Uma extensão a um modelo de tráfego auto-similar com aplicações em controle de admissão e tarifação em redes ATM

Tese submetida para a obtenção do título de
Mestre em Ciências
no Departamento de Engenharia Elétrica
da COPPE/UFRJ
por
Jorge Roberto Mendes Filho

Uma extensão a um modelo de tráfego auto-similar com aplicações em controle de admissão e tarifação em redes ATM

Jorge Roberto Mendes Filho

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

Aprovada por :

_____

Prof. Luis Felipe Magalhães de Moraes, Ph.D.
(Presidente)

_____

Prof. Luiz Fernando Gomes Soares, Ph.D.

_____

Prof. Jorge Lopes Leão de Souza, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
JULHO DE 1998

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

Uma extensão a um modelo de tráfego auto-similar com aplicações em controle de admissão e tarifação em redes ATM

Jorge Roberto Mendes Filho

Julho de 1998

Orientador: Luis Felipe Magalhães de Moraes
Programa: Engenharia Elétrica

As Redes Digitais de Serviços Integrados estão tornando-se realidade. Como tais redes tem a finalidade de prover, dentro de uma única infra-estrutura, o suporte a serviços diferenciados como voz, vídeo de alta definição e dados o seu gerenciamento e operação requerem cuidados especiais. O entendimento dos padrões de tráfego, do controle de congestionamento exercido pela rede e de mecanismos mais eficientes de tarifação são alguns dos muitos tópicos que devem ser estudados e analisados para que se faça o uso racional da infra estrutura de rede. Dentro deste contexto, o presente trabalho apresenta uma extensão de um modelo de tráfego utilizado em redes e a partir daí verifica a importância da reserva de recursos para que se faça um aproveitamento racional destes na rede de comunicação e suas implicações nos preços que podem ser cobrados aos clientes ou usuários.

Capítulo 1

Introdução

1.1 Introdução

A rede RDSI-FL (Rede Digital de Serviços Integrados) foi idealizada para o suporte de serviços de diversos tipos com os mais diferentes requisitos de QoS (*Quality of Service).* Conceitualmente a RDSI-FL deverá suportar não somente as aplicações e serviços existentes atualmente mas também fornecer toda a infra-estrutura para o suporte de aplicações futuras. Logo o principal desafio, na implementação de tais redes, é integrar as atuais aplicações com as futuras, de uma maneira economicamente viável. Dentro deste contexto, serviços como interconexão de redes locais, transmissão de fax outros atualmente utilizados deverão conviver, na mesma rede de comunicação, com aplicaçãoes de HDTV e teleconferência (entre outros) sem que a qualidade de serviço de um desses seja prejudicado em conseqüência do uso de outro e, ao mesmo tempo, fazendo a utilização dos recursos da rede ao máximo, sem que haja desperdício.

Para que haja uma utilização racional da rede é necessário o total entendimento do tráfego a ser transportado por ela. Desta forma pode-se dimensionar a rede, bem como o controle que ela vai submeter ao tráfego oferecido, de modo que as expectativas dos usuários sejam atendidas de maneira satisfatória.

A tecnologia ATM *(Asynchronous Transfer Mode)* foi a escolhida como a tecnologia de implementação da RDSI-FL. Esta tecnologia visa suprir as necessidades das redes existentes (redes especializadas em transmissão de dados, redes telefônicas e redes de televisão) ao mesmo tempo tenta eliminar suas limitações, oferecendo apenas uma infra-estrutura. Basicamente as redes ATM possuem as seguintes características [2 e 62]:

- A informação é transmitida em pequenos pacotes, denominadas células. A flexibilidade necessária ao suporte de taxas variáveis de transmissão é oferecida pela transmissão do número necessário de células por unidade de tempo.
- O cabeçalho de cada célula contém dois campos denominados identificador de canal virtual e identificador de caminho virtual, denotando os endereços utilizados para roteamento e para multiplexação. A célula é comutada, no interior da rede, baseada nos valores destes dois campos de identificação, cujos valores são determinados durante a fase de estabelecimento da conexão. A função de comutação é implementada em *hardware*. A célula é a unidade de comutação e de transmissão.
- A transmissão de células na sub-rede é baseada em um protocolo sem controles de fluxo e de erros de dados. Apenas há a verificação de integridade dos dados de cabeçalho.

Uma das principais vantagens que as redes ATM oferecem é a possibilidade da integração de vários serviços em uma única rede. Contudo o uso racional dos recursos da rede deve ser sempre procurado. O uso racional da rede é conseqüência do entendimento completo do tráfego que ela suporta a da aplicação apropriada dos mecanismos de controle que rejem a maneira pela qual a rede deverá ser utilizada. Os controles devem ter poder sobre a aceitação ou rejeição de conexões, sobre o descarte de células durante períodos de congestionamento e sobre os incentivos fornecidos aos usuários da rede. Estes últimos podem ser implementados através de uma política de tarifação dos serviços ou dos recursos oferecidos pela rede.

A seguir passaremos a discutir alguns aspectos fundamentais relacionados à modelagem de fontes de tráfego, ao controle de admissão em redes ATM e a aspectos de tarifação em redes.

1.1 Modelagem de Fontes de Tráfego

A modelagem de fontes de tráfego tem por finalidade a criação de modelos matemáticos que capturam o comportamento estatístico da geração de informação de uma fonte de tráfego, tentando formar um padrão para cada fonte. Uma fonte de tráfego pode ser entendida como um dispositivo capaz de gerar tráfego (bits, células ATM, datagramas IP, etc...) a partir de um terminal que fornece um certo tipo de serviço que pode variar desde a emissão de um documento de FAX até a transmissão de televisão de alta definição.

O entendimento destes processos de geração de informação tem fundamental importância na análise de desempenho da rede. Através dos modelos pode-se estudar o impacto que a rede sofrerá quando receber o tráfego que é aproximado por tais modelos (um impacto que pode ser estudado é a probabilidade de perda de células por exemplo). Normalmente o estudo matemático deste impacto é fornecido pela Teoria de Filas. Para ser capaz de fornecer a (QoS) aos usuários da rede, é necessário a estimativa destes indicadores, como por exemplo a probabilidade de perda de células devido a transbordo de filas compartilhadas por vários usuários quando há interação entre o tráfego gerado e a rede. Um entendimento melhor deste problema é essencial para a interação com os controles de mais alto nível como admissão de conexões, roteamento de chamadas, alocação de posição em filas compartilhadas e de largura de faixa e de controle de congestionamento.

Quanto mais realísticos forem os modelos de fontes de tráfego mais precisas serão as estimativas sobre o comportamento da rede e com mais precisão poderá ser garantida a qualidade de serviço fornecida ao usuário. Outro ponto importante a ser colocado sobre a precisão dos modelos de tráfego é quanto ao grau de utilização da rede. O ideal para a rede é oferecer o maior número de conexões possíveis fazendo o uso racional de seus recursos, obtendo o maior grau de utilização possível. Quanto maior o conhecimento sobre o tráfego sendo oferecido à rede, maior será o conhecimento do impacto da variação de algum parâmetro de tráfego relevante sobre o comportamento da rede. Deste modo é possível conhecer a consequência provocada pelo aceite de uma nova conexão específica no desempenho geral da rede.

Ainda está em aberto o problema da modelagem do tráfego em redes de serviços integrados. Ainda é desconhecido, por exemplo, o impacto que um determinado tráfego (de vídeo por exemplo) provoca sobre um outro (de voz por exemplo). No caso, uma vez que consideramos redes que suportam vários tipos de tráfego, seria de suma importância o estudo deste impacto.

Inicialmente os estudos de modelagem partiam dos modelos de tráfego para que posteriormente fosse feita a adequação ao tráfego real [5 e 6 por exemplo]. Na maioria das vezes esta metodologia mostrou ser falha uma vez que alguns comportamentos do tráfego não eram capturados. Posteriormente, particularmente através do modelos de tráfego auto-similares, o tráfego real foi estudado e a partir daí estão sendo criados modelos que tentam capturar propriedades relevantes do tráfego real.

1.2 Controle de Admissão de Conexões

O controle do tráfego oferecido às redes de comunicação tem a função de controlar o uso da rede para evitar que esta tenha o seu funcionamento prejudicado. Em particular quando uma rede tem seus recursos utilizados por mais conexões (tráfego) que ela pode suportar os seus usuários observam a degradação dos serviços prestados. Logo, é necessário que se estabeleça mecanismos de controle de tráfego tal que a rede opere a níveis aceitáveis mesmo que por determinados períodos de tempo no qual a carga oferecida a ela ultrapasse a sua capacidade.

O controle de tráfego pode ser ainda dividido em duas partes : uma preventiva e outra reativa. A primeira tem por finalidade fazer o aceite ou a rejeição de uma nova conexão. Esta decisão está diretamente relacionada às características de tráfego das fontes geradoras, às características das fontes das conexões que já estão em progresso e aos recursos totais da rede. A segunda tem por finalidade manter o grau de qualidade de serviço negociados pelas conexões que já estão em andamento, fazendo o descarte ou a marcação de células para um possível descarte. O presente trabalho faz um estudo apenas da primeira parte.

Em redes telefônicas atuais (comutação de circuitos) este controle também existe porém, é mais simples do aqueles necessários em redes de pacotes que integram vários serviços. Esta simplicidade está relacionada à reserva estática que é feita para cada conexão que é aceita. Se aquela quantidade fixa de largura de faixa está disponível entre origem e destino a conexão será aceita, sem que haja o risco de que a admissão desta nova conexão interfire no desempenho das outras em progresso.

Em redes de pacotes que oferecem integração de serviços (mais especificamente redes ATM) a tarefa de admissão de conexões é mais complexa uma vez que a natureza aleatória do tráfego oferecido à rede não sugere uma reserva determinística e igualmente distribuída de recursos para cada conexão, quando se pensa em uso eficiente dos recursos da rede. Daí vem a necessidade de se conhecer o comportamento estatístico das fontes de tráfego. Dentro deste contexto existem duas questões que precisam ser respondidas [3] :

- Como determinar a quantidade de largura de faixa exigida por uma nova conexão;
- Como assegurar que a qualidade de serviço das conexões em progresso não será afetada quando uma nova for multiplexada juntamente com as anteriores.

Qualquer técnica projetada para a resolução destes problemas deve ser feita em tempo real e deve atentar para a maximização da utilização dos recursos da rede. O primeiro passo seria determinar o conjunto de parâmetros requeridos para descrever estatisticamente a atividade das fontes com a finalidade de se prever as métricas de desempenho da rede.

1.3 Tarifação em Redes Multiserviço

Com o aumento do uso comercial da Internet (anteriormente restrita ao meio acadêmico) e da tendência da integração de serviços em uma mesma rede, mecanismos de tarifação aplicados a tais cenários tornam-se tão importantes quanto os aspectos tecnológicos. Neste contexto os mecanismos de tarifação não devem ter como objetivo somente os lucros dos provedores e operadoras de redes de comunicação mas também devem ser pensados como uma maneira de indução aos usuários para que estes façam uso racional da rede. O uso racional da rede envolve, por parte dos usuários, as escolhas das qualidades de serviços contratadas apropriadas às aplicações que eles pretendem utilizar.

Para o estudo de mecanismos de tarifação, sempre que possível, faz-se uma associação à Internet como ela é hoje. Este paralelo pode ser útil pelo fato de que esta rede apresenta tedências de integração de serviços e que hoje podemos dizer que o mecanismo de tarifação aplicado hoje resume-se apenas à taxa nominal de acesso à rede (para circutos dedicados) ou de tempo de acesso para os usuários domésticos.

Podemos identificar hoje três deficiências básicas destes mecanismos de tarifação :

1) O usuário não é taxado pelo uso que faz dos recursos da rede. Deste modo podemos dizer que um usuário que tenha um circuito dedicado de 512 Kbps até um provedor Internet pode gerar mais tráfego que um outro que tenha um circuto de 2 Mbps durante um certo período de tempo. Da mesma forma um usuário conectado durante 15 minutos com o seu provedor pode gerar um tráfego maior que um usuário que fica conectado durante 30 minutos. Nestes dois casos, apesar de números menores, os usuários geram um maior congestionamento à rede e nem por isso pagam mais.

2) A informação (datagramas IP) são tratados da mesma forma. Em uma rede que se propõem a integrar serviços o tratamento dos diferentes tipos de tráfego deve ser diferente, o que hoje não acontece. Além disso, mesmo pagando mais, os datagramas gerados por um usuário são tratados da mesma forma que um usuário que paga menos, uma vez que não é utilizada nenhuma funcionalidade de prioridades nos datagramas (seja a nível de usuário ou de aplicação) , que faça a diferenciação dos tráfegos gerados.

3) Os usuários não são penalizados pelo congestionamento que geram. Desta forma os usuários não tem conhecimento do congestionamneto provocado aos outros usuários. Uma vez que os recursos são compartilhados, uma ação tomada por um usuário (por exemplo fazer a transferência de um grande arquivo) pode gerar a insatisfação de outros e nem por isso ele é penalizado por isso. No caso de uma rede descongestionada este efeito é quase desprezível mas a medida que a rede fica congestionada este efeito fica mais grave.

Em outras palavras, os usuários atualmente não pagam os custos do congestionamento (perda e retardo de pacotes). Um mecanismo de tarifação irá converter os retardos e os custos de enfileiramento em custos financeiros. Se os preços refletem os custos do fornecimento de serviços, eles irão forçar que os usuários comparem o valor do tráfego gerado aos custos que este tráfego gera ao sistema [63].

As três deficiências básicas indicadas anteriormente geraram algumas propostas de mecanismos de tarifação em redes. Podemos dividi-las em :

1) Tarifação Baseada no Uso (recurso utilizado);

2) Tarifação Baseada em Prioridades;

3) Tarifação de Custo Marginal .

1.4 Motivação

No que se refere a redes de alta velocidade o laboratório RAVEL (Redes de Alta Velocidade) visa o estudo de problemas relacionados ao desempenho de tais redes entre outros. Sendo assim, este trabalho tem os seguintes objetivos:

- estudo de modelos estatísticos de tráfego para proporcionar um melhor conhecimento do impacto provocado pelo tráfego à rede;

- estudo dos principais mecanismos de CAC utilizados em RDSI-FL;

- apresentação de uma extensão a um modelo de tráfego, avaliação do desempenho de um sistema de fila quando alimentada por tal extensão;

- comparação de duas aproximações assintóticas para o comportamento do sistema de fila com aplicações em CAC;

- estudo de alguns mecanismos de tarifação com aplicação a partir das aproximações obtidas no item anterior;

Esta tese está organizada em seis capítulos. No capítulo 2 faz-se um resumo dos principais modelos de fontes de tráfego apresentados na literatura especializada. Neste capítulo fontes de voz, dados e vídeo são estudas através de alguns modelos, uns mais recentemente utilizados para vídeo e dados, como a modelagem Auto-Similar e outros tradicionais, chamados Modelos Markovianos. No capítulo 3 faz-se um resumo sobre algumas propostas de reserva de recursos em redes ATM. Normalmente estas propostas podem ser mapeadas em problemas de sistemas de filas onde o tráfego de entrada é modelado estatisticamente, de acordo com a idéia apresentada no capítulo 2. Ainda, independentemente do modelo de tráfego, faz-se uma divisão de tais mecanismos em determinísticos e estatísticos, apresentando casos dos dois. No capítulo 4 será apresentada uma extensão de um modelo de tráfego Auto-similar com aplicações em Controle de Admissão de Conexões. Serão utilizadas duas aproximações para prever o comportamento da fila. No capítulo 5 identifica-se e estuda-se algumas características relavantes da Internet, no seu estágio atual, para os mecanismos de tarifação. Cita-se também alguns requisitos que devem ser levados em consideração no projeto destes mecanismos, partindo da idéia que a integração de serviços será possível na Internet. E por último apresenta-se alguns dos principais mecanismos existentes na literatura especializada e uma aplicação a partir da extensão do modelo de tráfego apresentada no capítulo 4.

Capítulo 2

Revisão dos Principais Modelos de Tráfego e de Algoritmos de Controle de Admissão de Conexões

O presente capítulo tem por finalidade fazer um estudo dos principais modelos de tráfego e controle de admissões em redes ATM. Os dois temas estão intimamente relacionados, uma vez que o controle aplicado na rede é função do que se pretende controlar. Deste modo, a modelagem estatística correta das fontes de tráfego leva ao controle de admissão eficiente, ou seja, àquele que reserva recursos de rede racionalmente, sem que haja desperdício deles. Serão vistos modelos utilizados em caracterização de fontes de voz, vídeo e modelos auto-similares (utilizados para modelagem de dados e de vídeo).

Será também discutida o controle de admissão de conexões (CAC) e a importância que a modelagem tem sobre ele. Serão vistos o modelo estatístico (através da aproximação da capacidade equivalente) e determinísticos. A aproximação da capacidade equivalente será utilizada nos capítulos 3 e 4 para comparação com outra, em termos de reserva de recursos de rede e tarifação respectivamente.

2.1 Introdução :

Os modelos de fontes de tráfego para RDSI-FL podem ser definidos como modelos matemáticos que capturam da melhor maneira possível o comportamento estatístico das fontes de tráfego. Através da modelagem de fontes de tráfego podemos levantar parâmetros essenciais que refletem o comportamento de uma ou várias fontes de tráfego quando estas alimentam um sistema de filas. Os resultados podem ser então utilizados para a previsão do impacto que ocorre em uma rede de comunicação que recebe tais fontes de tráfego.

Uma das grandes vantagens ofereciadas pelas redes ATM é a possibilidade de termos ganho de multiplexação estatística das fontes VBR. Neste tipo de rede os usuários geram um certo número de células necessárias para fazer a transferência da quantidade de informação que eles querem transferir. A quantidade de recursos de rede (por exemplo largura de faixa e espaço em dispositivos de armazenamento de células) requerida pelos usuários mudam constantemente em proproção ao número de células geradas na unidade de tempo. Quando os recursos são compartilhados entre os usuários, as quantidades requeridas pelos usuários individuais não necessariamente atingem o seu valor de pico simultaneamente, logo a rede pode reduzir a quantidade total de recursos requerida para uma determinada carga oferecida à rede (ou, da mesma forma, pode acomodar uma carga maior com a mesma quantidade de recursos). Este fenômeno é chamado de ganho de multiplexação estatística e é uma das principais características de redes ATM. Por exemplo, a estratégia de administração de um circuito virtual ou a medida de eficiência de uma codificação VBR podem ser calculadas (ou parametrizadas) em termos deste ganho. Uma vez que as fontes são caracterizadas corretamente é possível a obtenção de estimativas mais acuradas sobre o ganho de multiplexação estatística e consequentemente do grau de aproveitamento dos recursos da rede, importante informação da operadora de telecomunicações proprietária da rede.

Uma outra aplicação da modelagem de fontes de tráfego RDSI-FL está no controle de admissão de conexões. Podemos definir o controle de admissão como um conjunto de ações executadas pela rede na fase de tentativa de estabelecimento de uma conexão para determinar se esta nova conexão deve ser aceita ou não. A nova conexão será aceita se a rede de comunicação for capaz de suportar esta com um determinado grau de serviço (negociado na tentativa de estabelecimento), e não deteriorar o grau de serviço das

conexões já estabelecidas. Por Grau de Serviço ou Qualidade de Serviço (QoS- *Quality of Service*) podemos entender que seja o efeito coletivo do desempenho do serviço e que determina o grau de satisfação do usuário deste serviço. Alguns parâmetros de desempenho são :

- Taxa de perda de células (*CLR- Cell Loss Rate*) : Em se tratando de um ambiente de multiplexação estatística, as células provenientes de várias fontes de tráfego competem por recursos comuns limitados (espaço de buffer no equipamento multiplex). Consequentemente algumas células podem ser perdidas por não encontrarem espaço em buffer (espaço de armazenamento). Algumas modalidades de serviço podem tolerar um número moderado de perdas (como serviços de voz), enquanto outras são mais sensíveis à perda de informação (como serviços de transmissão de dados).

- Atraso, ou Retardo de transferência de células : Também, neste caso, podemos identificar serviços que são mais sensíveis ao atraso do que outros. Neste contexto podemos destacar os serviços de voz, uma vez que células deste tipo de serviço devem chegar ao destino dentro de um certo intervalo de tempo, caso contrário serão inúteis. Por outro lado podemos podemos destacar o serviço de transmissão de dados que são insensíveis ao atraso. O requisito de atraso restringe o tamanho máximo dos buffers.

- Variação do atraso da célula (*Cell Delay Variation*), jitter : Descreve a variabilidade do atraso de transferência de células. Quando células de várias conexões são multiplexadas, células de uma dada conexão podem ser atrasadas enquanto são inseridas células de uma outra conexão na frente das primeiras.

Como podemos concluir, apartir dos parágrafos acima, a caracterização analítica dos processos de geração de células devem ser o mais próximo possível da realidade para que o projeto de redes ATM atendam as espectativas dos usuários e das operadoras de telecomunicações. Quanto mais precisa a caracterização mais precisa serão as estimativas dos parâmetros de desempenho, e o comportamento da rede será mais previsível.

2.2 Critérios de seleção de modelos de fontes de tráfego :

A seleção de modelos de fontes de tráfego apropriados pode ser baseada em um conjunto de critérios sumarizados a seguir, de acordo com [1] :

- Proximidade com as fontes de tráfego reais : Um modelo para ser selecionado deve representar da melhor maneira possível a fonte real. A importância deste critério de seleção é mais que óbvia. As principais características estatísticas que influenciarão o comportamento da rede quando alimentada pelas fontes de tráfego devem ser bem representadas pelas estatísticas correspondentes dos modelos. Por exemplo, um importânte parâmetro para fontes de vídeo é a função autocorrelação do número de células geradas durante períodos de tempo sucessivos. Logo um bom modelo para fonte de vídeo deve produzir uma função autocorrelação que aproxime satisfatoriamente uma medida feita na fonte de vídeo real.

- Generalidade : O modelo deve ser o mais geral possível. O ideal seria que o modelo aproximasse uma grande variedade de serviços (como voz, transferência de dados, etc...) , mudando apenas alguns parâmetros dentro do modelo para atingir a todo um espectro de características que existem entre os diferentes serviços.

- Simplicidade : O modelo deve ser descrito por um pequeno número de parâmetros. Além disso estes parâmetros devem ser representativos dos fenômenos físicos de uma maneira mais intuitiva possível com por exemplo a taxa média de geração de bits de uma fonte.

- Facilidade de aproximar a fonte real : O modelo pode representar a fonte real por uma certa seleção de parâmetros que compoem o modelo. Isto é realizado normalmente expressando os momentos de certas variáveis aleatórias relacionadas ao modelo em termos de seus parâmetros, fazendo com que estes parâmetros assumam valores tais que os valores dos momentos resultantes aproximem-se o máximo possível dos momentos do experimento relacionado.

- Tratamento analítico e acurácia : Um modelo ideal deve ser fácil de ser analisado. Por exemplo quando analisamos fontes de tráfego multiplexadas (normalmente para estimar a taxa com que células são perdidas), a análise exata normalmente é muito complexa de ser feita, a partir das características individuais das fontes. Logo o modelo utilizado para representar o tráfego agregado deverá ser acurado e tratável analiticamente. Para outras estimativas de desempenho (como por exemplo o retardo médio experimentado por uma célula que chega ao multiplex no mesmo sistema) estas considerações devem ser as mesmas porém o modelo utilizado normalmente é diferente.

- Facilidade de implementação : O modelo deve ser fácil de ser implementado em experimentos, seja através de simulação computacional seja através de geradores de tráfego baseados em hardware.

- Adequação para a modelagem tráfego agregado e/ou tráfego de saída : Normalmente é interessante modelar o tráfego agregado de várias fontes visando a análise do ganho de multiplexação estatística (para estudo de economia de largura de faixa e "buffers"). Também é necessária a adequação dos modelos que tentam representar o tráfego emergem como saída de outros links. Isto tem grande importância quando estamos analisando redes com vários nós comutadores. Quanto mais adequados estes modelos mais precisa será a estimativa do comportamento da rede, informação fundamental para aqueles que operam a rede.

2.3 Classificação de tráfego em redes ATM :

As redes ATM estão sendo projetadas para o provimento de um grande número de serviços com com características estatísticas variadas e com diferentes requesitos de qualidade de serviço. Com o propósito de classificar as diversas classes de serviço podemos dividi-las em:

- Seviço de taxa constante (CBR - *Constant Bit Rate*) : geram tráfego a uma taxa constante. Foi idealizado para suportar aplicações com exigências definidas de retardo ponto a ponto.

- Serviço de taxa variável (VBR - *Variable Bit Rate*) : geram tráfego a uma taxa variável. Também foi idealizado para suportar aplicações com exigências definidas de retardo ponto a ponto, taxa de perda de células e variação de jitter.

- Serviço de taxa indefinida (UBR - *Undefined Bit Rate*) : foi definido para suportar tráfego que não possui exigência definida de serviço e aceitar a qualidade de serviço oferecida pela rede.

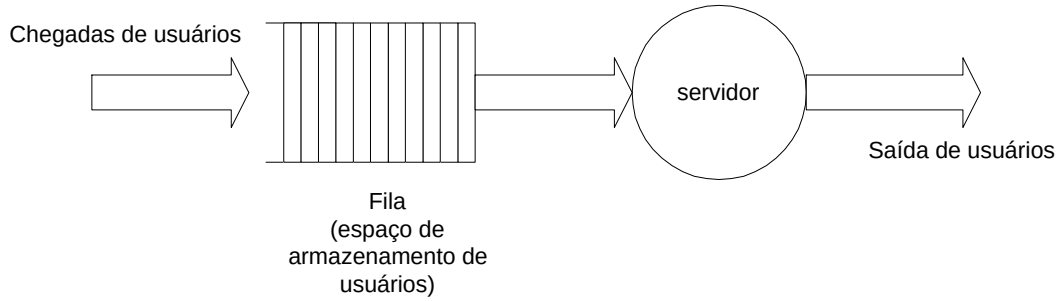
- Serviço de taxa disponível (ABR - *Available Bit Rate*) : foi definido para suportar aplicações com exigências mínimas de vazão e retardo na transferência de células.

O presente trabalho concentra-se no estudo do tráfego VBR, onde encontram-se os maiores problemas de caracterização estatística do tráfego, garantia da qualidade de serviço negociada entre rede e usuário e permite o uso racional dos recursos oferecidos pela rede.

A demanda precisa para comunicações em largura de faixa larga permanece desconhecida e alguns serviços ainda não imaginados poderão tornar-se reais somente quando a infra-estrutura para o transporte e manutanção de tais aplicações estiver operacional. Apesar da dificuldade de projetar controles adaptáveis a todos os serviços possíveis, é desejável que encontremos soluções de controle para os serviços já identificados.

Sistema de Fila com servidor único :

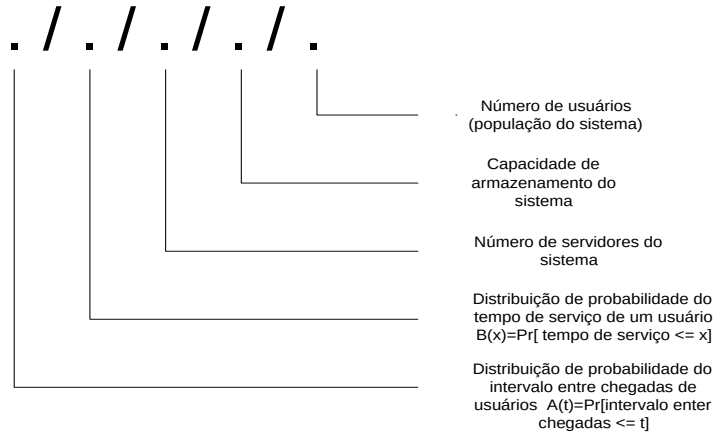
Basicamente um sistema de fila pode ser representado pela figura **XXXXXX** . O elemento central do sistema é o servidor, que provê serviço aos usuários do sistema. Os usuários, provenientes de alguma população, chegam ao sistema para serem servidos. Se o servidor está desocupado um usuário é servido imediatamente. Caso contrário o usuário é armazenado na fila e aguarda seu serviço. Quando o servidor completa o serviço de um usuário então este parte. Se existem usuários esperando na fila aquele na primeira posição vai para o serviço (considerando a disciplina primeiro que chega - primeiro a ser servido). A disciplina que indica qual o usuário que está na fila será o próximo a ser atendido é chamada de Disciplina de Despacho. No presente trabalho, salvo outras indicações, considera-se a disciplina primeiro que chega - primeiro a ser servido.



No caso de redes de computadores podemos associar os usuários a datagramas IP, células ATM, fluxo de bits dentre outros. O servidor pode ser representado por um enlace de comunicação que transportam tais usuários enquanto que a fila representa o espaço de armazenamento para os usuários nos equipamentos que compõem a rede.

Normalmente as medidas de desempenho em redes são adquiridas a partir de um modelo similar a este. Medidas como tempo médio de espera na fila, tamanho médio da fila, percentagem de usuários descartados (quando considera-se fila de tamanho finito) entre outros, são utilizados para a análise de desempenho da rede. Estas medidas são diretamente relacionadas às características dos usuários, tais como distribuição de probabilidade do intervalo entre duas chegadas consecutivas e distribuição de probabilidade do tempo de serviço dos usuários. Uma outra medida importante a ser analizada em um sistema de fila é a sua Utilização. A Utilização é a fração do tempo em que o servidor está ocupado, servindo os usuários. Assumindo-se que o espaço de armazenamento da fila é infinito (fila de tamanho infinito) então nenhum usuário é perdido pelo sistema; então eles são somente atrasados até o momento que podem ser servidos. À medida que a taxa de chegada de usuários aumenta, sem que haja mudança no serviço oferecido, a Utilização do sistema aumenta, sendo observado maiores quantidades de usuários armazenados na fila. Considerações práticas, como tempo de resposta e restrição no tamanho do tamanho da fila normalmente limitam a taxa de chegada de usuários.

Formalmente o estudo de sistemas de filas é feito através da notação de Kendall, onde cada campo é representativo de uma característica do sistema de fila. Esta notação é representada na figura **XXXXXX.**



Normalmente, em redes, o tráfego que alimenta o sistema de fila, isto é, o processo de geração de usuários para o sistema, é obtido através da agregação de várias fontes individuais. Muitas vezes será mais simples a análise do sistema considerando-se apenas um processo representativo da agregação de fontes. A análise considerando-se fonte a fonte

pode ser muito mais complexa. Isto pode ser pensado da seguinte forma: a cada nova fonte individual que se deseja analisar, novos parâmetros são adicionados ao problema.

Os modelos de tráfego estão diretamente relacionados à geração dos padrões de tráfego que alimentam os sistemas de filas no presente trabalho. Quanto mais precisos forem os modelos de tráfego, mais precisos serão as medidas de desempenho retiradas do sistema de filas e mais confiáveis serão as informações, obtidas a partir da análise do sistema, sobre o impacto de alguma mudança nos padrões de tráfego.

2.4 A divisão do problema modelando-se várias escalas de tempo envolvidas com os modelos de tráfego

Uma tarefa importante a ser feita é a identificação qualitativa das características de tráfego de serviços RDSI-FL que precisam ser levadas em consideração para projeto e avaliação de desempenho de redes que suportam tais serviços. Logo devemos nos concentrar na relação entre as características de tráfego de fontes VBR e o desempenho de sistemas de filas quando alimentadas por tais fontes de tráfego. Esta relação é fundamental para a definição de controle de tráfego em redes que suportam tráfego RDSI-FL, determinando quais serviços podem ser economicamente oferecidos e quais não podem. Um estudo de características de desempenho de um sistema de tráfego ATM serve para ilustrar as principais opções com relação ao controle de tráfego RDSI-FL. No presente trabalho, na grande maioria das vezes, consideramos uma única fila (muitas vezes chamada de multiplex ATM) com um certo espaço de armazenamento ("buffer") que recebe tráfego de um certo número de fontes VBR. Este ambiente pode ser representado pela figura 2.1 representada a seguir.

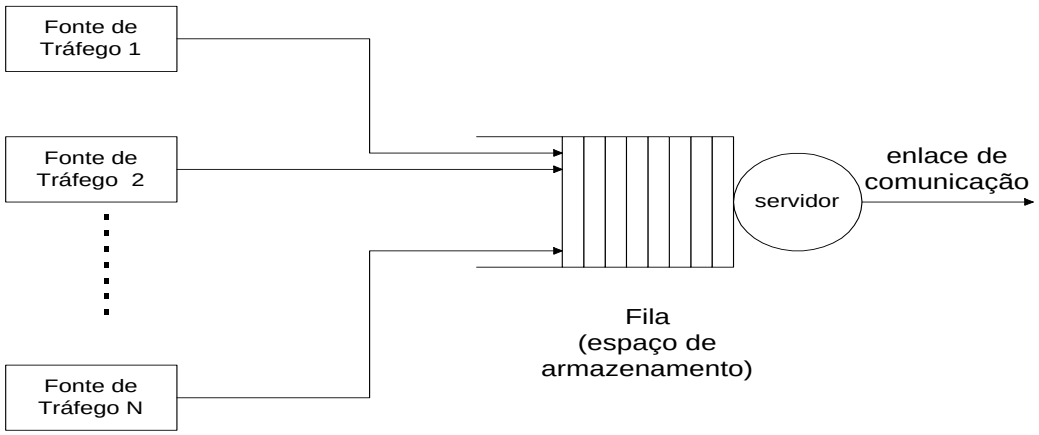


Figura 2.1

Tendo em mente este esquema de estudo nos concentraremos na relação entre :

- Tráfego : a natureza do tráfego agregado (obtida normalmente a partir de características de fontes individuais) oferecido expressa por um certo número de descritores de tráfego, provenientes do modelo de tráfego adotado.

- Capacidade : o tamanho do buffer e a ocupação / utilização do servidor presente no sistema de filas (enlace de transmissão que serve o multiplex).

- Desempenho : probabilidade de saturação do espaço de armazenamento (transbordo), distribuição do retardo experimentado pelas células, distribuição do tamanho da fila em termos de usuúrios ou outras medidas que podem ser importantes.

Normalmente podemos dividir o tráfego oferecido à rede em várias escalas de tempo relevantes [2]. São elas : de chamadas, de rajadas e de células. Cada uma delas têm suas próprias características particulares que discutiremos a seguir. Esta classificação de escalas de tempo não é geral; para modelos mais específicos de fontes de tráfego podem ser consideradas outras escalas com diferentes comportamentos. Por exemplo, fontes de vídeo podem ser modeladas através escalas de tempo de células, que agrupadas formam a escala de tempo de fatias (*slice*), que agrupadas formam escalas de tempo de quadros (*frames*), que agrupados formam as escalas de cenas [

2.4.1 Escala de tempo de chamada :

Esta escala de tempo tem seu comportamento determinado de acordo com os costumes humanos. Como exemplo podemos citar o tráfego telefônico que apresenta horas de maior movimento, determinadas por circunstâncias que ocorrem na vida cotidiana. Para a análise a nível de chamadas começaremos fazendo uma pequena análise do sistema telefônico atual utilizado para dimensionamento redes atuais e consideraremos posteriormente características de outros serviços diferentes.

A intensidade de tráfego telefônico é medida pelo número de chamadas que ocorrem em um determinado período e pelo número de chamadas que estão simultaneamente em progresso. O primeiro é expresso em Erlangs dado um determinado período de tempo no qual ocorrem as medidas. O tráfego medido pode variar dia a dia, hora a hora ou mês a mês, sempre refletindo variações no comportamento dos usuários.

A rede telefônica é dimensionada para seus usuários tenham um certo grau de serviço (como em redes do tipo B-ISDN) medido em probabilidade de bloqueio de chamadas em um período de "pico" de tráfego, ou hora de maior movimento. O tráfego neste período de pico é modelado de acordo com um processo estocástico descrito pelo processo de chegada de chamadas e o tempo de duração das chamadas.

Em redes de computadores, a noção de chamada não é necessariamente bem definida e depende se a rede é orientada a conexão ou não. Em redes orientadas a conexão podemos considerar uma chamada quando o circuito virtual está sendo utilizado. Durante esta chamada o tráfego é tipicamente em rajadas seguidas de período de inatividade. Neste caso a maior diferença que existe entre as redes telefônica e de comunicação de dados é o tempo de duração da chamada.

Em redes do tipo sem conexão, a noção de chamada não existe. Entretanto é conveniente notar que se os recursos da rede, como largura de faixa e espaço de armazenamento, precisam ser reservados o período de reserva deve ser o menor possível, mais ou menos igual ao tempo necessário para transmissão de um pacote. A interconexão de LANs, por outro lado, pode originar longas chamadas se o serviço é realizado através do estabelecimento de conexões virtuais entre os gateways que interligam elas.

O tráfego de vídeofone, ao nível da chamada, deve ser equivalente ao tráfego telefônico dada a definição do serviço. Por outro lado a vídeoconferência pode ter estatísticas diferentes : tempo de duração de chamadas de algumas horas, períodos preferenciais de início (por exemplo 10 a.m. ou 2 p.m.), reserva anterior para evitar congestionamento, de acordo com o serviço prestado atualmente.

Para aplicações convensionais baseadas em TCP/IP é apresentado um estudo em [3] sobre intervalo entre chegadas de conexões requesitadas. Neste estudo é mostrado que, em intervalos de uma hora, as chegadas de conexão TELNET e FTP são bem modeladas por um processo de Poisson. A chegada, no caso, reflete um usuário individual começando uma nova sessão. Por outro lado as chegadas de sessões WWW, SMTP(email) e NNTP(network news) não formam um processo bem definido.

2.4.2 Escala de tempo de rajada :

No nível de rajada estamos interessados no fenômeno ocorrendo em uma escala típica cujo comportamento é do tipo on/off (como por exemplo é a duração de um frame de vídeo) ao invés de uma escala de tempo que leva em consideração o intervalo de chegada entre células. Neste caso podemos ignorar a natureza discreta das células e considerar a chegada das células como um fluxo de taxa variável caracterizada por uma taxa instantânea. Para observarmos a dependência do comportamento da fila em relação a escala de tempo de rajada devemos considerar os instantes em que a taxa instatânea de chegada do fluxo é maior que a taxa instantânea de serviço, ou seja momentos em que o conteúdo da fila continua crescendo até que haja perda de informações por transbordo de espaço de armazenamento. Neste caso é de grande valia o conhecimento das distribuições de probalidade que rejem os tamanhos da rajada e seus momentos [21].

2.4.3 Escala de tempo de célula :

Aqui temos que considerar a natureza discreta da célula. A componente de célula surge devido às chegadas simultâneas de células das diversas fontes quando a taxa de chegadas agregada é menor que a taxa de serviço. Esta componente depende somente da distribuição estacionária da taxa de chegada total (sendo independente, por exemplo, das distribuições de probabilidade dos tamanhos dos períodos ativos e de silêncio quando consideramos fontes on-off) [21].

2.4.4 Considerações qualitativas sobre múltiplas escalas de tempo :

Mas qual a influência de múltiplas escalas de tempo sobre o comportamento do sistema de multiplex que recebe o tráfego gerado por várias fontes? Em [21] são apresentadas considerações qualitativas a partir de análise da superposição de fontes on-off.

No capítulo 3 faremos maiores considerações sobre o comportamento de uma fila alimentada por fontes de tráfego. No entanto, para o entendimento do problema de múltiplas escalas de tempo consideramos aqui o seguinte modelo para uma fila:

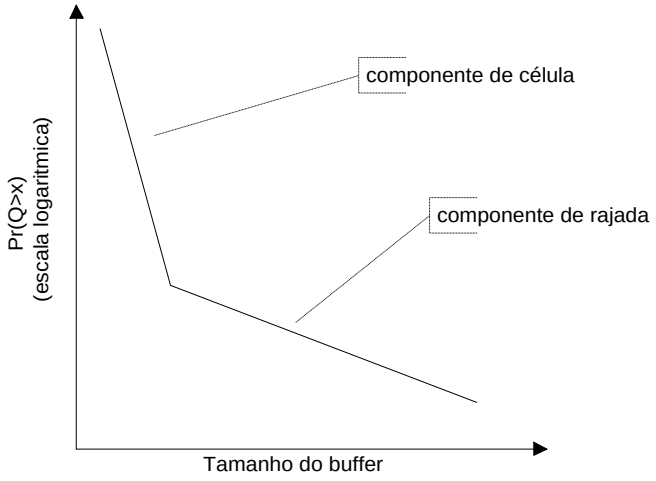
$$Q_{t+1} = (Q_t + A_t - C_t)^+$$

onde $x^+ = \max(x,0)$,sendo :

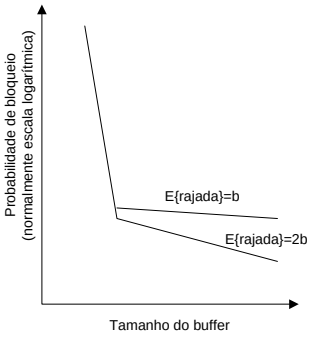
- A_t número de células que chegam durante o t-ésimo intervalo de tempo (slot);
- C_t : número de células que são transmitidas em um intervalo de tempo (slot);

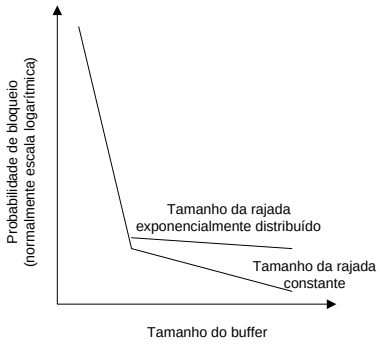
- Q_t conteúdo do buffer (em número de células) ao final do t-ésimo intervalo de tempo (slot);

Quando o tráfego oferecido é fixo a probalidade de overflow do buffer (ou uma aproximação para o seu cálculo) tem o seguinte comportamento, generalizado, descrito pela figura a seguir :



As duas componentes, já explicadas anteriormente, podem ser expressas através de probabilidades de tem relação direta : overflow de buffer com taxa de chegadas menor que a capacidade do multiplex e overflow de buffer com taxa de chegadas instantânea maior que a capacidade do multiplex. Na figura à esquerda abaixo podemos observar que a inclinação do componente de rajada é inversamente proporcional ao tamanho médio da rajada, enquanto que o componente de células permanece constante. O mesmo comportamento, representado na figura da direita, é observado em relação à variância da distribuição do tamanho da rajada. Uma análise mais detalhada destes comportamentos pode ser encontrada em [22].





A análise destes comportamentos pode ser utilizada no controle do tráfego em redes. Assumimos que temos o objetivo é termos a probabilidade de overflow de 10^{-9}. Desta forma temos duas opções básicas :

- O tráfego oferecido é restrito de tal forma que a probabilidade que objetivamos corresponde ao componente de célula.
- O tráfego é maior que na opção anterior levando a uma maior ocupação porém a probabilidade só é atingida no componente de rajada.

No primeiro caso o objetivo é manter probabilidade de que a taxa de chegadas seja menor que a capacidade do enlace fique menor que 10^{-9}. Levando-se em consideração fontes "on-off" esta probabilidade é determinada através de sua taxa de pico e da probabilidade que a fonte esteja ativa (verificar [2] pag. 150 para maiores detalhes). Por outro lado, permitindo-se o congestionamento a nível de rajada com mais freqüência, o tamanho do buffer correspondente à probabilidade da QoS adotada será função de outros parâmetros como tamanho médio da rajada, distribuição dos tamanhos dos períodos de atividade e de silêncio e possivelmente, descritivos de correlação entre sucessivos intervalos de silêncio e de atividade. Um simples exemplo desta análise pode ser vista em [23, pag. 187] onde o tamanho médio da fila M/G/1 cresce linearmente com a variância da distribuição do tempo de serviço.

2.5) Alguns modelos de fontes de tráfego VBR :

2.5.1) Modelos de fontes de voz :

Nos sistemas típicos de voz por pacotes o sinal é digitalizado, codificado e em seguida a informação é transmitida. O método modulação mais comum é o PCM, onde um sinal analógico é convertido ao formato digital codificado, que representa a amplitude

quantizada do sinal analógico original. As técnicas mais tradicionais para sistemas de transmissão de voz recaem em codificação CBR porque tais sistemas não permitem a variação da largura de faixa alocada para tais aplicações. A qualidade do som depende da taxa de amostra do sinal analógico (número de amostras por unidade de tempo) bem como da resolução da amostra (bits por amostra). A capacidade necessária seria então o produto destas duas quantidades [4].

Uma fonte de voz apresenta períodos ativos e períodos de silêncio. A voz CBR transmite os períodos de silêncio (que não tem nenhum tipo de informação) bem como os períodos ativos, fazendo, consequentemente, o uso ineficiente dos recursos providos pelo sistema. Para uma utilização eficiente do sistema a detecção dos períodos ativos é necessária para que os pacotes sejam gerados somente quando as fontes estão ativas. Ainda, para melhorar a eficiência da transmissão, novas técnicas de modulação são utilizadas como o DPCM (só é transmitida a diferença entre amostras consecutivas, com esta diferença sendo codificada por um número constante de bits) e o ADPCM (só é transmitida a diferença entre amostras consecutivas, com esta diferença sendo codificada por um número adaptativo de bits, conforme a diferença entre as amostras).

Desde que as fontes de voz VBR geram pacotes "periodicamente", as propriedades estatísticas do processo de chegadas dos pacotes precisam ser conhecidas e o processo precisa ser modelado para que o projeto de tais sistemas esteja de acordo com os requisitos de retardo e perda de pacotes para a reconstrução da voz no receptor. Normalmente podemos considerar que uma fonte de voz tem características do tipo "on/off" , quando no estado on há geração de células e quando no estado off não há geração de células. O que vai particularizar a fonte de voz, em relação a outros modelos do tipo on/off são as distribuições de probabilidade que rejem os tamanhos dos períodos de atividade e de silêncio.

Propriedades Estatísticas de uma fonte de voz :

Em [5] são descritas algumas propriedades estatísticas de uma fonte de voz . A partir destas propriedades foram propostos alguns modelos de tráfego agregado de fontes de voz como em [5,6] . Tais modelos constroem processos que tem por objetivo capturar propriedades observadas das fontes de voz e ao mesmo tempo fazer um estudo sobre um sistema onde várias fontes agregadas dividem recursos (como espaço de armazenamento e enlace de transmissão) de acordo com a figura 2.1.

O processo de chegada de pacotes ao multiplex é um tanto complexo e pode possuir correlação no número de chegadas em intervalos de tempo adjacentes o que afetará de forma significante o desempenho do multiplex. Até mesmo se o processo de geração de pacotes é aproximado por um processo de renovação, com pacotes deterministicamente espaçados durante o período ativo da fonte seguido por um período de silêncio exponencialmente distribuído, o processo resultante da superposição não é um processo de renovação e a sua análise exata seria intratável, especialmente se o sistema contém buffer finito. Para a análise do multiplex então são feitas aproximações para modelar a taxa agregada por processos mais simples que serão vistos a seguir.

Em [5] é construído um modelo representado por um Processo de Poisson Modulado por uma Cadeia de Markov (*MMPP - Markov Modulated Poisson Process*). Este processo é representado abaixo, figura 2.2 com suas variáveis.

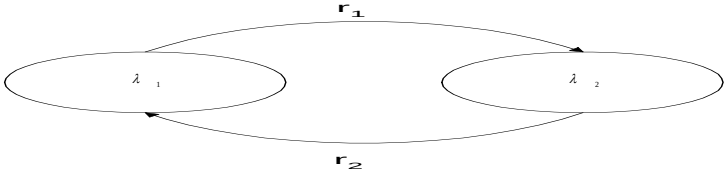


Figura 2.2

Os parâmetros r_1^{-1} e r_2^{-1} representam o tempo médio de permanência do processo nos estados 1 e 2 respectivamente enquanto que os parâmetros λ_1 e λ_2 representam as taxas de Poisson quando o processo encontra-se no estado 1 e 2 respectivamente. Em outras palavras quando a cadeia encontra-se está no estado *j* (*j=1,2*) o processo de chegadas é de Poisson com taxa λ_j. A partir de algumas características da fonte de voz individual (média, razão variância média e terceiro momento do número de chegadas de pacotes de voz em um determinado intervalo de tempo) são obtidas algumas características do modelo agregado (taxa média, razão variância média, terceiro momento para um intervalo de tempo finito e razão média variância em um intervalo de tamanho infinito). A partir destas quatro medidas são obtidos os parâmetros $r_1, r_2, \lambda_1, \lambda_2$. No mesmo trabalho é apresentado um algoritmo para

resolução da fila *MMPP/G/1*. Em [cookbook] são apresentados outras metodologias de resolução de fila *MMPP/G/1*.

Em [6] são apresentados outros três modelos de superposição de fontes de voz. Neste trabalho ainda são consideradas as propriedades estatísticas da fonte de voz, como apresentada em [5]. Considemos o modelo da fonte como sendo "on/off", ou seja, só há geração de pacotes durante o período ativo. Consideraremos *N* diferentes fontes de voz independentes, cada uma gerando um pacote a cada *1/V* s quando ativa. Ao ser gerado, o pacote entra na fila e o tempo de transmissão por pacote é de *1/VC* s. Logo *C* fontes ativas são necessárias para saturar o sistema. As distribuições dos tamanhos dos períodos ativo e de silêncio serão considerados exponenciais, com parâmetros α e λ respectivamente. Logo o número de fontes ativas pode ser modelado por uma cadeia de Markov de tempo contínuo (phase process) onde o número de fontes ativas é dado por γ. Esta cadeia é mostrada na figura 2.3 a seguir. As considerações feitas acima serão válidas para os três modelos apresentados em [6] , UAS, SMP e CMTC.

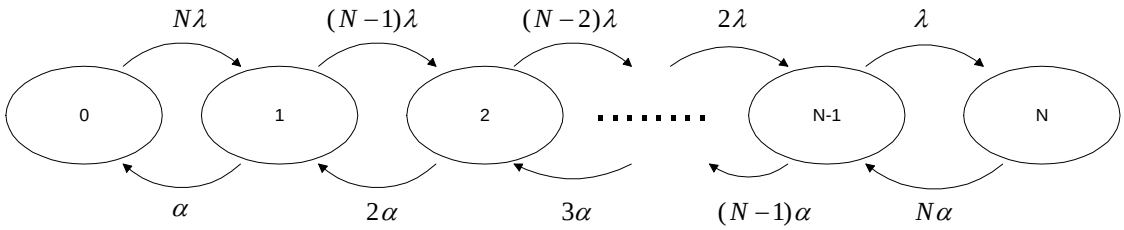


Figura 2.3

Para o multiplex descrito desta forma podemos dizer que três condições são possíveis :

- Para este processo temos que se $\gamma = j = C$ o tamanho da fila não muda (a taxa de mudança do tamanho da fila é zero). Só acontecerá se *C* for inteiro.

- Se temos $\gamma = j \leq C$ então, se a fila está ocupada por pacotes, ela esvazia-se de um pacote a cada $1 / [V(C - j)]$ segundos começando do ponto em que o processo entrou no estado *j* .

• E por último, se temos $\gamma = j \geq C$ o tamanho da fila aumenta de um pacote a cada $1/[V(j-C)]$segundos começando do ponto em que o processo entrou no estado j .

Os modelos SMP, CTMC e UAS foram gerados de forma a capturar o comportamento da fila de acordo com as condições apresentadas acima.

No modelo SMP (Semi-Markov Process) encontramos as seguintes aproximações: se $\gamma = C$ não há mudança no tamanho da fila, se $\gamma < C$ qualquer mudança no tamanho da fila é um decréscimo e se $\gamma > C$ qualquer mudança no tamanho da fila é um acréscimo. Em um sistema real, em qualquer instante de tempo em que um pacote é gerado durante a transmissão de outro, o tamanho da fila deverá aumentar independentemente da relação entre γ e C. Logo esta modelagem ignora as flutuações de alta freqüência que ocorrem na fila superestimando a probabilidade que fila esteja vazia. Na descrição matemática deste modelo o estado do processo é dado por $(l(t), \gamma(t))$ onde $l(t)$ representa o número de pacotes na fila no instante t e $\gamma(t)$ representa o número de fontes ativas no instante t. Se assumirmos a resincronização do processo nos instantes em que há mudança do número de fontes ativas as probabilidades de transição para este processo dependem somente do estado corrente. Então existe uma cadeia de Markov embutida nos instantes de mudança do número de fontes ativas, aumentos e decréscimos do tamanho da fila. A figura 2.4 a seguir representa o processo semi-Markoviano. Podemos observar que à direita da linha em negrito, representando a taxa de transmissão do servidor, só podemos ter aumento do número de pacotes na fila enquanto que à esquerda só podemos ter decréscimo do número de pacotes na fila. Uma vez que as probabilidades de equilíbrio para a cadeia forem determinadas a distribuição do tamanho da fila pode ser determinada. Para maiores detalhes sobre a resolução deste sistema consultar [6].

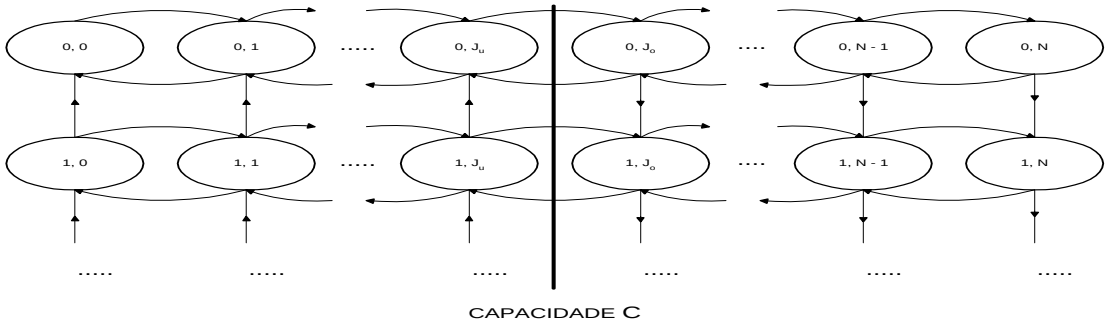


Figura 2.4

No modelo CTMC (Continous Time Markov Chain) consideraremos que os pacotes, durante os períodos ativos, são gerados de acordo com um processo de Poisson de parâmetro β ao invés da taxa constante apresentada no modelo SMP. A distribuição dos tamanhos dos períodos de silêncio e ativo continuam sendo exponenciais. Como a superposição de Processos de Poisson consiste de um novo Processo de Poisson, quando observarmos j fontes ativas, teremos uma taxa agregada de $j\beta$ para o processo de chegada de pacotes. Similarmente assumiremos que os tempos de serviço são exponencialmente distribuídos com parâmetro ν. O comportamento do modelo sobre um dado período de tempo durante o qual o processo está no estado $\gamma = j$, $j = 0, 1, ..., N$ é idêntico ao comportamento transiente do sistema *M/M/1* com taxa de chegadas de $j\beta$, taxa de serviço de ν e ocupação inicial igual a àquela do início do período. O diagrama de estados para o modelo CTMC pode ser visto a seguir, na figura 2.5. Para maiores informações sobre a solução do sistema consultar [6].

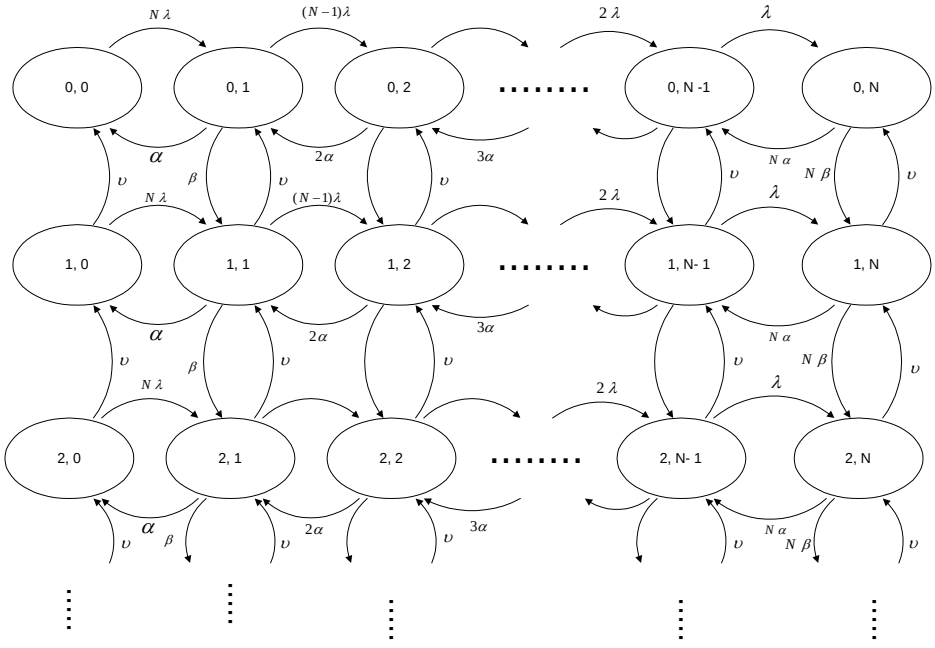


Figura 2.5

Para o modelo UAS (Uniform Arrival and Service) consideraremos que cada fonte, quando ativa, gera informação à taxa uniforme de uma unidade de informação por unidade de tempo e o servidor remove a informação a uma taxa uniforme que não excede a

capacidade do canal *C* unidades de informação por unidade de tempo. Quando o processo que descreve o número de fontes ativas em um determinado instante de tempo γ é maior que a taxa de transmissão do servidor ($\gamma=j>C$), o conteúdo do buffer aumenta a uma taxa constante de $j-C$ unidades de informação por unidade de tempo. Se o espaço de armazenamento está não vazio enquanto ($\gamma=j<C$) o conteúdo da fila é reduzido a uma taxa constante de $C-j$ unidades de informação por unidades de tempo. Em um sistema real a unidade de informação não entra no espaço de armazenamento de transmissão (e conseqüentemente não pode ser transmitida) até que uma fonte particular complete a geração deste pacote. Neste modelo, no entanto, é possível que a transmissão esteja sendo feita enquanto a geração ainda não foi concluída. O desempenho deste modelo tende a ser menos preciso a medida que o conteúdo do espaço de armazenamento é menor e que o número de fontes ativas é menor que a capacidade do sistema, por outro lado tende a melhorar à medida que a utilização do sistema de fila aumenta. Para a descrição matemática do sistema consideramos que este é representado por $B(t)$ e $\gamma(t)$ denotando o conteúdo do espaço de armazenamento e o número de fontes ativas em um determinado instante de tempo *t*. A solução matemática deste sistema é baseada na resolução de um conjunto de equações diferenciais com o objetivo de obter-se a distribuição de probabiliade do tamanho da fila. Para maiores informações sobre a solução consultar [6].

Algumas conclusões sobre os modelos de fontes de voz :

Os modelos apresentados em [6] (UAS, CTMC e SMP) tratam o tráfego agregado a partir do comportamento individual das fontes, ou seja em determinado período de tempo podemos ter a idéia de quantas fontes estão ativas, a taxa de serviço que produzem. Já no trabalho realizado em [5] é construído o modelo agregado tentando-se capturar as características estatísticas do processo de fontes agregadas.

Os modelos apresentados em [6] de um modo geral superestimam o tamanho da fila quando comparados com simulações por motivos apresentados anteriormente. Também é concluido em [6] que a covergência dos modelos CTMC e SMP para as filas M/M/1 e M/D/1 é bastante lenta em termos de previsão do comportamento do tamanho da fila devido à correlação do processo que gera os pacotes.

2.5.2) Modelos de fonte de vídeo :

Especula-se que o vídeo digital venha a ser o maior componente de tráfego em redes digitais de faixa larga. Aplicações como video-conferência, videofone, HDTV exigem destas redes uma quantidade relativamente grande de largura de faixa. A vantagem da tecnologia ATM na implementação de tais serviços é que esta consegue disponibilizar diversas taxas de transmissão sendo determinada uma quantidade que esteja de acordo com o serviço oferecido e com a qualidade de serviço a ser oferecida. O uso racional e econômico da largura de faixa requerida pelos serviços está intimamente ligada e dependente do desenvolvimento de técnicas de compressão de vídeo e dos modelos para as fontes de tráfego de vídeo[3].

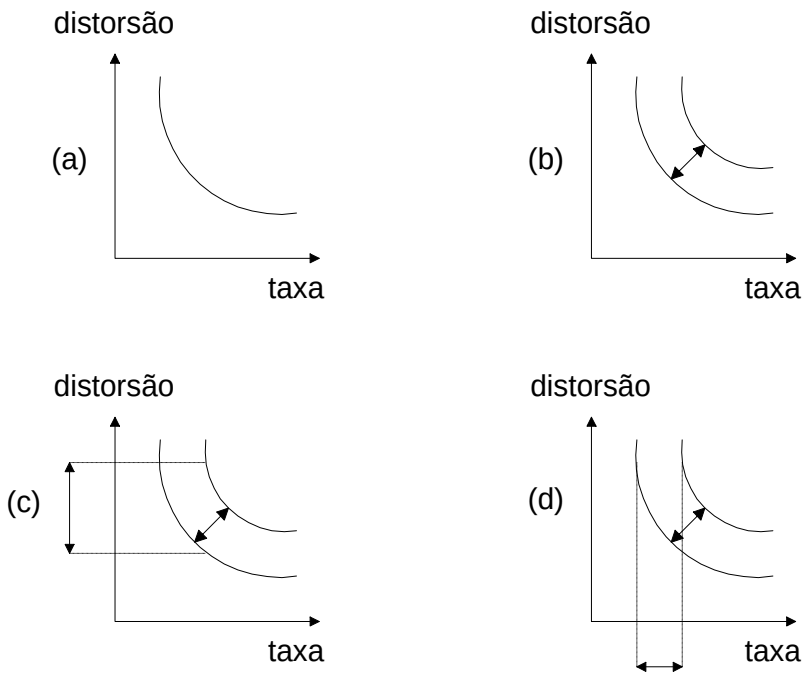
Para fazer uso do ganho de multiplexação estatística usando-se tráfego VBR é necessário ter total domínio das características estatíticas dos sinais de vídeo em redes de pacotes, em ATM ou outra tecnologia. O sinal original de vídeo apresenta uma grande correlação, que pode variar no tempo, entre imagens consecutivas. O grau de compressão do sinal de vídeo depende da correlação associada entre estas imagens (frames). Logo para um nível fixo de distorsão, a quantidade de informação produzida após a codificação irá variar ao longo do tempo, de acordo com o conteúdo do sinal de vídeo. As características estatísticas do tráfego gerado pelo vídeo podem ser obtidas observando-se a variabilidade da quantidade de informação gerada, bem como as características do codificador que gera esta informação. Dentro deste contexto podemos resumir alguns compromissos na transferência de informações de vídeo em redes de transferência assíncrona [7]. A tabela abaixo sumariza estes compromissos :

	A maximizar	A minimizar
Sessão de vídeo	Qualidade	Custo
Codificação	Qualidade	Taxa de transmissão
Controle de taxa de transmissão	Em concordância com a qualidade	Variabilidade da taxa de transmissão
Transferência	Utilização dos recursos	tempo gasto em filas
Controle de erro	recuperação de erro	"overhead"

A importância da codificação é maximizar a qualidade com uma menor necessidade de taxa binária de codificação, o que não é dependente diretamente da maneira de transmissão. A rede, por sua vez, deve minimizar o tempo de fila para reduzir o retardo fim a fim e a perda de informação utilizando os recursos da rede de uma maneira mais eficente

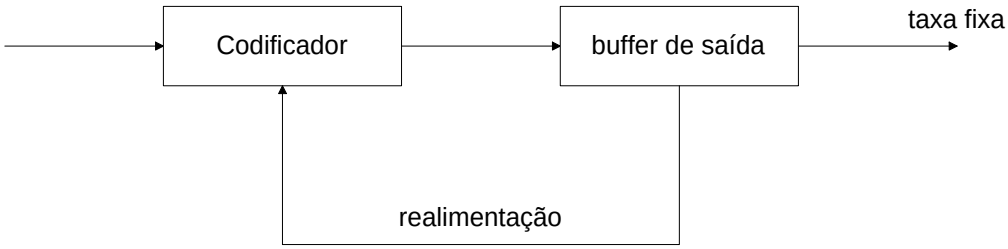
possível. Neste ponto é que a importância do controle da taxa de transmissão é vista : o fluxo deve ser suavizado antes da multiplexação ao mesmo tempo que é mantida a uniformidade da qualidade. O controle de erro é necessário para a recuperação de informação perdida devido a perdas e erros na rede ao mesmo tempo que minimiza a quantidade de "overhead" de informação na sinal. Podemos concluir que a principal questão é como controlar o codifcador para garantir uma boa qualidade e ao mesmo tempo permitir um bom desempenho de multiplexação na rede.

Para uma fonte de vídeo podemos ter uma taxa de geração de informação variável adaptando a taxa gerada à complexidade da imagem, tentando manter a qualidade da imagem constante. Isto pode ser visualizado na figura abaixo, reproduzida apartir de [8].



A figura (a) expressa uma típica função distorsão-taxa para um dado algoritmo de codificação e uma dada complexidade de imagem. Esta função é gerada pela variação o tamanho do degrau de quantização e medindo a taxa de geração binária em função de um critério de dirtorsão (para maiores detalhes consultar [impact]). A partir do momento que a complexidade das imagens geradas variam com o tempo a função taxa distorsão desloca-se como representado pela figura (b). Se temos uma taxa de geração/transmissão constante

(menor que um certo valor que garanta a qualidade fixa da imagem), a variação de complexidade de imagem acarretará variações na qualidade da imagem, como representado em (c). Isto pode ser parcialmente evitado utilizando-se um buffer adicional de acordo com a figura abaixo (novamente reproduzida de [impact].



Neste esquema é possível haver alguma flutuação na taxa de geração de bits, resultando em uma qualidade de imagem mais estável. O mecanismo de realimentação é incluído para prevenir "overflow" ou "underflow" no buffer, forçando o codificador a gerar uma taxa menor se buffer tende a um estado de overflow, e forçando uma taxa maior se o buffer tende a um estado de underflow. Ainda com esta solução teremos variações na qualidade do vídeo e um aumento de complexidade do sistema, o que é indesejável. Se uma taxa de transferência de informação varável é disponível uma qualidade de imagem constante e boa pode ser mantida como indicado em (d). Observamos que, se os parâmetros que descrevem a fonte são tratados no estabelecimento da conexão entre rede e usuário, não há a necessidade de controle da taxa do codificador e a transmissão do sinal em uma taxa ideal (podendo variar no tempo) é possível.

Alguns Modelos de fontes de vídeo :

Os modelos a seguir foram criados para a princípio modelarem tráfego de aplicações de vídeo-conferência, por isso consideraremos que não há mudança brusca na taxa gerada pelos codificadores.

1) Processo Auto-regressivo de primeira ordem :

Neste primeiro caso, é feita a aproximação estatística de uma única fonte de vídeo através de um processo auto-regressivo [maglaris]. É feita a modelagem dos valores que assume a taxa gerada a partir do codificador de vídeo por um processo estocástico em tempo discreto e estado contínuo. Neste processo a taxa atual da fonte é função dos valores

de taxas passadas, daí o nome de processo auto-regressivo. Seja λ(n) a taxa de uma fonte durante o n-ésimo frame. O modelo autorregressivo de primeira ordem é dado por :

$$\lambda(n) = a \cdot \lambda(n-1) + b \cdot \omega(n)$$

onde $\omega(n)$ é uma sequência de variáveis aleatórias Gaussianas independentes e a e b são constantes. Assumiremos que $\omega(n)$ tem média η e variância igual a 1 e que |a|<1, para que o processo seja estacionário para altos valores de n .O valor esperado de $\lambda(n)$ e a autocovariância discreta C(n) são dadas por [ross pag 465]:

$$E[\lambda(n)] = \frac{b}{1-a}\eta$$

$$C(n) = \frac{b^2}{1-a^2} a^n, n \geq 0, \quad .$$

A partir de dados medidos, como foi feito em [magladis] podemos determinar os valores das constantes a, b tendo como informação o valor esperado de $\lambda(n)$ e a função autocovariância que foi aproximada por uma exponencial em [magladis]. Este é utilizado principalmente para fins de simulação uma vez que a resolução analítica da fila alimentada por tal processo é complexa.

2) Cadeias de Markov com parâmetro (tempo) contínuo :

O modelo de Markov com estados discretos para a representação da taxa agregada gerada por várias fontes suporpostas possibilita um tratamento analítico mais simples. Neste caso a taxa agregada das fontes será discretizada em finitos níveis (estados). As transições entre os estados ocorrem com taxas exponenciais que dependem dos níveis correntes. Os níveis podem ser obtidos por amostragem do processo contínuo em intervalos de tempo aleatórios durante todo o processo, e em seguida, quantizados. A aproximação pode ser melhorada diminuindo intervalo entre os níveis de quantização e aumentando a taxa de amostragem. Como não estamos considerando imagens com mudanças abruptas (grandes variações de taxas) de cena, um processo nascimento e morte (transições permitidas somente entre estados adjacentes ou níveis de quantização) seria apropriado. Considera-se a tendência da taxa aumentar quando estamos em níveis mais baixos e diminuir quando o processo encontra-se em níveis mais altos . Para modelar o processo utilizaremos a figura 5, representada abaixo :

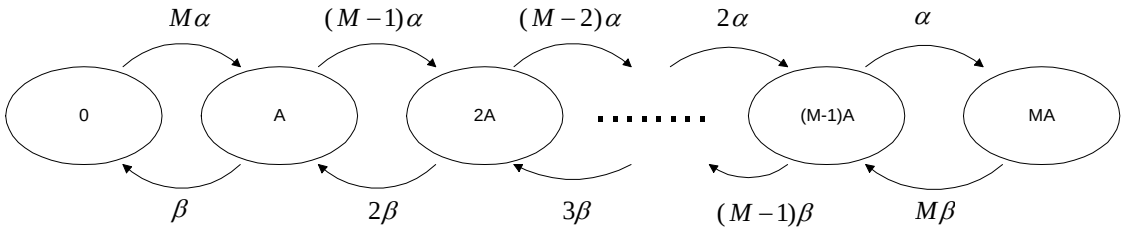


Figura 5

Onde : **A** representa o degrau de quantização (bits/pixel) e **M** +1 níveis possíveis (0, A, … , MA). As taxas exponenciais de transição $r_{i,j}$ do estado iA para o estado jA são dadas por :

$$r_{i,i+1} = (M-i)\alpha, \quad i < M$$
$$r_{i,i-1} = i\beta \quad i > 0$$
$$r_{i,i} = 0$$
$$r_{i,j} = 0 \quad |i-j| > 1.$$

Pode ser mostrado em [5 pag. 107] que $\lambda_N(t)$ no estado estacionário terá distribuição binomial com média $E(\lambda_N)$, variância $C_N(0)$ e autocovariância $C_N(\tau)$ dados por :

$$P\{\lambda_N(t) = kA\} = \binom{M}{k} p^k (1-p)^{M-k}, \quad p = \frac{\alpha}{\alpha+\beta}$$
$$E(\lambda_N) = MAp$$
$$C_N(0) = MA^2 p(1-p)$$
$$C_N(\tau) = C_N(0)\, e^{-(\alpha+\beta)\tau}$$

Os parâmetros do modelo *M*, *A*, α e β são obtidos através do conjunto de equações (6)-(9) e dos dados medidos.

Análise matemática do sistema de fila alimentado pelo tráfego da cadeia de Markov de tempo contínuo.

Consideremos uma fila sendo alimentada pelo tráfego representado analiticamente no sub-item anterior. Os valores de geração de informação desta fonte de tráfego assume seus valores $\lambda_N(t)$, em bits por segundo, de acordo com uma cadeia de Markov de tempo contínuo. Estes valores podem ser os seguintes valores discretos (*0, A, 2A, ..., MA*). Seja $r_{i,j}$ a taxa de transição exponencial discreto da taxa *i* para o nível *j* . A taxa de serviço é de *c* também em bits por segundo. Denotaremos o tamanho da fila no instante *t* por *q(t).* A descrição completa do sistema de fila requer um estado bidimensional { $q(t),\lambda_N(t)$ }. As estatísticas de um estado podem ser descritas por :

$$P_i(t,x) = \Pr[\lambda_N(t) = iA, q(t) \le x] \quad (10)$$

As equações de transição são dadas por :

$$P_i(t+\Delta t, x) = \sum_{j \neq i} r_{j,i} \Delta t P_j(t,x) + (1 - \Delta t \sum_{j \neq i} r_{i,j}) \cdot P_i\{t, x - (iA - c)\Delta t\} + O(\Delta t^2)$$

À medida que $\Delta t \to 0$, e ignorando os termos de segunda ordem Δt^2, a evolução do processo é governada pelo sistema de equações lineares diferenciais lineares.

$$\frac{\partial P_i(t,x)}{\partial t} + (iA - c)\frac{\partial P_i(t,x)}{\partial x} = \sum_{j \neq i} r_{j,i} P_j(t,x) - P_i(t,x) \sum_{j \neq i} r_{i,j}, \ 0 \le i \le M$$

Se o fator de utilização ρ é menor que 1 (um), ou seja :

$$\rho = \frac{E(\lambda_N)}{c} < 1$$

o processo atinge o seu estado estacionário com a distribuição limite $\lim_{t \to \infty} P_i(x,t) = F_i(x)$. Então o sistema é descrito por um sistema de equações diferenciais lineares dadas por :

$$(iA - c)\frac{dF_i(x)}{dx} = \sum_{j \neq i} r_{j,i} F_j(x) - F_i(x) \sum_{j \neq i} r_{i,j}, \ 0 \le i \le M$$

$F_i(x) = 0 \quad x < 0$

$F_i(0) = 0$ para $iA > c$.

As condições iniciais são obtidas de algumas observações : se a taxa instantânea *iA* é maior que a taxa de serviço *c* o buffer não pode estar vazio. Outra condição é obtida quando fazemos $x \to \infty$. Neste caso temos :

$$F_i(\infty) = \Pr\{\lambda_{\mathrm{N}}(t) = iA\}$$

O conjunto de equações [nn] podem ser escritas em forma matricial usando o vetor $\mathbf{F}(x) = (F_0(x), \ldots, F_M(x))^T$. Assim teremos :

$$\mathbf{D}\dot{\mathbf{F}}(x) = \mathbf{R}\mathbf{F}(x) \text{ com}$$

$$\mathbf{D} = \begin{pmatrix} A-c & 0 & 0 & \ldots & 0 \\ 0 & 2A-c & : & \ldots & 0 \\ 0 & : & 3A-c & \ldots & : \\ 0 & : & : & \ldots & : \\ 0 & 0 & 0 & \ldots & MA-c \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} -\sum_{j \neq 1} r_{1,j} & r_{1,2} & r_{1,3} & \ldots & r_{1,M} \\ r_{2,1} & -\sum_{j \neq 2} r_{2,j} & \ldots & \ldots & r_{2,M} \\ r_{3,1} & r_{3,2} & -\sum_{j \neq 3} r_{3,j} & \ldots & r_{3,M} \\ : & : & \ldots & \ldots & : \\ r_{M,1} & r_{M,2} & \ldots & \ldots & -\sum_{j \neq M} r_{M,j} \end{pmatrix}$$

Seja $\mathbf{\Phi}_i$ e z_i os autovetores e autovalores de $\mathbf{D}^{-1}\mathbf{R}$. A solução de tal sistema é dada em termos do seu valor quando $x=\infty$, dos autovalores e correspondentes autovetores de $\mathbf{D}^{-1}\mathbf{R}$ como se segue :

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_i k_i \Phi_{\mathrm{i}}\, e^{z_i x}$$

A soma na equação acima é feita considerando-se todos os autovalores no semi-plano lateral esquerdo para a solução ser uma distribuição de probabilidade (para isto devemos ter $\rho < 1$). As constantes k_i ainda na equação acima é determinada através das condições iniciais :

$$F_j(0) = 0 = F_j(\infty) + \sum_i k_i \phi_{ij} \text{ para } c/A < j \leq M$$

onde ϕ_{ij} denota o *j*-ésimo elemento de Φ_i . A distribuição em estado estacionário que estamos interessados é :

$$F(x) = \mathbf{Pr}\{q(t) \le x\} = \sum_{i=0}^{M} F_i(x)$$

e a probabilidade de que o conteúdo do buffer exceda um certo tamanho, *survivor function*, é $\bar{F}(x) = 1 - F(x)$.

Para o nosso caso em especial, fonte representada como na figura xxx, as taxas dadas em cccc simplificam a expressão dada em mmm.

$$(M-c)\frac{dF_i(x)}{dx} = (M-i+1)\alpha F_{i-1}(x) + (i+1)\beta F_{i+1}(x) - [\alpha + \beta + (M-i)\alpha]F_i(x), \quad 0 < i < M$$

Modelagem de fontes de vídeo baseadas no histograma :

Em geral a modelagem ideal seria aquela capaz de lidar com uma grande quantidade de seqüências independentemente de alguns parâmetros como conteúdo de cena e algoritmo de compressão adotado. Cosideraremos neste item que as taxas de bits de uma determinada fonte estão de forma que seja um certo número de bits é gerado para cada frame dentro de uma seqüência de frames. Dentro de um mesmo frame podemos fazer várias considerações (modos) sobre como as células distribuiem-se dentro dele, ou seja dentro de um período de frame consideraremos várias distribuições que regem a geração de células dentro deste período. Ao início de um frame a fonte apresenta um número *N* de células que devem ser transmitidas durante o próximo periodo de frame $1/f$.

Os modos (distribuições) de Poisson, uniforme e deterministico são modos mais suaves quando comparados com modo em rajada. Fontes nestes três primeiros modos estão sempre ligadas (gerando células) e sua geração de células varia frame a frame. Considerando o modo rajada em que quando a fonte está ligada há uma geração de bits a uma taxa constante de λ_p .

No modo Poisson a fonte gera células com intervalo de chegadas entre elas exponencialmente distribuidos (logo o numero de células gerado durante um período de frame não é exato mas é distribuido de acordo com um processo de Poisson). No modo

uniforme o numero de células gerado dentro de um frame é exato e o intervalo de chegadas entre as células tem uma distribuição aproximada de Poisson. Finalmente a fonte no modo determinístico gera células com espaçamento deterministico de tal modo que para aquele frame a última célula é gerada de modo que seu fim coincida com o fim do frame. A tabela a seguir sumariza a idéia acima :

Modo	Periodo ligado (de "on")	Número de células geradas	Distribuição do intervalo entre chegadas
rajada	N/λ_p	N	determinístico igual a $1/\lambda_p$
Poisson	$1/f$	Poisson com media N	exponencial com média igual a $1/(f \cdot N)$
uniforme	$1/f$	N	$\approx$ exponencial com média igual a $1/(f \cdot N)$
determinístico	$1/f$	N	determinístico com média igual a $1/(f \cdot N)$

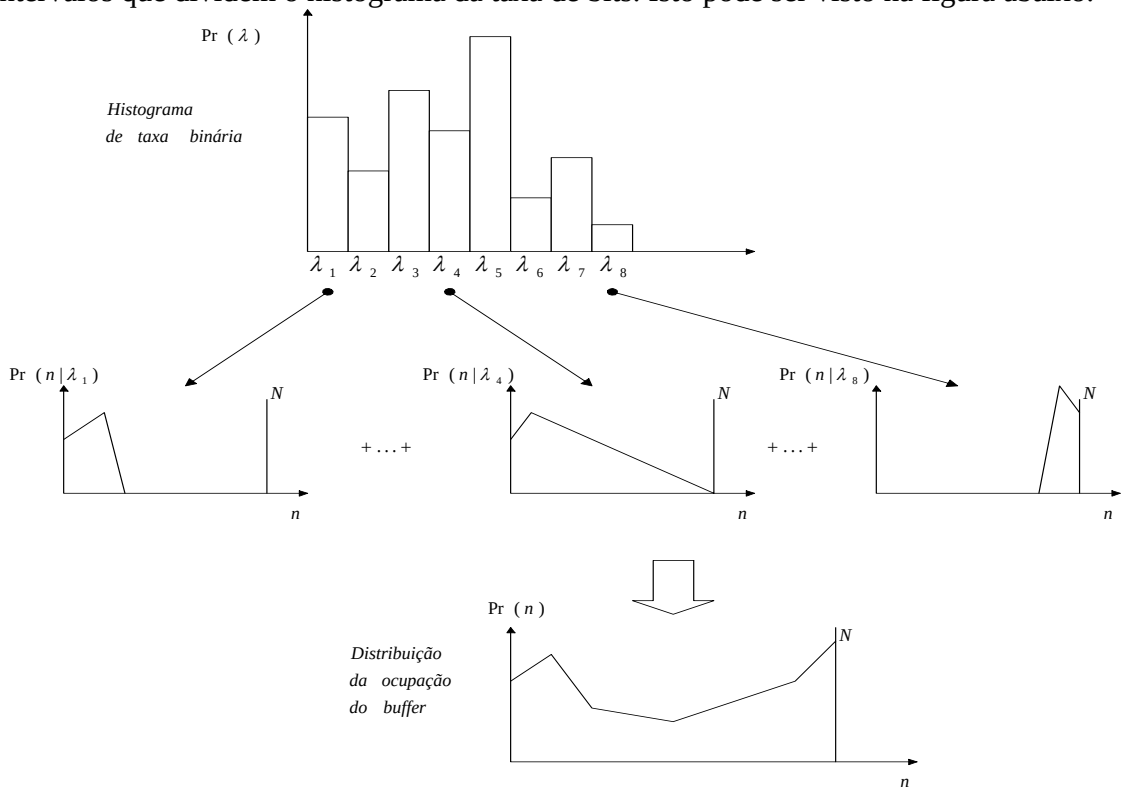
Intuitivamente podemos estimar que, para uma única fonte alimentando um *buffer* de tamanho finito, o modo em rajada seja o pior caso pois deve requerer um tamanho de buffer maior do que aquele requerido pelos modelos mais suaves para absorver a rajada de dados do frame corrente antes que estes dados sejam transmitidos através da rede. Para situações em que temos várias fontes multiplexadas a situação pode ser pior caso haja uma correlação inerente às fontes de vídeo. Para o caso de rajada caso tenhamos todas as fontes sincronizadas e com os mesmos tamanhos de frame, a sobrecarga no multiplex irá ocorrer sempre, ao início de cada frame. Para questões de modelagem e simplicidade consideremos as fontes não sincronizadas, ocasionando a suavização do tráfego fazendo com que seja menos freqüente a sobrecarga no buffer.

Cosideremos o modo uniforme. Notemos que este tipo de suavização aleatória não pode ser considerada como realista entretanto podemos ter uma visão qualitativa do que

ocorre na realidade. Sabemos que se as células de vídeo são distribuidas aleatoriamente dentro do frame com uma distribuição uniforme elas terão um intervalo de chegadas exponencialmente distribuidos aproximadamente. Desde que as células ATM tem tamanho fixo consideremos que elas terão serviço determinístico. Logo podemos considerar o sistema como uma fila *M/D/1/K* frame a frame , com *K* sendo o tamanho do buffer. Olhando para um frame de uma determinada fonte o processo considerado pode ser de Poisson com taxa λ , porém, considerando-se uma seqüência, λ terá uma distribuição $f_\lambda(\lambda)$. Desde que as taxas de chegada e saída das células são muito maiores que a taxa de frames, o sistema chega rapidamente ao estado estacionário, em relação ao tamanho do frame. Através deste sistema podemos observar as características de termo-longo na ocupação do buffer. Por exemplo a ocupação do buffer (dada uma fonte particular alimentando o multiplex) pode ser dada por :

$$P(n) = \sum_{I=1}^{N} \Pr(n \mid \lambda = \lambda_I) \cdot \Pr(\lambda = \lambda_I),$$

onde $\Pr(\lambda = \lambda_I)$ é a aproximação do histograma de $f_\lambda(\lambda)$ para a fonte de vídeo, $\Pr(n \mid \lambda = \lambda_I)$ é a ocupação do buffer dado que a taxa de chegada é λ_I e *N* é o número de intervalos que dividem o histograma da taxa de bits. Isto pode ser visto na figura abaixo.



2.5.3) Modelos de fonte de dados / Modelos Auto-Similares :

Estudos recentes de medidas de tráfego com alta qualidade e alta resolução têm revelado um fenômeno com ramificações em potencial para modelagem, projeto e controle de redes de banda larga. Estes estudos tem origem, e são baseados na análise de alguns milhões de pacotes observados em uma LAN Ethernet em um ambiente de pesquisa e desenvovimento [24]; e análise e observação de alguns milhões de frames de dados provenientes de serviços de vídeo VBR [25]. No primeiro caso uma rede LAN é monitorada e são colhetados o *timestamp*, o tamanho e o cabeçalho de cada pacote juntamente com o estado da rede (atividade ou não). No segundo caso é analisado o tamanho, em bytes, dos frames gerados a partir da codificação e compressão da informação. Nestes estudos o tráfego de pacotes, e de frames, parece ser estatisticamente auto-similar. O fenômeno auto-similar ou fractal faz com que o tráfego medido exiba uma estrutura similar quando visto em várias escalas de tempo. Em tráfego de pacotes, a auto-similaridade manifesta-se independentemente do tamanho das rajadas de dados: em todas as escalas de tempo consideradas (de alguns milisegundos até minutos ou horas), a natureza de rajada do tráfego mostra-se similar. Particularmente pode-se notar, quando observamos a plotagem do tráfego original, que, em qualquer escala de tempo, de milisegundos a minutos ou horas, não observa-se um tamanho natural da rajada, apenas períodos de tráfego mais intenso, separados por sub-períodos de tráfego menos intenso, mas todos eles sendo muito variáveis.

Implicações em potencial do tráfego real colhetado em [27] são colocadas em [28]. Tais implicações são relacionadas com projeto, controle e desempenho de redes de alta velocidade. Alguns resultados verificados são obtidos ser em termos de comparação entre medidas de tráfego real e tráfego gerado por modelos tradicionais (modelos Markovianos), notando-se a discrepância entre eles.

Não devemos confundir as expressões "dependência de termo longo" (*long range dependence*) e "auto-sililaridade" (*self-similarity*), as duas não são equivalentes. A primeira está relacionada com o comportamento de cauda de função auto-correlação de uma série de variáveis aleatórias enquanto que a segunda está relacionada ao comportamento de escalonamento de uma séria de variáveis aleatórias [36]. Em termos de modelagem estocástica os processos auto-similares são quase que exclusivamente utilizados em situações em que tenta-se levar em consideração a estrutura de correlação de termo longo de forma que se faça o uso de um pequeno número de parâmetros. Em [35] foi introduzido

o termo "auto-similaridade de segunda ordem exato" para seqüências estacionárias para processos que, quando agregados, geravam processos com a mesma estrutura de auto-correlação que o processo original. Tendo em vista esta definição utilizamos os termos "dependência de termo longo" e "auto-similaridade de segunda ordem exato (ou assintótico)" da mesma forma pois as duas fazem referência ao comportamento de cauda da função auto-correlação.

Podemos definir matematicamente um processo Auto-Similar exato da seguinte forma: dado um processo estacionário $X = (X_t : t = 0,1,2,\ldots)$ com média μ e parâmetro H para todo m=1, 2, 3,... e k=1 ,2 ,3 ,... temos que o processo formado por $1/m^H (X_{km-m+1} + \ldots + X_{km})$ tem mesma distribuição que o processo original X. Se o processo é Auto-Similar de segunda ordem então o processo agregado tem a mesma variância e auto-correlação que o original.

As diferenças entre os modelos de tráfego tradicionais (Markovianos) e as medidas de tráfego obtidos pode ser verificada de diversas formas. Uma das formas que podemos verificar estas diferenças é através da função auto-correlação do processo $X = (X_t : t = 0,1,2,\ldots)$ com média μ. A função autocorrelação é definida por:

$$r(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{E[(X_t - \mu)^2]}$$

No nosso estudo o processo X é representativo de uma variável relevante à análise do problema de teoria de filas que nos interessa, podendo representar, por exemplo, os intervalos entre chegadas usuários ou o número de usuários chegando ao sistema de fila em um dado intervalo de tempo. Para os processos tradicionais (chamados de *short range dependence SRD*) temos que a função autocorrelação decai exponencialmente, ou seja :

$$r_X(k) \approx |k|^{-\beta}, \text{ a medida que } |k| \to \infty$$

Para os processos com propriedades auto-similares temos o seguinte comportamento para tal função :

$$r_X(k) \approx |k|^{-\beta}, \text{ a medida que } |k| \to \infty, 0 < \beta < 1$$

Uma comparação matemática que podemos fazer entre as duas famílias de modelos é sobre o *IDC* (*index of dispersion for count*) [26], uma medida para avaliação da variabilidade do tráfego, que pode ser utilizado em diferentes escalas de tempo. O *IDC* em um intervalo de tamanho *t* é definido como a razão entre a variância do número de chegadas no intervalo e o número médio de chegadas no mesmo intervalo. Para o tráfego colhetado o *IDC* cresce monotonicamente em relação aos vários intervalos de tempo que foi medido. Isto está em contrate com modelos Markovianos, para os quais esta medida é uma constante ou converge para um dado valor rapidamente. Este comportamento pode ser verificado em [28].

O parâmetro chave para caracterizar o fenômeno da auto-similaridade é o parâmetro de Hurst *H*, que é utilizado para capturar o grau de auto-similaridade de uma dada seqüência, ou seja, um grau de rajada ("*burstiness*"). Este parâmetro pode ser estimado, dentre outras formas, através do próprio *IDC* através da seguinte relação:

$$IDC(t) = \frac{\mathrm{var}(\sum_{j=1}^{j=t} X_j)}{E(\sum_{j=1}^{j=t} X_j)} \approx k \cdot t^{2H-1}$$

onde *k* é uma constante positiva que não depende de *t*. X_j representa o número de usuários que chegam no *j*-ésimo intervalo. Para o caso da função autocorrelação temos que $H = 1 - \beta/2$.

2.5.3.1) Uma análise experimental com tráfego auto-similar :

Em [30] é feito um estudo a partir do tráfego medido em agosto de 1989 no laboratório Bellcore, verificar [24]. Este estudo tem por finalidade a análise de um sistema de fila quando alimentado pelo tráfego coletado e por alguns outros obtidos através de manipulações do original. É mostrado que a LRD não é o único fator determinante para o desempenho da fila, mas o determinante.

A fila, neste estudo, é modelada da seguinte forma : buffer infinito, serviço determinístico e único e o intervalo entre chegadas dos usuários obtidos a partir da medida original. As manipulações sobre o tráfego original são feitas da seguinte forma:

considerando a seqüência original do tráfego ethernet esta é dividida em blocos de tamanho *m* para um total de *N* intervalos entre chegadas. Logo exitem *N/m* blocos. As manipulações consistem da mistura dos intervalos internas aos blocos ou através da desordenação dos blocos originais. O estudo dos diversos efeitos provocados por estas manipulações são estudados através da plotagem do retardo médio versus utilização. Começaremos o estudo pela figura abaixo na qual a curva (A) representa o trágego original, a curva (B) representa a curva prevista pelo simulador QNA (*Queueing Network Analyser*) que usa um conjunto de aproximações G/G/1, baseadas no segundo momento do tráfego do tráfego de entrada. A curva (C) é obtida do tráfego gerado pelo "embaralhamento" de série de intervalos entre chegadas, mantendo-se a distribuição de probabilidade deles e destruindo as correlações entre eles. A discrepância entre as curvas (A) e (C) indica que um processo de renovação, por melhor que seja, não consegue estimar precisamente o retardo médio da fila considerando-se moderados e altos níveis de utilização, indicando que o aumento repentino observado no tráfego original é causado pela correlação existente no processo de chegadas. Um exemplo de modelagem auto-similar utilizando-se processos de renovação pode ser verificado em [31].

Outros experimentos são obtidos através de "embaralhamento externo", mantendo a correlação de termo curto e destruindo a de termo longo. A curva (E) na figura XXX é obtida para o valor de m=25 e observamos a discrepância em relação aos dados originais, para o "embaralhamento externo". A curva (F) é obtida através do "embaralhamento interno", quebrando-se a correlação de termo curto e mantendo-se a correlação de termo longo, ainda com m=25, uma vez que os blocos são mantidos na original, neste caso podemos observar a proximidade com o tráfego original. A curva (D) é obtida através do tráfego obtido da correlação de um passo, ou seja para um certo intervalo entre chegadas do tráfego original, o próximo intervalo é escolhido como um valor que corresponda a um daqueles que satisfaz o tráfego original. Logo através da figura podemos dizer que a correlação de um passo não é responsável pelo aumento repentino observado no gráfico correspondente ao do tráfego original. A conclusão que podemos tirar através da observação desta figura é que a *LRD* não é somente relevante para o desempenho da fila, mas sim a característica dominante.

Obviamente um "embaralhamento externo" com *m* suficientemente grande não irá afetar demasiadamente o resultado em relação ao original, uma vez que as correlações entre estes blocos maiores será mais fraca. Isto pode ser observado através da curva (G) para *m*=500. Podemos observar que a característica de retardo é diferente do resultado obtido através do tráfego original, enfatizando o fato que que a correlação em escalas de tempo maiores tem conseqüências práticas. Isto pode ser verificado na figura abaixo.

Os resultados similares podem ser obtidos quando outras formas de desempenho são consideradas como a distribuição do tamanho da fila. A figura a seguir mostra o logaritmo natural da distribuição complementar do tamanho da fila, $\ln[\Pr(Q > x)]$, versus *x*, obtida com utilização de 50% nas simulações experimentais descritas anteriormente. Este valor é particularmente importante porque nesta região ocorre o aumento abrupto da curva representativa da experiência com os dados originais. Para qualquer modelo Markoviano com estados finitos esta plotagem é assintoticamente linear indicando uma função complementar da forma $\exp(-\delta \cdot x)$, como podemos verificar para as curvas (C), (D) e (E); característica da aproximação da capacidade equivalente. Estas funções são particularmente importantes para controle de admissão de conexões como veremos posteriormente. As curvas são identificadas através das letras (A), (B), (C), (D), (E) e (F), da mesma forma que anteriormente. Podemos observar que as distribuições decaem mais rapidamente que a plotagem gerada a partir do tráfego original e a partir de (F). Podemos verificar um comportamento similar para as curvas (C), (D) e (E), todas com a correlação de termo longo removidas.

De acordo com o mesmo trabalho é esperado que resultados similares sejam obtidos se o conjunto de dados disponíveis é formado por uma série de tempos de contagens, ou seja, número de usuários que entram no sistema de filas em intervalos de tempos consecutivos e de mesmo tamanho.

2.5.3.2) Modelos de tráfego auto-similares :

Modelo $M/G/\infty$:

Este modelo é obtido gerando-se usuários de acordo com processo de Poisson de tempo discreto e oferecendo-os a um grupo de número infinito de servidores, assumindo-se que distribuição de serviço para os usuários é comum e é dada por uma distribuição de Pareto com parâmetro, ou seja, . O processo que conta o número de servidores ocupados

no início de um slot é o que chamamos modelo de entrada *M/G/∞*. O número de servidores ocupados representa o número de usuários que chega ao sistema em um slot de tempo considerado. Este processo é assintoticamente auto-similar com parâmetro de Hurst $H = (3-\alpha)/2 > 0.5$, de acordo com [32].

Modelo *Fractional Brownian Motion - FBM* :

O tráfego *fractional Brownian*, onde Z_t é um FBM normalizado e A_t é o tráfego oferecido no intervalo de tempo *[0,t)* é representado pela seguinte expressão [32]:

$$A_t = m \cdot t + \sqrt{a \cdot m} \cdot Z_t, \; t \in (-\infty, +\infty).$$

O processo tem três parâmetros, *m, a* e *H* com as seguintes interpretações:

- *m > 0,* é taxa média de chegada ;
- *a > 0,* é um coeficiente de variância ;
- $H \in [1/2, 1]$ é o parâmetro de Hurst associado a Z_t.

Z(t) é um objeto matemático que não tem dimensão física e o parâmetro *t* também não. Se a quantidade de trabalho (quantidade de informação a ser transmitida) é dada em bits e o tempo é medido em segundos então *a* tem dimensão de bits×segundos. O parâmetro de Hurst, como já dito anteriormente não tem dimensão.

O fator $\sqrt{m}$ é motivado pela seguinte propriedade de superposição : A soma $A_t = \sum_{i=1}^{K} A_t^i$ de *K* fontes de tráfego com parâmetros *a* e *H* mas taxas individuais m_i pode ser escrita como $A_t = m \cdot t + \sqrt{a \cdot m} \cdot Z_t, \; t \in (-\infty, +\infty)$, onde $m = \sum_{i=1}^{K} m_i$ e Z_t é um FBM com parâmetro *H*. Logo podemos fazer a separação dos parâmetros do tráfego, sendo que *a* e *H* representam "qualitativamente" o tráfego enquanto que *m* representa o tráfego "quantitativamente".

2.5.3.3) Análise do tráfego ao nível de fontes individuais :

Em [36] é realizado um estudo do tráfego de rede real, ainda [24], ao nível de fontes individuais, através dos pares fonte-destino. Essencialmente é mostrado que a superposição

de várias fontes *on-off* independentes e identicamente distribuídas, cada qual exibindo o fenômeno chamado efeito Noah, resulta um tráfego agregado auto-similar. Particularmente as fontes *on-off* são considerados os pares fonte-destino do tráfego de rede colhetado. O efeito Noah para fontes *on-off* é conseqüência da possível presença de grandes tamanhos de períodos de *off* (período em que não há geração de tráfego) e de *on* (período de geração de tráfego a uma taxa constante). Grandes tamanhos aqui quer dizer que estes valores podem ser grandes com probabilidades não negligenciáveis. Este efeito garante que o modelo *on-off* exiba características que cobrem um grande intervalo de escalas de tempo. Matematicamente usamos distribuições de probabilidade *heavy-tail* com variância infinita (por exemplo a distribuição de Pareto) para a representação do efeito Noah. Uma distribuição é heavy-tail se :

$$\Pr[X > x] \approx x^{-\alpha}, \text{ a medida que } x \to \infty,\ 0 < \alpha < 2.$$

Isto é : independentemente do comportamento da distribuição para pequenos valores da variável aleatória, se o formato assintótico da distribuição é hiperbólico então ela é *heavy-tail*. O valor de α fornece o grau de *heaviness* da cauda da distribuição e do grau do efeito Noah. Se $\alpha \leq 2$ então a distribuição tem variância infinita; se $\alpha \leq 1$então a distribuição tem média infinita. Logo a medida que o valor de α diminui maior a contribuição da cauda da distribuição. Neste ponto fica claro que uma variável aleatória que tem uma distribuição *heavy-tail* pode gerar grandes valores com probabilidades não negligenciáveis. O efeito Joseph (grau de auto-similaridade do tráfego agregado) é função direta do valor de α e é representado pelo parâmetro de Hurst H :

$$H = \frac{(3-\alpha)}{2} \quad .$$

Quando temos distribuições *heavy-tail* para os períodos de *on* e de *off* com parâmetros α_1 e α_2 respectivamente, temos a seguinte relação :

$$H = \frac{(3-\min(\alpha_1, \alpha_2))}{2}.$$

2.5.3.4) Possíveis Causas da Auto-Similaridade

No trabalho realizado em [37] é mostrado que, em alguns casos, a Auto-Similaridade presente no tráfego de rede pode ser explanada em termos das características

de sistemas de arquivos e comportamento de usuário. Neste trabalho os dados colhetados estão a nível de aplicação resultando na capacidade de exame da relação entre os períodos de transmissão (e de inatividade) e tamanhos de arquivos. Ainda é mostrando que a natureza *heavytail* de tais períodos não é resultado direto dos protocolos de rede ou de preferência dos usuários mas sim de propriedades do armazenamento de informação e de intervalos de tempo que o usuário pensa antes de tomar uma ação (clique no *mouse* por exemplo).

Para a obtenção dos dados, a URL (*uniform resourse locator*) de cada arquivo acessado pelo usuário, foi utilizado o *Mosaic* uma vez era de domínio público o seu código. Os dados capturados consistem de uma seqüência de pedidos de arquivos durante cada execução do *Mosaic*. Cada pedido de arquivo é definido por sua URL e tem a ele associado o *timestamp* que indica o início do pedido, o tamanho do documento (incluindo o *overhead* do protocolo) e o tempo de recuperação deste objeto. Por último os *bytes* transferidos de cada pedido são armazenados em reservatórios relativos à duração da transferência. Para a explicação do fenômeno auto-similar observado no tráfego WWW é utilizado o modelo de superposição de fontes *on-off* com ambos os períodos tendo distribuição *heavy-tail* e faz-se a análise dois períodos de transmissão e de inatividade.

É observado que a distribuição dos períodos de transmissão de arquivos de Web mostram probabilidades não negligenciáveis em um grande intervalo de tamanhos de arquivos, sugerindo uma variância infinita para tal distribuição. Isto seria função da distribuição de probabilidade dos tamanhos dos arquivos requisitados e mais ainda, dos tamanhos dos arquivos disponíveis na Web. Para maiores detalhes verificar [37].

Os períodos de inatividade são considerados aqueles entre a conclusão de transmissão de uma URL e o pedido de uma próxima dada uma sessão. Contudo, para a distribuição dos períodos de inatividade pode-se observar dois regimes : um devido aos períodos já definidos e outra devido aos curtos períodos de inatividade observados durante os períodos de inatividade. Estes podem ser explicados como a espera pela estação de um componente de página WWW ou da atividade de formatação e exibição da informação recebida antes de pedir um novo componente do documento.

Como as distribuições dos períodos de *on* e de *off* apresentam características *heavy-tail* a estimativa do parâmetro de Hurst fica de acordo com a expressão ( ). Neste caso os períodos de atividade ficam responsáveis pela Auto-Similaridade 1observada no tráfego

WWW. Isto está mais ou menos de acordo com a observação das transferências de FTP observadas em [3].

Capítulo 3

Controle de Admissão de Conexões

3.1 Introdução

Uma das áreas de grande importância em redes ATM é aquela relativa ao conjunto de procedimentos de controle de congestionamento. O controle de congestionamento tem por objetivo fazer com que usuário e rede tenham seus objetivos de desempenho atingidos, de forma que se faça o uso otimizado dos recursos da rede. Pode-se classificar o controle de congestionamento em preventivo e reativo. O primeiro previne a ocorrência de congestionamento (sendo este aplicado em tráfego CBR e VBR). O segundo é baseado em informações de funcionamento de rede que retratam o grau de congestionamento experimentado por esta, sendo aplicado ao tráfego ABR. A operação adequada de uma rede ATM parte do princípio que estas duas formas de controle de congestionamento são complementares. O controle de admissão de conexões, que é uma parte integrante do controle de congestionamento preventivo, será estudado no presente capítulo.

3.2 Controle de Admissão de Conexões

O controle de admissão em uma rede de comunicação pode ser definido como uma série de ações que a rede toma quando é solicitada uma nova conexão entre dois ou mais pontos que fazem parte desta rede. Pode-se verificar que o controle de admissão de conexões não é uma particularidade das redes ATM.

Em uma rede que utiliza a comutação de circuitos, como por exemplo a rede telefônica atual, os recursos (segmentos de tempo específicos em um quadro TDM) são alocados estaticamente para cada conexão. Durante a tentativa de conexão, a rota da fonte ao destino é calculada e os slots livres dentro deste caminhos são alocados à nova conexão. Se os recursos oferecidos pela rede, naquele momento, são insuficientes para o estabelecimento da nova conexão, ela é rejeitada. Este tipo de rede é ideal para suportar tráfego CBR com restrições de retardo e *jitter*. A reserva estática provoca a subutilização dos recursos da rede quando temos tráfego de dados, que apresentam características de rajada.

Em redes de comutação de pacotes tradicionais (datagrama ou circuito virtual) os recursos são alocados dinamicamente em termos de espaço de armazenamento em filas. Técnicas de realimentação ou controle de fluxo baseado em janelas são utilizados para garantir que não haja transbordo nas filas de equipamentos que fazem parte da rede. Devido às grandes variações de retardo, as redes de pacotes tradicionais não suportam serviços sensíveis a retardo e sensíveis a grandes valores de *jitter* com a qualidade aceitável .

Em redes ATM, a largura de faixa é alocada dinamicamente e virtualmente, além disso o grau de serviço oferecido é garantido de uma forma probabilística ou determinística. Considerando-se uma estratégia de alocação de recursos bem projetada podemos dizer que redes ATM podem combinar as vantagens de redes que usam comutação de pacotes e de circuitos, ao mesmo tempo evitando suas desvantagens.

O sucesso ou o fracasso das redes ATM dependem do desenvolvimento de um esquema de controle eficiente. Este controle eficiente ainda é uma questão sem solução adequada para as particularidades de tais redes. Podemos citar alguns aspectos que complicam este problema [3] :

- Várias fontes VBR geram tráfego a diferentes taxas. Isso implica em diferenciação de tratamento para as várias fontes possíveis.

- Uma única fonte pode gerar múltiplos tipos de tráfego (por exemplo, voz e dados) com diferentes características.

- Em adição às métricas de desempenho de bloqueio de chamadas (utilizada em tráfego telefônico) e taxa de perda de pacotes (redes de dados), o desempenho das redes ATM também tem que saber lidar com variação do retardo de células e retardo máximo.

- Diferentes serviços tem diferentes qualidades de serviços em diferentes níveis. A multiplexação estatística de diversos tipos de tráfego dificulta a previsão do desempenho de rede para todos os tipos.

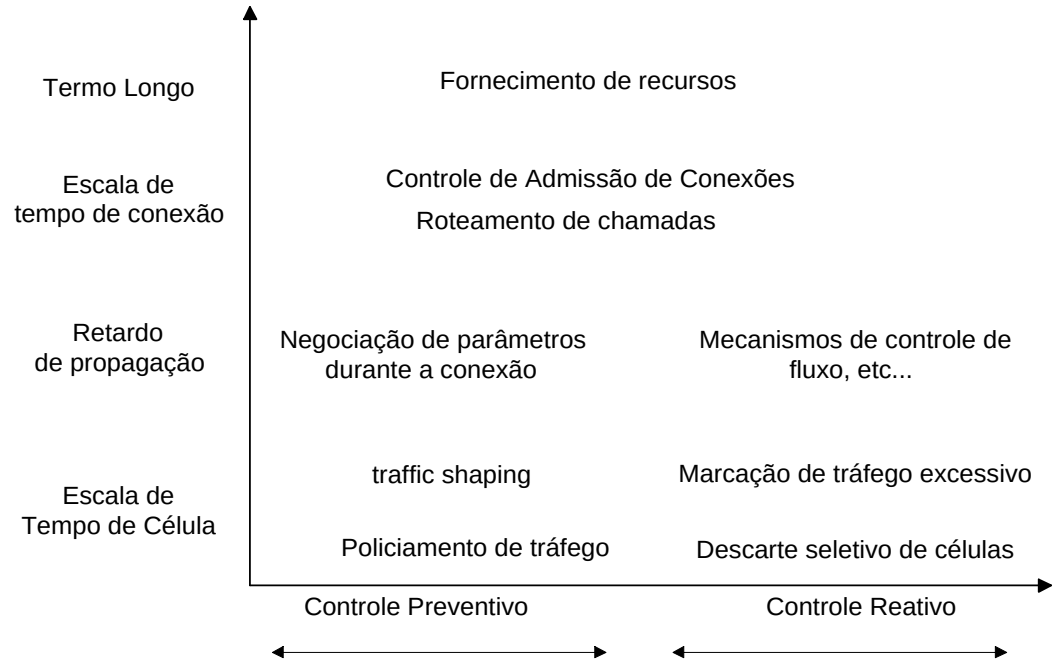
- As características de vários tipos de tráfego ainda não são bem conhecidas (conforme visto no capítulo 2).

- À medida que a velocidade aumenta, a razão entre a duração da chamada e o tempo de transmissão das células aumenta, contribuindo para o aumento da dificuldade do problema. Para melhor entendimento consideremos o seguinte exemplo [38] : Imagine um comutador ligando pontos A e B a uma distância de 100 Km. Assuma uma célula de 500 bits (imaginária) e um retardo de propagação de 5 μs por Km. Assumindo um canal de velocidade de 1 Mbps o tempo de transmissão de uma célula é de 0.5 ms (500 bits / 1 Mbps). O nó A começa a transmitir. Após 500 μs o primeiro bit chega a B enquanto que A está transmitindo o último bit da célula. Aumentando a velocidade do canal para 1 Gbps o tempo de transmissão da célula se reduz a 0.5 μs enquanto que a propagação no meio permanece a mesma. O nó A começa a transmitir continuamente. Quando o primeiro bit da primeira célula transmitida chega a B já existem 1000 células no meio de transmissão sem que o nó B tenha conhecimento, podendo ocasionar o transbordo da fila deste nó. Logo o aumento das taxas de transmissão exige novos mecanismos de controle de congestionamento.

Uma das etapas de controle do congestionamento em redes ATM, que dos requisitos de QoS dos usuários, é a implementação do Controle de Admissão de Conexões. Este é o processo pelo qual a rede deve decidir se uma nova conexão deve ser aceita ou não. Não fosse a necessidade de transportar uma grande variedade de tráfegos, com diversos requisitos de desempenho e características, os controles de admissão tradicionais poderiam ser aplicados. Em uma rede ATM, quando uma nova conexão é requesitada, ela é aceita somente se a rede tem recursos suficientes para suportar esta nova conexão sem que haja degradação do QoS das outras conexões que já estavam em progresso. Qualquer técnica projetada para resolver esta questão deve produzir seus devidos efeitos em tempo real e ao mesmo tempo tomar decisões precisas de tal forma que as exigências de QoS sejam respeitadas e que a rede atinja altos graus de utilização dos recursos da rede.

É importante enfatizar que o Controle de Admissão de Conexões é apenas um dos exigidos para o perfeito funcionamento de redes ATM. Os controles preventivos tem sua ação em diversas escalas de tempo, desde célula até o fornecimento dos recursos enquanto que os reativos estão restritos às escalas de retardo de propagação [3]. A figura 3.1 retirada de [3] exemplifica esta idéia.

Figura 3.1 : Controles Preventivos



Redes ATM fazem uso da multiplexação estatística. Dentro desta filosofia a quantidade de largura de faixa alocada para uma conexão, cuja fonte de tráfego é VBR, é menor que a sua taxa de pico mas necessariamente maior que a sua taxa média. Logo a soma das taxas de pico das conexões multiplexadas em um enlace pode ser maior que a capacidade do enlace, ao mesmo tempo que sua soma estatística pode ser menor ou igual à capacidade do enlace. Para estimar o valor da largura de faixa estatística de uma nova conexão devemos atentar para as seguintes questões :

- As exigências de QoS desta nova conexão deve ser satisfeita.

- As QoS das conexões pre-existentes não podem ser degradadas, assumindo níveis não aceitáveis, quando elas são multiplexadas com a nova conexão.

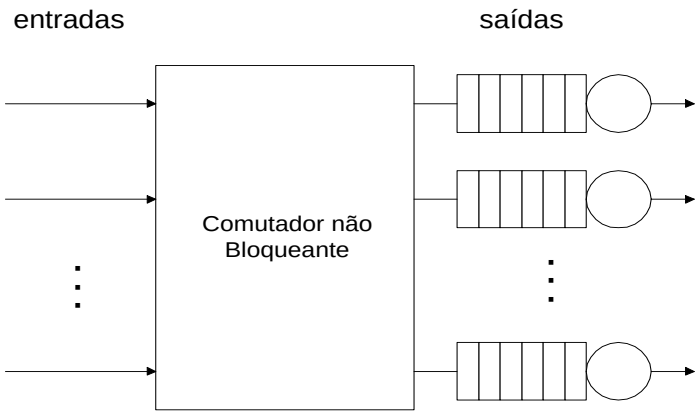
Este valor de largura de faixa estatística de uma conexão depende não somente das suas próprias características estatísticas, mas também depende das características estatísticas das conexões já em progresso. Isto mostra a dependência existente entre a modelagem de fontes de tráfego e o controle de admissão de conexões. Ou seja quanto mais precisa for a modelagem das fontes de tráfego, o controle de admissão de conexões baseada nela será mais preciso e fará o uso racional dos recursos da rede.

Em redes que apresentam fontes de tráfego VBR, como em redes ATM, um dos principais interesses relacionados à sua administração, é a economia de capacidade dos enlaces de comunicação, tirando-se proveito da multiplexação estatística. Para uma administração eficiente de uma rede com tráfego combinado de fontes VBR é necessário o poder de expressão da eficiência da multiplexação de diferentes fontes de uma maneira simples e útil.

3.3 Modelo de Comutador Utilizado

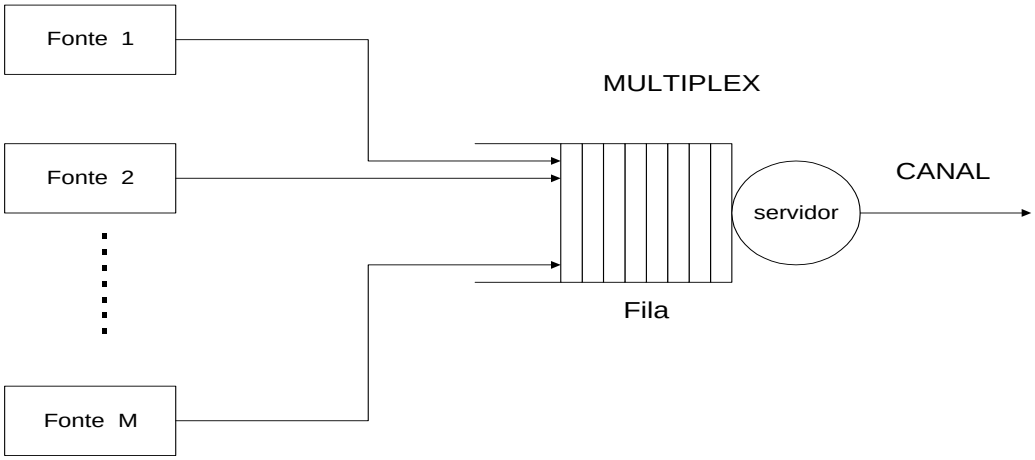
No presente trabalho considera-se o modelo de um comutador ATM não bloqueante. Neste tipo de comutador o congestionamento ocorre nas portas de saída, sendo que cada uma das portas de saída é provida de uma fila, de acordo com a figura 3.2.

Figura 3.2 : Esquema do Comutador não Bloqueante



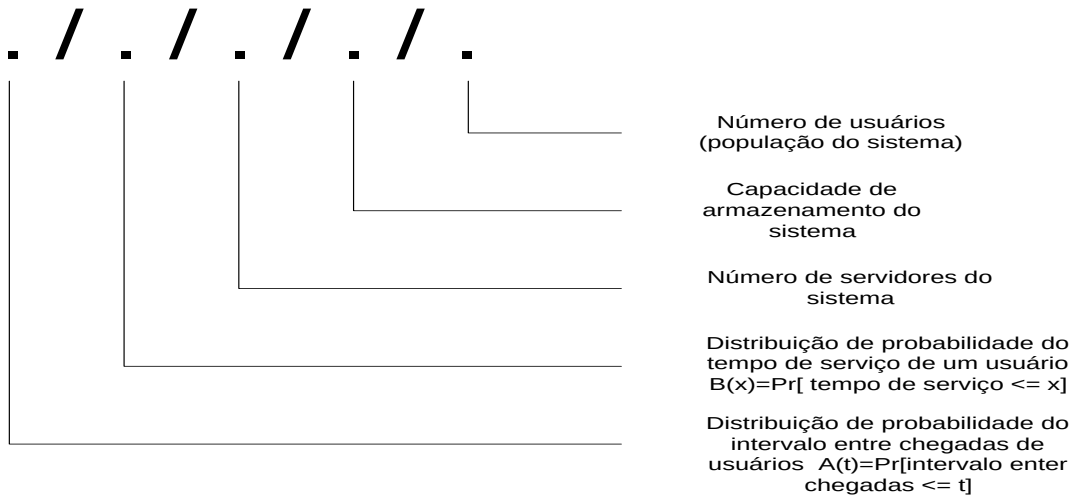
O algoritmo de controle de admissão de conexões deve ser aplicado a cada porta de saída e fila correspondente. Isolando uma porta de saída e sua fila de saída correspondente do comutador obtemos um modelo de fila. Esta estrutura isolada é chamada de multiplex ATM e considera-se que várias fontes alimentam este sistema de filas. A figura 3.3 mostra o esquema de um multiplex estatístico / multiplex ATM que será referenciado no presente trabalho.

Figura 3.3 : Esquema do Multiplex



Para formalizar este problema novamente utiliza-se a notação de Kendall (Figura 3.4). Este modelo serve como formalização da modelagem do comportamento do multiplex.

Figura 3.4 : notação de Kendall



A pergunta seria relativa ao CAC: É possível estabelecer uma quantidade de largura de faixa (tempo de serviço de uma célula / usuário), dado um valor de capacidade de armazenamento do sistema (tamanho da fila), a partir de poucas características estatísticas da fonte , que garantisse a qualidade de serviço negociada na fase de estabelecimento da

conexão e o alto grau de utilização dos recursos da rede ?. Sendo atribuído um valor para esta largura de faixa para uma fonte (ou para um conjunto delas) este representa a capacidade na qual o servidor deve trabalhar (transmitir um certo número de células em um determinado intervalo de tempo) para que seja garantida a qualidade do serviço contratado.

A importância relacionada ao cálculo desta largura de faixa está intimamente ligada às funções de controle da rede, como controle de congestionamento, controle de fluxo, roteamento (determinação de circuitos virtuais e caminhos virtuais) e, como já dito, controle de admissão de conexões. De um modo geral podemos listar vários requisitos que um mecanismo de controle de admissão de chamadas deve atender, dentre os quais podemos citar [11]:

- como já mencionado, uma resposta em tempo real do mecanismo. Isto terá como conseqüência o uso de mecanismos simples, aproximados ou baseados em tabelas, uma vez que não haveria tempo suficiente para realização dos cálculos numéricos exatos;

- deve existir uma margem de segurança de modo a garantir que a QoS seja satisfeita quando todas as conexões estiverem se comportando como declarado durante o contrato feito no estabelecimento da conexão;

- o mecanismo de admissão deve possuir associado a ele um mecanismo de policiamento para verificar a conformidade entre o tráfego declarado e o real, de modo a evitar que um tráfego excessivo na rede possa prejudicar a QoS do restante das conexões;

- os mecanismos devem ser válidos para os casos de conexões de tipos variados;

- é necessário que o mecanismo de controle de admissão de chamadas faça uso do efeito de multiplexação estatística (por exemplo não alocando uma capacidade superior para cada conexão do que aquela necessária).

3.4 Reserva não Estatística de Largura de Faixa

Considera-se uma fonte com taxa média de 20 Mbps e taxa de pico de 45 Mbps. A reserva pela taxa de pico, ou reserva não estatística, exige que a largura de faixa de 45 Mbps seja reservada na porta de saída do comutador independentemente se a fonte transmite continuamente a uma taxa de 45 Mbps. Esta modalidade de alocação é utilizada em serviços CBR.

A vantagem deste tipo de alocação é que é fácil decidir se uma nova conexão deve ser aceita ou não, isto porque a rede precisa ter conhecimento somente da nova conexão que é requisitada. A nova conexão será aceita se a soma das taxas de pico de todas as conexões existentes com a taxa de pico da nova conexão requisitada é menor que a capacidade total do enlace de saída. A desvantagem deste tipo de alocação é que a não ser que as fontes transmitam nas respectivas taxas de pico o enlace de saída será subutilizado.

3.5 Modelos de fontes de tráfego e controle de admissão de conexões determinísticos

Devido à complexidade ou à limitação estatística de modelos de fontes tráfego, e à dificuldade de se policiar o tráfego que entra na rede, alguns trabalhos propõem modelos determinísticos de tráfego para redes de serviços integrados.

Um modelo de tráfego determinístico é aquele que descreve parametricamente o comportamento de pior caso de uma fonte de tráfego. Este tipo de caracterização de tráfego tem a vantagem que ele pode ser policiado pela rede [18]. Por exemplo, se uma fonte garante que o intervalo mínimo de geração de duas células consecutivas é de $X_{\min}$, então isto pode ser verificado e forçado pela rede.

Em [20], por exemplo, a fonte de tráfego tem um compromisso de tal forma que o espaçamento mínimo entre pacotes é de $X_{\min}$, o tamanho máximo do pacote é de $S_{\max}$, e que em todo intervalo de tamanho I , a fonte não pode enviar mais que I/X_{ave} pacotes, onde X_{ave} é o espaçamento médio entre os pacotes.

Em [19] a fonte de tráfego é dita satisfazer o modelo do balde furado (*leaky -bucket*) se durante qualquer intervalo de tamanho t , o número de células que a fonte transmite é

menor que $\sigma+\rho\cdot t$. O modelo (σ,ρ) pode ser visto também em termos de policiamento de tráfego. Neste modelo de policiamento a fonte deve ter um crédito (ou *token*) para transmitir uma célula na rede. Neste caso a rede obtém os créditos a uma taxa de ρ e recebe até σ créditos.

Em [18] é apresentado o modelo D-BIND (*Deterministic Bounding INterval-lenght Dependent*) onde o usuário deve informar à rede as suas características de tráfego através de pares (R_K, I_K) onde o primeiro termo representa a pior taxa (taxa máxima de transmissão) de transmissão de células durante o intervalo representado pelo segundo termo.

Uma vez que os modelos determinísticos informam condições que ocorrem no pior caso não há pacotes perdidos ou atrasados acima de determinados valores. Neste caso a rede precisa reservar recursos de acordo com o cenário de pior caso e deve ser capaz de determinar como a coleção conexões irão interagir quando multiplexadas na rede, novamente considerando o pior caso. Especificamente, um modelo de tráfego determinístico é determinado por uma função restrição. Uma função crescente $b_j(\cdot)$ é chamada função restrição da conexão *j* se durante qualquer intervalo de tempo considerado de tamanho *t* , o número de células chegando, provenientes de conexão *j*, não é maior que $b_j(t)$. Formalmente seja $A_j[t_1,t_2]$ o número de células que chegam da conexão *j* no intervalo $[t_1,t_2]$; então $b_j(\cdot)$ é a função restrição de conexão *j* se $A_j[s,s+t]\le b_j(t),\nabla s,t>0$.

Para o serviço determinístico, as vantagens do policiamento é extendido aos usuários : este pode verificar se há células perdidas ou retardadas em um valor acima daquele negociado. Por outro lado a verificação do desempenho estatístico quando o desempenho estatístico é garantido em um grande intervalo de tempo é mais complexa.

Os controles de admissão de conexões determinístico baseiam-se nas técnicas de análise de retardo de pior caso. Conceitualmente, um limite superior no retardo, definido por $q(\tau)$, pode ser derivado a partir de uma expressão do tamanho da fila em um certo instante de tempo τ em termos das chegadas expresso em termos do tráfego gerado e da velocidade do enlace de comunicação *c*. Por exemplo para uma fila do tipo FCFS temos , de acordo com [18], que :

$$q(\tau)=\max_{s\le\tau}\{\sum_{j=1}^{N}A_j[s,\tau]-c(\tau-s)\}. \qquad (3.1)$$

Usando o fato que $A_j[s, s+t] \leq b_j(t)$, juntamente com outras manipulações de acordo com [18], para a política FCFS (o primeiro que chega é o primeiro que é servido), o limite superior para o retardo é dado por :

$$d = \frac{1}{c} \max_{t \geq 0} \{ \sum_{j=1}^{N} b_j(t) - ct + \bar{s} \}. \tag{3.2}$$

Para o CAC, de acordo com [18], uma conexão deverá ser admitida se esta nova conexão não interferir no retardo máximo negociado pelas outras conexões já existentes. Desta forma o número máximo de conexões aceitas é dado por :

$$N(d) = \max\{n | \frac{1}{c} \max_{t \geq 0} \{ \sum_{j=1}^{n} b_j(t) - ct \} \leq d \}. \tag{3.3}$$

Em [18] é feito um estudo comparativo entre vários modelos de CAC determinísticos.

3.6 Reserva de Largura de Faixa Estatística

Na alocação estatística a de largura de faixa determinada para uma conexão é feita de tal forma que a soma de todas as taxas de pico de cada uma das fontes pode ser maior que a capacidade total do enlace de saída. Neste caso podemos perceber que é possível a economia de recursos quando comparando-se aos métodos de reserva de largura de faixa determinística. Por outro lado, o desempenho desta modalidade está intimamente ligada à caracterização do tráfego. Quanto mais realísticos os modelos de tráfego utilizados, mais realísticos serão os cálculos realizados para a determinação da capacidade equivalente. Isso permitirá a utilização otimizada dos recursos providos pela rede.

Apesar de serem aplicadas em redes ATM, os algoritmos de CAC atuais podem ser vistos como soluções intermediárias para o corrente problema [34]. Em particular tais soluções alocam mais recursos que aqueles exigidos para que a QoS seja satisfeita, causando sub-utilização dos recursos da rede. Normalmente o problema relacionado ao CAC é a determinação do valor da largura de faixa estatística. Matematicamente este problema pode ser modelado de tal forma que o valor da taxa agregada binária, proveniente

da multiplexação de várias fontes, ultrapasse, dentro de um determinado valor e por um certo intervalo de tempo, o valor da capacidade do enlace. A existência da fila junto ao enlace manteria o transbordo de células em níveis aceitáveis.

Normalmente estes resultados de desempenho do sistema de filas são expressos em comportamento assintótico. Tipicamente o tempo de espera na fila (tempo transcorrido da chegada da célula até sua transmissão no enlace) é utilizado para aproximar a probabilidade de perda de células. A reserva de recursos (largura de faixa equivalente, espaço em fila ou outros), problema de grande relevância para o sucesso dos algorítmos de CAC, é realizada através de aproximações do comportamento de sistemas de filas quando alimentadas por fontes de tráfegos.

Em [17,33] são realizados estudos sobre os vários mecanismos de CAC. Em tais trabalhos são feitos estudos comparativos entre vários algoritmos de CAC, dentre eles as aproximações de Capacidade Equivalente, Aproximação de tráfego intenso, Aproximação Não Paramétrica, Alocação Rápida de Fila e de Largura de Faixa e Janelas de Tempo. Estas duas referências indicam onde podem-se encontrar maiores detalhes sobre tais algoritmos. As comparações realizadas entre mecanismos de CAC normalmente são feitas levando-se em consideração o número máximo de conexões que podem ser aceitos para um grupo determinado de QoS a ser respeitado.

Tais aproximações tentam tratar o problema de uma forma geral, ou seja não são específicas para um certo modelo de tráfego. Estas aproximações normalmente sugerem como entradas do problema as taxas média e de pico da fonte, tamanho médio da rajada. Parâmetros como estes são impróprios para a descrição estatística de fontes uma vez que é de nosso conhecimento que diferentes distribuições de probabilidade podem gerar taxas média e de pico e tamanho médio de mesmo valor e diferentes graus de rajada e correlação entre os intervalos de chegada de usuário, o que pode ter participação fundamental na determinação do QoS a ser praticado. [42]

Finalmente, assumindo que o processo de chegadas de células a um comutador pode ser totalmente caracterizado por um modelo de tráfego, a próxima questão que aparece é de como o grau de rajada e a correlação dos intervalos entre chegadas interferem o comportamento das filas dos comutadores ao longo das conexões. Este é um problema aberto que precisa ser estudado.

3.6.1 Aproximação da Capacidade Equivalente

Pode-se afirmar que é necessário estimar a quantidade de recursos que serão usados durante a duração das conexões em uma rede ATM, levando-se em consideração as propriedades estatísticas das fontes de tráfego e as exigências destas conexões com relação à qualidade de serviço contratada. Durante vários anos a noção de capacidade equivalente vem sendo desenvolvida para tentar estimar esta quantidade de recursos. [14 e 15]

Para o estudo do controle de admissão de conexões faz-se o uso da modelagem do sistema multiplex, Figura 3.3, em tempo discreto, através da seguinte equação :

$$Q_{t+1} = (Q_t + A_t - C_t)^+ \tag{3.4}$$

onde $x^+ = \max(x,0)$,sendo :

- A_t número de células que chegam durante o t-ésimo segmento de tempo;
- C_t : número de células que são transmitidas em um segmento de tempo;
- Q_t conteúdo do fila (em número de células) ao final do t-ésimo segmento de tempo;

Filas deste tipo representam um modelo natural para multiplex ATM. De acordo com [43] se $E(A_t) < E(C_t)$ (e $\{A_t, C_t \text{ para } t > 0\}$ são estacionários e ergóticos) então a solução do sistema ($\Pr(Q_t \geq x)$ por exemplo) é única.

Dado um modelo de fonte de tráfego (modelo matemático que determina o comportamento do processo A_t) a taxa na qual as células devem ser transmitidas para garantir uma determinada qualidade de serviço é chamada de capacidade equivalente. Para a teoria da capacidade equivalente esta qualidade de serviço é definida como :

$$\Pr(Q > x) \leq e^{-\delta \cdot x} \tag{3.5}$$

3.6.1.1 Apresentação Matemática

Esta apresentação é baseada em [10]. Para um sistema de filas considera-se um servidor de capacidade igual a *C* células servidas (transmitidas) por segmento de tempo. Consideraremos *B* certo valor que representa um referencial do conteúdo de células na fila e *X* o conteúdo (em células) também da fila em um instante de tempo arbitrário, para um

dado processo de chegadas de células *A*. A aproximação de capacidade equivalente sugere que a qualidade de serviço (critério de desempenho) seja de acordo com (3.5)

Apesar da desigualdade na expressão (3.5), muitos trabalhos consideram a igualdade em substituição [14 e 15].

Esta aproximação tende a ser mais precisa à medida que *B* tende a infinito e serve para aproximar a probabilidade de transbordo da fila. Para uma taxa *m* (maior que a taxa média agregada das fontes), a probabilidade de que sejam geradas $n \cdot M$ células em *n* slots tal que a fila comece a transbordar a partir do *n*-ésimo slot (considerando-se inicialmente o fila vazio) e dada por :

$$\exp(-B \cdot \frac{H(m)}{m-C}) \tag{3.6}$$

onde *H(m)* depende do critério de desempenho adotado. Esta afirmativa tem como base a Teoria dos Grandes Desvios [9,10]. Assumindo que o processo *A* tenha função geradora de momento logarítmico dada por :

$$h(\delta) = \lim_{t\to\infty} 1/t \cdot \ln E[e^{\delta \cdot A(t)}] , \tag{3.7}$$

sendo ela real para todo $\delta > 0$ e diferenciável em δ . Pelo teorema de Gartner-Ellis, *H*[10] é a transformada de Legendre de $h(\delta)$ dada por :

$$H(M) = \sup_{\delta\in\Re} \{\delta \cdot M - h(\delta)\} \quad \text{com } h(\delta) = \ln E[e^{A}] \tag{3.8}$$

Definimos $\alpha(\delta)$ como a capacidade equivalente de tal forma que temos :

$$\frac{H(M)}{M - \alpha(\delta)} = \delta \quad . \tag{3.9}$$

o valor $\alpha(\delta)$ representa o valor da taxa de serviço para que a QoS (3.6) deva ser satisfeita. Sendo assim a fórmula de capacidade (largura de faixa) equivalente é dada por :

$$\alpha(\delta) = \frac{1}{\delta \cdot t} \ln E\left[e^{\delta X[0,t]}\right] \quad 0 < \delta, t < \infty \tag{3.10}$$

O trabalho apresentado em [11] faz o estudo de capacidade equivalente para vários modelos de fontes de tráfego dentre eles MMPP, cadeias de Markov, fontes do tipo ativo/inativo, etc ..., fornecendo um detalhamento maior sobre o assunto.

Algumas propriedades da capacidade equivalente :

(i) Se $X[0,t]$ tem incrementos independentes - um processo de contagem tem incrementos independentes se o número de eventos que ocorrem em tempos distintos são independentes - então $\alpha(s,t)$ não depende de *t*.

(ii) Se existe uma variável aleatória *X* tal que $X[0,t]$=Xt para t>0, então $\alpha(s,t)=\alpha(st,1)$, logo, neste caso $\alpha(s,t)$ depende de *s* e *t* através de seu produto.

(iii) Se $X[0,t] = \sum_i X_i[0,t]$ onde $(X_i[0,t])_i$ são independentes então :

$$\alpha(s,t) = \sum_i \alpha_i(s,t) \tag{3.11}$$

(iv) Para qualquer valor fixo de t , $\alpha(s,t)$ é crescente com *s* e permanece entre a média e o pico da taxa de chegada medidas em um intervalo de tamanho *t* : isto é

$$\frac{E[X[0,t]]}{t} \le \alpha(s,t) \le \frac{\hat{X}[0,t]}{t} \tag{3.12}$$

onde $\hat{X}[0,t]$ (possivelmente infinita) é o pico da taxa de chegada.

A capacidade equivalente pode ser utilizada para o controle de admissão da seguinte forma : Supomos a existência da capacidade equivalente para uma fonte do tipo *j* (j= 1,2, ..., *J*), onde o tipo de fonte pode estar relacionada às características estatísticas das fontes de tráfego, sendo N_j o número de fontes do tipo *j*. Consideramos uma fila de tamanho infinito com taxa de serviço de *c* células por segmento de tempo. Então o compromisso descrito por :

$$\sum_{j=1}^{J} N_j \cdot \alpha_j \le c \tag{3.13}$$

deverá ser respeitado para que todas as qualidades de serviço negociadas sejam satisfeitas, onde α_j é a capacidade equivalente para a fonte *j*. O critério de admissão deverá ser tal que a desigualdade seja sempre respeitada, ou seja, o somatório das capacidades equivalentes das fontes individuais deve ser menor que a capacidade do canal.

3.6.2 Considerações sobre múltiplas escalas de tempo

Normalmente pode-se dividir o tráfego oferecido à rede em várias escalas de tempo relevantes [2]. De uma maneira mais global poderiam ser: de chamadas, de rajadas e de células. Cada uma delas têm suas próprias características particulares que serão discutidas a seguir. Esta classificação de escalas de tempo não é específica; para modelos mais específicos de fontes de tráfego podem ser consideradas outras escalas com diferentes comportamentos. Por exemplo, fontes de vídeo podem ser modeladas através escalas de tempo de células, que agrupadas formam a escala de tempo de fatias (*slice*), que agrupadas formam escalas de tempo de quadros (*frames*), que agrupados formam as escalas de cenas.

3.6.2.1) Escala de tempo de chamada

Esta escala de tempo tem seu comportamento determinado de acordo com os costumes humanos. Como exemplo podemos citar o tráfego telefônico que apresenta horas de maior movimento, determinadas por circunstâncias que ocorrem na vida cotidiana. Para a análise a nível de chamadas começaremos fazendo uma pequena análise do sistema telefônico atual utilizado para dimensionamento redes atuais e consideraremos posteriormente características de outros serviços diferentes.

A intensidade de tráfego telefônico é medida pelo número de chamadas que ocorrem em um determinado período e pelo número de chamadas que estão simultaneamente em progresso. O primeiro é expresso em Erlangs dado um determinado período de tempo no qual ocorrem as medidas. O tráfego medido pode variar dia a dia, hora a hora ou mês a mês, sempre refletindo variações no comportamento dos usuários.

A rede telefônica é dimensionada para seus usuários tenham um certo grau de serviço (como em redes do tipo B-ISDN) medido em probabilidade de bloqueio de chamadas em um período de "pico" de tráfego, ou hora de maior movimento. O tráfego

neste período de pico é modelado de acordo com um processo estocástico descrito pelo processo de chegada de chamadas e o tempo de duração das chamadas.

Em redes de computadores, a noção de chamada não é necessariamente bem definida e depende se a rede é orientada a conexão ou não. Em redes orientadas a conexão podemos considerar uma chamada quando o circuito virtual está sendo utilizado. Durante esta chamada o tráfego é tipicamente em rajadas seguidas de período de inatividade. Neste caso a maior diferença que existe entre as redes telefônica e de comunicação de dados é o tempo de duração da chamada.

Em redes do tipo sem conexão, a noção de chamada não existe. Entretanto é conveniente notar que se os recursos da rede, como largura de faixa e espaço de armazenamento, precisam ser reservados o período de reserva deve ser o menor possível, mais ou menos igual ao tempo necessário para transmissão de um pacote. A interconexão de LANs, por outro lado, pode originar longas chamadas se o serviço é realizado através do estabelecimento de conexões virtuais entre os gateways que interligam elas.

O tráfego de vídeofone, ao nível da chamada, deve ser equivalente ao tráfego telefônico dada a definição do serviço. Por outro lado a vídeoconferência pode ter estatísticas diferentes : tempo de duração de chamadas de algumas horas, períodos preferenciais de início (por exemplo 10 a.m. ou 2 p.m.), reserva anterior para evitar congestionamento, de acordo com o serviço prestado atualmente.

Para aplicações convensionais baseadas em TCP/IP é apresentado um estudo em [3] sobre intervalo entre chegadas de conexões requesitadas. Neste estudo é mostrado que, em intervalos de uma hora, as chegadas de conexão TELNET e FTP são bem modeladas por um processo de Poisson. A chegada, no caso, reflete um usuário individual começando uma nova sessão. Por outro lado as chegadas de sessões WWW, SMTP(email) e NNTP(network news) não formam um processo bem definido.

3.6.2.2 Escala de tempo de rajada

No nível de rajada estamos interessados no fenômeno ocorrendo em uma escala típica cujo comportamento é do tipo on/off (como por exemplo é a duração de um frame de vídeo) ao invés de uma escala de tempo que leva em consideração o intervalo de chegada entre células. Neste caso podemos ignorar a natureza discreta das células e

considerar a chegada das células como um fluxo de taxa variável caracterizada por uma taxa instantânea. Para observarmos a dependência do comportamento da fila em relação a escala de tempo de rajada devemos considerar os instantes em que a taxa instatânea de chegada do fluxo é maior que a taxa instantânea de serviço, ou seja momentos em que o conteúdo da fila continua crescendo até que haja perda de informações por transbordo de espaço de armazenamento. Neste caso é de grande valia o conhecimento das distribuições de probalidade que rejem os tamanhos da rajada e seus momentos [21].

3.6.2.3 Escala de tempo de célula

Aqui temos que considerar a natureza discreta da célula. A componente de célula surge devido às chegadas simultâneas de células das diversas fontes quando a taxa de chegadas agregada é menor que a taxa de serviço. Esta componente depende somente da distribuição estacionária da taxa de chegada total (sendo independente, por exemplo, das distribuições de probabilidade dos tamanhos dos períodos ativos e de silêncio quando consideramos fontes atividade/inatividade) [21].

3.6.2.4 Considerações qualitativas sobre múltiplas escalas de tempo

Mas qual a influência de múltiplas escalas de tempo sobre o comportamento do sistema de multiplex que recebe o tráfego gerado por várias fontes? Em [21] são apresentadas considerações qualitativas a partir de análise da superposição de fontes ativo/inativo.

No capítulo 3 faremos maiores considerações sobre o comportamento de uma fila alimentada por fontes de tráfego. No entanto, para o entendimento do problema de múltiplas escalas de tempo consideramos aqui o seguinte modelo para uma fila:

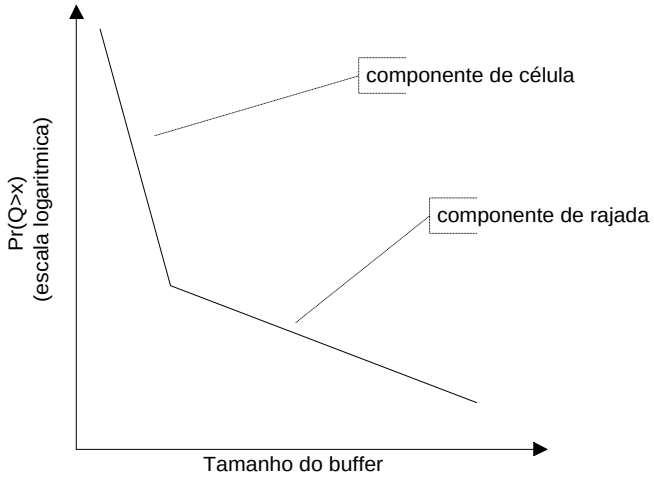
$$Q_{t+1} = (Q_t + A_t - C_t)^+ \tag{3.15}$$

onde $x^+ = \max(x,0)$,sendo :

- A_t número de células que chegam durante o t-ésimo intervalo de tempo (slot);
- C_t : número de células que são transmitidas em um intervalo de tempo (slot);
- Q_t conteúdo do fila (em número de células) ao final do t-ésimo intervalo de tempo (slot);

Quando o tráfego oferecido é fixo a probalidade de transbordo da fila (ou uma aproximação para o seu cálculo) tem o seguinte comportamento, generalizado, descrito pela figura 3.5 :

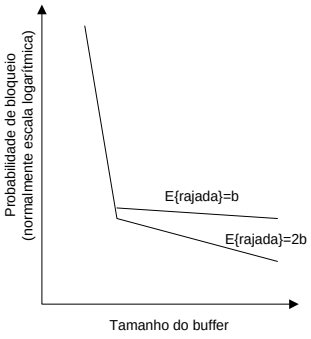
Figura 3.5

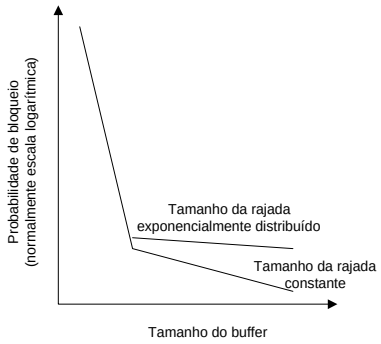


As duas componentes, já explicadas anteriormente, podem ser expressas através de probabilidades de tem relação direta : transbordode fila com taxa de chegadas menor que a capacidade do multiplex e transbordode fila com taxa de chegadas instantânea maior que a capacidade do multiplex. Na figura à esquerda abaixo podemos observar que a inclinação do componente de rajada é inversamente proporcional ao tamanho médio da rajada, enquanto que o componente de células permanece constante. O mesmo comportamento, representado na figura da direita, é observado em relação à variância da distribuição do tamanho da rajada. Uma análise mais detalhada destes comportamentos pode ser encontrada em [22].

Figura 3.6

Figura 3.7





A análise destes comportamentos pode ser utilizada no controle do tráfego em redes. Assume-se que objetivo é a probabilidade de transbordo de 10^{-9}. Desta forma temos duas opções básicas :

- O tráfego oferecido é restrito de tal forma que a probabilidade foca-se diretamente no componente de célula.

- O tráfego é maior que na opção anterior levando a uma maior ocupação porém a probabilidade só é atingida no componente de rajada.

No primeiro caso o objetivo é manter probabilidade de que a taxa de chegadas seja menor que a capacidade do enlace fique menor que 10^{-9}. Levando-se em consideração fontes ativo/inativo esta probabilidade é determinada através de sua taxa de pico e da probabilidade que a fonte esteja ativa (verificar [2] pag. 150 para maiores detalhes). Por outro lado, permitindo-se o congestionamento a nível de rajada com mais freqüência, o tamanho do fila correspondente à probabilidade da QoS adotada será função de outros parâmetros como tamanho médio da rajada, distribuição dos tamanhos dos períodos de atividade e de silêncio e possivelmente, descritivos de correlação entre sucessivos intervalos de silêncio e de atividade. Um simples exemplo desta análise pode ser vista em [23, pag. 187] onde o tamanho médio da fila M/G/1 cresce linearmente com a variância da distribuição do tempo de serviço.

3.6.3) A importância da aproximação para a previsão do comportamento da fila

A previsão do comportamento de fila é de fundamental importância para o controle de admissão de conexões, indicando o grau de utilização dos recursos da rede, fornecendo

informações sobre a possibilidade do aceite de novas conexões ou não. Dentro deste contexto verifica-se que a aproximação da capacidade equivalente indica a presença de apenas uma taxa de decaimento, não fazendo a consideração de várias escalas de tempo.

Em [40] é feito um estudo sobre os resultados de algumas aproximações do comportamento da fila (capacidade equivalente e duas outras alternativas) para mecanismos de CAC. As aproximações são utilizadas para estimar o comportamento de uma fila do tipo MAP/G/1 (MAP : Markovian Arrival Processes), isto porque a superposição de MAPs independentes (utilizadas para a representação de fontes individuais) ainda forma um MAP. Para as fontes do tipo *ativo/inativo* é utilizado o modelo IPP (*Interruted Poisson Process*) [4]. O IPP é um processo do tipo ativo/inativo. Por um período de tempo exponencialmente distribuído o processo permanece ativo. Da mesma forma acontece para o período inativo. Durante o período ativo o intervalo entre chegadas de usuários é exponencialmente distribuído, enquanto que no período ativo não há chegadas. Todos os processos são mutuamente independentes. O estudo é baseado em três aproximações :

$$\begin{aligned} &\Pr(Q > x) \approx e^{-\delta \cdot x} \\ &\Pr(Q > x) \approx \alpha \cdot e^{-\delta \cdot x} \\ &\Pr(Q > x) \approx \alpha_1 \cdot e^{-\eta_1 \cdot x} + \alpha_2 \cdot e^{-\eta_2 \cdot x} + \alpha_3 \cdot e^{-\eta_3 \cdot x} \end{aligned} \tag{3.16}$$

onde a primeira corresponde a da capacidade equivalente. A constante α na segunda aproximação está relacionada ao número de fontes que estão sendo superpostas, sendo mostrado numericamente no mesmo trabalho que se tem o seguinte comportamento :

$$\alpha_n \approx \beta \cdot e^{-n \cdot \gamma} \text{ a medida que } n \to \infty \tag{3.17}$$

sendo n o número de fontes, com β>1, $\gamma > 0$ para fontes com grau (utilizado o IDC) de rajada maiores que o processo de Poisson e $\gamma < 0$ para menores graus de rajada. O valor de $|\gamma|$ tende a ser maior à medida que o grau de rajada se afasta do processo de Poisson.

Na terceira aproximação os parâmetros α_1 e η_1 são obtidos da seguinte forma : o primeiro da mesma forma que o fator α na segunda aproximação e o segundo é calculado através da aproximação da capacidade equivalente. Os parâmetros $\alpha_2, \alpha_3, \eta_2$ e η_3 são determinados de forma que sejam igualados os valores de $\Pr[Q > 0]$ e dos três primeiros momentos $E[Q], E[Q^2]$ e $E[Q^3]$ de forma que $\eta_1 \leq \min(\eta_2, \eta_3)$. Através destas três aproximações estima-se o comportamento do algoritmo numérico "exato" que representa a

fila MAP/G/1. Pode-se perceber que, através das três taxas de decaimento, é feita uma tentativa de se modelar várias escalas de tempo.

A comparação entre as três aproximações é feita através do número de fontes que podem ser aceitas para uma qualidade de serviço fixo, $\Pr(Q > x)$. Cada fonte, IPP, utilizada na comparação é caracterizada por :

- período médio de atividade : 436.36;

- período médio de inatividade : 4363.63;

- taxa de pico durante o período de *atividade* : 0.1375;

com isso temos a taxa média da fonte igual a 0.1375 · 436.36 / (436.36 + 4363.63) = 0.0125 e intervalo médio entre chegadas de 1/0.0125=80. A razão pico média é dada por 0.1375/0.0125 = 11.0. A média do número de células durante o período de *atividade* é 0.1375 · 436.36 = 60.0. O serviço é considerado determinístico. É considerado um fila de tamanho 600 e considera-se $\Pr(Q > 600) \approx 10^{-9}$ como a qualidade de serviço desejada. O cálculo "exato" indica que 24 fontes podem ser admitidas (utilização de 30%). Com isso pode ser feita a comparação entre as várias metodologia de reserva de recursos através da seguinte tabela :

Figura 3.8

Método de Computação	Número de fontes permitidas para fila de tamanho = 600	Tamanho de fila para suportar *n* = 24 fontes
exato	24	600
$\Pr(Q > x) \approx e^{-\delta \cdot x}$	12	1146
$\Pr(Q > x) \approx \alpha \cdot e^{-\delta \cdot x}$ e $\Pr(Q > x) \approx \alpha_1 \cdot e^{-\eta_1 \cdot x} + \alpha_2 \cdot e^{-\eta_2 \cdot x} + \alpha_3 \cdot e^{-\eta_3 \cdot x}$	25	530
alocação pela taxa média	80	não aplicável
alocação pela taxa de pico	7	não aplicável

Este estudo é importante por explorar o impacto provocado pelo método de reserva de recursos. Por exemplo uma operadora que utiliza um modelo impreciso poderá fazer o

uso equivocado de seus recursos, aceitando menos conexões que a rede pode realmente aceitar. Por outro lado uma operadora, fazendo o uso de uma aproximação mais precisa pode aceitar um número maior de conexões dispondo dos mesmos recursos. Uma vez que esta aceita mais conexões que a outra, mantendo a mesma qualidade de serviço e utilizando os mesmos recursos, o seu lucro será maior.

3.6.3 Considerações sobre múltiplas escalas de tempo no processo de modelagem

Como verificado anteriormente, os processos de chegadas utilizados para o modelo de fontes de tráfego exibem uma estrutura em múltiplas escalas de tempo (por exemplo as escalas de células, rajadas e chamadas). Mais especificamente isto pode ser verificado através de informações associadas ao comportamento da distribuição do tamanho da fila (por exemplo $\Pr(Q > x)$) de um sistema do tipo $Q_{t+1} = (Q_t + A_t - C_t)^+$. A consideração de várias escalas de tempo corresponde a várias taxas de decaimento quando verificamos o comportamento de $\Pr(Q > x)$.

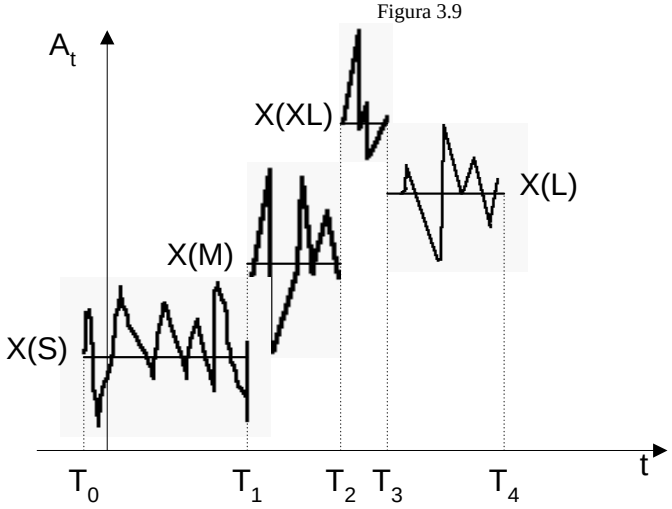
Por exemplo uma fonte de vídeo pode ser vista como composta de vários blocos hierárquicos : grupos de células formam uma fatia (*slice*), grupos de fatias que formam um quadro (*frame*) e grupos de frames que formam uma cena. Cada um destes componentes correspondem a uma diferente escala de tempo que possuem diferentes propriedades estatísticas. Pode-se ainda identificar a escala de chamada que tem características mais variáveis durante o dia e que tem grande influência do comportamento humano.

Um exemplo de consideração de várias escalas no problema de estimação do comportamento da fila é dado em [41, 42]. Em [42] uma amostra de vídeo MPEG o conjunto de 12 frames forma 1(um) grupo de imagens (gdi) e uma cena é composta por vários gdis. A mudança de uma cena para outra é obtida empiricamente de acordo com [45]. No caso a cena será a unidade básica da escala de tempo lenta e a escala de tempo mais rápida é formada pelos gdis (mais especificamente caracterizada pelos tamanhos dos gdis). As cenas (regimes) são caracterizadas pelos tamanhos médios gdis durante elas e são classificadas em **S**,**M**,**X**,**XL**. Sendo assim podemos definir uma matriz de transição entre os regimes :

$$P = \begin{bmatrix} 0 & p_{S,M} & p_{S,L} & p_{S,XL} \\ p_{M,S} & 0 & p_{M,L} & p_{M,XL} \\ p_{L,S} & p_{L,M} & 0 & p_{L,XL} \\ p_{XL,S} & p_{XL,M} & p_{XL,L} & 0 \end{bmatrix}$$

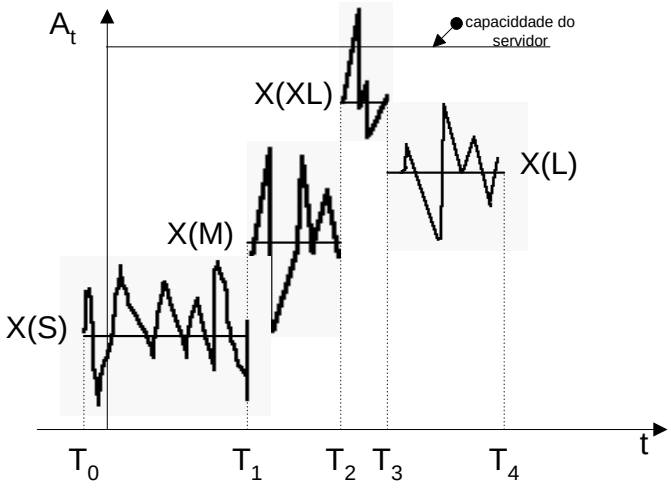
Seja $T = \{T_0, T_1, \ldots\}$ uma seqüência de instantes de renovação, sendo este processo representativo dos instantes em que a sequência de vídeo muda de um regime para o outro. Seja $M = \{M_n, n \geq 0\}$ uma cadeia de Markov com espaço de estados {**S**,**M**,**X**,**XL**} e matriz de taxa de transição entre estados definida por *P* (dada anteriormente). Define-se também o processo $B_t, t = 0,1,2,\ldots$ que assume os valores dentro do conjunto {**S**,**M**,**X**,**XL**}e representa o indicador do regime no qual o stream de vídeo está no instante *t* , isto é, $B_t = M_n, T_{n-1} < t \leq T_n, n \geq 1$. B_t é um processo de renovação Markoviano e denota-se por $\pi_i = \Pr[B_t = i]$ a probabilidade, no estado estacionário, que o processo esteja no regime *i*. Para finalizar define-se quatro processos independentes $X(i) = \{X_t(i), t \geq 0\}$ com média λ_i cada, cada *X(i)* sendo independente de *T* e *M* , $i \in$ {**S**,**M**,**X**,**XL**}.Cada processo *X(i)* modela a escala de tempo rápida para o regime *i*. Finalmente modela-se a sequência de vídeo como um processo da forma $A_t = X_t(B_t)$ onde o processo do regime *B* constitui o componente de escala de tempo lenta, e que modula o componente de escala de tempo rápida *X* . Estas idéias podem ser visualizadas na figura 3.9.

Figura 3.9



O comportamento do sistema de filas estritamente estável ocorre quando a capacidade do multiplex é maior que a taxa média de chegada no regime que apresenta o "pior" caso λ_{XL}. Neste caso o tamanho da fila aumenta somente quando a taxa de chegada instantânea é maior que a capacidade do servidor, o que tende a acontecer quando surgem grandes picos ocasionais como pode ser verificado na figura 3.10 .

Figura 3.10



Logo, para situações deste tipo, o problema principal é modelar estes picos (escala de tempo mais rápida) enquanto a modelagem da duração dos regimes (escala de tempo mais lenta) perde a importância. Neste caso o comportamento da fila pode ser aproximado pelo seguinte resultado de superposição [41]:

$$\Pr(Q_t = x) \approx \sum_{i \in \{S,M,L,XL\}} \pi_i \cdot \Pr(Q_t(i) = x) \tag{3.18}$$

onde $Q_t(i)$ representa o tamanho da fila obtido pela alimentação da fila somente com o regime i . Para este caso de fila estritamente estável a concentração pode ser focada nos regimes separadamente, considerando uma média ponderada deles sem se preocupar com a estrutura de dependência de longo termo (distribuição de probabilidade dos tempos em que o processo global passa em cada regime). Com esta formulação e condição percebemos que o comportamento do sistema é determinado pelas escalas de tempo mais rápidas, não sendo consideradas as mais lentas (distribuições de probabilidade dos tamanhos dos regimes).

Para o comportamento da fila com estabilidade fraca (capacidade do servidor da fila menor que a taxa média de chegadas no regime de pior caso e maior que a média total) a situação não é a mesma. Apesar deste cenário não ser apropriado no contexto de serviços em tempo real e interativos é interessante entender o comportamento em um cenário com longas filas e tolerâncias a grandes retardos de tempo.

Capítulo 4

Extensão a um Modelo de Tráfego Auto-Similar

O presente capítulo apresenta uma extensão a um modelo de tráfego auto-similar. Inicialmente é apresentado o modelo auto-similar original [49] e posteriormente a extensão proposta para ele. O modelo original parte da análise de uma fila, que pode ser considerada como um multiplex, considerando que os usuários do sistema são rajadas, provenientes de várias fontes de tráfego quando multiplexadas. Uma vez obtido o modelo através da extensão este é utilizado para alimentar um sistema de filas representativo de um multiplex ATM. Em seguida é feita a análise do sistema gerando-se resultados de desempenho da fila que podem ser utilizados em controle de admissão de conexões e mecanismos de tarifação.

As contribuições do presente trabalho são as seguintes :

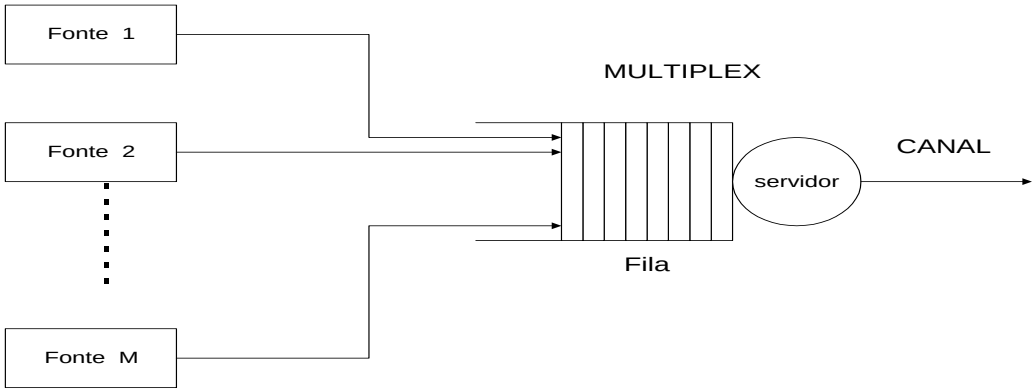
1) A análise original, em [49], é feita em termos de rajadas enquanto que a extensão considera células. Isto é importante uma vez que os parâmetros de desempenho em redes ATM (probabilidade de perda, distribuição do tempo de espera nas filas, etc...) são obtidos em termos células. Também pode-se dizer que rajadas não são muitos representativas quando se faz a análise de um sistema de filas, uma vez que estas tem tamanhos variados. Por exemplo, qual seria o sentido de se dizer que em média existem dez rajadas em uma fila ?

2) Análises de um sistema de filas quando este é alimentado pelo modelo de tráfego obtido em termos de células. Uma análise é feita observando-se o sistema somente quando ocorre a saída de uma célula dele. Uma segunda análise comparativa é feita utilizando-se duas aproximações assintóticas. É verificada a importância da precisão das aproximações, comparando-se as quantidades de largura de faixa e de espaço em fila que são reservadas para uma mesma qualidade de serviço.

4.1 O Modelo do Multiplex

Considera-se *M* fontes homogêneas de tráfego, alimentando uma fila em comum. Este modelo é representado na figura 4.1.

Figura 4.1 : esquema do multiplex



O objetivo neste momento é construir o modelo de tráfego agregado, formado pela superposição dos modelos das fontes individuais. Neste sentido queremos que o modelo agregado reflita as seguintes características :

i) A fonte individual comporta-se da seguinte forma : em momentos aleatórios de tempo ela começa a gerar células a uma taxa constante *R* , em bits/segundo (estado *ativo*). Uma rajada corresponde ao número de células (de tamanho fixo em bits) geradas durante o estado ativo. A distribuição de probabilidade de cada um dos tamanhos dos intervalos de tempo em que uma fonte encontra-se gerando células (rajada) é *heavy-tail*. Ao final do período de geração de células a fonte permanece inativa por um período aleatório de tempo (estado *inativo*).

ii) O número de fontes individuais *M* é grande, sendo considerado infinito, mas a intensidade de tráfego gerada pela agregação de fontes pode ser considerada finita.

iii) A superposição de fontes de tráfego, que gera o tráfego agregado é um processo assintoticamente auto-similar com parâmetro de Hurst $H > 0.5$.

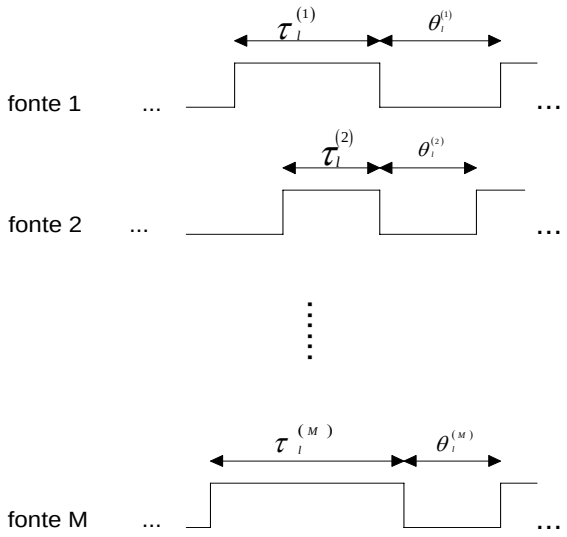
A partir destas considerações deseja-se construir um modelo de tráfego matematicamente tratável. Uma vez que é obtido o modelo este é utilizado para alimentar

um sistema de fila, para que sejam obtidas diversas medidas de desempenho. O sistema de serviço no sistema de filas será oportunamente explicado.

4.2 Modelo da Fonte Individual

As diversas fontes individuais (fonte 1, fonte 2, ..., fonte *M*) são consideradas independentes. Cada fonte, em qualquer instante de tempo $t \in I_{-\infty} = \{\ldots,-1,0,1,2,\ldots\}$, pode estar em um dos estados : ativo ou inativo. Quando no estado ativo a fonte gera células a uma taxa constante R . No estado inativo não há geração de células. O tempo gasto pela fonte *i* no *l*-ésimo estado ativo é chamado período ativo *l* e é denotado por $\tau_l^{(i)} \in \{1,2,3,\ldots\}$ com $l \in I_{-\infty}$. Da mesma forma, para o período inativo que se segue temos $\theta_l^{(i)} \in \{1,2,3,\ldots\}$ com $l \in I_{-\infty}$. Estes períodos ativo e inativo irão formar o *l*- ésimo ciclo da fonte *i* de tamanho $\tau_l^{(i)} + \theta_l^{(i)}$. As variáveis aleatórias $\tau_l^{(i)}$ e $\theta_l^{(i)}$ são supostas independentes, sendo $i \in \{1,2,\ldots, M\}, l \in I_{-\infty}$. Esta descrição pode ser visualizada na Figura 4.2.

Figura 4.2 : representação das *M* fontes individuais nos estados ativos e inativos .



As variáveis aleatórias $\tau_l^{(i)} \in \{1,2,3,\ldots\}$ com $l \in I_{-\infty}$ tem distribuição idêntica tal que:

$$\Pr\{\tau > t\} \approx t^{-\alpha},\ t \to \infty,\ 1 < \alpha < 2 \qquad (4.1)$$

onde *t* é a variável genérica $\tau_i^{(l)}$. A equação acima quer dizer $\tau_i^{(l)}$ que tem distribuição de Pareto com média finita, $a_\tau = E[\tau] < \infty$ e variância infinita. As variáveis aleatórias $\theta_l^{(i)}$,

$i \in \{1,2,...,M\}, l \in I_{-\infty}$ são identicamente distribuídas de acordo com alguma distribuição genérica $\Pr\{\theta \le t\}$,com média finita, $a_\theta = E[\theta] < \infty$.

A distribuição de Pareto faz parte da família de distribuições *heavy-tail*. Tais distribuições podem ser usadas para caracterizar densidades de probabilidades que descrevem processos de tráfego como intervalos entre chegadas de pacotes em redes de dados e tamanhos de rajadas. A distribuição de uma variável aleatória *X* é dita *heavy-tail* se :

$$\Pr\{X > x\} \approx \frac{1}{x^\alpha}, \text{ a medida que } x \to \infty, \ \alpha > 0$$

4.3 Modelo do Tráfego Agregado

Denota-se por

$$\omega^{(i)} = (..., \omega_{-1}^{(i)}, \omega_0^{(i)}, \omega_1^{(i)}, ...) \tag{4.2}$$

as seqüências de instantes de início dos períodos ativos da fonte, logo $\omega_{l+1}^{(i)} - \omega_l^{(i)} = \tau_l^{(i)} + \theta_l^{(i)}$. A partir da hipótese de independência das fontes, os processos $\omega^{(1)}, \omega^{(2)}, ..., \omega^{(M)}$ são mutuamente independentes e identicamente distribuídos. Mais especificamente $\omega^{(i)}$ é um processo de renovação, tendo os intervalos entre chegadas como pontos de renovação distribuídos com a seguinte regra :

$$\Pr\{\omega_{l+1}^{(i)} - \omega_l^{(i)} = k\} = \Pr\{\tau_l^{(i)} + \theta_l^{(i)} = k\} \tag{4.3}$$

não dependendo de *l* nem de *i*, uma vez que as fontes são superpostas são estatisticamente idênticas. Levando-se em consideração que $a_\tau + a_\theta < \infty$, pode-se assumir que o processo de renovação $\omega^{(i)}$ é estacionário.

Denota-se a superposição dos processos de renovação $\omega^{(1)}, \omega^{(2)}, ..., \omega^{(M)}$ por

$$\omega(M) = (..., \omega_{-1}(M), \omega_0(M), \omega_1(M), ...).. \tag{4.4}$$

A seqüência $\omega(M)$consiste dos $\omega_l^{(i)}$, nos mesmos instantes que estes últimos aparecem. Seja um componente $\omega_S(M)$ o instante de início de um período ativo *l* para a fonte *i* (isto é, $\omega_S(M) = \omega_l^{(i)}$) com tamanho de período ativo $\tau_l^{(i)}$. $\tau_l^{(i)}$ pode ser considerado

como marca da componente $\omega_S(M)$ e pode ser denotado como $\omega_S(M)$. Forma-se então o seguinte processo agregado :

$$\{\omega(M), \tau(M)\} = [\ldots, \omega_{-1}(M), \tau_{-1}(M), \omega_0(M), \tau_0(M), \ldots] \tag{4.5}$$

com todas as marcas $\tau_S(M)$ sendo mutuamente independentes. O modelo do tráfego agregado que está sendo construído será descrito pelo processo $\{\omega(M), \tau(M)\}$ com $M \to \infty$.Denota-se o processo no limite, ou seja, com $M \to \infty$, por (ω_S, τ_S), $s \in I_{-\infty}$. Não se alterando as distribuições dos períodos ativos e inativos quando $M \to \infty$ a intensidade de tráfego tende a infinito. Para evitar isso aumenta-se a_θ de tal forma que a intensidade de tráfego $\lambda = M/(a_\theta + a_\tau)$ mantenha-se constante.

Denota-se o número de períodos ativos que aparecem em qualquer segmento de tempo $[t_i,t_{i+1}), \mathrm{i} \in I_{-\infty}$ no processo de superposição $\omega(M)$ por $\xi_{t_i}(M)$. A seqüência $\xi(M) = (\ldots, \xi_{t_{-1}}(M), \xi_{t_0}(M), \xi_{t_1}(M), \ldots)$ é um processo aleatório no eixo dos tempos $I_{-\infty}$ tendo as componentes aleatórias $\xi_{t_i}(M)$ com valores no conjunto $I_0 = \{0,1,2,3,\ldots\}$.

Pode ser mostrado [51], que para qualquer $t_1, t_2, \ldots, t_n \in I_{-\infty}, t_i \neq t_j$, $n \in \{1,2,3,\ldots\}$, quando $M \to \infty$,fazendo com que o valor $\lambda = M/(a_\theta + a_\tau)$ e a distribuição $\Pr\{\tau \le t\}$ não se alterem, mas $\Pr\{\theta \le t\} \to 0$ para qualquer $t < \infty$ então :

$$\Pr\{\xi_{t_i}(M) = k_i\} = \Pr\{\xi_{t_i}(\infty) = k_i\} = \frac{e^{-\lambda} \cdot \lambda^{k_i}}{k_i!} \tag{4.6}$$

ou seja, quando $M \to \infty$, a seqüência $\xi = \xi(\infty)$ no limite é um processo de Poisson no eixo discreto dos tempos.

Seja Y_t a taxa total de geração de células para o tráfego agregado no tempo $t \in I_{-\infty}$. Cosidere a fonte *i*. Denota-se por $y_t^{(i)}$ a geração de taxa de células para a fonte *i* no instante *t* , então :

$$y_t^{(i)} = \begin{cases} P & \text{se a fonte está ativa,} \\ 0 & \text{se não está ativa.} \end{cases}$$

Como o processo é estacionário, a probabilidade que a fonte *i* esteja no estado ativo é $p = a_\tau/(a_\tau + a_\theta)$. Logo tem-se para uma fonte individual :

$$E\{y_t^{(i)}\} = R \cdot p, \quad E\{(y_t^{(i)})^2\} = R^2 \cdot p \tag{4.7}$$

A partir da consideração de independência entre as fontes pode-se escrever :

$$E\{\sum_{i=1}^{M} y_t^{(i)}\} = M \cdot R \cdot p \quad , \; E\{\sum_{i=1}^{M} (y_t^{(i)})^2\} = M \cdot R^2 \cdot p \tag{4.8}$$

No limite, quando $M \to \infty$ e quando $\lambda = M/(a_\theta + a_\tau)$, verifica-se que :

$$\begin{aligned} E\{Y_t\} &= R \cdot \lambda \cdot a_\tau, \\ E\{(Y_t - E\{Y_t\})^2\} &= R^2 \cdot \lambda \cdot a_\tau \end{aligned} \tag{4.9}$$

Em [49] é mostrado que o processo de tráfego agregado $Y = (\ldots, Y_{-1}, Y_0, Y_1, \ldots)$ é assintoticamente auto-similar com parâmetro de Hurst $H = (3-\alpha)/2 > 0.5$. Para isso é achada a função autocorrelação *r(k)* de $Y^{(m)} = (Y_1^{(m)}, Y_2^{(m)}, \ldots)$. O processo $Y^{(m)} = (Y_1^{(m)}, Y_2^{(m)}, \ldots)$, *m*=1,2,..., é definido por $Y_k^{(m)} = (Y_{km-m+1} + \ldots + Y_{km})/m$, $k > 0$ quando $k \to \infty$. A função autocorrelação de Y_t é dada por :

$$r(k) = \frac{E\{(Y_t - E\{Y_t\})(Y_{t+k} - E\{Y_t\})\}}{E\{(Y_t - E(Y_t))^2\}}$$

Mais especificamente, é provado em [49], que a função autocorrelação do processo $Y^{(m)}$, $r^{(m)}(k)$, tem o segunte comportamento :

$$\lim_{m \to \infty} r^{(m)}(k) = \frac{1}{2}\left[(k+1)^{3-\alpha} - 2 \cdot k^{3-\alpha} + (k-1)^{3-\alpha}\right] \tag{4.10}$$

Este comportamento é característico de processos assintoticamente auto-similares.

4.4 A análise da fila

Para a análise de fila alimentada pelo tráfego agregado obtido na seção anterior considera-se a disciplina FIFO para os usuários do sistema (rajadas) mas não para as células (que compõem as rajadas). O sistema de fila que pode descrever este sistema de

transmissão, a cada segmento de tempo, é do tipo *M/G/1* onde *M* representa a distribuição de Poisson ξ_t , o número de fontes que ficam ativas no intervalo de tempo *t*, e *G*, que é independente de ξ_t, e representa a distribuição de probabilidade dos tempos de serviço τ_S para as diferentes rajadas, no caso, os usuários do sistema. O servidor do sistema tem seu tempo de serviço a uma determinada rajada determinado pelo tamanho dela; quanto maior a rajada maior o tempo para a realização do serviço.

Considera-se o referido sistema de fila *M/G/1* nos períodos de tempo $\ldots, t_{-1}, t_0, t_1, \ldots$quando as sucessivas fontes concluem suas transmissões, ou seja, quando os usuários têm seus serviços concluídos. Seja N_k o número de fontes que permanecem na fila no instante t_k , ao final de um serviço fornecido a uma rajada. Consideremos a cadeia de Markov $\ldots, N_{-1}, N_0, N_1, \ldots$. Se

$$E[\xi_t] \cdot \lambda < \mu = \frac{1}{E[\tau_S]} \tag{4.11}$$

a cadeia de Markov N_k tem distribuição estacionária $q_i = \Pr[N_k = i]$, $i = 0,1,2,\ldots$, com $Q(z) = \sum_{i=0}^{\infty} q_i \cdot z^i$ (Transformada *z* da distribuição q_i) . Seja $\rho_j = \Pr\{\tau = j\}, j = 1,2,3,\ldots$a distribuição dos tempos de serviço dos usuários do sistema. As probabilidades q_i podem então ser encontradas como soluções das equações :

$$\begin{aligned} q_j &= \sum_{i=0}^{j} p_i \cdot q_{j-i+1} + p_j \cdot q_0, j = 0,1,2,\ldots \\ \sum_{j=0}^{\infty} q_j &= 1, \end{aligned} \tag{4.12}$$

onde p_i é a probabilidade de que *i* novas fontes entrem na fila durante a transmissão (serviço) de um usuário (rajada), dada por :

$$p_i = \sum_{j=1}^{\infty} \rho_j \cdot \frac{(j \cdot \lambda)^i}{i!} e^{-j\lambda} \tag{4.13}$$

ainda temos que $q_0 = Q(0) = 1 - \lambda \cdot a_\tau$. A partir daí podemos calcular as probabilidades q_i restantes.

4.5 Extensão ao Modelo Auto-Similar

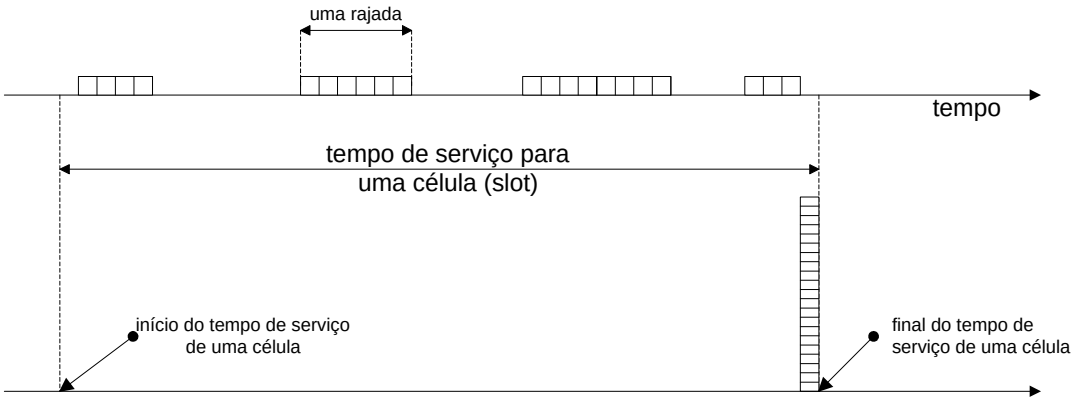
Para realizar uma extensão do estudo da fila, cujo o tráfego de entrada foi descrito na seção anterior, utililiza-se um modelo de fila discreto (ou seja o tempo é dividido em segmentos discretos de tempo). Considera-se que o tráfego de entrada composto de usuários (células de tamanho fixo em bits), um espaço para armazenar estes usuários (espaço na fila em comum) e um servidor que serve os usuários, se estes estiverem presentes na fila.

Considera-se aqui o estado do sistema como sendo o número de unidades encontradas nele, no momento de saída de um usuário do sistema. Cosidera-se ainda que, quando um usuário tem seu serviço finalizado ele sai do sistema e, imediatamente, o primeiro usuário encontrado no início da fila será servido, se este usuário existe de fato.

Inicialmente, determina-se que o tempo de serviço para um usuário (no caso uma célula) seja determinístico e igual a um segmento discreto de tempo. Durante o intervalo de tempo utilizado para servir uma célula podem chegar várias outras ao sistema de fila. Estas chegadas são reguladas por uma certa distribuição de probabilidade. Como indicado no item anterior, considera-se que, durante um segmento de tempo, o número de períodos ativos (rajadas) que surgem neste intervalo é dado por uma distribuição de Poisson. Ou seja durante o serviço de uma célula o número de rajadas que chegam ao sistema de fila é dado pela distribuição de Poisson.

As rajadas, por sua vez, são compostas por um certo número de células. Dentro desta idéia uma rajada gera um número de células que é proporcional ao seu tamanho. Como a distribuição do tamanho da rajada é *heavy-tail* o número de células que compoem as rajadas também é regulado por uma distribuição *heavy-tail*. Logo, durante um período de serviço de uma célula, podem chegar várias outras decorrentes das chegadas das várias rajadas. Considerando-se os períodos de observação do sistema, para efeito de cálculos, como o final de um serviço de célula, pode-se abstrair das rajadas em si e considerar-se, somente, o grupo de células (somatório das células componentes das várias rajadas). Esta idéia pode ser melhor observada na Figura 4.3.

Figura 4.3 : esquema de visualização do sistema de filas do nível de rajadas para o nível de células



Dentro deste contexto, onde já é conhecido que o número de rajadas dentro de um segmento de tempo é dado por uma distribuição de Poisson e que os tamanhos das rajadas (através do número de células geradas pelas rajadas) são dados por uma distribuição *heavy-tail*, pode-se calcular a distribuição de probabilidade do número de células que chegam ao sistema em um segmento de tempo.

Sendo h_i a seqüência representativa da distribuição de probabilidade do tamanho das rajadas em células, ou seja, $h_i = \Pr\{I = i\}$ onde a variável aleatória I representa o tamanho da rajada em células, sabe-se que a probabilidade de chegarem *i* células dado que chegaram *n* rajadas é dada por y_i , onde:

$$y_{i|n} = h_i * h_i * \ldots * h_i \quad n \text{ vezes.} \tag{4.14}$$

O símbolo $*$ representa a convolução discreta entre as seqüências. Ou seja, a distribuição de probabilidade da soma de variáveis aleatórias independentes e idênticamente distribuídas é dada pela convolução discreta das seqüências representativas das distribuições individuais, h_i $(i = 0,1,2,\ldots)$. Sabendo-se que a distribuição de probabilidade da variável aleatória que representa o número de rajadas (grupo de células) durante um segmento de tempo é dada pela distribuição de Poisson, dada pela seqüencia a_n $(n = 0,1,2,\ldots)$, determina-se a seqüência b_i $(i = 0,1,2,\ldots)$, distribuição de probabilidade do número de células que chegam ao sistema de fila em um segmento de tempo, provenientes das diversas fontes, através da seguinte expressão:

$$b_i = \sum_{n=1}^{\infty} y_{i|n} \cdot a_n \tag{4.15}$$

ou seja, a seqüência b_i $(i = 0,1,2,...)$ representa a modelo de tráfego, baseado em células, que alimentará o sistema de fila que representa o multiplex. Passa-se então à metodologia de resolução da fila alimentada por tal modelo de tráfego.

4.5.1 Apresentação matemática

O modelo matemático utilizado para esta seção é baseado em [50]. Em um segmento de tempo chegam *i* novas células ao sistema com probabilidade b_i $(i = 0,1,2,...)$. Nesta seção calcula-se a probabilidade p_i (estado *i*) de que, ao final de um segmento de tempo em que houve serviço, a célula que sai "deixe" *i* células no sistema. Por exemplo se o presente estado é *i* , com *i* diferente de zero, então o próximo estado será (*i+j-1*) com probabilidade b_j se *j* células, provenientes das várias fontes, chegam ao sistema durante o próximo período de serviço. Se o sistema está vazio em um dado segmento de tempo então a probabilidade de que o próximo estado seja *i* é dado por :

$$\text{Prob}[0 \to i] = \sum_{k=1}^{i+1} c_k \cdot b_{i-k+1} \tag{4.16}$$

onde c_k é a seqüência que representa a distribuição de probabilidade do número de usuários que chegam durante um segmento de tempo dado que ocorreram chegadas. $c_k \cdot b_{i-k+1}$ representa a probabilidade conjunta de que o número de células que chegam ao sistema no próximo segmento de tempo seja *k* e que (*i-k+1*) células cheguem durante o período do serviço sucessivo. A partir das probabilidades de transição podemos escrever as equações diferenciais para obtenção dos valores de $p_i, i = 0,1,2,...$:

$$\begin{aligned} p_0 &= p_0 c_1 b_0 + p_1 b_0 \\ p_1 &= p_0 (c_1 b_1 + c_2 b_0) + p_1 b_1 + p_2 b_0 \\ p_2 &= p_0 (c_1 b_2 + c_2 b_1 + c_3 b_0) + p_1 b_2 + p_2 b_1 + p_3 b_0 \\ &... \end{aligned} \tag{4.17}$$

onde $p_0 = \frac{1-\rho}{c_{med}}$ com ρ sendo o fator de utilização do sistema (número médio de usuários que chegam durante um intervalo de serviço-slot) e c_{med} representa o tamanho médio da distribuição de probabilidade c_k. A descrição do modelo é completada com a obtenção da distribuição de probabilidade b_k que já foi mencionada anteriormente. Vale ainda ressaltar que temos a seguinte relação entre b_k e c_k dada por :

$$c_k = \frac{b_k}{1-b_0}. \tag{4.18}$$

4.5.2 Comportamento Assintótico

Pode-se identificar duas limitações do sistema de fila apresentado na seção anterior:

1) Os estados do sistema (número de células deixados na fila por uma célula que sai do sistema de fila) só são obtidos se há a saída de uma célula do sistema.

2) A incapacidade de transmissão de várias células durante um mesmo segmento de tempo (anteriormente apenas uma célula poderia ser servida durante um mesmo segmento de tempo).

Uma situação mais realista seria capaz de observar o sistema em todos os segmentos de tempo independentemente da saída de células ou não, e considerar o caso em que é possível a transmissão de várias células durante um mesmo segmento de tempo (com isto seria possível aumentar a capacidade do servidor). Um sistema que pode representar estas duas situações pode ser modelado pela seguinte equação :

$$Q_{t+1} = (Q_t + A_t - C_t)^+ \tag{4.19}$$

onde $x^+ = \max(x,0)$,sendo :

- A_t número de células que chegam durante o t-ésimo segmento de tempo;
- C_t : número de células que são transmitidas em um segmento de tempo;
- Q_t conteúdo da fila em comum (em número de células) ao final do t-ésimo segemento de tempo;

Sistemas de filas deste tipo representam um modelo mais realístico do multiplex ATM. De acordo com [43] se $E(A_t) < E(C_t)$ (e $\{A_t, C_t$ para $t > 0\}$ são estacionários e ergóticos) então a solução do sistema ($\Pr(Q_t \geq x)$ por exemplo) é única. Como já mensionado no capítulo 3 informações deste tipo tem aplicações diretas em controle de admissão de conexões.

A motivação para a adoção do modelo de tráfego, apresentado até este momento, parte da obtenção de vários resultados que mostram que as funções distribuição dos processos de chegadas que parecem em redes podem apresentar propriedades *heavy-tail*. Para tais processos as condições de Cramer não são satisfeitas. Um processo A_t satisfaz às condições de Cramer se sua função geradora de momento logaritmico de A_t ,

$$\alpha(\delta) = \lim_{t\to\infty} \frac{1}{t} \ln\{E(e^{A_t \cdot \delta})\} \tag{4.20}$$

existe, é finita para para todo real $\delta > 0$ e $\alpha(\delta)$ é diferenciável. Para processos que satisfazem as condições de Cramer, a distribuição de probabilidade do tamanho da fila em número de usuários, assintoticamente, tende a uma exponencial simples, conforme foi verificado na teoria da largura de faixa equivalente no capítulo 3.

Para o modelo de tráfego apresentado em 4.5, observa-se que será necessária a truncagem do tamanho máximo do número de células. Isto ocorre porque é utilizada uma distribuição de probabilidade *heavy-tail* (subexponencial) para reger o comportamento do tamanho da rajada, em células. Com esta truncagem, a função geradora de momento logaritmico existe e, consequentemente, o comportamento assintótico da distribuição do tamanho da fila é exponencial. Uma distribuição de probabilidade subexponencial tem um decaimento mais lento que uma distribuição exponencial em relação a variável aleatória. Esta característica será indicada em figuras apresentadas posteriormente.

Entretanto, de acordo com [51], o intervalo em que a assíntota exponencial torna-se precisa, pode estar fora do intervalo de interesse para o projeto de redes (como por exemplo considerando uma fila de tamanho que não seja realístico). Pode ser possível que, até mesmo nos casos de chegadas limitadas (em que a aproximação da largura de faixa equivalente existe), a parte relevante para o projeto, da distribuição do tamanho da fila, apresente um comportamento diferente do exponencial. Com isso a aproximação exponencial não será uma maneira correta de representar o comportamento do sistema de filas. Por outro lado, o caso em que apresentam-se chegadas *heavytail* com chegadas limitadas, pode apresentar uma maneira mais precisa de representar o comportamento do sistema em regiões relevantes ao projeto. Este caso será estudado neste capítulo posteriormente.

Dentro deste contexto podemos definir, como em [43], os processos de chegadas como sendo subexponenciais e exponenciais. Informalmente, a diferença básica entre as duas é que: a categoria exponencial é representada por variáveis aleatórias cujas funções geradoras de momento logarítmico são finitas enquanto que a categoria subexponencial tem a mesma função com valor infinito.

Em [43] é realizado um estudo que tem por objetivo a obtenção de uma expressão que reflete o comportamento de uma fila do tipo $Q_{t+1} = (Q_t + A_t - C_t)^+$ quando o processo de chegadas é subexponencial. Como o modelo de tráfego apresentado em 4.5 é a soma de vários grupos de células, cada um com a distribuição subexponencial (*heavytail*) fica a pergunta: a soma de várias variáveis aleatórias subexponencialmente distribuídas é uma v.a. subexponencial ? De acordo com [52] sim. Mais especificamente : sejam os processos X_1 e X_2 auto-similares de segunda ordem com seus parâmetros de Hurst dados por $H_1 = H_2 = H$. Então $X_1 + X_2$ é auto-similar de segunda ordem com parâmetro de Hurst *H*.

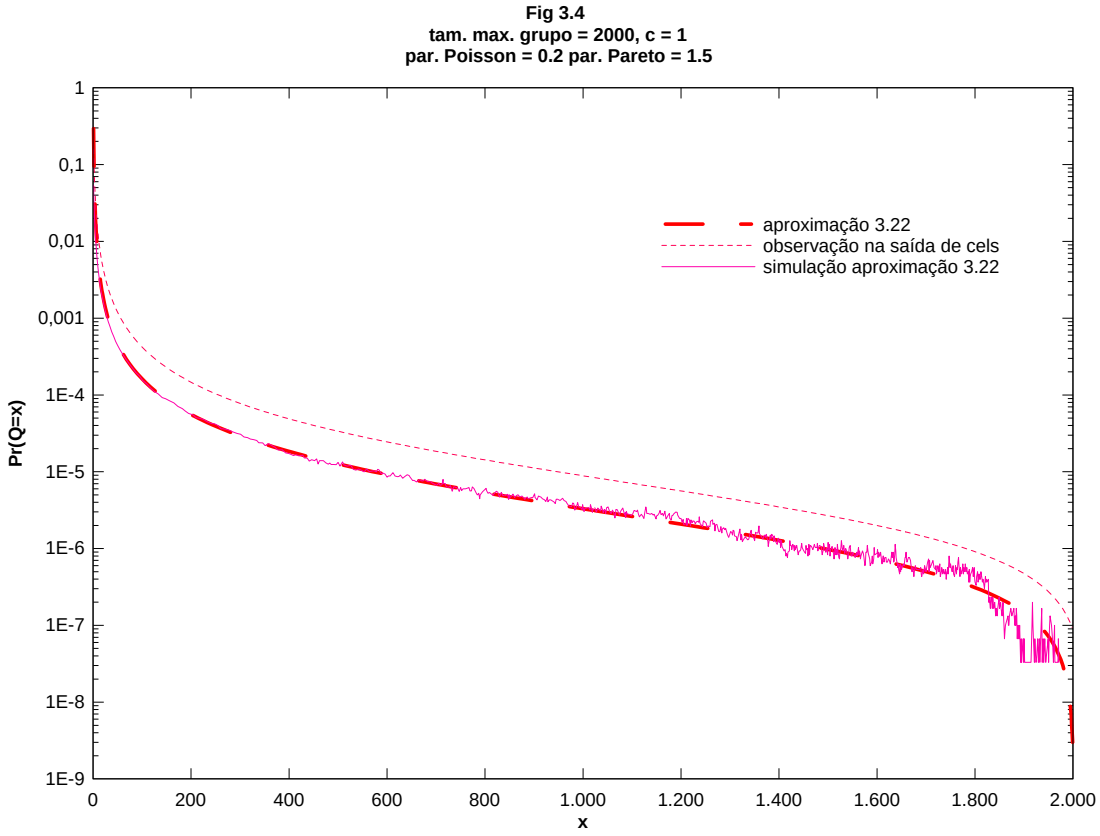
A expressão usada para representar o comportamento assintótico da fila, de acordo com [43], cosiderando-se processo de tráfego de entrada no sistema subexponencial, é dada por :

$$\Pr[Q_t > x] \approx \frac{1}{E(C_t) - E(A_t)} \int_x^\infty \Pr[A_t > u]\, du, \text{ a medida que } x \to \infty. \qquad (4.21)$$

Como temos um modelo discreto escrevemos :

$$\Pr[Q_t > x] \approx \frac{1}{E(C_t) - E(A_t)} \sum_{i=x}^{\infty} \Pr[A_t > i] \text{ a medida que } x \to \infty. \qquad (4.22)$$

A Figura 4.4 mostra a plotagem dos resultados até aqui. Para o parâmetro do processo de Poisson o valor é de 0.2, o valor de α (da distribuição de Pareto) é de 1.5. A capacidade do servidor é uma célula por segmento de tempo (*C=1*). Os gráficos representam as probabilidades (ordenadas) de encontrarmos os valores das abicissas de usuários no sistema. São representadas as curvas provenientes do esquema apresentado em 4.6.1 (observação somente nos instantes de saídas de células), proveniente da equação 4.22 (aproximação subexponencial) aplicada ao modelo de tráfego estendido e à simulação correspondente. Para estes três casos é considerado um tamanho máximo de grupo de 2000.



Pode-se verificar que o uso da equação (4.22) para prever o comportamento de uma fila quando alimentada pelo modelo de tráfego apresentado em 4.6.1 é viável. A curva, representada pela observação na saída das células, fica um pouco acima das demais pois ela leva em consideração somente os instantes em que há saída de um usuário (uma célula) do sistema enquanto que as duas outras consideram todos os segmentos de tempo, havendo saída de usuários ou não.

4.5.3 Capacidade Equivalente do Modelo apresentado em 4.5.1

A capacidade equivalente do modelo apresentado em 4.5.1 pode ser calculada a partir da fómula geral que é dada por :

Um processo $X[0,t]$ com incrementos independentes estacionários é chamado Processo de Lévy se $\alpha(\delta,t)$ não depende de t, ou seja, $\alpha(\delta,t)=\alpha(\delta)$. Um exexplo deste processo é o Processo Composto de Poisson. Se

$$X[0,t]=\sum_{n=1}^{N(t)} Y_n \qquad (4.23)$$

onde $Y_1, Y_2, \ldots$ são variáveis aleatórias independentes e identicamente distribuídas com função distribuição de probabilidade *F*, e *N(t)* é um processo de Poisson independente com taxa λ então :

$$\alpha(\delta) = \frac{1}{\delta} \int_{x} (e^{\delta x} - 1)\lambda \, dF(x) \tag{4.24}$$

Por exemplo, se $Y_1, Y_2, \ldots$ são exponencialmente distribuídos com taxa μ então teremos :

$$\alpha(\delta) = \frac{1}{\delta} \int_{0}^{\infty} (e^{\delta x} - 1)\lambda \, \mu e^{-\mu x} \, dx = \frac{\lambda}{\mu - \delta} \text{ para } \delta < \mu \tag{4.25}$$

Aplicando este resultado ao nosso modelo temos :

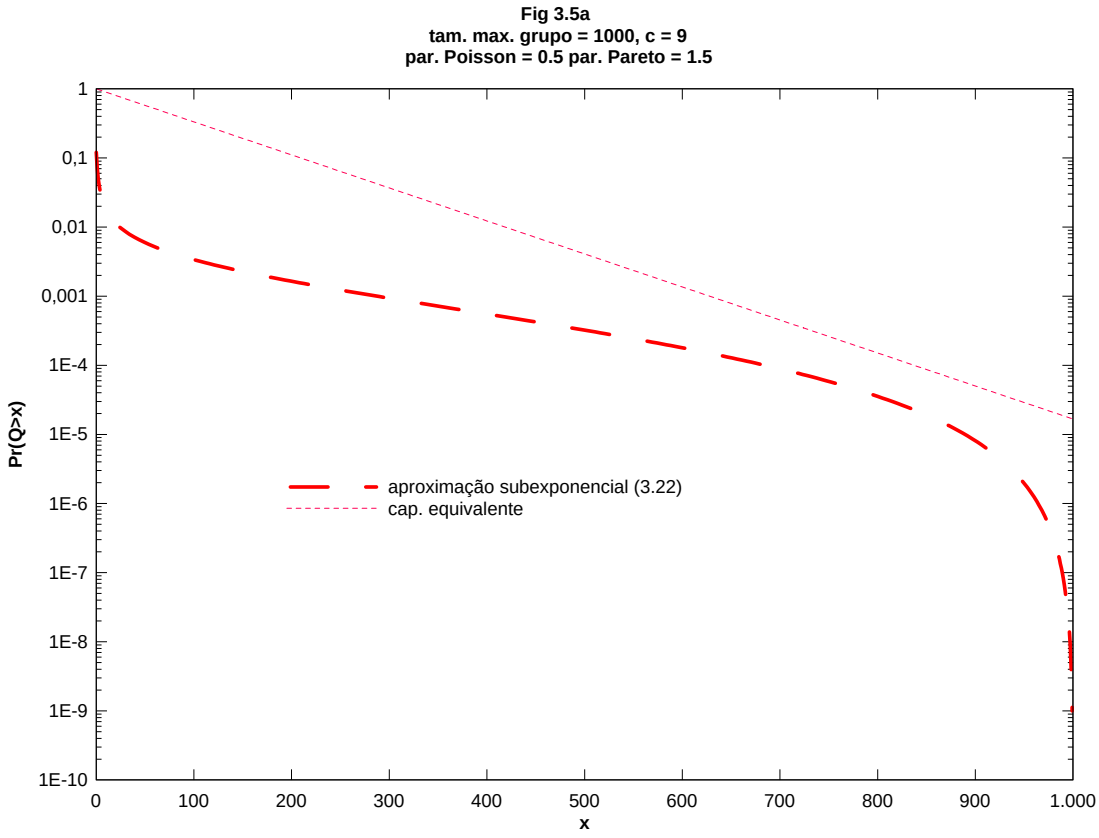
$$\alpha(\delta) = \frac{1}{\delta} \sum_{y} (e^{\delta y} - 1)\lambda \, \frac{y^{-\beta}}{k} \tag{4.26}$$

Ou seja para que haja garantia da qualidade de serviço $\Pr[Q_t > x] \leq e^{-x \cdot \delta}$ temos que ter uma servidor de capacidade dada pela expressão (4.26). É importante observar que este resultado é gerado através da truncagem do tamanho máximo de grupo de células geradas por uma fonte dentro de um segmento de tempo. Pelo fato de haver esta truncagem podemos assegurar que a capacidade equivalente do modelo existe. Em [7] são encontradas expressões para o comportamento assintótico dos modelos de tráfego auto-similares (M/G/∞ e Fractional Gaussian Noise) sem que haja truncagens deste tipo.

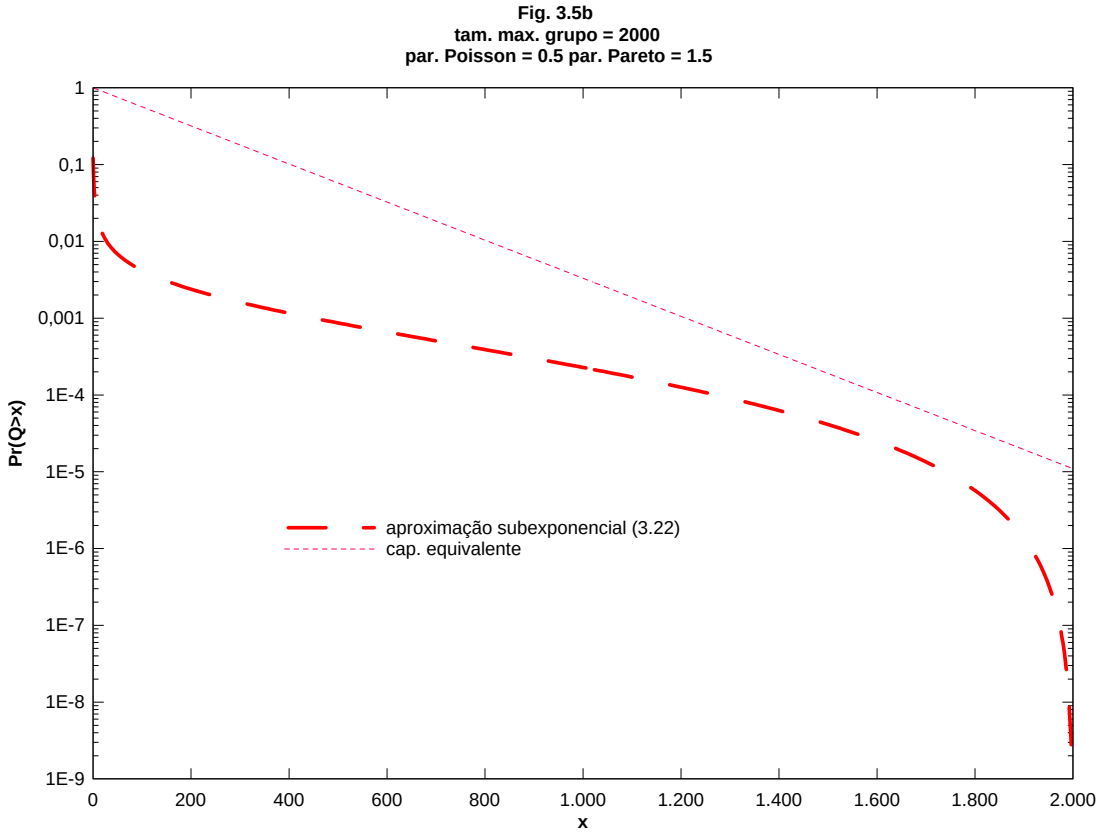
4.6 Comparações entre a aproximação da capacidade equivalente e do comportamento assintótico

De posse das aproximações utilizadas para alocação de recursos pode-se comparar os resultados fornecidos pelas fórmulas apresentadas em (4.26) e (4.22). A comparação mais simples a ser feita é a plotagem das curvas representativas dos valores $\Pr[Q_t > x]$ em função de *x* (representativo de um certo valor de conteúdo da fila). Para isso considera-se os seguintes valores de parâmetros de tráfego: parâmetro de Hurst H=0.75, parâmetro de intensidade de tráfego λ = 0.5 e tamanho máximo de grupo = 1000 e 2000. Consideremos ainda o valor da capacidade do servidor *C*=9. Os valores das capacidades equivalentes

correspondentes são 8.98 e 8.94 (correspondente aos valores de $\delta = 0.011$ e $\delta = 0.00571$). Sendo assim temos as figuras 4.5a e 4.5b representativa das curvas.



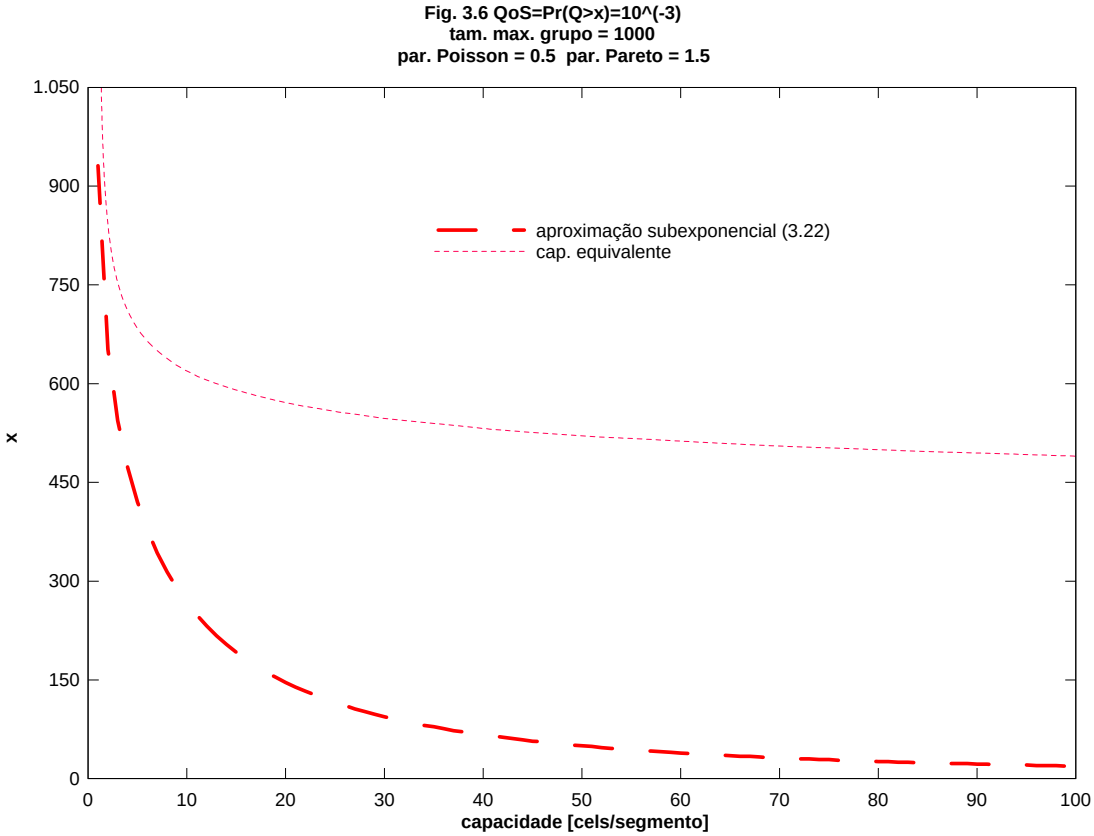
Pode-se verificar o efeito da truncagem do tamanho máximo do grupo pelo decaimento repentino das curvas representativas da aproximação 4.22. A truncagem é função do limite máximo de geração de células de uma fonte dentro de um segmento de tempo considerado. A determinação deste parâmetro pode ser vista como função do intervalo de tamanho da fila que deseja-se analisar e do grau de precisão que deseja-se ter em relação aos modelos auto-similares que não apresentam a truncagem como um passo intermediário de análise. Quanto maior o tamanho da fila a ser analisada e maior precisão para a comparação, maior deve ser o tamanho máximo de grupo devemos considerar. A figura 4.5b faz a mesma análise que a anterior com excessão do tamanho máximo de grupo que passa a ser de 2000.



A interpretação que pode-se fazer da figura 4.5b é que para um mesmo tamanho de *x*, a aproximação da capacidade equivalente prevê uma probabilidade de que o conteúdo da fila seja maior que um certo valor, maior em relação ao resultado fornecido pela aproximação em (4.22), específica para o modelo considerado.

Outra comparação a ser feita se dará pelos conjuntos formados pela reserva de espaço em fila e capacidades que satisfazem à mesma qualidade de serviço. Na figura 4.6 são apresentadas curvas que representam reservas para a QoS, representada por $\Pr[Q_t > x]$, de 10^{-3}, H (parâmetro de Hurst) = 0.75, parâmetro de Poisson = 0.5 e tamanho máximo de grupo = 1000. Pode-se observar que a aproximação representativa da capacidade equivalente prevê o uso de uma quantidade maior de recursos (espaço em fila e capacidade do servidor) que a da aproximação dada por (4.22) para uma mesma qualidade de serviço.

Podemos observar também que as aproximações tendem a ter o mesmo comportamento à medida que o tamanho do fila aumenta. Neste caso, como observado em [11], as aproximações tendem a alocar um valor de largura de faixa correspondente à taxa média gerada pelo modelo de tráfego.



Outro resultado que pode-se obter é a comparação entre a intensidade de tráfego suportada para mesmo valores de capacidade de servidor (células por segmento de tempo), qualidade de serviço, tamanho máximo de grupo e tamanho de fila. Esta comparação permite a avaliação da quantidade excessiva de recursos reservados pela aproximação da capacidade equivalente. Considera-se que para valores de qualidade de serviço de 10^{-2},10^{-3},10^{-4} e 10^{-5}, tamanhos máximos de grupo de 1000 e 2000 e parâmetro de Poisson de 0.5 (medida de intensidade de tráfego). Com isso pode-se levantar curvas como na figura 4.5, comparando as duas aproximações conforme as figuras 4.7 abaixo. O conjunto de curvas com tamanhos máximos de grupo diferentes irá fornecer informação sobre a quantidade excessiva de recursos reservados em relação a este parâmetro.

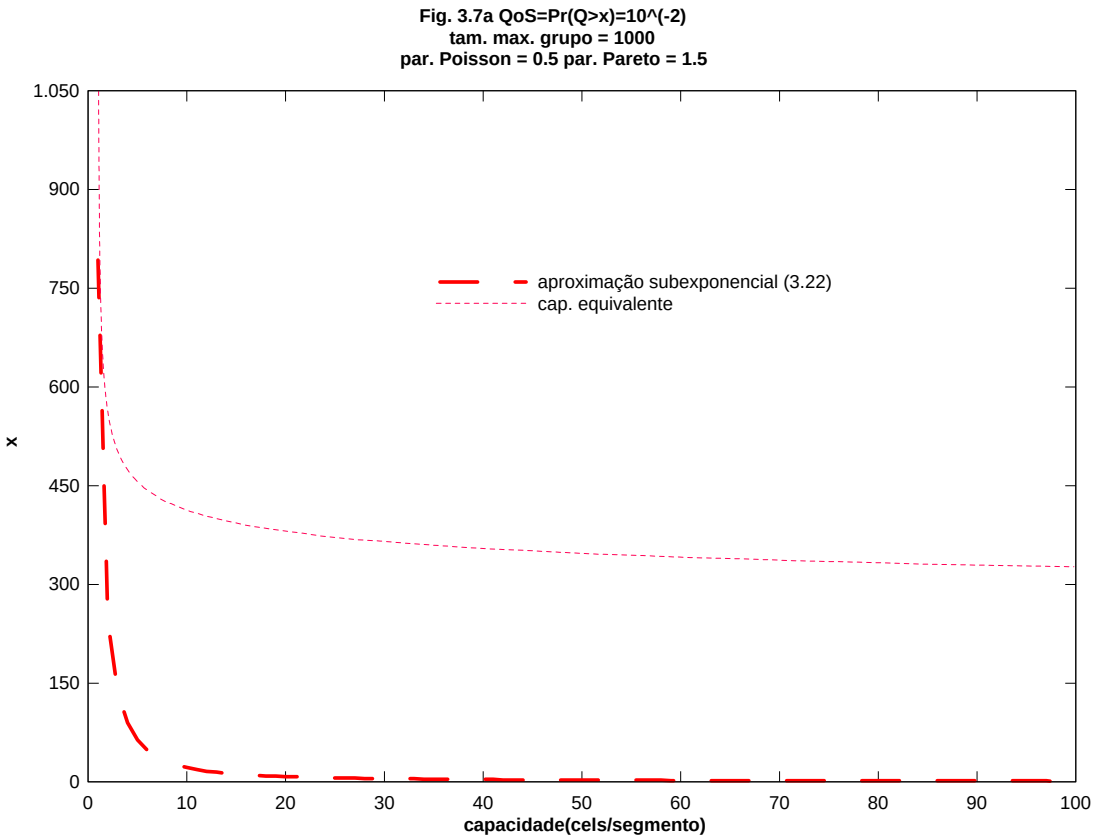


Fig. 3.7a QoS=Pr(Q>x)=10^(-2)
tam. max. grupo = 1000
par. Poisson = 0.5 par. Pareto = 1.5
1.050
900
750
600
450
300
150
0
x
0
10
20
30
40
50
60
70
80
90
100
capacidade(cels/segmento)
aproximação subexponencial (3.22)
cap. equivalente

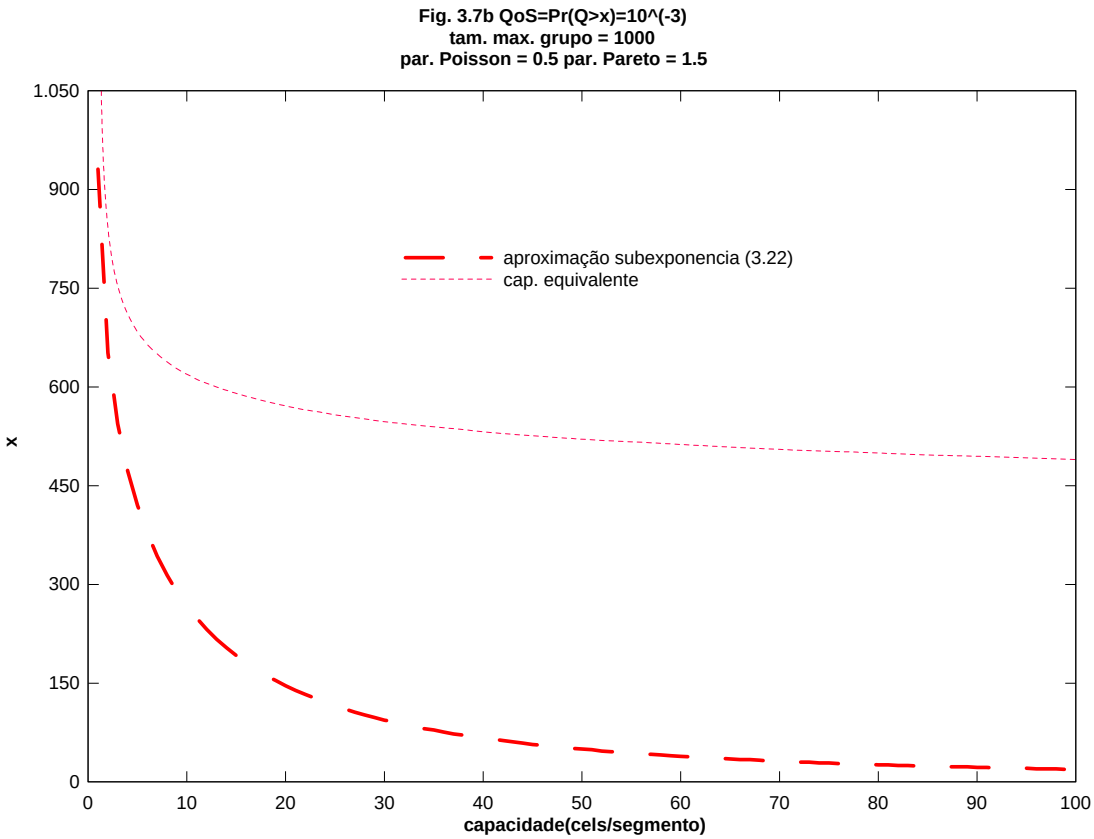


Fig. 3.7b QoS=Pr(Q>x)=10^(-3)
tam. max. grupo = 1000
par. Poisson = 0.5 par. Pareto = 1.5
1.050
900
750
600
450
300
150
0
x
0
10
20
30
40
50
60
70
80
90
100
capacidade(cels/segmento)
aproximação subexponencia (3.22)
cap. equivalente

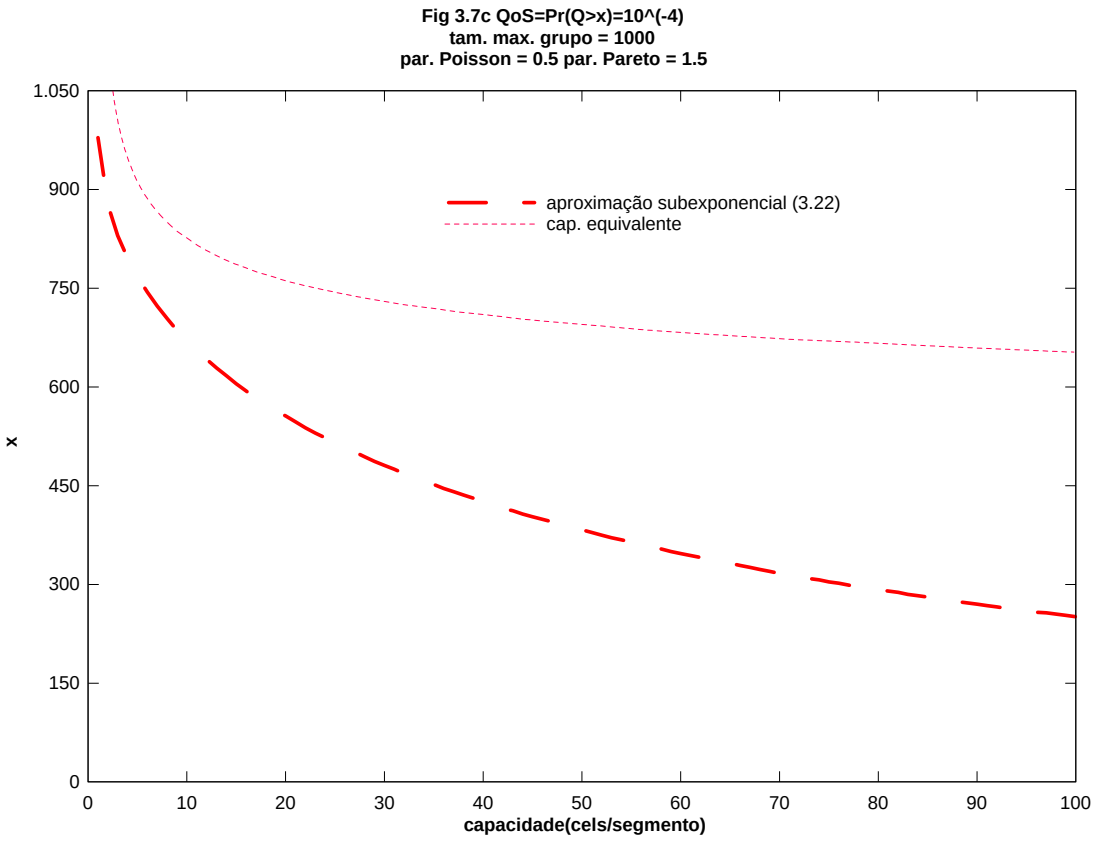
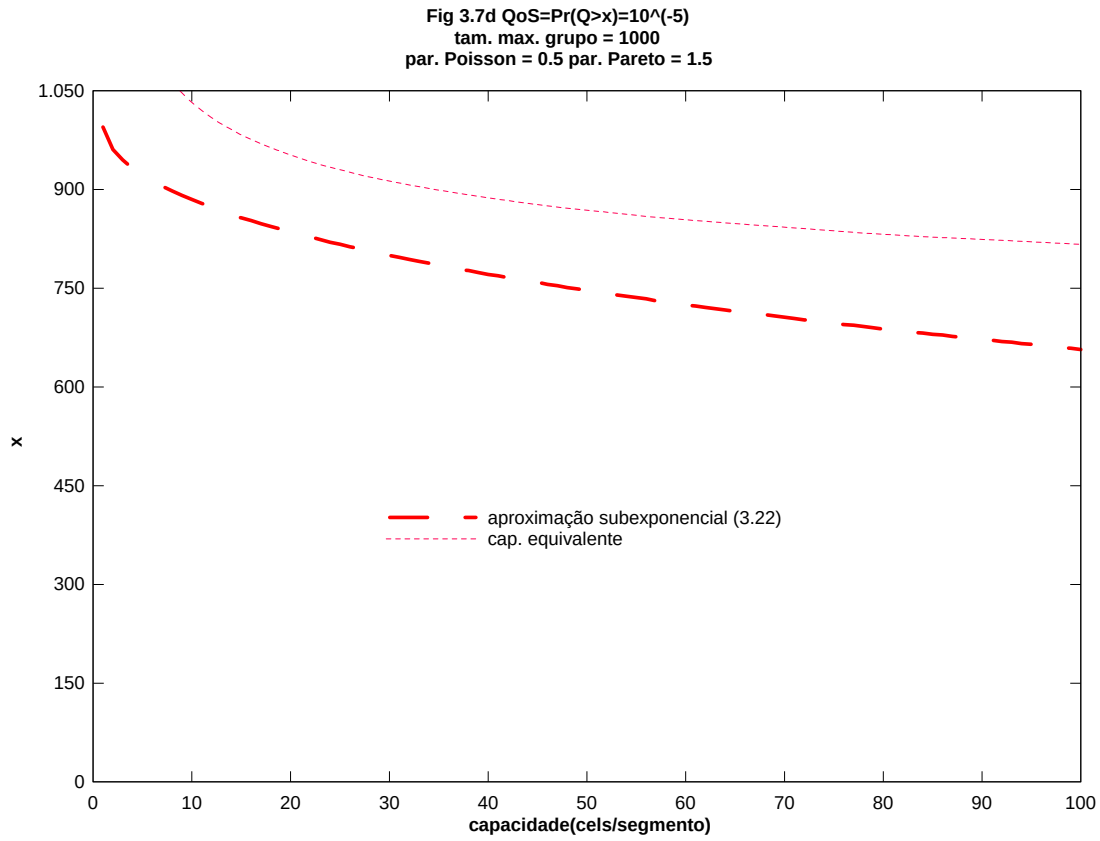


Fig 3.7c QoS=Pr(Q>x)=10^(-4)
tam. max. grupo = 1000
par. Poisson = 0.5 par. Pareto = 1.5
aproximação subexponencial (3.22)
cap. equivalente
x
capacidade(cels/segmento)
1.050
900
750
600
450
300
150
0
0
10
20
30
40
50
60
70
80
90
100



Pode-se dar a seguinte interpretação para o conjunto de figuras 4.7: os conjuntos de alocações de recursos para as aproximações de capacidade equivalente e da expressão (4.22) tem sua proximidade em função da qualidade de serviço planejada, sendo que à medida que a probabilidade relacionada diminui a proximidade aumenta. Não se deve deixar de considerar que esta aproximação é relacionada à truncagem feita. Pode-se observar que a distância entre os conjuntos de alocações aumenta à medida que o tamanho máximo do grupo aumenta. Pode-se verificar isso comparando-se o conjunto de figuras 4.7 e 4.8.

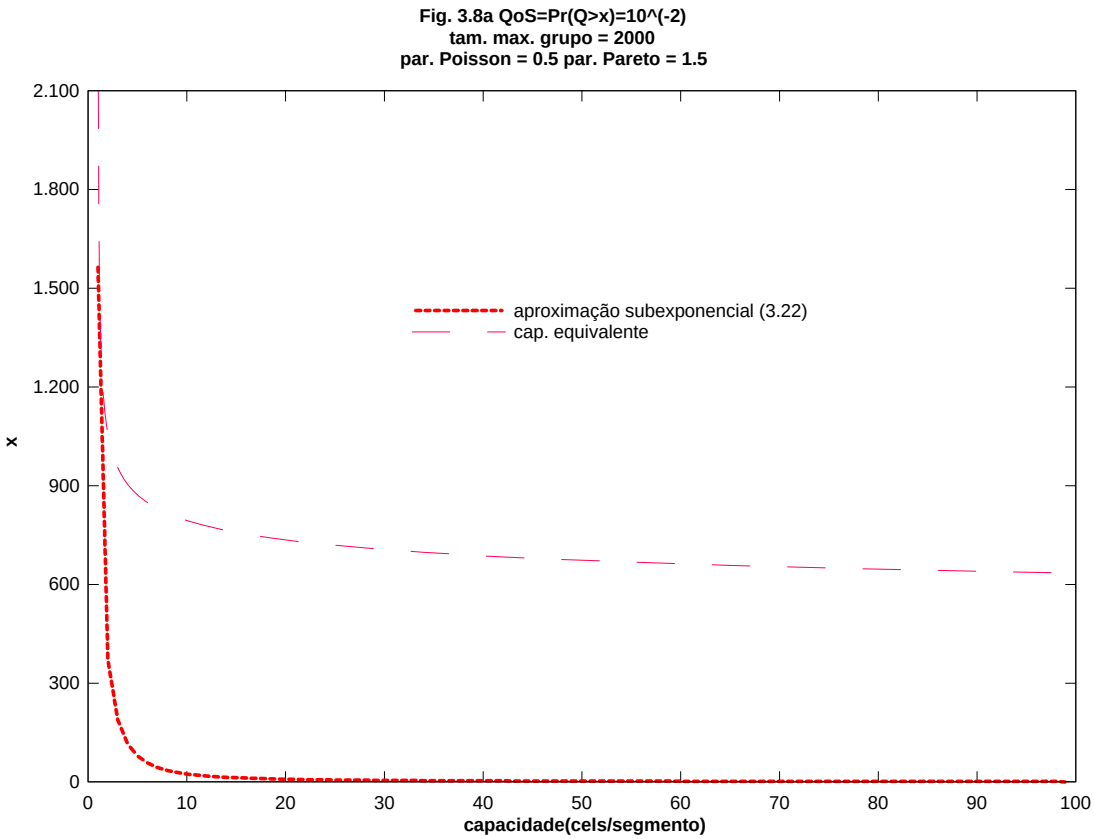


Fig. 3.8a QoS=Pr(Q>x)=10^(-2)
tam. max. grupo = 2000
par. Poisson = 0.5 par. Pareto = 1.5
2.100
1.800
1.500
1.200
900
600
300
0
x
0
10
20
30
40
50
60
70
80
90
100
capacidade(cels/segmento)
aproximação subexponencial (3.22)
cap. equivalente

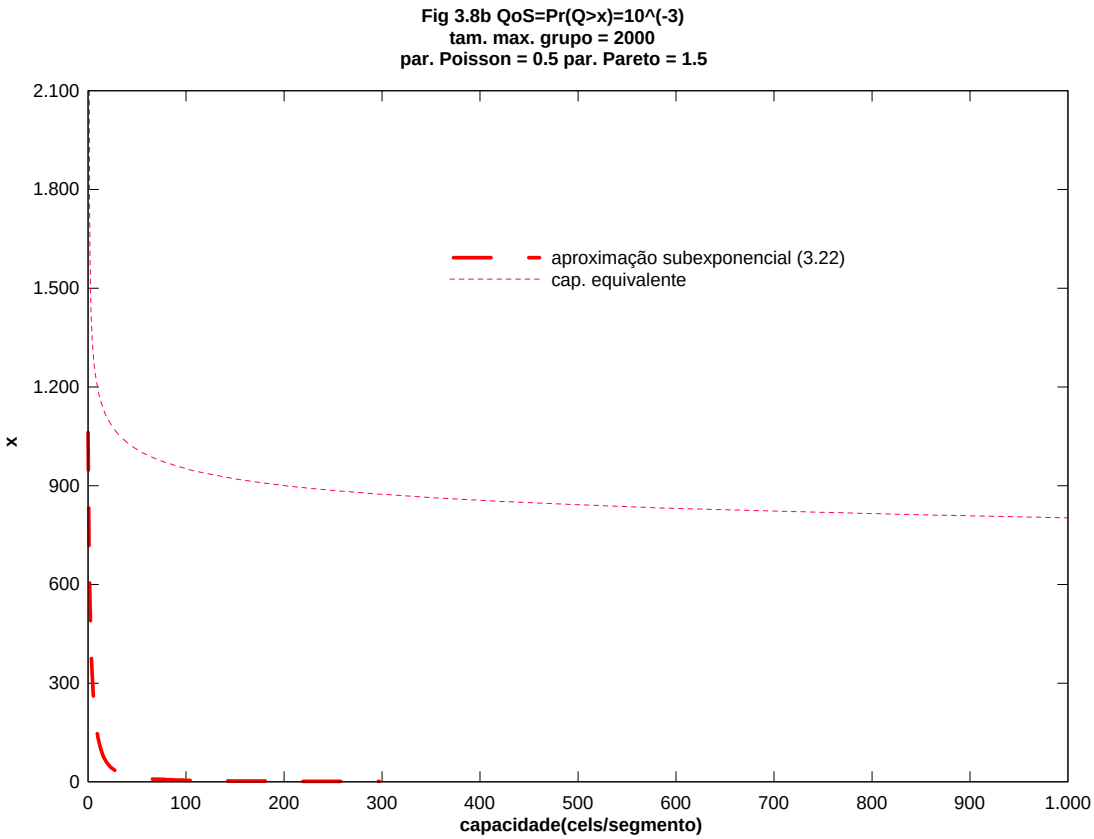


Fig 3.8b QoS=Pr(Q>x)=10^(-3)
tam. max. grupo = 2000
par. Poisson = 0.5 par. Pareto = 1.5
2.100
1.800
1.500
1.200
900
600
300
0
x
0
100
200
300
400
500
600
700
800
900
1.000
capacidade(cels/segmento)
aproximação subexponencial (3.22)
cap. equivalente

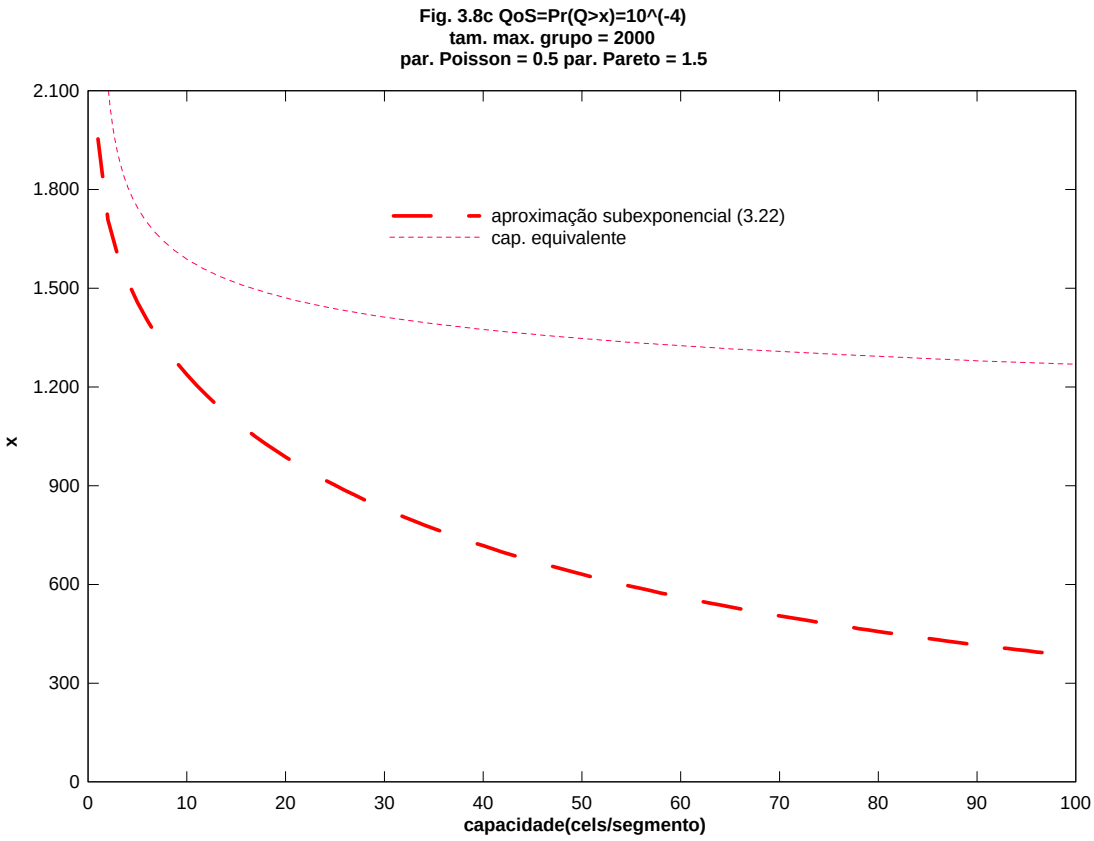
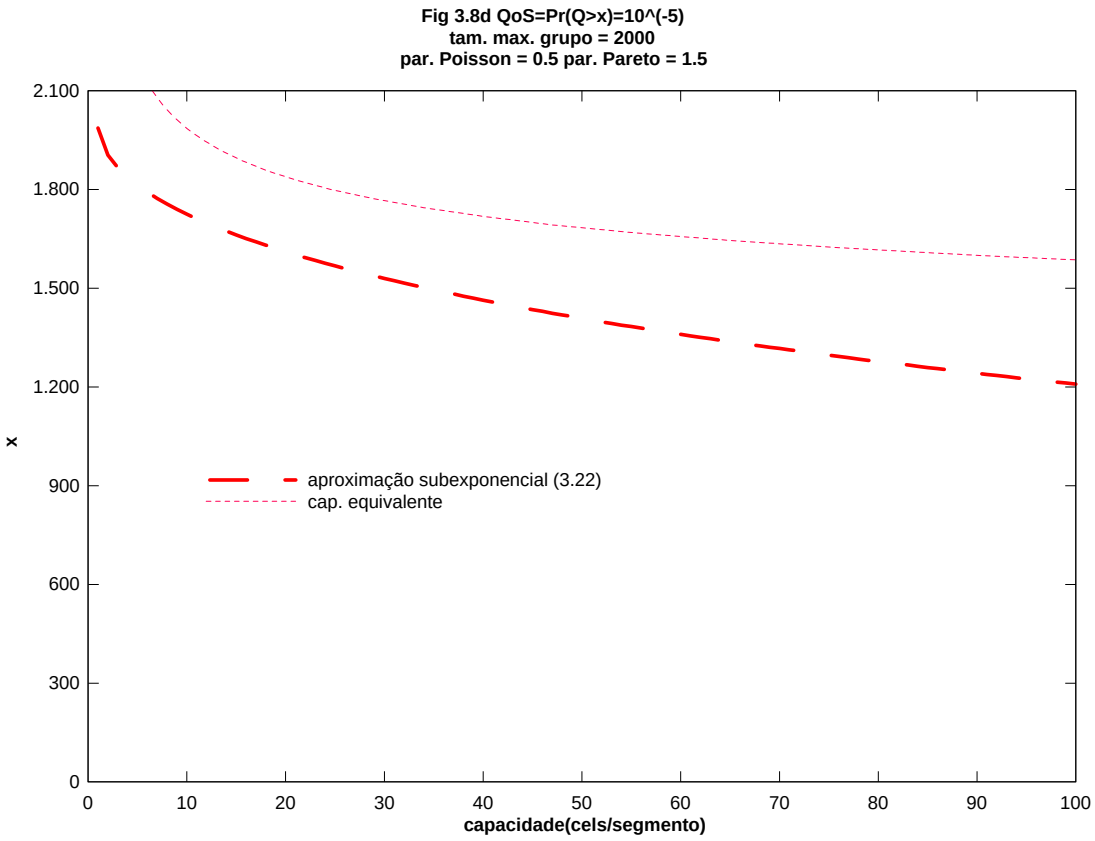


Fig. 3.8c QoS=Pr(Q>x)=10^(-4)
tam. max. grupo = 2000
par. Poisson = 0.5 par. Pareto = 1.5
2.100
1.800
1.500
1.200
900
600
300
0
x
0
10
20
30
40
50
60
70
80
90
100
capacidade(cels/segmento)
aproximação subexponencial (3.22)
cap. equivalente



Através da aproximação 4.22 obtém-se vários valores de capacidade e de tamanho de fila que satisfazem à qualidade de serviço. Agora deseja-se levantar os valores de intensidade que a aproximação da capacidade equivalente permite, mantendo-se os outros parâmetros constantes. Para efeitos de comparação considera-se a razão entre os valores de intensidade usados na aproximação 4.22 e os calculados através da aproximação da capacidade equivalente. Os valores próximos a 1 (um) indicam a igualdade entre os valores e à medida que este valor aumenta a reserva passa a ser excessiva. Sendo assim, a reserva exagerada da aproximação da capacidade equivalente fica caracterizada. Isto pode ser verificado nas figuras 4.9a e 4.9b nas quais as razões são plotadas para os valores de qualidade de serviço de 10^{-3},10^{-4},$10^{-4,5}$e 10^{-5} com tamanho máximo de grupo 1000 e 2000 respectivamente, intensidade de tráfego $\lambda = 0.5$ para a aproximação (4.22) e H=0,75. Pode ser verificado que o fato da utilização de dois tamanhos máximos de grupo distintos pouco influi na precisão da aproximação da capacidade equivalente.

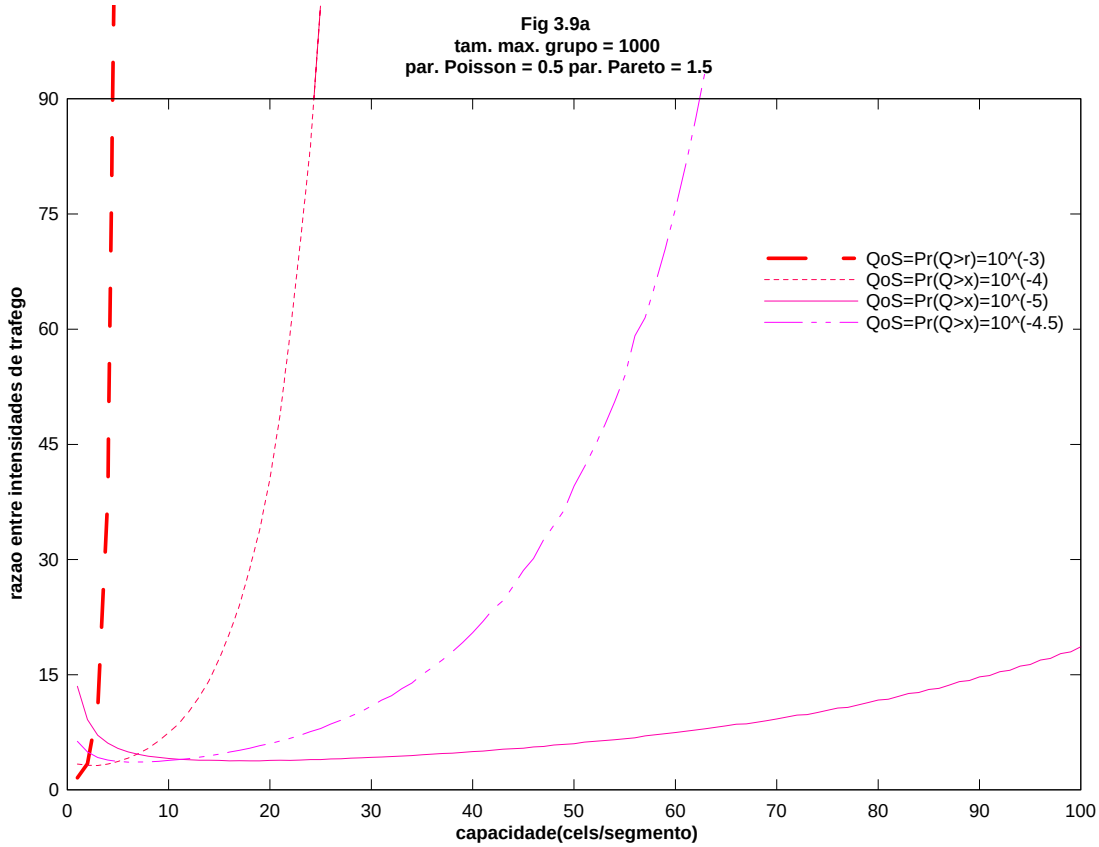
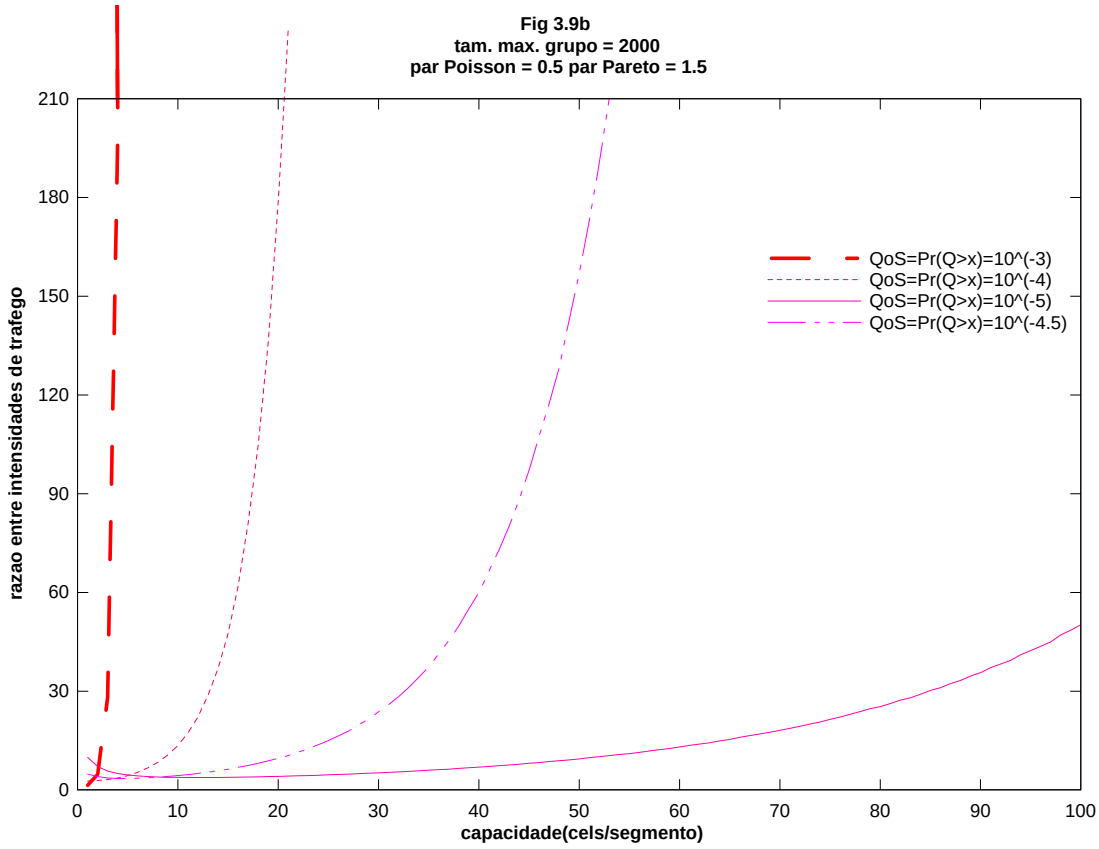


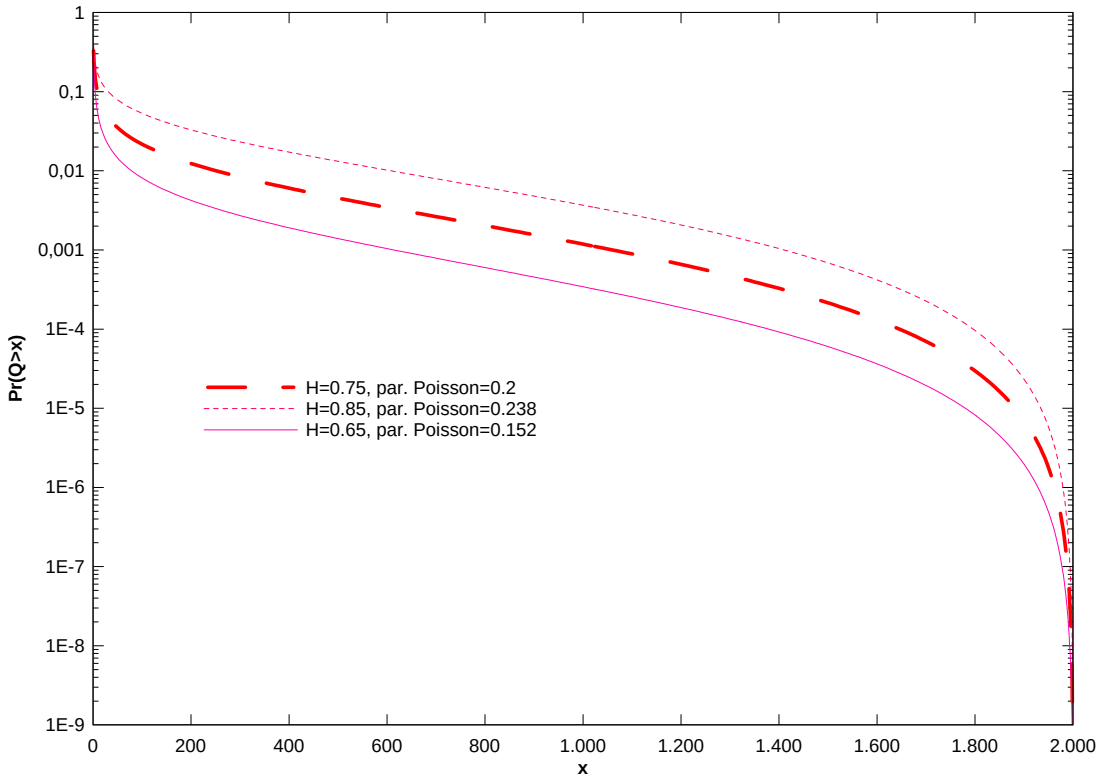
Fig 3.9a
tam. max. grupo = 1000
par. Poisson = 0.5 par. Pareto = 1.5
razao entre intensidades de trafego
capacidade(cels/segmento)
QoS=Pr(Q>r)=10^(-3)
QoS=Pr(Q>x)=10^(-4)
QoS=Pr(Q>x)=10^(-5)
QoS=Pr(Q>x)=10^(-4.5)
0
15
30
45
60
75
90
0
10
20
30
40
50
60
70
80
90
100



4.6.1 Comportamento em relação ao parâmetro de Hurst

Nesta parte do trabalho analisa-se o comportamento deste sistema em relação à variação do parâmetro de Hurst. Este parâmetro,conforme [24], mede o grau de rajada (*burstiness*) do modelo de tráfego. Para fazer esta análise gera-se três resultados representativos de filas alimentadas pelo modelo de tráfego apresentado em 4.5. Os três tráfegos gerados tem a mesma média de geração de células por segmento de tempo mas diferenciam-se no parâmetro de Hurst e na intensidade de tráfego (parâmetro de Poisson). Na figura 4.10, similar às figuras 4.5, tem-se resultados com intensidade de tráfego de 0.152 e H=0.85, com intensidade igual a 0.238 e H=0.65 e com intensidade de 0.2 e H=0.75. Quanto maior o valor de H maior o grau de rajada, sendo que estes resultados estão de acordo com o estudo feito em [21] no qual a translação vertical, de baixo para cima das curvas é proporcional ao tamanho médio da rajada, conseqüência do valor de H=0.85. Para maiores valores deste parâmetro maior o comportamento *heavy-tail* do tráfego gerado e do resultado apresentado.

Fig. 3.10
tam. max. grupo = 2000



Capítulo 5

Tarifação em Redes Multiserviço

5.1 Introdução

A principal intenção dos provedores de serviços em redes e operadoras de telecomunicações é a maximização da diferença entre os ganhos obtidos e os custos decorrentes da manutenção e operação de rede. Os ganhos obtidos podem ser pensados como a quantidade dos serviços que podem ser oferecidos e que são vendidos por um determinado preço. Os custos podem ser pensados como os recursos necessários para oferecer tais serviços. A maximização desta diferença está diretamente relacionada a fatores tecnológicos e de mercado (competição, preço que os usuários estão dispostos a pagar, etc...).

O método pelo qual é determinado o esquema de tarifação não é importante somente pelo retorno financeiro proporcionado por tal prática. A tarifação também deve ser utilizada para fins de controle de acesso e aproveitamento dos recursos da rede. Atualmente há uma grande demanda para serviços que consomem grandes quantidades de recursos de rede como aplicações multimídia. Considere a situação de estabelecimento de conexão em uma rede ATM. Neste momento é firmado um contrato entre rede e usuário. O usuário declara características dos tráfego que deverá ser transportado e a rede se propõem a transportar tal tráfego com a qualidade de serviço negociada. Logo a rede deverá cobrar um preço, ao usuário, que reflita a quantidade de recursos que a conexão originada pelo usuário irá consumir.

Podemos dizer que o desempenho de uma rede, pelo menos pela perspectiva das aplicações (usuários), não é completamente determinado pelas características técnicas da rede. O desempenho da rede também pode ser determinado pelo tráfego oferecido à rede. O tráfego agregado oferecido pela rede é o resultado de decisões individuais dos usuários sobre quando e como usar a rede. Estas decisões são afetadas diretamente pelos incentivos que os usuários encontram quando usam a rede. Logo, para a análise de desempenho da rede, em adição às especificações técnicas, a questão dos incentivos dados aos usuários

deve ser considerada. Estes incentivos podem ser dados de várias formas : incentivos de desempenho, incentivos monetários, incentivos administrativos, etc.... Existem várias questões relacionadas à tarifação em redes de computadores. Por exemplo, a tarifação será afetada pela estrutura de mercado do serviço prestado pela rede, pelo ambiente que regula a utilização da rede, e os custos das tecnologias empregadas na rede. A estrutura de tarifação deve levar em consideração também a demanda pelos serviços prestados pela rede, em termos de elasticidade de preço (como varia o preço de acesso a uma determinada aplicação) e de variabilidade de preços (variação do preço em termos da qualidade de serviço contratada), além da natureza dos serviços que a rede presta.

5.2. Motivação para tarifação em redes de computadores

A tarifação em redes de computadores torna-se uma questão de maior relevância quando comparamos características da atual Internet e discutimos como estas características tendem a mudar em um futuro próximo.

Hoje podemos identificar quatro características básicas da Internet relevantes à nossa discussão. Primeiramente a largura de faixa total é limitada inibindo a utilização de aplicativos que demandam uma quantidade maior deste recurso (como aplicações de HDTV, telemedicina, etc ...). Uma segunda característica que podemos apontar é o seu acesso restrito. Há alguns anos atrás somente entidades educacionais e instituições de pesquisa tinham acesso à Internet. Esta restrição ao acesso, além do controle do tamanho da população de usuários, ajudam a conservar a natureza coesiva da comunidade de usuários. A terceira característica é que a estrutura atual oferece apenas uma qualidade de serviço, sendo todos os datagramas IP servidos de maneira *best-effort* com disciplina FIFO. Esta característica limita a natureza das aplicações que podem ser suportadas adequadamente. Por último podemos dizer que não há controle sobre a utilização dos usuários, ou seja, os usuários não são taxados com base na quantidade de pacotes que eles transmitem, ou seja, o usuário individual não recebe uma taxa individual pelo uso da rede.

Baseado nestas características podemos perceber que a Internet, no futuro, será diferente da atual a partir do momento que haja demanda para novas aplicações e serviços a serem oferecidos por uma rede de serviços integrados. Considerando-se que este fato irá se concretizar, podemos enumerar algumas mudanças necessárias. Primeiramente, podemos apontar o aumento da largura de faixa disponível, necessidade indispensável para o uso de aplicações sensíveis a este recurso. Segundo, este aumento de largura de faixa disponível,

combinado com aumento do uso de PCs e terminais RDSI residenciais fazem com que a rede seja cada vez mais utilizada. Talvez não demore muito que a rede imponha restrições artificiais ao acesso (como acontece atualmente em outros serviços públicos, como por exemplo o telefônico), objetivando um uso comunitário mais cooperativo. Mesmo assim é esperado que se faça uma seleção baseada no seu próprio interesse, sem se preocupar com a funcionalidade total da rede.

Por último, os mecanismos de controle de tráfego que serão utilizados no futuro (como algoritmos de filas em comutadores ATM e política de descarte de células quando ocorre congestionamento) devem ser mais sofisticados que os atuais mecanismos utilizados na Internet atual (família de protocolos TCP/IP), que suportam somente uma classe de serviço.

Um ponto de consenso é que várias aplicações tem diferentes tipos de exigências de serviço. Por exemplo, aplicações como *e-mail* podem suportar um retardo de transferência tolerável sem que os usuários sintam uma queda de desempenho considerável. Por outro lado aplicações como voz por pacotes tem seu desempenho degradado com retardo de entrega de alguns milisegundos. A quantidade de aplicações e a diversidade de exigências de qualidade de serviço irão crescer rapidamente em um futuro próximo.

Outro ponto de consenso é que as redes devem, de uma maneira eficiente, fornecer esta variedade de exigências de desempenho para que um usuário possa escolher uma qualidade de serviço que se adapte à aplicação que está sendo utilizada por ele. A rede poderia, em um período de contenção de recursos, focalizar seus recursos em aplicações mais sensíveis à qualidade de serviço (voz, vídeo por exemplo) e ao desempenho e evitar o disperdício de recursos em aplicações nem tão sensíveis (*e-mail* por exemplo). Tipicamente, nem todos os serviços tem seu desempenho melhorado considerando-se esta estrutura de múltiplas classes de serviços. Por exemplo, tráfego em uma classe de serviço mais baixa pode experimentar, em alguns instantes, um desempenho pior que receberia em redes de apenas uma classe de serviço. O propósito de múltiplas classes de serviço seria degradar o desempenho para as aplicações menos sensíveis para melhorar o desempenho para aquelas que são mais sensíveis.

Um dos pontos principais é relacionado ao pagamento pelos serviços prestados pela rede. A possibilidade de múltiplas classes de serviço (vários tipos de QoS que podem ser fornecidos pela rede) está diretamente relacionada ao desempenho da rede em termos de

reserva de recursos para as diversas aplicações. É natural que os usuários desejem um bom desempenho da rede. Uma vez que eles agora são capazes de tomar ações que influenciam o desempenho que eles irão receber, existe, imediatamente, um incentivo a pedir o uso de uma classe de serviço que maximize a sua satisfação com um menor custo possível para o usuário. Em uma comunidade pequena e cooperativa de usuários um comportamento normalizado pode ser adotado para que a classe de serviço seja contratada apropriadamente. Porém em uma rede pública, ou seja uma comunidade grande de usuários quase anônimos, a classe de serviço contratada pode não ser contratada apropriadamente. Logo com a tarifação apropriada das classes de serviço as operadoras podem oferecer incentivos monetários para que os usuários reduzam a classse de serviço pedida. Logo a tarifação pode ser um veículo que permita aos usuários fazerem suas escolha de uma maneira coerente. Daí vem a importância de um projeto de política de tarifação que faça com que a tecnologia disponível seja aproveitada da melhor maneira possível tanto para a rede (interessada em fornecer um bom serviço a seus usuários atingindo alto grau de utilização de seus recursos) quanto para o usuário (interessado em pagar exatamente o necessário para a qualidade de serviço que foi contratada).

5.3 Algumas propriedades desejáveis da tarifação aplicada

A tarifação em redes pode ser pensada como um mecanismo de controle e *feedback*, além do objetivo financeiro. Isto acontece através da reação de cada usuário quando este se depara com o preço a ser pago pelo serviço prestado à rede e tenta minimizar este preço através de meios próprios (como por exemplo minimizar a tráfego gerado durante uma conexão). Vencido este obstáculo o próximo problema é atingir um ponto de equilíbrio de operação da rede (altos níveis de utilização da rede sem que seja grande a probabilidade de congestionamento e consequentemente degradação dos serviços prestados aos usuários). Ou seja, a tarifação deve ser vista como um meio de fazer com que os usuários minimizem os recursos exigidos da rede (para que não haja desperdício) e que a rede utilize todos os recursos de modo a produzir a satisfação para toda população de usuários.

Uma estrutura de tarifação deve ser também honesta (*fairness*). Isto significa que o usuário deve pagar um preço proporcional àquele dos recursos que ele irá utilizar. Isto pode ser utilizado como meio de comparação entre esquemas de tarifação, ou seja, a comparação é feita baseada na utilização dos recursos e não nos valores absolutos cobrados. Se o preço reflete os custos com precisão então os usuários podem comparar os benefícios proporcionados pelas suas ações e tomar decisões de posse de informações coerentes.

Diretamente relacionada ao parágrafo anterior está a questão da complexidade de estrutura de tarifação. A estrutura de tarifação deve ser baseada em medidas realísticas sendo que os usuários tomem suas decisões de alterar suas características de tráfego certos dos impactos que estas decisões irão ter no preço cobrado pela rede.

5.4 A dificuldade de tarifar os serviços de redes

Os esquemas de tarifação são necessários para que a rede recupere seus custos, relacionados à manutenção e operação em um ambiente competitivo, dos diversos usuários que utilizam os seus serviços. Os esquemas de tarifação também podem ser utilizados para a reserva eficiente de recursos de rede, administrando eficientemente o grau de congestionamento da rede. As taxas cobradas aos usuários serão determinadas por estratégias competitivas, entretanto, podemos generalizar que as taxas cobradas aos usuários refletirão a quantidade de recursos de rede utilizados por estes. A economia teórica sugere que a tarifação Baseada no Uso será empregada pelas redes no caso de competição perfeita. Neste cenário a rede define o preço a ser cobrado de acordo com parâmetros obtidos dos usuários (seja verificando-se o comportamento de pior caso ou verificando-se a capacidade equivalente). Uma vez que o preço está definido, e conhecidas suas necessidades de tráfego, o usuário escolhe o contrato mais barato, dentre os vários contratos com vários operadores, que esteja de acordo com sua expectativa sobre a qualidade desejada.

(i) A qualidade de serviço afeta a tarifação

A RDSI-FL, por um lado, simplifica o fornecimento de múltiplos serviços através de uma única interface que trata de todas as necessidades dos usuários. Mas por outro lado, faz com que a tarifação baseada no uso seja difícil de ser realizada a partir do momento em que os recursos são compartilhados entre os vários usuários, dificultando a decisão de como dividir o custo pela utilização da rede entre os vários serviços e usuários da rede. Os serviços agora passam a ser classificados também pelo QoS relacionado a eles, com isso a tarifação irá sofrer influência do QoS adotado uma vez que para garantir esta qualidade a rede deverá reservar uma certa quantidade de recursos.

(ii) Os usuários desejam flexibilidade

Um grau de complexidade adicional é originado do fato que os padrões propõem o uso de técnicas de policiamento e suavização (*shaping*) do tráfego que entra na rede gerado pelo usuário para proteger a rede do uso incorreto. Claramente que quanto mais o usuário aceita ser policiado mais o seu tráfego torna-se previsível sendo mais fácil para a rede a administração deste tráfego. Quanto mais os usuários são policiados menos eles pagarão e a tarifação irá se voltar à quantidade de policiamento aplicada. A questão torna-se mais interessante quando o usuário tem poder sobre as mudanças de estatísticas do seu tráfego usando de seus próprios mecanismos (como filtros por exemplo) para atingir um custo menor com o compromisso próprio de que a QoS permanecerá adequada.

(iii) O preço da tarifação

Até agora a tarifação poderia ser implementável de uma maneira satisfatória. Entendemos por isso que a informação exigida pelos mecanismos de tarifação poderia ser facilmente obtida e manipulada. Um pré-requisito para um mecanismo de tarifação realístico é o seu baixo custo de implementação. Os equipamentos mais modernos permitem medidas de tráfego sofisticadas feitas em *hardware* que fornecem informações que podem ser utilizadas em fins de tarifação. Mas o que devemos medir para termos um mecanismo de tarifação eficiente ? Neste caso podemos pensar na medida da largura de faixa estatística, uma medida que sumariza a quantidade de recursos utilizada pela fonte, as características estatísticas do tráfego, parâmetros de QoS e a capacidade de multiplexação do tráfego.

5.5 Alguns Mecanismos de Tarifação/Negociação propostos

No início da negociação, para o estabelecimento de uma conexão, ambos os lados sabem seus objetivos. O usuário tem informação sobre o tráfego a ser gerado e a possibilidade de avaliar o serviço, contudo, não tem informação sobre a capacidade disponível da rede (medida em espaço em filas e taxas de transmissão) e da demanda total de mercado para a capacidade disponível da rede. Por outro lado, a rede tem informações sobre a capacidade disponível mas não tem informações sobre a real expectativa feita pelo usuário e do QoS desejado, antes do estabelecimento da conexão. Dentro deste contexto o mecanismo de negociação pode ser resumido como se segue. Pelo estabelecimento de preços a rede passa ao usuário a capacidade disponível bem como a demanda de mercado. Em um mercado livre, deve haver concordância entre rede e o usuário sobre os preços

cobrados, quantidades de recursos utilizados em cada conexão e QoS. Os mecanismos de tarifação apresentados a seguir refletem a dinâmica entre rede e usuário, tentando-se atingir um ambiente de mercado eficiente para ambos os lados.

5.5.1 Tarifação Baseada em Prioridade

O trabalho realizado em [53] propõem uma política de tarifação baseada em prioridade em redes de computadores utilizando o conceito da Implementação de Nash, originária da teoria econômica. Um Equilíbrio de Nash é um ponto de equilíbrio em um jogo onde a estratégia utilizada por cada jogador é a melhor estratégia dada a escolha de todos os outros jogadores. Nenhum jogador tem incentivo de desviar do ponto de equilíbrio. Apesar disso o ponto de equilíbrio não necessariamente gera o melhor desempenho global de todos os jogadores. Uma Implementação de Nash de uma política social ótima existe se o ponto de operação social ótimo é também o único ponto de equilíbrio de Nash.

Quatro tipos de serviços simples (e-mail, FTP, Telnet e voz) são considerados neste trabalho, sendo que as funções benefício (funções que relacionam o QoS com os custos pagos por este QoS) são computadas utilizando-se o retardo total, probabilidade de perda e vazão média. As prioridades são escolhidas por cada aplicação baseadas nos preços de transmissão de pacotes de alta e de baixa prioridade. O objetivo do usuário é atingir um grau maior de QoS com menores gastos pagos à rede. O objetivo da rede é maximizar a função benefício, estabelecendo preços de tal forma que os usuários tenham um comportamento individual e social ótimo através do estabelecimento da política de prioridades de pacotes que devem ser transmitidos. O problema é o estabelecimento de preços e das prioridades a serem utilizadas. A decisão é atingida quando usuários e rede atingem um Equilíbrio de Nash deste "jogo". Através de simulações, com várias configurações de rede, é comprovada a existência de um intervalo de preços para cada nível de prioridade que irá resultar em uma Implementação de Nash.

Mais precisamente, para algumas configurações de rede, são feitas simulações comparando duas formas distintas de política da tarifação : uma "*flat pricing*", onde um preço uniforme é cobrado independentemente da aplicação do usuário (insensível à classe de serviço) e outra "*priority pricing*" onde um preço maior é cobrado à medida que a prioridade pedida pelo usuário aumenta (sensível à classe de serviço). Mantendo a renda recebida pela rede e medindo apropriadamente a satisfação do usuário em função do custo

oferecido ao usuário e da qualidade de serviço recebida por ele, é deduzido que : (1) os usuários de todos os tipos de aplicação são mais satisfeitos com o esquema de "*priority pricing*" que é sensível às várias classes de serviço. (2) estes usuários, quando maximizam sua satisfação, escolhem prioridades que maximizam a eficiência geral da rede (em termos de oferecimento de qualidade de serviço para os usuários) e difundem o benefício de múltiplas classes de serviço entre os usuários de todos os tipos de aplicação. Para uma melhor compreensão sobre a definição de nível de satisfação do usuário, de custo oferecido ao usuário e de outros consultar [53] .

Em [61] não é apresentado especificamente um mecanismo de tarifação mas um controle de tráfego baseado em prioridades que pode ser utilizado como tal. Parte-se do princípio da utilização do campo de prioridades, atualmente não utilizado, encontrado na cabeçalho dos datagramas IP. A escolha destas prioridades seria do usuário e dos administradores de rede e quotas teriam que ser negociadas entre as diversas redes. No caso de um provedor de serviço Internet este estabeleceria as várias quotas a serem distribuídas entre os seus clientes. Este quota seria uma soma ponderada dos valores de prioridades nos datagramas. A sugestão tem a seguinte forma neste trabalho :

$$Q = \sum_{i=2}^{6} x_i \cdot \omega^{i-2} \tag{5.1}$$

onde :

- Q é o total de quota (digamos por mês) de direito do usuário;

- x_i é o número de pacotes enviados com prioridade i durante o período estabelecido (digamos por mês);

- ω é um parâmetro maior que 1.

Podemos observar que o somatório é tomado entre 2 (dois) e 6 (seis). Prioridades 0 (zero) e 1 (um) não seriam levadas em consideração na quota total; pacotes com estas prioridades são aceitos na rede porém tem seu serviço com menor prioridade. A prioridade 7 (sete) é utilizada para administração da rede. Algumas questões sobre abuso da quota por parte do usuário, tráfego entre operadoras concorrentes, tratamento de serviço FTP anônimo dentre outros são discutidos no mesmo trabalho. A idéia de tarifação pode ser

pensada como colocando-se preço nos pacotes conforme a prioridade escolhida, como em [53].

Em [60] é apresentado um modelo de tarifação para redes baseado em prioridades. Este modelo está diretamente associado a um processo de estabelecimento de conexão. Este está dividido em dois estágios separados : caracterização dos usuários e da rede, e negociação do contrato. As caracterizações e objetivos são dadas por :

- usuário: deseja maximizar o valor entre a diferença entre o valor que ele está dispondo a pagar (benefício) e o valor que é cobrado a ele. Neste caso o benefício é definido em termos de demanda e de QoS : $ben(QoS_1, \lambda_{i1}, QoS_2, \lambda_{i2}, ..., QoS_M, \lambda_{iM})$ onde M é o número de classes de prioridade e λ_{ij} é a demanda do usuário *i* pela classe de prioridade *j* (em número de pacotes).

- rede : deseja maximizar o benefício total dos usuários reservando, de forma ótima, os recursos da rede entre as classes de prioridade e escolhendo a melhor combinação de QoS.

A rede tarifa os usuários por pacotes transmitidos. Logo a cobrança total a um usuário por uma conexão é dada por $\sum_j P_j \lambda_{ij}$ onde P_j é o preço cobrado pelos pacotes das classes de prioridade *j*. Inicialmente a rede comunica sua política de prioridade (preços) aos usuários e estes, baseados nestas informações, determinam suas demandas para cada classe de prioridade. Novamente, do lado da rede, dada a demanda dos usuários por cada classe de prioridade, a rede precisa caracterizar o QoS de cada classe de prioridade.

O processo de negociação do contrato é um processo de reserva de recursos distribuído e de maximização eficiente. O estabelecimento de preços é importante no sentido de dar aos usuários incentivos para escolher as demandas e QoS certos para eles e para a rede. Na negociação do contrato a rede precisa inicialmente estabelecer os preços para cada prioridade estabelecendo o QoS para cada uma delas. A partir daí, de maneira interativa, os usuários declaram suas demandas para cada classe de prioridade à rede. A rede calcula o QoS, dada a demanda agregada para cada classe de prioridade, e anuncia um novo conjunto de preços e QoS. Este processo interativo continua até que o QoS projetado seja o mesmo a ser fornecido aos usuários e até que o benefício total dos usuários seja maximizado. O desenvolvimento matemático é apresentado no mesmo trabalho.

5.5.2 Tarifação de custo marginal

Existem vários recursos que podemos classificar como usados pelos usuários : capacidade de roteamento dos roteadores, a largura de faixa dos enlaces de comunicação, capacidade de disco e CPU de servidores. Quando os usuários acessam estes recursos eles normalmente levam em consideração seus próprios custos e benefícios relacionados a este uso, porém ignoram o congestionamento, retardo que este uso impõe aos outros usuários. Para uma análise geral de alguns mecanismos de tarifação em redes multiserviço introduz-se a seguinte notação [59] :

- x_i denota o uso do recurso comum pelo usuário *i*;

- $X = \sum_{j=1}^{n} x_j$ denota o uso total do serviço;

- *K* denota a capacidade total do recurso;

- $Y = X/K$ denota a utilização total do sistema;

- $u_i(x_i, Y)$ é uma função que representa as preferências dos usuários, diferenciável e côncava em x_i e em *Y*, representando o que eles estariam dispostos a pagar para receber um determinado serviço;

- *c(K)* é uma medida do custo de provimento do recurso;

- *W(K)* é a função representativa do benefício agregado fornecido aos usuários;

Esta notação é geral o bastante para capturar a essência de alguns recursos de rede. Por exemplo, considere um servidor de ftp (*file transfer protocol*). Dentro deste contexto x_i seria o número de bytes transferidos ao usuário *i*, *K* poderia ser a capacidade do servidor em termos de quantos bytes ele pode transferir em um dado intervalo de tempo e *X* poderia ser a quantidade total de bytes transferidos para todos os usuários. Por definição o padrão de eficiência é a maximização da soma de benefícios (funções utilidades) menos custos. Desta forma tem-se :

$$W(K) = \max_{x_i} \sum_{j=1}^{n} u_j(x_j, Y) - c(K) \tag{5.2}$$

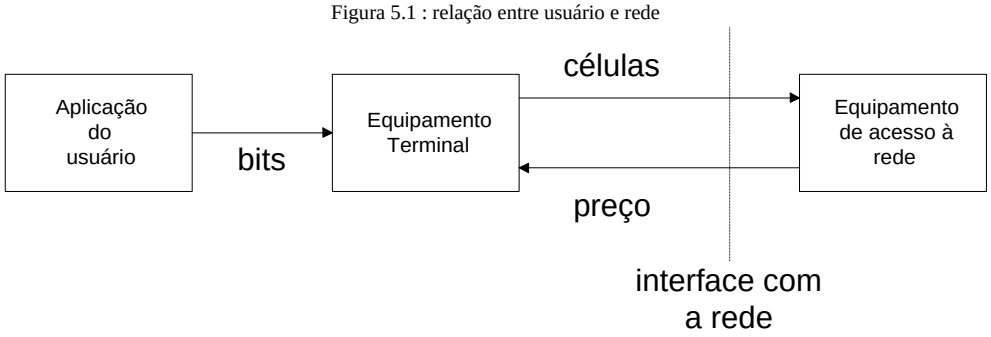
A solução ótima deve satisfazer a seguinte condição :

$$\frac{\partial u_i(x_i, Y)}{\partial x_i} = -\frac{1}{K} \sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y} \tag{5.3}$$

o que quer dizer que um incremento no uso do recursos por parte de um usuário não pode influir na otimização do sistema. Em outras palavras o incremento no benefício que um usuário obtém no uso do recurso se iguala ao incremento de custo que ele impõe aos outros usuários.

Em [56] é apresentado um esquema de tarifação de largura de faixa para um dado tamanho de fila compartilhado por vários usuários. Neste esquema os preços de largura de faixa são calculados em função da ocupação do fila. Neste esquema os usuários tem a responsabilidade de requisitar a largura de faixa, que, nas suas visões, irá satisfazer os QoS desejados. A função benefício descreve a relação entre o preço que o usuário estaria disposto a pagar por uma unidade de largura de faixa. A idéia da rede é cobrar ao usuário um valor correspondente ao custo de manutenção do tráfego da rede. A rede ajusta os preços da rede dinamicamente baseada nas condições da rede. Nesta configuração o usuário programa seu equipamento terminal com sua função benefício. A largura de faixa, em células por unidade de tempo (intervalo de tarifação), reservada ao usuário pela rede depende da função benefício declarada no equipamento terminal. O preço é enviado da rede para o equipamento terminal. Este esquema pode ser visualizado na Figura 5.1.

Figura 5.1 : relação entre usuário e rede



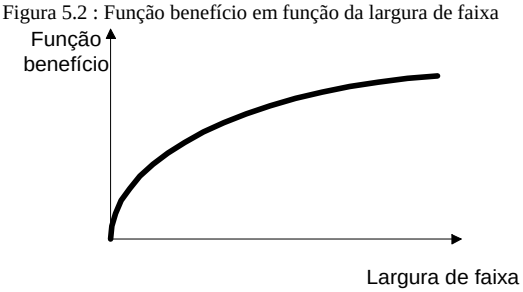
Este algoritmo de tarifação é distribuído pela rede e local aos nós de acesso ATM, sugerindo que ele pode ser capaz de resolver o problema da rapidez exigida na tomada de decisões do controle de admissão de conexões. Os usuários tem mais flexibilidade no acesso aos recursos da rede do que em outros esquemas de CAC, mas a eles é passada a responsabilidade em determinar a quantidade de recursos que o usuário deve ter para atingir suas necessidades.

5.5.4.1 Visão detalhada

Dentro deste esquema de tarifação vamos considerar que os caminhos virtuais são transportados por troncos físicos, cujas capacidades são consideradas constantes. Assumiremos que a multiplexação estatística seja feita entre os circuitos virtuais dentro dos caminhos virtuais (considerados de capacidade fixa).

Os usuários conectados a um comutador ATM desejam comunicar-se com vários destinos. Assume-se que suas funções benefício versus largura de faixa são côncavas crescentes como mostrado na figura 5.2 .

Figura 5.2 : Função benefício em função da largura de faixa



A figura está de acordo com a sugestão econômica de que inicialmente o usuário tem um ônus maior quando obtem um certo bem e que à medida que ele deseja mais recursos há uma diminuição incremental no seu valor.

O preço de uma unidade de largura de faixa no caminho virtual V_q é π_q . Cada usuário usa sua função benefício própria ben_q para chamada r para decidir quantas unidades de largura de faixa serão necessárias (preço total = b_{rq}). Este processo de decisão pode ser formulado matematicamente como :

$$\max \; ben_{rq}(b_{rq}) - \pi_q \cdot b_{rq} \tag{5.4}$$

O usuário resolve este problema de maximização para uma chamada específica e sua condição ótima é :

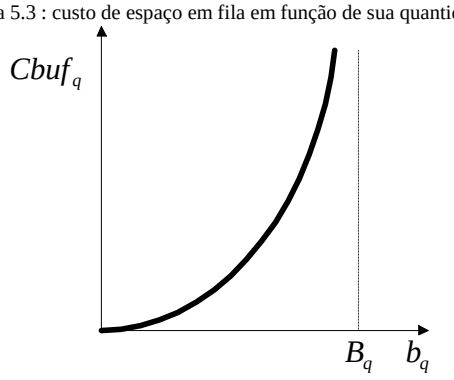
$$\frac{\partial ben_{rq}}{\partial b_{rq}} - \pi_q = 0, \text{ qualquer que seja } V_q \tag{5.5}$$

Um outro compromisso em relação à ocupação do fila deve ser estabelecido, assegurando que seu conteúdo b_q será menor que o seu tamanho B_q. Esta condição pode ser aproximada por uma função especial (custo de espaço de fila) que tem como objetivo o transbordo do fila. Por exemplo :

$$Cbuf_q(b_q) = \frac{1}{B_q - b_q} - \frac{1}{B_q} \tag{5.6}$$

O formato deste termo de custo, que assegura que à medida que a ocupação máxima do fila é atingida o seu preço tende a infinito, é mostrado na figura 5.3.

Figura 5.3 : custo de espaço em fila em função de sua quantidade



O problema do sistema pode ser formulado como se segue :

$$\max \sum_{todos\,V_q's} ben_{rq}(b_{rq}) \quad Cbuf_q \tag{5.7}$$

As condições ótimas do sistema são :

$$\frac{\partial ben_{rq}}{\partial b_{rq}} - \frac{\partial CBuf_q}{\partial b_q} \cdot \frac{\partial b_q}{\partial (\sum_r b_{rq})} = 0, \text{ quaisquer que sejam } r \text{ e } V_q \qquad (5.8)$$

A comparação entre as equações (5.7) e (5.8) mostra que estabelecendo

$$\pi_q = \frac{1}{(B_q - b_q)^2} \cdot \frac{\partial b_q}{\partial (\sum_r b_{rq})} \qquad (5.9)$$

a maximização do benefício do usuário (5.7) coincide com a otimização do sistema (5.8). Logo, se os preços são estabelecidos de acordo com (5.9), os usuários irão selecionar suas demandas ótimas de largura de faixa quando competem pelo caminho virtual. A rede pode induzir o comportamento ótimo do usuário estabelecendo preços de acordo com a equação 5.9 sem um conhecimento prévio da função benefício dos usuários.

5.5.4.2 Algoritmo de Tarifação Distribuído :

Seja o tempo dividido em intervalos de tarifação sucessivos de tamanho *T*. Dentro de cada intervalo as funções benefício de cada usuário são consideradas fixas podendo variar de intervalo para intervalo. O preço por unidade de largura de faixa no caminho virtual V_q , π_q é anunciado pela rede no início de cada intervalo de tarifação e permanece fixo durante todo o intervalo. Logo cada usuário resolve seu problema de maximização para decidir o valor de sua taxa média durante este período, transmitindo exatamente $T \cdot b_{rq}$ células durante o intervalo. Em cada nó da rede participante a conexão, as ocupações dos filas são medidas, permitindo à rede o cálculo do custo marginal de cada fila em relação ao seu tráfego total de entrada na rede ($\sum_r b_{rq}$), dado pela equação (5.3).

Se estes custos marginais são iguais aos preços que foram anunciados então o sistema opera em condições ótimas, de acordo com a equação (5.9). Caso contrário os preços precisarão ser modificados para corrigir a diferença. A seguir é feito um resumo sobre este método de reserva de largura de faixa / tarifação :

Passo 1 : a rede escolhe os valores iniciais para os preços $\{\pi_q\}$.

Passo 2 : a rede anuncia estes preços aos usuários no início do intervalo corrente de tarifação.

Passo 3 : a rede calcula os custos marginais dos filas em relação às larguras de faixa como no lado direito da equação (5.9).

Passo 4 : a rede ajusta os preços $\{\pi_q\}$ para reduzir a diferença entre os dois lados da equação (5.9). Volta ao Passo 1.

5.5.3 Tarifação Baseada no Recurso Utilizado

Um dos principais objetivos do estabelecimento de preços em redes de comunicação é que eles podem representar aos usuários os custos provenientes de suas ações. Se os preços refletem precisamente estes custos então os usuários podem comparar os benefícios proporcionados pelas suas ações com os custos relativos. A Tarifação Baseada no Recurso Utilizado pode ser utilizada para controlar o acesso por exemplo a um servidor WWW de maneira que aquele que mais valoriza o acesso tem a sua disposição uma maior quantidade de recursos.

A Tarifação Baseada no Recurso Utilizado não é somente uma maneira de obtenção de lucros. Se o provimento de serviço é desempenhado por empresas que competem entre si então os lucros são determinados pelo grau de competição. Os custos da implementação de um mecanismo mais elaborado podem ser diminuídos uma vez que esta modalidade de tarifação aumenta a funcionalidade e eficiência da rede visto que agora os usuários pagam proporcionalmente ao próprio uso.

Não é recomendável pensar que o aumento da oferta de largura de faixa seria uma medida suficiente para que o congestionamento seja minimizado. Os equipamentos de comutação (roteadores e *switches*) na realidade são computadores e à medida que novas tecnologias são implementadas sua capacidade de geração de tráfego na rede aumenta. Aliado a isso o aumento do número de usuários da rede tende a aumentar, utilizando para o acesso computadores cada vez mais capazes de gerar um grande quantidade de tráfego em curtos períodos de tempo. Dentro deste contexto fica ainda mais caracterizada uma necessidade de se controlar o acesso à rede, sendo a tarifação uma das formas de controle. Passa-se agora à decrição de alguns mecanismos de Tarifação Baseada no Recurso Utilizado.

Em [54] é apresentado um mecanismo de negociação interativo entre usuário e rede para o estabelecimento da conexão. O objetivo da rede é maximizar a função benefício total

e do usuário é maximizar o sua relação custo/benefício (representado pelo preço cobrado pela rede e pelo QoS fornecido pela rede ao tráfego gerado). Neste caso o usuário paga pela melhor combinação para ele de espaço em fila e largura de faixa mantendo-se o seu QoS desejado. A rede sugere preços para a largura de faixa e espaço em fila em cada enlace e em seguida o usuário decide a quantidade de cada recurso a ser usado para todos os enlaces que fazem parte da rota a ser utilizada. O processo de interação continua à medida que a rede estabelece novos preços baseados na demanda e no grau de utilização da rede (dos recursos da rede), e que os usuários mudam seus requisitos por recursos baseados nos novos preços propostos pela rede. Este processo interativo irá convergir para uma solução ótima onde o bem-estar de todos é maximizado.

A vantagem deste esquema de política de tarifação é que a rede não precisa saber a caracterização detalhada do tráfego a ser gerado pelo usuário e conseqüentemente não há necessidade de política de policiamento. A rede fornece os recursos que são utilizados pelos usuários, que os utilizam de acordo com suas necessidades.

A largura de faixa estatística juntamente com o valor de fila utilizado para o seu cálculo podem ser utilizados na medida de recursos utilizados. Esta medida pode ser utilizada como componente dos recursos utilizados na cobrança feita aos usuários. Dado que a largura de faixa estatística faz parte da taxa cobrada pode-se indentificar dois métodos extremos que representam esta situação [58].

Considere-se fontes do tipo j , onde o 'tipo' é diferenciado por parâmetros do contrato de tráfego e possivelmente alguma outra informação estática. A rede poderia fazer uma estimativa empírica da largura de faixa estatística, como por exemplo (fórmula da capacidade equivalente) , baseada nas características de conexões do tipo j passadas e taxar os usuários baseada neste valor. Pode-se entender melhor este esquema comparando-se com o sistema de restaurante rodízio. Dentro deste esquema o consumidor é cobrado não pela comida consumida, mas pelo conhecimento da quantidade que consumidores anteriores comeram. Entretanto este esquema beneficia àqueles que comem acima do normal enquanto que aqueles que comem pouco se sentem prejudicados.

O problema em adotar um esquema de cobrança, no qual, a taxa a ser paga é proporcional à largura de faixa estatística, que é totalmente determinada em função dos parâmetros que são fornecidos no momento de estabelecimento da conexão, é que não

existe nenhum mecanismo de realimentação para penalizar usuários que usam mais recursos do que o típico utilizado. Dentro deste contexto o usuário pode usar a quantidade máxima de recursos que o contrato para usuários daquele tipo permite, dentro dos limites de policiamento negociado. Por outro lado, os usuários do mesmo tipo que não utilizam todos os recursos oferecidos se sentem injustiçados uma vez que pagam mais por aquilo que não utilizam.

No outro extremo, os usuários podem ser cobrados levando-se em conta só as medidas que são feitas durante as conexões. Isto teria uma falha conceitual como se segue. Suponha que um usuário requer uma conexão que será policiada pela taxa de pico, mas durante esta ele transmite uma pequena quantidade de células, logo a estimativa inicial (no processo de negociação da conexão) da largura de faixa estatística baseada será maior que este cálculo feito com as medidas durante a duração da conexão. Com isso o conceito de se reservar recursos durante a negociação de estabelecimento da conexão não se aplica e neste caso a rede reservaria recursos que não seriam utilizados, o que seria desvantajoso para ela.

Um esquema de tarifação baseado na medida da largura de faixa estatística pode ser definido em função de parâmetros estáticos (como taxa de pico por exemplo) e dinâmicos (como o volume real transferido durante uma conexão juntamente com sua duração). Neste caso seriam policiados os parâmetros estáticos e medidos os parâmetros dinâmicos; a largura de faixa estatística seria limitada por uma função linear dos parâmetros medidos e coeficientes que dependeriam dos parâmetros estáticos. Estas funções lineares serviriam como base do mecanismo de tarifação. No presente trabalho concentra-se somente nos parâmetros estáticos, não sendo utilizado qualquer tipo de policiamento de tráfego.

5.5.3.1 Formulação matemática

Como já visto a largura de faixa efetiva é bastante útil na modelagem de fontes de tráfego bem como no controle de admissão. Através do cálculo correto da largura de faixa equivalente pode-se garantir que a qualidade de serviço será garantida sem o desperdício de recursos. Em [3,5] é apresentado um mecanismo de tarifação no qual o usuário é aconselhado a declarar precisamente seus parâmetros de tráfego para que não seja preciso o uso de policiamento da sua fonte. A rede representa a largura de faixa equivalente de um usuário por uma função $B(Z)$, que é obtida através de parâmetros passados pelo usuário. No caso o usuário passa para a rede uma estimativa z do parâmetro Z. O usuário então é tarifado por um valor $aT + bV$, onde T é o tempo de duração da conexão e V é volume

transferido durante a conexão. Os parâmetros *a* e *b* dependem da declaração de parâmetros feita pelo usuário . As declarações de *a* e *b* podem ser definidas como de tal forma que a tarifa cobrada ao usuário é dada por $\alpha \cdot T$ onde α é uma aproximação da largura de faixa equivalente. Em [55] é dado um exemplo em que a aproximação de largura de faixa equivalente $\alpha(m)$ é côncavo em *m* (a taxa média da fonte). Se *a* e *b* são tais que a reta *a* + *bm* é tangente à aproximação $\alpha(m)$ no ponto m_0 , então o usuário irá minimizar o valor a ser pago se ele selecionar o par *(a,b)* que corresponde à tangente a $\alpha(m)$ no ponto $m_0 = M$, onde *M* (=V/T) é a taxa média, de fato, do usuário. Neste o valor correspondente a ser pago é dado por $(a + b \cdot \frac{V}{T})T$, sendo que o valor cobrado por unidade de tempo será dado por $a + bM = \alpha$. Nesta análise a única exigência é que $\alpha(m)$ seja côncavo em *m*.

Pode-se notar que a tarifa paga será de $\alpha \cdot T$, considerando que o usuário sabe a sua taxa média. Se ele é incorreto na declaração de seus parâmetros, dele será cobrada uma taxa maior.

O valor do parâmetro declarado pelo usuário não necessariamente precisa ser a média do seu tráfego gerado. O importante é que a aproximação da capacidade equivalente seja côncava no parâmetro declarado. Por exemplo, para uma fonte "on-off" tem-se que a expressão da capacidade equivalente é dada por:

$$\frac{1}{st}\ln[1 + \frac{m}{h}(e^{sht} - 1)] \tag{5.10}$$

onde *m* é a taxa média, *h* é a taxa de pico, *s* é o parâmetro relacionado à qualidade de serviço. Considerando $z = 1 + \frac{m}{h}(e^{sht} - 1)$ tem-se que a aproximação da capacidade equivalente é côncava em z , então a formulação utilizada no parágrafo anterior pode ser utilizada. Neste caso o usuário que escolher o par *(a,b)* será cobrado na quantidade de :

$$\begin{aligned}(a + bz)T &= \{a + b[1 + \frac{V}{Th}(e^{sht} - 1)]\} \cdot T \\ &= (a + b)T + b \cdot V\frac{e^{sht} - 1}{h} \ .\end{aligned} \tag{5.11}$$

Se assume-se que a taxa de pico é independente do volume de tráfego transferido tem-se que o valor esperado do preço por unidade de tempo é dada por $(a + b) + bm\sigma$, onde

m é a taxa média e $\sigma = E\left\{\frac{(x-1)^+}{h}\right\}$. Logo para declaração de parâmetros que levariam à melhor tarifa o usuário deve ter conhecimento, neste caso, da média "escalonada" $m\sigma$.

5.5.3.2 Aplicação ao modelo de tráfego apresentado em 3.5.1

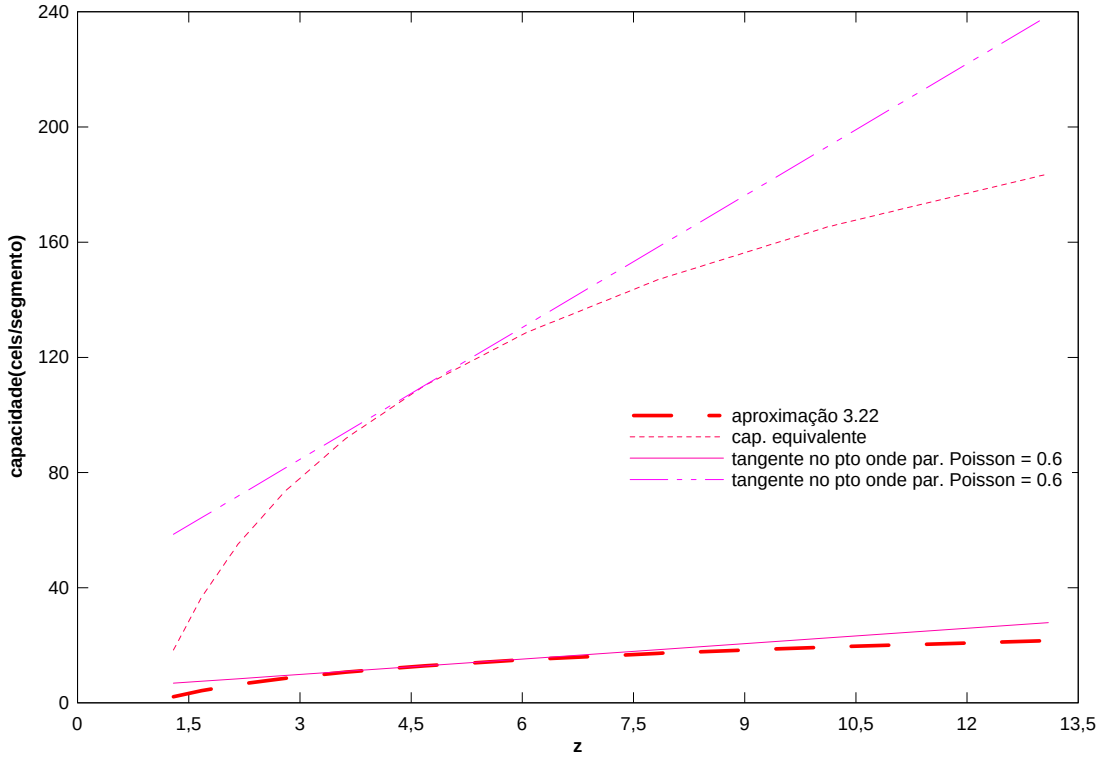
O modelo de tráfego apresentado em 3.5.1 permitiu que fosse deduzida a fórmula da capacidade equivalente e a adequação à uma formula de comportamento assintótico. Estes dois mecanismos permitem acomparação entre as aproximações de reservas de fila e largura de faixa. Foi verificado que a aproximação de capacidade equivalente reserva mais recursos que o necessário para a obtenção de uma certa qualidade de serviço. Como este modelo reserva mais recursos que o necessário, é razoável pensar que, em se usando tal modelo para esquemas de tarifação, o preço cobrado quando usa-se esta aproximação é maior que o preço cobrado quando usa-se modelos mais realísticos, como aquele apresentado em 3.5.1 fórmula (3.22).

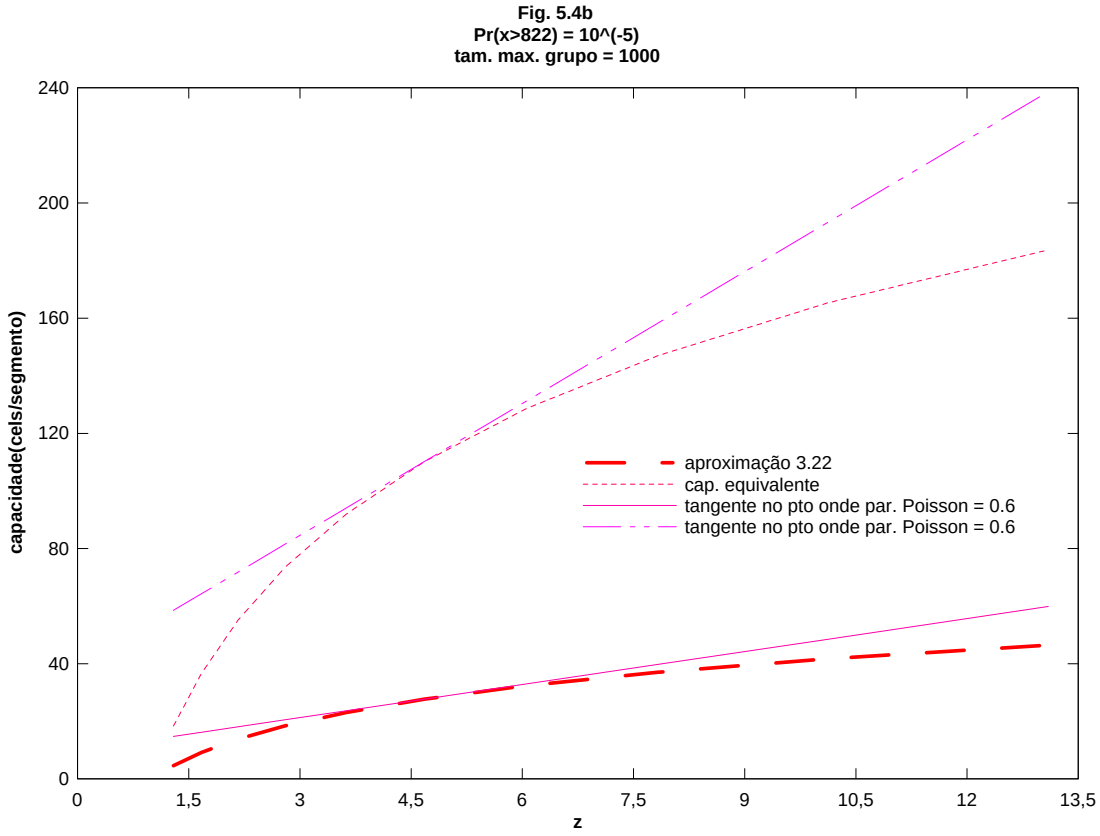
A comparação entre os dois modelos é feita com base nos valores de largura de faixa (preço cobrado) em função dos parâmetros de tráfego tratados e do tamanho do fila utilizado. Primeiramente é calculado o valor de $z = e^{\lambda \cdot E\left[e^{\delta Y}\right]-\lambda}$, onde λ é a intensidade de tráfego e Y é a variável aleatória que representa o tamanho do grupo de células gerado por um usuário no slot de tempo. Mais uma vez ocorre a truncagem deste tamanho de grupo. A capacidade equivalente calculada para o nosso modelo é dada por :

$$\frac{1}{\delta \cdot t} \ln \left\{ z^{\left[1 - E\left[e^{\delta Y}\right]\right]} \right\} \qquad (5.12)$$

Logo pode-se verificar que a capacidade equivalente calculada é convexa em relação ao valor de z. Para a qualidade de serviço de 10^{-4}, com $\delta = 0.014$ o valor do fila é dado por B = 658, ou seja tem-se $\Pr(x > 658) \approx 10^{-4}$. O tamanho máximo do grupo de células geradas por uma fonte é de 1000. Para os mesmos parâmetros e mesmo QoS é calculado o valor da largura de faixa correspondente para a estimativa feita pela aproximação dada pela equação (3.22). A comparação pode ser visualizada nas Figuras 5.4 e 5.5.

Fig. 5.4a
Pr(x>658) = 10^(-4)
tam. max. grupo = 1000





Pode-se afirmar que a diferença de preço cobrado entre as duas aproximações é função de como elas reservam os recursos para os usuários. Pode-se então concluir que além de fazer o uso irracional dos recursos da rede, o preço cobrado por uma aproximação exagerada é maior. Para um caso prático pode se dizer que uma operadora que utiliza uma boa aproximação para a reserva de recursos, além de poder suportar muito mais clientes pode também cobrar preços mais baratos para uma mesma qualidade de serviço, em relação a uma concorrente que faça a utilização de uma aproximação errônea.

Capítulo 6

Conclusão

Este trabalho teve por finalidade a realização de alguns estudos de suma importância na área de redes de comunicação projetadas para suportar diversos tipos de serviço. Estes estudos são fundamentais para uma melhor utilização e eficiência de tais redes. As principais dificuldades encontradas, para que os objetivos de utilização e eficiência sejam atingidos, são decorrentes do fato de que tais redes devem tratar diferentes padrões de tráfego com diferentes requisitos de desempenho, tais como atraso máximo de transferência e probabilidade de perda de células (caso para rede ATM).

No estudo de modelagem de fontes de tráfego o enfoque principal foi dado aos modelos de fontes VBR com requisitos de desempenho. Como consequência deste problema alguns modelos de fontes de tráfego de voz, vídeo e dados foram apresentados. Foram verificadas também algumas das principais diferenças entre os modelos tradicionais (chamados Markovianos) e modelos mais recentemente estudados (chamados modelos Auto-Similares). Dentro deste contexto são verificadas algumas implicações do desempenho do sistema de fila quando alimentada por fontes de tráfego auto-similares.

No estudo de Controle de Admissão de Conexões (CAC) verificou-se que este é apenas um dos controles que objetiva o uso racional dos recursos de rede e ao mesmo tempo foca a garantia de Qualidade de Serviço que é fornecida ao usuário. Este tipo de controle foi classificado em Reserva não Estatística, Controle determinístico (no qual considera-se o comportamento da fonte no "pior" caso) e o Controle Estatístico, que permite que a Qualidade de Serviço não seja respeitada dentro de alguma pequena faixa de probabilidades. Foi dada ênfase ainda ao problema de múltiplas escalas de tempo no comportamento do desempenho de um sistema de fila, indicando a importância da construção de processos e modelos analíticos de resolução de filas que atentem para este problema.

A maior contribuição deste trabalho recai sobre uma extensão de um modelo de tráfego auto-similar e suas aplicações em controle de admissão de conexões e tarifação em

redes multiserviço. O modelo de tráfego original fazia a análise de uma fila considerando os usuários do sistema como rajadas. Apesar de ser uma análise válida e contribuinte, em sistemas reais, como roteadores ou comutadores utilizados em redes de computadores, tal usuário não tem o seu sentido, Que interpretação poderia-se dar a frases do tipo: "o comutador perdeu duas rajadas" ? A idéia da extensão do modelo, considerando agora as células que formam as rajadas, é aproximar mais o modelo do mundo real. Uma vez que faz mais sentido dizer que a fila do comutador tem capacidade de armazenar 500 células, ou que foram perdidas 10 células durante uma sessão de vídeo conferência, o modelo extendido a células parece ser mais útil.

Uma vez que o modelo de tráfego foi definido, o próximo passo foi analisar um sistema de fila que recebe tal tráfego. Inicialmente foi realizada uma análise exata do modelo observando-se que considerando-se o modelo durante um segmento de tempo seguinte à saída de um usuário, verifica-se que o número de usuários no sistema contém todas as informações do passado e do presente que determinam os estados futuros do sistema (o estado do sistema durante uma saída é um processo de Markov). A limitação observada nesta primeira análise é que somente são observados os instantes em que há partida de usuários do sistema. Uma situação mais completa seria aquela que retratasse um sistema de fila do tipo $Q_{t+1} = (Q_t + A_t - C)^+$.

Para a análise da fila do tipo $Q_{t+1} = (Q_t + A_t - C)^+$ foram utilizadas duas aproximações: a capacidade equivalente, mais geralmente utilizada e uma segunda que considera chegadas subexponenciais.

Esta duas aproximações permitiram a verificação do desempenho da fila estimando-se a distribuição de probabilidade de seu tamanho (Pr(Q>x)). Anteriormente, através de teste de simulação foi verificado que a aproximação subexponencial é uma boa aproximação para a estimativa do desempenho da fila. De posse destas duas aproximações foram realizadas algumas comparações de desempenho entre elas. De um modo geral, a aproximação da capacidade equivalente, dentro de intervalos reais de tamanho de fila, tende a fazer uma reserva de uma quantidade maior de recursos de rede (no caso largura de faixa e tamanho de fila) quando comparada com a aproximação 4.22 (subexponencial).

Outra parte tocante ao projeto de redes foi tratada no trabalho: a aplicação da tarifação. Foi verificado que atualmente não existe uma política justa de tarifação, ou seja, uma política que garanta uma cobrança ao usuário proporcional ao tráfego que ele gera, ao

congestionamento que ele impõe a outros usuários e à qualidade de serviço que ele contrata. Foi realizado o estudo de algumas propostas encontradas na literatura, sendo que elas foram classificadas em três grupos: Tarifação baseada em Prioridade, Tarifação de Custo Marginal e Tarifação baseada no Recurso Utilizado. Dentro desta última, utilizando resultados provenientes das aproximações subexponencial e capacidade equivalente, foi verificado que, para uma mesma QoS, o preço cobrado pela aproximação da capacidade equivalente é maior que a da aproximação subexponencial. Isto é consequência direta dos resultados obtidos na reserva de recursos.

Este trabalho, como um todo, não tem o objetivo de ser completo nem tão pouco definitivo. Muitas questões relacionadas aos assuntos aqui mencionados ainda existem. Dentre elas podemos citar algumas:

- multiplexação estatística entre classes de tráfego com características e requisitos distintos, observando-se a influência entre elas;

- toda a análise apresentada neste trabalho é feita porque há a truncagem do tamanho máximo do grupo do modelo de tráfego. Há indícios [32] que essa limitação pode ser ultrapassada através de um estudo mais completo de Teoria de Processos Auto-Similares, Teoria de Grandes Desvios e Programação Linear;

- análises de filas mais complexas onde fosse possível a adição de novos usuários ao sistema e ao mesmo tempo observando-se o desempenho da fila.

Referência Bibliográfica

[1] G. D. Stamoulis, M. E. Anagnostou e A. D. Georgantas, "Traffic source models for ATM networks : a survey ", Computer Communications, vol. 17 , número 6, junho de 1994, pags 428-438.

[2] Commission of European Communities, "COST 224 Performance evaluation and design of multiservice networks", outubro de 1991.

[3] V. Paxson e S. Floyd, "Wide Area Traffic : The Failture of Poisson Modeling", IEEE/ACM Trasactions on Networking, vol. 3, número 3, junho de 1995.

[4] Onvural, Raif O. , "Asynchronous Transfer Mode Networks : Performence Issues", Artech House, inc ,1994.

[5] H. Heffes e D. M. Lucantoni, "A Markov modulated characterization of packet voice and data traffic and related statistical multiplexer performance", IEEE Journal on Selected Areas in Communications, vol 4 pags 856-868, setembro de 1986.

[6] J. N. Diagle e J. D. Langford , "Models for Analysis of Packet Voice Communication Systems", IEEE Journal on Selected Areas in Communications, vol 6, setembro 1986, pags 847-855.

[7] Gunnar Karlsson, "Asynchronous Transfer of Video", Swedish Institute of Computer Science, SICS Research Report R95:14.

[8] W. Verbiest, Luc Pinnoo e Bart Voeten, "The Impact of the ATM Concept on Video Coding", IEEE Journal on Selected Areas in Communications, vol 6, num. 9, dezembro de 1988, pags. 1623-1632.

[9] Alan Weiss, "An Introduction to Large Deviations for Communication Networks", IEEE Journal on Selected Areas in Communications, vol 13, num. 6, agosto de 1995, pags. 938-952.

[10] G. Kesidis, J. Walrand e C. S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", IEEE/ACM Trasactions on Networking, vol. 1, número 4, agosto de 1993.

[11] F. Kelly, "Notes on Effective Bandwidths", disponível em http://www.statslab.cam.ac.uk/~frank/

[12] B. B. Mandelbrot e J. W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Apllications", SIAM Review, vol. 10, 1968, pags 422-437.

[13] C. W. Grange e R. Joyeux, "An Introduction to Long-Memory Time Series Models and Fractional Differing", Times Series Anal.,vol 1,1980, pags 15-29.

[14] R. Guerin, H. Ahmadi e M. Naghshineh, "Equivalent capacity and its application to bandwidth application in high-speed networks", IEEE Journal on Selected Areas in Communications, vol 9, pags 968-981, setembro de 1991.

[15] Anwar I. Elwalid and Debasis Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks", IEEE/ACM Trasactions on Networking, pags 329-343, junho de 1993.

[16] Gagan L. Choudury, David M. Lucantoni, and Ward With, "Squeezing the Most Out ATM" , IEEE Trans. on Commun., vol. 44,pags 203-217, fevereiro de 1996.

[17] M. F. Abdalla, "Análise de Mecanismos de Controle de Admissão de Conexão para Redes ATM", Tese de Mestrado, Universidade Federal do Rio de Janeiro, PEE/COPPE/UFRJ, 117 p, Rio de Janeiro, setembro de 1996.

[18] E. W. Knightly and Hui Zhang, "D-BIND : An Accurate Traffic Model for Providing QoS Guarantees to VBR Traffic, IEEE/ACM Transactions on Networking.

[19] R. Cruz, "A calculus for network delay, part I : Network elements in isolation, Transaction

[20] D. Ferrari and D. Verma, "A scheme for real-time channel establishment in wide-area networks", IEEE JSAC, abril de 1990, pags 368-379.

[21] J. W. Roberts, "Variable-Bit-Rate Traffic Control in B-ISDN", IEEE Communications Magazine, setembro de 1991, pags 50-56.

[22] I. Norros, J. W. Roberts, A. Simonian e J. T. Virtamo, "The Superposition of Variable Bit Rate Source in an ATM Multiplex", IEEE JSAC, abril de 1991, pags 378-386.

[23] L. Kleinrock, Queuering Systems, vol 1 , New York : Willey 1975.

[24] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", IEEE/ACM Trasactions on Networking , vol 2, pags 1-16, fevereiro de 1994.

[25] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, "Long-Range Dependence in Variable-Bit-Rate Video Traffic", IEEE Trans. on Commun., vol. 43, pags 1566-1579 fevereiro/março/abril de 1995.

[26] Gusella, R. , "Characterizing the Variability of Arrival Processes with Indexes of Dispersion", IEEE Journal on Selected Areas in Communications, vol 9, pags 203-211, fevereiro de 1991.

[27] W. E. Leland e D. V. Wilson, "High time-resolution measurement and analysis of LAN Traffic : Implications for LAN interconnection", Proc. de INFOCOM 1991, pags. 1360-1366.

[28] H. J. Fowler e W. E. Leland, "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management", IEEE JSAC, setembro de 1991, pags 1139-1149.

[29] I. Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks", IEEE Journal on Selected Areas in Communications, vol. 13, número 6, agosto de 1995, pags 953-962.

[30] A. Erramilli, O. Narayan e W. Willinger, "Experimental Queuering Analisys with LongRange Dependent Packet Traffic", IEEE/ACM Trasactions on Networking, abril de 1996.

[31] D. Veitch, "Novel Models of Broadband Traffic", Proc. IEEE Globecom 93, Houston, dezembro de 1993, pags. 1057-1061.

[32] M. Parulekar e A. Makowski, "Tail probabilities for a multiplexer with self-similar traffic", Proc. IEEE Globecomm 1996 pags. 1452-1459.

[33] H. G. Perros e K. M. Elsayed, "Call Admission Control Schemes : A Review", IEEE Communications Magazine, novembro de 1996, pags 82-91.

[34] E. Gelenbe, X. Mang e Raif Önvural, "Bandwidth Allocation and Call Admission Control in High-Speed Networks", IEEE Communications Magazine, maio de 1997, pags 122-129.

[35] D. R. Cox, "Long-range dependence : a review", Statistics : An Appraisal H. A. David e H.T. David, Eds. Ames, IA : Iowa State University Press, 1984, pags 55-74.

[36] W. Willinger, M. S. Taqqu, R. Sherman e D. V. Wilson, "Self-similarity through high-variability : Statistical analysis of Ethernet LAN traffic at the source level", IEEE/ACM Trasactions on Networking, vol. 5 pags 71-86, fevereiro de 1996.

[37] M. E. Crovella, A. Bestavros, "Self-Similarity in World Wide Web Traffic : Evidence and Possible Causes", IEEE/ACM Transactions on Networking, vol. 5, número 6, dezembro de 1997, pags. 835-846.

[38] J. J. Bae e T. Suda, "Survey of traffic control schemes and protocols in ATM networks", Proceeding of IEEE, vol. 79, num. 2, pags. 170-189, fevereiro de 1991.

[39] A. A. Lazar, G. Pacifini e D. E. Pendarakis, "Modeling video sources for real-time scheduling", Multimedia Systems, vol. 1 número 6 pags 253-266. Disponível em http://www.ctr.columbia.edu/comet/publications.

[40] Gagan L. Choudury, David M. Lucantoni, and Ward With, "Squeezing the Most Out ATM" , IEEE Trans. on Commun., vol. 44,pags 203-217, fevereiro de 1996.

[41] Aurel A. Lazar, Predrag R. Jelenkovic, "On Dependence of Queue Tail Distribution on Multiple Times Scales of ATM Multiplexers"

[42] P. R. Jelenkovic, A. A. Lazar, "The Effect of Multiple Time Scales and Subexponentiality in MPEG Video Streams on Queueing Behavior", IEEE JSAC, vol. 15, número 6, agosto de 1997.

[43] P. R. Jelenkovic, A. A. Lazar, "The Asymptotic Behavior of a Network Multiplexer with Multiple Time Scale and Subexponential Arrivals", Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York.http://ctr.columbia.edu/comet/publications

[44] W. Fischer e K. Meier-Hellstern, "The markov-modulated poisson process (MMPP) cookbook", Performance Evaluation, vol. 18, pags. 149-171, março 1991.

[45] D. P. Heyman e T. V. Lakshman, "Source Models for vbr broadcast-video traffic", IEEE/ACM Trans. Networking, vol. 4, pags. 40-48, fevereiro de 1996.

[46] B. Maglaris, D. Anastassiou, P. Sen, N. Rikli, "Models for Packet Switching of Variable-Bit-Rate Video Sourses", IEEE Journal on Selected Areas in Communications, vol. 7, junho 1989, pags 865-869.

[47] Ross, Sheldon M. , "Introduction to Probability Models", Fourth Edition, Academic Press, Inc 1989.

[48] P. Skelly, M. Schwartz e S. Dixit, " A Histogram-Based Model for Video Traffic Behavior in an ATM Network Node with an Application to Congestion Control " , IEEE INFOCOM 1992 pags 95-104, maio de 1992.

[49] N. Likhanov, B. Tsybakov e N. D. Georganas, "Analysys of an ATM Buffer with Self-Similar ("Fractal") Input Traffic, Proc. IEEE INFOCOM 95, Boston, pags. 985-991, abril de 1995.

[50] M. S. Schmookler, "Limited Capacity Discrete Time Queues with Single or Bulk Arrival", IBM Corporation, Systems Development Division, junho de 1970.

[51] P. R. Jelenkovic, A. A. Lazar, "Subexponential Asymptotics of a Network Multiplexer", Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York.http://ctr.columbia.edu

[52] B. Tsybakov e N. D. Georganas, "On Self-Similar Traffic in ATM Queues : Definitions, Overflow Probability Bound and Cell Delay Distribution",

[53] R. Cocchi, D. Estrin, S. Shenker and L. Zhang, "Pricing in Computer Networks : Motivation, Formulation and Example" , IEEE/ACM Transaction on Networking, vol. 1, num. 6, dezembro de 1993, pags. 614-627.

[54] S. H. Low and P. P Varaiya, "A New Approach to Service Provisioning in ATM Networks" , IEEE/ACM Transaction on Networking, vol. 1, num. 3, dezembro de 1993, pags. 547-553.

[55] F. P. Kelly, "On tariffs, policing and admission control for multiservice networks", Operational Research Letters 15,1994, pags 1-9 .

[56] J. Murphy and L. Murphy, "Bandwidth Allocation By Pricing In ATM Networks"

[57] F. P. Kelly, "Charging and Accounting for Bursty Connections", disponível em www.statlab.com.ac.uk/~frank/papers.

[58] C. Courcoubetis, F. P. Kelly, V. Siris e R. Weber, "A study of simple usage-based charging schemes for broadband networks", ICS-Forth Techinical Report número 193, maio de 1997.

[59] J. K. Mackie-Mason and H. R. Varian, "Princing Congestionable Network Resources", University of Michigan, julho de 1994, disponível em www.sims.berkley.edu/resourses/infoecon/Pricing.html

[60] H. Jiang, I. Sidhu e S. Jordan, "A Pricing Model for Networks with Priorities" disponível em www.sims.berkley.edu/resourses/infoecon/Pricing.html

[61] R. Bohn, H. Braun e S. Wolff, "Mitigating the coming Internet crunch: multiple service levels via Precedence" disponível em www.sims.berkley.edu/resourses/infoecon/Networks.html

[62] H. Saito, "Teletraffic Technologies in ATM Networks", Boston, Lodon : Artech House, primeira edição, 1994.

[63]J. K. Mackie-Mason e H. R. Varian, "Pricing the Internet", Technical report, Universidade de Michigan ,maio de 1993. Disponível em http://www.sims.berkley.edu/resourses/infoecon/Networks.html