APROXIMAÇÕES PARA TEMPOS DE ESPERA EM SISTEMAS DE MÚLTIPLAS FILAS COM MÚLTIPLOS SERVIDORES E PRIORIDADES

Renato Souza Silva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Luís Felipe Magalhães de Moraes

Rio de Janeiro
Março de 2015

APROXIMAÇÕES PARA TEMPOS DE ESPERA EM SISTEMAS DE MÚLTIPLAS FILAS COM MÚLTIPLOS SERVIDORES E PRIORIDADES

Renato Souza Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Luís Felipe Magalhães de Moraes, Ph.D.

Prof. Felipe Maia Galvão França, Ph.D.

Prof. Carlos Alberto Vieira Campos, D.Sc.

Bruno Astuto Arouche Nunes, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2015

Silva, Renato Souza

Aproximações para Tempos de Espera em Sistemas de Múltiplas Filas com Múltiplos Servidores e Prioridades/Renato Souza Silva. – Rio de Janeiro: UFRJ/COPPE, 2015.

XVII, 108 p.: il.; 29, 7cm.

Orientador: Luís Felipe Magalhães de Moraes

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2015.

Referências Bibliográficas: p. 99 – 104.

1. Redes Sem Fio. 2. Multicanal. 3. *Polling*. 4. Prioridade. 5. Protocolo. I. Moraes, Luís Felipe Magalhães de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedico este trabalho e todo meu esforço à minha Família. À minha Esposa e Filhas pela paciência e por terem me ensinado a "nunca desistir". Pelas muitas vezes que estivemos juntos, mas a minha cabeça "girava em torno dos estudos". Dedico também à minha Mãe e ao meu Pai por terem me ensinado a "ousar ir além".

Agradecimentos

Gostaria de agradecer primeiramente ao meu Professor Orientador Luís Felipe Magalhães de Moraes, que tanto me ajudou a seguir no "melhor caminho"e aos membros da minha banca pela disponibilidade e compreensão.

Agradeço ao meu Diretor Walkyrio José de Faria Tostes pelo incentivo e ao meu Gerente Leandro Henz pelo suporte.

Um agradecimento especial aos meus colegas de trabalho na Oi, que me ajudaram a dividir um pouco da carga, e aos colegas do Ravel pelas relevantes contribuições.

Quero agradecer especialmente ao amigo Evandro Luis Cardoso Macedo pela ajuda e companheirismo.

Ao Pesc, à Coppe e à Oi pelo apoio que recebi em mais esta caminhada.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

APROXIMAÇÕES PARA TEMPOS DE ESPERA EM SISTEMAS DE MÚLTIPLAS FILAS COM MÚLTIPLOS SERVIDORES E PRIORIDADES

Renato Souza Silva

Março/2015

Orientador: Luís Felipe Magalhães de Moraes

Programa: Engenharia de Sistemas e Computação

A crescente "invasão"das redes sem fio é hoje uma inexorável realidade social e tecnológica. Esta grande utilização se deve sobretudo à praticidade destas redes, em quase todos os ambientes. Uma conseqüência imediata de toda esta praticidade é o surgimento de novas aplicações, aproveitando-se do novo patamar de conectividade proporcionado. Um bom exemplo recente, endereçado pelo novo cenário de conectividade, é o conceito de "Internet das Coisas".

Como esperado, neste caso também há um preço a se pagar. Ao mesmo tempo que a demanda de utilização destas redes aumenta, aumentam também os desafios tecnológicos para torná-las cada vez mais eficientes, considerando ambientes com alta densidade de redes sem fio e fontes interferentes. Neste sentido, a utilização simultanea de múltiplos canais pode ser encarada como uma opção promissora para suportar as novas demandas.

Apesar da comprovada eficiência no aumento da vazão agregada da rede, os modelos analíticos de protocolos com múltiplos canais não são tão comuns quanto os modelos com um único canal. Isto se deve principalmente às dificuldades no processo de modelagem, que envolve complexidade matemática e adoção de novas hipóteses de aproximação.

Este trabalho propõe modelos matemáticos e de simulação, desenvolvidos à partir de um sistema com múltiplas filas, com multiplos servidores e com diferenciação por prioridades. O modelo principal é baseado num protocolo de *polling*, cuja estrutura é amplamente conhecida e documentada, o que credencia a utilização dos modelos propostos na análise de outros protocolos mais complexos.

Os diferentes cenários de testes apresentados, considerando situações normais e assintóticas, mostram resultados relativamente precisos e coerentes.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

WAITING TIME APPROXIMATION IN MULTIPLE QUEUES SYSTEMS WITH MULTIPLE SERVERS AND PRIORITIES

Renato Souza Silva

March/2015

Advisor: Luís Felipe Magalhães de Moraes

Department: Systems Engineering and Computer Science

The growing invasion of wireless networks is an inexorable technological reality due to its usability among several environments. As a consequence, new ever-connected applications have emerged, taking advantage over this new connectivity level improvement. A good example of that new scenario refers to the new concept of "Internet of Things".

As expected, the high density of wireless networks has a cost. As the demand increases, technological challenges increase proportionally so as to keep providing efficiency, speed and availability, even considering the previous mentioned heavily interfering environment. In this sense, the simultaneous use of multiple channels can be viewed as a promising option to support new demands.

Despite the proven effectiveness in increasing the aggregate network throughput, analytical models of protocols with multiple channels are not as common as models with a single channel. This happens mainly due to the difficulties in modeling process, which involves mathematical complexity and the adoption of new hypothesis approach.

This work proposes some mathematical and simulation models, developed from a multiqueue system with multiple servers and different priority classes. The main model is based on a polling protocol, whose modeling process is widely documented, which permits the use of the proposed models in order to analyze other complex protocols.

Different tests scenarios have shown relatively accurate and consistent results, including usual and asymptotic situations.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

C_X	Coeficiente de variação do tempo de serviço dos usuários, no sistema de filas considerado, p. 55
K	Número máximo de servidores no sistema de filas, p. 37
L	Número máximo de usuários de uma fila qualquer, que poderão ser atendidos durante a visita do servidor, p. 60
N	Número total de filas no sistema, p. 51
P	Número total de classes de prioridades no sistema de filas, p. 60
R_p	Tempo médio resisual de um intervalo de *polling*, no instante da chagada de uma determinada mensagem i, p. 53
R_x	Tempo médio residual do usuário encontrado em serviço, no instante da chagada de uma determinada mensagem i, p. 53
V	Tempo que o servidor demora para comutar o controle de uma fila para a próxima fila na sequência (*walk time*). Neste trabalho V foi considerado constante., p. 53
Λ	Taxa média total de chegada de usuários no sistema de filas, p. 60
λ	Taxa média de chegada de usuários em cada fila, p. 47
λ_p	Taxa média de chegada de usuários de prioridade p em cada uma das filas do sistema, p. 58
μ	Taxa média de atendimento de cada servidor em *usuários*/segundo, p. 48
$\overline{W}$	Valor médio da variável aleatória W_i, que representa o tempo de espera em fila do *i-ésimo* usuário ($i \to \infty$), com variância σ^2_W, p. 51

$\overline{W}_{M/D/K}$ Tempo médio de espera dos usuários numa estrutura de filas $M/D/K$, p. 55

$\overline{W}_{M/G/1}$ Tempo médio de espera dos usuários numa estrutura de filas $M/G/1$, p. 50

$\overline{W}_{M/G/K}$ Tempo médio de espera dos usuários numa estrutura de filas $M/G/K$, p. 55

$\overline{W}_{M/M/K}$ Tempo médio de espera dos usuários numa estrutura de filas $M/M/K$, p. 55

$\overline{X}$ Valor médio da variável aleatória $[X_i]_0^\infty$, independentes e identicamente distribuídas (iid), que seguem uma distribuição geral e representam o tempo de serviço do *i-ésimo* usuário ($i \to \infty$), com variância σ_X^2, p. 47

$\overline{Y}$ Valor médio da variável aleatória Y_i, que representa a soma dos intervalos de *polling*, que o *i-ésimo* usuário ($i \to \infty$) deve aguardar, desde o instante da sua chegada até a próxima visita do servidor (ciclo), com variância σ_Y^2, p. 53

ρ Segundo [4], o valor de ρ representa o fator de utilização do sistema, que corresponde à razão entre a taxa de trabalho que chega no sistema (por unidade de tempo) e a taxa máxima que sistema consegue atender este trabalho (por unidade de tempo), p. 47

ρ_p Fator de utilização do sistema, com que corresponde à razão entre a taxa de trabalho de prioridade p que chega ao sistema (unidade de tempo) e a taxa máxima que o sistema em executar este trabalho, p. 60

Lista de Abreviaturas

AC	*Access Category*, p. 30
AIFS	*Arbitration Interframe Space*, p. 30
AMPS	*Advanced Mobile Phone System*, p. 18
CDMA	*Code Division Multiple Access*, p. 16
CSMA/CA	*Carrier Sense Multiple Access/Collision Avoidance*, p. 17, 20
CSMA/CD	*Carrier Sense Multiple Access/Collision Detection*, p. 9, 17, 20, 33
CSMA	*Carrier Sense Multiple Access*, p. 17
CTS	*Clear To Send*, p. 18
CW	*Contention Window*, p. 30
DCF	*Distributed Coordination Function*, p. 20, 22
FCFS	*First Come First Served*, p. 27, 58, 59
FDMA	*Frequency Division Multiple Access*, p. 16
FIFO	*First In First Out*, p. 59
GSMA	*Global Scheduling Multiple Access*, p. 19
HOL	*Head Of Line*, p. 59, 74
IEEE	*Institute of Electrical and Electronics Engineers*, p. 35
IP	*Internet Protocol*, p. 29
LAN	*Local Area Network*, p. 21, 32
LCFS	*Last Come First Served*, p. 28
MANET	*Mobile Ad Hoc Network*, p. 5

NAV	*Network Allocation Vector*, p. 18
OP	Operador de Prioridades do Modelo de *polling*, p. 61
PAN	*Personal Area Network*, p. 32
PA	Ponto de Acesso, p. 10, 36
PCF	*Point Coordination Function*, p. 22
QoS	*Quality of Services*, p. 2, 30
RTS	*Request To Send*, p. 18
SJF	*Shortest Job First*, p. 26
SNIR	*Signal Noise plus Interference Ratio*, p. 43
SRMA	*Split-channel Reservation Multiple Access*, p. 19
TDMA	*Time Division Multiple Access*, p. 16
VIP	*Very Important Person*, p. 26
WLAN	*Wireless Local Area Network*, p. 1

Capítulo 1

Introdução

O termo "Floresta de *WiFi*"foi utilizado na referência [5] para ilustrar o cenário de grande densidade de redes sem fio, considerado pelo autor. Este termo pode ajudar a entender a dimensão e a permeabilização alcançada pelas redes sem fio no dia a dia das pessoas. Por exemplo: se estiver visitando um condomínio com muitos apartamentos, tendo cada apartamento seu próprio acesso à Internet, disponibilizado através de uma rede local sem fio (*WLAN*), poderá perceber que estará literalmente dentro de uma densa "Floresta de *WiFi*". Esta floresta é composta de pontos de acesso dos mais variados tipos e modelos e de clientes se utilizando destes pontos de acesso, com as suas mais diversas aplicações. O tema também pode ser abordado de uma forma um pouco diferente, analisando os diversos equipamentos que dependem da conectividade proporcionada pelas redes sem fio para funcionarem adequadamente. Figurativamente, pode-se dizer então que ambientes outrora "desertos", estão sendo "invadidos"por estes equipamentos conectados. Esta invasão em si, bem como as novas aplicações que surgem da exploração deste novo ambiente, é analisada na referência[6].

É claro que este novo contexto traz consigo novos desafios para os sistemas de comunicação. Seja para lidar com a garantida interferência mútua entre os pontos de acesso, seja para proporcionar o desempenho exigido pelas novas aplicações. Assim, esta nova relação de compromissos remete imediatamente à necessidade de protocolos de controle de acesso ao meio mais eficientes.

Um dos protocolos de controle de acesso mais estudados ao longo dos tempos é o protocolo **reserva implícita**, doravante chamado de protocolo de *polling* neste trabalho. O *polling*, como será detalhado nas seções a seguir, é um protocolo livre de contenção, que além de oferecer amplo suporte à operação com diferentes níveis de prioridade (por tráfego e por terminal), também pode operar com múltiplos canais. Estas características fazem do protocolo do tipo *polling* uma excelente opção na composição de uma solução para novos protocolos, mais eficientes dentro do cenário considerado.

O processo de desenvolvimento de novos protocolos passa necessariamente pela etapa de modelagem. O principal objetivo nesta etapa é antecipar as análises iniciais de comportamento do novo protocolo, de acordo com as especificações de projeto.

Apesar de existir muita literatura sobre os protocolos do tipo *polling*, não existem tantos trabalhos assim quando se trata de *polling* com múltiplos canais. Principalmente em se tratando de redes assimétricas[1] (vide [7]), as propostas existentes são geralmente complexas, requerendo ainda muitas aproximações para se viabilizarem matematicamente.

De uma forma resumida, este trabalho propõe a análise de um protocolo de controle de acesso de reserva implícita (*polling*), com capacidade de prover "Qualidade de Serviços"(QoS) para diferentes tipos de tráfego, baseado na classe de prioridade das mensagens. No protocolo considerado, o *polling* em cada um dos terminais da rede pode ser feito de forma cíclica ou aleatória. Além disto, cada terminal está limitado a transmitir um determinado número máximo de mensagens (L) por ciclo.

O protocolo analisado foi escolhido em face da sua simplicidade, sua capacidade de suportar funcionalidades (exemplo: múltiplos canais, diferenciação por prioridades) e principalmente pela sua flexibilidade para compor protocolos mais complexos, que podem inclusive ser aplicados para redes sem fio.

1.1 Controle de Acesso ao Meio

Em termos práticos, o conceito de redes de comutação por pacotes nasceu junto com o projeto Arpanet em 1969, cujos detalhes podem ser estudados no relatório [8]. O principal objetivo do projeto *Arpanet* era o desenvolvimento uma rede para interligar os computadores das bases militares americanas, que funcionasse independentemente de uma estrutura centralizada, no caso o Pentágono. Neste objetivo, a rede foi projetada de forma a garantir múltiplas opções de caminhos para se interligar uma determinada origem a cada um dos possíveis destinos. Um novo paradigma acabara de ser quebrado e surgia então uma nova escala de conectividade mundial, que veio a se tornar a Internet de hoje, seguindo as projeções citadas na referência [9].

Em se tratando mais especificamente das redes de computadores, o conceito de controle de acesso ao meio surge mais intensamente com o advento do compartilhamento de recursos. No trabalho de [9], os autores discorrem sobre este tema,

[1] Ao contrário das redes simétricas, nas redes assimétricas cada terminal da rede recebe uma taxa de chegada de mensagens diferente

abordando principalmente a questão da redução de custos de processamento numa escala global. A referência [9] enfatiza também o papel fundamental do protocolo de controle de acesso ao meio de organizador, otimizando a utilização compartilhada dos recursos pelos usuários da rede.

Um outro ponto interessante do trabalho de [9], que serviu inclusive como ponto de partida nesta dissertação, foi a análise comparativa dos tipos de controle de acesso que podem ser implementados nos protocolos. A Tabela 1.1 mostra uma análise comparativa, considerando as vantagens de cada método e o "preço"que é pago, de acordo com cada metodologia. Por exemplo: os protocolos que utilizam um método de controle dinâmico (*polling* ou de reserva) são livres de contenção e aproveitam bem os recursos de rede. Entretanto, estes protocolos "pagam"o preço do *overhead*, que são as mensagens trocadas na rede, para o controle na utilização dos recursos.

Tabela 1.1: Métodos de controle de acesso ao meio.

Método	`Colisão`	`Desperdício de recursos`	*Overhead*
Sem controle (ex: *ALOHA*)	SIM	NÃO	NÃO
Controle estático (ex: *FDMA*)	NÃO	SIM	NÃO
Controle dinâmico (ex: *polling*)	NÃO	SIM	SIM

1.1.1 Ambiente Sem Fio

A utilização de meio de transmissão sem fio é o principal assunto de muitos estudos desde a década de 70. Além da implícita facilidade que traz à mobilidade, a comunicação em *broadcast*, proporcionada pelo ambiente sem fio, é bastante aproveitada em muitos protocolos de controle de acesso ao meio. Dentre os muitos trabalhos relacionados, destacam-se [10–14].

Apesar das grandes vantagens já listadas no parágrafo anterior, as redes sem fio também oferecem alguns desafios, tais como: sensibilidade às condições climáticas (umidade), baixas taxas de transmissão, obstáculos físicos, interferências, etc. Todos estes desafios são potencializados com o fator mobilidade, proporcionado pelas redes sem fio.

Um dos principais desafios enfrentados pelos sistemas de transmissão sem fio atuais é a restrição dos rádios em operar somente no modo *half-duplex*. Os rádios *half-duplex* não conseguem receber e transmitir ao mesmo tempo, impossibilitando a detecção de uma colisão no momento de uma transmissão. Desta forma, protocolos como o *CSMA-CD* (*Carrier Sense Multiplo Access - Colision Detect*), que operam com detecção de colisão durante o processo de transmissão, tornam-se inviáveis.

O protocolo *CSMA-CA* (*Carrier Sense Multiplo Access - Colision Avoidance*) foi desenvolvido para contornar o problema causado pela restrição do rádio, implementando um mecanismo para evitar a colisão (*CA - Collision Ovoidance*). Apesar de minimizar a perda de desempenho por colisão, dois problemas permanecem particularmente importantes nas redes sem fio:

- Terminal Escondido: (Figura 1.1).

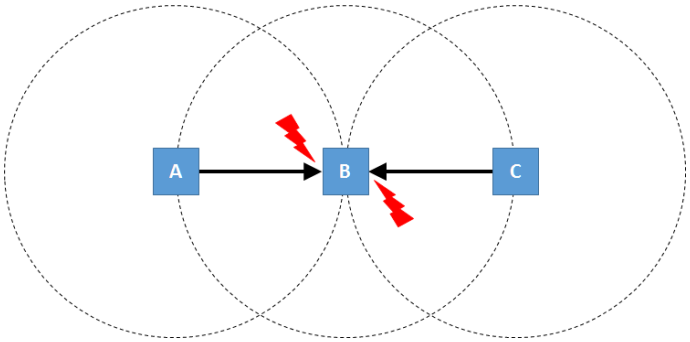

Figura 1.1: Colisão no terminal B por que o terminal C não consegue identificar a transmissão do terminal A.

- Terminal Exposto: (Figura 1.2).

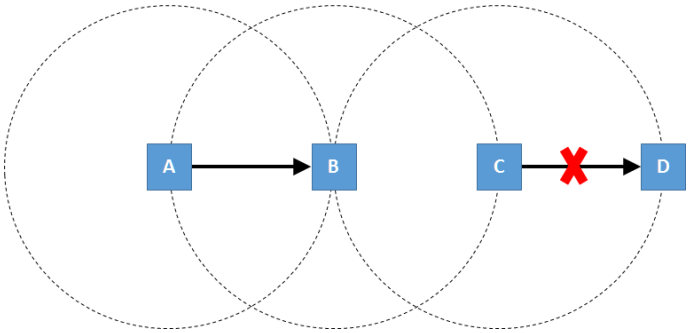

Figura 1.2: O terminal C fica impedido de transmitir para o terminal D por que se previne incorretamente de colidir com a transmissão em curso de A para B.

As redes sem fio ainda podem ser classificadas em: estruturada e não-estruturada. As redes estruturadas são aquelas que dependem de uma infra-estrutura de controle para funcionar. Um bom exemplo de rede estruturada é o sistema móvel celular. As redes não-estruturadas, ao contrário das redes estruturadas, dispensa a existência de um sistema de controle central para que dois terminais da mesma rede possam se comunicar. Um exemplo de rede não-estruturada é a rede *Ad Hoc*, que será apresentada com mais detalhes na próxima seção.

1.1.2 Redes *Ad hoc*

Na referência [15] as redes *Ad Hoc* são definidas como um conjunto de elementos com capacidade de comunicação sem fio, que podem se organizar dinamicamente em rede, formando topologias arbitrárias e temporárias, sem a necessidade de qualquer infra-estrutura pré-existente. Quando os elementos desta rede ainda têm a capacidade de se movimentar, a rede *Ad hoc* passa a ser conhecida como *MANET*.

As principais vantagens da rede *Ad Hoc* são: flexibilidade, baixo custo e robustez. Estas características fazem desta rede uma excelente opção para aplicações militares, operações de emergência, recuperação de desastres, etc.

A despeito das suas vantagens em relação às redes estruturadas, as rede *Ad Hoc* também trazem a tona uma série de desafios importantes, desde o projeto até a sua implementação. São eles:

- Restrições de banda nos canais: tentativas de acesso concorrentes, esvanecimento de sinal de rádio-freqüência, ruídos e interferências, etc. Estes fatores, que podem variar no tempo, acabam por limitar bastante a vazão efetiva dos canais de comunicação.
- Mobilidade: a topologia da rede está em constante alteração ao longo do tempo. Ou seja, um destino "alcançável"num determinado instante de tempo pode não ser mais no instante seguinte.
- Consumo de energia: os nós dependem de baterias internas para estarem ativos na rede.
- Segurança: numa rede que qualquer terminal pode entrar ou sair a qualquer momento, é preciso garantir a confidencialidade e/ou a integridade dos dados transmitidos.
- Alcançabilidade: muitas vezes a transmissão de uma origem para um destino só é possível passando por outros elementos, entre esta origem e este destino. A este tipo de comunicação que ocorre com múltiplos saltos, dá-se o nome *multi-hop*.
- Problema de terminal escondido e terminal exposto: como se sabe, a mitigação do problema de terminal escondido também acaba por provocar o problema do terminal exposto. No entanto, em se tratando de redes *Ad Hoc*, tanto as condições de ocorrência, como os impactos destes problemas são bem diferentes em relação a uma rede estruturada. Estes problemas são bem analisados no trabalho de [16].

Significa afirmar que, além da lista de objetivos mostrada na Seção 1.1, o controle de acesso para as redes *Ad hoc* ainda deve endereçar todas estas dificuldades listadas acima.

1.2 Utilização de Modelos

A necessidade cada vez maior de informações e serviços inteligentes transformam o dia a dia das pessoas numa verdadeira "corrida"contra o tempo. Nesta corrida, quase sem perceber, uma pessoa comum lida diariamente com um número muito grande de sistemas. Por exemplo: ao acordar pela manhã e ligar a televisão para assistir o noticiário, a pessoa automaticamente se conecta com um grande sistema de transferência de conteúdo; ao utilizar o seu cartão de transporte no ônibus para o trabalho, o usuário utiliza um sistema complexo de transferência de dados até um banco de créditos centralizado, etc. A evolução tecnológica atua como parceira da humanidade de forma a simplificar e tornar possível tarefas outrora impensadas. Entretanto, esta simplificação, proporcionada externamente pela evolução tecnológica, deve ser assimilada internamente pelos sistemas, que ao contrário, tornam-se cada vez mais complexos.

É claro que estes sistemas não são desenvolvidos sem que seja cumprido um longo e meticuloso processo. Primeiramente, para que possam ser entendidos realmente como uma solução viável, os sistemas devem ser projetados de acordo com as necessidades e com objetivos bem especificados. Além disto, as respostas esperadas destes sistemas precisam ser suficientemente rápidas e precisas. Estas questões fazem com que o esforço necessário para o desenvolvimento de grandes sistemas sejam enormes e caros. Neste contexto, surgem os modelos.

Segundo [23], o desenvolvimento de qualquer sistema passa necessariamente por três fases distintas, mas que se interdependem mutuamente:

- Projeto: nesta fase são tomadas decisões fundamentais que irão determinar o rumo do projeto e o seu sucesso.
- Planejamento de capacidade: nesta etapa são estimadas a carga e a capacidade do sistema, na sua interação com o seu ambiante.
- Evolução: nesta fase o comportamento do sistema é avaliado considerando um eventual crescimento da demanda e/ou com novas demandas que possam ser assimiladas.

Mesmo cumprindo adequadamente estas etapas, é preciso antecipar a análise do comportamento do sistema funcionando, interagindo normalmente com as entradas

e as saídas; e sob condições especiais. Há basicamente três maneiras diferentes de se analisar um determinado sistema:

- Utilizar a intuição de especialistas para extrapolar situações conhecidas e projetar o comportamento do sistema. Esta abordagem é difícil por que exige algo que é cada vez mais caro e raro.

- A avaliação experimental do sistema é uma abordagem interessante que permite predizer com boa precisão o comportamento de um sistema. Entretanto, esta opção, quando possível, é onerosa e pode ser até proibitiva, dependendo da complexidade e da amplitude do sistema. Uma outra desvantagem é a dificuldade de se obter resultados mais genéricos a partir de um conjunto de considerações assumidas para viabilizar o experimento.

- Entre estas duas abordagens, existe a modelagem. Diferentemente da intuição, que proporciona resultados rápidos mas com pouca confiabilidade; e do experimento, que apresenta resultados precisos mas que pode não ser suficientemente flexível, a modelagem representa um meio termo interessante.

A modelagem representa a abstração de um sistema, fugindo um pouco da massa de detalhes que o caracteriza como um todo e se apegando àqueles aspectos que realmente são interessantes no contexto estudado.

Neste trabalho, é estudado o comportamento de um protocolo de comunicação sem fio, onde os terminais se comunicam entre si através de múltiplos canais. Cada terminal recebe uma taxa contínua de mensagens, que ficam armazenadas nos terminais, até que sejam transmitidas aos seus destinos. A utilização destes múltiplos canais pelos terminais é controlada, para evitar que comunicações simultâneas possam se interferir mutuamente.

Para o sistema descrito brevemente acima, será utilizado um modelo baseado na teoria básica de filas, onde os terminais são representados pelas filas, as mensagens são representadas por usuários, que aguardam na fila para serem atendidos; e os canais são representados por servidores, destinados a os atender.

Também será utilizado um modelo de simulação, que foi desenvolvido com base numa arquitetura do tipo “produtor x consumidor”. Neste sistema, as mensagens são produzidas, armazenadas nas filas dos terminais virtuais e depois de um determinado tempo (tempo de serviço) as mensagens são consumidas (apagadas).

1.3 Motivação

Este trabalho é motivado pela análise do modelo analítico proposto em [1], onde os autores apresentam um mecanismo baseado num sistema de *polling* para controlar o problema de "destino ocupado", numa rede sem fio com múltiplos canais.

O problema de destino ocupado se caracteriza basicamente pela perda da oportunidade de acesso ao meio pelo terminal A, que deseja transmitir suas mensagens ao terminal B, que por sua vez está ocupado, transmitindo ou recebendo de um outro terminal.

À despeito da aplicação específica do mecanismo de controle proposto em [1], a flexibilidade, a simplicidade e a precisão dos resultados numéricos obtidos no modelo proposto em [1], chamaram a atenção como uma oportunidade para se ampliar a sua aplicação, na análise de protocolos de controle de acesso mais complexos, que podem ser modelados a partir de sistemas de *polling*, com e sem diferenciação por prioridades.

1.4 Contribuição

Apesar desta dissertação ter sido motivada pelo trabalho em [1], as contribuições propostas por ambos são bem diferentes. Na referência [1] é proposto um protocolo, com seu respectivo modelo analítico e de simulação, cujo objetivo é reduzir o impacto do problema de destino ocupado no tempo médio de espera em fila, conforme descrito na Seção 1.3.

A principal contribuição deste trabalho é a apresentação de uma proposta de modelo analítico de uso geral, baseado num protocolo de controle de acesso de *polling* com múltiplos canais, com e sem diferenciação por prioridades. Complementarmente, é apresentado também um modelo de simulação, que foi utilizado neste trabalho para validação de resultados; além de alguns outros modelos analíticos mais específicos, desenvolvidos a partir do modelo analítico principal.

A ideia de se construir um modelo analítico baseado no protocolo de *polling*, surgiu em virtude da simplicidade dos modelos propostos na vasta literatura existente; mas sobretudo do fato de que o protocolo de *polling* também é largamente empregado como base de outros protocolos de controle de acesso mais complexos, aumentando assim a aplicabilidade dos modelos propostos neste trabalho.

Ao todo, são então apresentados 5 modelos analíticos diferentes, que estão listados logo abaixo:

- Modelo para o tempo médio de espera em fila de um sistema de *polling*, com um único canal/servidor, operando segundo uma regra de atendimento cíclica - (3.12).
- Modelo para o tempo médio de espera em fila de um sistema de *polling*, com múltiplos canais/servidores, operando segundo uma regra de atendimento cíclica - (3.17).
- Modelo para o tempo médio de espera em fila de um sistema de *polling*, com múltiplos canais, operando segundo uma regra de atendimento cíclica, com diferenciação por prioridades - (3.21).
- Modelo para o tempo médio de espera em fila de um sistema de *polling*, com múltiplos canais, operando segundo uma regra de atendimento randômica - (3.24).
- Modelo para o tempo médio de espera em fila de um sistema de *polling*, com múltiplos canais, operando segundo uma regra de atendimento cíclica, com diferenciação por prioridades - (3.25).

A comparação entre os resultados obtidos dos modelos analíticos propostos e os resultados obtidos nas simulações, mostram níveis de precisão bastante razoáveis, considerando a simplicidade e a flexibilidade dos modelos. Esta flexibilidade inclusive é explorada nos diversos cenários de testes, que foram pensados com o objetivo de analisar o comportamento do protocolo em algumas situações características e de interesse experimental.

1.5 Trabalhos Relacionados

A utilização de múltiplos canais para aumentar a vasão agregada de uma rede não pode ser considerada uma ideia nova, extrapolando inclusive o tipo de rede considerada (com fio e sem fio) e o tipo de protocolo utilizado.

Existem vários estudos com resultados expressivos, que mostram o quão promissora é a abordagem de múltiplos canais. Um destes trabalhos é a referência MARSAN e ROFFINELLA [26], que analisa o desempenho de vários protocolos utilizados em redes locais, através de modelos matemáticos modificados para múltiplos canais.

Um outro trabalho interessante, apresentado em MARSAN e NERI [27], analisa especificamente o comportamento do protocolo *CSMA/CD*, considerando a utilização de múltiplos canais *broadcast*, ao invés de apenas um único canal. Os resultados realmente demostram que a utilização de múltiplos canais aumenta consideravelmente a vazão da rede, além de diminuir a variância no tempo médio de espera em fila.

A partir dos resultados obtidos dos trabalhos sobre os ganhos com a utilização de múltiplos canais, vários trabalhos e muitas propostas de protocolos com múltiplos canais, desenvolvidos a partir de diferentes abordagens. A referência WANG *et al.* [28], consolida as principais propostas, de acordo com estas diferentes abordagens e metodologias empregadas:

- Canal de Controle Dedicado.
- Saltos de Canal.
- Divisão de Tempo.
- Múltiplos Rádios.

De acordo com a referência ANTHONY e WATSON [29], a utilização de modelos matemáticos na análise de sistemas complexos oferece ao projetista informações essenciais para o desenvolvimento de um projeto, principalmente na fase inicial, quando se faz necessário tomar decisões determinantes. Ainda dentro deste tema, a referência LAZOWSKA *et al.* [23], discute os principais desafios para se obter a máxima aderência do modelo analítico ao sistema real, utilizando sistemas de filas. Mesmo com as vantagens mencionadas e mesmo com o grande número de propostas de novos protocolos, não é tão simples encontrar modelos analíticos para protocolos de controle de acesso ao meio com múltiplos canais. Esta escassez se deve sobretudo à complexidade envolvida no desenvolvimento dos modelos com múltiplos servidores e às aproximações necessárias para sua viabilização matemática.

Um exemplo de modelagem de sistemas com múltiplos servidores a partir de estruturas de filas é apresentado na referência MARSAN *et al.* [7]. Os autores de [7] consideram um sistema de *polling* em regime estacionário, operando com múltiplos servidores. As análises são feitas sobre dois modelos, com disciplinas de serviço diferentes: exaustiva e *gated*. A partir destas análises, são apresentados resultados sobre o tempo de ciclo, da estabilidade do sistema e do tempo médio de espera em fila. Os modelos ainda consideram diferentes tipos de atendimento às filas:

1. No um servidor atende cada fila ($1 \times Q$).
2. Qualquer número de servidores (S) podem atender simultânea e cooperativamente cada fila ($S \times Q$).

Um outro trabalho de modelagem com múltiplos servidores é apresentado na referência DE MORAES e NOBREGA [1], onde é proposto um mecanismo de controle para minimizar os efeitos do problema de "destino ocupado"numa rede estruturada, com múltiplos canais, controlado por um sistema de *polling* A partir de um Ponto de Acesso (PA).

No mecanismo proposto, o PA pergunta a cada terminal (na sequência) sobre a sua intenção de transmitir mensagens para a rede. Se a resposta do terminal (terminal A) perguntado for positiva, isto é, se ele responder que deseja transmitir, o terminal informa, na própria resposta, a identificação do terminal de destino das suas mensagens (terminal B). Considerando que o PA tem informações atualizadas sobre o estado de ocupação de cada um dos terminais da rede, se o terminal B estiver ocupado, ao invés de simplesmente negar a oportunidade de transmissão ao terminal A naquele ciclo, o PA insere um registro numa fila de adiamento, que tem prioridade sobre a fila normal da sequência do *polling*.

Assim, sempre depois que acontecer uma transmissão com sucesso e o *polling* for se deslocar para o próximo terminal na sequência, esta operação é interceptada pelo PA, que cede o controle do *polling* ao primeiro terminal da fila de adiamento. Consequentemente, o tempo que o terminal A esperaria para receber novamente a oportunidade de transmissão (que poderia não acontecer novamente se o terminal B estivesse novamente ocupado), é reduzido, diminuindo então o tempo médio de espera na fila.

Para modelar o mecanismo descrito acima, os autores de [1] se basearam num sistema de *polling*, com algumas aproximações a saber:

- Aproximação para limitar em L o número máximo de mensagens transmitidas, proposta por [30].
- Aproximação do termo da expressão que decorre do cálculo do tempo médio residual de serviço, proposto por [31] para uma estrutura de fila $M/G/K$.
- Aproximação do cálculo da probabilidade de uma determinada mensagem encontrar todos os servidores ocupados, proposta por [32].

Para avaliar a precisão das aproximações utilizadas em [1], dentro dos objetivos iniciais deste trabalho, foram analisadas outras propostas. Uma destas propostas está contida na referência KIMURA [33]. De acordo com KIMURA [33], a aproximação do tempo médio de espera em fila do sistema $M/G/K$ apresentada em [31] e utilizada no modelo proposto na referência [1], subestima o valor do tempo médio residual, quando o sistema está operando com utilização alta ($\rho \lesssim 1$). Neste mesmo artigo, o autor de [33] apresenta uma proposta de aproximação para $M/G/K$ e a analisa, comparando-a com os resultados obtidos de outras propostas e valores exatos. O resultado desta análise comprova a precisão do modelo (3.20), proposto em [33], principalmente quando o número de servidores é pequeno ($K < 10$).

A partir do modelo base, definido para o tempo médio de espera em fila com múltiplos servidores, buscou-se aumentar a aplicabilidade da proposta, adicionando

a funcionalidade de diferenciação por prioridades. Com este objetivo, buscou-se informações básicas, mas que pudessem orientar estudos mais profundos na busca do modelo objetivo. Um destes trabalhos está contido no livro de KLEINROCK [34], que analisa a disciplina de fila *HOL - Head Of Line*. Nesta disciplina, as mensagens possuem diferentes classes de prioridades e são servidas de acordo com esta classe. As mensagens de prioridades mais altas são servidas primeiro. As mensagens dentro de uma mesma classe de prioridade são servidas na ordem de chegada na fila (*FCFS-First Come First Served*).

Para uma maior aproximação do modelo objetivo, também foram buscadas informações básicas de diferenciação por prioridades, voltadas para sistemas de *polling*. Na referência DE MORAES [35] é apresentada uma visão geral de diversos modelos de protocolos de comunicação multi-acesso, que empregam mecanismos de priorização por tráfego e por terminal. O autor também apresenta alguns resultados do desempenho de alguns destes modelos, além de disponibilizar uma extensa lista de referências relacionadas ao assunto principal.
O trabalho lista as principais propriedades operacionais, que são comumente requeridas de um sistema de prioridades:

- Independência hierárquica de performance: capacidade de manter o tráfego de mensagens de mais alta prioridade, imune ao tráfego de mensagens de classes de prioridades mais baixas.

- Isonomia de tratamento dentro de uma mesma classe (*fairness*): capacidade de tratar de forma igual (justa) as mensagens que pertencem a uma mesma classe de prioridade.

Considerando as características de flexibilidade do modelo base definido neste trabalho, junto com o conhecimento de diferenciação por prioridades, adquirido nos trabalhos estudados, foram analisados alguns modelos matemáticos que pudessem orientar a implementação desejada, a partir do modelo base. No livro de TAKAGI [36] é apresentada uma análise completa da teoria de filas, iniciando com o estudo da estrutura mais geral $M/G/1$ e finalizando com a análise dos sistema de *polling* com prioridades. Destaca-se neste livro o Capítulo 3.6, que analisa os sistemas de *polling* com a disciplina de serviço do tipo *gated*. Na Seção *Vacation for Each Batch*, do mesmo capítulo, é proposto um modelo considerando um período de férias do servidor, que é imediatamente reiniciado, no caso de não haver mensagens para serem servidas nas filas. Algumas simplificações matemáticas no modelo de [36] permitiram a sua a utilização na adaptação do modelo base, para prover diferenciação por prioridades ao mesmo.

Ainda se aproveitando da flexibilidade do modelo base, proposto neste trabalho, pensou-se em mais uma adaptação, no sentido de se viabilizar análises de protoco-

los com múltiplos canais, que podem ser utilizados para redes *Ad Hoc*. Conforme explicado na Seção 1.1.2, uma das características das redes *Ad Hoc* é o alto nível de independência entre os terminais e a consequente aleatoriedade das tentativas de acesso ao meio pelos mesmos. Esta característica, aliada com a disponibilidade de múltiplos canais, fazem do protocolo de *polling* cíclico uma opção pouco aderente ao funcionamento da rede. No artigo de KLEINROCK e LEVY [37] é apresentado um modelo para sistemas de *polling* com disciplina exaustiva e *gated*, onde a próxima fila a ser servida é escolhida aleatoriamente com distribuição de prioridades uniforme. Esta escolha aleatória é representada no modelo final proposto pela adição de mais um termo que aumenta o tempo médio de espera em fila, em comparação com a escolha cíclica.

Um exemplo de protocolo para redes *Ad Hoc*, que pode ser analisado com o modelo de *polling* aleatório apresentado neste trabalho, é proposto na referência CHEN *et al.* [24]. O referido protocolo em [24] define um canal de controle e K canais de dados. O acesso ao canal de controle é disputado aleatoriamente pelos terminais em regime de contenção. Depois de ganhar a competição pelo canal de controle, o terminal vencedor deve negociar a ocupação de um dos K canais de dados com o terminal de destino, para poder transmitir. No fim da transmissão, tanto o terminal de origem quanto o terminal de destino, voltam a sintonizar seus rádios no canal de controle.

A fim de avaliar a precisão dos modelos matemáticos propostos neste trabalho, foi desenvolvido um programa de simulação com base numa estrutura "produtor x consumidor". A referência ALQAHTANI [3] analisa o desempenho do sistema de *polling*, operando em com diferentes disciplinas de serviço. As análises foram feitas através de um simulador próprio, cujo fluxograma proposto foi utilizado no desenvolvimento do simulador apresentado neste trabalho.

A última parte deste trabalho compara os resultados obtidos dos modelos matemáticos, com os resultados obtidos do modelo de simulação. A inferência estatística, utilizada a apresentação dos resultados da simulação, foi feita a partir da referência LEON-GARCIA [38].

1.6 Organização

Este trabalho está organizado da seguinte forma:

1. O Capítulo 2 contém uma visão geral sobre os protocolos de controle de acesso ao meio, mostrando inicialmente as categorias e propondo uma taxonomia de classificação, baseada nos métodos de controle de acesso e estratégias de alocação de recursos. Na Seção 2.1, é enfatizada a amplitude de utilização

do protocolo de *polling*. Dentro de algumas condições, que também foram explicadas, o protocolo de *polling* é apresentado como uma boa opção de plataforma para suportar as diferentes funcionalidades requeridas para operação com múltiplos canais e com diferenciação por prioridades. A Seção 2.2 oferece uma visão geral sobre as estratégias de diferenciação por prioridades, suas aplicações e uma teoria básica envolvendo conservação de energia e equidade. Na Seção 2.3 aborda a utilização de múltiplos canais, como uma proposta promissora no sentido de aumentar a vazão agregada da rede sem fio, potencializando o reuso de canais e diminuindo a interferência. Também são apresentadas algumas propostas de protocolos de controle de acesso com múltiplos canais. A Seção 2.3.2 detalha um pouco melhor a utilização de múltiplos de canais no protocolo de *polling*.

2. O Capítulo 3 apresenta noções teóricas sobre o processo de modelagem. Em seguida o capítulo evolui para a especificação do protocolo que será modelado, medidas de interesse, hipóteses adotadas e definição dos modelos. A partir da definição dos modelos analíticos e de simulação, é mostrada todo o embasamento conceitual e matemático para se chegar aos mesmos.

3. O Capítulo 4 descreve a estrutura básica dos programas de simulação e os detalhes em relação a cada uma das classes do programa.

4. No Capítulo 5 os resultados numéricos são comparados e analisados em cada uma das situações importantes, definidas no trabalho.

5. No Capítulo 6 é feita uma análise final de todo o trabalho, ressaltando os principais resultados e indicando possíveis trabalhos futuros.

Capítulo 2

Protocolos de Controle de Acesso ao Meio

Relembrando a citação feita na seção anterior, o controle de acesso é responsável pela organização na utilização compartilhada dos recursos pelos usuários de um sistema, seja ele qual for: pode ser uma fila de banco, uma fila de processos dentro do computador ou uma rede de telecomunicações. Neste sentido, muito esforço tem sido gasto no desenvolvimento de protocolos de comunicação que consigam atender as necessidades de cada sistema. Apesar de muitos requisitos funcionais dos protocolos serem definidos a partir das necessidades específicas das aplicações suportadas, pode-se dizer que a maioria dos protocolos de controle de acesso ao meio trabalha para reduzir o tempo máximo de espera em fila ou reduzir o número médio de usuários aguardando até serem atendidos.

De uma forma geral e dependendo de sua aplicação, os protocolos de controle de acesso são especificados de acordo com algumas premissas. Estas premissas podem ser:

- Inicialização: espera-se que os terminais entrem no estado requerido para operação de forma automática e logo após serem ligados.
- Equidade (ou *Fairness*): a rede deve compartilhar seus recursos de forma justa entre todos os terminais dentro de uma mesma classe prioridade.
- Prioridade: a rede deve ser capaz de lidar com diferentes prioridades de tráfego e/ou de terminal, de modo a permitir a convivência entre diferentes tipos de serviços.
- Controle: a rede deve ser capaz de exercer controle sobre os terminais para evitar ou diminuir a probabilidade de colisões.

- Recepção: as mensagens devem ser todas entregues ao seu destino sem duplicação e na seqüência correta.
- Detecção de erro: a técnica deve ser capaz de detectar a ocorrência de um erro.
- Recuperação: a técnica deve ser capaz de se recuperar de uma situação anormal, por exemplo: se duas mensagens colidirem, a rede deve ser capaz de reordenar os terminais para nova transmissão sem colisão.
- Reconfiguração: a rede deve ser capaz de se reconfigurar automaticamente com a saída ou a entrada de um novo terminal.
- Compatibilidade: a técnica deve acomodar equipamentos de fornecedores diferentes, mas que operem dentro da sua especificação funcional.
- Robustez: a rede deve continuar funcionando, mesmo se um ou mais terminais falharem.

Com relação à metodologia empregada, no trabalho de [17] é apresentada uma análise bastante ampla sobre os diversos tipos de protocolos de acesso ao meio. Desta análise é possível montar uma boa sugestão de taxonomia, com o objetivo de facilitar o entendimento dos capítulos posteriores.

Com relação à metodologia de alocação de recursos, o protocolo pode ser dividido em:

- Alocação Fixa.

  A técnica de alocação fixa se caracteriza principalmente pela divisão e distribuição estática dos recursos para os usuários, independentemente da demanda de cada um. Tem a vantagem de garantir o recurso para o usuário sempre, mas ocasiona desperdício de recursos, quando um usuário não tem demanda. O *FDMA*-(*Frequency Division Multiple Access*), o *TDMA*-(*Time Division Multiple Access*) e o *CDMA*-(*Code Division Multiple Access*) são exemplos de técnicas de alocação fixa.

  A Figura 2.1 ajuda a entender as diferenças entre os exemplos de sistema de alocação fixa.

- Alocação Aleatória (ou Acesso Aleatório).

  De acordo com [9], o tráfego de dados entre dois computadores é caracterizado pela sua disposição em forma de "rajada". Isto é, períodos grandes de inatividade seguidos de períodos curtos com alta concentração de tráfego.

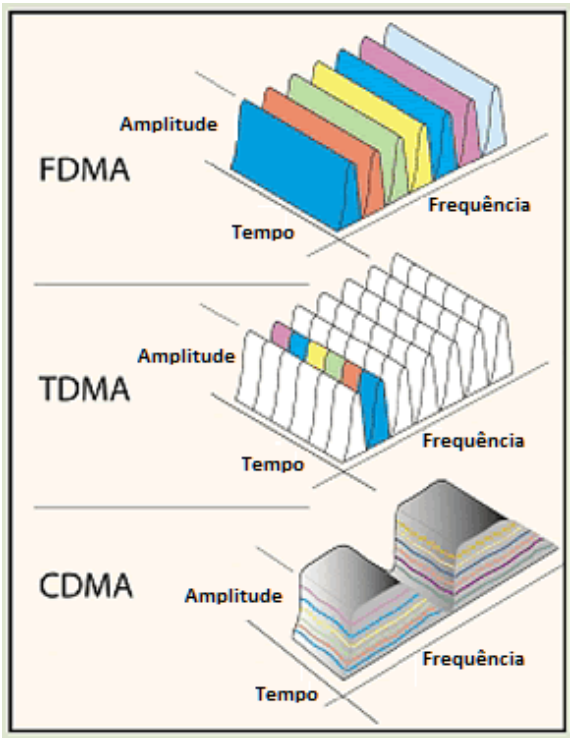

Figura 2.1: Diferenças entre os sistemas de alocação fixa *FDMA, TDMA, CDMA*.

A técnica de alocação aleatória aproveita esta característica de "rajada"do tráfego entre computadores para alocar todo o recurso de uma só vez a um único usuário, quando este precisar, diminuindo o tempo de resposta do sistema. De acordo com [4], a "Lei dos Grandes Números"garante, com uma probabilidade muito alta, que a demanda de transmissão num determinado instante é igual a soma das demandas médias de toda a população de usuários no sistema, quando este número é muito grande. Assim, a técnica de acesso aleatório consegue otimizar bastante o compartilhamento do meio em relação à técnica anterior de alocação fixa, considerando uma rede com tráfego baixo ($\rho < 0,2$ no caso do protocolo *Slotted Aloha*).

O protocolo *Aloha* foi desenvolvido em 1970 como precursor nesta técnica de controle de acesso. Depois deste, vários protocolos foram desenvolvidos, sendo o mais comum o *CSMA* e suas variações *CSMA/CD* e o *CSMA/CA*, este último utilizado em redes sem fio. Apesar dos muitos protocolos existentes, nenhum deles consegue evitar por completo o problema das colisões.

Uma colisão ocorre quando dois ou mais terminais tentam usar o mesmo meio de comunicação no mesmo instante de tempo. Nesta situação, geralmente todas as tentativas de transmissão são perdidas e a rede perde desempenho. Existem muitos mecanismos para minimizar a ocorrência de colisões. O protocolo *CSMA/CA* no modo *DCF-Distributed Coordination Function*

implementa um mecanismo conhecido como *Collision Avoidance - CA*. Este mecanismo utiliza um sistema de *backoff* e janelas de contenção (CW) para diminuir a probabilidade de colisões (Vide [18]).

Uma outra forma de controle de colisões possível no modo *DCF* do protocolo *CSMA/CA* é através de sinalização de controle (*RTS-Request To Send/CTS-Clear To Send*) e portadora virtual (*NAV-Network Allocation Vector*).

- Alocação Dinâmica.

  As técnicas anteriores representam os dois extremos em relação à metodologia de alocação de recursos. O sistema de alocação dinâmica, representa um meio termo que consegue evitar as colisões e ao mesmo tempo otimizar a utilização de recursos da rede. Isto é feito através da troca de informações de controle entre os membros da rede, garantindo com que o compartilhamento dos recursos aconteça de forma organizada.

Os sistemas de alocação dinâmica ainda podem ser classificados de acordo com tipo de controle que é exercido na rede:

- Controle Centralizado.

  O sistema centralizado de alocação dinâmica compartilha os recursos da rede a partir de um único controlador central, com base na informação da demanda de transmissão de cada um dos terminais do sistema. Se um terminal tem demanda de transmissão, ele a informa à estação centralizada, que a organiza frente às outras demandas e distribui o recurso de forma escalonada e livre de contenção. Este sistema possui duas vulnerabilidades importantes que são: o funcionamento de toda rede depende do controlador central e o *overhead* entre o controlador e os terminais, principalmente quando os tempos de propagação são grandes. Este tipo de sistema possui algumas variações:

  - Sistemas Orientados a Circuitos: neste método, cada terminal tem um sub-canal de controle permanentemente estabelecido com a entidade central, por onde explicita a sua demanda. A desvantagens deste sistema são principalmente: a própria necessidade e condição permanentemente do canal de controle, que pode ser considerado um *overhead*. Um bom exemplo deste tipo de método é o protocolo *AMPS-Advanced Mobile Phone System*, que foi muito utilizado no Brasil na década de 90, no início do sistema móvel celular. maiores detalhes sobre o protocolo *AMPS* podem ser encontrados na referência [19].
  - Sistemas de *Polling* ou Reserva Implícita: a referência [20] explica o *polling* como um sistema de de multiplexação baseado em comutação

por pacotes, onde uma entidade central pergunta implicitamente a cada terminal da rede se este tem alguma demanda de transmissão. Se o terminal responde positivamente, a entidade central aloca todo o recurso possível para este terminal. Se a resposta for negativa, a entidade central segue na **ordem** de perguntas para o próximo terminal.

Este sistema, apesar de bastante eficiente, apresenta algumas desvantagens que limitam sua utilização. A principal delas é o *overhead* causado pela necessidade constante de troca de mensagens de controle entre a unidade central e os terminais. Além desta desvantagem, o sistema de *polling* também desperdiça recursos no ciclo de "perguntas"para terminais sem demanda, ou cujo destino esteja ocupado (vide [1]).

- Sistemas de *Probing*: o sistema de *probing* minimiza uma das principais desvantagens do sistema de *pooling*, que são as perguntas desnecessárias para terminais sem demanda de transmissão. Neste sistema, a unidade central emite uma pergunta geral em *broadcast* para todos os terminais e repete a mesma pergunta, enquanto não receber uma resposta positiva de algum dos terminais. Assim que receber uma confirmação de demanda de algum terminal perguntado, a unidade central vai dividindo o conjunto de terminais em sub-conjuntos e repetindo a pergunta aos sub-conjuntos, até que o terminal com demanda seja isolado e identificado. O protocolo proposto na referência [21] é um bom exemplo de sistema de *pobing*.

- Sistema de Reserva: neste sistema o terminal explicita sua demanda diretamente à unidade central, que controla o compartilhamento dos recursos. Cada terminal tem acesso com contenção a um canal de controle, que fica permanentemente conectado à unidade central. Quando um terminal tem uma demanda ele simplesmente tenta **tomar posse** do canal de controle e envia sua solicitação de reserva à unidade central. A unidade central organiza as solicitações de reserva dos terminais e aloca recursos sem contenção, de acordo com o algorítimo de alocação.

  A probabilidade de colisões no canal de controle existe, mas é pequena por causa do tamanho reduzido dos pacotes de reserva. Um bom exemplo deste tipo de protocolo é o *SRMA* (*Split-channel Reservation Multiple Access*), utilizado principalmente em sistemas de comunicação por satélite.

- Sistema de Agendamento Global: o *GSMA* (*Global Scheduling Multiple Access*), como é conhecido, é um sistema de reserva baseado em comutação de pacotes, que opera sem contenção[1]. Todos os terminais são permanentemente associados a um sub-canal exclusivo, por onde trans-

[1] A primeira fase do protocolo, do envio das reservas, opera com contenção.

mitem as solicitações de reserva e por onde também recebem as confirmações. Como os canais são fixamente associados à cada terminal, não há a necessidade de identificadores, assim o *overhead* é reduzido. Um bom exemplo deste sistema é o modelo proposto na referência [22].

- Controle Distribuído.

  O sistema de alocação dinâmica com controle distribuído pode ser considerado uma evolução do sistema centralizado, uma vez que o controle é feito pelos próprios terminais da rede, de forma descentralizada. Entretanto, a necessidade do controle distribuído do meio de acesso, através do próprio meio de acesso, representa um novo desafio para o protocolo. Apesar das dificuldades, os ganhos também são importantes: principalmente com relação à confiabilidade de não mais depender de um único controlador central e com relação ao retardo na negociação, especialmente significante em se tratando de altos tempos de propagação (satélite).

  Neste tipo de sistema, todos os terminais da rede precisam receber a mesma informação atualizada sobre a utilização do meio. Isto é fundamental para que o algoritmo de controle de acesso, presente em cada terminal, possa funcionar corretamente, aumentando a eficiência do protocolo.

  Existem vários exemplos de protocolos que implementam o controle distribuído, dependendo do tipo do meio de comunicação envolvido: CSMA/CD, CSMA/CA (modo *DCF*), e o *Token-ring*. Neste último, o terminal ganha acesso ao meio quando recebe o *token* de um terminal anterior na seqüência. Assim que acaba de transmitir, o terminal que está com o *token* o repassa adiante na mesma seqüência, que é comum a todos os terminais.

2.1 Protocolo de *Polling*

O protocolo de *polling* foi introduzido no Capítulo 2 como um tipo de protocolo de controle de acesso ao meio, onde os terminais são implicitamente "convidados"a transmitir, por um único canal, através de uma consulta ordenada, que pode ser cíclica ou aleatória, a partir de uma entidade central. A Figura 2.2 ajuda a entender o cenário em questão.

De acordo com [46], muito mais que um tipo de protocolo, o sistema de *polling* é um modelo que pode ser aplicado para estudar uma vastidão de situações diferentes. Por exemplo: processo de verificação e reparo de máquinas numa indústria, nos processos de visitas médicas aos doentes num hospital, transferência

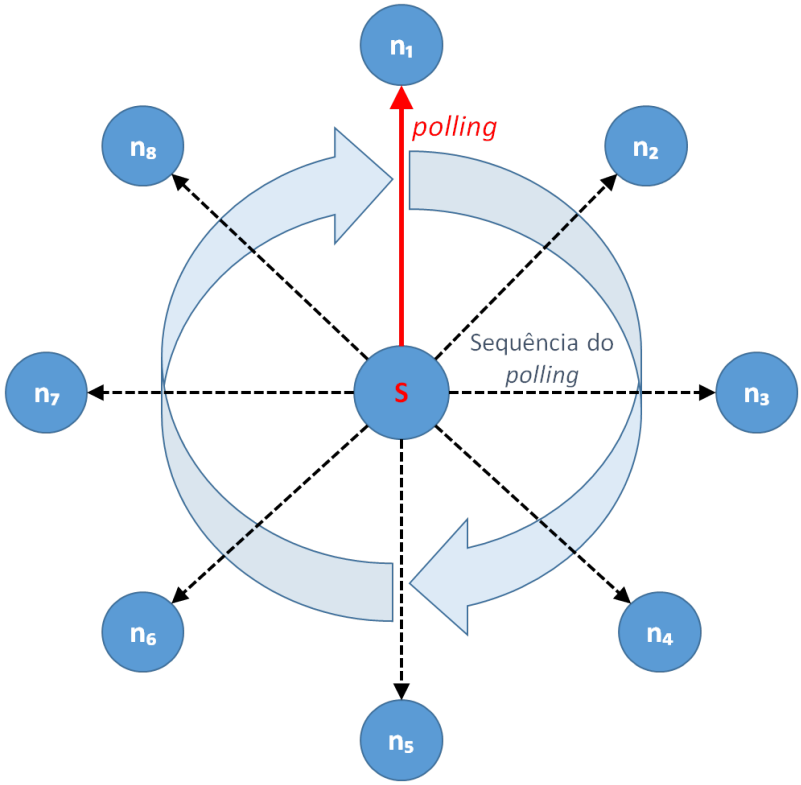

Figura 2.2: Entidade de controle central S inquirindo os terminais n_i.

de dados a partir de terminais remotos para um computador central, estudos de protocolos para redes locais (*LANs*) à base de *token*; e mais recentemente, no processo de compartilhamento de recursos de balanceamento de carga em computadores com múltiplos processadores.

Apesar de ter sido iniciado na Seção 2 como um protocolo de controle de acesso centralizado, o modelo de *polling* também pode ser utilizado para estudar protocolos de controle distribuído, que é feito pelos próprios terminais da rede, de forma descentralizada ou distribuída.

Para que este controle distribuído funcione adequadamente, é essencial que todos os terminais possuam as informações necessárias da rede. Por exemplo: o terminal que acaba de transmitir a sua mensagem, deve necessariamente saber qual será o próximo terminal que irá possuir este "direito", no caso de um protocolo que utiliza o *token*.

Segundo [17], os protocolos que utilizam sistemas de *polling* com controle centralizado são eficientes dentro de algumas condições:

- Tempo de propagação pequeno. Num sistema de comunicação qualquer, o tempo de propagação pode ser entendido como o tempo necessário para que um sinal eletromagnético percorra todo o meio de transmissão, até chegar ao seu

destino, na velocidade da luz (ou bem próxima). Como o tempo de propagação é diretamente proporcional a distância percorrida pelo sinal, quanto maior a distância, maior será o tempo de propagação do sinal.

No caso de uma rede sem fio com controle de acesso de *polling*, se o tempo de propagação for muito grande, o tempo gasto com o *overhead* será excessivamente grande, mesmo sendo estas mensagens de *polling* bem pequenas em relação às mensagens de dados.

- Considere a seguinte relação $\beta = (X_m)/(X_p)$ entre o tamanho médio das mensagens de dados e o tamanho médio das mensagens de *polling*. A eficiência do sistema é diretamente proporcional ao valor de β.
- Se o tráfego de cada terminal não se comportar como "rajadas"muito intensas.
- Se a população de terminais não for muito grande.

De uma forma geral, o trabalho em [47] mostra que os protocolos à base de *polling* apresentam um desempenho superior[2] em relação à outros protocolos com contenção, quando submetidos à cargas de tráfego muito intensas.

Um bom exemplo de utilização do sistema de *polling* é o próprio protocolo IEEE 802.11, operando no modo *PCF* - (*Point Coordination Function*).

Em [18] é apresentado um tutorial sobre o protocolo IEEE 802.11, incluindo uma explicação detalhada sobre os modos de operação *DCF* e *PCF*, que são definidos na camada MAC do protocolo. Segundo [18], o modo *PCF* opera em regime livre de contenção, voltando-se principalmente para tráfego em tempo real e compartilhando recursos com a operação no modo *DCF* - (*Distributed Coordination Function*).

Na referência [2] é feita uma análise comparativa de desempenho do protocolo IEEE 802.11, nos modos *PCF* e *DCF*. A análise relaciona o tempo médio de espera em fila em função do parâmetro *goodput*[3]. O gráfico mostrado na Figura 2.3, comprova a superioridade do modo *PCF*, quando a taxa útil (*goodput*) ultrapassa 60%.

2.1.1 Regras de Atendimento e Disciplinas de Serviço

Numa abordagem modelística geral, um sistema de *polling* pode variar com relação à ordem na qual as filas são visitadas (Regra de Atendimento) e no número de

[2]Considerando o quesito de tempo de espera em fila.

[3]O termo *goodput* é definido na referência [2] como a vazão de dados "úteis"na rede, do ponto de vista da aplicação.

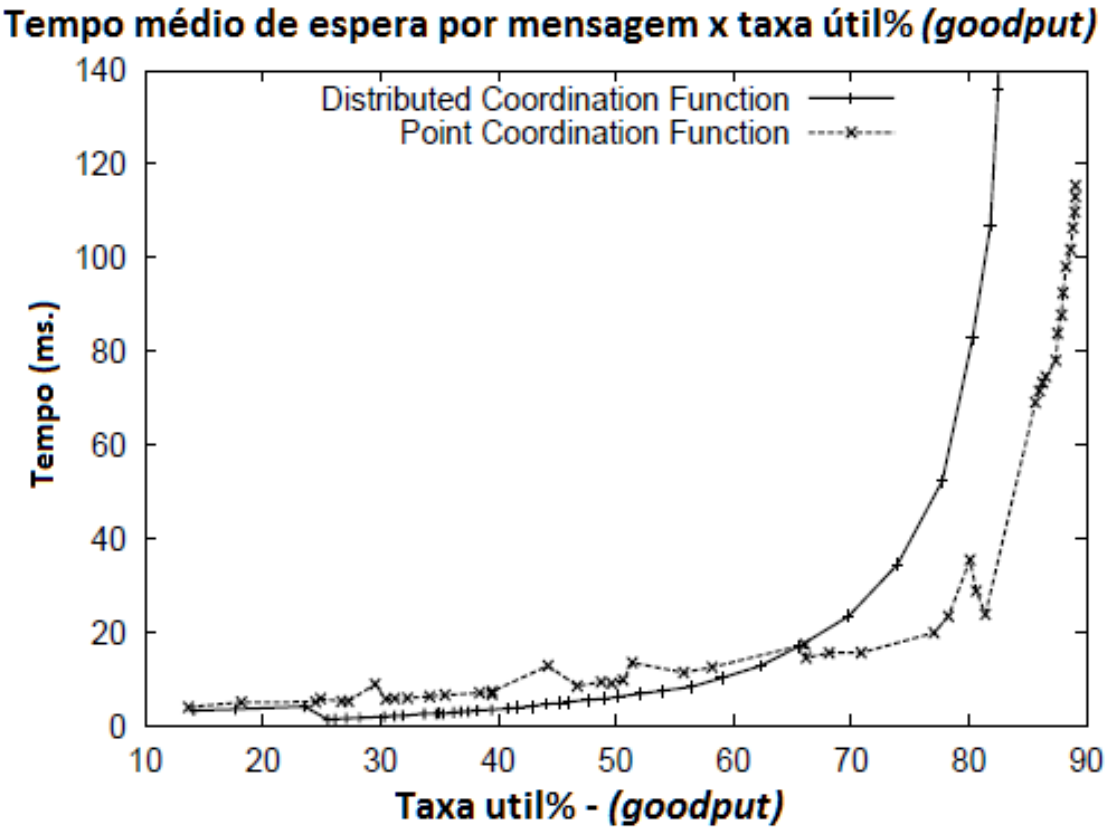

Figura 2.3: Comparação do desempenho dos modos *PCF* e *DCF* em função da taxa útil percentual (*Goodput*) em 2Mbps, extraída de [2].

usuários, que podem ser atendidos durante uma visita do servidor (Disciplina de Serviço).

Estas variações, junto com a sua própria estrutura, definem a estratégia de compartilhamento dos recursos do sistema, e portanto têm grande importância no desempenho do mesmo.

A **regra de atendimento** descreve a ordem na qual o servidor visita as filas.

- Ordem cíclica: o servidor visita as filas obedecendo uma ordem pré-determinada (ciclo), retornando à primeira fila assim que finalizar a visita à última fila do ciclo. Neste tipo de sistema, cada fila receberá uma nova visita, depois de completado o ciclo.

- Elevador: após servir a última fila n_N, o servidor volta atendendo o restante das filas na ordem inversa $(..., n_1, n_2, ..., n_N, n_{N-1}, n_{N-2}, ...)$.

- Ordem randômica: a próxima fila a ser visitada é escolhido aleatoriamente pelo servidor. Este modelo pode ser aplicado na análise de sistemas de controle

distribuído, onde a regra de atendimento é resultado de algum processo de escolha aleatória e sem memória (vide [37]).

- Ordem de prioridade: quando a forma de atender à uma determinada fila varia de acordo com a prioridade da desta mesma fila frente as outras[4]. Esta regra de atendimento pode ser implementada de duas maneiras diferentes nos sistemas de *polling*

  - Dependendo da prioridade de uma determinada fila no sistema, o número de visitas do servidor à fila num mesmo ciclo, pode ser maior que nas demais filas de mais baixa prioridade[5].
  - O tempo que o servidor fica alocado para atender a uma determinada fila de alta prioridade pode ser maior que o tempo para atender filas de baixa prioridade.

A **disciplina de serviço** descreve o número de usuários que serão atendidos em cada fila. Existem muitas disciplinas de serviço híbridas ou de maior complexidade, mas estas são as mais básicas:

- Exaustiva: o servidor permanece servindo enquanto houver usuários na fila. Considerando o tempo médio de espera em fila, o desempenho desta disciplina de serviço supera o desempenho de todas as outras.

- *Gated* (represada)[6] apenas os usuários que estiverem na fila antes da chegada do servidor serão atendidos. É como se uma "porteira"se fechasse atrás do último usuário selecionado, assim que o servidor chega na mesma. Somente os usuários do lado de dentro da porteira serão servidos. Os que ficaram do lado de fora terão que aguardar que o servidor percorra toda a sequência de atendimento às demais filas e retorne à fila original, cumprindo seu ciclo de serviço, para serem atendidos.

- *Limitado* (ou não-exaustiva): ao chegar na fila, o servidor está limitado a atender um número máximo de usuários, independente do número de usuários que encontrar aguardando. Esta disciplina pode funcionar consorciada com as duas primeiras (exaustiva e *gated*) e geralmente é utilizada para melhorar o nível de equidade (*fairness*) da rede, distribuindo de forma mais justa o tempo de alocação do servidor entre os terminais da rede.

[4]Esta prioridade pode ser pré-definida ou pode ser uma função de algum outro parâmetro do sistema. Exemplo: o número de mensagens de alta prioridade aguardando na fila.

[5]Vide Seção 3.2.7.

[6]Em razão da dificuldade de tradução exata deste termo para o português e da sua intensa utilização neste trabalho, utilizaremos o termo em inglês.

  - Exaustiva-*Limitado*: o servidor irá servir um número L máximo de usuários, mas se o número de usuários aguardando na fila for menor que L, o servidor continuará servindo até completar o serviço dos L usuários ou esvaziar a fila, o que acontecer primeiro.
  - *gated/L-limitado*: o servidor irá servir um número L máximo de usuários que ficaram do lado de dentro da porteira. Se este número de usuários do lado de dentro for menor que L, o servidor servirá apenas todos os L usuários.

- *L-Decrementado*: o servidor continuará servindo a fila do terminal até restarem apenas L usuários aguardando, ou a fila se esvaziar, o que acontecer primeiro.

- *T-Limitado*: o servidor continuará atendendo quantos usuários puder, durante um intervalo de tempo pré-determinado T, ou até a fila se esvaziar.

- *Gated-global*: as disciplinas anteriores podem ser consideradas disciplinas "locais", uma vez que a decisão do número de usuários a serem servidos numa determinada fila n é tomada a partir das informações locais da mesma fila. A disciplina *gated-global* usa informações mais amplas, acerca das demais filas, para decidir quantos usuários serão servidos em cada visita. Por exemplo: durante a visita à fila n, o servidor irá atender somente os usuários que já estavam dentro da porteira desta fila n, no momento em o servidor visitou a fila 1. Esta disciplina considera implicitamente que o servidor possui uma visão ampla sobre todas as filas do sistema.

Um bom exemplo de utilização de diferentes disciplinas de serviço é o modelo híbrido, proposto em [53]. A proposta considera mensagens de dois diferentes níveis de prioridade, que são separadas em filas distintas, dentro de um mesmo terminal. A fila de alta prioridade é servida de forma exaustiva e a fila de baixa prioridade é servida de acordo com a disciplina *gated/L-limitado*.

2.2 Diferenciação por Prioridades

A diferenciação de tratamento por prioridade é aplicada em muitas situações práticas e comuns no dia a dia, exemplo: atendimento prioritário à idosos nas filas de bancos, caixas de supermercados exclusivos para clientes com poucas compras, etc. Entretanto, apesar da diversidade de situações práticas, o objetivo essencial permanece o mesmo: otimizar a utilização de um determinado recurso compartilhado pelos clientes, de acordo com as suas necessidades específicas.

Segundo [34], a escolha do próximo usuário a ser atendido num sistema de

filas é determinada pela disciplina de fila utilizada no modelo. Esta escolha pode ser baseada em:

- A ordem de chegada do usuário nesta fila. Por exemplo: na fila da padaria, quem chega primeiro é servido antes (*FCFS - First Come First Served*). Esta é talvez a disciplina mais conhecida no nosso dia a dia, mas existem muitas outras.

- A quantidade de tempo necessária para atender um usuário. Um bom exemplo deste método de escolha é a disciplina conhecida como (*SJF - Shortest Job First*) que é bastante comum nos supermercados. É representada pelo caixa rápido, que atende numa fila exclusiva os clientes (usuários) que possuem um número máximo de volumes, ou seja, que vão demorar menos tempo de serviço com o caixa.

- Uma função qualquer, relacionada ao grupo que o usuário pertence. Um exemplo desta estratégia é o caixa destinado à pessoas idosas nos bancos, ou a fila de clientes *VIP* nos aeroportos.

Este terceiro método endereça os estudos de diferenciação por prioridades, com base no tipo de usuário que está aguardando na fila.

2.2.1 Conservação de Trabalho

Segundo [34], num sistema conservativo, o privilégio dado a um determinado usuário, deve se dar às custas do prejuízo causado aos outros. Ou seja, considerando um sistema de filas, a redução do tempo médio de espera em fila de uma determinada classe p de usuários, causa um aumento no tempo médio de espera em fila dos usuários de prioridades mais baixas.

Ainda segundo [34], sistema de filas é dito como conservativo dentro de algumas condições:

- O servidor nunca fica "ocioso"enquanto existir ao menos um usuário para ser servido em qualquer fila.

- Um usuário nunca sai do sistema sem que seja atendido pelo servidor.

As Leis de Conservação são obtidas a partir de um processo estocástico $U(t)$, que representa a quantidade de trabalho remanescente no sistema, medido num determinado instante t; ou o tempo necessário para o sistema ser esvaziado, a partir de t, se não ocorrerem mais chegadas. Como $U(t)$ não se altera, independente da ordem de serviço, as leis de conservação permanecem válidas e úteis para analisar sistemas com disciplinas de fila mais complexas, pois permitem que alguns cálculos

sejam baseados em resultados existentes, de disciplinas mais comuns, como a *FCFS - First Come First Served*, que será melhor detalhada na Seção 2.2.3.

De acordo com [36], considerando um sistema de filas conservativo sem interrupção (*nonpreemptive*), a disciplina de fila também será conservativa se:

- Não alterar a ordem de chegada de usuários na fila.
- Não alterar a quantidade de trabalho (tempo) necessário para atender um usuário.

Considerando as condições de conservação de trabalho definidas, pode-se concluir então que um sistema de *polling* com tempo de comutação entre as filas positivo ($V > 0$) é não conservativo, uma vez que há a chance do servidor ficar ocioso, mesmo quando existirem usuários aguardando nas filas para serem atendidos.

Para analisar os sistemas não-conservativos, foram propostas algumas extensões para as leis de conservação, conhecidas como leis pseudo-conservativas. Uma generalização da lei de conservação *M/G/1* é apresentada em [50]. Esta generalização parte da decomposição da quantidade remanescente de trabalho em t $[U(t)]$ na soma de duas variáveis aleatórias independentes:

$$U(t) = T + I$$

Onde T é o montante de trabalho dentro de um sistema de filas conservativo $M/G/1$, I é o montante de trabalho presente no intervalo ocioso do sistema original.

2.2.2 Equidade

Um outro fator que geralmente causa até algumas discussões no dia a dia das pessoas é a equidade (ou *Fairness*). Isto é, se todos os clientes na fila pertencerem a um mesmo grupo, por exemplo: grupo de pessoas com menos de 60 anos de idade, espera-se que todos sejam atendidos de forma igual, justa e sem qualquer privilegio a qualquer um do mesmo grupo.

Numa rede de comunicação sem diferenciação por prioridades, o conceito de equidade pode ser definido como uma medida da justiça com a qual são compartilhados os recursos da rede entre os terminais e/ou entre as mensagens armazenadas nas filas dos terminais, aguardando para serem transmitidas.

O trabalho em [51] considera uma estrutura de filas para propor uma medida do "índice de discriminação"dos usuários. Esta outra abordagem é justamente

o oposto do conceito de equidade, explicado no parágrafo anterior. Ou seja, o "índice de discriminação"mede então o quão mais rápido os usuários de maior prioridade são atendidos, em relação aos outros, de menor prioridade.

O nível de equidade de uma rede tem então um significado relativo, dependendo da sua aplicação. Por exemplo, se considerarmos um protocolo onde se pretende compartilhar de forma mais justa os recursos de rede entre os terminais, quanto maior o nível de equidade, melhor o sistema neste quesito. Se considerarmos no entanto, um sistema com diferenciação por prioridades, pode ser desejável que haja uma maior discriminação entre as mensagens de diferentes classes de prioridades.

2.2.3 Disciplinas de Fila

A **disciplina de fila** define a ordem de serviço dos usuários que estão aguardando na fila para serem atendidos, durante a visita do servidor.

Apesar de existirem muitas variações, as principais disciplinas de fila são:

- *FCFS - First Come First Served* (o primeiro que chega é o primeiro a ser servido): é conhecida como política padrão. Será servido primeiro o usuário que estiver aguardando na fila a mais tempo. Esta política tem uma desvantagem importante, quando a variância do tempo de serviço é grande. Isto acontece por que usuários que podem ser atendidos rapidamente deverão aguardar o serviço dos usuários que irão demorar muito, mas que chegaram antes.

- *LCFS - Last Come First Served* (o último a chegar é servido primeiro): será servido primeiro o usuário que chega na fila mais recentemente. É exatamente o inverso da disciplina anterior (*FCFS*), porém apresenta a mesma desvantagem.

- *SJF - Shortest Job First* (o que demora menos tempo para ser atendido é servido primeiro): do ponto de vista do tempo médio de espera em fila, a *SJF* é a melhor política possível. Apesar de apresentar os melhores resultados, esta política requer processamento rápido e constante para aproximar o tempo de serviço de cada usuário e reordenar na fila em tempo de processamento.

- *HOL - Head of Line*: nesta disciplina é servido primeiro o usuário de maior **prioridade**. Os usuários na fila são primeiramente classificados de acordo com a sua classe de prioridade. As classes de prioridades mais altas são atendidas primeiro. Dentro de uma mesma classe de prioridade, os usuários são atendidos de acordo com a disciplina *FCFS*. Apesar de eficiente, esta disciplina pode trazer um problema conhecido como "inanição"(*starvation*). Ou seja, se a

taxa de entrada de usuários de alta prioridade for muito grande em relação às taxas de prioridades mais baixas, os usuários de prioridades mais baixas podem demorar excessivamente para serem servidos.

O comportamento do sistema ainda pode variar em relação à ação, em função da chegada de um usuário de alta prioridade, no mesmo instante que um usuário de baixa prioridade está em atendimento. Tomando como exemplo a disciplina de serviço *HOL*, se um pacote de alta prioridade chega na fila e encontra um outro usuário de mais baixa prioridade em serviço, duas ações diferentes podem ser implementadas, dependendo da **regra de interrupção de serviço** em vigor:

- Sem-interrupção (*nonpreemptive*) o serviço é concluído sem interrupção, mesmo se chegar um outro usuário de mais alta prioridade na fila.
- Com-interrupção (*preemptive*) o serviço em curso é interrompido para dar lugar a qualquer eventual usuário de mais alta prioridade que chega na fila.

2.2.4 Estratégias de Priorização em Sistemas de Comunicação

Numa rede de computadores que desejam trocar mensagens entre si, utilizando um determinado protocolo de controle de acesso ao meio, a diferenciação por prioridade significa fazer com que as mensagens com prioridade mais altas, de aplicações que exigem um determinado nível de serviço (exemplo: tráfego de voz sobre *IP*), sejam transmitidas mais rapidamente que as mensagens de aplicações menos exigentes (exemplo: tráfego de *email*). De acordo com [35], existem basicamente três diferentes estratégias para se implementar diferenciação por prioridades num sistema de comunicação:

- Priorização local por tráfego: quando o terminal observa a prioridade de cada mensagem que recebe em sua fila e ordena as mensagens para transmissão de acordo com a classe de prioridade de cada mensagem.
- Priorização por terminal: quando os terminais de uma rede possuem prioridades diferentes para acessar a rede para transmitir suas mensagens, independentemente da prioridade destas mensagens.
- Priorização por tráfego e por terminal: quando a rede consegue observar de forma global a prioridade das mensagens que são recebidas nas filas de cada terminal e controla o acesso de cada terminal, de acordo com algum padrão, em função das mensagens ordenadas na fila de cada um deles.

Ainda segundo [35], as mais importantes propriedades esperadas para o funcionamento de um sistema de comunicação com diferenciação por prioridades são:

- Independência hierárquica de desempenho: significa que o tráfego produzido por mensagens de mais alta prioridade não deve ser afetado pelo tráfego produzido por mensagens de prioridades mais baixas.
- Igualdade (*fairness)* dentro de uma mesma classe de prioridade: as mensagens que pertencem a uma mesma classe de prioridade devem ser processadas de forma justa, dentro das mesmas condições de atendimento.
- Descriminação entre classes de prioridades diferentes. É exatamente o contrário do que se espera no item anterior. Considerando o tempo médio de espera em fila como parâmetro de comparação, deseja-se patamares diferentes e bem definidos entre classes de prioridades diferentes.

Dependendo do tipo de rede e do protocolo de controle de acesso utilizado, as formas para se implementar diferenciação por prioridades podem variar bastante, com relação à sua eficiência e ao grau de complexidade envolvida.

Em se tratando de redes sem fio, fazer diferenciação por prioridades pressupõe desafios ainda maiores aos protocolos de controle de acesso, advindos principalmente das próprias características destas redes, comentadas na Seção 1.1.1.

Na referência [52] é feita uma análise completa sobre os diferentes aspectos relacionados ao desempenho e à Qualidade de Serviços (QoS) de aplicações de rede em ambiente sem fio, utilizando o protocolo IEEE 802.11e. Neste padrão, a implementação de diferenciação por prioridades é feita em função da variação controlada dos parâmetros $AIFS[AC]$, $CW_{min}[AC]$ e $CW_{max}[AC]$. Em linhas gerais, estes parâmetros regulam competição pelo acesso ao meio único, mas também estão relacionados com a probabilidade de colisão. Vide Figura 2.4

2.2.5 Sistemas de *Polling* com Prioridades

Uma das mais importantes características do sistema de *polling* é a sua flexibilidade para suportar funcionalidades como: operar com múltiplos canais e/ou prover diferenciação por prioridades.

Além da flexibilidade, por ser livre de contenção, o protocolo de *polling* consegue níveis de discriminação (vide Seção 2.2.2) maiores que protocolos com contenção, o que é justamente uma das principais propriedades esperadas no funcionamento do sistema de prioridades, listadas na seção anterior.

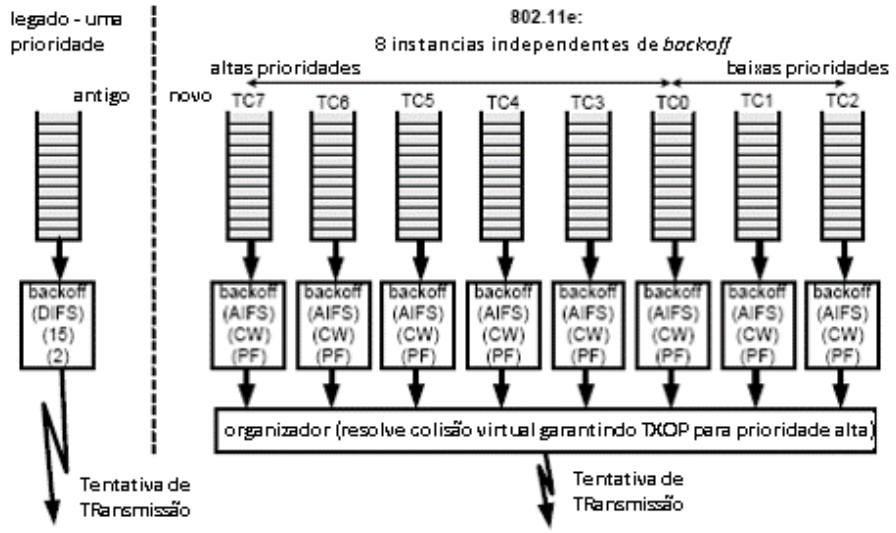

Figura 2.4: *Backoff* virtual com 8 diferentes categorias de tráfego, extraída de [52].

Estas características, entre outras, fazem dos sistemas *polling* uma das principais opções de protocolo, para sistemas de comunicação com diferenciação por prioridades.

De acordo com [35], as três diferentes estratégias de diferenciação por prioridades (vide 2.2.4) podem ser implementadas em sistemas de *polling* da seguinte forma:

1. Priorização por terminal.
    - Alterar a seqüência de atendimento do servidor, aumentando a freqüência das visitas nas filas dos terminais de maior prioridade.
2. Priorização por tipo de tráfego.
    - Alterar a ordem de serviço das mensagens dentro das filas, transmitindo primeiro aquelas de mais alta prioridade.
3. Priorização por terminal e por tráfego.
    - Combinar as duas formas anteriores de modo a aumentar o nível de diferenciação entre as diferentes prioridades.

Além das já mencionadas facilidades funcionais oferecidas pelos sistemas de *polling* para prover diferenciação por prioridades nas diferentes estratégias, a parte de modelagem destes sistemas é bastante ampla e abrangente.

Em [36] é apresentada uma extensa análise sobre vários modelos envolvendo diferenciação por prioridades, considerando sistemas operando em regime estacionário.

A análise de [36] compreende desde os modelos mais básicos do tipo *M/G/1* sem-interrupção (*nonpreemptive*) até modelos para sistemas de reserva, citado no início do Capítulo 2.

Para cada um dos modelos analisados é extraída uma expressão do tempo médio de espera em fila (W_p) para uma determinada prioridade p.

2.3 Múltiplos Canais

A conectividade é um dos vetores de evolução dos sistemas de informação. O surgimento de novas aplicações, cada vez mais integradas e distribuídas, faz surgirem redes de comunicação mais avançadas e com maior poder de conectividade, principalmente nos ambientes sem fio.

Dois exemplos práticos são as redes *Bluetooth*, aplicadas às comunicações pessoais e de curta distância (*PAN - Personal Area Network*) e as redes *WiFi*, voltadas para redes locais (*LAN - Local Area Network*).

Estas novas redes, além da evolução tecnológica implícita dos seus componentes físicos, requerem novos protocolos, que sejam capazes de prover respostas rápidas, grande capacidade de vazão agregada e lidar com diferentes tipos de tráfego (prioridade).

A utilização de múltiplos canais é uma das várias abordagens sobre o problema da interferência mútua, para aumentar a vazão agregada de uma destas redes. Para suportar estas novas necessidades, o protocolo de controle de acesso ao meio à base de *polling* é sempre considerado uma opção interessantes como base no desenvolvimento de protocolos avançados.

O artigo de [39] estabelece uma relação inversa da vazão agregada da rede em função do número de terminais na mesma rede. Ou seja, a vazão agregada da rede tende a zero, quando o número de terminais é muito grande. Nesta mesma linha, o trabalho de [40] analisa os efeitos da interferência (*SINR - Signal Interference plus Noise Ratio*) sobre a vazão agregada de uma rede sem fio qualquer.

A partir dos resultados das referências [39] e [40] é possível concluir que a densidade de terminais por si só já representa um sério desafio a ser enfrentado na busca de eficiência máxima das redes sem fio.

A análise dos efeitos da densidade de terminais nas redes sem fio é abordada na literatura de diversas formas diferentes, sendo que as principais são:

- Antenas Direcionais: diferentemente do espectro de energia das antenas omnidirecionais, que se espalha de uniformemente no espaço, as antenas direcionais conseguem concentrar a energia, na direção de um determinado ponto no

espaço. O direcionamento minimiza o espalhamento do espectro para outros terminais vizinhos, fora do cenário da transmissão, minimizando a interferência e melhorando a possibilidade de reuso da frequência. O trabalho de [41] analisa a capacidade de vazão por terminal fim a fim de uma rede de sensores com antenas direcionais. A análise comparativa é feita em relação à mesma rede com antenas omnidirecionais. Os resultados principais mostram que a utilização de antenas direcionais aumenta notavelmente a vazão por terminal.

- Diversidade de Antenas: neste método, o aumento da capacidade da rede é obtido aumentando-se o número de antenas receptoras. Isto é feito para explorar a diversidade de caminhos, que fazem com que os sinais transmitidos cheguem no terminal receptor em instantes/fases diferentes. No trabalho de [42] é demonstrado que para N terminais na rede que se interferem mutuamente, $K + N$ antenas conseguem anular $N - 1$ interferências, enquanto que $K + 1$ sinais são melhorados.

- Múltiplos Canais: neste último exemplo, que é o tema principal desta seção, o aumento da capacidade de vazão da rede é obtido através da utilização de diferentes canais ortogonais, possibilitando comunicações simultâneas não interferentes e potencializando o reuso espacial de frequências.

2.3.1 Protocolos de Controle de Acesso com Múltiplos Canais

A utilização de múltiplos canais para aumentar a capacidade de vazão de uma rede se aplica tanto para redes sem fio quanto para redes com fio. Considerando o ambiente com fio, o trabalho apresentado em [26] avalia o desempenho de algumas variantes do protocolo CSMA, nos dois cenários: com um único canal e utilizando múltiplos canais. A análise comparativa é feita através de modelos matemáticos adaptados para múltiplos canais. Os resultados mostram um notável aumento de desempenho, em comparação com os resultados com um único canal.

Na referência [27] é avaliado o comportamento do protocolo CSMA/CD, comparando os cenários com múltiplos canais em relação a um único canal. A medida de interesse avaliada no trabalho é o tempo médio de espera na fila. Os resultados mostram uma notável redução, tanto no valor médio do tempo de espera quanto na sua variância, em favor dos modelos com múltiplos canais.

No ambiente sem fio, a utilização de múltiplos canais ataca principalmente o tema da interferência mútua entre os terminais da rede, transmitindo ao mesmo tempo. Este problema, bem como seus principais impactos é analisado no trabalho

[43]. Esta referência se baseia em resultados numéricos, obtidos através de simulações, para mostrar com precisão como e quanto a interferência reduz a vazão de uma rede. O trabalho inclusive propõe um método (*Rapid Channel Hopping*) para tornar a rede mais tolerante às interferências, às custas de um pequeno aumento no *overhead* da rede.

Considerando o tema do espectro de frequencias disponíveis para a utilização de múltiplos canais, percebe-se que as oportunidades existem e são palpáveis. Um bom exemplo é o próprio padrão IEEE 802.11 b/g, sabe-se que a banda total disponível de 72 MHz é dividida em 11 canais de 22 MHz de banda cada. Todos estes canais são sobrepostos, exceto os canais 1, 6 e 11 que são ortogonais entre si (Figura 2.5). Isto significa que, três comunicações simultâneas não interferentes, ao invés de somente uma, poderiam acontecer na rede, se o protocolo IEEE 802.11 b/g pudesse utilizar os três canais, ao invés de somente um.

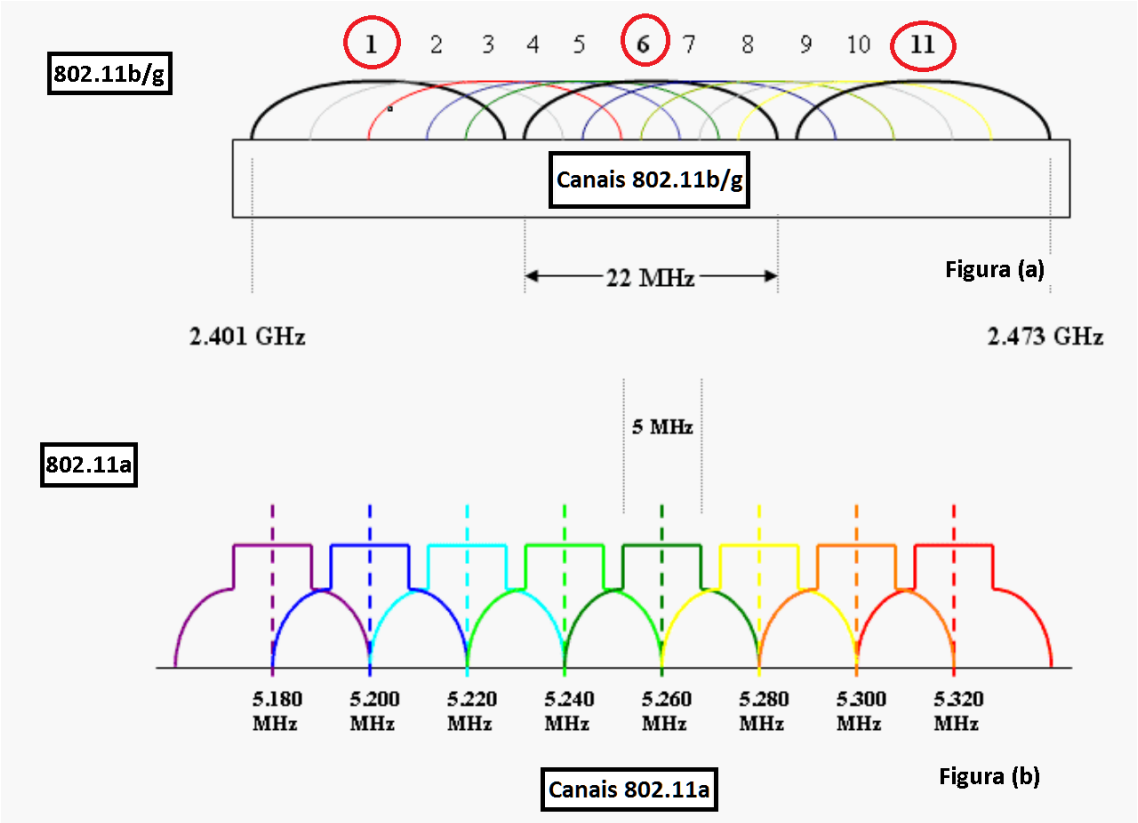

Figura 2.5: Configuração de canais IEEE 802.11 a/b/g.

O artigo [44] analisa o aumento de capacidade de uma rede IEEE 802.11 com múltiplos canais, onde o número de canais é menor que o número de interfaces. Nesta análise é obtida uma relação ótima entre o número de canais e o número de interfaces para conseguir a máxima capacidade de vazão na rede. O trabalho também propõe um protocolo de roteamento que seleciona rotas com alta vazão, aproveitando o ambiente de múltiplos canais e múltiplas interfaces.

Com base nestes bons resultados esperados, muitas propostas de novos protocolos tem surgido, inclusive algumas adaptações do padrão IEEE 802.11 para operar em múltiplos canais. Um bom exemplo nesta linha é a proposta de [24], que divide o protocolo em duas fases: uma primeira fase de negociação, operando em regime de contenção; e uma segunda fase de transmissão que opera sem contenção, sobre um canal de serviço exclusivo, escolhido na primeira fase.

Apesar da grande quantidade de propostas e das muitas diferentes abordagens sobre o tema da utilização de múltiplos canais, os desafios enfrentados nos cenários analisados são essencialmente os mesmos. Na referência [28], estes desafios são consolidados e classificados da seguinte forma:

- Canal de Controle Dedicado: nesta abordagem a utilização dos canais de serviço é negociada num único canal, que é comum a todos os terminais da rede e exclusivo para controle. A principal vantagem deste sistema é que ele dispensa sincronização. Entretanto, o canal de controle já representa perda de eficiência espectral e pode se tornar um ponto de "gargalo", dependendo do número de terminais na rede.

- Saltos de Canal: as principais vantagens deste modelo são: a necessidade de apenas um rádio *half-duplex* por terminal e a possibilidade de utilizar todos os canais simultaneamente, eliminando os "gargalos"do modelo anterior. Os terminais ficam continuamente "saltando"de canal em canal, obedecendo um determinado padrão cíclico. Quando dois terminais entram em acordo para transmitir, ambos param de "saltar"até que a transmissão seja finalizada, quando retornam novamente ao padrão anterior. As principais desvantagens são a necessidade de sincronização de toda a rede e os atrasos decorrentes das constantes comutações dos rádios.

- Divisão de Tempo: nesta proposta existe um canal comum a todos os terminais, que alterna períodos de controle e serviço. Quando não estão em serviço, os terminais se sintonizam no canal comum e negociam suas transmissões no período de controle. Depois da negociação finalizada, sintonizam no canal negociado e permanecem até o final da transmissão. As principais vantagens desta proposta são a possibilidade de operar com apenas um rádio *half-duplex* por terminal e a utilização integral de todos os canais. As desvantagens são a necessidade de sincronização e a mesma possibilidade de "gargalo"no canal comum da primeira proposta.

- Múltiplos Rádios: neste sistema, cada terminal tem 3 ou mais rádios. Há a

possibilidade de transmitir/receber dados em todos os canais simultaneamente, aumentando bastante a vazão da rede. A principal desvantagem desta proposta é o *hardware*, que é mais complexo e caro; e o consumo de energia, que pode ser muito elevado por terminal.

Um dos principais desafios que surge com a utilização de múltiplos canais é a perda da transmissão por "destino ocupado". Este problema se caracteriza quando um terminal n_1 recebe o direito de transmitir para um outro terminal n_2, que por sua vez, já está ocupado transmitindo para um terceiro terminal n_3. Ou seja, n_1 tenta transmitir para o destino n_2, que está ocupado. Nesta condição, a perda de tempo sofrida pelo terminal n_1 é bastante significativa, uma vez que perderá a sua oportunidade de transmitir naquele ciclo. Este problema é endereçado na referência [1], que propõe um mecanismo de adiamento para minimizar o atraso sofrido pelo terminal n_1. O protocolo proposto em [1] funciona para uma rede estruturada, cujo controle de acesso ao meio é feito através de um sistema de *polling* cíclico, controlado pelo PA - (Ponto de Acesso). O "adiamento"é feito com a introdução de uma nova fila no PA, que é preenchida com a identificação dos terminais que não puderam transmitir por causa de destino ocupado. Sempre que uma nova transmissão for negociada com sucesso, a ordem normal do ciclo de *polling* é interrompida e interceptada pela fila de adiamento, minimizando o tempo de espera dos terminais nesta fila. Os resultados obtidos através de modelagem analítica e de simulações, comprovam uma notável redução no tempo médio de espera em fila.

Uma oportunidade importante que pode ser explorada nos protocolos de controle de acesso com múltiplos canais ortogonais é poder escolher o canal, que irá servir o terminal da "vez". Um bom algoritmo de escolha do canal é determinante para aumentar ainda mais o desempenho da rede (vazão agregada), principalmente quando o cenário em questão permite reuso de canais. Neste contexto, o trabalho de [45] contém uma análise sobre os algoritmos mais comuns de escolha dinâmica de canais. O trabalho também apresenta uma proposta de algoritmo de escolha, que reduz a probabilidade de ocorrer interferência co-canal e aumenta a vazão agregada da rede. Os resultados numéricos mostram o ganho em relação aos outros mecanismos comparados.

2.3.2 *Polling* com Múltiplos Canais

A utilização de múltiplos canais em sistemas de comunicação com controle de acesso de *polling* tem como principais objetivos reduzir o tempo médio de espera em fila das mensagens e aumentar a vazão agregada da rede. Estes objetivos podem ser atingidos, através da utilização simultânea de canais ortogonais ou não interferen-

tes entre si, dentro de uma mesma rede. Ou seja, até K diferentes comunicações simultâneas podem coexistir sem interferência mútua, numa mesma rede, desde que utilizem canais ortogonais diferentes.

Apesar das grandes possibilidades de aplicações práticas, estudos de sistemas de *polling* com múltiplos canais não são tão antigos, nem tão vastos quanto os sistemas que compartilham um único canal. Segundo [7], a explicação para este fato é baseada na complexidade da modelagem destes sistemas e nas hipóteses adicionais, adotadas para viabilizar os cálculos.

Nesta mesma referência [7] são propostos modelos aproximados para o cálculo do tempo médio de espera em fila, considerando múltiplos servidores. Nos modelos apresentados, as filas podem ser atendidas de duas formas diferentes, com relação ao número de servidores envolvidos:

- Uma fila só pode ser atendida por apenas um único servidor por ciclo[7].
- Uma fila pode ser atendida cooperativamente por mais de um servidor no mesmo ciclo.

Levando-se em conta o controle de apresentação do *pool* de servidores para o atendimento às filas, os modelos com múltiplos servidores podem variar da seguinte forma:

- Se existirem K servidores no sistema, existirão K processos exclusivos de *polling*, independentes entre si, um para cada servidor. O trabalho em [48] analisa este modelo, considerando que os processos de *polling* "visitam"as filas de forma cíclica.

  Na referência [48] são obtidas aproximações para o tempo médio de espera em fila, considerando um sistema assimétrico (ou desbalanceado). Isto é, as N filas podem receber diferentes taxas médias de chegada de usuários e podem ser atendidas com diferentes taxas médias de serviço (μ). Os resultados obtidos das aproximações são comparados com resultados obtidos numericamente. A comparação mostra que a diferença aumenta, à medida que o sistema fica mais carregado e/ou mais assimétrico.

- Todos os K servidores **juntos** são controlados por um único processo de *polling*, que "visita"as filas de acordo com uma ordem qualquer (ex: cíclica ou randômica).

  Este modelo pode representar uma característica interessante nos sistemas de *polling* com múltiplos servidores, que foi citada em [49]: existe a tendência dos

[7]Na referência [7], um ciclo é definido como o intervalo de tempo entre duas visitas consecutivas a uma mesma fila do sistema

servidores "caminharem"juntos (*cluster*), se a ordem de serviço utilizada por todos os servidores for a mesma, especialmente em situações de alto tráfego. Este fenômeno ocorre por que os últimos servidores tendem a passar mais rapidamente pelas filas já atendidas, enquanto que os primeiros servidores, tendem a passar mais devagar, à medida que encontram as filas ainda por atender.

2.4 Conclusão

O Capítulo 2 teve como objetivo aprofundar o conhecimento do leitor a respeito dos principais protocolos de controle de acesso, classificando-os de acordo com o método utilizado e a estratégia de alocação. Nesta análise, enfatizou-se a ampla capacidade do protocolo de *polling* para suportar funcionalidades, principalmente: operação com múltiplos canais e diferenciação por prioridades.

Na diferenciação por prioridades, foram apresentadas as principais estratégias de priorização, envolvendo um suporte teórico sobre equidade e conservação de energia, disciplinas e regras de serviço.

Na utilização de múltiplos canais, foram relacionadas e referenciadas as principais metodologias utilizadas pelos protocolos, de acordo com os problemas e desafios enfrentados em cada ambiente, principalmente nas redes sem fio.

Capítulo 3

Modelagem

De acordo com a Seção 1.2, a modelagem de um determinado sistema tem por objetivo reproduzir com a maior acurácia possível as características mais importantes do sistema, de forma que seja possível uma análise e avaliação objetiva do comportamento do mesmo, antes da sua implementação propriamente. A definição das características que serão analisadas, bem como o nível de detalhe que se pretende chegar sobre cada uma destas, indicam a complexidade do modelo a ser utilizado.

Segundo ANTHONY e WATSON [29] o processo de desenvolvimento de qualquer modelo passa por duas etapas distintas:

- Caracterização do Sistema

  Caracterizar um determinado sistema significa levantar as informações essenciais, que serão exploradas na definição do modelo. Para se caracterizar um sistema, é necessário cumprir quatro tarefas:

  - Definir a composição do Sistema: relacionar os atores e os recursos do sistema.
  - Identificar e enfatizar as suas principais características funcionais: as características funcionais de um sistema definem o comportamento do mesmo e a sua resposta aos estímulos de entrada.
  - Definir as principais medidas de interesse do sistema: aquela medida que está relacionada com o objetivo do seu projeto.

- Definição do Modelo

  - Depois de cumpridas as três etapas iniciais de caracterização, o sistema precisa ser comparado com os diversos tipos de modelos conhecidos, considerando a acurácia das informações necessárias ao projeto do sistema e a viabilidade na construção do modelo.

– Esta viabilidade está diretamente relacionada com as hipóteses adotadas para tornar o modelo tratável matematicamente, sem no entanto ferir significativamente a sua integridade, prejudicando a aderência do modelo ao sistema original.

O resultado desta comparação indicará o(s) tipo(s) de modelo(s) a ser(em) utilizado(s) e as adaptações a serem executadas no(s) mesmo(s), considerando as hipóteses de viabilização que virão a serem assumidas.

3.1 Caracterização do Sistema - Protocolo de Controle de Acesso ao Meio

O sistema considerado neste trabalho é um protocolo de controle de acesso ao meio, que coordena o compartilhamento de K canais, sem contenção, pelos terminais da rede. Este controle é exercido através de um sistema de *polling*, que percorre todos os terminais, oferecendo a cada um deles uma oportunidade de transmitir suas mensagens armazenadas, através de um único canal, escolhido num conjunto dos K canais disponíveis.

Como explicado na Seção 1.3, a utilização deste protocolo no trabalho foi inspirada na necessidade de se criar um modelo simples, mas que também pudesse servir como base na análise de outros protocolos de múltiplos canais mais complexos.

3.1.1 Composição do Sistema

Nesta seção são definidos os "atores"do sistema e a configuração dos mesmos (*setup*), dentro das suas principais características individuais.

A rede é composta por N terminais, que podem trocar mensagens entre si através de uma rede sem fio com múltiplos canais. A Figura 3.1, ajuda a ilustrar a composição do sistema.

Cada terminal é equipado com um rádio *Half-Duplex* sintonizável. Isto é, o rádio consegue sintonizar em diferentes canais, mas não pode transmitir e receber simultaneamente. Além disso, cada terminal possui uma área de armazenamento (*buffer*), onde as mensagens são agrupadas e ficam aguardando até serem transmitidas.

A rede dispõe de $K+1$ canais de comunicação ortogonais, não interferentes entre si. Um dos canais é utilizado para tráfego de mensagens de controle (Cnc na Figura 3.1), enquanto que os outros K canais podem ser utilizados para transmissão das mensagens de dados (Cn1 a CnK na Figura 3.1).

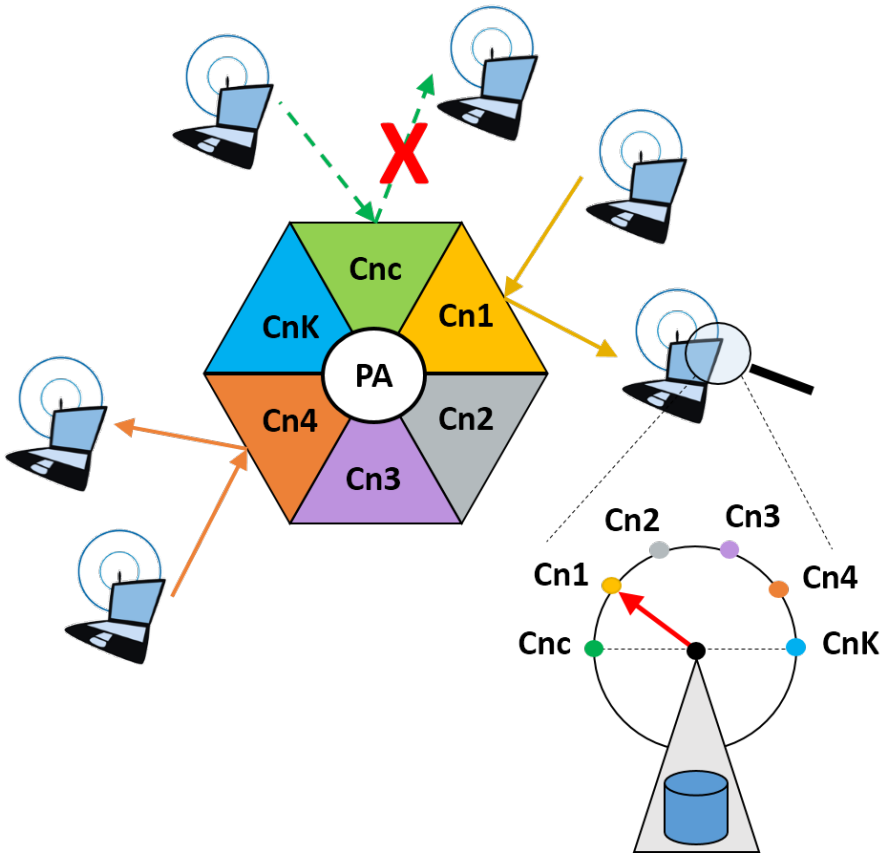

Figura 3.1: Composição do sistema de comunicação considerado.

O rádio dos terminais permanece constantemente sintonizado no canal de controle, exceto quando está transmitindo ou recebendo dados.

Para coordenar a utilização dos canais entre os terminais, a rede dispõe de um protocolo de controle de acesso ao meio do tipo *polling*. A rede pode ser controlada de forma centralizada, através de um ponto de acesso - PA (*AP - Access Point*), ou de forma distribuída, onde cada terminal que acabou de transmitir suas mensagens, conhece antecipadamente o próximo terminal na sequência do *polling*[1].

Se considerarmos que a rede é controlada por um ponto de acesso, este PA é equipado com $K + 1$ rádios e cada rádio fica constantemente sintonizado em cada um dos seus respectivos $K + 1$ canais. Da mesma forma que nos terminais, os rádios do PA também não conseguem transmitir e receber simultaneamente (*Half-Duplex*).

3.1.2 Características Funcionais do Protocolo

Esta seção especifica como os "atores" do sistema de comunicação, definidos na seção anterior (3.1.1), interagem entre si; e os resultados destas interações, dentro das especificações de projeto.

[1] O protocolo do tipo *token* é um exemplo de controle distribuído que pode ser analisado de forma similar.

A comunicação entre os terminais se dá em duas fases diferentes e bem definidas:

- Fase de Controle: fase na qual são trocadas mensagens de controle. Seja para a rede consultar se existe demanda no terminal de origem (*polling*), seja na negociação de um canal de serviços entre os terminais de origem e destino, decorrente da consulta.

  A fase controle acontece exclusivamente no canal de controle único e que é comum a todos os terminais da rede. As trocas de mensagens durante a fase de controle representam basicamente o *overhead* do protocolo de comunicação.

- Fase de Transmissão: pode acontecer logo em seguida à fase de controle, caso exista demanda de transmissão no terminal de origem consultado e caso a negociação de um dos K canais de serviço, entre os terminais de origem e destino, tenha sido bem sucedida.

  Ou seja, é somente na fase de transmissão que acontecem as trocas efetivas de dados, entre os terminais de origem e destino na rede.

Se, durante a fase de controle, os terminais não conseguirem entrar num acordo (por exemplo: se o terminal de destino estiver ocupado recebendo ou transmitindo) sobre o canal de serviço a ser utilizado, o terminal de origem perde a oportunidade de transmissão naquele ciclo e o processo de *polling* segue para o próximo terminal, de acordo com a regra de atendimento da rede.

A escolha do próximo terminal a ser consultado pode ser feita de várias formas. Neste trabalho serão modeladas as regras de atendimento cíclica: onde o próximo terminal a ser consultado faz parte de uma lista pré-configurada; e aleatória: onde o próximo terminal é escolhido aleatoriamente, com probabilidade uniforme[2] $(1/N)$.

A consulta supra citada, que pode ser doravante chamada de *polling*, é enviada para cada um dos terminais escolhidos via canal de controle. O terminal consultado responde positivamente, se houver demanda de transmissão, ou negativamente, se não houver mensagens a serem transmitidas em sua fila; ou não responde, se já estiver sintonizado num canal de serviço, enviando ou transmitindo dados. Neste caso, como já mencionado, o terminal consultado perde a sua oportunidade de transmissão no ciclo e o *polling* segue na sequencia.

Se o terminal responder positivamente, somente um dos K canais de serviços pode ser escolhido e o processo de comunicação entra na Fase de Transmissão. Nesta fase, os terminais de origem e destino sintonizam os seus rádios no canal de serviço negociado e iniciam o processo de comunicação.

[2] N é o número de terminais no sistema de comunicação considerado.

A escolha do canal de serviços pode ser de qualquer forma e depende da estratégia adotada no protocolo. O artigo de SARKAR *et al.* [54] apresenta um algoritmo simples para escolha do melhor canal disponível, considerando um ambiente de rede do tipo *Ad hoc*. O algoritmo se baseia nas informações de *SNIR - Signal Noise plus Interference Ratio*, para decidir o melhor canal a ser utilizado pelo par de terminais de origem e destino.

Como pode ser visto em SARKAR *et al.* [54], a escolha do canal de serviço tem impacto significativo no desempenho do protocolo como um todo.

Depois de escolhido o canal de serviço, o terminal de origem seleciona as mensagens que estão armazenadas na fila, até o momento da chegada do *polling*. Deste conjunto de mensagens que estavam na fila, antes do fechamento da porteira (*Gate*), somente um número limitado de mensagens (L) poderá ser transmitido.

Finalizado o processo de transmissão, o canal de serviço é liberado pelos terminais, tornando-se disponível para a rede novamente.

Como a rede conta com K canais, é importante ressaltar que o processo de *polling* não precisa aguardar que todas as mensagens selecionadas sejam transmitidas, para seguir adiante na regra de atendimento. Ou seja, o processo de *polling* continua saltando de um terminal para o próximo, independentemente da Fase de Transmissão de um par de terminais e independentemente da existência de canais de serviço livres para serem utilizados pelo próximo par.

Além disto, como a rede não tem conhecimento prévio do estado de cada terminal, as mensagens de *polling* são enviadas aos mesmos, mesmo que estes não tenham mensagens ou estejam inaptos para transmissão.

Se considerarmos uma rede não estruturada (*Ad Hoc*), onde os terminais não dependem de uma estrutura para se comunicarem, a regra de atendimento aleatória poderia representar um processo randômico de competição entre terminais da rede pela oportunidade de controle do meio (*token*), para então poderem transmitir as mensagens da fila pelo canal de serviço negociado.

O protocolo proposto em [24] pode ser considerado um exemplo, onde se poderia pensar na utilização do modelo com regra de atendimento aleatória, proposto neste trabalho.

A escolha do próximo terminal a ser "consultado"no processo de *polling* foi analisada em ambas as regras de atendimento:

- Cíclica: onde o próximo terminal a ser escolhido está dentro de uma sequência

pré-definida na rede, voltando ao primeiro na sequência após completar um ciclo.

- Aleatória: a escolha do próximo terminal a ser consultado é aleatória, onde todos os terminais têm a mesma probabilidade $(1/N)$ de serem escolhidos.

3.1.3 Medida de Interesse do Sistema

Um sistema de comunicação pode ter diversas medidas de interesse, dependendo da aplicação e do protocolo de controle de acesso ao meio utilizado. Entretanto, na avaliação da eficiência do próprio protocolo de controle de acesso no compartilhamento dos recursos entre os terminais, o tempo médio de espera que uma mensagem qualquer experimenta, desde o instante t_0, no qual chegou ao terminal de origem, até o instante t_1, no qual se iniciou o processo de transmissão para o destino, é uma das principais medidas de interesse.

Do ponto de vista da mensagem aguardando no *buffer* do terminal de origem, a medida do tempo médio de espera representa basicamente o quão rapidamente o protocolo disponibiliza o recurso, para a transmissão da mensagem até o terminal de destino.

Se os terminais da rede considerada estiverem distribuídos numa área pequena e os canais de transmissão forem de alta capacidade, como no caso de uma rede local (*LAN*, Figura 3.2), o tempo médio de espera no *buffer*, pode ser bastante representativo no desempenho geral da rede.

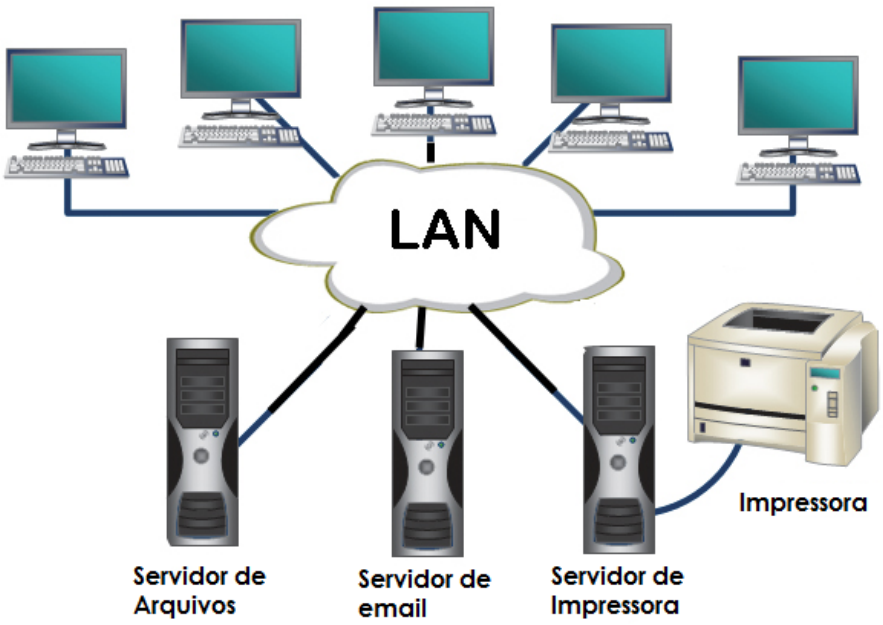

Figura 3.2: Rede local comum.

3.2 Definição do Modelo

Neste trabalho serão utilizados dois modelos, baseados em estruturas de filas:

- Modelo Analítico: baseado numa estrutura de filas $M/G/K$ com férias, considerando que os servidores entram novamente em férias, caso não exista nenhum usuário aguardando para ser atendido.
- Modelo de Simulação: será utilizado um modelo de simulação baseado numa estrutura de filas, controlada por um sistema do tipo "produtor x consumidor".

O objetivo da utilização destes dois modelos juntos é poder confrontar os resultados de ambos para validar os resultados numéricos obtidos e ampliar a análise de alguns cenários de testes com o sistema saturado. Neste último caso, será utilizado somente o modelo de simulação, uma vez que foram adotadas as hipóteses do sistema operando em regime estacionário e de forma estável, para se chegar no modelo analítico.

3.2.1 Representação no Modelo

Nesta fase, os principais atores e processos do sistema de comunicação são representados na modelagem.

- Cada um dos N terminais é representado por uma única fila n, dentro de uma estrutura de N filas.
- Cada um dos K canais ortogonais de comunicação *broadcast* livre de erros, é representado por um único servidor, dentro de um grupo de K servidores.
- O processo de chegada de mensagens nos terminais é representado pelo processo de chegada de usuários nas filas do sistema, com taxa média igual a λ usuários/segundo.
- O processo de transmissão de mensagens de um terminal de origem para um outro terminal de destino, dentro da mesma rede, é representado pelo processo de atendimento de cada servidor a cada um dos usuários aguardando nas filas da estrutura, com capacidade média de μ usuários/segundo.
- O tempo de transmissão de uma mensagem é o intervalo de tempo decorrido entre o início da transmissão do primeiro *bit* da mensagem pelo canal de transmissão, alocado no terminal de origem, até o recebimento do último

bit da mesma mensagem pelo terminal de destino, desprezando-se o tempo de propagação[3].

- O tempo que uma mensagem aguarda na fila do terminal, até que se inicie o seu processo de transmissão, é representado por pelo tempo que um usuário i aguarda na fila, antes de ser atendido pelo servidor.

3.2.2 Medidas de Interesse no Modelo

Nesta seção é especificada a medida de interesse principal do sistema de comunicação, que será analisada no modelo.

De acordo com ADAN e RESING [55], é possível extrair algumas medidas de interesse bastante relevantes dos modelos de filas:

- Distribuição do tempo de espera em fila.
- Distribuição do tempo de permanência do servidor na fila, também conhecido como *sojourn time*.
- Distribuição do número de usuários no sistema, incluindo ou não os usuários em serviço.
- Distribuição do total de trabalho no sistema. Este medida é a soma dos tempos de serviço dos usuários aguardando na fila e dos tempos residuais dos usuários em serviço.
- Distribuição do período que o servidor fica ocupado, atendendo às filas.

A medida de interesse escolhida para ser analisada neste trabalho é o tempo de espera na fila W de cada usuário i, que pode ser representada por uma variável aleatória $[W_i]_{i=0}^{\infty}$, com média $\overline{W}$ e variância σ_W^2. Esta medida representa o intervalo de tempo desde a chegada do usuário na fila de origem, até o início do seu processo de atendimento. A Figura 3.3 ajuda a entender esta grandeza.

Na Figura 3.3 os usuários chegam na fila numa taxa média igual a λ usuários/segundo e são atendidos pelo servidor, cuja capacidade média de atendimento é igual a μ usuários/segundo.

Um usuário i, que acaba de chegar na fila no instante t_0, deverá esperar $W_i = t_1 - t_0$, até ser atendida no instante t_1.

[3]O tempo de propagação numa rede local, com distância máxima entre os terminais é 100 metros, é de aproximadamente $50\mu s$, cerca de 15 vezes menor que o tempo médio de transmissão de uma mensagem ($727\mu s$).

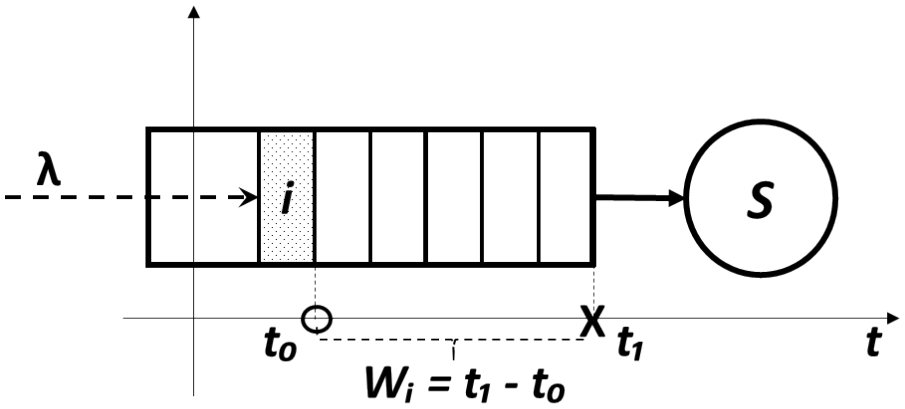

Figura 3.3: Tempo médio de espera numa fila qualquer.

Desta forma, pode-se dizer que o tempo médio de espera em fila mede basicamente a eficiência do protocolo de controle de acesso ao meio, considerando que o sistema está em regime estacionário, ou $i \to \infty$.

3.2.3 Hipóteses Adotadas

Antes de iniciar a construção dos modelos, faz-se necessário adotar algumas hipóteses especiais sobre o modelo escolhido para facilitar o desenvolvimento matemático das expressões. São elas:

- Em cada fila, o número de usuários que chega num dado intervalo de tempo t é representado por uma variável aleatória de *Poisson*, com média λt.
- Os tempos de atendimento dos usuários em cada uma das N filas são representados por variáveis aleatórias $[X_i]_{i=1}^{\infty}$, independentes e identicamente distribuídas (iid), que seguem uma distribuição geral $B(x) = P\{X_i \leq x\}$, onde X_i é o tempo de serviço do *i-ésimo* usuário; com média $\overline{X}$, variância σ_X^2.
- O sistema de filas está em regime estacionário ($i \to \infty$).
- O sistema opera em regime estável. Isto é a utilização $\rho < 1$. Segundo [4], ρ representa o fator de utilização do sistema, que corresponde à razão entre a quantidade de trabalho que chega no sistema (unidade de tempo) e a capacidade do sistema em executar este trabalho.
- Todas as filas possuem capacidade de armazenamento infinita. Ou seja, nenhum usuário sai da fila sem que seja atendido pelo servidor.
- Os processos de chegada de usuários no sistema e do tempo de serviço são independentes entre si.

- O intervalo de tempo para comutar o controle de uma fila n para a próxima fila $n+1$, também conhecido como *walk time*, é constante e denotado por V.
- O número de servidores da estrutura de filas é denotado por K.
- Todos os K servidores são estatisticamente idênticos e sem falhas[4].
- Cada um destes K servidores tem uma taxa média de atendimento (serviço) igual a μ usuários/segundo .
- A estrutura de filas é simétrica. Isto é, as filas são estatisticamente idênticas e cada uma das N filas recebe a mesma taxa média de chegada de usuários (λ). Assim, a taxa média total de chegada de usuários no sistema de filas é dado por $\Lambda = N\lambda$.
- A ordem de atendimento dos usuários é independente dos seus tempos de serviço (atendimento).

3.2.4 Modelo Analítico

Os modelos analíticos têm como principal vantagem a flexibilidade em prover resultados precisos rapidamente ou em condições assintóticas, difíceis de serem obtidos através de outros tipos de modelo. Por exemplo: imagine que se deseja analisar o comportamento do sistema com taxas de serviço muito baixas e com taxas de chegadas elevadas. Nesta condição, o tempo necessário para se obter dados de um modelo de simulação seria muito elevado, comprometendo a análise do sistema.

Geralmente, nas situações mais práticas de projeto, os resultados obtidos dos modelos analíticos são utilizados para embasar decisões iniciais ou de definição de metodologia.

O modelo analítico adotado neste trabalho provém de uma estrutura de filas do tipo $M/G/K$ com férias, onde os servidores retornam imediatamente às férias se não houver nenhum usuário aguardando para ser atendido nas filas.

A partir da estrutura $M/G/K$, proposta para construir o modelo analítico, é necessário estabelecer mais algumas definições, considerando o mecanismo de controle de acesso, disciplinas e regras de serviço do sistema de comunicação modelado, de acordo com as especificações funcionais do protocolo (Seção 3.1.2).

- O controle de acesso a estes servidores é feito através de um sistema de *polling*, que visita as filas segundo uma regra de atendimento cíclica[5].

[4]Considerando que os servidores representam canais de comunicação estatisticamente idênticos e sem erros.

[5]O modelo com regra de atendimento aleatória será desenvolvido a partir do modelo cíclico.

- Somente um dos K servidores pode ser escolhido para atender a uma determinada fila, visitada pelo *polling*.

- O período de férias do servidor, do ponto de vista de uma determinada fila, representa o ciclo do processo de *polling*, que é o tempo decorrido entre duas visitas consecutivas nesta mesma fila.

- Quando um dos K servidores é escolhido[6] para atender uma determinada fila n, os demais servidores (máximo $K-1$) continuam disponíveis; e portanto seguem prontos para eventuais novas alocações, à medida que o processo de *polling* avança na seqüência de visitas às demais $(N-1)$ filas.

- Atender a uma fila significa servir no máximo os L primeiros usuários, selecionados dentre o conjunto de usuários que estavam na fila, antes da visita do servidor (*gated*). O serviço então é finalizado quando qualquer uma das três condições abaixo acontecer primeiro:

  - Todos estes L usuários forem servidos.
  - Todos o conjunto de usuários antes da "porteira"(*gate*) forem servidos.
  - Todos os usuários da fila forem servidos.

  A Figura 3.4 ajuda a entender o modelo proposto.

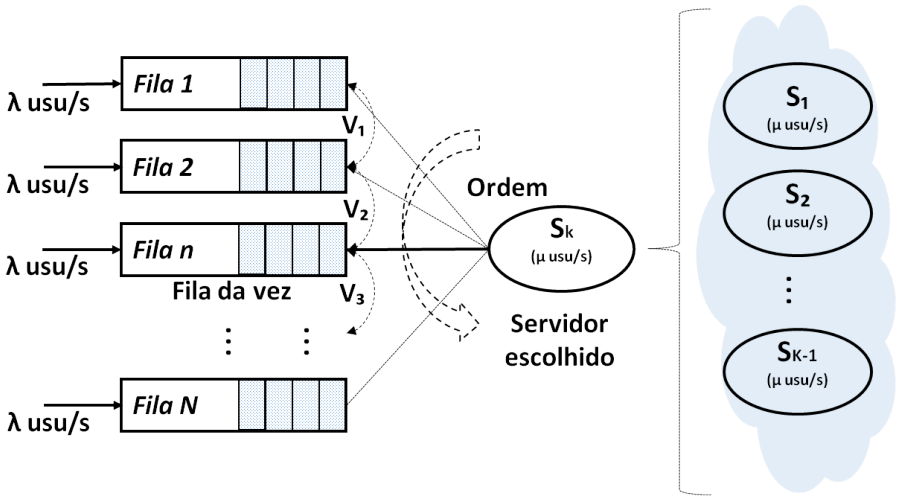

Figura 3.4: Sistema de Filas com Múltiplos Servidores considerado.

Nas próximas Seções 3.2.4 - $M/G/1$, 3.2.4 - $M/G/1$ com férias e 3.2.5, serão apresentados inicialmente alguns modelos de filas básicos, que foram utilizados para se chegar nos modelos com múltiplos canais finais, propostos neste trabalho.

[6]Neste trabalho o servidor é escolhido aleatoriamente.

Modelos *M/G/1*

A análise dos modelos mais complexos, que serão estudados neste trabalho, será iniciada a partir do modelo $M/G/1$, com taxa média de chegada de usuários λ, tempo médio de serviço $\overline{X}$, segundo momento $\overline{X^2}$ e variância σ_X^2.

O modelo de filas $M/G/1$ é o na verdade um caso particular do modelo $M/G/K$, quando $(K = 1)$. Ou seja, existe apenas um servidor para atender à estrutura de filas.

Utilizando-se o resultado da referência [56], chega-se na seguinte expressão para o valor do tempo médio de espera numa fila *M/G/1* em regime estacionário:.

$$\overline{W}_{M/G/1} = \frac{N\lambda\overline{X^2}}{2(1-\rho)} \tag{3.1}$$

Segundo [56], a Equação (3.1) é conhecida como fórmula de *Pollaczek-Khinchin*. Esta expressão está demostrada na Seção A.1 do Apêndice A.

Modelo *M/G/1* com Férias

O modelo *M/G/1* com férias (parâmetros λ e $\overline{X}$) considera que o servidor fica indisponível para atender a uma determinada fila durante um tempo aleatório, logo depois do período de serviço na mesma. Este período inativo pode ser comparado figurativamente com um período de "férias", daí se origina a designação deste modelo.

Nesta nova condição, um outro componente residual de tempo é adicionado no cálculo do tempo médio de atraso na fila *M/G/1*. Este novo componente é calculado em [56], a partir da possibilidade de um usuário qualquer chegar numa das filas e encontrar o servidor **durante** um período de férias.

Juntando-se este novo componente de tempo na Equação (3.1), obtém-se a expressão para o tempo médio de espera numa fila *M/G/1* com férias, em regime estacionário:

$$\overline{W} = \overline{W}_{M/G/1} + \frac{\overline{V^2}}{2\overline{V}} \tag{3.2}$$

A demonstração deste resultado pode ser analisada com mais detalhes na Seção A.2 do Apêndice A.

3.2.5 Modelo de *Polling*

No sistema de comunicação considerado, cada terminal da rede recebe uma mensagem de *polling* de tempos em tempos (ciclos). Esta mensagem, como já explicado na Seção 3.1.2, tem como principal finalidade verificar se o terminal consultado tem mensagens de dados para transmitir para outros terminais da rede. Se o resultado desta consulta for positivo, uma oportunidade de transmissão é concedida ao terminal consultado.

Traçando um paralelo entre o sistema de comunicação considerado e o modelo de *polling*, cada terminal da rede é representado por uma fila, o canal é representado pelo servidor, as mensagens de dados, que chegam no *buffer* dos terminais, são representadas pelos usuários, chegando nas suas respectivas filas; e o processo de *polling* é representado pelo processo de visitas em cada fila.

Este modelo considera que a regra de atendimento é cíclica. Isto é, a próxima fila a ser visitada pelo servidor é escolhida numa seqüência definida e após a visita à última fila, o processo se reinicializa, voltando à primeira fila.

Modelo com Disciplina de Serviço *gated*

Na disciplina de serviço do tipo *gated*, somente os usuários que estavam aguardando na fila até a visita do servidor serão atendidos. Os usuários que chegarem a partir deste momento deverão aguardar o próximo ciclo para serem atendidos. Voltando na comparação paralela com o sistema de comunicação modelado, serão transmitidas somente as mensagens que já estiverem aguardando na fila, no momento que o seu terminal recebe a mensagem de *polling* do Ponto de Acesso.

Um modelo matemático para cálculo do tempo médio de espera em fila ($\overline{W}$), considerando um sistema de *polling* com disciplina de serviço do tipo *gated*, é apresentado em [46]. Esta proposta provém de uma análise em tempo contínuo do sistema de filas, de acordo com as características postuladas no parágrafo anterior.

$$\overline{W} = \frac{N\lambda\overline{X^2}}{2(1-\rho)} + \frac{(N+\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \tag{3.3}$$

A Equação 3.3 foi baseada ainda em algumas hipóteses adicionais, utilizadas para facilitar o desenvolvimento do modelo:

- Operação com um único servidor (equivalente a um único canal no sistema de comunicação modelado).

- Regra de atendimento com ciclos regulares.
- Sem diferenciação por prioridades.
- A taxa total de usuários que chega nas N filas, considerando a simetria da rede, pode ser representado por:

$$\Lambda = \sum_{n=1}^{N} \lambda_n = N\lambda \tag{3.4}$$

O modelo da Equação 3.3, com algumas adaptações, pode ser utilizado como base para o desenvolvimento dos outros modelos, de acordo com os objetivos deste trabalho. Para se adaptar o modelo básico dentro dos objetivos propostos, é preciso analisar a metodologia utilizada para se chegar na Equação 3.3.

Analisando separadamente os termos da Equação 3.3 é possível perceber que o primeiro e o último termos provém dos resultados das Equações (3.1) e (3.2), que são os resultados dos modelos $M/G/1$ e $M/G/1$ com férias, respectivamente.

Sobre este resultado da Equação 3.2, adiciona-se um terceiro componente, que decorre do valor médio esperado da soma de todos os intervalos de *polling*[7], que um determinado usuário i deve aguardar até ser atendido, considerando que existe a possibilidade deste usuário i chegar num determinado terminal, depois da visita do servidor; e portanto terá que esperar todo ou uma parte do ciclo de *polling* para ser atendido.

Sendo a taxa total de chegada de usuários no sistema de acordo com a Equação 3.4, a fração de utilização do servidor pode ser expressa por[8]:

$$\rho = N\lambda\overline{X} = \frac{N\lambda}{\mu} \tag{3.5}$$

Para que o sistema como um todo se comporte de forma estável, é necessário que a utilização se mantenha abaixo do seu valor máximo. Ou seja: $\rho < 1$.

Modelo com Disciplina de Serviço *gated/L-limitado*

Na disciplina de serviço *gated/1-limited*, o servidor servirá apenas **um** dos usuários que já estavam na fila no momento da chegada do servidor (se houver algum).

[7]Este intervalo também é conhecido como *walk time* e representa o tempo necessário para o servidor migrar de uma fila n para a próxima fila $n+1$.

[8]Lembrando que neste modelo está sendo considerado o sistema com um único servidor (somente um canal de comunicação).

Nesta condição, para cada usuário atendido, haverá um acréscimo de um novo ciclo de *polling*. Portanto, o valor de $\overline{Y}$ deve ser acrescido de $NW\overline{V}$.
Onde $\overline{Y}$ é o valor médio da variável aleatória $[Y_i]_{i=0}^{\infty}$, representando a soma dos intervalos de *polling*, que o *i-ésimo* usuário ($i \to \infty$) deve aguardar, desde o instante da sua chegada até a próxima visita do servidor (ciclo)[9].

A equação do valor do tempo médio de espera em fila fica:

$$\overline{W} = R + \rho\overline{W} + \overline{Y} + N\lambda\overline{WV} \tag{3.6}$$

$R = R_x + R_p$ é a componente de tempo médio residual, calculada a partir da soma do tempo médio residual de serviço (R_x) com o tempo médio residual de *polling* (R_p).

ou seja:

$$\overline{W} = \frac{R + \overline{Y}}{1 - \rho - N\lambda\overline{V}} \tag{3.7}$$

O trabalho apresentado em [30] analisa um sistema de *polling* simétrico em estado estacionário, operando segundo uma regra de atendimento cíclica com um único servidor. Nesta análise é mostrado que o número de usuários no sistema pode ser calculado a partir da soma de três diferentes variáveis aleatórias, sendo uma delas a própria quantidade de usuários numa fila *M/G/1* padrão. A partir deste resultado são obtidas equações do tempo médio de espera em fila para as disciplinas de serviço: exaustiva, *gated* e *L-limitado*.

Para o contexto deste trabalho, o principal resultado da análise de [30] é a definição de um limite superior para o tempo médio de espera em fila, considerando uma disciplina de serviço *L-limitado*. Este resultado pode ser aplicado na (3.7) para formar a Equação 3.8 abaixo:

$$\overline{W} \leq \frac{R + \overline{Y}}{1 - \rho - \frac{N\lambda\overline{V}}{L}} \tag{3.8}$$

Onde L representa o número máximo de usuários na fila que poderão ser atendidos pelo servidor em cada visita.

Analisando o resultado da Equação 3.8 é possível estabelecer uma outra relação para o tempo médio de espera em fila da disciplina *L-limitado* em relação à

[9] A dedução completa deste resultado pode ser encontrada em [56].

disciplina *gated*.

$$\overline{W}_{L-limitado} = \overline{W}_{gated} \times \left(\frac{1-\rho}{1-\rho-\frac{N\lambda\overline{V}}{L}} \right) \tag{3.9}$$

Juntando os resultados das Equações 3.9 e 3.3, chega-se na seguinte expressão para o tempo médio de espera em fila com disciplina de serviço *gated*/L-limitado:

$$\overline{W} = \frac{N\lambda E[X^2]}{2(1-\rho-\frac{N\lambda\overline{V}}{L})} + \frac{(N+\rho)\overline{V}}{2(1-\rho-\frac{N\lambda\overline{V}}{L})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\frac{N\lambda\overline{V}}{L})} \tag{3.10}$$

Aproveitando a relação definida na Equação 3.1, pode-se reescrever a Equação 3.10 da seguinte forma:

$$\overline{W} = \frac{\overline{W}_{M/G/1}(1-\rho)}{(1-\rho-\frac{N\lambda\overline{V}}{L})} + \frac{(N+\rho)\overline{V}}{2(1-\rho-\frac{N\lambda\overline{V}}{L})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\frac{N\lambda\overline{V}}{L})} \tag{3.11}$$

Considerando que os intervalos de *polling* (*walk time*) são fixos, implica que $\overline{V} = V$ e $\sigma_V^2 = 0$. Assim a Equação 3.11 pode ser reescrita como:

$$\overline{W} = \frac{\overline{W}_{M/G/1}(1-\rho)}{(1-\rho-\frac{N\lambda V}{L})} + \frac{(N+\rho)V}{2(1-\rho-\frac{N\lambda V}{L})} \tag{3.12}$$

3.2.6 Aproximação para o Modelo *M/G/K*

Uma primeira aproximação para o modelo com K servidores é a utilização dos servidores. Em [57] também considera-se uma rede completamente simétrica, operando em regime estacionário, no cálculo da carga de cada um dos K servidores do sistema, considerando que os K servidores dividem igualmente a carga total do sistema, na média.

$$\rho = P\{Servidor\ k\ Ocupado\} = \frac{N\lambda\overline{X}}{K} = \frac{N\lambda}{K\mu} \tag{3.13}$$

Apesar de ser um cenário bastante comum, principalmente nas redes de telecomunicações, o cálculo do tempo médio de espera em fila para sistemas com múltiplos servidores do tipo $M/G/K$ não é simples. Muitas vezes a complexidade matemática é tanta que o problema se torna "intratável". Para contornar alguns destes problemas, tem sido propostas muitas aproximações, que geralmente se baseiam em interpolações e heurísticas. Uma destas aproximações com excelentes resultados é apresentada em [33].

Aproximações do Cálculo do Tempo Médio de Espera Básico para K Servidores ($\overline{W}_{M/G/K}$)

A aproximação do cálculo do tempo médio de espera em fila, proposta em [33], para modelos do tipo $M/G/K$, utilizada nesta dissertação é:

$$\overline{W}_{M/G/K} \simeq \frac{1 + C_X^2}{\frac{2C_X^2}{\overline{W}_{M/M/K}} + \frac{1-C_X^2}{\overline{W}_{M/D/K}}} \tag{3.14}$$

Onde o termo C_X é o coeficiente de variação do tempo de serviço, que é dado por:

$$C_X = \frac{\sigma_X}{\mu}$$

$\overline{W}_{M/M/K}$ corresponde ao valor do tempo médio de espera em fila, considerando que o tempo de serviço dos usuários obedece a uma distribuição exponencial. Este valor pode ser calculado através da seguinte equação:

$$\overline{W}_{M/M/K} = \frac{(K\rho)^K}{K!K\mu(1-\rho)^2} \left\{ \sum_{j=0}^{K-1} \frac{(K\rho)^j}{j!} + \frac{(K\rho)^K}{K!(1-\rho)} \right\}^{-1} \tag{3.15}$$

$\overline{W}_{M/D/K}$ corresponde ao valor do tempo médio de espera em fila, considerando que o tempo de serviço é uma grandeza determinística. O cálculo de $\overline{W}_{M/D/K}$, assim como o cálculo de $\overline{W}_{M/G/K}$, não é simples e também requer aproximações.

Para o cálculo de $\overline{W}_{M/D/K}$ utilizou-se uma aproximação proposta na referência [58], que tem com base no valor de $\overline{W}_{M/M/K}$. Esta aproximação, de acordo com [33], funciona bem quando o número de servidores $K \leq 10$, o que está de acordo com os objetivos desta dissertação.

$$\overline{W}_{M/D/K} = \Psi(\theta, \rho)\overline{W}_{M/M/K} \tag{3.16}$$

Onde:

$$\theta = \frac{K-1}{K+1} \quad \text{e} \quad \Psi(\theta, \rho) = \frac{1}{2}\{1 + F(\theta)g(\rho)\}$$

$$F(\theta) = \frac{\theta}{8(1+\theta)} \left(\sqrt{\frac{9+\theta}{1-\theta}} - 2 \right) \quad \text{e} \quad g(\rho) = \frac{1-\rho}{\rho}$$

É importante ressaltar que todas as aproximações apresentadas nesta seção se aplicam apenas à sistemas não saturados, ou seja, $\rho < 1$.

A análise matemática de sistemas saturados ($\rho \geq 1$) não faz parte do escopo

deste trabalho, por que o sistema é considerado estável na Seção 3.2.3. Maiores informações sobre análise de sistemas saturados podem ser obtidas em [59].

Aproximações Funcionais do Modelo com K Servidores

As aproximações funcionais podem ser definidas neste trabalho como adaptações matemáticas aplicadas à Equação 3.12, considerando o cenário de múltiplos servidores e as características operacionais do sistema de comunicação modelado, apresentadas na Seção 3.1.2. Estas alterações no modelo provêm de considerações assumidas ou de ponderações, baseadas no comportamento esperado da rede de comunicação, sob condições especiais.

- No sistema de comunicação considerado, o PA (Ponto de Acesso), que controla o processo de *polling*, não tem conhecimento nenhum sobre o estado dos terminais da rede. O PA não sabe por exemplo, se um determinado terminal tem mensagens aguardando para serem transmitidas, nem se está transmitindo ou recebendo. No modelo de filas, as visitas às filas acontecem continuamente, independentemente do estado ou da existência de usuários nas mesmas.

- Relembrando um dos parágrafos da Seção 3.1.2 que, apesar de existirem múltiplos canais na rede, cada terminal está restrito a utilizar somente um deles para ser atendido. No modelo de filas, a fila visitada será atendida somente por um dos K servidores disponíveis.

- O servidor escolhido é imediatamente alocado para atender a fila visitada, ficando indisponível (ocupado) para atender outras filas, durante o seu período de serviço. Entretanto, a sequência de visitas continua para o restante das filas. Assim, se houver pelo menos um servidor livre para servir os usuários da próxima fila visitada, um determinado usuário aguardando nesta fila não precisa aguardar pelo fim do serviço na fila anterior, considerando que o processo de *polling* é contínuo e independente das condições da rede.

Aplicação das Aproximações no Modelo *M/G/K*

As aproximações descritas nas seções anteriores serão aplicadas à Equação (3.12), que se encontra listada logo abaixo para facilitar o entendimento passo a passo:

1. O numerador do primeiro termo da Equação (3.12) contém o tempo médio de espera na fila $M/G/1$ ($\overline{W}_{M/G/1}$). No caso de K servidores, este termo pode ser substituído pelo tempo médio de espera na fila do tipo $M/G/K$ ($\overline{W}_{M/G/K}$), calculado pela Equação (3.14).

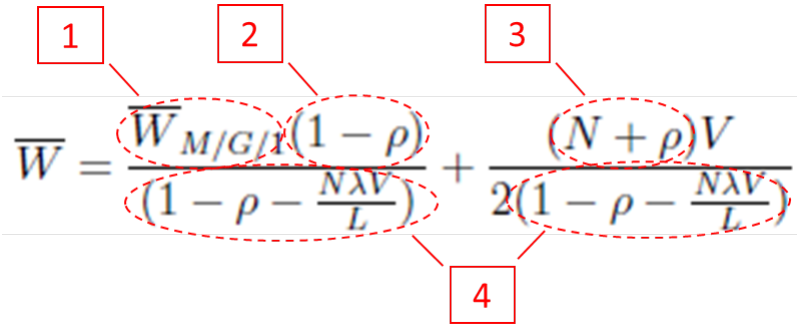

Figura 3.5: Equação (3.12) comentada.

2. Ainda no numerador do primeiro termo, aparece o fator $(1-\rho)$. Este fator é utilizado na adaptação do modelo *gated/L-limitado*, oriundo da Equação (3.9). No cenário com K servidores, o valor da fração de utilização a ser utilizado é o da Equação (3.13). Esta alteração se deve ao fato de que este primeiro termo decorre do cálculo do tempo médio de serviço residual em cada fila e as filas são sempre atendidas por um único servidor (vide segundo item da Seção 3.2.6.

3. De acordo com a referência [56], o numerador do segundo termo da Equação (3.12) $(N+\rho)V$ decorre da soma do tempo médio residual de um período de férias (R_p) com o tempo médio total de férias de um ciclo de *polling* $(\overline{Y})$. No cálculo de $\overline{Y}$ apresentado em [56], a fração de utilização que pondera a probabilidade de um usuário chegar ao sistema, durante um intervalo de dados ou durante um intervalo de férias, considera todos os K servidores. Consequentemente, este fator precisa ser alterado para $K\rho$.

4. O cálculo do denominador dos dois termos anteriores $(1-\rho-\frac{N\lambda\overline{V}}{L})$ é baseado originalmente nos usuários esperando nas filas, que serão servidos antes ($\overline{QX}$ - onde $\overline{Q}$ é o número médio de usuários esperando na fila e $\overline{X}$ o tempo médio de atendimento de cada usuário na fila); e no tempo médio adicional, decorrente da limitação do número máximo de usuários servidos por fila numa visita (L), de acordo com [56].

   Conforme o terceiro item da Seção 3.2.6, com múltiplos servidores no sistema de *polling*, um determinado usuário esperando pelo serviço numa fila a ser visitada não precisa aguardar o serviço da fila anterior, desde que haja servidores disponíveis na sua vez. Desta forma, o tempo decorrente dos usuários que serão servidos antes deve ser ponderado pela probabilidade de todos os servidores estarem ocupados (ρ^K).

Aplicando as aproximações acima descritas, a equação final do tempo médio de espera em fila do sistema de *polling* com regra de atendimento cíclica, disciplina de

serviço *gated*/L-limitado, operando com K canais pode ser escrita como:

$$\overline{W} = \frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K - \frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K - \frac{N\lambda V}{L})} \tag{3.17}$$

3.2.7 Modelo com Prioridades

O sistema de comunicação com diferenciação por prioridades, presume que as mensagens chegam nos dos terminais com uma marca (*tag*), indicando a sua prioridade frente às outras mensagens, aguardando pelo atendimento. A prioridade de cada mensagem pode variar de 1, que é a prioridade mais alta (por exemplo: uma mensagem contendo dados de uma aplicação de voz), até P, que é a prioridade mais baixa (por exemplo: uma mensagem contendo dados de uma aplicação de envio de *e-mail*). Desta forma, se consegue otimizar a utilização dos meios de transmissão (recursos), sem comprometer significativamente o desempenho das aplicações.

No modelo proposto, os usuários que adentram ao sistema de filas, se diferenciam uns dos outros por uma marca equivalente, indicando a sua classe de prioridade. Ao chegarem nas filas, os usuários são organizados para o atendimento, de acordo com a sua classe de prioridade. Isto é, as prioridades mais altas à frente dos usuários de prioridades mais baixas, até se completar o número máximo de usuários atendidos (L), durante a visita do servidor.

A Figura 3.6 mostra o esquema gráfico das chegadas (taxa média λ_p) e a classificação dos usuários numa determinada fila (*Terminal n*), de acordo com a sua prioridade. Depois de classificados e ordenados (*HOL*), o servidor escolhido S_i processa no máximo o número de usuários L, das que estavam presentes na fila no instante da chegada do servidor (*gated*).

Dentro das filas, a ordenação dos usuários **dentro da mesma classe de prioridade** obedece à disciplina *FCFS*. Isto é, os usuários da mesma classe de prioridade, que chegam primeiro na fila, são atendidos primeiro.

Hipóteses Adotadas no Modelo com Diferenciação por Prioridades

Assim como foi feito na Seção 3.2.3, serão listadas logo abaixo as considerações especiais, que foram assumidas para viabilizar o modelo com prioridades:

- Cada usuário que entra no sistema de filas possui uma marca, que indica a sua classe de prioridade.

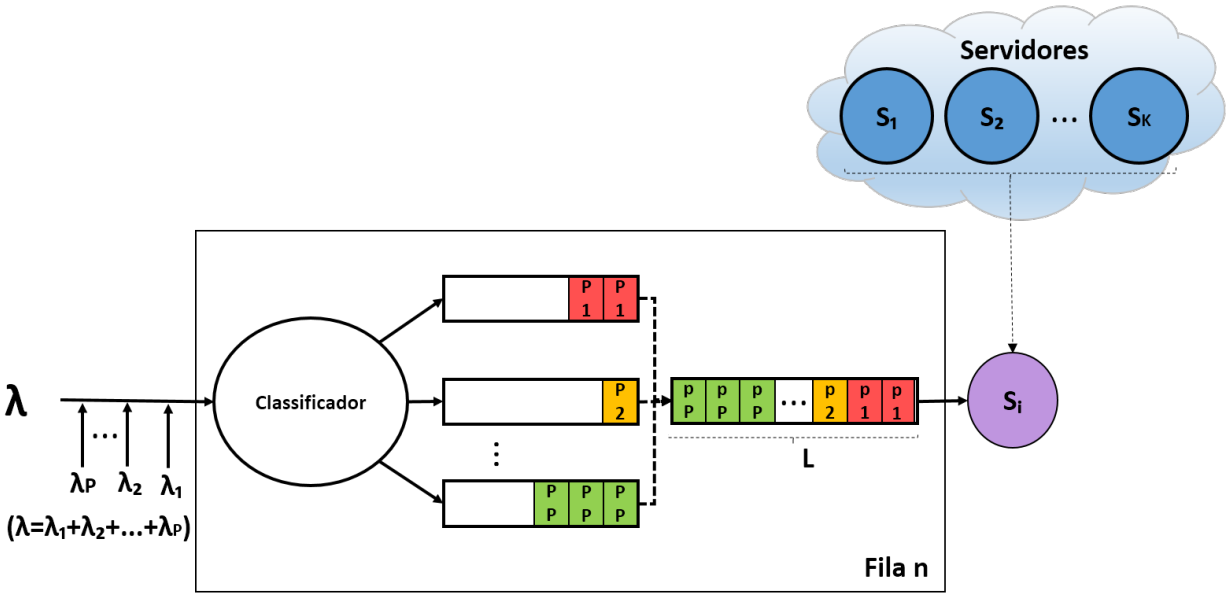

Figura 3.6: Chegada e saída do terminal n.

- Todos os usuários no sistema serão eventualmente servidos, mais cedo ou mais tarde.
- A rede não têm qualquer conhecimento da quantidade nem da prioridade dos usuários que estão aguardando nas filas.
- Enquanto houver usuários nas filas para serem atendidos, sempre haverá no mínimo um servidor ocupado.
- Os usuários são organizados dentro das filas, de acordo com as suas respectivas classes de prioridade. Isto é, os usuários de uma classe de prioridade mais alta serão agrupados na frente dos usuários de menor prioridade. Esta disciplina de fila, aplicada ao modelo, é conhecida como *HOL*.
- Os usuários pertencentes a uma mesma classe de prioridade são sub-agrupadas de acordo com a sua ordem de chegada na fila. Este tipo de disciplina de fila é conhecida como *FCFS* ou *FIFO*.
- A regra de interrupção é sem interrupção (*nonpreemptive*). Isto é, uma vez iniciado o serviço num usuário, este não pode ser interrompido pela chegada de um outro usuário de mais alta prioridade.
- As regras que controlam o início e o fim das férias dos servidores, de acordo com o modelo analítico escolhido, são independentes das chegadas de usuários nas filas (*Poisson*).

- O servidor escolhido irá atender um número máximo de L usuários por visita, do conjunto de usuários presentes na fila até o instante da chegada do servidor (*gated/L-limitado*).

Modelo Final com Prioridades e Notação Matemática

No modelo com diferenciação por prioridades, cada fila ($n = 1, 2, ..., N$) pode conter usuários de uma ou mais classes de prioridade. Considerando P como o número total de classes de prioridade possíveis numa determinada fila n; e ($p = 1, 2, ..., P$) como uma determinada classe de prioridade entre 1 e P, define-se 1 como a mais alta prioridade e P como a mais baixa prioridade.

A fração de utilização do sistema como um todo, considerando P diferentes classes de prioridade, é dada por:

$$\rho = \sum_{p=1}^{P} \rho_p = \frac{\Lambda}{K\mu}$$

Onde K representa o número total de servidores do sistema e Λ é a taxa média total de chegada de usuários de qualquer classe de prioridades no sistema, dada pela Equação (3.4). Considera-se ainda a simetria de todas as N filas e a identidade estatística dos servidores (vide 3.2.3).

Separando a taxa de entrada em cada fila (λ) por classe de prioridade, pode-se definir λ_p como a taxa média de chegada de usuários de classe de prioridade p numa determinada fila.

$$\lambda = \sum_{p=1}^{P} \lambda_p$$

Considerando que todos os K servidores são estatisticamente idênticos, a taxa média de serviço é sempre μ, para qualquer classe de prioridade. Desta forma, a fração de utilização de cada um dos K servidores, atribuída a uma determinada classe de prioridade p, é dada por ρ_p:

$$\rho_p = \frac{N\lambda_p}{K\mu}$$

Em [36] é proposto um modelo para o tempo médio de espera em fila num sistema de *polling* com um único servidor, operando segundo a disciplina de serviço *gated*. O modelo proposto em [36], considera que o período de férias do servidor é reiniciado,

quando no retorno, não existirem usuário para serem atendidos no sistema de filas.

Reescrevendo a equação do tempo médio de espera em fila de [36], já substituindo a notação do livro pela notação adotada neste trabalho, considerando que o período de férias V é fixo, tem-se:

$$\overline{W}_p = \left(1 + \rho_{p-1}^{+} + \rho_p^{+}\right) \times \left[\frac{\lambda\overline{X^2}}{2(1-\rho^2)} + \frac{\overline{V^2}}{2(1+\rho)V}\right] \tag{3.18}$$

Onde: $\rho_p^{+} \triangleq \sum_{i=1}^{p} \rho_i$

Substituindo o primeiro termo da expressão entre colchetes da Equação (3.18) pela Equação (3.1) e colocando o termo $(1+\rho)$ em evidência no denominador, chega-se:

$$\overline{W}_p = \frac{(1 + \rho_{p-1}^{+} + \rho_p^{+})}{(1+\rho)} \times \left[\overline{W}_{M/G/1} + \frac{\overline{V^2}}{2V}\right] \tag{3.19}$$

Analisando a parte da Equação (3.19), que está entre os colchetes, verifica-se que é idêntica à Equação (3.2) do tempo médio de espera em fila para uma estrutura $M/G/1$ com férias, proposto em [56] e donde partiu o desenvolvimento dos demais modelos de *polling*.

Por indução, chega-se então que a parte da Equação (3.19), que está fora dos colchetes, pode ser definida como o "Operador de Prioridades"do modelo de *polling* (OP). É como se fosse um fator multiplicativo geral, que pode ser aplicado aos modelos de *polling* sem diferenciação de prioridades (entre colchetes), para calcular o tempo médio de espera em fila de usuários da classe de prioridade p.

$$OP = \frac{(1 + \rho_{p-1}^{+} + \rho_p^{+})}{(1+\rho)} \tag{3.20}$$

Então, para se obter o cálculo do tempo médio de espera na fila, com diferenciação por prioridades, para o modelo com múltiplos servidores, basta aplicar o Operador de Prioridade (OP) sobre a Equação (3.17).

A equação final para o tempo médio de espera em fila com prioridades e múltiplos canais fica:

$$\overline{W}_p = \frac{(1+\rho_{p-1}^{+}+\rho_p^{+})}{(1+\rho)} \times \left[\frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K-\frac{N\lambda V}{L})} \right] \tag{3.21}$$

3.2.8 Modelo com Regra de Atendimento Aleatória

Este novo modelo considera que a regra de atendimento é aleatória. Ou seja, O próximo terminal a receber o *polling* é escolhido de forma aleatória (sem memória).

Conforme [37], os modelos com regra de atendimento cíclica tem sido utilizados com sucesso para analisar sistemas de *polling* com um controlador central. Um típico exemplo prático é um computador central que compartilha algum recurso com terminais remotos.

Contrastando com o controle centralizado, o esquema com regra de atendimento randômica tem por objetivo modelar sistemas de controle distribuído.

Os sistemas distribuídos se caracterizam basicamente pela ausência de um controlador central, decidindo qual será o próximo terminal a receber o *polling*. Geralmente, nos sistemas de controle distribuído, a ordem de *polling* está relacionada a alguma condição externa (ex: demanda de mensagens nas filas dos terminais do sistema, aguardando para serem transmitidas) e de algum algoritmo interno nos terminais, que lida com esta informação.

A rede *Ad hoc* representa um exemplo clássico de rede de controle distribuído. Conforme explicado na Seção 1.1.2, as redes *Ad hoc* dispensam a existência de uma estrutura centralizada para a comunicação direta entre dois terminais da mesma rede.

Um bom exemplo prático para utilização do modelo com regra de atendimento randômica é o protocolo proposto por [24]. O funcionamento deste protocolo pode ser dividido em duas fases distintas:

- Uma fase com contenção, que ocorre no canal de controle, quando os terminais que possuem mensagens nas suas filas competem pela oportunidade de transmitir (*token*) e em seguida escolhem um canal de serviços.
- Uma segunda fase sem contenção, onde o terminal que conseguiu ganhar o *token* e transmite suas mensagens através do canal escolhido.

O modelo de *polling* com ciclo randômico se caracteriza principalmente pela inexistência de uma sequência fixa pré-definida, segundo a qual as filas são visitadas. Isto é, se a fila n acabou de ser servida, a próxima fila a ser servida será a fila

$(j = 1, 2, ..., n, ..., N)$ com probabilidade p_j.

Considerando que todas as filas recebem a mesma taxa média de usuários (condição de simetria da rede conforme Seção 3.2.3), a probabilidade de uma fila j receber o *polling* é distribuída uniformemente da forma: $p_j = p_1 = p_2 = p_n = ... = p_N = p = \frac{1}{N}$.

Em [37] é proposto um modelo matemático para o tempo médio no sistema (T_S), numa estrutura *M/G/1*, com *polling*, com regra de atendimento randômica e disciplina de serviços *1-Limited*. A proposta, elaborada à partir de uma análise em tempo discreto (*slotted*), pode ser escrita como:

$$T_S = \frac{\delta^2}{2r} + \frac{(1 + Nr)\sigma^2}{2\mu(1 - N\mu - Nr\mu)} + \frac{N\delta^2\mu}{2(1 - N\mu - Nr\mu)} + \frac{(N - 1)r}{2(1 - N\mu - Nr\mu)} \quad (3.22)$$

Fazendo uma tradução na notação utilizada na Equação (3.22):

- δ^2 representa a variância do intervalo de *polling* ou σ_V^2 na notação utilizada neste trabalho.
- r representa o valor médio do intervalo de *polling* ou $\overline{V}$.
- N representa o número de terminais.
- σ^2 representa o segundo momento do tempo de serviço ou $\overline{X^2}$.
- μ representa o número médio de chegadas na fila no tempo t.

Considerando o tempo $t = 1/K\mu$ e multiplicando este valor pela taxa de chegadas λ, pode-se representar o valor de μ da Equação (3.22) da seguinte forma:

$$\mu = \frac{\lambda\overline{X}}{K} \quad (3.23)$$

Onde o valor de $\overline{X}$ e K do lado direito da Equação (3.23) representam respectivamente o tempo médio de transmissão no canal e o número de canais de serviço disponíveis na rede.

Substituindo a notação da Equação (3.22) pela notação utilizada neste trabalho e o valor de μ pelo seu cálculo equivalente na Equação (3.23), percebe-se que a Equação (3.22) é idêntica à Equação (3.10), exceto pelo último termo a mais e pela adaptação da Equação (3.10) para *L-limitado*.

Segundo [37] este último termo da Equação (3.22) representa justamente o acréscimo no tempo de serviço, devido à operação com *polling* randômico, em relação ao *polling* cíclico.

Juntando os resultados anteriores com o resultado apresentado na Equação (3.17), é possível chegar na equação do tempo médio de espera em fila para um sistema de *polling* com ciclo randômico e com múltiplos canais:

$$\overline{W} = \frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K - \frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K - \frac{N\lambda V}{L})} + \frac{(N-1)V}{2(1-\rho^K - \frac{N\lambda V}{L})} \tag{3.24}$$

A Equação (3.24) é basicamente o cálculo do tempo médio de espera em fila de um sistema com *polling* cíclico ((3.17)), acrescida da componente de tempo decorrente da regra de atendimento randômica, que está explicitada abaixo:

$$\frac{(N-1)V}{2(1-\rho^K - \frac{N\lambda V}{L})}$$

Importante relembrar que a Equação (3.24) já considera que o intervalo tempo para comutar de uma fila para a próxima (V), também conhecido como *walk time*, é mantido constante.

Uma outra abordagem para o cálculo da componente de tempo aleatória é analisando o número de intervalos de *polling* que a mensagem deverá aguardar, até que o seu terminal seja escolhido. Este número é na verdade uma variável aleatória geométrica de parâmetro $p = 1/N$, onde são contabilizados os primeiros fracassos antes do primeiro sucesso. Ou seja:

$$P(X = j) = (1-p)^j p$$

O valor médio esperado ($E\{j\}$) pode ser calculado da seguinte forma:

$$E[j] = \frac{1-p}{p} = N - 1$$

Isto significa que o número médio de intervalos de *polling* que uma determinada mensagem deve aguardar, até que o seu terminal seja sorteado para ser servido, é $N-1$. Considerando que no *polling* cíclico, este número médio é $(N-1)/2$, basta então acrescentar ao tempo médio de espera em fila do modelo *L-limitado* a outra metade deste acréscimo, referente ao número de intervalos de *polling*. Ou seja,

adicionando mais $(N-1)V/2$, chega-se finalmente na Equação (3.24).

A adaptação da Equação (3.24), incluindo diferenciação por prioridades, pode seguir a mesma estratégia utilizada na Seção 3.2.7. Ou seja, aplicando o Operador de Prioridade (OP) da Equação (3.20), é possível se chegar na equação que representa o tempo médio de espera em fila para um sistema de *polling*, com regra de atendimento aleatória e com diferenciação por prioridades:

$$\overline{W}_p = \frac{(1+\rho_{p-1}^{+}+\rho_p^{+})}{(1+\rho)} \times \left[\frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N-1)V}{2(1-\rho^K-\frac{N\lambda V}{L})} \right] \tag{3.25}$$

3.3 Tabela de Resultados

Para facilitar a leitura, é apresentada logo abaixo a Tabela 3.1, consolidando todos os resultados importantes obtidos neste capítulo.

Tabela 3.1: Resumo consolidado dos resultados do Capítulo 3.

Ref.	Equação
(3.12)	$\overline{W} = \frac{\overline{W}_{M/G/1}(1-\rho)}{(1-\rho-\frac{N\lambda V}{L})} + \frac{(N+\rho)V}{2(1-\rho-\frac{N\lambda V}{L})}$
Modelo básico proposto por [56], para cálculo do tempo médio de espera em fila numa estrutura de *polling* cíclica, com um único servidor e com disciplina de serviço *gated, L-limitado*.	
(3.17)	$\overline{W} = \frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K-\frac{N\lambda V}{L})}$
Modelo proposto neste trabalho, para cálculo do tempo médio de espera em fila numa estrutura de *polling* cíclica, com K servidores e com disciplina de serviço *gated/L-limitado*.	
(3.21)	$\overline{W}_p = \frac{(1+\rho^+_{p-1}+\rho^+_p)}{(1+\rho)} \times \left[\frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K-\frac{N\lambda V}{L})}\right]$
Modelo proposto neste trabalho, para cálculo do tempo médio de espera em fila numa estrutura de *polling* cíclica, com K servidores, com disciplina de serviço *gated/L-limitado* e com diferenciação por prioridades, baseada no usuário (*HOL*).	
(3.24)	$\overline{W} = \frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N-1)V}{2(1-\rho^K-\frac{N\lambda V}{L})}$
Modelo proposto neste trabalho, para cálculo do tempo médio de espera em fila numa estrutura de *polling* randômica, com K servidores e com disciplina de serviço *gated/L-limitado*.	
(3.25)	$\overline{W}_p = \frac{(1+\rho^+_{p-1}+\rho^+_p)}{(1+\rho)} \times \left[\frac{\overline{W}_{M/G/K}(1-\rho)}{(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N+K\rho)V}{2(1-\rho^K-\frac{N\lambda V}{L})} + \frac{(N-1)V}{2(1-\rho^K-\frac{N\lambda V}{L})}\right]$
Modelo proposto neste trabalho, para cálculo do tempo médio de espera em fila numa estrutura de *polling* randômica, com K servidores, com disciplina de serviço *gated/L-limitado* e com diferenciação por prioridades, baseada no usuário (*HOL*).	

Capítulo 4

Descrição do Simulador

A principal motivação em se criar os modelos de simulação foi a necessidade de validação dos resultados dos modelos analíticos propostos neste trabalho. Com este objetivo, foram desenvolvidos três modelos distintos, mas que contam com uma mesma arquitetura básica:

- Modelo multicanal com *polling* cíclico, sem prioridades.
- Modelo multicanal com *polling* randômico, sem prioridades.
- Modelo multicanal com *polling* cíclico, com prioridades.

Ao invés de se utilizar uma plataforma de simulação de rede já conhecida, como por exemplo o NS ou o OPNET, preferiu-se desenvolver um programa de simulação específico na linguagem Java.

A escolha pelo desenvolvimento se justifica pela simplicidade do sistema a ser simulado e também por causa da necessidade de modificações estruturais nos protocolos dos simuladores de mercado, que dificultariam o desenvolvimento.

A escolha pela linguagem Java se atribui sobretudo pelo fato ser operável em diferentes plataformas de *software/hardware* e também pela simplicidade trazida pela orientação a objeto.

O desenvolvimento do modelo de simulação deste trabalho foi baseado no fluxograma de um sistema de *polling*, proposto na referência [3]. A Figura 4.1, mostra o fluxograma citado.

4.1 Estrutura Básica

A estrutura básica dos programas de simulação foi desenvolvida de acordo com uma estrutura do tipo "Produtor x Consumidor". Nesta estrutura, são gerados eventos diferentes, dependendo de cada função no programa, que ficam

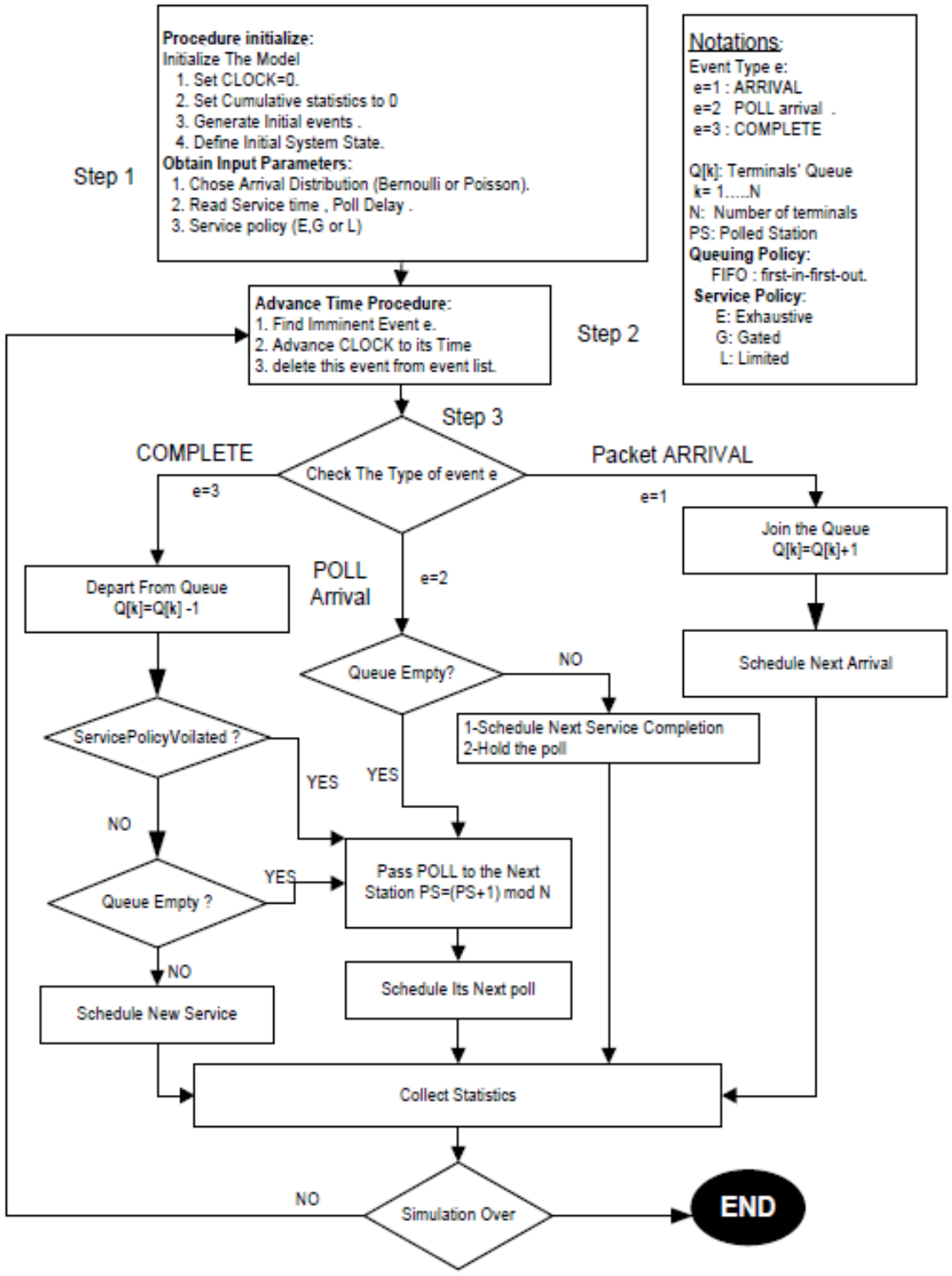

Figura 4.1: Fluxograma de simulação de um protocolo de *polling* proposto em [3].

armazenados numa fila. Os eventos são então processados, desencadeando ações correspondentes e depois são eliminados da fila de eventos. Por exemplo: os eventos de chegadas de usuários são transformados em usuários gravados nas filas do sistema. Estes usuários são atendidos e eliminados em seguida das filas de origem.

A estrutura básica dos programas de simulação pode ser melhor entendida através do diagrama de blocos da Figura 4.2.

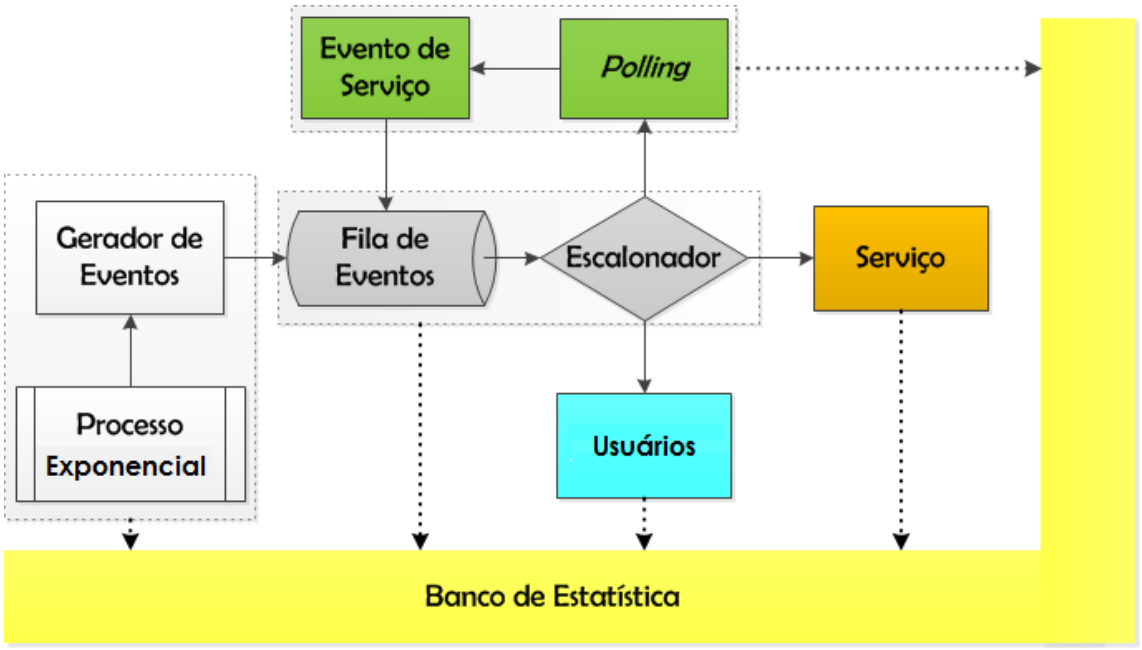

Figura 4.2: Blocos Funcionais da Simulação.

Na Figura 4.2 é possível diferenciar 9 blocos funcionais distintos, que por sua vez são agrupados em apenas 6:

- Gerador de Eventos (cor branca): responsável pela geração dos eventos de chegada de usuários e *polling*, de acordo com o Processo Exponencial (subprograma).
- Escalonador (cor cinza): é a parte central da estrutura. É responsável em receber os eventos na fila de eventos, analisar e processar, de acordo com cada tipo de evento.
- *Polling* (cor verde): o processo de *polling* é chamado pelo Escalonador e é responsável pela pesquisa de demanda (usuários aguardando para serem atendidos) em cada uma das filas do sistema, de acordo com a regra de serviço considerada (Cíclica ou Aleatória). Se o resultado de uma determinada pesquisa for positivo, o processo de serviço aloca um dos K servidores disponíveis (se houver) e gera um evento de serviço para atender a fila pesquisada.
- Serviço (cor laranja): este processo é disparado pelo processo de *polling*, quando verifica que a fila pesquisada possui usuários aguardando atendimento. O processo de serviço seleciona os usuários, de acordo com a disciplina de serviço utilizada, atende-os (apaga da fila) e libera o servidor no final.

- Usuários (cor azul): o processo de usuários é disparado pelo escalonador. É responsável em gerar as chegadas de usuários nas filas do sistema, de acordo com as características pré-definidas no evento original (prioridade).
- Banco de Estatística (cor amarela): todos os demais blocos enviam dados (setas pontilhadas) para o banco de estatísticas, que os consolida e gera as informações de interesse do trabalho.

Os detalhes de funcionamento de cada um destes blocos funcionais podem ser melhor visualizados na Figura 4.3.

4.2 Descrição das Classes

Foram definidas 14 classes para o programa. Estas classes foram agrupadas em 4 pacotes, de acordo com a função de cada uma delas.

- classes
  - Principal
- core
  - Canal
  - Escalonador
  - Evento
  - Gerador
  - GeradorChegada
  - GeradorPoll
  - Pacote
  - Servico
  - Terminal
- estat
  - Distribuicao
  - Estatistica
  - Exponencial
- util
  - FileUtils

4.2.1 Classe Principal

A Classe Principal funciona como a entrada à todas as outras classes e também funciona como "portal"para os parâmetros de simulação do programa. Os parâmetros de simulação são aquelas informações de entrada, que definem uma rodada de simulação. São eles:

- Quantidade de Filas: é o número total de terminais (N) que estarão compondo a rede. Esta informação é utilizada num *loop* para popular as N filas do sistema na simulação.
- Número de Visitas: corresponde ao número mínimo de visitas que fila deve receber do *polling*, até que a simulação possa ser finalizada. Quanto mais visitas cada fila receber, mais próxima da condição estacionária estará a simulação.
- Número de Servidores: corresponde à condição de operação da rede com múltiplos servidores (K), que serão utilizados para atender às filas.
- Limitação: é um parâmetro ligado à disciplina de serviço utilizada pelo protocolo de controle de acesso. Corresponde ao número máximo (L) de usuários que poderão ser atendidos na fila visitada.

O código da Classe Principal é bastante similar para os dois programas de simulação desenvolvidos (com e sem diferenciação por prioridades). A única diferença é que no caso do programa com processamento de prioridades, são iniciados 5 geradores de eventos de chegada de usuários independentes, um para cada classe de prioridade definida.

4.2.2 Classe Canal

A Classe Canal é exatamente a mesma para todos os modelos de simulação (com e sem prioridade). A classe Canal representa as características funcionais de cada um dos K servidores. Como cada servidor corresponde a um canal no sistema de comunicação modelado, foram definidas as seguintes características:

- Taxa de Serviço: corresponde ao valor de μ nos modelos analíticos propostos no Capítulo 3, que é justamente a taxa média de serviço de cada um dos K servidores do modelo.
- Estado: existem somente dois estados: livre e ocupado.

4.2.3 Classe Escalonador

Esta classe é o que se pode chamar de "unidade central de processamento"do código. A classe Escalonador controla o acionamento de quase todas as outras classes do programa (exceto as classes Principal e GeradorChegada), a partir da análise do tipo de evento lido da lista de eventos. Além de controlar a lista de eventos do programa, a classe Escalonador ainda insere os dados finais no banco de estatística.

A classe escalonador também é responsável pela configuração do sistema de filas, com base nos parâmetros de simulação, recebidos pela classe Principal, ex: número de filas, número de servidores, número máximo de usuários servidos por visita e o "fechamento do portão"(*gate*), característico da disciplina de serviço utilizada.

Pelo fato de ser um recurso compartilhado por diferentes classes, a lista de eventos (vetor de eventos) tem uma trava de sincronização de acesso, imposta para evitar eventuais travamentos e/ou corrupções nos dados.

Nesta classe também estão contidos os métodos de "escolha da próxima fila"(escolherTerminal), que são utilizados no processo de inserção de novos usuários nas filas (aleatória) e no processo de escolha da próxima fila a ser visitada (*polling*). Esta última pode ser em ordem cíclica (opção 1) ou aleatória (opção 2).

4.2.4 Classe Evento

A classe Evento é responsável pela configuração do evento que será inserido na fila de eventos. Os eventos que irão popular a lista de eventos possuem campos internos importantes, que irão determinar o seu processamento no programa. São eles:

- Tipo: define o tipo de evento dentro do código. São eles:
  - Chegada de Usuários
  - Chegada de *Polling*
  - Serviço
- Ocorrência: marca a hora que o evento foi criado na fila de eventos. Esta informação será utilizada para fins estatísticos.
- Prioridade: em se tratando de um evento de chegada de usuário, deve ser informada também a sua prioridade[1].

[1]Se o código for sem diferenciação por prioridades, todos os eventos de prioridade terão prioridade 1.

4.2.5 Classe GeradorChegada

Com o próprio nome indica, a classe GeradorChegada é responsável pela geração dos eventos de usuários na fila de eventos, cujo intervalo é calculado a partir da classe FileUtils (será detalhada posteriormente). A prioridade do usuário que será gerado na fila, pelo evento de chegada de usuários, é determinada antecipadamente na classe Principal.

4.2.6 Classe GeradorPoll

A classe GeradorPoll é responsável pela geração dos eventos de visita a cada uma das filas do sistema, de acordo com a regra de serviço definida (cíclica ou aleatória). O intervalo de geração entre cada evento de *polling* é fixo e compõe a lista de parâmetros de configuração do sistema de comunicação em simulação.

4.2.7 Classe Pacote

A classe Pacote[2] é responsável pela configuração do usuário no formato correto na fila do terminal. É possível observar que são configurados ao todo 4 campos no usuário que será gravado na fila:

- Prioridade
- *Bytes*: diz respeito ao tamanho da mensagem em *Bytes*, considerando o sistema de comunicação modelado.
- timeStamp: grava o instante de criação da usuário na fila.
- tsHystory: marcações de instantes importantes que serão utilizados para cálculos diversos no banco de estatística.

4.2.8 Classe Serviço

Depois da classe Escalonador, a classe Serviço é uma das mais importantes e complexas dentro do código. Esta classe é responsável pelo cumprimento da disciplina de serviço em cada terminal e pelo "consumo"(apagamento) da mensagem final da fila atendida.

Entenda-se como cumprimento da disciplina de serviço a seleção dos usuários que serão atendidos (consumidos), de acordo com o instante da sua gravação e o instante

[2]O nome "Pacote"da classe se dá por causa da configuração original do programa antigo, onde todos os usuários tinham o mesmo tamanho fixo. Não é o caso do simulador utilizado neste trabalho.

da "visita"do servidor; e da limitação do número máximo de usuários servidos por visita (L).

Nesta classe há uma pequena diferença entre os códigos com e sem diferenciação de prioridade. A diferença está basicamente na seleção dos usuários que serão atendidos. No caso do código sem prioridade, observa-se apenas a ordem de gravação dos usuários na fila ($FCFS$). Já no código com prioridade, observa-se antes a prioridade do usuário que está aguardando. A classe de prioridade mais alta é atendida primeiro que as outras classes (HOL) e dentro de uma mesma classe é que se observa a ordem de chegada dos usuários ($FCFS$).

A classe de Serviço também faz a liberação do servidor, depois de "consumidas"(transmitidas) os usuários selecionados da fila visitada.

O método de escolha do canal livre, que foi utilizado nos códigos, é crescente. Isto é, é feita uma varredura na lista de canais à procura do primeiro canal livre, iniciando-se pelo índice mais baixo. Este método de escolha de canal foi utilizado pela simplicidade e pelo escopo geral do código, que não contempla a análise de métodos de escolha mais elaborados (como o algorítimo proposto por [45]).

4.2.9 Classe Terminal

A classe Terminal é responsável pela configuração das N filas no sistema. No caso da simulação sem diferenciação de prioridade, cada fila recebe todas os usuários com a mesma prioridade (prioridade 1). No caso da simulação com diferenciação de prioridade, cada fila principal possui 5 filas internas, uma para cada classe de prioridade utilizada na simulação. Desta forma, já é possível concluir que os códigos com e sem prioridades são diferentes, apesar de apresentarem a mesma estrutura básica.

Na classe Terminal existem três verificações importantes para o funcionamento do código:

1. isServindo: informa se a fila está sendo atendida no momento da consulta.
2. temDemanda: informa se há usuários na(s) fila(s)[3] aguardando para serem atendidas.
3. verificaMinVisitas: informa se a fila já recebeu o número mínimo de visitas, especificado como parâmetro de entrada da simulação.

No código com diferenciação por prioridades está programado o método para simular a disciplina de serviço HOL, que transmite primeiro as mensagens de maior

[3]No caso do código com diferenciação por prioridades, cada fila principal tem 5 filas internas.

classe de prioridade; e a gravação da informação dos momentos das visitas a cada fila, que será utilizada na classe Serviço para "fechar o portão"para os usuários da fila (disciplina de serviço *gated*).

4.2.10 Classe Distribuição

A classe Distribuição abstrai uma distribuição de probabilidades qualquer. Neste código, esta classe é utilizada em conjunto com a classe Exponencial, para gerar a distribuição exponencial do intervalo entre os eventos de chegada de usuários e no tempo de serviço.

O código da classe Distribuição já inclui também um gerador de números aleatório entre 0 (inclusive) e 1 (inclusive), cuja semente é o resultado da função *time* no sistema operacional.

4.2.11 Classe Exponencial

Assim como a classe anterior, a classe Exponencial é idêntica para os três modelos de simulação.

O objetivo desta classe é a geração de uma variável aleatória exponencial, para ser utilizada na classe Distribuição, a fim de gerar uma distribuição de probabilidades exponencial, escolhida para representar o intervalo entre as chegadas de usuários e o tempo de serviço.

4.2.12 Classe Estatística

A classe Estatística apresenta algumas diferenças importantes entre o código com prioridades e sem prioridades. Como todas as outras classes anteriores, as diferenças residem basicamente na consolidação dos resultados finais, considerando as diferentes classe de prioridades, no caso do modelo com processamento de prioridades.

Esta classe consolida todos os dados colecionados, para gerar as informações, com os resultados de interesse do trabalho. O principal resultado é o tempo médio de espera em fila, que é mostrado separadamente, para cada uma das prioridades, no caso do modelo com prioridade.

4.3 Simulação

Neste trabalho foram feitas simulações do tempo médio de espera em fila para os três modelos analíticos referenciados abaixo, com 5 canais e com no máximo 5 usuários atendidos por ciclo:

1. Modelo de *polling* cíclico sem prioridades (3.17).

2. Modelo de *polling* randômico sem prioridades (3.24).

3. Modelo de *polling* cíclico com 5 classes de prioridade (3.21).

Para cada um dos modelos enumerados acima, foram obtidos valores do tempo médio de espera em fila, em função do número de filas no sistema (N).

O programa de execução das simulações (*shell script*) foi configurado para repetir a mesma simulação 200 vezes para cada valor de N (de 1 a 31).

Foram considerados três critérios para a definição do momento de parada de cada simulação:

- Esvaziamento da fila de eventos.
- Esvaziamento de todas as filas de todos os terminais da rede.
- Número mínimo de visitas em cada terminal da rede igual ou maior que 1000.

O escalonador para de funcionar, quando as três condições acima acontecerem simultaneamente.

O gerador de eventos de mensagens para de funcionar, quando o número mínimo de visitas em todos os terminais da rede é 1000.

O gerador de *polling* para de funcionar, quando o número mínimo de visitas em todos os terminais da rede é 1000; **e** quando não restarem mais usuários nas filas, para serem atendidos (apagados).

4.3.1 Definição dos Parâmetros

A maioria dos parâmetros de entrada no modelo de simulação são especificados com base na referência [1], na intensão de comparar os resultados de alguns cenários simulados.

- A limitação do número de filas no sistema (N) é em função da comparação dos modelos analíticos e de simulação. Com valores de N acima de determinados limites (dependendo do modelo), o sistema se aproxima da saturação ($\rho = 1$) e as simulações tendem a demorar excessivamente.
- Taxa de serviço ($\mu = 1375$) usuários/segundo do servidor. Baseado na taxa de transmissão máxima de $11Mbps$ do protocolo IEEE 802.11b e no tamanho médio de mensagem de 1000 *Bytes*, definido em [1].

- Taxa média total de chegada de usuários em cada fila ($\lambda = 250$ usuários/segundo). Perfaz um total de $3Mbps$ de entrada em cada fila. A escolha deste valor, relativamente alto, intenciona testar uma das principais vantagens do sistema de *polling*, na operação com alta carga de tráfego.

- Taxa média de chegada por prioridade, em cada fila do sistema ($\lambda_p = 50$ usuários/segundo). Perfaz um total por prioridade de $0,6Mbps$, não ultrapassando o limite ($\lambda = 250$ usuários/segundo). A escolha de taxas de chegadas (λ_p) iguais, para todas as classes de prioridade, facilita a análise da discriminação entre as classes.

- Número máximo de classes de prioridade ($P = 5$). As operadoras geralmente oferecem 5 níveis de diferenciação por prioridades nos seus produtos de comunicação de dados:

  - Voz (*priority queue*)
  - Dados críticos (sinalização de voz)
  - Vídeo
  - Sistemas corporativos (SAP, outras aplicações cliente x servidor)
  - Melhor esforço (*email*, *web*)

- Número máximo de usuários atendidos por fila em cada visita ($L = 1 - 5$). Esta variável, junto com o tamanho dos usuários, define um limite de tempo, em que um servidor permanece atendendo uma determinada fila do sistema. É mais ou menos representativo, dependendo do número de filas do cenário simulado.

- Número máximo de servidores do sistema ($K = 1 - 5$). Esta definição foi baseada na intensão de comparar os cenários com apenas 1 servidor/canal (IEEE 802.11b normal) e com 3 ou mais servidores, considerando os três canais ortogonais do IEEE 802.11b.

- O tempo de comutação de controle de uma fila para a próxima (V - *walk time*) de $139\mu s$, é baseado no tempo necessário para a troca de 3 mensagens de *polling* de 64 *Bytes* (*handshake*).

- O Coeficiente de variação do tempo de serviço (C_X) foi definido em 1, por causa da distribuição exponencial, utilizada na simulação.

4.3.2 Inferência Estatística

Segundo [38], a distribuição normal, conhecida também como distribuição gaussiana, é uma das mais importantes distribuições contínuas. Sua importância se deve principalmente devido ao teorema do limite central. Este teorema tem muita utilidade em aplicações práticas e teóricas, pois garante que a média dos dados converge para uma distribuição normal, à medida que o número de amostras aumenta, mesmo que estes dados não sejam distribuídos segundo uma normal.

Muitas pesquisas práticas têm como resultado uma distribuição normal. Por exemplo: a distribuição das medidas de altura de uma determinada população segue, em geral, uma distribuição normal.

Uma variável aleatória contínua qualquer X tem distribuição normal se sua função densidade de probabilidade $f(x)$ for dada por:

$$f(x) = \frac{1}{2\pi\sigma^2} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$x \in (-\infty, +\infty).$$

Pode-se utilizar a seguinte notação $X \sim N(\mu, \sigma^2)$.

Considerando que o número de amostras de cada simulação (A) foi fixado em 200 [4], utiliza-se a aproximação pela distribuição normal $N(0,1)$ para a inferência estatística do tempo médio de espera na fila estimado ($\overline{W_{E,N}}$).

Assim, para cada configuração de simulação com N variando de 1 a 31, calcula-se a média amostral da seguinte forma:

$$\overline{W}_{S,N} = \frac{\sum W_{R,N}}{A}$$

Onde $W_{R,N}$ é o tempo médio[5] de espera em fila em cada repetição de simulação.

O desvio padrão ($\sigma_{S,N}$) de cada simulação, para cada uma das configurações com N terminais, é calculado como:

$$\sigma_{S,N} = \sqrt{\frac{\sum(W_{R,N} - \overline{W}_{S,N})^2}{A}}$$

O tempo médio de espera em fila estimado, é finalmente calculado através da

[4] Optou-se por se fixar um número grande de repetições para facilitar o programa de execução das simulações, mesmo sabendo que demoraria mais tempo em cada simulação.

[5] Este valor é na verdade a média dos tempos de espera em fila de cada mensagem transmitida nos modelos de simulação.

seguinte expressão:

$$W_{E,N} = \overline{W}_{S,N} \pm Z \times \frac{\sigma_{S,N}}{\sqrt{A}} \tag{4.1}$$

O valor da variável de padronização da normal (Z) é calculado a partir do valor do nível de confiança, que foi definido em 95% para todas as simulações. Assim, utilizando a tabela Z para $0,475$ $(0,95/2)$, tem-se:

$$Z = 1,96$$

Assim, a Equação (4.1) pode ser finalmente escrita como:

$$W_{E,N} = \overline{W}_{S,N} \pm 1,96 \times \frac{\sigma_{S,N}}{\sqrt{200}} \tag{4.2}$$

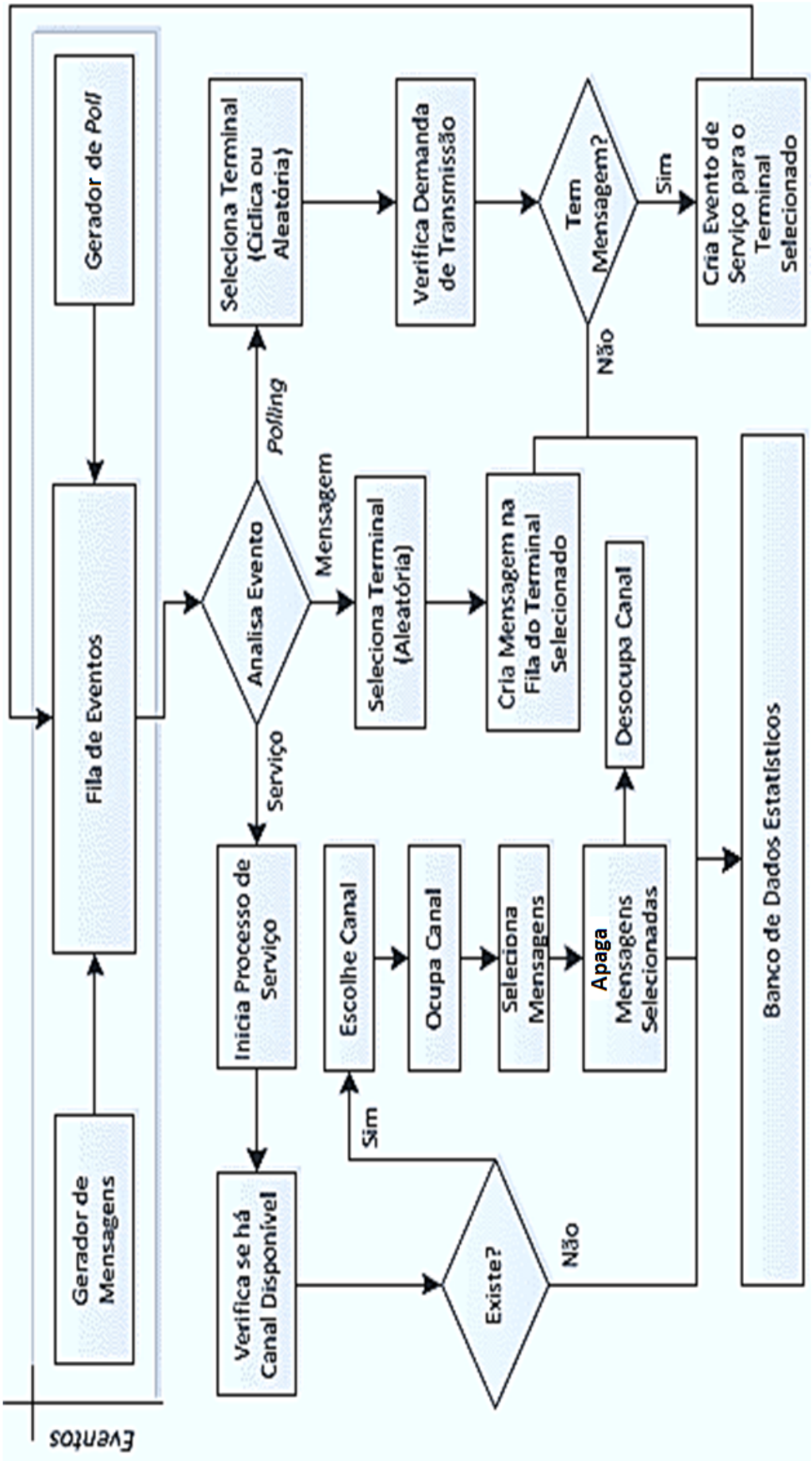

Figura 4.3: Diagrama funcional da estrutura de programação.

Capítulo 5

Análise dos Resultados

O objetivo deste capítulo é analisar os resultados numéricos, obtidos dos diversos cenários de testes nos modelos analíticos e de simulação, que foram propostos neste trabalho. A configuração dos cenários de testes foi direcionada para enfatizar as principais características operacionais do sistema modelado e projetar o seu comportamento em algumas situações assintóticas.

Os cenários iniciais de testes se voltaram para a validação dos pares de modelos analíticos e de simulação. Esta validação é feita graficamente, através da comparação mútua dos resultados dos pares; e depois comparando com os resultados de outros modelos similares, propostos na literatura[1].

Após a etapa de validação, são explorados outros cenários de testes, analisando o comportamento dos modelos em determinadas situações de interesse no sistema modelado. Alguns destes cenários são testados exclusivamente com modelos analíticos ou de simulação, dependendo de cada caso. Por exemplo: como os modelos analíticos foram desenvolvidos a partir da premissa de estabilidade, os resultados dos testes com o sistema saturado ($\rho \geq 1$) foram obtidos exclusivamente dos modelos de simulação.

5.1 Validação dos Resultados

O principal objetivo desta seção é a validação dos principais modelos propostos neste trabalho. A validação será feita em duas etapas:

1. Validação inicial, com base na comparação gráfica dos resultados numéricos obtidos nos modelos analíticos, com os resultados obtidos de cada dos seus respectivos pares de simulação.

[1]Neste caso a comparação foi feita com o modelo proposto em [1], que foi a principal motivação deste trabalho.

2. Validação final, comparando graficamente os resultados numéricos do modelo principal[2], com os resultados numéricos do modelo analítico proposto por [1].

5.1.1 Comparação entre os Modelos Sem Prioridades (Analítico e Simulação) Cíclico x Randômico

Nesta primeira análise, compara-se as curvas do tempo médio de espera em fila dos modelos cíclico (Equação (3.17)) e randômico (Equação (3.24)), com as respectivas curvas obtidas dos modelos de simulação.

Os parâmetros de entrada da Figura 5.1 são mostrados na Tabela 5.1. A escolha destes parâmetros é explicada na Seção 4.3.1.

Tabela 5.1: Parâmetros de entrada na Figura 5.1.

Parâmetro	`Valor`	`Descrição`
N	$1 - 27$	Número de terminais
K	5	Número de canais(servidores)
λ	250 mensagens/s	Taxa média de chegada de mensagens na fila
μ	1375 mensagens/s	Taxa média de serviço
L	5 mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço

Analisando as quatro curvas é possível perceber claramente o retardo adicional provocado pela componente de escolha randômica sobre o tempo médio de espera em fila. Pode-se perceber ainda que o valor deste retardo é aproximadamente o mesmo, nos modelos analíticos e de simulação, o que contribui na validação dos modelos propostos.

Observa-se também que as curvas dos modelos cíclicos (analítico x simulação) apresentam resultados mais próximos entre si, do que seus pares randômicos (analítico x simulação), principalmente quando a utilização do sistema se passa de $0,7$ ($N = 20$ e $\rho = 0,727$). A menor similaridade entre os modelos randômicos (analítico x simulação) ocorre principalmente por causa da aproximação adicional para a regra de serviço aleatória, onde é incluída uma nova componente randômica ao modelo com regra de serviço cíclica.

A partir de $\rho \geq 0,7$ ($N \geq 20$), nota-se um maior distanciamento entre as curvas, quando os valores dos modelos analíticos superam os valores de simulação. Este distanciamento é causado pela imprecisão da aproximação proposta em [33], que foi utilizada neste trabalho, para valores de utilização (ρ) mais altos[3].

[2]Considerado principal por que é a base para o desenvolvimento dos outros modelos propostos neste trabalho.

[3]Vide tabela comparativa na página 357 em [33].

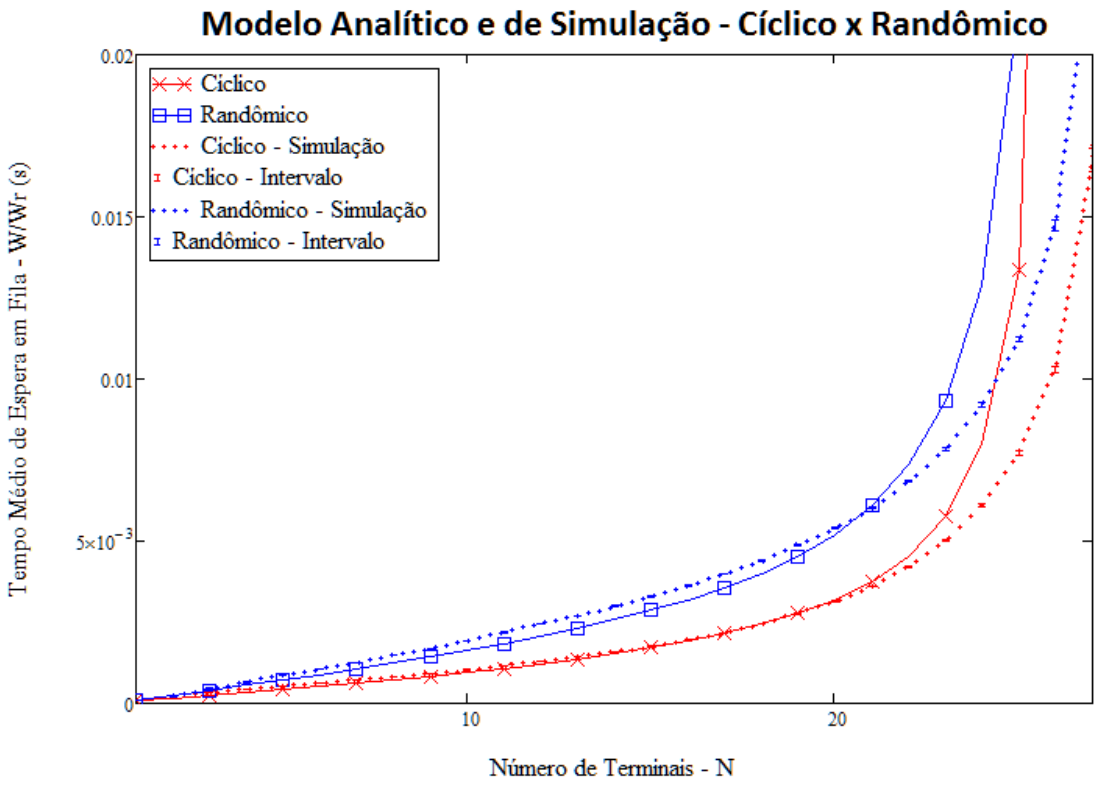

Figura 5.1: Comparação dos modelos analítico e simulação do tempo médio de espera em fila cíclico x randômico, Sem Prioridade - intervalo de confiança máximo de 2% e nível de confiança de 95%.

5.1.2 Comparação entre Modelos Analítico Cíclico Com Prioridades x Simulação

Neste cenário está sendo validado o resultado do modelo da Equação (3.21), que calcula o tempo médio de espera em fila para um sistema de *polling* cíclico, com disciplina de serviço do tipo *gated*/L-limitado, **com** diferenciação por prioridades.

Os parâmetros de entrada da Figura 5.2, tanto para o modelo analítico quanto para a simulação estão dispostos na Tabela 5.2.

Pode-se observar na Figura 5.2 os 5 diferentes níveis de prioridade, tanto para o modelo analítico (linhas contínuas), quanto para a simulação (linhas pontilhadas). Considerando somente as linhas contínuas (ou somente as pontilhadas), é possível notar ainda que as linhas de prioridades diferentes estão bastante próximas. Este é um comportamento esperado, uma vez que o mecanismo de diferenciação de prioridade, utilizado neste trabalho (*HOL*), começa a mostrar mais intensamente a sua eficiência quando o sistema está saturado ($\rho \geq 1$).

Tabela 5.2: Parâmetros de Entrada na Figura 5.2.

Parâmetro	`Valor`	`Descrição`
N	$1-27$	Número de terminais
K	5	Número de canais(servidores)
λ_1	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 1 na fila
λ_2	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 2 na fila
λ_3	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 3 na fila
λ_4	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 4 na fila
λ_5	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 5 na fila
μ	1375 mensagens/s	Taxa média de serviço
L	5 mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço
p	$1-5$	Classe de Prioridade da mensagem

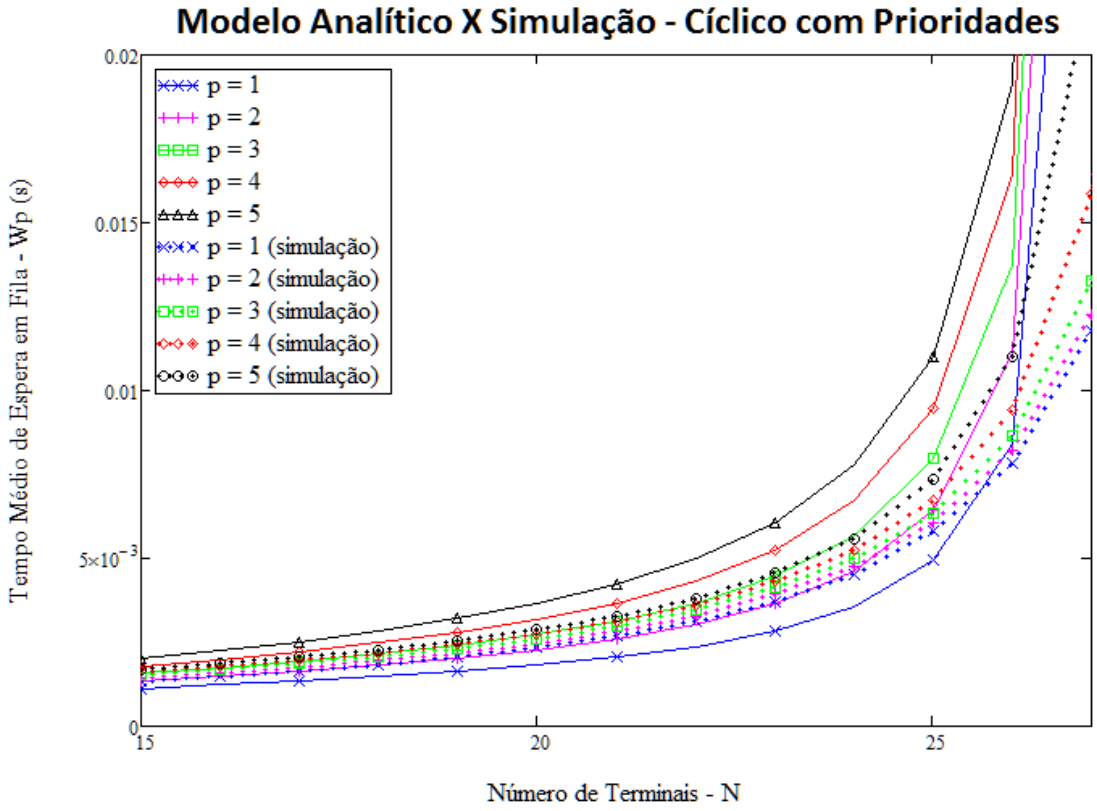

Figura 5.2: Comparação do modelo analítico x simulação - intervalo de confiança máximo de 5% e nível de confiança de 95%.

5.1.3 Comparação entre Modelos Analítico Cíclico Sem Prioridades x Modelo [1] x Simulação

O objetivo deste cenário de testes é aumentar o nível de validação dos modelos propostos neste trabalho, comparando os resultados obtidos dos mesmos, com os resultados de outro modelo, proposto na literatura. Para isto, os resultados da Equação (3.17) são simultaneamente comparados com os resultados obtidos do modelo de simulação (referência) e com os resultados de outro modelo similar,

proposto por [1][4]. Todos os modelos, considerados neste cenário, calculam o tempo médio de espera em fila para um sistema de *polling* cíclico, com disciplina de serviço do tipo *gated*/L-limitado, sem diferenciação por prioridades.

Considerando a curva dos resultados da simulação como referência, percebe-se que a curva dos resultados do modelo proposto em [1] não é tão próxima quanto a curva do modelo analítico proposto neste trabalho. Esta diferença ocorre pela aproximação para múltiplos servidores, proposta em [31] e que foi utilizada no modelo de [1].

Os parâmetros de entrada da Figura 5.3, tanto para os modelos analíticos quanto para a simulação estão dispostos na Tabela 5.3.

Tabela 5.3: Parâmetros de entrada na Figura 5.3.

Parâmetro	`Valor`	`Descrição`
N	$1-27$	Número de terminais
K	5	Número de canais(servidores)
λ	250 mensagens/s	Taxa média de chegada de mensagens na fila
μ	1375 mensagens/s	Taxa média de serviço
L	5 mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço

5.1.4 Comparação entre Modelos Analítico Cíclico Sem Prioridades x Modelo [1] x Modelo [57]

Neste último teste de validação, compara-se curvas do modelo cíclico sem prioridades, sem a limitação do número máximo de usuários servidos por visita (L). Isto é, a disciplina de serviço testada neste cenário é *gated* e não mais *gated*/L-limitado. Esta adaptação foi necessária para compatibilizar a análise das três curvas, incluindo os resultados do modelo proposto em [57], mostrando mais uma vez a flexibilidade dos modelos propostos neste trabalho.

Para adaptar o modelo proposto neste trabalho e o modelo proposto em [1] à disciplina *gated*, utilizada no modelo proposto em [57], retirou-se o último termo do denominador da Equação (3.17) ($N\lambda V/L$), que representa o acréscimo de tempo, em função da limitação do número máximo de usuários atendidos por visita.

Para adaptar o modelo de simulação à disciplina de serviço *gated*, foi retirada a condição de término de serviço na fila, quando o número máximo de usuários atendidos atinge o valor máximo (L).

[4]Conforme Seção 1.3, o trabalho proposto em [1] foi o principal motivador deste trabalho.

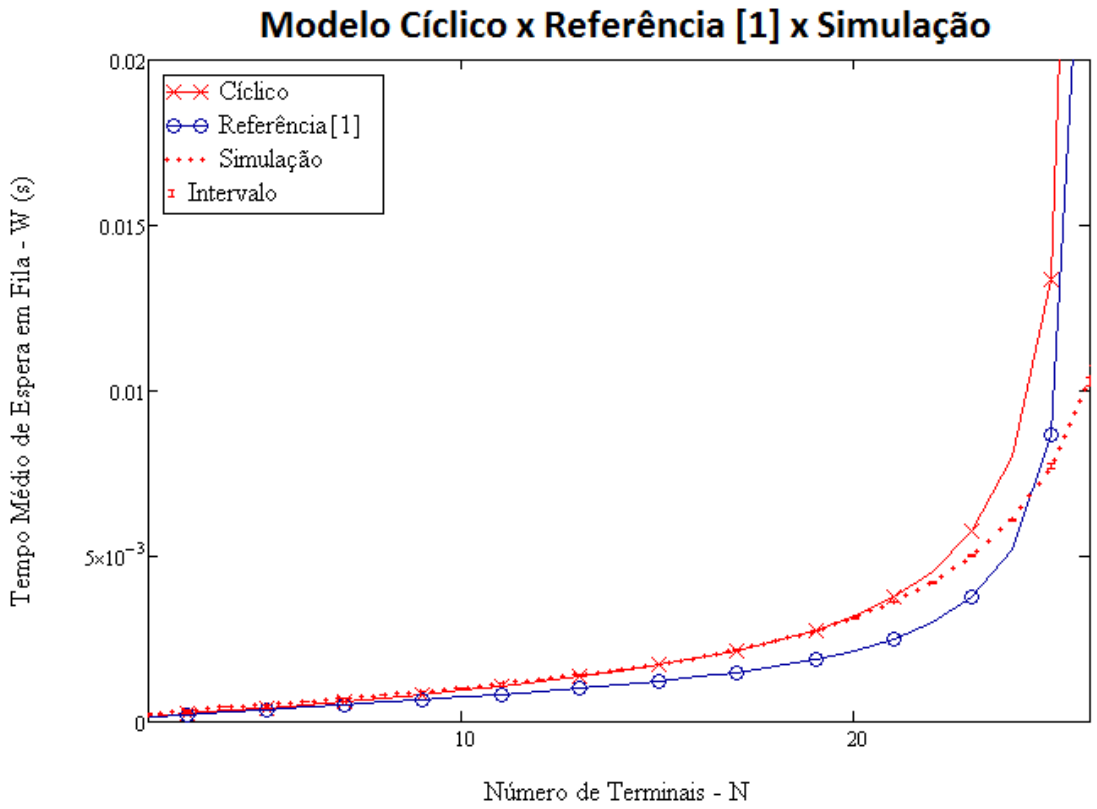

Figura 5.3: Comparação entre o modelo analítico cíclico x modelo analítico proposto por [1] x simulação, do tempo médio de espera em fila, Sem Prioridades - intervalo de confiança máximo de 2% e nível de confiança de 95%.

Concluídas as adaptações descritas acima, todas as curvas da Figura 5.4 passam a representar o tempo médio de espera na fila para a disciplina de serviço *gated*, de acordo com os parâmetros dispostos na Tabela 5.4.

Tabela 5.4: Parâmetros de entrada na Figura 5.4.

Parâmetro	`Valor`	`Descrição`
N	$1-30$	Número de terminais
K	5	Número de canais(servidores)
λ	250 mensagens/s	Taxa média de chegada de mensagens na fila
μ	1375 mensagens/s	Taxa média de serviço
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço

O modelo proposto em [57], utilizado nesta comparação, considera que somente 1 servidor pode atender uma determinada fila por visita, mesmo existindo mais servidores disponíveis ($1 \times K$).

Nesta comparação, é possível verificar que as curvas seguem relativamente próximas, até valores de $N = 20$, quando então a curva do modelo proposto em [57], cresce mais rapidamente que as demais.

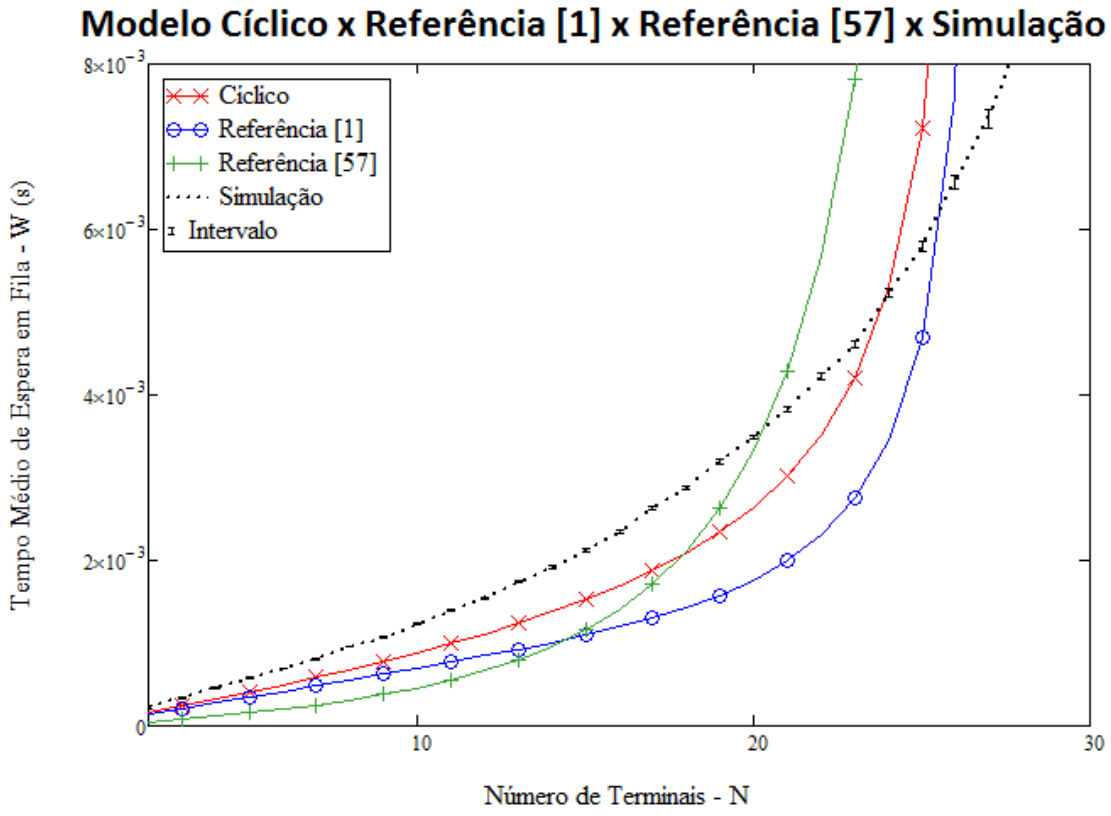

Figura 5.4: Comparação entre o modelo analítico cíclico (*gated*) x modelo analítico proposto em [1] (*gated*) x modelo analítico proposto em [57] ($1 \times K$) x simulação - intervalo de confiança máximo de 2% e nível de confiança de 95%.

Considerando a curva da simulação como referência (preta pontilhada), observa-se que a curva do modelo cíclico *gated* (vermelha), proposto neste trabalho, é a que mais se aproxima, considerando valores de N de 1 a 25 ($\rho = 0,909$).

5.2 Análise dos Modelos

Nesta seção são analisados outros cenários, explorando os modelos validados na seção anterior para caracterizar alguns comportamentos, de acordo com os interesses específicos deste trabalho, considerando as especificações do sistema modelado.

5.2.1 Simulação Cíclica Com Prioridades

Complementando a análise da Seção 5.1.2, esta seção mostra as curvas extraídas do modelo de simulação, com regra de serviço cíclica e **com** diferenciação por prioridades. Nas curvas apresentadas, o número de filas (N) prossegue aumentando, mesmo depois da saturação do sistema ($\rho \geq 1$)[5].

[5]Para ($N = 31$), a utilização ($\rho = 1,127$).

Neste cenário, foram utilizados os mesmos parâmetros de entrada da Tabela 5.2.

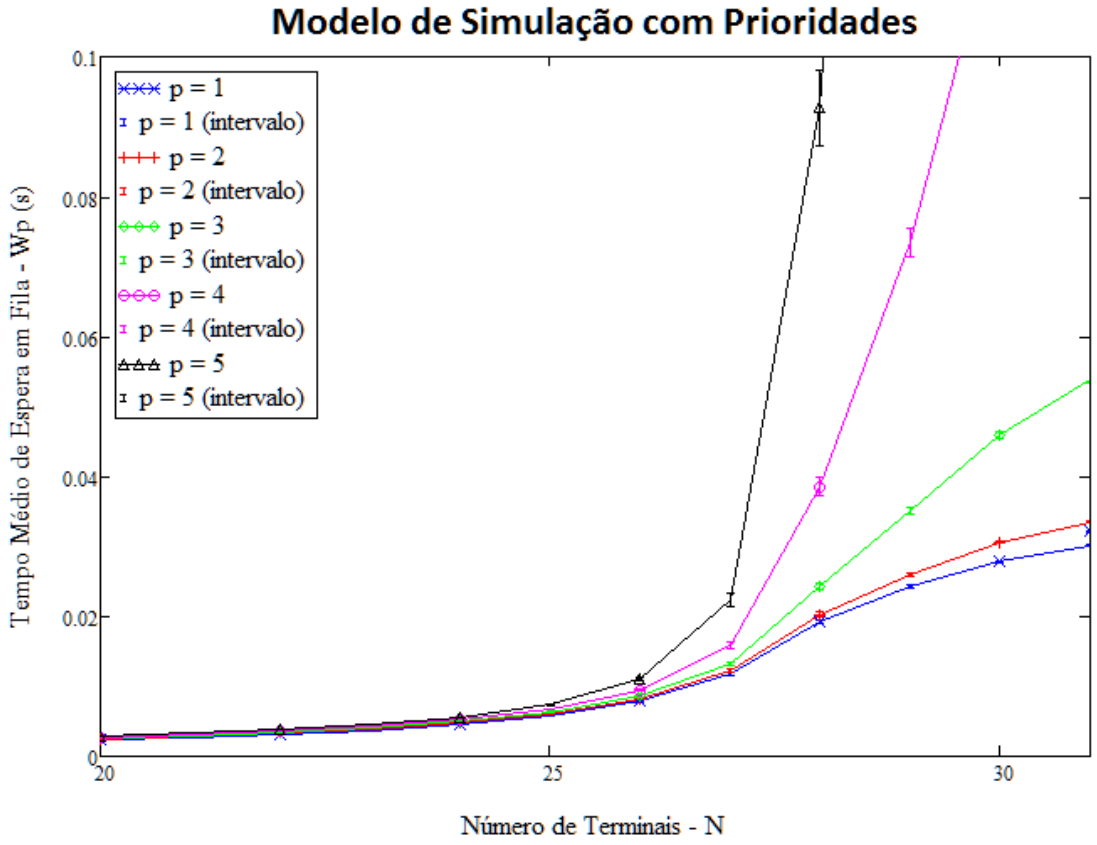

Figura 5.5: Modelo de simulação cíclico com prioridades - intervalo de confiança máximo de 5% e nível de confiança de 95%.

Observa-se claramente na Figura 5.5 que as curvas tendem a se distanciarem mais intensamente umas das outras, a partir de $N = 27$ e $\rho = 0,982$. Este distanciamento reflete o aumento da discriminação entre as diferentes classes, privilegiando as classes de prioridades mais altas ($p = 1$), em detrimento das classes de prioridades mais baixas ($p = 5$).

Como a disciplina de fila utilizada é *HOL - Head Of Line*, à medida que o número de filas no sistema (N) aumenta, aumenta também o tempo necessário para que o *polling* retorne à uma determinada fila recém atendida (ciclo), dando tempo suficiente para que os usuários de alta prioridade cheguem novamente na fila e obrigando aos usuários de baixas prioridades que aguardem mais.

Simulação Cíclica Com Prioridades - *Zoom*

A Figura 5.6 mostra com mais detalhes a parte mais condensada da Figura 5.5, que vai de $N = 1$ a $N = 19$. O objetivo desta figura é analisar se há sobreposição dos intervalos de confiança das curvas de diferentes prioridades.

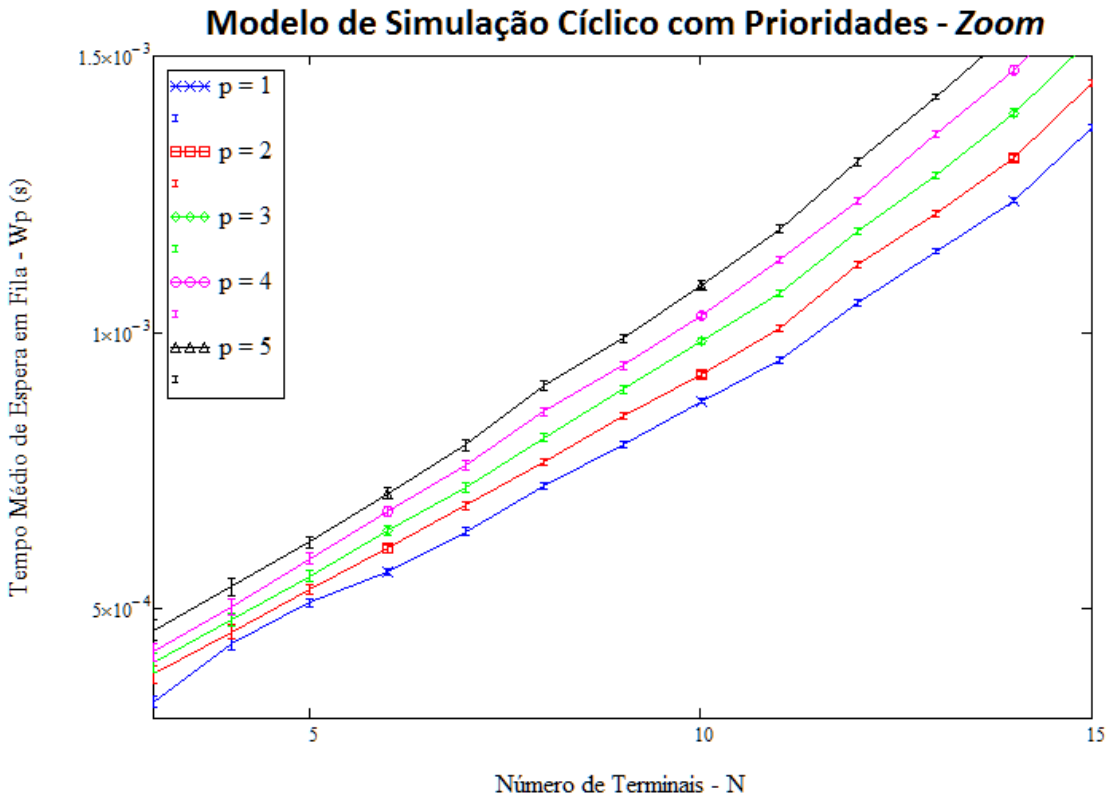

Figura 5.6: *Zoom* da Figura 5.5 de $N = 1$ até $N = 15$.

A sobreposição dos intervalos de confiança neste caso, indica que não há discriminação clara entre as classes de prioridade, considerando o intervalo de confiança.

Observando a Figura 5.6, é possível notar que até o valor de $(N = 6)$, praticamente não há discriminação entre as classes de prioridade.

Isto ocorre por que o tempo de ciclo é pequeno o suficiente para que sejam atendidos todos os usuários, de todas as prioridades na fila. Assim a diferença de tempo de espera em fila entre as classes diferentes é principalmente atribuída ao tempo de serviço dos usuários de mais alta prioridade, que serão atendidos antes.

5.2.2 Modelo Cíclico Sem Prioridades Variando o Número de Canais Disponíveis - K

O objetivo deste cenário é mostrar a influência do número de servidores (K), no tempo médio de espera no sistema de filas.

Na Figura 5.7 são apresentadas as curvas do modelo analítico da Equação (3.17), que mede o tempo médio de espera em fila para um sistema de *polling* cíclico, **sem** prioridades. Nesta equação o valor de K (número de servidores) varia de 1 a 5.

Os parâmetros de entrada da da Figura 5.7 estão dispostos na Tabela 5.5.

Tabela 5.5: Parâmetros de entrada na Figura 5.7.

Parâmetro	`Valor`	`Descrição`
N	$1 - 27$	Número de terminais
K	$1 - 5$	Número de canais(servidores)
λ	250 mensagens/s	Taxa média de chegada de mensagens na fila
μ	1375 mensagens/s	Taxa média de serviço
L	5 mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço

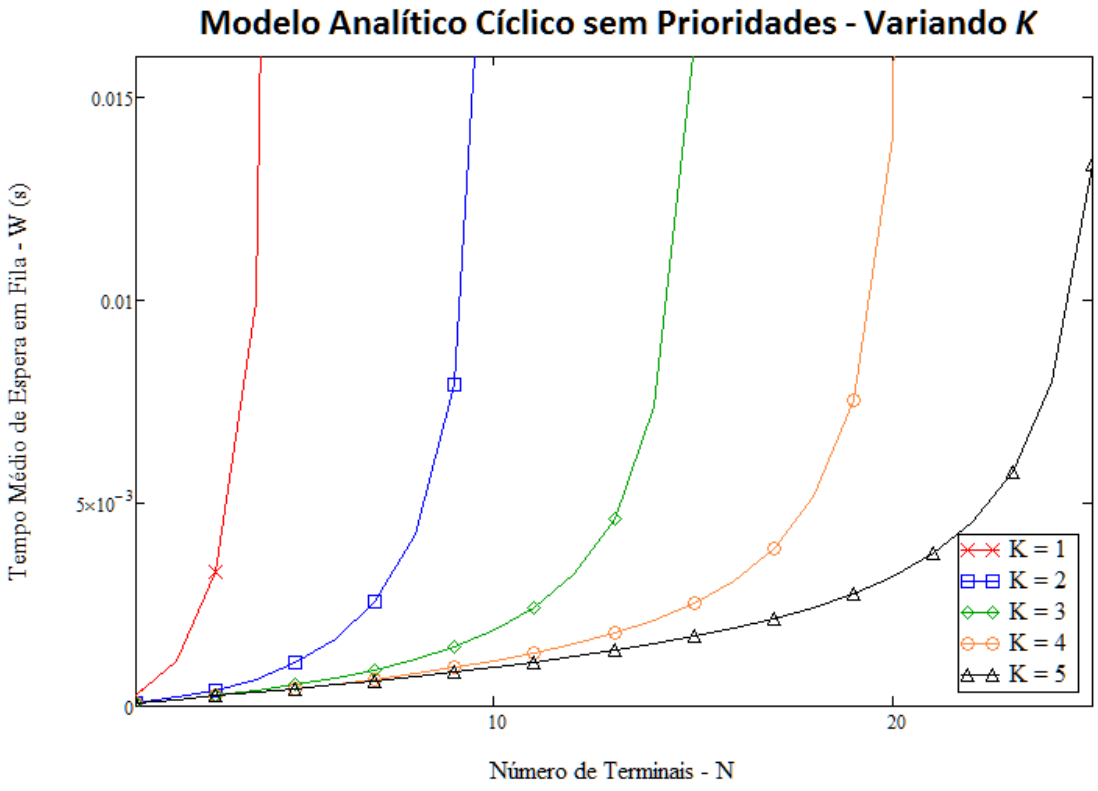

Figura 5.7: Tempo médio de espera em fila do modelo de simulação **Sem** Prioridade, variando-se K.

Analisando a Figura 5.7, é possível verificar que a variação no número de servidores (K) é bastante representativa no tempo médio de espera em fila. A cada 1 servidor incluído no sistema, consegue-se aumentar o número de filas (N) em aproximadamente em 5, mantendo-se os mesmos valores do tempo médio de espera em fila (aproximadamente).

5.2.3 Modelo Cíclico Sem Prioridades Variando o Número Máximo de Mensagens Servidas por Visita - L

Ao invés de se variar o número de servidores no sistema (K), esta seção analisa a variação do número **máximo** de usuários atendidos a cada "visita"do *polling* (L).

O estabelecimento de um determinado limite para o número de usuários servidos por visita (L), está relacionado com o nível de justiça do sistema, no compartilhamento dos recursos (servidores) (*fairness*). Ou seja, mesmo que uma determinada fila tenha mais mensagens para serem atendidas, no momento de uma visita, o servidor fica limitado a atender somente um número máximo de usuários (L), antes de ser liberado para o restante das filas.

O lado negativo desta disciplina de serviço, ocorre quando a rede é muito assimétrica. Isto é, quando as taxas de chegadas de usuários em cada fila forem muito diferentes. Nesta situação, o sistema pode ficar ocioso durante muito tempo, visitando uma determinada fila sem usuários suficientes ($Q_n < L$) no instante t, enquanto existem filas bem carregadas ($Q_n \geq L$) aguardando pelo *polling*. Onde Q_n é a quantidade de usuários aguardando numa determinada fila n, num instante t.

A tabela de parâmetros de entrada para a Figura 5.8 é a Tabela 5.6.

Tabela 5.6: Parâmetros de entrada na Figura 5.8.

Parâmetro	`Valor`	`Descrição`
N	$1-27$	Número de terminais
K	5	Número de canais(servidores)
λ	250 mensagens/s	Taxa média de chegada de mensagens na fila
μ	1375 mensagens/s	Taxa média de serviço
L	$1-5$ mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço

Analisando as curvas da Figura 5.8 é possível verificar que a alteração do número máximo de mensagens servidas (transmitidas) por fila (terminal) é menos significativa que a variação do número de servidores (K), principalmente para valores entre ($L = 3-5$). Entretanto, o distanciamento entre as curvas tende a aumentar à medida que aumenta também a taxa de utilização (ρ), com o aumento do número de terminais na rede (N).

5.2.4 Modelo Cíclico Sem Prioridades Variando o Coeficiente de Variação do Tempo de Serviço - C_X

O Coeficiente de variação do tempo de serviço (C_X) é uma grandeza adimensional, que mede a variabilidade dos dados em relação à sua média. Isto é, quanto menor for o coeficiente de variação de uma variável aleatória, mais homogênea é a mesma.

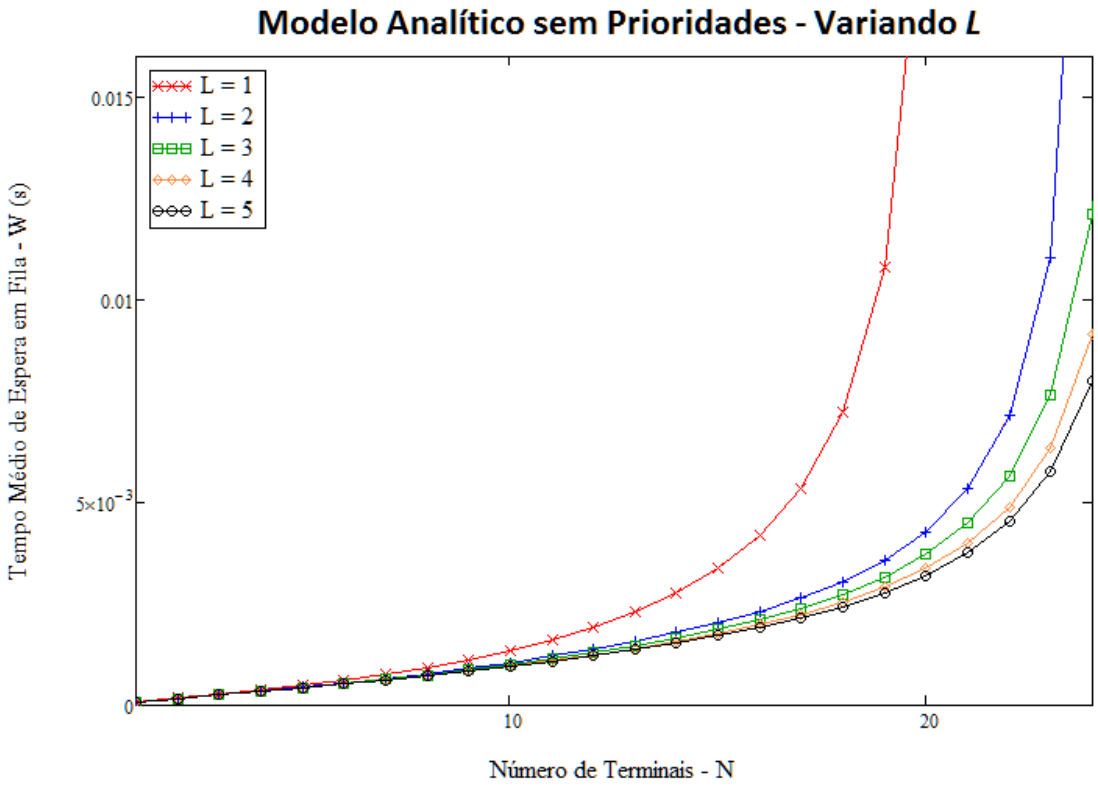

Figura 5.8: Tempo médio de espera em fila do modelo de simulação **Sem** Prioridade, variando-se K.

Em se tratando de tempo de serviço (ou tamanho de mensagem), o fato do coeficiente de variação ser elevado significa que existem usuários com tempo de serviço (X) muito maiores ou muito menores que o tempo médio de serviço ($\overline{X}$).

No atendimento aos usuários com tempo de serviço muito grande ($X_i >> \overline{X}$), o sistema entra em congestionamento, aumentando muito o tempo médio de serviço.

A tabela de parâmetros de entrada para a Figura 5.9 é a Tabela 5.7.

Tabela 5.7: Parâmetros de entrada na Figura 5.9.

Parâmetro	`Valor`	`Descrição`
N	$1 - 27$	Número de terminais
K	5	Número de canais(servidores)
λ	250 mensagens/s	Taxa média de chegada de mensagens na fila
μ	1375 mensagens/s	Taxa média de serviço
L	5 mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$0, \sqrt{1}, \sqrt{5}, \sqrt{25}$	Coeficiente de variação do tempo de serviço

Como pode ser observado nas curvas da Figura 5.9, à medida que o coeficiente de variação do tempo de serviço aumenta ($C_X^2 = 0, 1, 5, 25$), aumenta também o tempo médio de espera em fila.

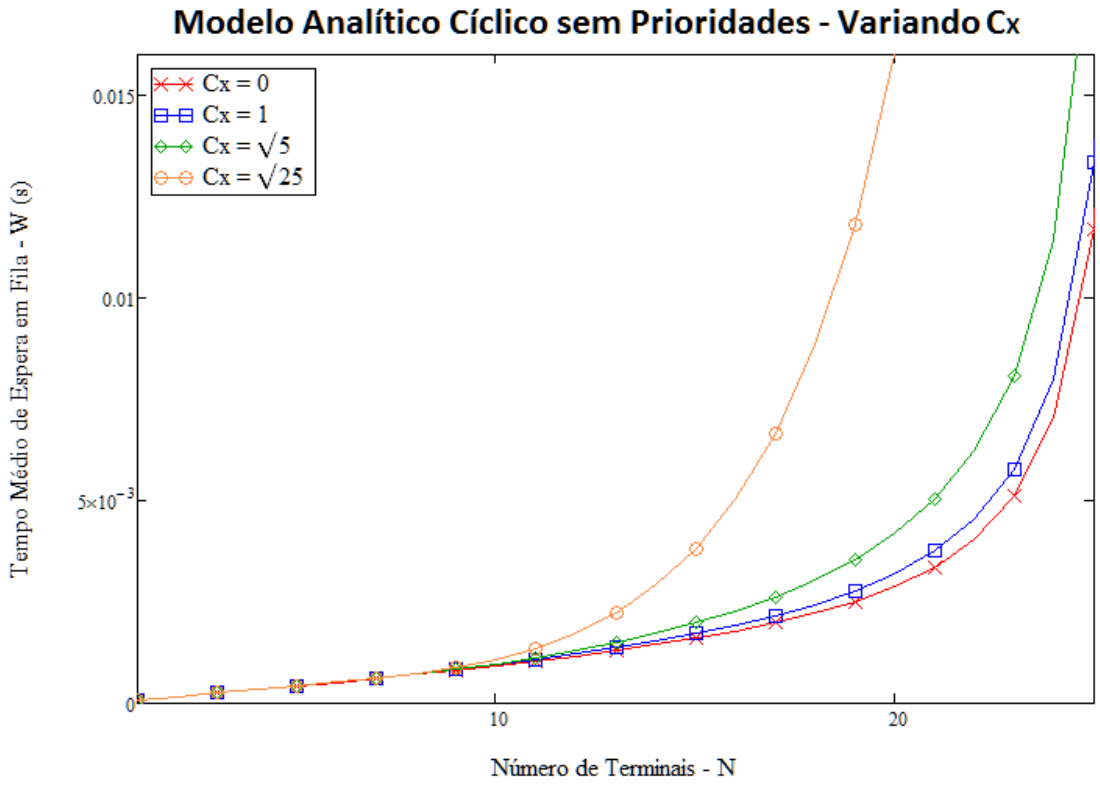

Figura 5.9: Tempo médio de espera em fila do modelo de simulação **Sem** Prioridade, variando-se C_X.

5.2.5 Modelo Cíclico Com Prioridades Variando a Taxa Média de Chegada de Mensagens de Prioridade 5 - λ_5

Nesta seção será testada a imunidade das classes de prioridade mais altas ($p = 1$) em relação ao aumento de tráfego das classes de prioridades mais baixas ($p = 5$). De acordo com o que já foi explicado na Seção 2.2.5, espera-se que o sistema com diferenciação por prioridades seja hierarquicamente independente no desempenho. Ou seja, variações na taxa de chegada de usuários de classes de prioridades mais baixas, não devem afetar o tempo médio de espera dos usuários de classes de prioridades mais altas.

Neste cenário de testes, a taxa de chegada de usuários de prioridade 5 (λ_5) será aumentada em relação às demais taxas. O resultado da nova curva do tempo de espera em fila dos usuários de prioridade 1 (W_1) será comparado com o resultado anterior de (W_1), plotado na Figura 5.2, onde as taxas de chegada de usuários de todas as classes de prioridade eram iguais a 50 mensagens/s.

A tabela de parâmetros de entrada para a Figura 5.10 é a Tabela 5.8.

Tabela 5.8: Parâmetros de entrada na Figura 5.10.

Parâmetro	`Valor`	`Descrição`
N	$1-27$	Número de terminais
K	5	Número de canais(servidores)
λ_1	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 1 na fila
λ_2	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 2 na fila
λ_3	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 3 na fila
λ_4	50 mensagens/s	Taxa média de chegada de mensagens de prioridade 4 na fila
λ_5	70 mensagens/s	Taxa média de chegada de mensagens de prioridade 5 na fila
μ	1375 mensagens/s	Taxa média de serviço
L	5 mensagens/visita	Número máximo de mensagens transmitidas por visita
C_X	$\sqrt{1}$	Coeficiente de variação do tempo de serviço
p	$1-5$	Classe de Prioridade da mensagem

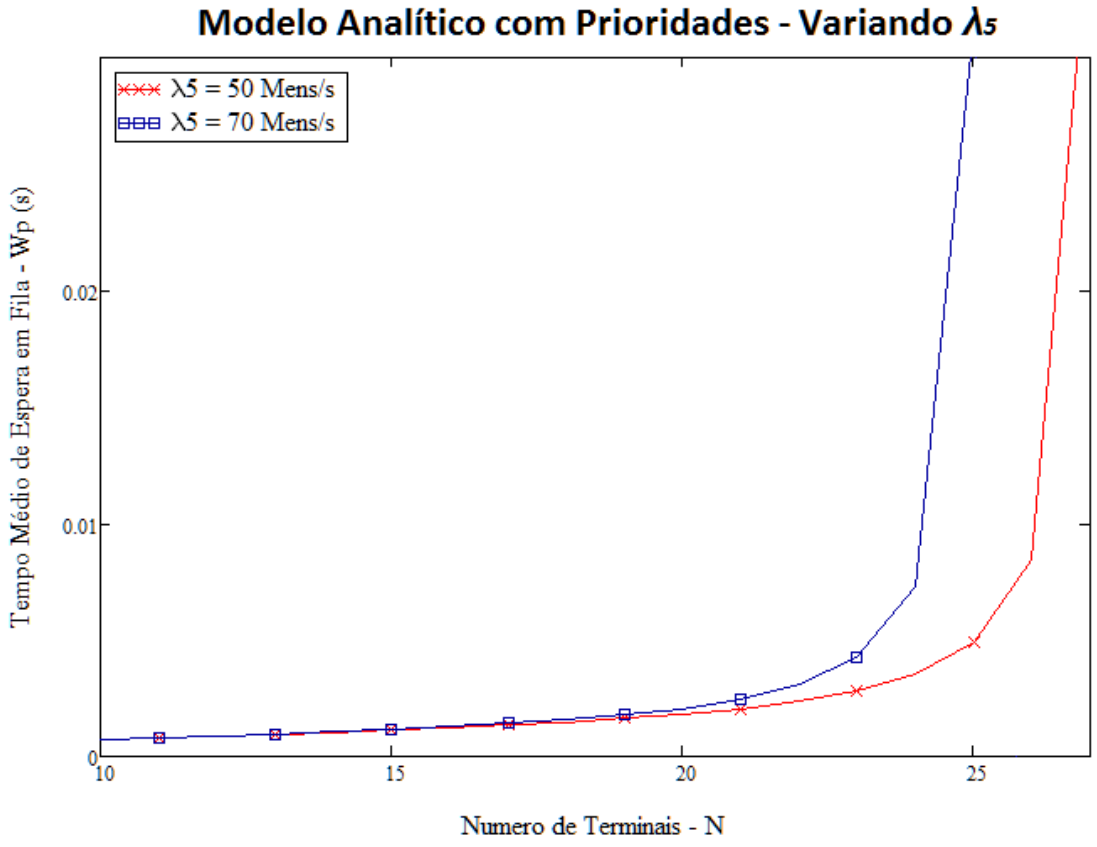

Figura 5.10: Comparação do Tempo Médio de Espera em Fila de Mensagens de Prioridades ($p = 1$), aumentando-se a taxa média de chegada de mensagens de prioridade ($p = 5$) de 50 mensagens/s para 70 mensagens/s.

Na Figura 5.10, pode-se notar uma diferença, aumentando o tempo médio de espera em fila (W_1) para a curva ($\lambda_5 = 70$). Esta variação ocorre em função da disciplina de interrupção utilizada neste trabalho, descrito na Seção 3.2.7, que é sem-interrupção (*nonpreemptive*)[6]. Desta forma, como a taxa de chegada de usuários de prioridade ($p = 5$) aumenta, aumenta também a probabilidade de um usuário de

[6]Relembrando que num sistema sem-interrupção (*nonpreemptive*), o serviço em curso não é interrompido quando da chegada de uma mensagem de mais alta prioridade.

prioridade alta ($p = 1$), encontrar um outro usuário de prioridade baixa ($p = 5$) em serviço, tendo que aguardar até o fim deste atendimento.

5.2.6 Modelos Cíclico Com e Sem Prioridades

Nesta seção é investigado o efeito da "Lei de Conservação do Trabalho", explicada na Seção 2.2.1. Apesar do sistema considerado neste trabalho não ser conservativo, pode ser verificado na Figura 5.11 que valores $N = 1$ até $N = 24$, o sistema se comporta como um sistema conservativo. Isto é, o privilégio dado para os usuários de classe de prioridade $p = 1$, é dado em função da depreciação dos usuários da classe $p = 5$.

Os parâmetros de entrada da Figura 5.11 são os mesmos da Tabela 5.2.

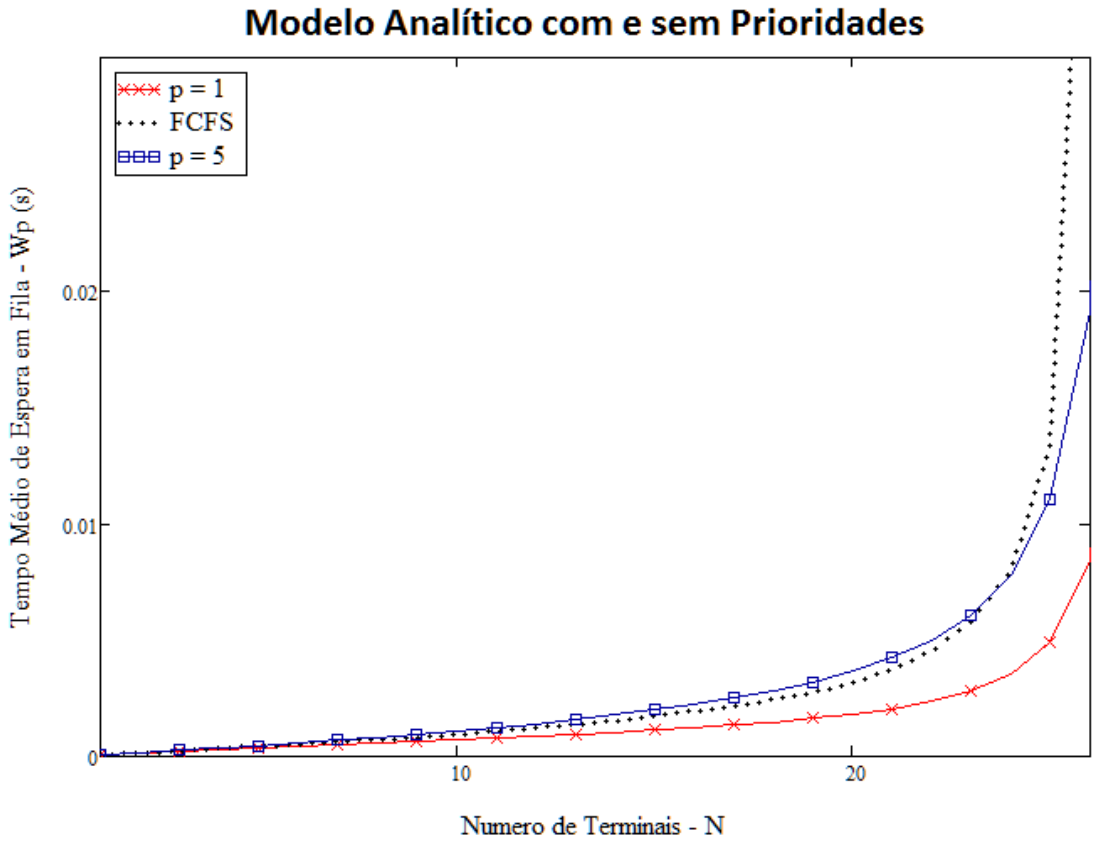

Figura 5.11: Tempo médio de espera em fila com e sem diferenciação por prioridades.

Analisando a Figura 5.11 em paralelo com o que foi explicado logo acima, é possível notar que o tempo médio de espera em fila, para os usuários de prioridade ($p = 5$), é um pouco mais alto que o tempo médio de espera em fila do modelo sem prioridades ($FCFS$), da Equação (3.17).

Esta pequena depreciação (elevação) no tempo médio de espera para usuários

de prioridade $p = 5$ é compensada pela grande diminuição do tempo médio de espera nas demais classes de prioridade, em especial para os usuários de classe de prioridade $p = 1$.

Capítulo 6

Conclusões

A principal conclusão deste trabalho se baseia na coerência e na precisão dos modelos propostos, em comparação com outros trabalhos existentes na literatura ([1] e [57]), referenciando-se aos resultados dos respectivos modelos de simulação implementados. Com isto, considerando a larga utilização dos sistemas de *polling* e a flexibilidade dos modelos propostos, conclui-se que estes modelos podem realmente ser utilizados na análise de outros protocolos de controle de acesso mais complexos, incluindo protocolos para redes sem fio.

Os primeiros cenários de testes tiveram como objetivo principal a validação dos modelos apresentados. Esta validação foi feita tomando-se os resultados dos modelos de simulação como referência e comparando-se os resultados numéricos obtidos dos modelos analíticos propostos, com resultados de outros modelos similares, propostos na literatura. Nesta análise foi possível observar inclusive que os modelos propostos neste trabalho apresentam resultados mais próximos dos modelos de simulação (referência), do que os resultados dos outros modelos testados.

A análise comprovou que os resultados dos modelos são bem próximos, principalmente para valores de utilização $\rho < 70\%$. A partir deste valor de utilização, nota-se um distanciamento, que vai aumentando mais rapidamente à medida que a utilização atinge valores próximos de 1.

Num dos cenários testados, com vários valores diferentes de K (número de canais/servidores), foi possível observar a eficiência da utilização simultânea de múltiplos canais, no aumento da vazão agregada do sistema, mantendo-se os níveis do tempo médio de espera em fila.

Os testes com diferentes classes de prioridades permitiram avaliar o comportamento do nível de discriminação entre as diferentes classes, em duas situações bem distintas: com $\rho < 1$ (estável) e com $\rho \geq 1$ (instável). Nesta análise, observou-se que o nível de discriminação do sistema aumenta muito com o aumento da carga do sistema, donde se pode concluir que o sistema de diferenciação de prioridades utilizado (HOL) não é tão eficiente em regime estável.

Destacam-se também os testes variando-se o número máximo de usuários servidos por ciclo (L). Apesar de não ser tão representativo no desempenho do sistema, quanto o variação do número de canais, o valor de L pode ser um parâmetro de ajuste fino, considerando as características de cada aplicação.

O restante dos cenários testados objetivaram avaliar o comportamento geral do sistema, em determinadas situações específicas, com base nos resultados teóricos previamente apresentados, por exemplo: teste de impacto do tráfego de baixa prioridade (λ_5) sobre o tráfego de alta prioridade (λ_1) e teste de conservação de trabalho, comparando os cenários com prioridades (HOL) e sem prioridades ($FCFS$) juntos. Os resultados apresentados permitiram concluir que os modelos propostos apresentam resultados coerentes com a teoria básica, que fundamentou o desenvolvimento dos modelos.

6.1 Trabalhos Futuros

A apresentação dos modelos analíticos e de simulação apresentados neste trabalho viabiliza uma eventual reanálise de alguns dos protocolos de múltiplos canais propostos, como por exemplo o protocolo proposto por [1], incorporando diferenciação por prioridades ou a implementação controle dinâmico para o número máximo de mensagens transmitidas por ciclo (L). O controle dinâmico poderia ser implementado à partir da informação da ocupação do *buffer* terminal no momento da visita e de informações estatísticas da rede como um todo, armazenadas ao longo do tempo. Esta nova funcionalidade poderia otimizar a utilização dos recursos da rede, diminuindo o tempo médio de espera em fila.

Um outro trabalho futuro, viabilizado com os resultados deste trabalho, seria a complementação da estratégia de diferenciação por prioridades, adicionando a diferenciação de prioridades por terminal[1]. Com a utilização da estratégia híbrida (terminal + tráfego), espera-se que a discriminação entre as classes de prioridade aumente, para valores de utilização (ρ) mais baixos.

[1]Neste trabalho é implementada somente diferenciação de prioridades por tráfego.

Referências Bibliográficas

[1] DE MORAES, L. F. M., NOBREGA, J. H. S. "Um Protocolo de Acesso ao Meio para Redes Locais sem Fio Infra-estruturadas utilizando Múltiplos Canais Ortogonais", *27º SBRC*, v. 1, n. 1, pp. 497–506, 2009.

[2] KÖPSEL, A., WOLISZ, J.-P. E. A., PIERRE EBERT, J., et al. "A Performance Comparison of Point and Distributed Coordination Function of an IEEE 802.11 WLAN in the Presence of Real-Time Requirements". 2000.

[3] ALQAHTANI, S. A. "Performance Modeling and Analysis of Distributed-Based Polling Networks", *IJCSI International Journal of Computer Science Issues*, v. 8, n. 2, pp. 57–62, Sep. 2011.

[4] KLEINROCK, L. *Queueing Systems*, v. I: Theory. Wiley Interscience, 1975.

[5] BURCHFIELD, R., NOURBAKHSH, E., DIX, J., et al. "RF in the Jungle: Effect of Environment Assumptions on Wireless Experiment Repeatability". In: *Proceedings of the 2009 IEEE International Conference on Communications*, ICC'09, pp. 4993–4998, 2009.

[6] SATYANARAYANAN, M. "Pervasive computing: vision and challenges." *IEEE Personal Commun.*, v. 8, n. 4, pp. 10–17, 2001.

[7] MARSAN, M. A., DE MORAES, L. F. M., DONATELLI, S., et al. "Analysis of Symmetric Nonexhaustive Polling with Multiple Servers". In: *INFOCOM*, pp. 284–295, 1990.

[8] HEART, F., MCKENZIE, A., MCQUILLIAN, J., et al. *ARPANET Completion Report*. Relatório técnico, Bolt, Beranek and Newman, Burlington, MA, 1978.

[9] KLEINROCK, L. "On Resource Sharing in a Distributed Communication Environment", *IEEE Communications Magazine*, v. 17, n. 1, pp. 27–34, January 1979.

[10] KLEINROCK, L., TOBAGI, F. "Packet Switching in Radio Channels: Part I–Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics", *Communications, IEEE Transactions on*, v. 23, n. 12, pp. 1400–1416, Dec. 1975.

[11] TOBAGI, F., KLEINROCK, L. "Packet Switching in Radio Channels: Part II–The Hidden Terminal Problem in Carrier Sense Multiple-Access and the Busy-Tone Solution", *Communications, IEEE Transactions on*, v. 23, n. 12, pp. 1417–1433, Dec. 1975.

[12] TOBAGI, F., KLEINROCK, L. "Packet Switching in Radio Channels: Part III–Polling and (Dynamic) Split-Channel Reservation Multiple Access", *Communications, IEEE Transactions on*, v. 24, n. 8, pp. 832–845, Aug. 1976.

[13] TOBAGI, F., KLEINROCK, L. "Packet Switching in Radio Channels: Part IV–Stability Considerations and Dynamic Control in Carrier Sense Multiple Access", *Communications, IEEE Transactions on*, v. 25, n. 10, pp. 1103–1119, Oct. 1977.

[14] KLEINROCK, L., SCHOLL, M. "Packet Switching in Radio Channels: New Conflict-Free Multiple Access Schemes", *Communications, IEEE Transactions on*, v. 28, n. 7, pp. 1015–1029, Jul. 1980.

[15] RUBINSTEIN, M., MORAES, I., CAMPISTA, M., et al. "A Survey on Wireless Ad Hoc Networks". In: Pujolle, G. (Ed.), *Mobile and Wireless Communication Networks*, v. 211, *IFIP The International Federation for Information Processing*, Springer US, pp. 1–33, 2006.

[16] JAYASURIYA, A. "Hidden vs exposed terminal problem in ad hoc networks". N. 0000022619, Australian Telecommunication Networks Applications Conference (ATNAC), pp. 52–59. ATNAC 2004, 2004.

[17] TOBAGI, F. "Multiaccess Protocols in Packet Communication Systems", *Communications, IEEE Transactions on*, v. 28, n. 4, pp. 468–488, Apr. 1980.

[18] VIEIRA, D. L. F. G. "IEEE 802.11". Dec. 2005.

[19] TANENBAUM, A. *Computer Networks*. Prentice Hall Professional Technical Reference, 2002. ISBN: 0130661023.

[20] TAKAGI, H., KLEINROCK, L. *A Tutorial on the Analysis of Polling Systems*. Relatório Técnico UCLA Report No. 850005, Computer Science Department, UCLA, February 1985.

[21] CAPETANAKIS, J. "Generalized TDMA: The Multi-Accessing Tree Protocol", *Communications, IEEE Transactions on*, v. 27, n. 10, pp. 1476–1484, Oct. 1979.

[22] NG, S., MARK, J. W. "A Multiaccess Model for Packet Switching with a Satellite Having Some Processing Capability", *Communications, IEEE Transactions on*, v. 25, n. 1, pp. 128–135, Jan. 1977.

[23] LAZOWSKA, E. D., ZAHORJAN, J., GRAHAM, G. S., et al. *Quantitative System Performance: Computer System Analysis Using Queueing Network Models*. Prentice-Hall, Inc., 1984.

[24] CHEN, J., SHEU, S.-T., YANG, C.-A. "A new multichannel access protocol for IEEE 802.11 ad hoc wireless LANs". In: *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, v. 3, pp. 2291–2296, Sep. 2003.

[25] FANG, X., HU, X., ZHANG, J., et al. "A Priority MAC Protocol for Ad Hoc Networks with Multiple Channels". In: *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, pp. 1–5, Sep. 2007.

[26] MARSAN, M., ROFFINELLA, D. "Multichannel Local Area Network Protocols", *Selected Areas in Communications, IEEE Journal on*, v. 1, n. 5, pp. 885–897, Nov. 1983.

[27] MARSAN, M., NERI, F. "A simulation study of delay in multichannel CSMA/CD protocols", *Communications, IEEE Transactions on*, v. 39, n. 11, pp. 1590–1603, Nov. 1991.

[28] WANG, M., CI, L., ZHAN, P., et al. "Multi-channel MAC Protocols in Wireless Ad Hoc and Sensor Networks". In: *Computing, Communication, Control, and Management, 2008. CCCM '08. ISECS International Colloquium on*, v. 2, pp. 562–566, Aug. 2008.

[29] ANTHONY, A., WATSON, H. "Techniques for developing analytic models", *IBM Systems Journal*, v. 11, n. 4, pp. 316–328, 1972.

[30] FUHRMANN, S. "Symmetric Queues Served in Cyclic Order", *Oper. Res. Lett.*, v. 4, n. 3, pp. 139–144, 1985.

[31] MARTIN, J. *Systems analysis for data transmission*. Prentice Hall series in automatic computation. Prentice-Hall, 1972.

[32] BOLCH, G., GREINER, S., DE MEER, H., et al. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Wiley-Interscience, 1998.

[33] KIMURA, T. "Approximatios for multi-server queues: system interpolations", *Queuing Systems*, , n. 17, Nov .

[34] KLEINROCK, L. *Queueing Systems: Volume 2: Computer Applications*. John Wiley & Sons New York, 1976.

[35] DE MORAES, L. F. *Priority scheduling in multiaccess communication*. Relatório Técnico RZ 1886, IBM. Zurich Research Laboratory (Zurich, CH), 1989.

[36] TAKAGI, H. *Queueing analysis : a foundation of performance evaluation, vol. 1 : vacation and priority systems, part 1*. Elsevier, 1991.

[37] KLEINROCK, L., LEVY, H. "The Analysis of Random Polling Systems", *Oper. Res.*, v. 36, n. 5, pp. 716–732, Sep. 1988.

[38] LEON-GARCIA, A. *Probability and random processes for electrical engineering*. Addison-Wesley, 1994.

[39] GUPTA, P., KUMAR, P. R. "The capacity of wireless networks", *IEEE TRANSACTIONS ON INFORMATION THEORY*, v. 46, n. 2, pp. 388–404, 2000.

[40] CHAFEKAR, D., ANIL KUMAR, V., MARATHE, M., et al. "Capacity of wireless networks under SINR interference constraints", *Wireless Networks*, v. 17, n. 7, pp. 1605–1624, 2011. ISSN: 1022-0038.

[41] ZHANG, J., JIA, X., ZHOU, Y. "Analysis of Capacity Improvement by Directional Antennas in Wireless Sensor Networks", *ACM Trans. Sen. Netw.*, v. 9, n. 1, pp. 3:1–3:25, Nov. 2012.

[42] WINTERS, J., SALZ, J., GITLIN, R. "The impact of antenna diversity on the capacity of wireless communication systems", *Communications, IEEE Transactions on*, v. 42, n. 234, pp. 1740–1751, Feb. 1994.

[43] GUMMADI, R., WETHERALL, D., GREENSTEIN, B., et al. "Understanding and Mitigating the Impact of RF Interference on 802.11 Networks". In: *Proceedings of the 2007 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '07, pp. 385–396. ACM, 2007.

[44] KYASANUR, P. N. *Multi-Channel Wireless Networks: Capacity and Protocols*. Relatório técnico, University of Illinois, 2001.

[45] NASIPURI, A., DAS, S. R. "Performance of multichannel wireless ad hoc networks", *International Journal of Wireless and Mobile Computing*, v. 1, pp. 191–203, 2006.

[46] TAKAGI, H. "Queuing Analysis of Polling Models", *ACM Comput. Surv.*, v. 20, n. 1, pp. 5–28, Mar. 1988.

[47] VAN DER MEI, R. "Towards a unifying theory on branching-type polling systems in heavy traffic", *Queueing Systems*, v. 57, n. 1, pp. 29–46, 2007.

[48] BORST, S., VAN DER MEI, R. "Waiting-time approximations for multiple-server polling systems", *Performance Evaluation*, v. 31, n. 3–4, pp. 163–182, 1998.

[49] BORST, S. "Polling systems with multiple coupled servers", *Queueing Systems*, v. 20, n. 3-4, pp. 369–393, 1995.

[50] BOXMA, O. J. "Workloads and Waiting Times in Single-server Systems with Multiple Customer Classes", *Queueing Syst. Theory Appl.*, v. 5, n. 1-3, pp. 185–214, Nov. 1989.

[51] TAKAGI, H., MURATA, M. "Queueing Analysis of Nonpreemptive Reservation Priority Discipline", *SIGMETRICS Perform. Eval. Rev.*, v. 14, n. 1, pp. 237–244, May 1986.

[52] VERISSIMO, F. C. A. *PROPOSTAS E AVALIAÇÕES DE PROTOCOLOS DE ACESSO ALTERNATIVOS AO PADRÃO IEEE 802.11E*. Tese de Mestrado, Universidade Federal do Rio de Janeiro, The address of the publisher, Oct. 2005. Ravel.

[53] BOON, M. A. A., ADAN, I. J. B. F. "Mixed gated/exhaustive service in a polling model with priorities." *Queueing Syst.*, v. 63, n. 1-4, pp. 383–399, 2009.

[54] SARKAR, M., NAGARAJ, S., BALSANIA, I. "A SINR based MAC layer protocol for multi-channel ad-hoc networks". In: *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International*, pp. 1889–1893, Jul. 2011.

[55] ADAN, I., RESING, J. *Queueing Theory: Ivo Adan and Jacques Resing*. Eindhoven University of Technology. Department of Mathematics and Computing Science, 2001.

[56] BERTSEKAS, D., GALLAGER, R. *Data Networks (2Nd Ed.)*. Prentice-Hall, Inc., 1992.

[57] MARSAN, M. A., DE MORAES, L. F. M., DONATELLI, S., et al. "Cycles and Waiting Times in Symmetric Exhaustive and Gated Multiserver Multiqueue Systems". In: *INFOCOM*, pp. 2315–2324, 1992.

[58] COSMETATOS, G. "Some Aproximate Equilibrium Results for the Multiserver Queue (M/G/r)", *Oper. Res. Quarterly*, v. 268, n. 271, pp. 615–620, 1976.

[59] GAVER, D. P., JACOBS, P. A. "On Inference Concerning Time-Dependent Queue Performance: The M/G/1 Example", *Queueing Syst.*, v. 6, n. 3, pp. 261–275, 1990.

Apêndice A

Teoria Básica de Filas

A.1 Modelo *M/G/1*

O cálculo do tempo médio de espera em fila do modelo *M/G/1* ($W_{M/G/1}$) leva em consideração dois fatores:

1. Tempo médio residual da mensagem encontrada em serviço (R_t), conforme Figura A.1.
2. Tempo médio de serviço das mensagens que serão servidas antes ($\overline{QX}$).

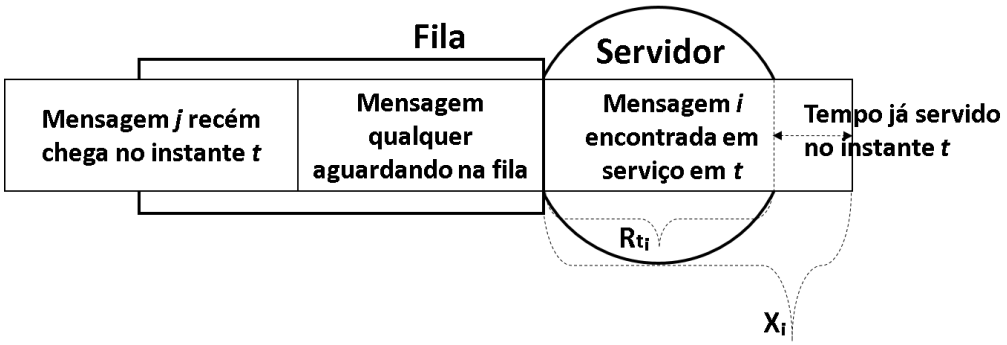

Figura A.1: Tempo residual da mensagem j encontrada em serviço na fila, observada no instante t da chegada da mensagem i

Onde $\overline{Q}$ é o número médio de mensagens numa fila *M/G/1* e $\overline{X}$ é o tempo médio de atendimento de cada usuário no sistema.

Juntando os dois fatores, pode-se escrever a expressão como:

$$W_{M/G/1} = R_t + \overline{QX}$$

Esta equação ainda pode ser rescrita como:[1]

[1]Considerando a aplicação da Lei de *Little* onde $Q = \lambda W$ e $\overline{X} = \frac{1}{\mu}$.

$$W_{M/G/1} = R_t + \rho W \tag{A.1}$$

De acordo com [56] o valor de R_t pode ser calculado como:

$$R_t = N\frac{\lambda\overline{X^2}}{2} \tag{A.2}$$

Juntando os resultados (A.1) com (A.2), obtém-se a equação do tempo médio de espera numa fila do tipo *M/G/1*:

$$W_{M/G/1} = \frac{N\lambda\overline{X^2}}{2(1-\rho)} \tag{A.3}$$

Ainda no livro de [56], é proposta uma discussão interessante a respeito da estabilidade de um sistema de fila $M/G/1$. Como pode ser observado na Equação A.3, o valor o tempo médio de espera ($W_{M/G/1}$) pode tender para infinito, mesmo se a utilização do sistema estiver abaixo de 1 ($\rho < 1$), se o segundo momento do tempo de serviço ($\overline{X^2}$) tender para infinito (∞). Isto pode acontecer no caso de existirem mensagens extremamente grandes, que demorarão muito tempo para serem servidas, causando grande enfileiramento das chegadas. Desta forma, a contribuição no tempo médio de espera em fila ($W_{M/G/1}$) é proporcional ao quadrado do tempo de serviço, quando se tem $\overline{X^2}$ tendendo para infinito (∞).

A.2 Modelo *M/G/1* com Férias

O modelo *M/G/1* com férias considera que o servidor fica indisponível para servir uma determinada fila durante um tempo aleatório, logo depois do período de serviço na mesma. Este período inativo pode ser comparado figurativamente com um período de "férias", daí se origina a designação deste modelo.

Nesta nova condição, um outro componente residual de tempo é adicionado no cálculo do tempo médio de atraso na fila *M/G/1*. Este novo componente, é calculado por [56] à partir da possibilidade de uma mensagem qualquer chegar numa das filas e encontrar o servidor **durante** um período de férias.

Este novo componente, chamado de tempo médio residual de férias (R_p), é calculado à partir da possibilidade de uma mensagem qualquer chegar numa das filas e encontrar o servidor **durante** um período de férias (V_n). Neste caso o valor total do tempo de espera em fila passa para:

$$W = W_{M/G/1} + R_p \tag{A.4}$$

O cálculo de R_p é pode ser melhor entendido à partir de Figura A.2 e também está demonstrado em [56].

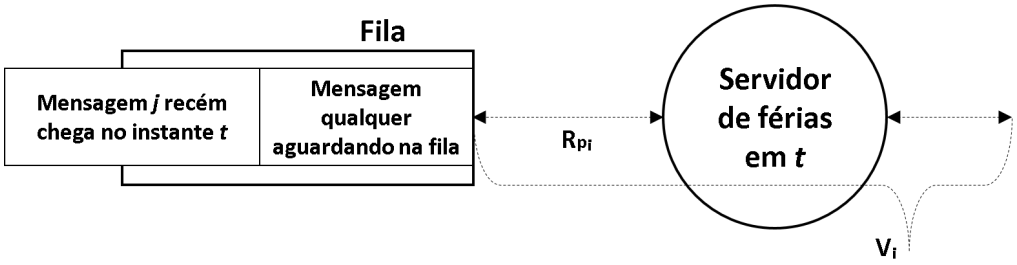

Figura A.2: Tempo residual de férias do servidor, observado no instante t da chegada da mensagem i

Por esta análise, o valor de tempo médio residual de férias (R_p) é dado por:

$$R_p = (1 - \rho)\frac{\overline{V^2}}{2\overline{V_n}} \tag{A.5}$$

Juntando este resultado com a Equação A.4, obtém-se a equação geral o tempo médio de espera em fila para o modelo *M/G/1* com "Férias"da forma:

$$W = W_{M/G/1} + \frac{\overline{V^2}}{2\overline{V_n}} \tag{A.6}$$

Onde $\overline{V_n}$

Apêndice B

Nomenclatura de Kendall

B.1 Modelo *M/G/K*

O termo *M/G/K*, de acordo com a nomenclatura de *Kendall* disponível em KLEINROCK [4], representa:

- *M* representa que o processo de chegadas de usuários na fila considerada obedece a uma distribuição de probabilidades de *Poisson* - (os intervalos de tempo entre as chegadas obedecem a uma distribuição exponencial).
  De acordo com MARTIN [31], a utilização do processo de *Poisson*, para representar a distribuição do número chegadas de mensagens num terminal, é bastante usual na modelagem de sistemas de comunicação, por causa da independência e da aleatoriedade dos eventos.

  Na distribuição de *Poisson*, a probabilidade de um determinado número de mensagens chegar num determinado terminal, durante um período de tempo, é alta. Ao invés disto, a probabilidade de uma chegada particular acontecer neste mesmo terminal, no mesmo período ou em qualquer outro período igual no futuro, é muito baixa. Ou seja, mesmo que a taxa chegadas por unidade de tempo seja alta, a probabilidade de uma chegada específica acontecer num segundo é bastante baixa e igual, independentemente do período de referência.

- *G* significa que o tempo de serviço de cada mensagem obedece a uma distribuição "Geral".
  Considerando que a rede de comunicação suportará diferentes tipos de aplicações, que geram diferentes tipos de tráfego, com tamanhos de mensagens bastante diferentes, é razoável não especificar uma determinada distribuição, como a exponencial por exemplo.

- *K* significa o número total de servidores que podem atender às filas. No cenário prático, representa o número total de canais ortogonais disponíveis na rede.