]	15,202	2,363 [2,095;2,631]	24,237	3,384 [2,915;3,853]

Capítulo 6

APLICAÇÕES

Para ressaltar a utilidade dos modelos como ferramentas de auxílio na engenharia de sistemas, nós aplicamos o modelo analítico à avaliação de desempenho de estruturas de interligação de processadores reais. O auxílio é prestado principalmente na fase de concepção de um sistema, quando são avaliadas e comparadas, sob a óptica de desempenho, as diversas soluções que atendem aos requisitos funcionais predeterminados.

Os sistemas de comutação de alto desempenho empregados em computadores de uso geral ou em telecomunicações se caracterizam pela alta taxa de troca de mensagens. No caso de um Ponto de Transferência de Sinalização (PTS) de uma rede de sinalização por canal comum, por exemplo, se estima um aumento de 10.000–20.000 mensagens/seg para a rede telefônica atual para 100.000–500.000 mensagens/seg para a rede inteligente.

O principal objetivo deste capítulo é investigar a capacidade de troca de mensagens interprocessadores na solução multi-servidores cíclicos, supondo que os sistemas usam uma grande quantidade de processadores, digamos acima de 100.

Na Seção 6.1, nós apresentamos um método de obtenção da taxa de sinalização interprocessadores em um sistema de comutação telefônica e aplicamos o modelo na avaliação do desempenho da estrutura de interligação do sistema TRÓPICO. Nas Seções 6.2 e 6.3, respectivamente, nós investigamos o desempenho da estrutura de interligação de um PTS e de um sistema de comutação de mensagens de alto desempenho. Nós empregamos o método da superposição dos ciclos condicionais, que foi o método que apresentou melhores aproximações nesses exemplos particulares.

6.1 Sistema de Comutação TRÓPICO

Os assinantes, quando conectados a um sistema de comutação, não podem perceber nenhum atraso significativo na execução das fases de uma chamada te-

lefônica. Dito de outra forma, eles devem ter uma ilusão de instantaneidade de serviço. Esta é conseguida estipulando-se que a probabilidade do tempo de execução de uma fase ultrapassar um dado valor predeterminado seja bem pequena. Por exemplo, o tempo de espera pelo tom de discar deve ser inferior a 1 seg em 99,5% dos casos e inferior a 3 seg em 99,9% dos casos, sob condições nominais. Sob condições de sobrecarga, ele deve ser inferior a 3 seg em 99% dos casos. Requisitos análogos são estipulados para outras fases da chamada telefônica [40]. Conseqüentemente, a estrutura de sinalização deve dar uma disponibilidade suficiente, tal que qualquer efeito de bloqueio possa ser considerado desprezível e os tempos médios de transferência de mensagens entre dois processadores sejam relativamente pequenos.

Os processadores de um sistema de comutação TRÓPICO são divididos em grupos (vide Apêndice B). Cada grupo desempenha uma função bem específica. O número de processadores de cada grupo em um sistema que deve escoar um determinado tráfego telefônico é dimensionado de acordo com padrões de qualidade de serviço internacionalmente aceitos. Os padrões são relativos a esperas pela execução de funções e probabilidades de perda de chamada. Os requisitos recomendados pelo CCITT[40] são normalmente adotados.

Sob o ponto de vista de troca de mensagens para a realização das funções telefônicas, podemos considerar que a atratividade de sinalização entre os processadores de um mesmo grupo é simétrica. No entanto, a atratividade entre os processadores pertencentes a grupos distintos é assimétrica. Existem, inclusive, grupos de processadores que nunca se comunicam para a realização de determinadas funções telefônicas. Além disso, a carga de trabalho não é distribuída eqüitativamente entre os grupos. Conseqüentemente, as taxas médias de geração de mensagens por grupo são distintas. Assim, o encaminhamento das mensagens entre os processadores do sistema e os processos de chegada das mensagens são assimétricos.

Denotando por ς_{ij} , $i,j = 1,2,\ldots,m$ a taxa média de sinalização originada em um processador i e destinada a um processador j, nós temos que $\varsigma_{ii} = 0\ \forall i$ e $\lambda_i = \sum_{j=1}^{m} \varsigma_{ij}$. A taxa média global de sinalização é $\lambda_T = \sum_{i=1}^{m} \lambda_i$.

Cada ς_{ij} é obtida por inspeção dos organogramas de tratamento de chamadas telefônicas e procedimentos de operação, manutenção e supervisão. Nos pontos onde existem um processador solicitante de um recurso e vários processadores que contêm o recurso solicitado, o número de tentativas para a obtenção desse recurso é uma variável aleatória. Assim, se há n processadores que tratam r recursos cada um e a ocupação dos recursos é aleatória, a probabilidade de encontrar p processadores com todos os recursos ocupados, $G_n(p)$, $p = 0,1,\ldots,n$, é dada por [43]

$$G_n(p) = \binom{n}{p} \sum_{k=0}^{n-p} (-1)^k \binom{n-p}{k} H(p \times r + k \times r)$$

onde

- $H(x) = \sum_{t=x}^{n\times r} \dfrac{\binom{n\times r - x}{t-x}}{\binom{n\times r}{t}} L(t)$ é a probabilidade que x recursos predeterminados estejam livres e

- $L(t)$, $t = 0, 1, \ldots, n \times r$, é a distribuição do número de recursos ocupados. Normalmente, em uma aplicação particular, nós assumimos a distribuição de Erlang, binomial ou Engset [32],[35],[43].

A distribuição do número de tentativas para a obtenção de um recurso é obtida em função de $G_n(p)$. Conseqüentemente, obtemos também a distribuição do número de mensagens trocadas entre o processador solicitante e os processadores tratadores, considerando que em cada tentativa há uma mensagem de solicitação de recurso e uma mensagem de resposta (positiva ou negativa). Através da distribuição do número de mensagens trocadas, a taxa ζ_{ij} é obtida diretamente. A partir das taxas médias de sinalização interprocessadores, nós obtemos as taxas médias de sinalização e as probabilidades de encaminhamento intergrupos.

Um sistema de comutação telefônica trata diversos procedimentos operacionais e diversas classes de chamadas, de acordo com os dados de especificação de características do projeto. A cada classe de chamada telefônica estão associados vários tipos de chamadas, em função do comportamento dos assinantes e do estado de suas linhas. Sem perda de generalidade, consideramos que a troca de mensagens interprocessadores para o processamento das chamadas telefônicas é preponderante em relação às demais funções operacionais na hora de maior movimento. Portanto, nós desprezamos a troca de mensagens relativa a essas funções operacionais. Assim sendo, uma questão que surge é: para um tráfego de voz comutado pela central, A_0 (erl), qual é a máxima taxa de troca de mensagens interprocessadores?

Supomos que há C classes de chamadas ($C \geq 1$), sendo $\delta_i \times 100\%$ do tipo i, $i = 1, 2, \ldots, C$. Os respectivos tempos médios de retenção, obtidos em função de todos os tipos associados a cada classe, são denotados por μ_i^{-1} (seg) e o número médio de mensagens trocadas interprocessadores são denotados por ℓ_i. Nessas condições, a taxa média global de geração de mensagens, λ_T, é dada por

$$\lambda_T = A_0 \sum_{i=1}^{C} \delta_i \mu_i \ell_i \quad \text{(mensagens/seg)}.$$

Os requisitos de projeto de uma central estabelecem que as proporções de cada classe de chamada e os respectivos tempos médios de retenção devem se situar entre valores máximos e mínimos pré-estabelecidos, que denotamos, respectivamente, por

$$\delta_{i,min}, \delta_{i,max} \quad \text{e} \quad \mu_{i,min}^{-1}, \mu_{i,max}^{-1}.$$

Assim, a fim de obter a taxa média global máxima de sinalização, devemos resolver o *problema de programação não linear*:

maximizar

$$\lambda_T = A_0 \sum_{i=1}^{C} \delta_i \mu_i \ell_i$$

sujeito a

$$\mu_{i,min}^{-1} \leq \mu_i^{-1} \leq \mu_{i,max}^{-1}$$

$$\delta_{i,min} \leq \delta_i \leq \delta_{i,max}$$

$$\sum_{i=1}^{C} \delta_i = 1$$

$$\mu_i^{-1} > 0$$

$$\delta_i \geq 0$$

$$\delta_{i,max} \geq \delta_{i,min} \geq 0$$

$$\mu_{i,max}^{-1} \geq \mu_{i,min}^{-1} > 0$$

$$A_0 > 0$$

$$\forall i = 1, 2, \ldots, C.$$

Este problema é resolvido através da aplicação da condição necessária de Kuhn-Tucker. Conforme a intuição, o valor máximo de λ_T é obtido para $\mu_i^{-1} = \mu_{i,min}^{-1}$ e $\delta_i = \delta_{i,min}$ ou $\delta_{i,max}$, dependendo dos valores relativos dos números de mensagens trocadas em cada classe, $i = 1, 2, \ldots, C$.

Como um exemplo de aplicação, consideremos um sistema de comutação tandem-local TRÓPICO, com as seguintes características básicas:

- $A_0 = 12.200$ erl
- 64.320 terminais de assinante
- 11.726 juntores
- capacidade de equipar até 768 processadores
- $g = 7$ grupos de processadores, sendo:

Grupo	Quantidade	Tipo
1	402	assinante
2	32	enviador/receptor de sinalização MFC
3	118	juntor de entrada
4	25	registrador
5	118	juntor de saída
6	32	comutação
7	1	operação e manutenção

Assumimos que todos os processadores são do tipo I, embora nos sistemas reais os processadores do grupo 4 sejam do tipo II (vide Apêndice A). O sistema deve tratar $C = 5$ classes de chamadas telefônicas. Aplicando o procedimento anteriormente descrito, nós determinamos que as porcentagens ($\delta_i \times 100\%$) e os tempos médios de retenção (μ_i^{-1}) ótimos relativos a cada classe são

Classe	$\delta_i \times 100\%$	μ_i^{-1} (seg)
intracentral	5	70
saída local	7,5	70
saída interurbana	40	80
entrada	40	70
trânsito	7,5	80

a matriz de probabilidades de encaminhamento entre os grupos, β_{kl}, $k, l = 1, 2, \ldots, 7$, é dada por

$$\begin{bmatrix} 0,07 & 0 & 0,10 & 0,65 & 0,03 & 0,10 & 0,05 \\ 0 & 0 & 0,35 & 0 & 0,51 & 0,14 & 0 \\ 0,04 & 0,43 & 0 & 0,48 & 0 & 0,03 & 0,02 \\ 0,42 & 0 & 0,20 & 0 & 0,33 & 0,05 & 0 \\ 0,06 & 0,44 & 0 & 0,50 & 0 & 0 & 0 \\ 0,21 & 0,29 & 0,14 & 0,36 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

e, denotando por Δ_k, $0 \leq \Delta_k \leq 1$, $\sum_{k=1}^{7} \Delta_k = 1$, a proporção do tráfego oferecido por servidor pelo grupo k em relação ao tráfego total oferecido por servidor, a, as proporções são

Grupo	$\Delta_k \times 100\%$
1	17,483
2	16,583
3	13,387
4	29,071
5	19,281
6	4,196
7	0,000

Assim, se a_k representa o tráfego oferecido por servidor por grupo, resulta que $a_k = \Delta_k a$, $k = 1, 2, \ldots, 7$. Repare que o processador do grupo 7 (operação e manutenção) não envia mensagens de processamento de chamadas telefônicas a nenhum grupo. Entretanto, ele recebe mensagens dos processadores dos grupos 1 e 3.

Para o sistema que estamos considerando as mensagens têm comprimento

variável, com a seguinte distribuição percentual:

Tamanho (*bytes*)	Porcentagem (%)
32	80
64	10
128	6
256	3
512	1

O tempo de serviço é composto pelo tempo de ocupação de uma via durante a transmissão da mensagem e pelo tempo de liberação do processador transmissor e do processador receptor. Considerando a capacidade de processamento dos processadores utilizados no sistema TRÓPICO e as taxas de transmissão das vias, para uma mensagem de comprimento igual a b *bytes*, nós temos que o tempo de serviço é dado por

$$[1010 + 81(b-3)] \times 133 \times 10^{-6} + 4(b-1) \times 10^{-3} \quad \text{ms.}$$

O processo de serviço é suposto simétrico. Nessas condições, nós temos que $h = 0,899$ ms e $h^{(2)} = 0,978$ ms^2. O processo de caminhada também é suposto simétrico. O tempo de caminhada é composto pelo tempo de varredura de um canal de solicitação e pelo tempo de habilitação do processador transmissor. Ele pode ser considerado constante, com $u = 8$ μs.

A taxa média global de sinalização máxima em condições nominais de operação é $\lambda_{T_{nom}} = 12.220$ mensagens/seg. Portanto, a taxa média global de sinalização máxima em condições de sobrecarga, que corresponde a um acréscimo de 40% em relação à taxa nominal, é $\lambda_{T_{sob}} = 17.108$ mensagens/seg. Nós definimos o tempo médio de espera ponderado, $\overline{w}$, por $\overline{w} = \sum_{k=1}^{7} \Delta_k w_k$, onde w_k denota o tempo médio de espera em fila no grupo k, $k = 1, 2, \ldots, 7$. Nós devemos ter que $\overline{w}$ em condições nominais e em condições de sobrecarga, $\overline{w}_{nom}$ e $\overline{w}_{sob}$, respectivamente, devem satisfazer às restrições: $\overline{w}_{nom} \leq 5$ ms e $\overline{w}_{sob} \leq 10$ ms.

Assumimos que cada plano de sinalização é implementado com três enlaces de comunicação, sendo que cada um desses enlaces é composto por duas vias de comunicação multiplexadas no tempo: uma via de transmissão e uma via de recepção. Através da aplicação do modelo analítico, onde empregamos o procedimento de superposição de Bhuyan, nós verificamos que o número mínimo de planos de um sistema, tal que que os requisitos de espera sejam satisfeitos, é igual a três. Nessas condições, nós temos que $s = 18$ servidores e os resultados para todos os grupos são apresentados na Tabela 6.1.

Donde, $\overline{w}_{nom} = 1,759$ ms e $\overline{w}_{sob} = 9,160$ ms, ou seja, os requisitos em condições nominais e em condições de sobrecarga são satisfeitos. A verificação do desempenho da estrutura de sinalização de qualquer outro sistema TRÓPICO é feita de maneira análoga. Basta seguir os passos:

Tabela 6.1:

Valores do Tempo Médio de Espera em Fila (w) Para o Sistema Trópico Com 728 Processadores
$h = 0,899$ ms $h^{(2)} = 0,978$ ms^2 $u = 8$ μs
Procedimento de Superposição: Bhuyan

Grupo	w_k	
	nominal	sobrecarga
1	1,598	8,321
2	1,785	9,295
3	1,530	7,967
4	1,989	10,357
5	1,649	8,586
6	1,972	10,268
7	0,000	0,000

1. calcular o número de processadores por grupo, de modo que o sistema comute um tráfego telefônico e satisfaça aos padrões de qualidade de serviço estipulados;

2. calcular as probabilidades de encaminhamento intergrupos de processadores;

3. para as classes de chamadas que o sistema deve tratar, determinar as proporções dos tráfegos oferecidos por grupo e as taxas médias globais máximas de sinalização em condições nominais e em sobrecarga;

4. para o perfil de distribuição dos comprimentos das mensagens de processamento de chamadas, determinar os parâmetros relativos aos tempos de serviço e de caminhada;

5. assumindo uma dada configuração de um plano de sinalização, aplicar o modelo analítico, empregando um dos métodos de superposição, a fim de determinar o valor mínimo de s, tal que os requisitos relativos aos tempos médios em fila ponderados em condições nominais e em sobrecarga sejam satisfeitos.

Este procedimento pode ser generalizado diretamente para a verificação do desempenho de qualquer sistema que se enquadre na classe dos sistemas de filas cíclicas que estamos considerando.

6.2 Ponto de Transferência de Sinalização

Um ponto de transferência de sinalização (PTS) é um importante componente de uma rede de sinalização por canal comum (SCC) [39], [41]. Uma SCC é utilizada para comutar informações sobre o estabelecimento e liberação de conexões entre circuitos, serviços, operação e manutenção de uma rede telefônica.

Um PTS deve operar com um alto desempenho, para ser capaz de tratar milhares de mensagens por segundo e satisfazer a requisitos estritos relativos a tempos de transferência de mensagens. Nós definimos o tempo de transferência de uma mensagem como sendo o intervalo de tempo decorrido desde o instante em que a mensagem é recebida de um enlace de sinalização de entrada até o instante em que ela é encaminhada a outro PTS, através de um enlace de sinalização de saída. Conforme especificação do sistema CCITT número 7 [41], o tempo médio de transferência de mensagem deve ser no máximo igual a 20 ms, em condições nominais de operação. Para um acréscimo de 15% na carga nominal, que corresponde ao primeiro nível de sobregarga, o tempo médio de transferência deve ser no máximo igual a 40 ms. Finalmente, para um acréscimo de 30% na carga nominal, que corresponde ao segundo nível de sobrecarga, o tempo médio de transferência deve ser no máximo igual a 100 ms.

Nós supomos que o PTS que desejamos analisar utiliza a mesma estrutura de interligação de processadores do sistema TRÓPICO. Um importante componente do tempo de transferência é o tempo de espera da mensagem no *buffer* de transmissão. Assumimos, também, que:

- o PTS é simétrico;
- são equipados 128 processadores do tipo I;
- a capacidade máxima de processamento de cada processador é igual a 100 mensagens/seg;
- cada plano de sinalização tem três enlaces de comunicação e cada enlace tem duas vias (uma de transmissão e uma de recepção);
- a distribuição percentual do comprimento das mensagens é

Comprimento (*byte*)	Percentual (%)
39	46,2
40	31,7
42	7,1
45	7,9
48	7,1

Tabela 6.2:

Valores do Tempo Médio de Espera em Fila (w) Para os 3 Níveis de Carga no PTS Com 128 Processadores
$h = 0,706$ ms $h^{(2)} = 0,500$ ms^2 $u = 8$ μs
Procedimento de Superposição: Bhuyan

Condição	λ_T (mensagens/seg)	w (ms)
nom	8.648	1,242
1*sob*	9.946	3,504
2*sob*	11.243	25,000

- o tempo médio de espera de mensagem no *buffer* de transmissão corresponde a $\frac{1}{4}$ do tempo de transferência de mensagem. Assim, denotando, respectivamente, o tempo médio de espera no *buffer* em condições nominais, no primeiro e no segundo nível de sobrecarga por w_{nom}, w_{1sob} e w_{2sob}, nós devemos ter que $w_{nom} \leq 5$ ms, $w_{1sob} \leq 10$ ms e $w_{2sob} \leq 25$ ms.

Nessas condições, temos um sistema simétrico com $m = 128$ nós, $h = 0,706$ ms, $h^{(2)} = 0,500$ ms^2 e $u = 8$ μs (constante). Portanto, fixado o número de planos de sinalização a serem implementados no PTS, devemos determinar as taxas médias globais de troca de mensagens interprocessadores em condições nominais, no primeiro e no segundo nível de sobrecarga, denotadas, respectivamente, por $\lambda_{T_{nom}}$, $\lambda_{T_{1sob}}$ e $\lambda_{T_{2sob}}$, tal que um, e apenas um, dentre os três seguintes conjuntos de restrições seja satisfeito:

1. $w_{nom} \leq 5$ ms, $w_{1sob} \leq 10$ ms e $w_{2sob} = 25$ ms
2. $w_{nom} \leq 5$ ms, $w_{1sob} = 10$ ms e $w_{2sob} < 25$ ms
3. $w_{nom} = 5$ ms, $w_{1sob} < 10$ ms e $w_{2sob} < 25$ ms.

Inicia-se a busca pelo primeiro conjunto de restrições. Caso ele não seja satisfeito, passa-se ao segundo. Se este também não é satisfeito, passa-se, finalmente, ao terceiro conjunto de restrições. Por exemplo, se há dois planos equipados, então $s = 12$ servidores e, aplicando o modelo analítico com o procedimento de superposição de Bhuyan, nós obtivemos os resultados apresentados na Tabela 6.2.

Neste caso, o primeiro conjunto de restrições é satisfeito e a restrição relativa à capacidade máxima de processamento de um processador não é violada. Assim, a capacidade nominal do PTS é $\lambda_{T_{nom}} = 8.648$ mensagens/seg. Se o PTS fosse assimétrico, teria que ser adotado um critério de desempenho associado ao tempo

médio de espera em fila ponderado, analogamente ao que foi feito no caso da estrutura de sinalização do sistema TRÓPICO. A determinação da capacidade nominal de sinalização de qualquer outro PTS é feita de acordo com o mesmo procedimento que foi aqui ilustrado.

6.3 Sistema de Comutação de Mensagens de Alto Desempenho

Nós consideramos, agora, a aplicação do modelo e as suas extensões para os casos de transmissão *full duplex* e acessibilidade total dos servidores a um nó à análise de um sistema de comutação de mensagens de alto desempenho. Usamos a nomenclatura: modelo com acessibilidade parcial e transmissão unidirecional ou bidirecional e modelo com acessibilidade total, conforme for o caso.

Supomos que se tenha um sistema simétrico com m unidades de processamento e, de acordo com as disponibilidades tecnológicas de momento, sejam conhecidos os parâmetros relativos aos tempos de transmissão das mensagens, tempos de varredura ou de passagem das fichas ou quadros, etc. e o número máximo de planos de sinalização que podem ser implementados. Este último é conseqüência direta do número máximo de portas de acesso aos planos que podem ser colocadas em uma unidade de processamento. Com relação ao modelo, nós temos, então, que são conhecidos os valores de h, $h^{(2)}$, u, $u^{(2)}$ e s_{max}, onde s_{max} denota o valor máximo de s.

Fixado um valor limite admissível para o tempo médio de permanência das mensagens no *buffer* de transmissão, digamos w_0, algumas questões a serem respondidas são:

1. para $s = 1, 2, \ldots, s_{max}$ quais as taxas globais médias de sinalização interprocessadores, λ_{T_0}, onde $\lambda_{T_0} = \lambda_T|_{w=w_0}$, para cada tipo de acesso e transmissão?

2. qual a taxa global de sinalização interprocessadores máxima para cada tipo de acesso e transmissão, ou seja, $\lambda_{T_0}|_{s=s_{max}}$?

3. para cada valor de s quais os ganhos relativos nas taxas de sinalização globais ao se mudar de um tipo de acesso e transmissão para outro? Definimos o ganho relativo à mudança de a para b por

$$\Delta_{b/a} = \frac{\lambda_{T_{0,b}} - \lambda_{T_{0,a}}}{\lambda_{T_{0,a}}} \times 100\% \ ,$$

onde a e b denotam o tipo de acesso e transmissão.

Como um exemplo de aplicação particular, consideremos um sistema com $m = 1.024$ unidades de processamento e onde é utilizada a mesma estrutura de

sinalização do sistema TRÓPICO. Supomos que a distribuição do comprimento das mensagens é Poissoniana, com média igual a 64 *bytes*, e $s_{max} = 24$. Nessas condições, nós temos que $h = 1,051$ ms, $h^{(2)} = 2,157$ ms^2 e $u = 8$ μs (constante), ou seja, o tempo de serviço tem distribuição exponencial negativa e o tempo de caminhada é constante. Nós utilizamos o procedimento de superposição de Bhuyan. Nas tabelas 6.3 e 6.4, respectivamente, nós apresentamos as taxas globais de sinalização e os ganhos relativos, para cada valor de s e $w_0 = 10$ ms. A partir dos resultados obtidos, nós observamos que:

- as taxas globais de sinalização máximas do sistema, $\lambda_{T_0}|_{s=24}$, para acessibilidade parcial e transmissão unidirecional e bidirecional e acessibilidade total são iguais, respectivamente, a 17.810, 21.650 e 21.900 mensagens/seg;
- não há ganho relativo significativo, para nenhum valor de s, ao se mudar da acessibilidade parcial e transmissão bidirecional para acessibilidade total ($\Delta_{acessibilidade\ total/bidirecional} < 1,5\% \quad s = 1,2,\ldots,24$);
- a mudança de transmissão unidirecional para bidirecional não é vantajosa para $s = 1,2,\ldots,9$. Para $s > 9$ o ganho relativo passa a aumentar à medida que s aumenta, atingindo valores significativos para $s > 17$ ($\Delta_{bidirecional/unidirecional} > 10\%$).

Estas e outras constatações análogas constituem uma orientação objetiva aos projetistas de sistemas, no sentido de se escolher o tipo de acesso e transmissão mais apropriado, sob o ponto de vista de desempenho. No caso particular em estudo, por exemplo, não valeria a pena adotar a acessibilidade total das unidades de processamento aos planos, desde que ela é a solução mais cara dentre as três e não apresenta ganhos relativos significativos em relação à acessibilidade parcial e transmissão bidirecional. Para aumentar o desempenho do sistema, considerando $w_0 = 10$ ms, seria suficiente adotar a transmissão bidirecional em vez da unidirecional.

Consideremos, agora, mensagens com comprimento médio igual a 32, 64, 128, 256 e 512 *bytes* e distribuição de Poisson, mantendo constantes os demais parâmetros do sistema. O objetivo é verificar o desempenho para s e w_0 fixados. Os dois primeiros momentos do tempo de serviço em cada caso são dados, então, por:

Comprimento (*bytes*)	h (ms)	$h^{(2)}$ (ms^2)
32	0,579	0,914
64	1,051	2,157
128	1,997	5,985
256	3,888	19,004
512	7,670	66,495

Na Tabela 6.5 nós apresentamos a taxa de sinalização média global interprocessadores, fixando-se $s = 18$ servidores e $w_0 = 10$ ms, para cada valor de comprimento médio de mensagem. Repare que:

- a taxa global de sinalização interprocessadores é reduzida drasticamente aumentando-se o comprimento médio das mensagens, para cada tipo de acesso e transmissão;
- não há ganho relativo significativo, para nenhum valor de comprimento médio de mensagem, ao se mudar de acessibilidade parcial e transmissão bidirecional para acessibilidade total;
- o ganho relativo à mudança de transmissão unidirecional para transmissão bidirecional diminui à medida que aumenta o comprimento médio de mensagem.

Também neste caso, para aumentar o desempenho do sistema, considerando $s = 18$ e $w_0 = 10$ ms, é suficiente passar de transmissão unidirecional para transmissão bidirecional, qualquer que seja o comprimento médio de mensagem dentre aqueles que estamos considerando.

Tabela 6.3:

Taxas de Sinalização Globais no Sistema Com 1.024 Processadores Para Tempo Médio de Espera em Fila (w_0) Igual a 10 ms
$h = 1,051$ ms $h^{(2)} = 2,157$ ms^2 $u = 8$ μs
Procedimento de Superposição: Bhuyan

	λ_{T_0} (mensagens/seg)		
s	unidirecional	bidirecional	acessibilidade total
1	561	561	561
2	1.370	1.371	1.378
3	2.240	2.254	2.255
4	3.140	3.159	3.170
5	4.020	4.067	4.080
6	4.910	4.976	4.997
7	5.826	5.896	5.927
8	6.661	6.850	6.880
9	7.536	7.782	7.830
10	8.373	8.704	8.745
11	9.209	9.630	9.677
12	9.990	10.506	10.615
13	10.809	11.500	11.545
14	11.586	12.390	12.460
15	12.344	13.346	13.420
16	13.087	14.300	14.377
17	13.774	15.200	15.288
18	14.469	16.100	16.240
19	15.012	17.000	17.200
20	15.699	17.980	18.119
21	16.373	18.883	19.080
22	16.940	19.784	20.044
23	17.295	20.736	20.900
24	17.810	21.650	21.900

Tabela 6.4:

Ganhos Relativos Percentuais Às Mudanças de Tipos de Acesso e Transmissão Para o Sistema Com 1.024 Processadores

s	$\Delta_{b/a}$ (%)		
	bidirecional/unidirecional	acessibilidade total/bidirecional	acessibilidade total/unidirecional
1	0,00	0,00	0,00
2	0,07	0,51	0,58
3	0,63	0,04	0,67
4	0,61	0,35	0,96
5	1,17	0,32	1,49
6	1,34	0,42	1,77
7	1,20	0,53	1,73
8	2,84	0,44	3,29
9	3,26	0,62	3,90
10	3,95	0,47	4,44
11	4,57	0,49	5,08
12	5,17	1,04	6,26
13	6,39	0,39	6,81
14	6,94	0,56	7,54
15	8,12	0,55	8,72
16	9,27	0,54	9,86
17	10,35	0,58	10,99
18	11,27	0,87	12,24
19	13,24	1,18	14,58
20	14,53	0,77	15,41
21	15,33	1,04	16,53
22	16,79	1,31	18,32
23	19,90	0,79	20,84
24	21,56	1,15	22,96

Tabela 6.5:

Taxas de Sinalização Globais Interprocessadores No Sistema Com 1.024 Processadores $s = 18$ servidores e $w_0 = 10$ ms

Comprimento ($bytes$)	λ_{T_0} (mensagens/seg)		
	unidirecional	bidirecional	acessibilidade total
32	25.640	29.000	29.200
64	14.469	16.100	16.240
128	7.752	8.500	8.600
256	4.045	4.390	4.420
512	2.070	2.230	2.240

Capítulo 7

CONCLUSÕES

Neste trabalho, nós apresentamos um modelo analítico aproximado e um modelo de simulação apropriados para a avaliação do desempenho de uma estrutura de interligação de processadores em barramentos múltiplos com alocador centralizado, mas que, juntamente com as suas respectivas extensões, se aplicam à análise de uma ampla classe de estruturas de interligação de processadores.

O caráter aproximado do modelo analítico é devido a certas hipóteses que foram adotadas (H_{10}-H_{26}). Nesse modelo, é desenvolvida explicitamente, em relação a uma fila de transmissão j, a função geradora das probabilidades de estado estacionárias, a transformada de Stieltjes-Laplace da distribuição do tempo de espera e o tempo médio de espera dos usuários em fila. Nós aplicamos o enfoque de cadeias de Markov imersas, o conceito de independência entre segmentos de tempos de ciclo dos servidores e a probabilidade de visita bem sucedida de um servidor a um nó. O modelo pode ser generalizado para tratar outras classes de sistemas, tais como sistemas de filas com capacidade limitada, acessibilidade total dos servidores a um nó, acessibilidade parcial com transmissão bidirecional ou servidores operando no modo espera.

O modelo analítico dá bons resultados para sistemas com um grande número de filas e servidores e tempo de caminhada pequeno em relação ao tempo de serviço, que são as mesmas características da classe dos sistemas que têm motivado as nossas investigações. O modelo de simulação é utilizado para propósitos de validação do modelo analítico. Ele se aplica bem a sistemas com pequeno número de filas e servidores, desde que, nessas situações, o tempo de execução de uma rodada é relativamente pequeno.

A principal idéia do modelo analítico é a criação de um servidor que se comporta de uma maneira equivalente ao conjunto dos servidores originais, visto por um nó particular. A partir dos conceitos de tempo de ciclo do servidor equivalente, que é o período de tempo que o servidor equivalente se ausenta de um nó após uma visita bem ou mal sucedida, e probabilidade de visita bem sucedida a um nó (ρ), que é a probabilidade de um servidor original prestar serviço quando ele

visita uma fila de transmissão, nós generalizamos ao caso multi-servidor o modelo de Hashida e Ohara [3] relativo a servidor único em férias e serviço do tipo um-por-vez. Nós determinamos, então, o tempo médio de espera em fila dado pela expressão em (3.9). A avaliação de ρ e dos dois primeiros momentos do tempo de serviço e do tempo de ciclo do servidor equivalente (h_e, $h_e^{(2)}$, c_e e $c_e^{(2)}$) é aproximada. Conseqüentemente, o tempo médio de espera também é aproximado.

O parâmetro ρ (avaliado de acordo com a hipótese H_{10}) não tem uma influência marcante na precisão dos resultados.

Para determinar os parâmetros h_e, $h_e^{(2)}$, c_e e $c_e^{(2)}$, nós desenvolvemos o método da equivalência entre as taxas de serviço e caminhada e o método da superposição dos ciclos condicionais.

O primeiro método é baseado em três hipóteses (H_{11}, H_{12} e H_{13}). Nas duas primeiras hipóteses, nós assumimos, respectivamente, que as taxas de serviço e caminhada do servidor equivalente são iguais a s vezes as correspondentes taxas de cada servidor individual (s denota o número de servidores). A terceira hipótese é relativa à independência entre os segmentos de ciclos do servidor equivalente. As duas primeiras hipóteses relativas ao segundo momento do tempo de serviço e do tempo de caminhada e a terceira hipótese são as mais influentes na precisão dos resultados.

O segundo método é baseado em treze hipóteses (H_{14}-H_{26}). Nesse método, nós generalizamos o modelo de Kuehn [4] ao caso multi-servidor, estendendo aqui o conceito de tempos de ciclo condicionais. Nós determinamos, então, o tempo médio de espera em fila dado pela expressão em (3.30). A avaliação de ρ e dos dois primeiros momentos dos tempos intervisitas condicionais (v', $v'^{(2)}$, v'' e $v''^{(2)}$) é aproximada. Para sistemas de filas cíclicas fracamente acopladas, geralmente são obtidos resultados relativamente próximos daqueles obtidos por simulação. Isto é estimulante, devido à extrema complexidade do comportamento dos sistemas de filas cíclicas com múltiplos servidores, o qual depende fortemente dos valores dos parâmetros e é difícil de ser tratado por técnicas simples de aproximação. Deve ser lembrado, inclusive, que mesmo no caso de sistemas muito mais simples, tal como a fila $M/G/s$, $s > 1$, a solução exata não é conhecida [20]. As hipóteses mais significativas com respeito à precisão dos resultados são as hipóteses relativas à determinação dos segundos momentos dos tempos intervisitas condicionais.

Na tentativa de melhorar a precisão do método da equivalência, nós desenvolvemos o método da equivalência modificada, onde nós supomos que o tempo de caminhada do servidor equivalente é igual ao tempo de caminhada de um servidor reduzido por um fator s apenas para valores pequenos de tráfego, e não é alterado à medida que o tráfego oferecido por servidor tende à unidade [20]. Esse método fornece resultados às vezes melhores ou iguais e às vezes piores do que os correspondentes resultados fornecidos pelo método da equivalência. Os resultados piores resultam do fato que ele tende a superestimar os primeiros dois momentos do tempo de ciclo do servidor equivalente. Para tráfegos da ordem de 0,6 erl, em geral ele funciona bem.

A grande vantagem dos métodos da equivalência sobre o método da superposição é que eles são explícitos e mais simples em termos de obtenção de resultados. Assim, são preferíveis ao método da superposição sempre que fornecerem precisões comparáveis em relação aos resultados de simulação.

A principal idéia do modelo de simulação é a criação de entidades que evoluem no tempo em seus respectivos estados discretos, de modo a reproduzir no modelo as mesmas características dos sistemas de filas cíclicas descritas no modelo analítico, assim como de outros sistemas ainda não tratados na literatura. Como um auxílio na descrição, nós introduzimos os diagramas de transição de estados das entidades do modelo. Esses diagramas se revelaram práticos e eficientes, uma vez que proporcionam uma visualização clara das mudanças de estados e das relações entre as entidades em uma rodada de simulação.

O modelo de simulação fornece parâmetros que podem ser usados como entrada para o modelo analítico. Assim, ao contrário dos outros trabalhos existentes na literatura, que só comparam o tempo médio de espera em fila obtido analiticamente e por simulação, nós somos orientados pelo modelo de simulação sobre quais parâmetros têm mais influência na determinação de um dado resultado do modelo analítico, como, por exemplo, o tempo médio de espera.

Nós aplicamos os modelos à avaliação de desempenho de importantes redes de interligação de processadores reais, tais como a rede do sistema de comutação TRÓPICO, de um ponto de transferência de sinalização (PTS) e de um sistema de comutação de mensagens de alto desempenho. Aqui, a palavra real tem o sentido de existente ou passível de se tornar existente. Através da aplicação dos modelos, nós podemos, por exemplo:

- verificar se os requisitos relativos ao tempo médio de permanência das mensagens nos *buffers* de transmissão são satisfeitos, em condições nominais e em condições de sobrecarga, para determinadas configurações sistêmicas e condições de tráfego;
- determinar a máxima taxa de sinalização interprocessadores, de modo a serem satisfeitos requisitos de qualidade de serviço predeterminados;
- determinar as condições de estabilidade de uma rede de processadores;
- em um sistema assimétrico, verificar quais os grupos de processadores que têm uma influência maior no tempo médio de permanência das mensagens nos *buffers* de transmissão;
- verificar a variação da taxa global de sinalização interprocessadores de um sistema em função de alguns parâmetros, como, por exemplo, o comprimento médio das mensagens;
- verificar se é vantajoso, sob o ponto de vista de desempenho, adotar uma transmissão do tipo *full duplex* ao invés de semiduplex ou, ainda, adotar

a acessibilidade total das unidades de processamento aos planos de sinalização, enfim, verificar o tipo de acesso e transmissão mais apropriado a uma aplicação particular.

Nossas investigações serão orientadas agora no sentido de aprimorar o modelo analítico nos pontos fracos detetados no nosso estudo. Assim sendo, nós devemos:

- melhorar a aproximação para os parâmetros h_e, $h_e^{(2)}$, c_e e $c_e^{(2)}$, no método da equivalência entre taxas, para os parâmetros v', $v'^{(2)}$, v'' e $v''^{(2)}$, no método da superposição, e para o parâmetro ρ, nos dois métodos, de modo a ampliar a gama de aplicações do modelo (parece ser mais adequado ao método da equivalência um fator de redução do tempo de caminhada situado entre s e $[1-a]s+a$);

- obter também, com relação ao modelo analítico, o segundo momento do tempo de espera, a fim de se obter aproximadamente um dado quantil desse tempo, através, por exemplo, do ajuste de uma função gama incompleta, conforme fizemos em [23] utilizando resultados de simulação;

- obter uma expressão aproximada para o tempo de espera pela liberação da parte recepção em uma rede em barramentos múltiplos com modo de escalonamento do tipo espera;

- verificar se em um sistema assimétrico com múltiplos servidores cíclicos ocorre o mesmo fenômeno observado em [12] para o caso de sistema assimétrico com servidor único. Nesse caso, Ibe e Cheng mostraram que se duas estações têm idêntica taxa de chegada, a mesma distribuição do tempo de serviço e de caminhada, elas não têm necessariamente o mesmo tempo médio de espera em fila. Este último depende da posição da estação em questão relativa à estação com a mais elevada taxa de chegada.

As estruturas de interligação de processadores que nós analisamos não atendem, de acordo com as disponibilidades tecnológicas e de custo atuais, aos requisitos relativos a taxas de mensagens previstos para o novo cenário em Telecomunicações com Redes Inteligentes, que são da ordem de 100.000–500.000 mensagens/seg. Portanto, nós deveremos também investigar o comportamento de outras estruturas de comutação de mensagens de alto desempenho que não se enquadram na classe de sistemas aqui apresentados.

Apêndice A

Estrutura de Sinalização do Sistema TRÓPICO

O sistema de comutação TRÓPICO [38],[39] é um sistema distribuído, onde o processamento de uma chamada telefônica e de uma função de operação, manutenção ou supervisão é realizado pela concatenação de fases bem definidas. O controle do processamento é feito por meio de fluxos de dados transportados por sinais *software* chamados mensagens. Cada mensagem é constituída por um conjunto de até no máximo 512 *bytes* com estrutura bem determinada. A primeira tarefa da primeira fase é ativada por um evento externo ao sistema. Ao final de qualquer tarefa, é gerada uma mensagem que ativa a próxima tarefa da fase. Por exemplo, se o evento externo é um dígito enviado por um assinante, a fase realizada é a análise do dígito e, eventualmente, a seleção da linha do assinante chamado ou do tronco, conforme se trate de uma chamada intracentral ou de saída, respectivamente. As tarefas são executadas por programas residentes nos processadores do sistema. O processo de comunicação entre os programas é chamado sinalização. Se duas tarefas consecutivas em uma fase são executadas por processadores distintos, há que ser utilizada uma estrutura física denominada *estrutura de sinalização.*

A estrutura de sinalização interprocessadores é do tipo multibarramento com alocador centralizado. Ela possui uma série de características desejáveis:

- é compatível com a filosofia orientada a barramento, implícita na maioria das famílias de microprocessadores atualmente existentes;
- é modular, permitindo um incremento relativamente fácil no número de unidades de processamento e de barramentos;
- é tolerante a falhas, no sentido de que se um barramento falha o sistema ainda funciona, supondo que continue pelo menos um barramento em operação, embora com desempenho degradado;

- utiliza tecnologias convencionais de mercado ao invés de componentes dedicados, o que torna o sistema altamente viável economicamente;

- os processadores apresentam internamente apenas o *hardware* essencial às funções de processamento e comunicação. Os circuitos específicos para as diferentes funções são implementados em placas separadas e interconectadas ao processador através de interface *hardware* padrão. Essa característica possibilita, se respeitadas as interfaces definidas, a utilização integral da estrutura no desenvolvimento de outros sistemas distribuídos.

Essa estrutura é descrita detalhadamente em [23] e [39]. Aqui, nós apresentamos uma descrição sumária. Por simplicidade, às vezes nós denotamos uma unidade de processamento por processador.

Os processadores se comunicam através da passagem de mensagens pela rede e seus ***buffers*** de transmissão e recepção operam no modo semiduplex, isto é, eles não podem transmitir e receber mensagens simultaneamente. Cada processador possui um certo número de rotas ou planos de sinalização operando em partição dinâmica de carga. Cada plano incorpora a cabeação e a estrutura *hardware* de interfaceamento e tem um alocador que controla um certo número de barramentos. Um sistema de comutação pode ter no máximo quatro planos de sinalização.

O acesso dos processadores aos planos é através de circuitos dedicados de sinalização (*CIDS*). Cada *CIDS* possui internamente uma memória mapeada em dois ***buffers***, um dedicado à transmissão e outro à recepção. Com respeito ao acesso, há dois tipos de unidades de processamento:

1. tipo I: com um único *CIDS* que acessa todos os planos;

2. tipo II: com um *CIDS* associado a cada plano.

Um alocador ativa, mantém e desativa as conexões físicas entre os processadores a fim de possibilitar a transmissão de dados. A troca de informações sobre o meio de transmissão compartilhado é disciplinada por um protocolo bem definido. Cada alocador controla a alocação de recursos do plano de sinalização correspondente através de um enlace de sincronismo, de um enlace de comunicação e de um enlace de controle.

O enlace de sincronismo distribui os sinais de relógio e sincronismo à rede.

O enlace de comunicação é o meio físico por onde se propagam as mensagens. Ele pode ser composto por uma via de comunicação ou, alternativamente, por duas vias de comunicação multiplexadas no tempo. Cada via é composta por uma via de transmissão e por uma via de recepção, que operam a uma taxa de 2 Mbit/s. Cada plano pode conter até três enlaces de comunicação.

O enlace de controle é composto por até quatro vias de solicitação e uma via de habilitação.

Cada via de solicitação consiste de uma linha com 256 canais seriais. Cada canal é associado biunivocamente a um processador do sistema, que o utiliza para

requerer a transmissão de uma mensagem. O número de vias de solicitação controladas por um alocador é, portanto, função da capacidade do sistema.

Cada alocador possui um dispositivo de varredura que consulta, um a um, os canais de solicitação, isto é, faz uma chamada seqüencial e cada processador responde se há ou não mensagem a transmitir. Os processos de varredura dos alocadores são defasados, de modo a impedir que uma mesma solicitação seja atendida por mais de um alocador.

Cada alocador utiliza a via de habilitação para controlar o acesso dos processadores através de uma palavra de 16 *bits*, que contém o endereço do processador selecionado, a identidade da via a ser utilizada e o modo de seleção: transmissão, recepção ou liberação.

Em cada unidade de processamento há uma fila de mensagens a transmitir. Em um instante de envio de mensagem, a unidade carrega a primeira mensagem dessa fila no *buffer* de transmissão do *CIDS* (disciplina *FIFO*), o qual solicita o envio a todos os alocadores, se a unidade é do tipo I, ou ao alocador correspondente, se a unidade é do tipo II. A partir desse ponto, a unidade se libera e a responsabilidade de transmissão da mensagem passa a ser exclusiva do *CIDS* em questão.

Ao ser habilitado para transmitir, o *CIDS* envia ao alocador um segmento de informação, que contém o tamanho da mensagem (até um máximo de 512 *bytes*) e a identificação da unidade receptora. O alocador, então, envia uma habilitação de recepção a essa unidade. Se ela estiver disponível para recepção, é estabelecida uma via de comunicação entre o processador transmissor e o processador receptor. A seguir o alocador liga uma temporização de tomada de via, cujo valor é função do tamanho da mensagem e da velocidade de transmissão. Ao fim desse tempo, ele libera a via, ou seja, desativa a conexão entre os *CIDS's* transmissor e receptor. Se ainda houver uma via disponível sob seu controle, ele retorna à varredura. Caso contrário, aguarda passivamente pela liberação de uma via.

Se o *CIDS* receptor está disponível ao ser habilitado para recepção, ou seja, não está enviando, nem recebendo, nem carregando e nem descarregando uma mensagem, o *CIDS* transmissor envia-lhe de uma só vez todo o pacote da mensagem. A transmissão é assíncrona a nível de caracteres e síncrona a nível de *bits*. Isto exige uma base de tempo na recepção sincronizada com a transmissão através de uma onda de relógio (enlace de sincronismo). A partir da identificação do elemento de início da mensagem (*start bit*), o receptor detecta os instantes de separação dos *bytes*. De posse do primeiro *byte* de informação, que fornece o número de *bytes* da mensagem, ele sincroniza o término da transmissão. Em cada transmissão bem sucedida, uma única mensagem é transmitida (serviço *um-por-vez*). Caso o *CIDS* receptor esteja indisponível para recepção (ocupado), o *CIDS* transmissor volta a solicitar via e é desencadeado um novo procedimento de tentativa de transmissão de mensagem (*modo repetição*).

Após uma transmissão ou recepção, a unidade de processamento em questão verifica se há mensagens na sua fila de mensagens a transmitir. Em caso positivo, tem início um procedimento de tentativa de transmissão de mensagem.

Para fins de detecção de erros, o *CIDS* transmissor faz um embaralhamento dos ***bits*** a serem transmitidos. Assim, é gerado um ***byte*** de redundância que é acrescentado ao final da mensagem. O *CIDS* receptor é sincronizado com o transmissor e recupera a seqüência original dos ***bits*** através de uma operação inversa de desembaralhamento. Desta forma, ele pode comunicar ao transmissor o bom ou o mal recebimento da mensagem. Em caso de insucesso, o *CIDS* transmissor solicita novamente uma via e desencadeia um novo procedimento de tentativa de transmissão (*modo repetição*).

O número de tentativas de transmissão relativo a uma mesma mensagem é limitado por uma certa temporização de transmissão, que é ligada pela unidade de processamento, assim que a mensagem é carregada no ***buffer*** do *CIDS*.

Apêndice B

Manual de Utilização do Programa do Modelo Analítico

O nome do arquivo fonte do programa de computador do modelo analítico é *Sismmc.sim*. Os dados de entrada são obtidos de maneira interativa. Em geral, nos sistemas assimétricos existem determinados grupos de nós afins, em relação aos quais os processos de chegada, caminhada e serviço e o encaminhamento são simétricos. Assim, para facilitar a entrada de dados e economizar espaço na memória do computador, em um sistema assimétrico com um grande número de nós, os m nós são divididos em g grupos de nós afins. Cada grupo tem g_i nós, onde $\sum_{i=1}^{m} g_i = m$. Nesta situação, nós definimos a probabilidade de encaminhamento intergrupos, β_{kl}, como sendo a probabilidade de um usuário da fila de transmissão de um nó do grupo k ser encaminhado à fila de recepção de um nó do grupo l. Nós temos que

$$\beta_{kl} \geq 0,\ k,l = 1,2,\ldots,g \quad \text{e} \quad \sum_{l=1}^{g} \beta_{kl} = 1,\ k = 1,2,\ldots,g.$$

As probabilidades de encaminhamento entre os nós são tais que

$$\alpha_{ij} = \begin{cases} \frac{\beta_{kl}}{g_l - 1}, & \text{se } k = l \text{ e } i \neq j, \\ 0, & \text{se } k = l \text{ e } i = j, \\ \frac{\beta_{kl}}{g_l}, & \text{se } k \neq l, \end{cases}$$

onde k é o grupo ao qual i pertence e l é o grupo ao qual j pertence, $i,j = 1,2,\ldots,m$. Assim, no caso de encaminhamento assimétrico, não é necessário fornecer ao programa as probabilidades α_{ij}. Elas são calculadas dinamicamente durante a execução do programa.

Após o pedido de execução, o programa solicita ao operador, inicialmente, os seguintes dados comuns aos casos simétrico e assimétrico:

- cabeçalho do relatório de resultados: texto limitado em 130 caracteres e que serve como identificação da particular execução;

- caracterização da simetria dos processos de chegada, encaminhamento, caminhada e serviço: em cada caso deve ser fornecido o texto *simetrico* ou *assimetrico*, conforme for o caso;
- método de obtenção do tempo de ciclo do servidor equivalente: equivalência entre taxas ou superposição dos ciclos condicionais, sendo que neste último caso deve-se fornecer também o procedimento de superposição dos processos recorrentes (Kuehn ou Bhuyan) e a precisão das aproximações(ϵ);
- acessibilidade dos servidores a um nó: total ou parcial, sendo que neste último caso deve-se fornecer também o tipo de transmissão (unidirecional ou bidirecional);
- nome do arquivo de saída de resultados;
- número de nós (m);
- número de servidores (s).

Se o sistema é simétrico, o programa solicita, adicionalmente, os seguintes dados: u, $VAR[U]$, h, $VAR[H]$, número de valores de tráfegos oferecidos por servidor (n_a) e os respectivos valores dos tráfegos oferecidos por servidor a_i, $i = 1, 2, \ldots, n_a$. O programa fornece os resultados para cada um desses valores.

Agora, se o sistema é assimétrico, o programa solicita, adicionalmente, os seguintes dados:

- número de grupos afins (g);
- número de nós por grupo (g_i, $i = 1, 2, \ldots, g$);
- se a chegada é simétrica: tráfego oferecido por servidor; caso contrário, tráfego oferecido por servidor em cada grupo;
- se o encaminhamento é assimétrico: probabilidades de encaminhamento entre os grupos β_{kl}, $k, l = 1, 2, \ldots, g$; caso contrário, não deve ser fornecido nenhum valor, desde que, neste caso, nós temos que $\beta_{kl} = \frac{1}{g}$, $k, l = 1, 2, \ldots, g$;
- se a caminhada é simétrica: u e $VAR[U]$; caso contrário, u_k e $VAR[U_k]$ para cada grupo k, $k = 1, 2, \ldots, g$;
- se o serviço é simétrico: h e $VAR[H]$; caso contrário, h_k e $VAR[H_k]$ para cada grupo k, $k = 1, 2, \ldots, g$.

No relatório de resultados são fornecidos os valores de λ, ρ, P_0, c', $c'^{(2)}$, c'', $c''^{(2)}$, v', $v'^{(2)}$, v'', $v''^{(2)}$, a e w, no método da superposição e λ, ρ, P_0, c_e, $c_e^{(2)}$, a e w, no método da equivalência. Se a condição

$$0 < \frac{\lambda v'}{1 - \lambda(v'' - v')} < \rho \leq 1 \,,$$

no método da superposição, ou a condição

$$0 < \frac{\lambda c_e}{1 - \lambda h_e} < \rho \leq 1 \ ,$$

no método da equivalência, é violada, o programa fornece os valores de λ, ρ e a e imprime a mensagem *Sistema Instável.*

Apêndice C

Manual de Utilização do Programa de Simulação

As entidades do modelo foram implementadas em SIMULA [30] como *Process Classes*. Uma declaração *Class* define um padrão de programa (dados e ações). Durante a execução do programa, objetos *Class* (como programas auto-suficientes) podem ser criados, tendo seus próprios dados e ações definidos pela declaração *Class* . Um objeto de uma *Class* prefixada por *Process* é chamado um objeto *Process*. Um objeto *Process* tem propriedades de elementos em filas (isto é, podem aparecer como usuários em uma fila) e, além disso, é caracterizado por um dos estados: *ativo, suspenso, passivo* e *terminado*.

O nome do arquivo fonte do programa de computador é *Redmmc.sim*. Cada rodada de simulação é caracterizada por um conjunto de dados de entrada, que são subdivididos em dados de configuração, de controle e de tráfego. Antes de uma execução do programa, o operador deve colocar os dados nos arquivos Cofmmc.dat, Cotmmc.dat e Trammc.dat, respectivamente.

Os dados pertencentes a cada arquivo são os seguintes:

COFMMC.DAT (configuração): número de nós do sistema (m), de servidores cíclicos (s), de caminhos de comunicação por servidor (l_s), do nó marcado (≥ 1 e $\leq m$) e do servidor marcado para estatística (≥ 1 e $\leq s$).

COTMMC.DAT (controle): sementes das variáveis aleatórias geradas pelas entidades: u_i, $i = 1, 2, \ldots, 7$,onde

- u_1: número do nó de transmissão do usuário;
- u_2: número do nó de recepção do usuário;
- u_3: tempo de ocupação de caminho (serviço);
- u_4: tempo entre geração de usuários;
- u_5: tempo de caminhada;

- u_6: número do caminho livre;
- u_7: tempo de esvaziamento;

número total de usuários a serem simulados na rodada, número de usuários simulados por relatório periódico e as seguintes variáveis texto:

- *cabecalho* (no máximo 130 caracteres): cabeçalho dos relatórios de resultados;
- *procechegada* (6 caracteres): tipo de processo de chegada dos usuários.
- *procecaminha* (6 caracteres): tipo de processo de caminhada dos servidores.
- *proceservico* (6 caracteres): tipo de processo de serviço.
- *procesvaziam* (6 caracteres): tipo de processo de esvaziamento das filas de recepção.
- *modatefilatx* (9 caracteres): modo de atendimento das filas de transmissão.
- *modesvfilarx* (9 caracteres): modo de esvaziamento das filas de recepção.
- *modbloq* (6 caracteres): modo de escalonamento devido a destino bloqueado.
- *compfiltx* (8 caracteres): capacidade das filas de transmissão.
- *compfilrx* (8 caracteres): capacidade das filas de recepção.
- *chegada* (9 caracteres): simetria das taxas de chegada nas filas de transmissão.
- *encaminha* (9 caracteres): simetria do encaminhamento dos usuários.

O cabeçalho é um texto qualquer que serve para identificar a particular rodada de simulação. Com relação à variável *procechegada*, deve-se entrar com o valor *Negexp*, desde que o processo de chegada é suposto Poissoniano. Se o modelo for estendido para tratar outros tipos de processos de chegada, essa variável poderá assumir outros valores. Quanto às variáveis *procecaminha, proceservico* e *procesvaziam,*deve-se entrar com um dos textos: *Negexp* (Poisson), *Erlang* (Erlang$-r$) ou *Consta* (constante). As variáveis *modatefilatx* e *modesvfilarx* devem ser associadas a um dos textos: *Exaustivo* (exaustivo), *Combarrei* (com barreira),*Limexaust* (limitado exaustivo) ou *Limbarrei* (limitado com barreira). A variável *modbloq* deve assumir um dos textos: *Repete* (repetição) ou *Espera* (espera). As variáveis *compfiltx* e *compfilrx* devem assumir um dos textos: *Limitada* (limitada) ou *Ilimitada* (ilimitada). Finalmente, a variável *chegada* deve assumir um dos textos: *Simetrica* (simétrica) ou *Assimetrica* (assimétrica) e a variável *encaminha* um dos textos: *Simetrico* (simétrico) ou *Assimetrico* (assimétrico).

TRAMMC.DAT (tráfego): Os dados desse arquivo são função dos dados dos arquivos anteriores.

- se *chegada=Simetrica*, entra-se com o valor de λ; caso contrário, entra-se com os valores de λ_j, $j = 1, 2, \ldots, m$;
- tempos médios:
  - de serviço (ocupação de caminho);
  - de comunicação usuário-servidor;
  - de caminhada;
  - de esvaziamento.

A entrada de dados a seguir é opcional. Se a condição é verificada, conforme foi estipulado no arquivo Cotmmc.dat, deve-se entrar com o parâmetro correspondente. Caso contrário, o operador não deve fornecer nenhum valor.

- se *proceservico=Erlang* $\Rightarrow$ parâmetro r_h da distribuição Erlang-r_h do tempo de serviço;
- se *procecaminha=Erlang* $\Rightarrow$ parâmetro r_u da distribuição Erlang-r_u do tempo de caminhada;
- se *procesvaziam=Erlang* $\Rightarrow$ parâmetro r_e da distribuição Erlang-r_e do tempo de esvaziamento;
- se *encaminha=Assimetrico* $\Rightarrow$ matriz de probabilidade de encaminhamento dos usuários α_{ij} , $i, j = 1, 2, \ldots, m$;
- se *compfiltx=Limitado* $\Rightarrow$ valor do comprimento máximo das filas de transmissão;
- se *compfilrx=Limitado* $\Rightarrow$ valor do comprimento máximo das filas de recepção;
- se *modatefiltx=Limbarrei* ou *Limexaust* $\Rightarrow$ limite do número de usuários da fila de transmissão atendidos por visita do servidor;
- se *modesvfilrx=Limbarrei* ou *Limexaust* $\Rightarrow$ limite do número de usuários da fila de recepção atendidos por visita do esvaziador.

Após o pedido de execução feito pelo operador, a rodada é iniciada, se os dados de entrada forem compatíveis. Nesse caso, são apresentados os relatórios periódicos na tela do terminal de vídeo ou em arquivo. No primeiro relatório periódico e no relatório final, são apresentados os dados de entrada.

Os seguintes resultados são apresentados nos relatórios: probabilidade de visita bem sucedida, probabilidade de fila vazia, probabilidade de bloqueio da fila de transmissão, primeiro e segundo momentos dos tempos de ciclo e intervisitas condicionais e absolutos, tráfego de comunicação usuário-servidor, tráfego de

ocupação dos caminhos, tráfego de ocupação total dos caminhos, primeiro e segundo momentos dos tempos de permanência nas filas de transmissão e recepção e do tempo de espera pela liberação da porta de recepção.

Para m grande e u pequeno, o programa pode se tornar computacionalmente ineficiente, pois, nesta situação, os servidores cíclicos são escalados um número muito grande de vezes durante uma rodada de simulação. Para contornar esse problema, nós desenvolvemos um outro modelo de simulação. Nesse modelo, um servidor cíclico vai ao estado *Aguardando Aviso de Solicitação (Agsoli),* após um ciclo de caminhada completo vazio, isto é, um ciclo onde todas as filas de transmissão dos m nós estão vazias. Ele permanece passivamente nesse estado, até receber a mensagem *Solicitação* do primeiro usuário que é gerado após a ocorrência de um ciclo completo vazio. A partir do tempo em que ficou no estado *Agsoli* e dos valores de u e $u^{(2)}$, o servidor cíclico obtém o número do nó por onde ele deve recomeçar a caminhada. Ele vai, então, ao estado *Aginvi*, a partir do qual as ações são idênticas às do servidor cíclico do outro modelo. Cada usuário, antes de entrar na fila de transmissão de um nó, envia aos servidores que eventualmente estiverem no estado *Agsoli* a mensagem *Solicitação.* Agora, os servidores cíclicos não enviam mais aos nós do modelo a mensagem *Visita.* Assim, no modelo de simulação modificado, os parâmetros ρ, c', $c'^{(2)}$, c'', $c''^{(2)}$, v', $v'^{(2)}$, v'' e $v''^{(2)}$ não são mais obtidos. Este é o preço a pagar para simular sistemas com um grande número de nós e tempo de caminhada relativamente pequeno.

Temos também que, para tratar o caso de sistemas assimétricos de grande porte, os m nós são divididos em g grupos de nós afins. Cada grupo tem g_i nós, onde $\sum_{i=1}^{g} g_i = m$. Neste caso, a probabilidade de encaminhamento intergrupos é β_{kl}, $k, l = 1, 2, \ldots, g$, tal como no programa do modelo analítico.

O nome do arquivo fonte do programa de computador do modelo de simulação modificado é *Sireci.sim.*

Bibliografia

[1] R. B. Cooper e G. Murray, "Queues Served in Cyclic Order", *The Bell System Technical Journal,* Março 1969, pp. 675-89.

[2] R. B. Cooper, "Queues Served in Cyclic Order: Waiting Times", *The Bell System Technical Journal,* Março 1970, pp. 399-413.

[3] O. Hashida e K. Ohara, "Line Accomodation Capacity of a Communication Control Unit", *Review of the Electrical Communication Laboratories,* Vol. 20, Números 3-4, Março-Abril 1972, pp. 231-9.

[4] P. J. Kuehn, "Multiqueue Systems With Nonexhaustive Cyclic Service", *The Bell System Technical Journal,* Vol. 58, Número 3, Março 1979, pp. 671-98.

[5] M. J. Ferguson e Y. J. Aminetzah, "Exact Results for Nonsymmetric Token Ring Systems", *IEEE Transactions on Communications,* Vol. Com-33, Número 3, Março 1985, pp. 223-31.

[6] S. W. Fuhrmann e R. B. Cooper, "Application of Decomposition Principle in M/G/1 Vacation Model to Two Continuum Cyclic Queueing Models-Especially Token-Ring LANs", *AT&T Technical Journal,* Vol. 64, Número 5, Maio-Junho 1985, pp. 1091-9.

[7] H. Takagi, "Mean Message Waiting Times in Symmetric Multi-Queue Systems With Cyclic Service", *Performance Evaluation,* Vol. 5, 1985, pp. 271-7.

[8] T. Raith e P. Tran-Gia, "Performance Analysis of Polling Mechanisms With Receiver Blocking", in:*P. Kuehn (ed.), Elsevier Science Publishers B.V. (North-Holland), ICCC 1986,* pp. 577-83.

[9] D. Everith, "Simple Approximations for Token Rings", *IEEE Transactions on Communications,* Vol. Com-34, Número 7, Julho 1986, pp. 719-21.

[10] H. Takagi, "Analysis and Applications of a Multiqueue Cyclic Service System With Feedback", *IEEE Transactions on Communications,* Vol. Com-35, Número 2, Fevereiro 1987, pp. 248-50.

[11] P. Tran-Gia e T. Raith, "Performance Analysis of Finite Capacity Polling Systems With Nonexhaustive Service", *Performance Evaluation,* Vol. 9, 1988, pp. 1-16.

[12] O. C. Ibe e X. Cheng, "Approximate Analysis of Asymmetric Single-Service Token-Passing Systems", *IEEE Transactions on Communications,* Vol. 37, Número 6, Junho 1989, pp. 572-7.

[13] D. Everith, "A Note on the Pseudoconservation Laws for Cyclic Service Systems With Limited Service Disciplines", *IEEE Transactions on Communications,* Vol. 37, Número 7, Julho 1989, pp. 781-3.

[14] L. F. M. de Moraes, "Comments on"Analysis and Applications of a Multiqueue Cyclic Service System With Feedback"", *IEEE Transactions on Communications,* Vol. 38,Número 2, Fevereiro 1990, pp. 148-9.

[15] H. Takagi, *Analysis of Polling Systems* (The MIT Press, Cambridge, MA, 1986).

[16] A. C. Lavelha e J. M. de Souza, "Simulação da Comunicação Interprocessadores na Central TRÓPICO-R", *in: N. Meisel e L. J. Braga-Filho(eds.) III Congresso da Sociedade Brasileira de Computação,* Campinas, SP, 23-29 de Julho de 1983.

[17] R. J. T. Morris e Y. T. Wang, "Some Results for Multiqueue Systems With Multiple Cyclic Servers", in:*H. Rudin e W. Bux(eds.), Performance of Computer-Communication Systems(North-Holland),* Amsterdam, 1984, pp. 245-58.

[18] T. Raith, "Performance Analysis of Multibus Interconnection Networks in Distributed Systems", in:*Minoru Akiyama(ed.), Teletraffic Issues in an Advanced Information Society, ITC 11, Elsevier Science Publishers, B.V.,* 1985, pp. 662-8.

[19] A. E. Kamal e V. C. Hamacher, "Approximate Analysis of Non-exhaustive Multiserver Polling Systems With Applications to Local Area Networks", *Computer Networks and ISDN Systems,* Vol. 17, 1989, pp. 15-27.

[20] M. A. Marsan, L. F. de Moraes, S.Donatelli e F.Neri, "Analysis of Symmetric Nonexhaustive Polling With Multiple Servers", in: *Proc. of IEEE INFOCON 90,* São Francisco, Califórnia, Junho 3-7.

[21] M. A. Marsan, S. Donatelli e F. Neri, "GPSN Models of Markovian Multiserver Multiqueue Systems", *Performance Evaluation,*Vol. 11, Número 4, Novembro 1990, pp. 227-240.

[22] L. N. Bhuyan, D. Ghosal e Q. Yang, "Approximate Analysis of Single and Multiple Ring Networks", *IEEE Transactions on Computers,* Vol. 38, Número 7, Julho 1989, pp. 1027-40.

[23] G. Fernandes e A. C. Lavelha, "Estrutura de Sinalização do TRÓPICO-RA", in: *VII Simpósio Brasileiro de Telecomunicações,* Florianópolis, SC, 3-6 de Setembro de 1989.

[24] M. A. Marsan, S. Donatelli e F. Neri, "Multiserver Multiqueue Systems With Limited Service And Zero Walk Time", submetido a *IEEE INFOCOM '91.*

[25] A. C. Lavelha, J. M. de Souza e J. B. R. do Val, "Approximate Analysis of Multiqueue Systems With Multiple Cyclic Servers", submetido a publicação.

[26] P. J. Kuehn, "Approximate Analysis of General Queueing Networks by Decomposition", *IEEE Transactions on Communications,* Vol. Com-27, Número 1, Janeiro 1979, pp. 113-26.

[27] W. Whitt, "Approximating a Point Process by a Renewal Process, I:Two Basic Methods", *Operations Research,* Vol. 30, Número 1, Janeiro-Fevereiro 1982, pp. 125-47.

[28] N. T. J. Bailey, "On Queueing Processes With Bulk Service", *Journal of the Royal Statistical Society,* series B, Vol. 16, pp. 80-87.

[29] N. D. Georganas, "Buffer Behavior With Poisson Arrivals And Bulk Geometric Service",*IEEE Transactions on Communications,* Agosto 1976.

[30] G. M. Birtwistle, O-J. Dahl, B. Myhrhaug e K. Nygaard, *Simula Begin* (Auerbach Publishers Inc., Filadélfia, Pensilvânia, 1973).

[31] A. M. Law, "Statistical Analysis of Simulation Output Data", *Operations Research,* Vol. 31, Número 6, Novembro-Dezembro 1983.

[32] R. B. Cooper, *Introduction to Queueing Theory,* segunda edição (Elsevier North Holland, Inc. 1981).

[33] J. L. Hammond e P. J. P. O'Reilly, *Performance Analysis of Local Computer Networks* (Addison-Wesley Publishing Company, 1986).

[34] M. Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis* (Addison-Wesley Publishing Company, 1987).

[35] W. Feller, *Introduction to Probability Theory and Its Applications,* III edição (John Wiley & Sons, Volume 1, 1970).

[36] L. Breiman, *Statistics With a View Toward Applications* (Houghton Mifflin Company, Boston, 1973).

[37] R. E. Shannon, *Systems Simulation the art and science* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975).

[38] *TRÓPICO R - Central Local/Tandem Digital do Sistema TRÓPICO*, Manual Interno do CPqD–TELEBRÁS, 1984.

[39] *Sistema TRÓPICO RA*, Manual Interno do CPqD–TELEBRÁS, 1988.

[40] CCITT Blue Book - Volume VI - Fascicle VI.5
Digital local, transit, combined and international exchanges in integrated digital networks and mixed analogue-digital networks. Recommendation Q.543

[41] CCITT Blue Book - Volume VI - Fascicle VI.8
Specifications of Signalling System No. 7. Reccomendation Q.706

[42] G. Hébuterne, *Écoulement du Trafic dans les Autocommutateurs* (Masson, Paris, 1985).

[43] B. Wallström, "Congestion Studies in Telephone Systems With Overflow Facilities", *Ericsson Technics,* Número 3, 1966, pp. 187-351.

Bibliografia Suplementar

Nota: Os artigos listados a seguir foram recomendados por um dos membros da comissão julgadora, o Prof. Dr. Luis Felipe M. de Moraes. Eles constituem um excelente complemento ao conjunto de referências sobre sistemas multifilas com múltiplos servidores cíclicos que foram arroladas no nosso trabalho.

[**1**] H. Takagi, "A Bibliography on the Analysis and Applications of Polling Models (Part I: A-K), (Part II: L-Z)", *Proceedings of the International Workshop on the Analysis of Polling Models, Kyoto*, Japão, Dezembro, 1988, atualizado em 12 de Julho de 1991.

[**2**] M. A. Marsan, S. Donatelli e F. Neri, "GSPN Models of Multiserver Multiqueue Systems", *The Third International Workshop on Petri Nets and Performance Models*, pp. 19-28, 11-13 de Dezembro, 1989, *Kyoto*, Japão.

[**3**] M. A. Marsan, S. Donatelli, F. Neri e U. Rubino, "GSPN Models of Random, Cyclic, and Optimal 1-Limited Multiserver Multiqueue Systems", *SIGCOMM'91 Conference, Zürich, Switzerland*, 3-6 de Setembro, 1991.

[**4**] B. D. Bunday e E. Khorram, "The Eficiency of Uni-directionally Patrolled Machines With Two Robot Repairmen", *European Journal of Operational Research*, Vol. 39, Número 1 (Março), pp. 32-39, 1989.

[5] —, "The Efficiency of N Machines Equally Spaced Round a Circular Drum Which Rotates Against r Fixed Robot Operatives", a aparecer em *The Arabian Journal for Science and Engineering.*

[6] S. Casale, V. Catania, A. Faro, N. Parchenkov e L. Vita, "Design and Performance Evaluation of an Optical Fibre LAN With Double Token Rings", *Computer Communications*, Vol. 12, Número 3 (Junho), pp. 158-66, 1989.

[7] C. H. Chen e L. N. Bhuyan, "Design and Analysis of Multiple Token Ring Networks", *IEEE INFOCOM'88*, pp. 477-86, *New Orleans, Louisiana*, 27-31 de Março, 1988.

[8] B. Gamse e G. F. Newell, "An Analysis of Elevator Operation in Moderate Height Buildings - II. Multiple Elevators", *Transportation Research, B,* Vol. 16B, Número 4 (Agosto), pp. 321-35, 1982.

[9] Q. Yang, D. Ghosal e L. N. Bhuyan, "Performance Analysis of Multiple Token Ring and Multiple Slotted Ring Networks", *Proceedings of the 1986 Computer Networking Workshop*, pp. 79-86, *Washington*, D. C., 1986.

[10] T. I. Yuk e J. C. Palais, "Analysis of Multichannel Token Ring Networks", *International Conference on Communication Systems (Singapore ICCS'88)*, pp. 907-11, 31 de Outubro - 3 de Novembro, 1988.

[11] M. Zafirovic-Vukovic e I. G. Niemegeers, "Performance Modelling of the Orwell Basic Access Mechanism", *Computer Communication Review*, Vol. 17, Número 5, pp. 35-48, (Proceedings of the ACM SIGCOMM '87 Workshop, *Stowe, Vermont*, 11-13 de Agosto, 1987).

[12] —, "Performance Modelling of a HSLAN Slotted Ring Protocol", *Performance Evaluation Review*, Vol. 16, Número 1 (Maio), pp. 37-46, (Proceedings of the 1988 ACM SIGMETRICS Conference, Santa Fé, Novo México, 24-27 de Maio, 1988).

[13] M. Zafirovic-Vukovic, I. G. Niemegeers e D. S. Valk, "Performance Modelling of Slotted Ring Protocols in HSLAN's", *IEEE Journal on Selected Areas in Communications*, Vol. 6, Número 6 (Julho), pp. 1001-24, 1988.

[14] M. A. Marsan, F. Neri e L. F. M. de Moraes, "Performance Analysis of Multichannel Token Protocols For High Speed Data Networks", *Proceedings of XXX TIMS*, Rio de Janeiro, Brasil, 17 de Julho, 1991.

[15] M. A. Marsan, L. F. M. de Moraes, S. Donatelli e F. Neri, "Cycles and Waiting Times in Symmetric Exhaustive and Gated Multiserver Multiqueue Systems", a aparecer em *INFOCOM'92.*