Pós-Graduação em Ciência da Computação

“Avaliação de desempenho de VoIP através de modelos estocásticos utilizando distribuições poli-exponenciais”

Por

Antonio Ricardo Pereira Cavalcanti

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao
Recife, Agosto de 2008

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANTONIO RICARDO PEREIRA CAVALCANTI

"AVALIAÇÃO DE DESEMPENHO DE VOIP ATRAVÉS DE MODELOS ESTOCÁSTICOS UTILIZANDO DISTRIBUIÇÕES POLI-EXPONENCIAIS"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADOR(A): PAULO ROMERO MARTINS MACIEL

RECIFE, AGOSTO/2008

Agradecimentos

Agradeço a Deus, por me fazer acreditar na realização deste mestrado.

Ao Professor Paulo Maciel, pela orientação, apoio e pela enorme paciência, que foram essenciais para realização deste trabalho.

Ao Professor Almir Guimarães, pela contribuição fundamental ao desenvolvimento do estudo de caso.

Aos professores Ricardo Massa e Ricardo Salgueiro, pelo convite aceito para compor a banca de defesa do mestrado.

Ao Departamento de Redes do DETRAN-PE, em especial a Bruno Bezerra, Carlos Silva e Sildomar Ivson que proporcionaram os meios para a realização deste trabalho.

Gostaria de registrar meus sinceros agradecimentos como reconhecimento da dedicação de muitas pessoas que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Finalmente, gostaria de agradecer a minha família, em especial a meus pais, Antonio e Dayse, minha princesa, Ana Paula, minha irmã e meu cunhado, pelo incentivo, ajuda e apoio moral.

Resumo

O crescimento e a popularização da internet aliados às novas tecnologias de voz e o surgimento dos enlaces de comunicação mais velozes proporcionam uma infra-estrutura convergente de dados e de voz conhecida como Voz sobre IP (VoIP – *Voice over Internet Protocol*). As principais vantagens são: redução de custo de comunicação e a utilização do serviço de voz com aplicações de dados. A internet não foi projetada com a intenção de transportar informações em tempo real, pois utiliza o modelo de melhor esforço onde não há garantias de entrega e de limite de atraso dos pacotes. Em contrapartida, as aplicações VoIP requerem que a entrega dos pacotes de voz sejam com atrasos e perdas limitadas. Portanto é preciso garantir um nível de conversação usando VoIP comparado as redes de telefonia tradicional. A necessidade de utilizar VoIP em diversas arquiteturas de rede requer que o ambiente não sofra degradação de desempenho. Neste trabalho, nós apresentamos um modelo de avaliação de desempenho de VoIP baseado em modelos estocásticos utilizando distribuições poli-exponenciais.

Palavras-Chave: VoIP, avaliação de desempenho, modelos estocásticos, distribuições poli-exponenciais.

Abstract

The growth and popularization of the Internet allied with new technologies and the emergence of voice communication links faster to provide an infrastructure converged voice and data known as Voice over IP (VoIP – *Voice over Internet Protocol*). The main advantages are: reduction in the costs of communication and the use of the service of voice with data applications. The internet was not designed with the intention of supporting real-time information, as it uses the model of best effort that offers no delivery guarantee and limit the delay of packages. However, VoIP applications that require delivery of packages of voice with delays and losses limited. Therefore, it is necessary a level of conversation using VoIP compared to traditional telephony networks. The need of using VoIP in different network architectures requires that the environment should not have any performance degradation. In this work, we present a VoIP performance evaluation model based on the stochastic models using distributions poly-exponentials.

Keywords: VoIP, evaluation of performance, stochastic models, distributions poly-exponentials.

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas, Siglas e Símbolos

3G	Terceira Geração
AP	Acess Point
FCFS	First-Come-First-Served
FIFO	First In First Out
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GSPN	Generalised Stochastic Petri Nets
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
ITU	International Telecommunication Union
LCFS	Last-Come-First-Served
MOS	Mean Opinion Score
NTP	Network Time Protocol
PCM	Pulse Code Modulation
PSTN	Public Switch Telephone Network
RdPs	Redes de Petri
RR	Round Robin
RTCP	RTP control protocol
RTP	Real Time Protocol
SIP	Protocolo de Inicialização de Sessão
SPN	Stochastic Petri Net
TCP	Transport Control Protocol
TMB	Tamanho Médio do Buffer
TMT	Tempo Médio de Transmissão
TTPD	Tempo de Transmissão dos Pacotes de Dados
TTPV	Tempo de Transmissão dos Pacotes de Voz
UDP	User Datagram Protocol
VoIP	Voice over Internet Protocol
WLAN	Wireless Local Area Network

Capítulo 1 - Introdução

Este capítulo apresenta o contexto no qual esta dissertação encontra-se inserida, apresentando as motivações, as justificativas para o seu desenvolvimento, seus principais objetivos, além da sua estrutura em capítulos.

O progresso tecnológico unido com a popularização da internet provocou profundo impacto no sistema de telefonia tradicional. O sistema tradicional de comunicação de voz se baseia na reserva de largura de banda e pela duração de uma chamada telefônica. Essa reserva garante uma boa qualidade nas interações telefônicas, mas, em contrapartida, tem um custo alto, assim como também não faz um uso eficiente dos recursos de rede, uma vez que a largura de banda fica reservada durante todo o período de duração da chamada em curso [12] .

A Internet apresentou um rápido crescimento nos últimos anos, possibilitando, a um custo baixo, a sua utilização ubíqua. Com o surgimento de enlaces de comunicação mais velozes, as aplicações de voz passaram a ser uma opção viável.

Uma das principais alternativas ao sistema de telefonia tradicional é a adoção de Voz sobre IP, ou apenas VoIP (Voice over Internet Protocol) [38] . VoIP é uma tecnologia multimídia que transmite voz em pacotes sobre uma rede de dados IP (Internet Protocol) que proporciona maior interação entre membros de uma determinada comunidade. Estamos falando de uma estrutura de comunicação convergente, um ambiente que pode transportar voz e dados pelo mesmo canal de comunicação. Vários outros benefícios provêm do uso da tecnologia de VoIP, dentre eles podemos citar:

- uma única infra-estrutura é capaz de transportar dados, correio eletrônico, voz, *voice mail;*
- redução dos custos das ligações;

- utilização de criptografia na comunicação. Os pacotes de voz serão criptografados;
- ferramentas de comunicação gratuitas;
- aplicações avançadas;
- simplificação.

Diversos formalismos para descrição de sistemas têm sido desenvolvidos, os quais permitem a avaliação de desempenho [40] . As redes de Petri estocásticas (Stochastic Petri nets - SPNs) [39] , [48] são um desses formalismos. Redes de Petri (RdPs) é um termo genérico associado a um conjunto de formalismos matemáticos adequados para a modelagem de sistemas concorrentes, assíncronos, paralelos e distribuídos. Além da capacidade de representação, as RdPs são suportadas por um vasto arcabouço de mecanismos para análise, verificação de propriedades e validação.

Generalised Stochastic Petri Nets (GSPN) são adequadas para avaliação de desempenho de VoIP. Além disso, conflitos, *buffers,* paralelismo, prioridade, pesos e probabilidades podem ser modelados de forma direta. Os modelos GSPN podem ser avaliados através de técnicas de análise numérica ou por meio de simulação estocástica. Permite também a avaliação transiente, (comportamento do sistema é avaliado a partir do instante inicial até um determinado momento específico) ou estacionária, na qual avaliamos o sistema após os efeitos do comportamento transiente.

A avaliação de desempenho de um sistema através do seu modelo desempenho se baseia na representação das características de desempenho por meio de expressões matemáticas que relacionam elementos do modelo (métricas). A avaliação de desempenho de sistemas através de modelos permite a análise de cenários que podem ser difíceis de se observar no sistema real. É um mecanismo, portanto, extremamente útil para avaliação de novos projetos, configurações, e análise de condições difíceis de serem configuradas em um sistema real e que esteja em operação.

Nesse trabalho serão avaliados cenários auxiliarão identificar gargalos[1] em uma arquitetura de rede.

A internet [8] trabalha com o paradigma do melhor esforço (*best effort*). Esta arquitetura não oferece garantias de qualidade de desempenho. Portanto, a tecnologia VoIP tem como infra-estrutura uma rede que não oferece garantias de qualidade de serviço. Quando os nós de comunicação com a internet não oferecem qualidade de serviço, o algoritmo de escalonamento é baseado em filas FIFO (*First In First Out*), ou seja, os pacotes que chegam primeiro são atendidos primeiro [1] .

Apesar de se basearem em uma rede que não oferece garantias de qualidade de serviço, nos últimos anos as aplicações de VoIP têm se mostrado eficientes e proporcionado diversos avanços. Em muitas organizações, os serviços de VoIP têm grande importância, e sua degradação pode levar a perdas consideráveis.

Devido à convergência das redes de dados com o tráfego de aplicações de voz, tráfegos particulares de aplicações devem ser avaliados e a infra-estrutura deve ser planejada para evitar degradação de desempenho destas aplicações. O tráfego tradicional das aplicações IP é caracterizado por rajadas [11] , assim como as aplicações não são fortemente sensíveis a atrasos e a variações de atraso [11] . Por outro lado, as aplicações de voz são caracterizadas por gerarem tráfego contínuo e constante. Por conseguinte, a comunicação é altamente afetada por longos períodos de atrasos, variações de atraso ou perdas de pacotes [11] .

Existem outros desafios a serem enfrentados tanto em relação à confiabilidade quanto à interoperabilidade destes sistemas [38] . Com relação à confiabilidade, diversos serviços têm sido desenvolvidos para aplicações de VoIP, como por exemplo, desenvolvimento de mecanismos que garantem

[1] O gargalo refere-se ao ponto do sistema onde a quantidade de pacotes recebidos é maior do que transmitidos.

qualidade de serviço, roteamento automático para aplicações de tempo real e gerenciamento de *buffer* [38] . Outro desafio é a interoperabilidade juntamente com a confiabilidade, o baixo custo da utilização das redes baseadas em IP para transmissão de voz, acompanhado pela falta de padrões [12] .

Entretanto, com a caracterização do tráfego de aplicações envolvidas em nosso ambiente e o conhecimento dos limites de nossos recursos, é possível dimensionar o tráfego para que não afete as aplicações de VoIP. Dentro desse contexto, modelos de avaliação de desempenho são mecanismos importantes para diagnóstico e planejamento de infra-estruturas de rede.

O ambiente de avaliação de VoIP é diverso. VoIP pode estar inserido em uma rede de telefonia móvel, em interconexões de rede de longa distância pela internet ou em arquiteturas de redes com qualidade de serviço. Dentro de cada ambiente existem algumas particularidades que precisam ser representadas para que avaliação do seu desempenho seja significativa.

Em uma rede de terceira geração da telefonia móvel [41] , o sinal de transmissão para o equipamento VoIP é um recurso preponderante na avaliação, pois está diretamente relacionado com o desempenho da comunicação de voz [41] .

A avaliação de desempenho de aplicações VoIP sobre conexões de longas distâncias envolve a avaliação sobre os pontos de interconexão, pelos quais os pacotes de voz transitam. Nesse caso, é importante avaliar não só o impacto que a falha de um ponto de interconexão provoca na comunicação, mas também o tempo de perda para transmitir através de um caminho alternativo [5] .

Equipamentos que utilizam o recurso de serviços diferenciados são bem mais caros do que os equipamentos sem esse recurso. VoIP em uma arquitetura com serviços diferenciados proporciona melhorias de desempenho que devem ser analisadas, pois existem algumas considerações, além do custo, nessa arquitetura [20] .

Explicaremos os objetivos de cada trabalho mencionado na seção de trabalhos relacionados. Independente do cenário proposto, o modelo de avaliação apresentado neste trabalho pode ser utilizado para avaliar o desempenho do tráfego de aplicações VoIP.

1.1 Objetivo

Para avaliar o desempenho da aplicação de VoIP em uma rede convergente [1] este trabalho tem por objetivo prover uma maneira para identificar os pontos críticos de tráfego que afetam a comunicação de VoIP. Dessa forma este trabalho se propõe a criar um modelo de avaliação de desempenho para análise de tráfego de voz em uma rede *ethernet*, e desse modo, almeja prover um meio para planejamento de capacidade.

Para o desenvolvimento do modelo de avaliação, defini-se uma metodologia que inclui as seguintes atividades:

- Definição do problema e dos componentes. Descrever a infra-estrutura a ser avaliada.
- Medição. Medir a carga de tráfego transmitida de cada componente envolvido na infra-estrutura. Todas as etapas de medição, coleta dos dados e ferramentas utilizadas são descritas nessa atividade.
- Geração do modelo abstrato.
- Analisar e validar as propriedades do modelo abstrato.
- Geração do modelo refinado e mapeamento das métricas. Incluir as medidas nos respectivos componentes do modelo abstrato. Inserir as métricas a serem avaliadas no modelo.
- Validar o modelo refinado.
- Avaliar o desempenho da aplicação de VoIP através do modelo refinado.

- Interpretação dos resultados.

De uma forma geral, os cenários do sistema são avaliados através de modelos GSPN [33] , [2] . Parâmetros inseridos no modelo foram obtidos através de medições em uma plataforma de sistema real ou são parâmetros cujas variações pretendemos avaliar o impacto no desempenho do sistema.

Neste trabalho, identificamos o limite de tráfego que provoca degradação de desempenho. Também avaliamos a aplicação de VoIP com outros tipos de tráfego para obtenção de um modelo que permite a avaliação de cenários que representam a arquitetura da rede. Assim, os dados obtidos através da avaliação do modelo nos permite dimensionar o tráfego de dados e das aplicações de VoIP que são suportados em um determinada infra-estrutura de rede IP.

1.2 Estrutura da Dissertação

Esta dissertação está organizada da seguinte forma: o Capítulo 2 apresenta os trabalhos relacionados com avaliação de desempenho de VoIP. O Capítulo 3 é dividido em duas partes. Na primeira parte é descrita a tecnologia de VoIP, destacando as suas características e o seu funcionamento. Na segunda parte são apresentadas, de forma breve, as Redes de Petri estocásticas, bem como aspectos relacionados à modelagem, técnicas de análise e validação. O Capítulo 4 descreve a metodologia proposta para avaliação de desempenho e descreve suas atividades. O Capítulo 5 descreve o modelo de desempenho de VoIP concebido e descreve seus componentes GSPN. O Capítulo 6 apresenta um estudo de caso baseado no modelo proposto e adotando a metodologia concebida para avaliar o de desempenho destes sistemas. Nesse capítulo também são apresentados os resultados da avaliação do estudo de caso. O Capítulo 7 apresenta as conclusões obtidas durante o desenvolvimento desta dissertação como também, as principais contribuições

do trabalho. Por fim, são apresentados trabalhos futuros que darão continuidade ao estudo desenvolvido.

Capítulo 2 - Trabalhos Relacionados

Neste capítulo descreve-se sobre trabalhos relacionados à avaliação de desempenho de voz sobre IP.

A primeira geração de telefonia móvel iniciou-se com a tecnologia analógica [19] . Posteriormente, vieram as redes de segunda geração (2G) que são chamadas de Global System for Mobile Communications (GSM) [19] . GSM é uma tecnologia no qual o sinal e os canais de voz são digitais. A terceira geração (3G) oferece serviços de dados por pacotes na rede de telefonia [19] .

Não iremos estender sobre a tecnologia 3G porque não é o foco deste trabalho, mas apenas descrever alguns pontos importantes para entendimento dos trabalhos mencionados abaixo.

General Packet Radio Service (GPRS) [44] é a nova geração das redes GSM *(Global System for Mobile communication)* [37] que oferece transporte IP. A tecnologia 3G tem o objetivo de fornecer os serviços de telefonia por voz e a transmissão de dados a longas distâncias com mobilidade. A informação a ser transmitida é dividida em pacotes e os mesmos são relacionados entre si antes de serem transmitidos e remontados no destinatário. Os recursos de rádio utilizados na rede GPRS serão utilizados apenas quando os usuários estiverem enviando ou recebendo dados. O recurso pode ser compartilhado concorrentemente entre vários usuários, ou seja, o número de usuários conectados concorrentemente depende da aplicação em uso e de quanta informação está sendo transferida.

O trabalho de Rajavelsamy envolve a avaliação de desempenho de VoIP sobre uma rede 3G-WLAN (*Wireless Local Area Network)* [41] . A Figura 2.1 [19] representa a integração da rede 3G-WLAN. Duas ferramentas foram usadas para avaliar o desempenho das aplicações de voz sobre IP, *Netperf: (http://www.netperf.org/netperf/)* e *pktstat (http://www.adaptive-enterprises.com.au/~d/software/pktstat/). Netperf* foi usado para medir o

throughput das aplicações de voz sobre IP e a latência fim-a-fim. A *pktstat* foi utilizada para identificar a largura de banda. O trabalho apresenta a avaliação do impacto das aplicações de VoIP sobre uma rede 3G-WLAN criptografada. Ele demonstra que o uso do túnel IPSec [46] aumenta o atraso e a largura de banda do tráfego de VoIP. Existe um aumento significativo no tamanho do pacote com a utilização do IPSec, o qual aumenta a largura de banda requerida para o tráfego de voz. Esse aumento corresponde a 19,62% no codec G.711, por exemplo.

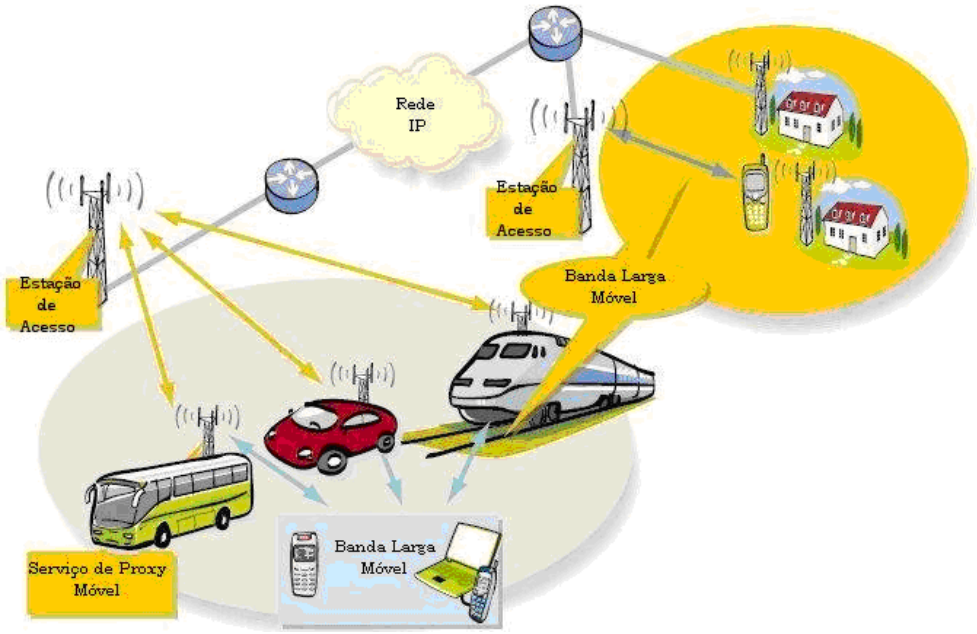


Figura 2.1 Rede 3G-WLAN

A variabilidade no atraso é um dos problemas a ser enfrentado e, uma observação importante na avaliação desse trabalho, é que não houve diferença no intervalo do tempo de chegada dos pacotes, com o uso ou não da criptografia, o que poderia afetar seriamente o desempenho das aplicações de VoIP.

VoIP sobre GPRS permite detectar o silêncio na comunicação e, nesse caso o tráfego de voz pode ser multiplexado para uma largura de banda menor, aumentando conseqüentemente a capacidade do serviço de voz. Os impactos sobre a rede GPRS estão no atraso e, principalmente, na variabilidade desse. Outro ponto analisado no trabalho mostra que o *throughput* para o tráfego de dados normal foi de 6.1 Mbps, mas, com o uso do

IPSec a taxa de transmissão de dados diminuiu para 16%. A taxa de transmissão para 802.11b é de fato 11 Mbps, mas devido ao CSMA/CA, protocolo de acesso ao meio, a taxa medida foi de 6.1 Mbps.

O desempenho das aplicações de voz na rede WLAN sem criptografia, demonstrou que o desempenho dos equipamentos dos clientes não foi afetado com o aumento do número de conexões VoIP. Porém o atraso fim-a-fim aumenta para um nível inaceitável de comunicação. Isso acontece porque o AP (*Acess Point*) 802.11b não consegue transmitir os dados na mesma quantidade que recebe, tornando-se um gargalo. A medição realizada no experimento mostrou que o número máximo de conexões permitidas para o codec G.711 foi de 28 conexões simultâneas.

Avaliações dos desempenhos das aplicações de voz em uma rede 3G-WLAN com mobilidade nas aplicações dos clientes, ou seja, os clientes se comunicando por diferentes WLAN AN (*Acess Network*), demonstram um aumento significativo no atraso e na qualidade da comunicação. O atraso maior acontece quando o cliente precisa desconectar-se do antigo ponto de acesso e automaticamente se conectar ao novo (atraso médio de 146 ms). Nesse caso a qualidade da comunicação é afetada, chegando a ter algumas falhas de comunicação.

Alguns trabalhos indicam que no futuro a tecnologia de VoIP será em uma rede fundamentalmente baseada em IP móvel [44] . Partindo desse pressuposto, [44] avalia o desempenho de VoIP sobre uma rede GPRS. Em uma rede GPRS o canal de acesso e o canal de tráfego são separados. As estações móveis enviam pedidos para acessar os canais. Se a requisição é aceita, então o canal de tráfego é garantido para ser usado pelas estações móveis. Em [44] o canal de acesso e o canal do tráfego de dados são avaliados.

A avaliação feita por [44] mostrou que VoIP baseado em GPRS provê transmissão de melhor qualidade do que as redes de voz baseadas em circuito compartilhado. O máximo número de conexões concorrentes de voz que uma freqüência GPRS pode suportar é de 34 e 12 conexões para taxas de codificação de 5.3 Kbps e 13 Kbps, respectivamente. Usando a taxa de codificação de 13 Kbps no sistema de circuito compartilhado, o número

máximo de conexões concorrentes caiu para 7 conexões. A medição foi feita no canal de tráfego de dados porque a utilização efetiva do canal de acesso é muito baixa, o que não afeta a comunicação de voz.

Alguns trabalhos mencionam que as três maiores causas de degradação de desempenho dos serviços de voz na internet enfrentadas pelos provedores de serviços são: congestionamento da rede, falha de conexão e instabilidade de roteamento [5] .

Avaliando o impacto sobre uma conexão em um *backbone* (concentrador principal para redes menores), quando a falha de conexão é detectada os pacotes são automaticamente encaminhados para um novo caminho alternativo. A degradação ocorrida na qualidade da voz é mínima quando ocorre essa mudança de roteamento. O atraso médio alcançado para mudança de roteamento foi de 100 ms e um pequeno jitter (variação de atraso) de 500 μs [5] . A degradação é mínima porque para atingir um bom nível de interatividade na conversação, o atraso médio não pode ser superior a 150 ms [11] . Se as falhas de conexões acontecerem em seqüência de tempo inferiores a 1 min, os pacotes de voz não são *bufferizados* pelos roteadores. Eles são simplesmente desprezados (*dropped*) porque são imediatamente encaminhamos por caminhos inválidos. A degradação na qualidade da voz é perceptível e a indicação de pacotes desprezados no equipamento de avaliação não é devido a eventos de congestionamentos, mas sim a algum tipo de falha de roteamento [5] .

O trabalho [13] avalia o impacto da perda dos pacotes de voz. Nesse trabalho, ele demonstra que o codec com baixa taxa de compressão apresenta melhor eficiência da utilização da largura de banda, ou seja, proporciona uma maior taxa de transmissão de informação. O trabalho mostra que G.711 tem um melhor desempenho que o G.729. Se os quadros de voz são perdidos durante os períodos silêncio, não se têm impacto na aplicação de voz. Esse trabalho usa o algoritmo de *Discontinuous Transmission* (DTX) para indicar os quadros que não são importantes. O DTX interrompe o fluxo constante de quadros até que os novos quadros contenham conteúdo de áudio.

O pacote de voz é formado por vários quadros e a perda de um quadro não tem relevância na qualidade da voz, mas sim uma grande quantidade de quadros perdidos [13] . Portanto, se um pacote de voz é perdido, um ou muitos quadros também são perdidos. Quando o tamanho da perda chega a 10 ms, o impacto na qualidade da voz não é verificado nos codec G.711 e G.729 [13] . Considerando a perda de dois pacotes de 20 ms do codec G.711, conclui-se que é melhor do que perder um pacote de 40 ms. Porém, a perda de dois pacotes de 40 ms tem um impacto muito maior do que a perda de um pacote de 80 ms [13] .

Avaliar o desempenho das aplicações de voz em uma arquitetura de redes com serviços diferenciados [20] exige que as diversas interligações da rede sejam implementadas com esse recurso (serviços diferenciados). O recurso marca o pacote de voz como prioritário e este passa a ser encaminhado com maior precedência sobre os outros pacotes. A análise feita por [20] mediu o atraso médio dos pacotes de voz em filas de roteamento com melhor esforço e com prioridade sobre diferentes (baixa, média e alta) cargas de tráfego. O cenário de avaliação consistiu na utilização de recursos (componentes de rede) separados por uma conexão de 2 mb. Nesse contexto, o tráfego baixo, médio e alto corresponde à utilização dessa conexão, ou seja, o tráfego baixo utiliza 50% da conexão, o médio 90% e o alto 100%. A avaliação consistiu na verificação do tempo médio de chegada dos pacotes de voz. A avaliação feita com tráfego baixo mostrou que o atraso médio dos pacotes foram os mesmos. O atraso com fila de melhor esforço foi bem superior ao de fila com prioridade utilizando o tráfego médio. Utilizando uma carga alta, o atraso da fila de melhor esforço chega a ser o dobro da apresentada com prioridade. A Tabela 2.1 demonstra o atraso médio para os pacotes de voz sobre as filas de roteamento [20] .

Tabela 2.1: Atraso médio para pacotes de voz (ms)

Fila\Carga	Baixa	Média	Alta
FIFO	35,7	53,3	67,0
Prioridade	32,8	33,6	33,2

A maioria das estruturas das redes de computadores das empresas já está implantada, então como avaliar o desempenho das aplicações de voz com as outras aplicações existentes?. Existe um custo elevado para avaliarmos VoIP em um sistema complexo, como por exemplo avaliar o desempenho de VoIP sobre conexões de longa distância ou sobre redes 3G. As avaliações feitas pelos trabalhos apresentados não são ricas em detalhes; existe, porém, uma série de recursos, falhas e gargalos que precisam ser identificados quando avaliamos um tráfego de voz porque os mesmos influenciam o desempenho da comunicação. Portanto, este trabalho tem o objetivo de criar um modelo de desempenho, que contemplam os detalhes dos componentes, para avaliação de desempenho de voz sobre ip da qual poderemos obter respostas rápidas, precisas e com um custo relativamente baixo.

Capítulo 3 - Fundamentos

Este capítulo está dividido em três partes. A primeira parte refere-se à tecnologia de voz sobre IP, na qual se descreve o processo de formação e encaminhamento da voz que inclui a parte de sinalização, transmissão, codificação e transporte. A segunda parte é sobre Redes de Petri, em que detalharemos as propriedades comportamentais e estruturais dos modelos, citando os métodos de análise e especificando GSPN. A terceira descreve uma breve visão sobre a teoria das filas.

3.1 Voz sobre IP

Durante mais de um século é transmitida voz humana à distância com uma qualidade razoavelmente inteligível. Com o passar dos anos houve um aperfeiçoamento das técnicas e dos meios de comunicação e uma série de pontos fortes como uma padronização estabelecida, transparência na interoperabilidade entre grande parte de seus elementos de hardware e software, estabilidade e aceitação. Essa rede é conhecida atualmente como Rede Pública de Telefonia Comutada (PSTN – *Public Switch Telephone Network*) [1] .

As redes PSTNs sofrem de algumas limitações. Uma de suas limitações é que elas não foram originalmente projetadas para transportar dados de forma eficaz. Outra limitação é o desperdício de recursos de rede, já que a largura de banda fica reservada durante toda a duração da chamada em curso. Em contrapartida, uma rede de transmissão de dados baseada em pacotes consegue usar de forma mais otimizada a largura de banda disponível [12] .

A comunicação de dados cresceu com grande intensidade, a internet evoluiu e, conjuntamente, os acessos a recursos disponíveis na internet e a respectiva comunicação também se intensificaram.

VoIP não é uma tecnologia que surgiu para competir com PSTN. A tecnologia de voz sobre IP pode facilitar tarefas e serviços que podem ser mais difíceis de executar ou mais caros do que PSTN. O tráfego de voz em tempo real pode ser utilizado sobre uma rede IP de várias formas, conforme vista na Figura 3.1.

Na comunicação de voz entre computadores, os usuários se comunicam através de softwares de aplicações multimídia. Para que seja possível a interligação das redes telefônicas convencionais com o VoIP, usa-se um equipamento denominado *Gateway*. O *Gateway* é responsável pela conversão do sinal analógico em digital (e vice-versa), além de executar os sinais de controle necessários para implementação das chamadas telefônicas. O Gateway Controller (ou Call Agent) é o responsável pelo controle das chamadas feitas pelo Gateway. O controle se dá pelo estabelecimento, supervisão e liberação das chamadas que trafegam pela rede IP [45] .

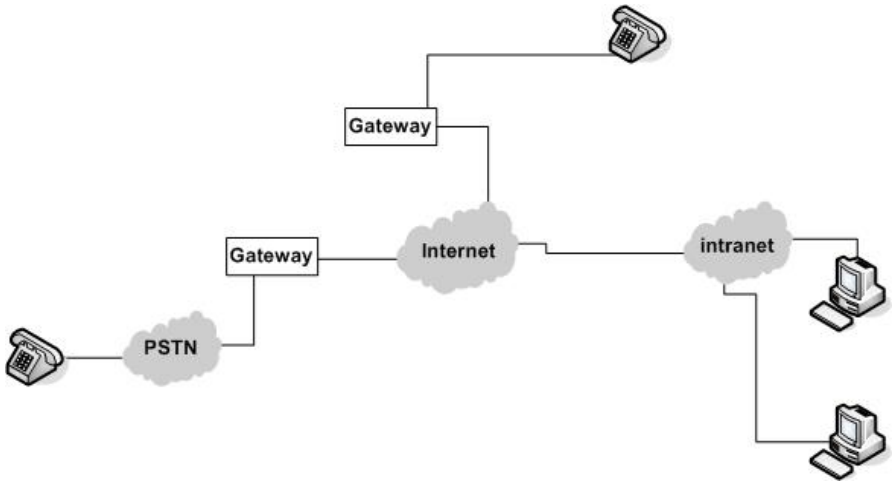


Figura 3.1 Infra-estrutura de VoIP

Essa tecnologia proporciona alguns serviços avançados, tais quais:

1. Integração de voz, dados e fax. A integração de dados inclui outros serviços (por exemplo, troca de arquivos) disponíveis na internet.

2. PBX (*Private Branch Exchange*) Remoto. Através do *gateway* de VoIP os usuários podem acessar remotamente o PBX da empresa para realizar e receber chamadas.

2. Ligação entre empresas. Uma ligação entre companhias remotas, principalmente entre países, pode diminuir consideravelmente os custos de ligações.

3. *Call Centers*. Os serviços de suporte e atendimento de voz pela internet.

4. Independência de localização. Somente uma conexão com a internet é necessária para conseguir conectar com um provedor de voz.

3.1.1 Sinalização

Um protocolo de sinalização para VoIP deve especificar a codificação da voz, a configuração das chamadas, o transporte de dados, o modo de autenticação, segurança, métodos utilizados na comunicação, cabeçalho, endereçamento, sintaxe da mensagem [15] . Sinalizar significa que a informação da chamada é carregada através dos limites da rede.

Com o propósito de apresentar de forma mais concreta a sinalização em voz sobre IP, serão apresentados os protocolos mais comuns: o H.323 e o SIP [15] .

3.1.1.1 H.323

O H.323 [21] é um padrão que se constitui de recomendações de procedimentos, protocolos, equipamentos e serviços que possibilitam o tráfego de aplicações de tempo real de áudio, vídeo e conferências de dados sobre redes, como a internet. O H.323 foi projetado para ser usado em cima da camada de transporte da pilha de protocolos do sistema de rede, podendo, portanto ser utilizado em cima dos protocolos UDP (*User Datagram Protocol*) ou TCP (*Transport Control Protocol*).

O padrão H.323 define quatro tipos de componentes que, juntos possibilitam a comunicação multimídia: *gatekeepers*, *gateways,* terminais *e Multipoint Control Units (MCUs)* [38] .

- *Gatekeeper* é um equipamento opcional que fornece um serviço de controle (admissão, controle e registro) de chamada para os terminais. Quando se tem um *gatekeeper* no sistema, todos os terminais devem se registrar no mesmo. Os principais serviços oferecidos pelo *gatekeeper* são usados para: autorizar e/ou intermediar a sinalização das sessões de áudio; traduzir nomes (alias) para endereços de transporte; controlar o número de terminais H.323 que podem ter acesso simultâneo à rede.

- *Gateways* permite que sistemas finais de redes diferentes se comuniquem. Por exemplo, permite que um sistema final em uma rede H.323 se comunique com um usuário da rede PSTN.

- Terminais são os sistemas finais. O padrão declara que todos os terminais H.323 devem obrigatoriamente suportar o serviço de voz, enquanto serviços de vídeo e dados são opcionais. Exemplos são telefones IP (hardphones) e computadores executando software de voz (softphones).

- *Multipoint Control Units* (MCUs): um MCU consiste de um *Multipoint Controller* (MC) e zero ou mais *Multipoint Processors* (MP). O MC

manipula as negociações entre todos os terminais para determinar capacidades comuns para processamento de áudio e vídeo. Já o MP é o responsável por mesclar, chavear e processar os bits de áudio, vídeo e/ou dados.

O padrão H.323 é completamente independente dos aspectos relacionados à arquitetura da rede. Dessa forma, podem ser utilizadas quaisquer tecnologias de enlace, como Ethernet, Fast Ethernet, FDDI, ou Token Ring. A adoção do padrão H.323 para aplicações multimídia em redes traz uma série de benefícios, entre os quais podemos citar [6] [9] [50] :

- Interoperabilidade de equipamentos e aplicações: o H.323 permite interoperabilidade entre dispositivos e aplicações de diferentes fabricantes.
- Independência de plataforma: o H.323 não determina o hardware ou sistema operacional a ser usado.
- Representação padronizada de mídia: o H.323 estabelece codificadores para compressão e descompressão de sinais de áudio e vídeo.
- Flexibilidade nas aplicações clientes: uma conferência H.323 pode envolver aplicações clientes com capacitações multimídia diferentes.
- Interoperabilidade entre redes: além da independência da rede citada anteriormente, é possível estabelecer conferências entre participantes localizados numa intranet e em outras redes completamente diferentes, como a rede telefônica pública ou ISDN.
- Suporte a gerenciamento de largura de banda: o padrão provê mecanismos de gerenciamento que permitem delimitar a quantidade de conferências simultâneas e a quantidade de largura de banda destinada às aplicações H.323.
- Suporte a conferências multiponto: o H.323 suporta conferências com três ou mais participantes simultâneos.

- Suporte a *multicast:* o H.323 suporta técnicas de *multicast* nas conferências multiponto. Uma mensagem *multicast* envia um único pacote a todo um subconjunto de destinatários na rede sem replicação.

Na Figura 3.2 representamos a pilha de protocolos H.323. Anteriormente, citamos que o H.323 faz uso de alguns protocolos. Os principais são o H.245 e o H.225.0. O H.245 *control signalling* é usado para gerenciar o fluxo de mídia, negociação dos codificadores de áudio e das portas de comunicação que serão usados na sessão de áudio. O H.225.0 tem o objetivo de definir as mensagens trocadas pelo H.323. As mensagens podem ser de sinalização (*call singnaling*), usadas para estabelecimento, controle e término de uma chamada H.323, ou podem ter a função de sinalização RAS (*Registration, Admission and Status*), usada na comunicação entre um terminal e um *gatekeeper*.

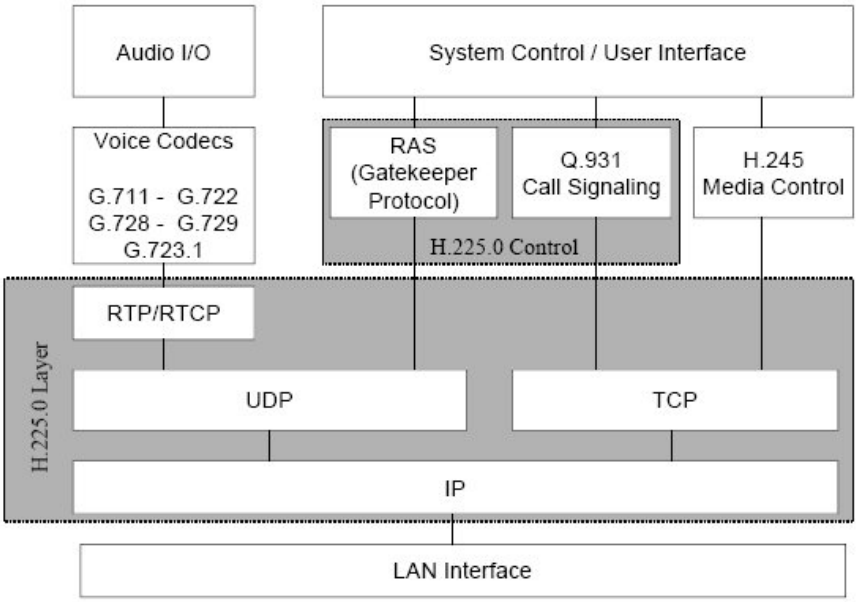


Figura 3.2 Pilha de protocolos H.323

3.1.1.2 SIP

O Protocolo de Inicialização de Sessão (SIP) é um padrão da internet, definido pelo IETF (*Internet Engineering Task Force*) como um protocolo de sinalização que trabalha na camada de aplicação e tem a função de criar, modificar e terminar sessões com um ou mais participantes [28] . O funcionamento do SIP é similar ao HTTP, no qual as requisições são geradas pelo cliente e enviadas ao servidor. O servidor, por sua vez, processa as requisições e envia a resposta de volta para o cliente. Cada requisição e resposta constitue uma transação. O processo de abrir um canal confiável nos quais as mensagens de controle de chamadas são passadas é feito por mensagens do tipo *INVITE* e *ACK*. O SIP define um conjunto de mensagens que são usadas na comunicação entre o cliente e o servidor:

INVITE: pedido de início de sessão.

BYE: para terminar uma conexão entre dois usuários.

ACK: confirmação de início de sessão.

OPTION: para obter informações sobre as capacidades de uma chamada.

REGISTER: para fornecer informações sobre a localização de um usuário ao servidor de registro.

CANCEL: cancelamento de pedido pendente.

O SIP trabalha independente de qualquer protocolo da camada de transporte. O SIP depende de outros protocolos para fornecer o serviço para os usuários. Um dos protocolos é o SDP (*Session Description Protocol*) usado para conduzir a negociação para identificação do codificador. O SDP faz a descrição do conteúdo das sessões multimídia, identificando o codec e as portas IP que serão usadas, por exemplo.

Os serviços oferecidos pelo SIP são:

- Localização do usuário: determinação do sistema final a ser usado na comunicação.

- Estabelecimento da chamada: estabelecimento dos parâmetros da chamada de ambas as partes.

- Disponibilidade do usuário: determinação da concordância da parte chamada de se juntar na comunicação.

- Capacidade do usuário: determinação da mídia e dos seus parâmetros.

- Gerenciamento da chamada: transferência e término das chamadas.

O SIP consiste de dois componentes, agentes de usuário e servidores de rede. O agente de usuário é composto por um UAC (*User Agent Client*) e um UAS (*User Agent Server*). Um UAC é uma entidade lógica que cria novas requisições. Um UAS é uma entidade lógica que gera uma resposta para uma requisição SIP. Os servidores de rede são representados por três tipos. Um deles refere-se ao servidor de registro que recebe atualizações sobre a localização atual dos usuários. O segundo refere-se ao servidor *proxy* que recebe as requisições e encaminha para outro servidor. O terceiro refere-se a um servidor de redirecionamento que recebe as requisições, determina qual o próximo servidor e retorna o endereço desse servidor ao cliente ao invés de encaminhar a requisição [28] .

3.1.2 Transmissão da voz

A voz humana é uma forma de onda mecânica com freqüências principais na faixa que vai de 300 a 3400 Hz, com alguns padrões de repetição definidos em função do timbre de voz e dos fonemas emitidos durante a conversação [15] . Freqüências menores correspondem a sons graves, enquanto freqüências altas a sons agudos. Para a comunicação telefônica foi estabelecido que o espectro de freqüência até 3,4 kHz seria suficiente para boa

conversação. Representamos na Figura 3.3 um emissor enviando um sinal analógico passando pela rede digital.

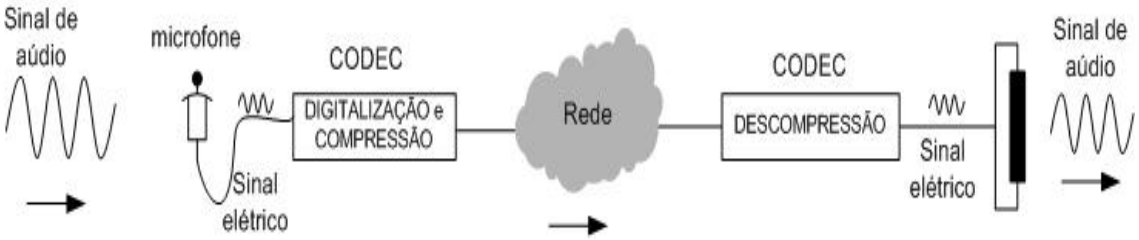


Figura 3.3: Conversão Analógico Digital

Para se digitalizar um sinal analógico são necessárias duas operações: amostragem e quantização. Amostragem é o processo de medir instantaneamente valores de um sinal analógico em intervalos regulares, ou melhor, para digitalizar um sinal mede-se a amplitude do sinal analógico em intervalos e a esse processo chamamos de amostragem. O intervalo de freqüências que compreende a diferença entre a maior e a menor freqüência que compõe o sinal é chamado de taxa de amostragem. Para representar fielmente um sinal, a taxa de amostragem deve ser, no mínimo, o dobro da mais alta freqüência presente no sinal [29] . Portanto, a amostragem é dita sem perdas se o sinal original pode ser reconstituído a partir das amostragens. Baseado no Teorema de *Nyquist* [32] , a voz humana requer 8000 amostras por segundo para uma freqüência máxima.

Quantização é a conversão das amostras com valores em escala contínuas para valores discretos. Os valores amostrados (em escala discreta, pois nos referimos a valores obtidos através de dispositivos físicos que têm restrições intrínsecas relativas a capacidade de mensurar sinais) que representam a amplitude do sinal no momento da amostragem são quantizados em um número discreto de níveis. O valor depende do número de bits do conversor analógico/digital [29] . Para um sinal de áudio com qualidade de chamada telefônica, considera-se adequado utilizar entre 8 a 16 bits, que correspondem a 256 e 65.536 níveis.

Os valores quantizados são codificados em seqüência de bits. Explicaremos na próxima seção o processo de codificação do sinal de voz.

3.1.3 Codificação

O codificador de voz, também chamado de codec (codificador-decodificador), é um mecanismo capaz de codificar e/ou decodificar um sinal analógico em um formato digital. A técnica de codificação é utilizada basicamente para reduzir o número de bits que representa uma informação e assim reduzir o uso da banda passante do canal necessário para transmissão da voz digitalizada. Todo esse processo deve obedecer às restrições temporais para possibilitar a comunicação e interação [32] . A compressão de sinais é baseada em técnicas de processamento que retiram informações redundantes, previsíveis ou inúteis. A compressão pode acontecer com ou sem perda de informação. Tudo depende da degradação que se admite para o sinal e do fator de compressão que se deseja atingir. Basicamente duas formas podem ser usadas para a codificação da voz humana:

- codificação da forma de onda: codifica diretamente a forma de onda gerada pelo sinal analógico de voz, através da amostragem, convertendo a amplitude de cada amostra para o valor mais próximo de um conjunto finito de valores discretos. Dentro dessa categoria se encontram os métodos mais simples de codificação de voz.

- codificação da fonte: esse método usa a entrada para produzir um sinal que remonta a fala original. No processo de codificação, o sinal original é mapeado em um modelo matemático de como o som é reproduzido na traquéia e na decodificação utiliza sintetizadores para reproduzir o som. A codificação da fonte permite uma redução considerável na taxa de transmissão, porém são perdidos fatores essenciais como timbre de voz, tornando-a mais impessoal. Dessa forma, sistemas de codificação da fonte são utilizados quando é necessária uma baixa taxa de transmissão.

O ITU (*International Telecommunication Union*) [16] , o TIA (*Telecommunication Industries Associtation*) [17] e o USFS (*United States Federal Standards*) [18] são os principais órgãos internacionais que padronizam os algoritmos de compressão de voz. Alguns critérios devem ser atendidos para serem aprovados, como [36] :

- recuperação do ruído do ambiente;
- menor degradação da qualidade de voz após vários processos sucessivos de codificação/decodificação;
- habilidade para facilmente codificar, para outros padrões de diferentes codificadores do ITU, os sinais que já foram anteriormente codificados; e
- qualidade satisfatória, mesmo depois da perda de quadros.

Diversos codificadores atendem esses requisitos. Na forma de onda temos o ITU G.711 [22] , o ITU G.726 [23] e o ITU G.722 [24] . Na fonte temos o ITU G.723 [25] , o ITU G.728 [26] e o ITU G.729 [27] .

O padrão G.711 do ITU-T, conhecido como PCM (*Pulse Code Modulation*), é um padrão de codificação de voz baseado na forma de onda e para a digitalização da voz esse é o padrão mais usado. Temos 8000 amostras /segundo e cada amostra codificada por uma seqüência de 8 bits, isto é, cada amostra pode ter 1 entre 256 valores possíveis [29] . Esse tipo de codificação necessita de um canal com banda de 64 Kbps para transmissão do sinal digitalizado, já que são geradas 8000 amostras de 8 bits por segundo.

A qualidade de sistemas de transmissão de voz é medida pelo *Mean Opinion Score (MOS) [51]* . O MOS é uma medida intuitiva, derivada do método ACR (*Absolute Category Rating*), vastamente usada para comparar a qualidade da transmissão da voz [51] . O valor do MOS varia entre 1 (ruim) e 5 (excelente) e dessa forma os ouvintes julgam a qualidade da voz. Uma qualidade excelente implica que a fala codificada é indistinguível da original e sem ruído perceptível. Por outro lado, uma má qualidade (inaceitável) implica na presença de um ruído extremamente incômodo e uma característica

artificial na fala codificada. Testes com o G.711 foram feitos e o mesmo possui o *MOS* de 4.1, enquanto que o G.729 possui 3.92 e o G.726, 3.85 [51] .

3.1.4 Transporte

O TCP (*Transmission Control Protocol*) e o UDP (*User Datagram Protocol*) são os principais protocolos que fazem parte da internet [1] .

O TCP [1] está presente na maioria das comunicações de pacotes, mas não suporta transmissão de voz em tempo real porque utiliza um mecanismo de recuperação dos dados perdidos por retransmissão. Nesse caso, a perda de um pacote leva a aplicação esperar por todas as retransmissões, acarretando atrasos inaceitáveis.

O UDP [1] é um protocolo sem conexão no qual os pacotes podem ser entregues fora de ordem ou sem garantias de que chegarão ao destino. Os quadros processados pelo *codec* são transmitidos em pacotes. Esses pacotes são transmitidos sobre o protocolo UDP. Portanto, para aplicações de VoIP o serviço de entrega de pacotes fornecido pelo UDP não é suficiente. É necessário saber a ordem e o tempo de geração dos pacotes, além de identificar a qualidade da conexão.

O RTP (*Real Time Protocol*) [14] é o principal protocolo usado para aplicações de VoIP. RTP foi criado para carregar informações em tempo real fim a fim, como áudio e vídeo. O transporte efetuado pelo RTP pode ser através de unicast ou multicast. O RTP é executado sobre o UDP e se propõe a facilitar a entrega, monitoração, reconstrução e a sincronização de fluxos de dados em tempo real.

O RTP não fornece qualquer mecanismo para garantir a entrega no tempo certo ou qualidade de serviço. O cabeçalho do RTP, como mostra a Figura 3.4, fornece o número de seqüência usado pelo receptor para reconstruir a seqüência de pacotes enviados pelo emissor. Esse número de

seqüência também pode ser usado para determinar a localização de um pacote. O *timestamp* fornece o tempo que o pacote de voz foi gerado.

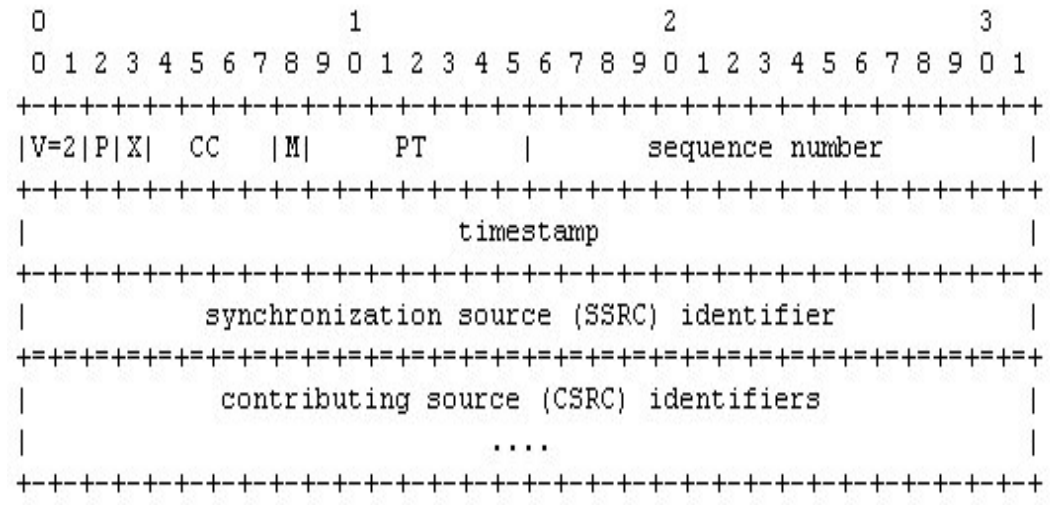


Figura 3.4: Cabeçalho RTP

O RTP faz uso de um protocolo de controle RTCP (*RTP control protocol*) [7] para monitorar a qualidade do serviço oferecido pelo RTP e para carregar informações sobre os participantes de uma sessão. Cada pacote RTCP contém um número de elementos, geralmente um relatório do transmissor (SR) ou um relatório do receptor (RR) seguido de descrição de fontes (SDES).

Os relatórios do transmissor descrevem a quantidade de dados enviados até o momento, bem como correlacionam o *timestamp* do RTP com o tempo absoluto para permitir a sincronização em diferentes mídias. Os relatórios do receptor são enviados pelos participantes da sessão RTP que estão recebendo os pacotes de voz sobre IP. Cada bloco contendo o relatório descreve a taxa de perda e o *jitter.*

A descrição de fontes são pacotes usados para controle de sessão. Contêm o CNAME (*Canonical Name*), um identificador único global similar em formato a um endereço de correio eletrônico. As aplicações clientes podem mostrar as informações de nome e email na interface do usuário. Isso possibilita aos participantes da sessão saber mais sobre os outros participantes.

3.1.5 Principais Desafios

Desde que a voz é pronunciada pela origem até chegar ao destino existem alguns desafios como o atraso, a variação do atraso e a perda de pacotes.

- Atraso

O atraso ocorrido em uma comunicação de voz sobre IP é gerado por uma série de pequenos atrasos. Esses pequenos atrasos são descritos da seguinte forma:

Atraso de codificação e decodificação

A qualidade da voz diminui quando aumenta a taxa de compressão. Isso ocorre porque quando aumenta a taxa de compressão aumenta também o atraso. O retardo do processamento e o retardo do quadro são os fatores que geram o atraso de codificação. O atraso de decodificação é tipicamente metade do atraso de codificação na origem [31] .

O retardo de processamento é o atraso para processar um único quadro de voz. As amostras de voz são analisadas quadro a quadro, portanto o retardo do quadro acontece quando o algoritmo de codificação analisa o quadro e o próximo quadro com o intuito de realizar uma correlação entre os quadros adjacentes para diminuir a taxa de transmissão.

Tabela 3.1: Atrasos de Codificação e Decodificação

CODEC	**G.711**	**G.729**	**G.723.1**
Taxa de Bit (Kbps)	64	8	6.3
Tamanho da Amostra (Bytes)	80	10	24
Retardo do Quadro (ms)	0	5	7,5
Retardo de Processamento (ms)	10	10	30
Atraso Total Codificação (ms)	10	15	37,5
Atraso Decodificação (ms)	5	7,5	18,75

Atraso de Empacotamento

Anteriormente vimos que os quadros processados pelo *codec* são transmitidos em pacotes. Um pacote de dados VoIP inicia com um cabeçalho IP, UDP e RTP, dando um total de 40 bytes [31] . Após o cabeçalho, estão os quadros de voz codificados (ver Figura 3.5).

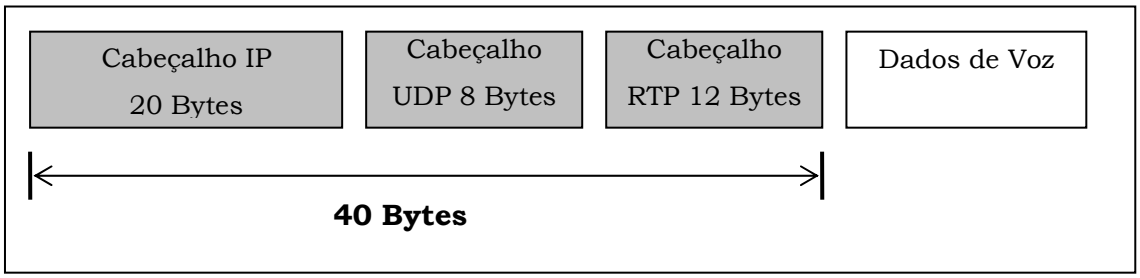


Figura 3.5: Cabeçalho do pacote de voz

O codec G.723.1 para transmissão dos pacotes de voz tem 40 bytes do cabeçalho mais 24 bytes de voz. O tempo para empacotamento é de 30 ms para cada 24 bytes, o que representa uma eficiência de 37,5%. O G.729 possui um tempo de empacotamento de 20 ms para cada 20 bytes, o que representa uma eficiência de 33,33 % considerando 40 bytes do cabeçalho mais 20 bytes de voz. Da mesma forma, o G.711 possui um tempo de 20 ms para cada 160 bytes de voz, o que representa uma eficiência de 80%.

Atrasos em filas

A comunicação de voz sobre IP passa por diversos equipamentos (roteadores, switches, gateways) de rede, ou seja, passa por diversos pontos de rede até chegar ao seu destino. Esses equipamentos possuem uma fila na qual os pacotes de voz têm que esperar para serem transmitidos na rede. Quando não existe priorização de tráfego, a política é: o primeiro pacote que chega será o primeiro a sair. Dessa forma, os pacotes recebem atrasos adicionais nas filas dos equipamentos e, até podem ser descartados, caso o mesmo não tenha condições de atender à demanda de tráfego.

- Variação do atraso

A variação do atraso é conhecida como *jitter* [30] . Essa variação refere-se ao intervalo de tempo entre chegadas de pacotes no destino. Se esse intervalo ultrapassa 25 ms, acontece o que chamamos de eco. A maioria dos aplicativos de voz sobre IP utiliza o cancelamento de eco. Os pacotes chegam com diferentes atrasos e, dessa forma, não podem ser enviados diretamente ao usuário de destino, pois a voz sofreria cortes maiores em função de variação de atrasos. Como a variação de atraso não tem um tempo constante, então se faz necessário utilizar um *buffer* no destino para armazenamento temporário dos pacotes [30] .

- Perda de Pacotes

A demanda de usuários da internet aumenta consideravelmente, assim como a possibilidade de ocorrer congestionamento na rede, resultando em perda de pacotes. A qualidade da voz pode ser altamente afetada com a perda de pacotes.

Em uma comunicação de voz sobre IP, um ou muitos quadros de voz são colocados juntos em um pacote. Portanto se um pacote é perdido, um ou muitos quadros de voz são perdidos. Conforme citado anteriormente, a

importância dessa perda dependerá do tamanho do pacote [14] . Os codificadores também possuem esquemas de recuperação de perdas que variam entre 3% e 5%. O G.723.1, por exemplo, usa as características do quadro anterior para compensar o quadro perdido.

3.2 Redes de Petri

Redes de Petri (RdPs) é um termo genérico que se refere a uma família de técnicas para especificação formal de sistemas que possibilita uma representação matemática e possui mecanismos de análise que permitem a verificação de propriedades e validação do sistema especificado [39] , [48] .

O conceito de RdPs foi apresentado por Carl Adam Petri, em 1962, na sua tese de doutorado intitulada Kommunikation mit automaten (comunicação com autômatos) [4] . Redes de Petri (RdPs) são ferramentas gráficas para descrição formal de sistemas caracterizados pela concorrência, sincronização, distribuídos, assíncronos, não-determinísticos e ou estocásticos.

As vantagens da utilização das RdPs na modelagem de sistemas são conhecidas [39] , entre elas podemos destacar:

- RdPs têm capacidade de representar sistemas concorrentes, sincronização de processos, paradigmas de comunicação e compartilhamento de recursos;
- RdPs fornecem um formalismo de modelagem que permite uma representação gráfica e é fundamentado matematicamente;
- existe uma grande variedade de ferramentas para o projeto e análise, além de ferramentas de *software* desenvolvidas para suportar essas atividades;

- RdPs provêem mecanismos para abstração e refinamento que são integrados ao modelo básico; e
- existem várias extensões ao modelo básico de RdP.

O estudo das propriedades de uma RdP permite uma análise detalhada do sistema modelado. As propriedades de uma RdP costumam ser divididas em comportamentais, que dependem do estado (ou marcação inicial) e da estrutura da RdP, e estruturais, que dependem apenas da estrutura da rede.

Dentre as propriedades comportamentais das RdPs, podem ser citadas a alcançabilidade, limitação e vivacidade.

A alcançabilidade indica a possibilidade de atingirmos uma determinada marcação pelo disparo de um número finito de transições, a partir de uma dada marcação. Uma marcação M' é dita acessível de M_0 se existe uma seqüência de transições que, disparadas, levam a marcação M'.

Alcançabilidade: *Seja Mi[tj > Mk e Mk[th > M1 então Mi[tjth > M1. Por recorrência designamos o disparo de uma seqüência s Є T* por M[s > M'. O conjunto de todas as possíveis marcações obtidas a partir da marcação* M_0 *na rede RM = (R;* M_0*) é denotado por CA(R;* M_0*) =* $\{M' \in IN^m \mid \exists\ s\ ,\ Mo[s > M'\}$*, onde m é a cardinalidade do conjunto de lugares da rede.*

Alguns problemas de análise podem ser observados em termos dessa propriedade. Por exemplo, se uma rede fica em *deadlock* em uma determinada marcação, pode-se querer saber se essa marcação é acessível.

Uma RdP é limitada se e somente se o número de marcas de cada lugar da rede não supera k. M(pi) ≤ k para qualquer marcação alcançável.

Limitação: *Seja um lugar pi Є P, de uma rede de Petri marcada RM = (R;* M_0*). Esse lugar é dito k-limitado (k-bounded) (k Є IN) ou simplesmente limitados se para toda marcação acessível M Є CA(R;* M_0*), M(pi) ≤ k.*

Rede Limitada: *Diz-se que uma rede RM = (R; M_0) é limitada (bounded) se k(pi) ≤ ∞, ∀ p Є P.*

O conceito de vivacidade está definido em função das possibilidades de disparo das transições. O termo vivacidade também é conhecido como *liveness*. Vivacidade é uma propriedade fundamental para sistemas do mundo real. Porém, muitas vezes é muito caro observar essa propriedade em alguns sistemas de grande porte. A ausência de bloqueio (*deadlock*) em sistemas está fortemente ligada ao conceito de vivacidade. *Deadlock* em uma RdP é a impossibilidade do disparo de qualquer transição da rede.

Rede viva: *Uma rede RM = (R; M_0) é dita viva (live) se para toda M Є CA (R; M_0) é possível disparar-se qualquer transição de RM através do disparo de alguma seqüência de transições.*

As propriedades estruturais são aquelas que refletem características independentes da marcação. Tais propriedades possibilitam a análise do comportamento em função da estrutura do modelo. Desde que as redes sejam puras, a estrutura da rede pode ser representada pela matriz de incidência. Serão visto os conceitos de limitação estrutural, conservação, consistência e repetitiva.

Limitação Estrutural: *Uma rede de Petri R = (P, T, 1, O, K) é classificada como estruturalmente limitada (structural bounded) se é limitada para qualquer marcação inicial.*

Uma rede de Petri é estruturalmente viva se é viva para qualquer marcação inicial finita.

Uma rede de Petri é conservativa se o somatório de pesos das marcas em todos os nós da árvore de alcançabilidade for constante, inclusive a marcação inicial. Uma rede conservativa é algumas vezes chamada de rede *S-invariant* ou *P-invariant.*

Conservação: *Uma rede marcada RM = (R; M_0) é dita conservativa com relação a um vetor de pesos W = (w_1, $w_{2,...,}$ w_n), se* $\sum w_1 . M_k(p_i) = \sum w_1 . M_0(p_i)$, *onde n= #P e* w_i *é um inteiro positivo,* $\forall p_i \in P$ *e* $\forall M_k \in A(R; M_0)$.

Um rede de Petri é estruturalmente consistente se disparando uma seqüência de transições a partir de uma marcação M_0 retorna-se a M_0, porém todas as transições da rede são disparadas pelo menos uma vez.

Consistência: *Seja RM = (R; M_0) uma rede marcada e s uma seqüência de transições. RM é dita consistente se* $M_0 [s > M_0$ *e toda transição ti, dispara ao menos uma vez em s.*

Repetição: *Uma rede de Petri é repetitiva se existe uma marcação* M_0 *e uma seqüência de ativações s tal que os elementos associados ao vetor de ativações v são infinitos.*

Para as redes de Petri os métodos de análises são classificados em três grupos: análise baseada na árvore de cobertura, os métodos baseados na equação de estado e as técnicas de redução. [48] , [49]

Árvore de Cobertura – esse método de análise se baseia na construção de uma árvore que possibilite a representação de todas as possíveis marcações de uma rede. Com a marcação inicial de uma rede de Petri obtêm-se diversas marcações para um grande número de transições potencialmente habilitadas. Para cada nova marcação, podem-se encontrar novas marcações alcançáveis.
A árvore de cobertura é utilizada para representar de forma finita um número infinito de marcações. Para uma rede de Petri limitada, a árvore de cobertura é denominada árvore de alcançabilidade, dado que esta contém todas as possíveis marcações da rede. Algumas propriedades, tais como limitação e transições mortas podem ser analisadas através da árvore de cobertura [39] .

Equação de Estado – o comportamento dinâmico de muitos sistemas pode ser descrito por equações diferenciais ou equações algébricas. A vantagem das

técnicas algébricas sobre as técnicas baseadas na análise das árvores de cobertura é que a análise de propriedades pode ser efetuada pela resolução de equações lineares simples. As equações desenvolvidas governam o comportamento concorrente dos sistemas modelado por RdPs. Todavia, a solução dessas equações é limitada, em parte devido à natureza não determinística dos modelos de RdP e por causa da restrição que soluções devem ser encontradas como inteiros não-negativos [48] .

Matriz de incidência: *a matriz de incidência A de uma RdP é uma matriz n x m de inteiros, definida como:*

$$A = [a_{ij}]$$

e

$$a_{ij} = a^{+}_{ij} - a^{-}_{ij}$$

onde $a^{+}_{ij} = w(i,j)$ *é o peso do arco da transição i para seu lugar de saída j e* $a^{-}_{ij} = w(i,j)$ *é o peso do arco do lugar de entrada j para a transição i.*

Seja m_k *a marcação de uma RdP após sua k-ésima execução, com* $k \geq 0$*. A próxima marcação* m_{k+1} *é determinada pela equação de estado definida por*

$$m_{k+1} = m_k + Au_k$$

onde A é a matriz de incidência e u_k *é um vetor de dimensão (m x 1) de inteiros não negativos, chamado vetor de disparo, no qual cada entrada representa o número de vezes que a respectiva transição disparou durante a k-ésima execução da rede.*

Reduções: reduções são transformações aplicadas ao modelo de um sistema com o objetivo de simplificá-lo, e ainda preservando as propriedades do sistema a ser analisado. Consistem em transformações que reduzem a dimensão do grafo de alcançabilidade, mas que asseguram parcialmente a conservação das propriedades a serem analisadas.

As técnicas de redução são baseadas nas transformações de redes originais em um modelo mais abstrato de tal maneira que propriedades como

liveness, *boundedness* e *safeness* são preservadas nos modelos obtidos por estas reduções. A transformação reversa (refinamento) pode ser usada para processos de síntese.

As regras de transformação das redes podem ser a partir de aplicação das fusões, tanto de lugares, quanto de transições. Serão citadas apenas as mais simples:

- Fusão serial de lugares, como mostrado na Figura 3.6 (a).
- Fusão serial de transições, como mostrado na Figura 3.6(b).
- Fusão paralela de lugares, como mostrado na Figura 3.6(c).
- Fusão paralela de transições, como mostrado na Figura 3.6(d).
- Eliminação de lugares auto-laço, como mostrado na Figura 3.6(e).
- Eliminação de transições auto-laço, como mostrado na Figura 3.6(f).

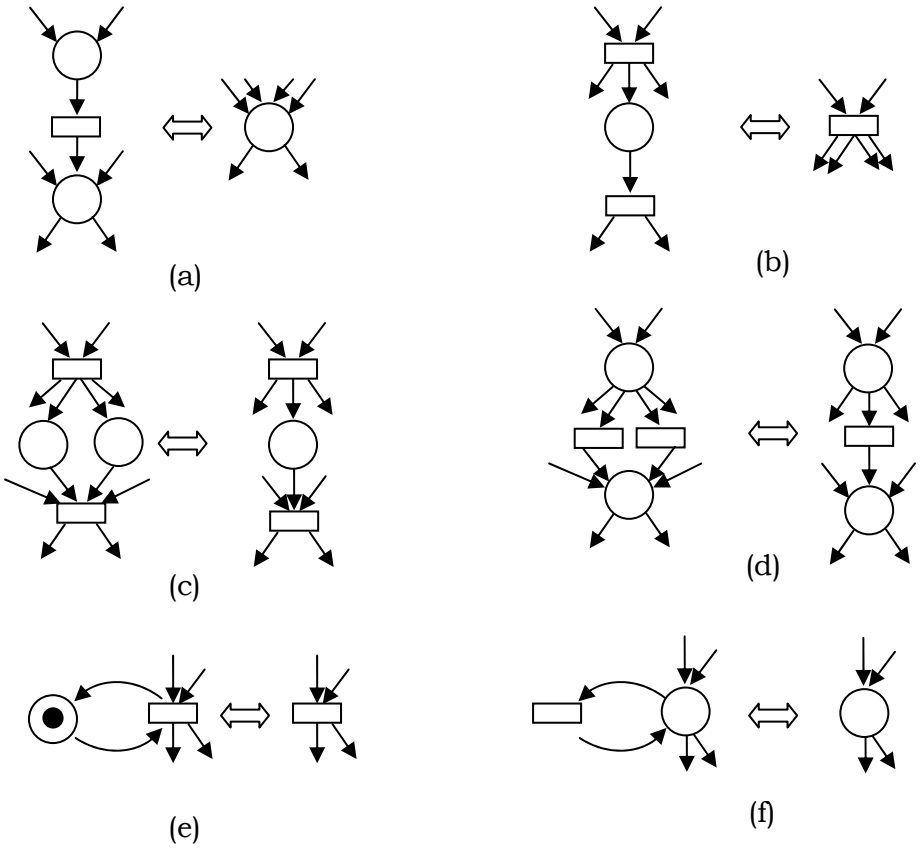


Figura 3.6: Seis transformações preservando Vivacidade e Limitação.

3.2.1 GSPN

GSPN foi originalmente definida em [34] . Modelos GSPN compreendem dois tipos básicos de transições: as transições temporizadas, as quais têm retardos (*delays*) exponencialmente distribuídos associados, e as transições imediatas, que disparam com tempo zero, e têm prioridade superior as transições temporizadas. Níveis diferentes de prioridade podem ser atribuídos às transições. As prioridades podem servir para solucionar situações de confusão. Associam-se pesos às transições imediatas, a fim de solucionar situações de conflito.

Formalmente um modelo GSPN é uma óctupla μ_{GSPN} *= (P, T, Π, I, O, H,* M_0*, W), onde (P, T, Π, I, O, H,* M_0*) é a rede não temporizada PN subjacente, que compreende:*

- *Um conjunto de lugares P;*
- *Um conjunto de transições T;*
- *As funções de entrada e de saída I,O;*
- *A função de arco inibidor H:* $T \rightarrow IN$*;*
- *Uma marcação inicial* M_0*;*
- *Adicionalmente, a definição GSPN compreende a função de prioridade* $\Pi : T \rightarrow N$*; a qual associa a menor prioridade (≥1) às transições temporizadas e prioridades mais alta (0) para transições imediatas:*

$$\Pi(t) = \begin{cases} \geq 1 \text{ se } t \text{ é temporizada} \\ 0 \text{ se } t \text{ é imediata} \end{cases}$$

- *A função W: T → ℜ {0}, que associa um valor real não negativo com transições w(t) é:*

• *Se t é uma transição temporizada, então w será o valor do parâmetro da função densidade probabilidade exponencial;*

• *Se t é uma transição imediata, então w será um peso, que é usado para o cálculo das probabilidades de disparo das transições imediatas em conflitos.*

.

Os arcos inibidores são usados para prevenir transições de serem habilitadas quando certa condição é verdadeira.

Nos modelos GSPN há dois tipos de estados (marcações) chamados de estados tangíveis (*tangible*) e os estados voláteis (*vanish*). Os estados voláteis são assim denominados, porque o seu tempo de vida é igual a zero. O estado volátil é criado em decorrência da marcação dos lugares que são pré-condições de uma transição imediata. Dessa forma, quando as marcas chegam a esses lugares, são instantaneamente consumidas. O tempo de permanência das marcas nesses lugares é zero. Essa é a razão de chamá-los de estados voláteis, pois são criados e instantaneamente destruídos.

O modelo GSPN faz uso da semântica *interleaving* de ações [33] . Assume-se que as transições são disparadas uma a uma, mesmo que o estado compreenda transições imediatas não conflitantes. A análise de um modelo GSPN requer a solução de um sistema de equações igual ao número de marcações tangíveis.

A Figura 3.7 apresenta um exemplo de geração do grafo de alcançabilidade de uma GSPN. Na primeira rede, existe um conflito entre duas transições imediatas (t1 e t2). Quando a transição T1 dispara, o sistema entra estado p2, habilitando as duas transições imediatas, t1 e t2, gerando o estado p2. Há uma mudança imediata (tempo zero) para o estado p3 ou p4, através do disparo da transição t1 ou t2, com probabilidades $\frac{\alpha}{\alpha+\beta}$ e $\frac{\beta}{\alpha+\beta}$ respectivamente.

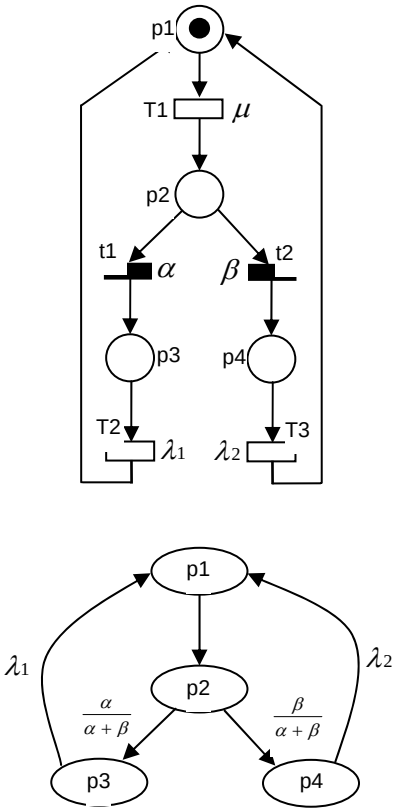


Figura 3.7: Gráfico de alcançabilidade GSPN

A taxa na qual o sistema se move do estado p1 para p3 ou p4 é obtida pelo produto da taxa λ da transição do estado p1 para o estado volátil p2, com a probabilidade de ir do estado p2 para o estado p3 ou p4. Qualquer rede de Petri estocástica marcada, com um número finito de lugares e transições, é isomórfica a uma cadeia de Markov [48] .

3.2.2 Moment Matching

Verificamos anteriormente que as GSPN estão restritas para transições imediatas e exponenciais. Porém, essa restrição não limita os resultados de desempenho quando o sistema contém transições de estados não exponenciais. As cadeias Semi-Markovianas fornecem uma estrutura matemática simples para inclusão de distribuições poli-exponenciais na estrutura do modelo Markoviano. Uma variedade efetiva de *delays* pode ser contruída em modelos GSPN, usando-se os construtores *throughput subnets e s-transitions.* O conceito de construtores GSPN surgiu para aumentar a abrangência do modelo para outros tipos de distribuição não-exponencial [2] .

Várias combinações de lugares, transições exponenciais e transições imediatas podem ser usadas entre dois lugares para alcançar diferentes tipos de distribuição. A Figura 3.8 a seguir descreve três *throughput subnet* [33] formados por conexões em série e paralelo.

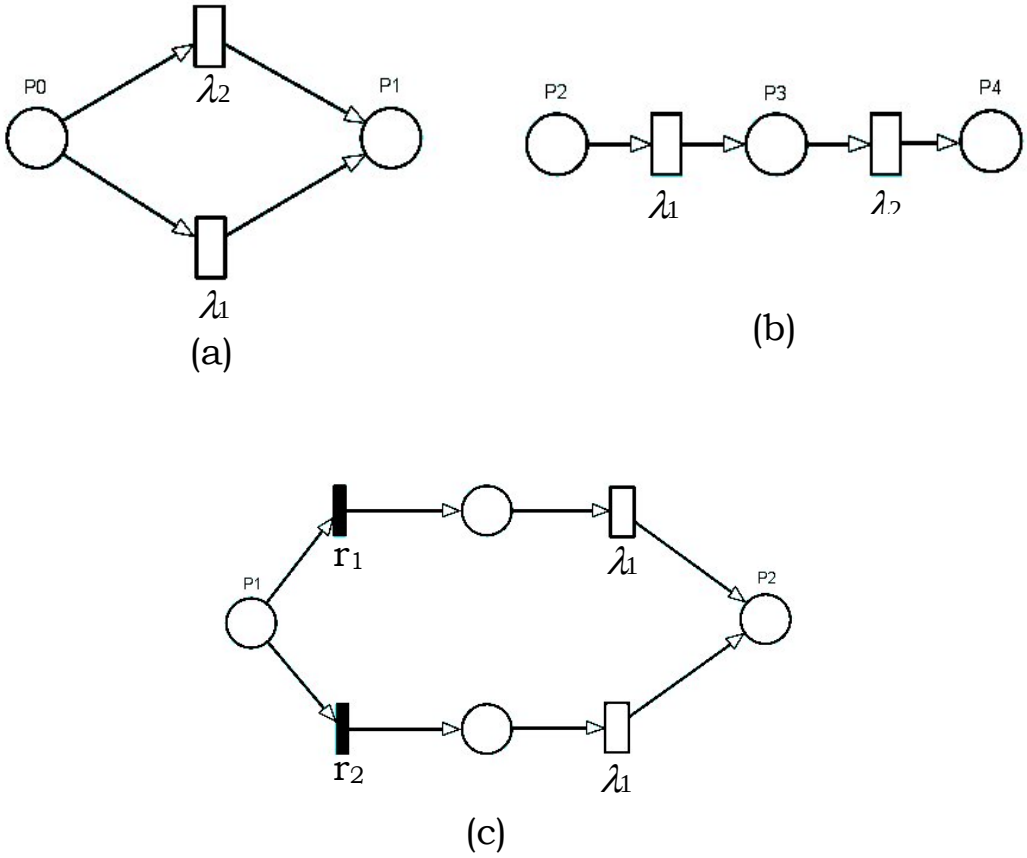


Figura 3.8: Diferentes tipos de *Throughput subnets*

A Figura 3.8a consiste em duas transições exponenciais em paralelo, com taxa de parâmetro λ_1 e λ_2, respectivamente. Um *token* aparecendo em P0 aparecerá em P1 depois que dois curtos *delays* r_1 + r_2 tiverem decorridos. O resultado da função de densidade com esse *delay* é dado por:

$r = \min(r_1 + r_2)$

$f_r(t) = (\lambda_1 + \lambda_2) e^{(\lambda_1 + \lambda_2)t}$, $t \geq 0$.

Dessa forma, esse tipo de interconexão é equivalente a uma transição exponencial com parâmetro $\lambda_1 + \lambda_2$ e, consequentemente, não fornece flexibilidade na modelagem com relação à análise de desempenho.

A Figura 3.8b consiste em duas transições exponenciais em série com parâmetro $\lambda_1 + \lambda_2$, respectivamente. O *delay* resultante é $t = t_1 + t_2$, resultando na função de densidade abaixo:

$$f_\tau(t) = (f_{\tau 1} * f_{\tau 2})(t)$$

$$= \lambda_1 \lambda_2 (e^{-\lambda_1 t} - e^{-\lambda_2 t}) / (\lambda_2 - \lambda_1), \quad t \geq 0.$$

Onde ***** é o operador de convolução. Esta expressão generaliza para mais que duas transições. Para o caso onde $\lambda_1 = \lambda_2 = \ldots = \lambda_n$, a função densidade vem:

$$f_\tau(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}, \quad t > 0$$

Isto é, uma distribuição do tipo *Erlang* de ordem N. Em particular a distribuição do tipo Erlang é especificada por dois parâmetros $\lambda > 0$ e $n > 0$.

A Figura 3.8c consiste em uma distribuição hiperexponencial que é modelada com duas ramificações paralelas, cada uma contendo uma transição imediata e outra exponencial. A transição imediata forma a probabilidade de mudar do lugar P1. Quando um *token* chegar em P1, a probabilidade de cada ramificação é determinada pelos pesos das transições r_1 e r_2. A função de densidade é:

$$f_\tau(t) = r_1 f_\tau(t) + r_2 f_\tau(t) = r_1 \lambda_1 e^{-\lambda 1 t} + r_2 \lambda_2 e^{-\lambda 2 t}, \quad t > 0$$

Esta *throughput subnet* implementa uma função de *delay* hiperexponencial. Uma particular distribuição hiperexponencial é caracterizada da seguinte forma:

n, a ordem

$r_j = 1 \ldots n, \sum r_j = 1$

λi, i = 1...n.

No entanto, muitos estudos têm fornecido meios de adaptação de modelos Markovianos (aproximações) [2] [10] para representar outras distribuições temporizadas. O ajuste da distribuição pode ser aplicado para descobrir a distribuição teórica que melhor se adequa à distribuição empírica. Infelizmente, essa distribuição pode ser muito complexa e quando considerar o número de variáveis estocásticas simultâneas, isso pode levar a um enorme complexo processo estocástico.

Em geral, para representar uma distribuição com variáveis aleatórias, é necessário representar uma combinação de variáveis aleatórias exponenciais (distribuições exponenciais)[10] . Um método bem estabelecido que considera distribuições expolinomiais é baseado na distribuição do *moment-matching.* O processo de *moment-matching* apresentado por Desrochers and Al-Jaar [10] leva em conta que distribuições Hipoexponencial e Erlangian têm um atraso médio (μ_D) menor do que o desvio-padrão (σ), e distribuições Hiperexponenciais têm um atraso médio (μ_D) maior do que o desvio-padrão (σ). De fato, para representar uma distribuição como uma Erlangian ou uma Hiperexponential, a *subnet* é conhecida como uma *s-transition*. Vale notar que em alguns casos nas quais essas distribuições têm um $\mu_D = \sigma$, eles são, de fato, equivalentes para uma distribuição exponencial com parâmetro igual a $1/\mu_D$.

Portanto, de acordo com o coeficiente de variação associado com atraso de uma atividade, uma apropriada implementação do modelo de *s-transition* pode ser escolhido. Para cada implementação do modelo de *s-transition*, um conjunto de parâmetros deve ser configurado para adequar seus primeiros e segundos momentos. Em outras palavras, o atraso médio (μ_D) e o respectivo desvio padrão (σ), que são calculados estatisticamente a partir da distribuição dos dados medidos, são combinados com o primeiro e segundo momentos da *s-transition* (distribuições exponenciais).

Dependendo do valor do coeficiente de variação (CV = σ / μ_D), a respectiva atividade é associada para uma dessas distribuições: Erlang, Hiperexponencial or Hipoexponencial.

No caso do coeficiente de variação ser maior do que 1 (CV > 1), ou o mesmo seja um valor inteiro, a aproximação deve ser feita por uma distribuição de Erlang, que é representada por uma seqüência de transições exponenciais γ, com taxa λ. O modelo de redes de Petri para uma aproximação da distribuição de Erlang é representado na Figura 3.9.

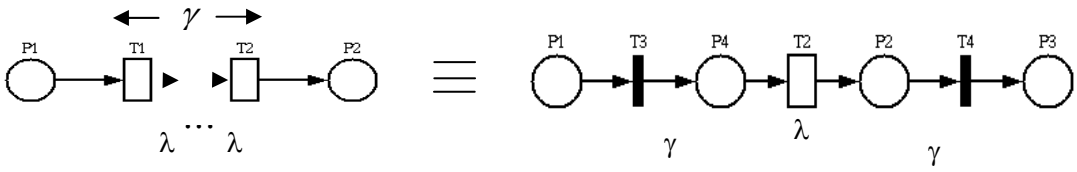


Figura 3.9: Distribuição de Erlang

Os parâmetros λ e γ são calculados de acordo com as equações abaixo:

$$\gamma = \left(\frac{\mu_D}{\sigma_D} \right)^2$$

$$\lambda = \frac{\gamma}{\mu_D}$$

Caso o coeficiente de variação seja menor do que 1 (CV < 1), sendo um inteiro ou não, a distribuição é aproximada por uma distribuição Hiperexponencial, cujos parâmetros são: λ, taxa da transição exponencial, e r_1 e r_2, pesos das transições imediatas. Tais parâmetros são calculados através das equações:

$$\lambda = \frac{2\mu_D}{(\mu_D^2 + \sigma_D^2)}$$

$$r_1 = \frac{2\mu_D^2}{(\mu_D^2 + \sigma_D^2)}$$

$$r_2 = 1 - r_1$$

O respectivo modelo em redes de Petri para distribuição Hiperexponencial é representado na Figura 3.10.

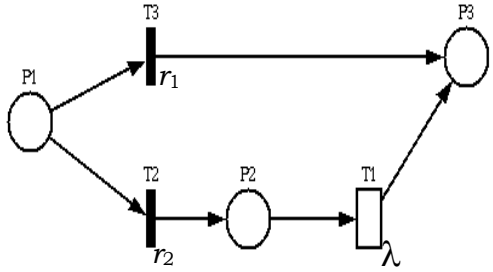


Figura 3.10: Modelo Hiperexponencial

Para o caso de CV > 1, a distribuição deverá ser aproximada à distribuição Hipoexponencial, composta de uma exponencial com taxa λ_1 e uma Erlang formada por γ exponenciais de taxas λ_2. Os parâmetros λ_1, λ_2 e γ são calculados a partir das equações seguintes:

$$\left(\frac{\mu}{\sigma}\right)^2 - 1 \leq \gamma < \left(\frac{\mu}{\sigma}\right)^2$$

$$\mu_1 = \frac{\mu_d \mp \sqrt{\gamma(\gamma+1)\sigma_d^2 - \gamma\mu_d^2}}{\gamma+1} \qquad \lambda_1 = \frac{1}{\mu_1}$$

$$\mu_2 = \frac{\gamma\mu_d \pm \sqrt{\gamma(\gamma+1)\sigma_d^2 - \gamma\mu_d^2}}{\gamma+1} \qquad \lambda_2 = \frac{1}{\mu_2}$$

O modelo em redes de Petri que representa a distribuição Hipoexponencial é representado na Figura 3.11.

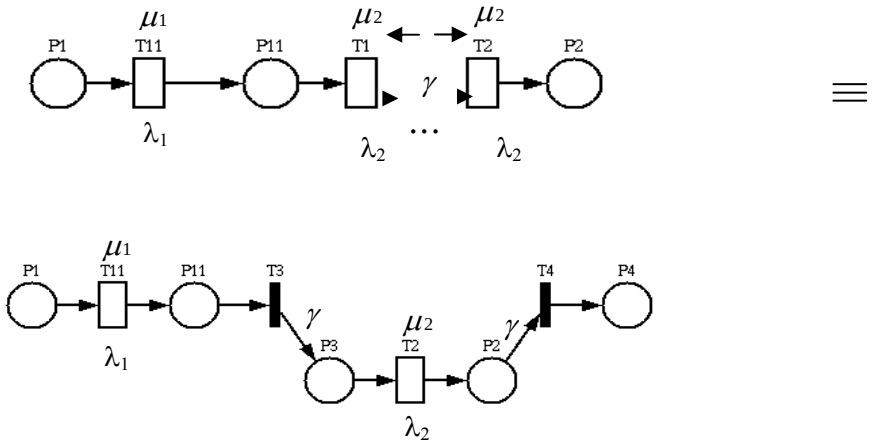


Figura 3.11: Distribuição Hipoexponencial

3.3 Visão Geral sobre Filas

Um sistema de filas, como mostrado na Figura 3.12, consiste de *buffer* de tamanho fixo ou infinito e um ou mais servidores [2] . Um servidor pode somente atender um cliente por vez e, portanto, ele pode estar ocupado ou ocioso. Se todos os servidores estão ocupados quando novos clientes estão chegando, então estes serão colocados no *buffer*, caso esteja disponível, para, posteriormente, serem atendidos. Quando o cliente deixa de ser atendido, ou seja, deixa o sistema, um cliente que estava esperando é atendido de acordo com a política (ou disciplina de atendimento) de fila [2] .

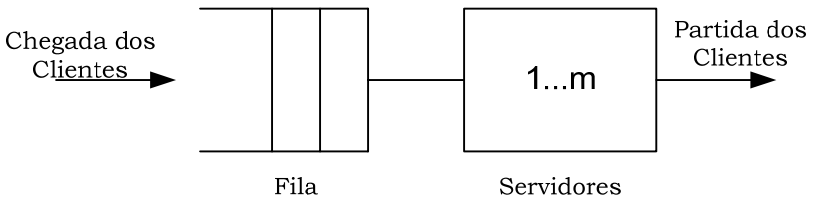


Figura 3.12 Representação de uma fila

A política de fila ou disciplina de atendimento determina qual cliente é selecionado da fila para ser atendido quando o servidor torna-se disponível. Algumas políticas de filas comumente usadas são:

- FCFS (*First-Come-First-Served*): os clientes são atendidos pela ordem de chegada.
- LCFS (*Last-Come-First-Served*): os últimos que chegam são os primeiros a serem atendidos.
- *RR* (*Round Robin*): cada cliente recebe uma fatia de tempo do servidor dentro da qual é atendido. Após o término do tempo, se a atividade não foi completada, o cliente volta para fila de acordo com a política FCFS e outro passa a ser atendido.
- *Preemption:* o cliente com maior prioridade é atendido imediatamente, interrompendo o atendimento ao cliente com menor prioridade. Ao terminar, o cliente de menor prioridade volta a ser atendido, podendo continuar o processo de onde parou ou então reiniciá-lo.

O conceito de "cliente" em teoria de filas é um termo genérico, aplicando-se não somente a seres humanos, mas também a pacotes que trafegam em uma rede VoIP, por exemplo. Da mesma forma o conceito de "servidores" não se limita apenas a recurso computacional, podendo ser estendido a switches ou roteadores.

A característica do sistema de filas consiste no tempo entre chegadas do cliente, existindo diferentes tipos de distribuições que representam o tempo entre chegadas. A distribuição de Erlang, Hipoexponencial, Hiperexponencial e Exponenciais são exemplos de uma dessas distribuições.

Diferentes sistemas de filas são analisados matematicamente para determinar as medidas de desempenho. As importantes medidas de desempenho são descritas a seguir.

Se um sistema de filas consiste de um único servidor, então a utilização ρ é a fração do tempo em que o servidor está ocupado. Caso não exista um

limite no número de clientes que podem ser atendidos pela fila do servidor, então a utilização do servidor é dada por [2] :

$$\rho = \frac{\text{tempo médio serviço}}{\text{tempo médio entre chegadas}} = \frac{\text{taxa chegada}}{\text{taxa serviço}} = \frac{\lambda}{\mu}$$

A utilização de uma estação de serviço com múltiplos servidores é dada pela fração da quantidade de servidores ativos, desde que $m\mu$ seja a taxa global do serviço:

$$\rho = \frac{\lambda}{m\mu},$$

O interesse maior de se avaliar um cenário com critérios de desempenho requer uma avaliação estacionária, quando todo o comportamento transiente tiver terminado. Mas sabe-se que dependendo da complexidade do cenário proposto isso nem sempre é possível. No caso de solução estacionária, ρ pode ser usado para formular a condição para comportamento estacionário. Essa condição é dada por:

$$\rho < 1,$$

Quando essa situação acima acontece, o sistema está em equilíbrio estatístico, isto é, a taxa na qual os clientes entram no sistema para serem atendidos é igual a taxa que os clientes deixam o sistema. Em outras palavras, o número de clientes que chegam em um determinado intervalo de tempo é menor do que o número de clientes que o sistema consegue processar.

O *throughput* do sistema é definido como a quantidade média de clientes que o sistema consegue processar por uma unidade de tempo. O *throughput* é dado por:

$\lambda = \mathrm{m}\rho\mu$

O número médio de clientes na fila do sistema pode ser expresso por $\overline{\mathrm{K}}$ e o tamanho médio da fila $\overline{\mathrm{Q}}$ pode ser calculado usando um dos mais importantes teoremas de *Little[42] :*

$\overline{\mathrm{K}} = \lambda\overline{\mathrm{T}}$ ($\overline{\mathrm{T}}$ é o tempo total gasto no sistema incluindo o tempo de espera e o tempo de serviço).

$\overline{\mathrm{Q}} = \lambda\overline{\mathrm{W}}$ ($\overline{\mathrm{W}}$ é o tempo gasto de espera para ser atendido).

Capítulo 4 - Metodologia de Avaliação

Este capítulo apresenta a metodologia usada para avaliação de desempenho do tráfego de voz em uma rede VoIP. Este capítulo descreve todas atividades necessárias para coleta dos dados, bem como os requisitos para avaliação, modelagem e validação do modelo e avaliação de cenários.

A metodologia de avaliação desempenho é baseada em modelos *GSPN* e para que se execute, com sucesso, alguns aspectos devem ser observados, tais como a complexidade dos modelos e a adequação de recursos computacionais disponíveis.

O ambiente a ser avaliado pode ser alterado devido as diferentes métricas que podem ser obtidas, no qual o cenário de avaliação é diverso, assim como os recursos envolvidos. No entanto, existe um conjunto de atividades comuns relativas a estes cenários e estas são definidas na metodologia.

4.1 Visão Geral

A metodologia proposta consiste de oito atividades: definição do problema e dos componentes, medição, geração do modelo abstrato, análise das propriedades e validação do modelo abstrato, mapeamento das métricas e refinamento do modelo, validação do modelo refinado, avaliação do desempenho da aplicação e interpretação dos resultados. Esta seção apresenta uma visão sucinta das atividades e têm por objetivo fazer com que o leitor tenha uma perspectiva geral das atividades executadas. A Figura 4.1 descreve o fluxo da metodologia.

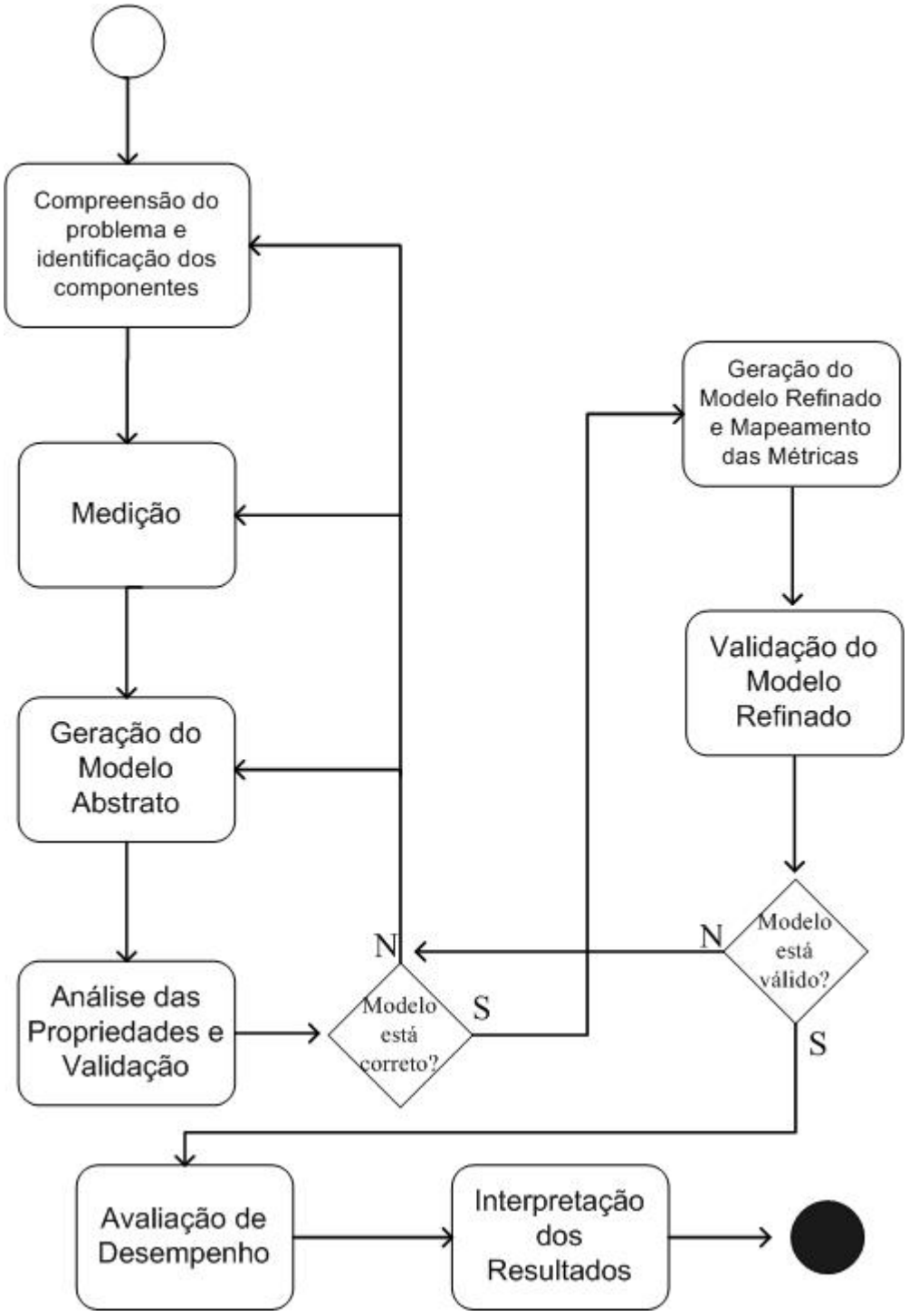


Figura 4.1 Fluxo da Metodologia

A primeira atividade da metodologia consiste na definição do contexto do sistema. Nesta atividade deve-se levar em consideração qualquer tráfego e componentes que interfiram no desempenho da aplicação de VoIP.

A atividade de medição corresponde à fase de coleta dos dados. Esta fase exige que os recursos disponíveis estejam dedicados para o experimento. Portanto, é necessário avaliar o tráfego do sistema considerando situações particulares para que se possam obter medidas "básicas"[2] antes para geração do modelo refinado e posterior realização dos experimentos.

O modelo abstrato é a base para geração de modelos que representarão os aspectos temporais específicos do sistema particular que se avalia. O modelo abstrato consiste numa representação de alto-nível dos componentes definidos na atividade de identificação dos componentes. Nesta fase nenhum aspecto temporal particular é atribuído a esses componentes.

Após a geração do modelo abstrato, este é analisado para se verificar as propriedades qualitativas definidas no Capítulo 3 (seção 3.2). Além da análise qualitativa formal, o modelo abstrato deve ser validado (qualitativamente) através do *token game*[3].

Após a obtenção do modelo abstrato e das métricas básicas associadas aos componentes do sistema (obtidas na fase de medição) o modelo refinado deve ser gerado, assim como as métricas de interesse devem ser representadas através dos elementos do modelo (marcações de lugares e transições). Este modelo deve ser validado quantitativamente através de um estudo comparativo entre aos valores das métricas de interesse (calculados do modelo) e os valores respectivos obtidos do sistema real em estudo. É importante salientar que o processo de validação quantitativa, embora realizado seguindo procedimentos estatísticos, possibilita afirmar que o modelo concebido fornece resultados equiparáveis ao sistema para os cenários avaliados. Portanto, quanto maior o

[2] Considera-se medidas "básicas" medidas intrínsecas associadas a dispositivos. Por exemplo, o tempo necessário que o determinado dispositivo leva para processar um pacote de dados.

[3] *Token game* é o termo utilizado para se referir a simulação qualitativa executada pelo disparo de transições de forma automática ou através da escolha das seqüências de disparo feitas pelo avaliador.

número de cenários (graus de liberdade) [42] e o nível de confiança observado, maior credibilidade terá o modelo.

Após a validação, o modelo pode ser configurado para representar o conjunto de cenários de interesse, as métricas de desempenho devem ser calculadas, os resultados interpretados e o diagnóstico apresentado utilizando-se uma linguagem adequada ao público a quem se dirige.

4.2 Atividades da Metodologia

A primeira atividade corresponde à fase de compreensão do sistema, ressaltando o contexto em que a aplicação está inserida e o tráfego de voz no cenário de interesse a ser avaliado. O cenário a ser avaliado é estudado de maneira a definir os componentes relevantes e suas interações internas com o ambiente. A execução desta atividade depende de um conjunto de ações que devem ser realizadas. As ações são:

- Compreender a estrutura do sistema a ser avaliada (é necessário um desenho que demonstre como estão organizados os componentes e as relações entre eles). Descrever a estrutura do ambiente a ser avaliado e as condições limitantes (a descrição compreende o fluxo de comunicação entre todos os componentes que influenciem o desempenho do sistema em avaliação).
- Configurar as ferramentas de software e hardware para execução das atividades.

Este trabalho utiliza a ferramenta Myphone (http://myphone.sourceforge.net/) para comunicação de voz entre os clientes. Esta é uma ferramenta que permite a utilização dos codecs para avaliação do sitema, além de transmitir voz com baixas larguras de banda na rede. Utiliza-

se, ainda, o software Tfgen (http://www.st.rim.or.jp/~yumo/pub/tfgen.html) para geração do tráfego de dados. Esse será o tráfego concorrente com o tráfego de voz. Para coleta dos dados, usa-se o Ethereal (http://www.ethereal.com/) e o What's up (http://www.ipswitch.com/Products/WhatsUp/), além da interface de linha de comando, própria dos recursos envolvidos que permitem toda a análise e coleta de tráfego.

As ferramentas de análise de tráfego são utilizadas para identificação do tráfego dos clientes de voz e dos componentes envolvidos no cenário. Para obtenção dos tempos médios de transmissão, os computadores devem estar sincronizados via *Network Time Protocol* (NTP).

A segunda atividade corresponde à atividade de medição. A primeira parte da medição concerne à coleta dos tempos de transmissão dos componentes que influenciam o desempenho do sistema. No caso particular dos pacotes de voz, como estes são transmitidos via o protocolo RTP (Real Time Protocol), utiliza-se o número de seqüência para determinar a localização (a ordem) de um pacote e o *timestamp* para identificação do tempo que o pacote de voz foi gerado. Portanto, sabendo o tempo de geração e o tempo de chegada dos pacotes de voz, têm-se o tempo médio entre chegadas. A taxa de chegada dos pacotes de voz é determinada pela taxa de envio.

A terceira atividade refere-se à geração do modelo abstrato. O modelo abstrato é a base para geração de modelos que possuirão aspectos temporais particulares, pois o modelo abstrato não tem nenhuma informação temporal específica associada. Esta atividade modela cada componente do sistema e os une através das regras de composição. A escolha da granularidade determina as análises que podem ser realizadas, dependendo do detalhamento dos componentes do modelo.

Todos os passos para geração do modelo abstrato foram feitos utilizando o software TimeNET (*http://pdv.cs.tu-berlin.de/~timenet/*) que é

uma ferramenta de software que fornece um ambiente de modelagem GSPN, permitindo a criação gráfica do modelo, a definição das métricas de desempenho, a análise estrutural e a avaliação estocástica do modelo.

A quarta atividade refere-se à análise das propriedades e validação do modelo. Esta análise diz respeito às duas classes de propriedades: propriedades comportamentais e estruturais. Definidas no Capítulo 3, as propriedades comportamentais são as que estão relacionadas com a estrutura do modelo e seu estado inicial (marcação inicial). Propriedades estruturais também dizem respeito a aspectos comportamentais, mas não estão ligadas a uma determinada marcação inicial. Entre as propriedades comportamentais de interesse podemos enfatizar: limitação, conservação, livre de *deadlock,* vivacidade e reversibilidade. As propriedades estruturais de interesse para nossa metodologia são: limitação estrutural, conservação estrutural (invariantes de lugar) e invariantes de transição (necessária condição para vivacidade). Além dessas propriedades a validação através do *token game* é muito importante, pois permite ao analista verificar se o modelo apresenta algum comportamento não esperado. Se o modelo abstrato foi validado e as propriedades foram atendidas, o modelo pode ser refinado.

A quinta atividade refere-se à geração do modelo refinado e nessa atividade o modelo abstrato é transformado com base nas estatísticas obtidas na atividade de medição. Os mapeamentos das métricas que representam as métricas de desempenho desejadas são caracterizados pelos seus elementos (conjunto de lugares e transições) no modelo.

Os dados coletados correspondem a uma amostra de uma distribuição desconhecida. Antes da coleta efetiva dos dados, uma análise exploratória dos dados é primordial. A análise exploratória permite a detecção de situações associadas ao sistema que gerem *outliers,* o correspondente diagnóstico, assim como o eventual ajuste no processo de medição. Dentre as estatísticas calculadas, são de interesse particular, para o processo de refinamento, a média e desvio padrão. Utiliza-se a técnica de *moment matching* para realizar a

aproximação por fases. Para isso, deve-se calcular o inverso do coeficiente de variação (CV):

$$CV = \frac{\sigma_D}{\mu_D}$$

Uma amostra representativa de distribuição empírica desconhecida deve ser representada em modelo de GSPN através das seguintes distribuições Expolinomiais: Erlang, Hiperexponencial ou Hipoexponencial. Isso torna possível representar a questão probabilística envolvida no problema através de distribuições que derivam de uma associação de transições exponenciais.

A sexta atividade é a validação do modelo refinado. Precisa-se validar o modelo dadas as métricas especificadas. Além de executar todos os passos relativos à validação do modelo abstrato, o modelo refinado é validado quantitativamente, ou seja, os resultados das métricas obtidos por meio do modelo e os mensurados no sistema são comparados para verificar se o modelo refinado reflete as características do sistema. Após a validação do modelo, cenários de interesse podem ser avaliados, possibilitando o diagnóstico e o planejamento de recursos, considerando métricas de estado estacionário ou transiente.

Na avaliação transiente, o comportamento do modelo é avaliado para uma dada condição inicial. A avaliação no estado estacionário, por outro lado, é independente do tempo. Dependendo das métricas de interesse e considerando as propriedades qualitativas do modelo refinado, a avaliação pode ser conduzida via análise numérica ou simulação.

A análise é significativa quando uma avaliação de desempenho numérica direta e exata se dá através do grafo de alcançabilidade, do qual se obtêm a Cadeia Markov. Para definir entre as técnicas a utilizar, devem-se considerar os recursos de memória disponíveis, dadas as necessidades de armazenamento do espaço de estados. Vale ressaltar, contudo, que os

resultados são exatos. A simulação, por outro lado, não demanda armazenamento substancial (quando comparado com a análise numérica), contudo seus resultados são aproximações, e para que se alcance uma precisão específica associada a um nível de confiança desejado, o tempo de simulação pode ser longo.

Após a avaliação, os resultados obtidos devem ser interpretados e um diagnóstico apresentado em função das necessidades enunciadas. Avaliações de cenários são realizadas e recomendações podem ser feitas de forma que os requisitos de desempenho e níveis de serviços previamente acertados sejam respeitados.

Capítulo 5 - Modelo de Desempenho

Este capítulo apresenta a descrição dos modelos GSPN que representam componentes de um sistema de VoIP. A última seção do capítulo apresenta a validação do modelo.

5.1 Descrição dos componentes

O primeiro componente descrito é o cliente. A Figura 5.1 representa um cliente que faz alguma requisição com tempos médios exponenciais. Essa representação pode ser de um cliente de voz, ou pode ser um gerador de tráfego, ou qualquer outro tipo de cliente que possua um tempo médio exponencialmente distribuído. No entanto, a transição T0 deste modelo, também pode ser considerada uma transição cuja distribuição temporal seja genérica. Neste caso, a transição T0 deve ser refinada. Neste trabalho, os refinamentos considerados são os relativos às distribuições expolonomiais apresentadas no Capítulo 3.

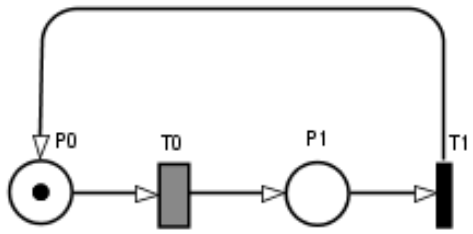


Figura 5.1 Cliente

A Figura 5.2 representa um cliente que possui um comportamento diferenciado em relação ao anterior. Esse comportamento diferenciado é o tráfego em rajadas. A representação desse cliente mostra que em determinados momentos existe períodos de inatividade e de atividade. A

transição genérica T3 (deve ser modelada conforme sua distribuição) representa a ativação do sistema para inatividade, assim como o lugar P2 representa o lugar onde o sistema está inativo. A transição T4 representa a ativação do sistema.

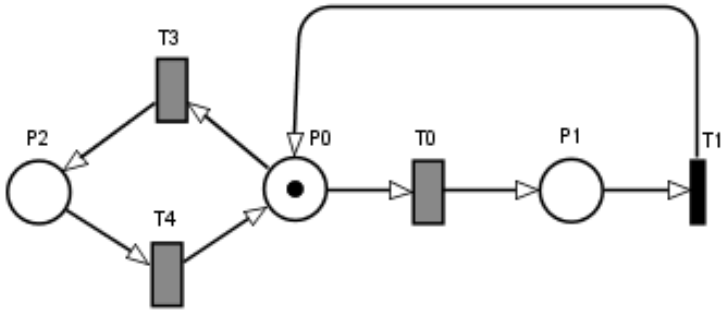


Figura 5.2 Cliente com rajada

Existem algumas ferramentas que permitem que as solicitações feitas pelo cliente sejam "*bufferizadas*" antes de serem enviadas. A rede da Figura 5.3 demonstra como modelar um cliente utilizando um *buffer*. O lugar P01 representa a marcação inicial do sistema, ou seja, a chegada da solicitação ou o início da solicitação.

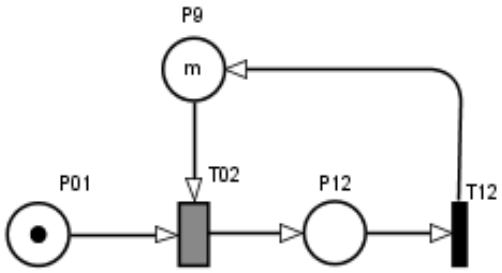


Figura 5.3 Cliente com buffer

Os clientes utilizam equipamentos de interconexão, sejam *switches*, roteadores ou algum concentrador. Se o equipamento de interconexão possuir um *buffer*, o que é normalmente implementado através de uma fila de transmissão, o modelo abstrato pode ser representado como o da Figura 5.4, no qual a marcação m do lugar P9 corresponde ao tamanho do *buffer*.

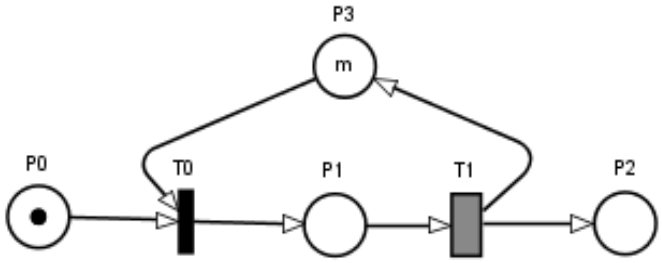


Figura 5.4 Interconexão com um buffer

Se o equipamento de interconexão possuir um *buffer* de entrada e um *buffer* de saída, o adotado é o representado na Figura 5.5.

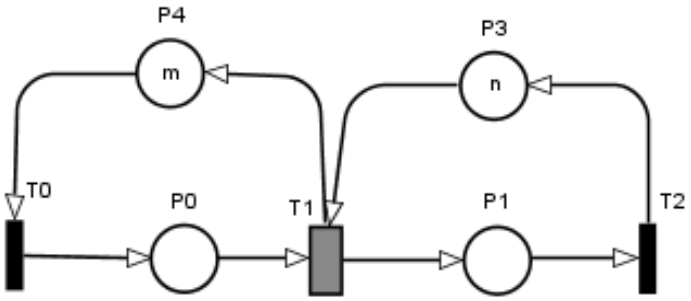


Figura 5.5 Interconexão com um buffer de entrada e buffer de saída

Os equipamentos de interconexão podem ter políticas de filas diferentes. A política se refere as prioridades de atendimento e tamanhos dos *buffers*. Um cenário muito comum em redes com serviços diferenciados é representado na Figura 5.6.

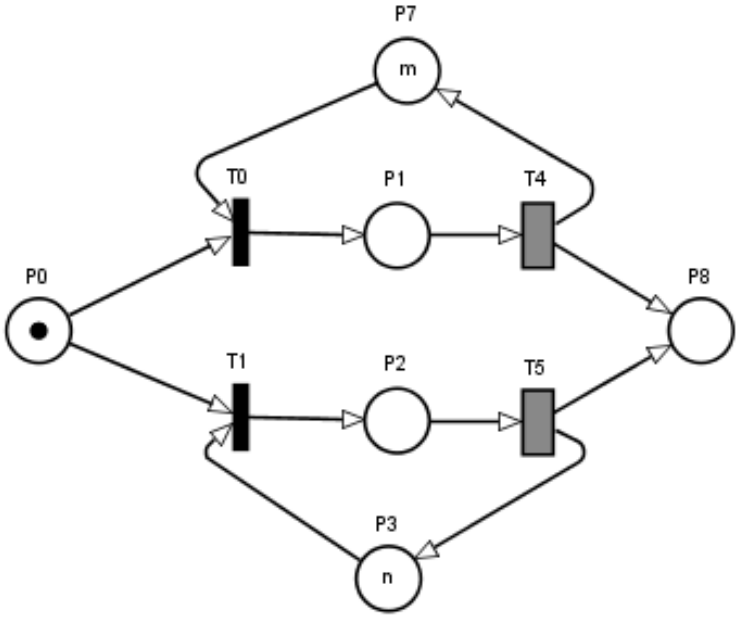


Figura 5.6 Modelo de interconexão com prioridades de atendimento

As características das transições imediatas referentes ao modelo da Figura 5.6 estão representadas na Tabela 5.1.

Tabela 5.1: Características das transições imediatas

Transição	Pesos	Prioridades
T0	w_1	Π_1
T1	w_2	Π_2

A prioridade no contexto do sistema de VoIP refere-se a ordem de ativação, ou seja, quem deve ser atendido primeiro (Π_1 ou Π_2) quando surge um situação conflitante. Com prioridades diferentes nada adianta ter pesos diferenciados nas transições conflitantes. Os pesos poderiam ser associados aos pacotes de VoIP. Exemplificando, pode-se marcar os pacotes de voz com pesos em 80%(w_1) e os que não forem em 20% (w_2). Nesse caso, os pesos de w_1 e w_2 são colocados em 8 e 2, respectivamente.

5.2 Modelo de Validação

Para validar o modelo proposto, um cenário foi criado em que dois clientes de voz, em redes diferentes, se comunicam via VoIP, e uma terceira máquina é utilizada para gerar tráfego de dados de forma a gerar interferência no desempenho da comunicação de voz. Este cenário conta com uma infra-estrutura de hardware composta por quatro computadores (Athlon XP 2500) e um *switch layer* 3. A Figura 5.7 mostra a organização da infra-estrutura de rede composta por duas máquinas clientes de voz, uma máquina de monitoramento, um gerador de tráfego e um *switch*..

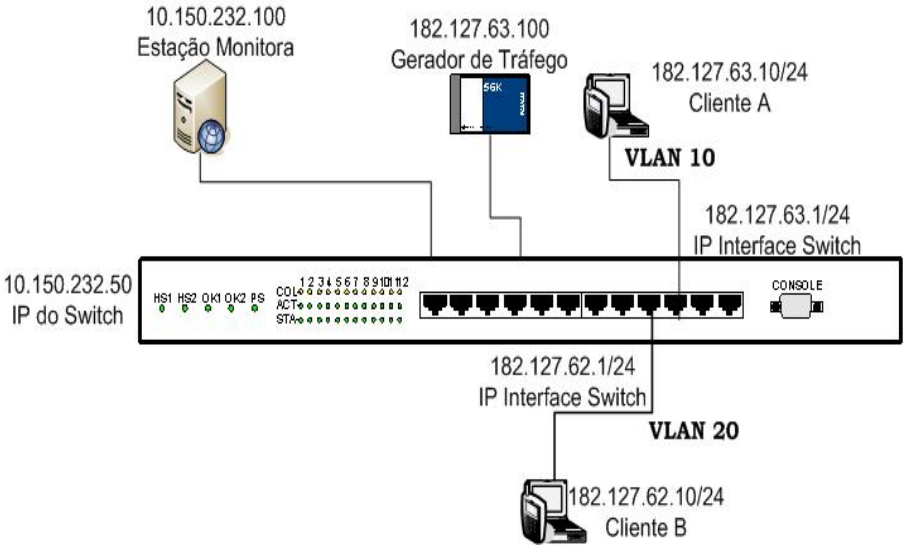


Figura 5.7 Infra-estrutura

Os clientes de voz estão em redes isoladas. O nome *virtual lan* (vlan), refere a uma rede virtual logicamente independente [47] . As vlan's foram usadas para segmentar a rede (Figura 6.2). O cliente A pertence à rede 182.127.63.0/24 (chamada de vlan 10) enquanto o cliente B pertence à rede 182.127.62.0/24 (chamada de vlan 20). Duas interfaces de roteamento foram configuradas para cada rede individualmente. Nesse trabalho utiliza-se o switch E1 *Fast Ethernet* da Enterasys.

A ferramenta de comunicação de VoIP utilizada pelos clientes foi o MyPhone. Esta ferramenta implementa o protocolo H.323 e não necessita de qualquer outro recurso para intermediar a comunicação. Configurou-se o MyPhone para trabalhar com o codec G.711 [22] . Um recurso utilizado foi a detecção de períodos de silêncio, pois se os pacotes de voz são perdidos durante este período, não se tem impacto na aplicação de voz.

O endereço correspondente ao *switch* pertence à outra VLAN e serve para conexão da máquina de monitoramento. É por esse endereço que a máquina de monitoramento coleta os dados dos respectivos componentes (clientes e gerador de tráfego). Cada porta do *switch* fornece informações sobre o número de pacotes que são transferidos. Além dessa coleta, diversos experimentos foram feitos "espelhando" o tráfego dos componentes. Essa técnica chama-se de *port mirror* e serve para reproduzir o tráfego de qualquer porta(s) do equipamento para outra, ou seja, capturamos o tráfego de qualquer componente da infra-estrutura através do espelhamento de portas. É importante ressaltar que antes do processo de medição deve-se verificar a ausência de tráfego na interface. Esta verificação é executada através do Ethereal ou com o gerenciador de tarefas do próprio sistema operacional.

Para geração de tráfego utilizamos o TfGen. O TfGen foi configurado para gerar um tráfego contínuo e constante (CBR – Constant bitrate)[35] . Utiliza-se uma carga contínua e constante porque o codec G.711 também se comporta dessa forma. A carga gerada tem como alvo o cliente B de voz.

Durante a atividade de medição, inicialmente são mensurados apenas os dados dos clientes de voz. Nesse contexto existe apenas o tráfego entre o cliente A e o cliente B. Portanto, nenhum tráfego de dados foi gerado para análise do impacto desse tráfego na qualidade da voz. Medindo o tráfego gerado pela aplicação de voz, obteve-se uma taxa de 74 pps, um tamanho médio de 284 bytes relativos ao pacote de voz e um tempo médio gasto para ir de cliente a outro de $5{,}76 \times 10^{-4}$s.

Posteriormente, outros experimentos são executados, agora utilizando o gerador de tráfego. O objetivo é analisar a qualidade da voz em função da variação do tráfego dados. Portanto, varia-se o tráfego de dados (enviada pelo gerador de tráfego) até o limite em que a comunicação de voz começa perder qualidade.

Coleta-se a média de pacotes transmitidos pelo cliente A e pelo gerador de tráfego, bem como a média de pacotes recebidos pelo cliente B durante um período especificado. Nestes estudos adotou-se o período de 5 minutos porque os valores médios com 10 ou 30 minutos não apresenta grande variabilidade na qual possa comprometer os resultados. Utilizando as ferramentas de medição e de monitoramento (*Ethereal, What's up* e interface de linha de comando), verifica-se que a comunicação de voz é audível até que o tráfego de dados do gerador atinge 4497 pps (somente de A para B). O tamanho médio do pacote de dados igual a 1514 bytes. O tamanho do pacote de dados é 5.33 vezes maior do o pacote de voz. Neste experimento também se verificou que o retardo relativo ao tráfego dos pacotes de voz é menor que 150 ms. A Tabela 5.2 mostra a variação do tráfego relativo a carga.

Tabela 5.2: Pacotes por segundos enviados e recebidos

Cliente A Pacotes/s	Gerador de Tráfego Pacotes/s	Cliente B Pacotes/s
74	265.43	340.36
74	504.34	588.29
74	883.79	965.56
74	1200.72	1271.31
74	1450.45	1516.45
74	2090.78	2139.35
74	4043.95	4106.46
74	4900.75	5008.26

Após a coleta dos dados, um modelo deve ser definido e validado. A validação consiste em comparar os resultados providos através do modelo e os mensurado no sistema real. A Figura 5.8 apresenta o modelo abstrato GSPN gerado que representa o sistema a ser analisado.

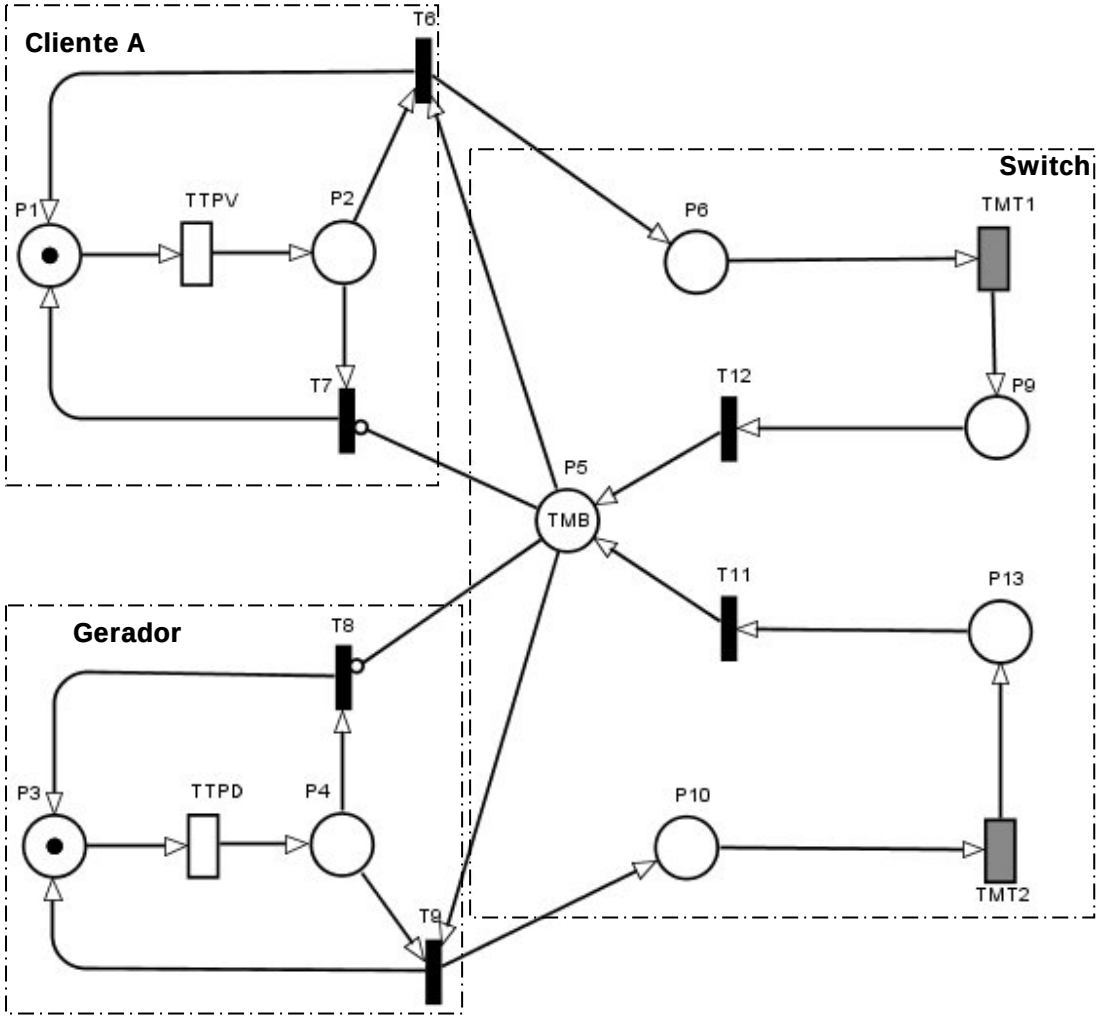


Figura 5.8 Modelo Abstrato

O modelo do cliente é composto por vários elementos: um lugar representado por P1 e uma transição que representa a transmissão do pacote de voz e que tem associado a ela o tempo de transmissão do respectivo pacote (TTPV). O arco inibidor saindo de P5 e chegando à transição T7, representa a inviabilidade de transmissão de pacotes de voz quando o *buffer* estiver cheio.

O gerador de tráfego de dados é representado pelo lugar P3 e uma transição que representa a transmissão do pacote de dados. O tempo atribuído a esta transição é o tempo de transmissão do respectivo pacote (TTPD). O arco inibidor de entrada da transição T8 representa a inviabilidade de transmissão de pacotes de dados quando o *buffer* respectivo estiver cheio.

As transições TMT1 e TMT2 representam distribuições expolinomiais (distribuição genérica associada a diversos tipos de distribuições)[3] . Por intermédio das estatísticas obtidas (média e desvio padrão), utiliza-se a técnica de *moment marching* para definir a distribuição adequada para o Modelo Abstrato. O coeficiente de variação foi de 0.107252553, a média de 0.000576948 e um desvio padrão de 0.0000618791755. Pelo coeficiente de variação a aproximação por fases deve ser por uma distribuição Hipoexponencial. A Figura 5.9 demonstra a aproximação para esta distribuição.

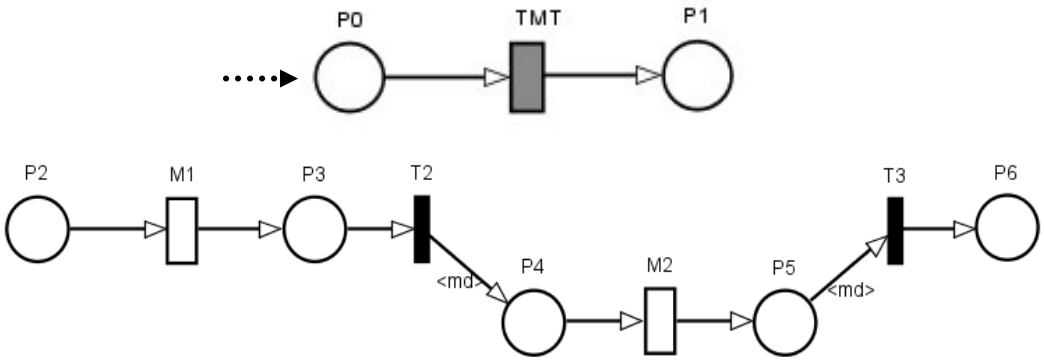


Figura 5.9 Modelo antes e depois da aproximação

Os valores obtidos de μ_1 , μ_2 e γ (Ver Capítulo 3) que representam o número de fases são os seguintes:

$$\mu_1 = 0.00056864$$

$$\mu_2 = 0.0000083077$$

$$\gamma = 86$$

Após gerar modelo refinado é preciso analisar as propriedades comportamentais e estruturais. Em função do resultado desta análise e da execução do *token game*, o avaliador pode: alterar o modelo para que venha ter propriedades que se julguem importantes (por exemplo, limitação), alterar a formulação das expressões que representam as métricas desempenho e escolher o tipo adequando de técnica para executar a avaliação (análise numérica ou simulação estocástica).

As propriedades comportamentais deste modelo em particular são: o modelo é limitado, é livre de *deadlock*, é reversível, pois é possível retornar a marcação inicial, que é alcançável de qualquer marcação do modelo. Com relação às propriedades estruturais o modelo é coberto por invariantes de lugar, ou seja, o modelo é limitado (espaço de estados finito) e consistente (propriedade necessária para *liveness*). É importante ressaltar que o *token game* ajuda identificar comportamentos indesejáveis do modelo. Pode-se verificar, por exemplo, se alguma transição pode ser habilitada antes de outra pré-determinada.

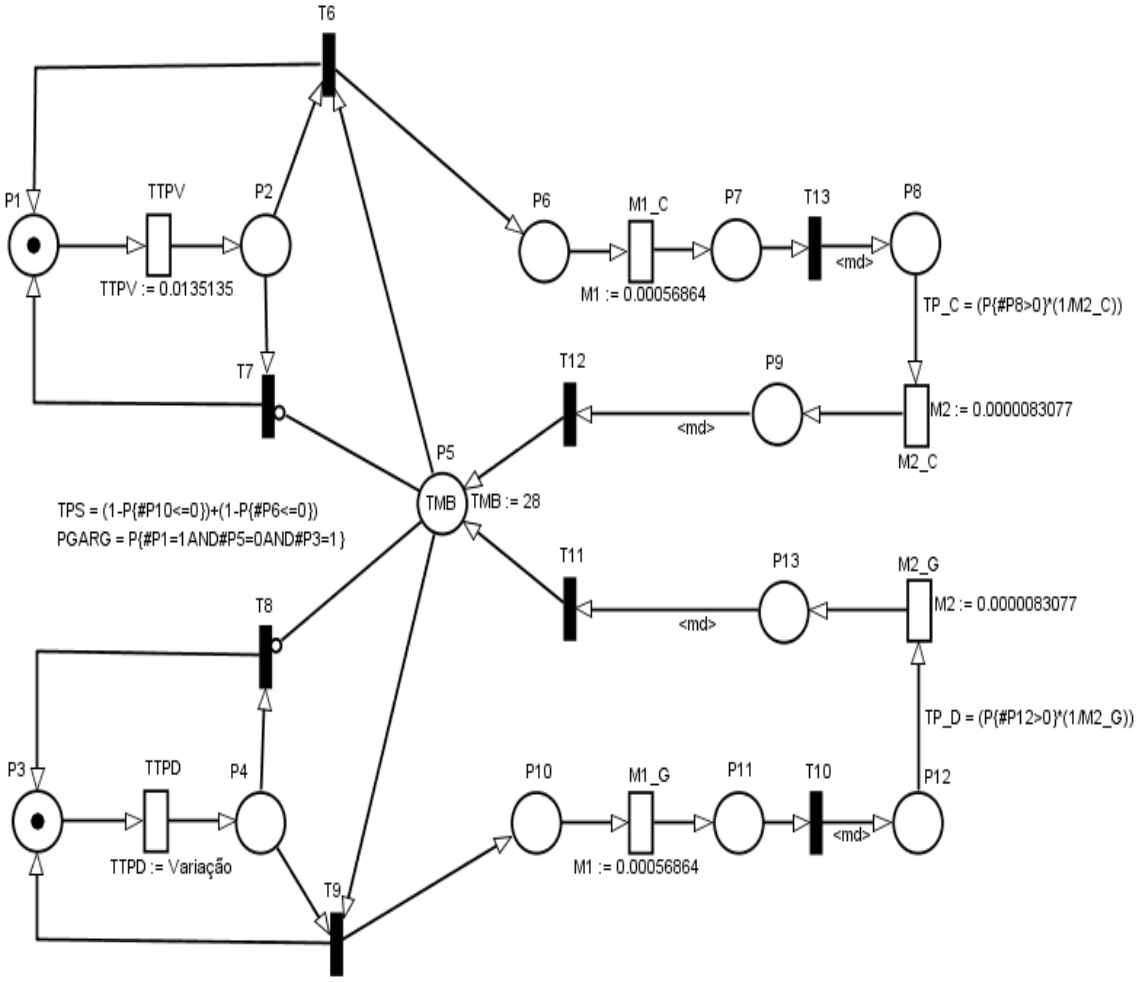


Figura 5.10 Modelo Refinado

Após a obtenção do modelo refinado, é necessário validá-lo. No modelo refinado (Figura 5.10), as marcas nos lugares P6, P7, P8 e P9 representam pacotes de voz. As transições M1_C e M2_C têm tempos referentes a μ_1 e μ_2. As marcas nos lugares P10, P11, P12 e P13 representam pacotes de dados. As transições M1_G e M2_G têm tempos referentes a μ_1 e μ_2, respectivamente.

No modelo não se usou lugares para o cliente B, pois os valores associados à saída do *switch* (saída do lugar P9 e P13) são de fato para o cliente B. A transição TTPD não possui um tempo de transmissão específico dado que esse tempo foi alterado para propósitos de validação, ou seja, alteramos TTPD conforme a variabilidade do tráfego gerado pelo gerador de tráfego.

Para o cliente de voz, o tempo de transmissão dos pacotes de voz (TTPV) corresponde a 0.0135135s. O tempo de transmissão dos pacotes de dados (TTPD) depende da variação do gerador de tráfego. A marcação do lugar P5 representa uma estimativa do tamanho médio do buffer (TMB).

O tamanho médio do buffer (TMB) depende da taxa total de transmissão (taxa de transmissão de voz + taxa de transmissão dos dados) e do tempo médio de transmissão (TMT). A taxa total refere-se à taxa de transmissão total por segundos sem descartes de pacotes.

TTOTAL = taxa de transmissão voz + taxa de transmissão de dados
= 74 + (4497) 2* 5.33*
= 48012 pps

Usa-se a lei de *Little* (Capítulo 3) para estimar o tamanho médio do buffer.

*TMB = TTOTAL * TMT*
*= 48012 * 0.00057694* ≈ 28

A probabilidade de gargalo é representada no modelo da seguinte forma:

PGARG = P{#P1=1AND#P5=0AND#P3=1}

A probabilidade de gargalo acontece quando:

- O lugar P1 tem uma marcação.
- O lugar P5 não tem nenhuma marcação.
- O lugar P3 tem uma marcação.

O *throughput* do sistema representa a soma do *throughput* de voz mais a soma do *throughput* de dados. Para calcular o *throughput* precisa-se obter a probabilidade de um determinado lugar, correspondente ao pacote de dados

ou de voz, ter uma marcação ou mais e dividi-lo pelo seu tempo de transmissão. Esta métrica é representada no modelo pela expressão:

$$TP = \frac{P(\#P8 > 0)}{M2\_C} + \frac{P(\#P12 > 0)}{M2\_G}$$

Nesse caso para calcular somente o *throughput* de voz representamos da seguinte forma:

$$TP\_C = \frac{P(\#P8 > 0)}{M2\_C}$$

Para avaliar a utilização máxima do sistema, e sabendo que o pacote é de dados ou de voz, verifica-se a probabilidade do lugar P10 (pacote de dados) ser menor ou igual a zero e a probabilidade do lugar P6 (pacote de voz) ser menor ou igual a zero. Se ambas as probabilidades forem iguais a 1, o sistema não está sendo utilizado. Vale observar que o modelo (Figura 5.10) é concorrente, portanto, a utilização máxima do sistema, representa o percentual dele não está sendo utilizado (1 representando 100%), menos o percentual do sistema está ocioso. Essa métrica é representada da seguinte forma no modelo:

TPS = (1-P{#P10<=0})+(1-P{#P6<=0})

O modelo refinado foi criado e agora é preciso validá-lo.

Na avaliação estacionária na qual avalia-se o modelo, percebe-se que o número de fases é muito grande e por isso o tempo de resposta na obtenção dos resultados é muito longo, além de consumir um grande recurso da máquina que a executa. Nesse caso, verifica-se que um número de fases menor (8) no modelo não interfere nos resultados e ganha-se tempo na avaliação.

A análise estacionária realizada no modelo manteve os valores correspondentes a todas as transições com exceção da transição TTPD, que teve variações entre 265.43 e 4900.75 pps, ou seja, entre os tempos de 0.00376 e 0.000204s. Utiliza-se o intervalo acima porque na medição o valor correspondente a degradação da qualidade da voz pertence ao intervalo. Os resultados mostrados na Tabela 5.3 são os valores obtidos através da medição e do modelo.

O objetivo é validar se a taxa de envio dos pacotes do cliente A e do gerador na medição corresponde à taxa de pacotes recebidos pelo cliente B na medição e na modelagem.

Tabela 5.3: Pacotes/s recebidos pelo cliente B na medição (MED) e no modelo (MOD)

Pacotes/s Enviados Cliente A	Pacotes/s Enviados Gerador	Pacotes Recebidos Cliente B Med	Pacotes Recebidos Cliente B Mod
74	265.43	340.36	339.95
74	504.34	588.29	578.36
74	883.79	965.56	957.86
74	1200.72	1271.31	1274.71
74	1450.45	1516.45	1523.21
74	2090.78	2139.35	2164.47
74	4043.95	4106.46	4041.90
74	4900.75	5008.26	4844.54

Um teste t-emparelhado [43] foi conduzido e o intervalo das médias das diferenças contém o zero. O resultado mostra que não se têm evidência para refutar o modelo concebido como sendo adequado para representar o sistema. Esta afirmação é feita com um nível de confiança de 95% e 7 graus de liberdade. Os resultados do teste t-emparelhado estão representados na Tabela 5.4. Usa-se a ferramenta xlstat (http://www.xlstat.com) para executar os procedimentos do teste t-emparelhado.

Tabela 5.4 Teste t-emparelhado

	Quantidade da Amostra	Média	Desvio Padrão	Erro Padrão
Cliente B MED	8	1992	1694	599
Cliente B MOD	8	1966	1643	581
Diferença	8	26,4	61,2	21,7

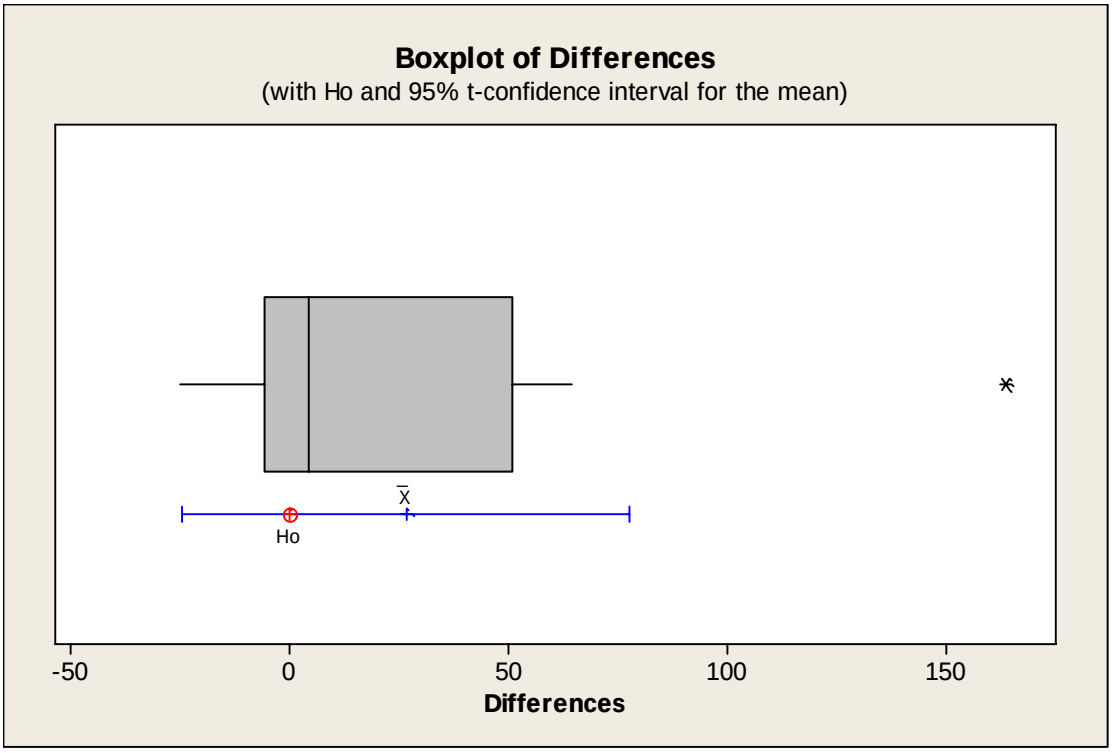


Figura 5.11 Gráfico de caixa das diferenças (cliente B MED e cliente B MOD)

As Figuras 5.11 e 5.12 apresentam o gráfico de caixa (*Box plot*) das diferenças entre cliente B (MED) e o cliente B (MOD). O diagrama de pontos (*dot plot*), respectivamente, mostra que a mediana (4,05), tendência central da distribuição, está um pouco acima de zero e 75% dos valores estão acima de -4,24, portanto, os valores medidos e os obtidos com o modelo estão bem próximos. Ambos os gráficos mostram visualmente que os resultados das diferenças estão dentro do intervalo de confiança, e apenas um ponto se encontra fora do intervalo. Utilizou-se o Minitab (http://www.minitab.com) para o cálculo das estatísticas.

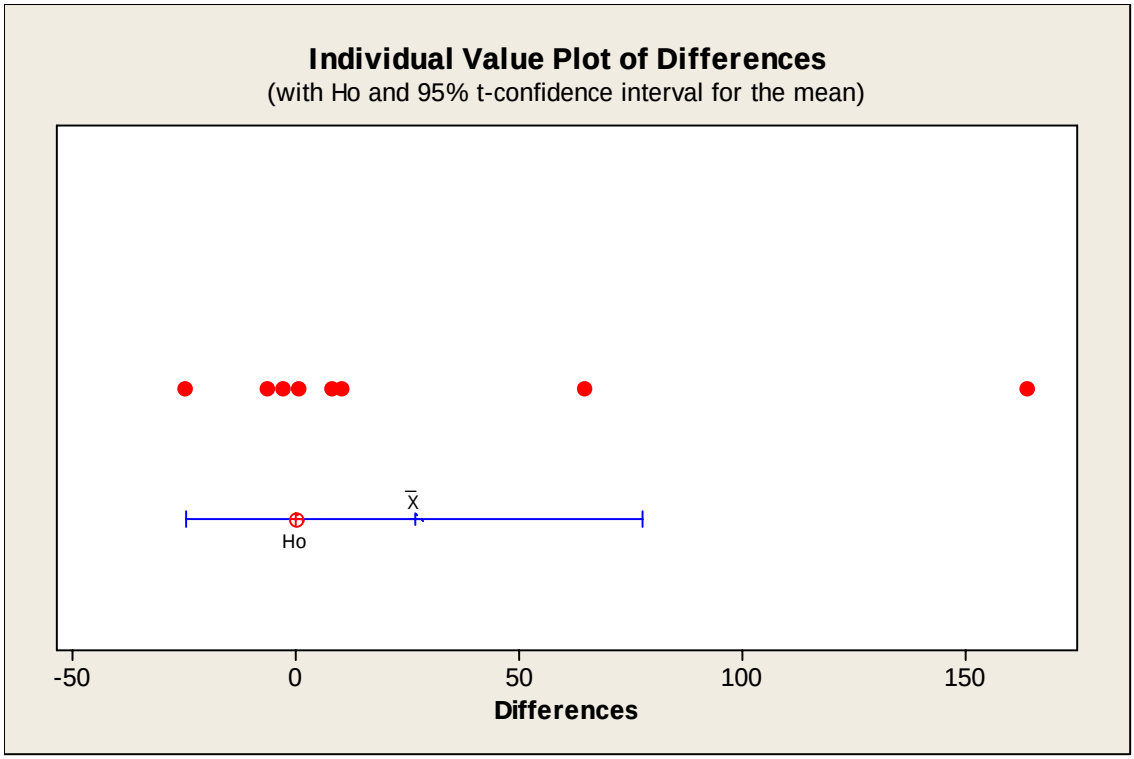


Figure 5.12 Valores das diferenças individuais

A medição mostrou que o sinal de voz sofre interferência significativa do tráfego de dados até a freqüência (de transmissão de pacotes de dados - TTPD) que corresponde ao tempo de transmissão de um pacote próximo a 1,6 ms (o tráfego de dados é criado e configurado no gerador de tráfego). Esta observação é corroborada pela ausência de perdas de pacotes. O modelo foi avaliado considerando-se tempos de transmissão de pacotes de dados com valores inferiores e superiores ao limite observado.

O gráfico da Figura 5.13 ilustra o *throughput* de voz recebido pelo cliente B. Os valores na abscissa correspondem ao TTPD que varia de 0,1 ms a 11,5 ms.

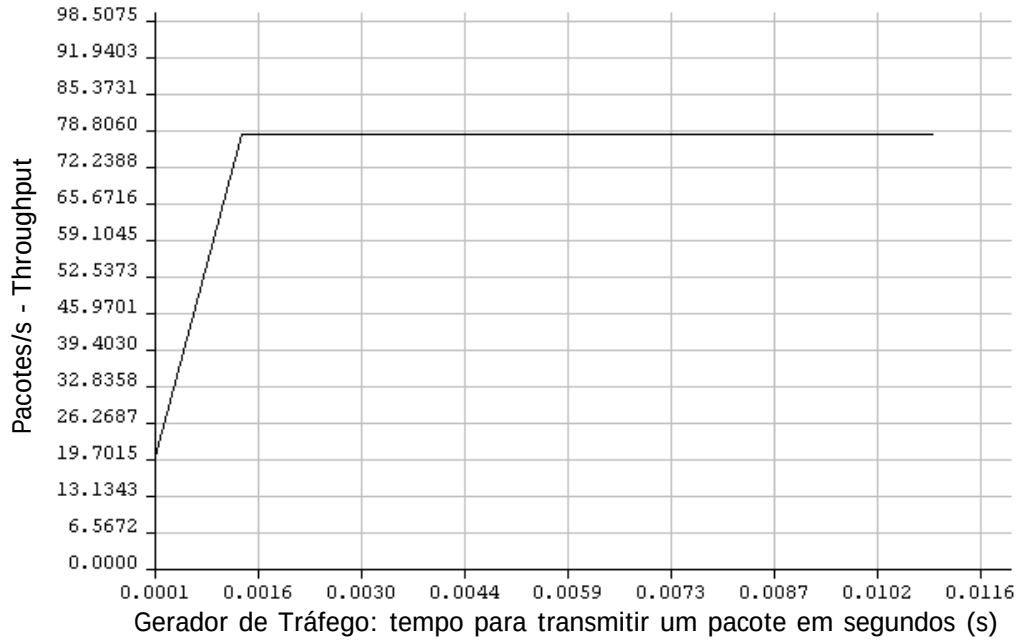


Figura 5.13 *Throughput* de voz recebido pelo cliente B

Pode-se verificar através do gráfico da Figura 5.13 que o *throughput* de voz calculado através do modelo é de 77.4 pps quando o tráfego de dados é de 1,6 ms. Na medição, o valor obtido foi 74 pps. A probabilidade de gargalo verificada neste contexto (TTPD superior a 1,6 ms) foi de 0% e o percentual de utilização dos recursos foi 0.04%. Ou seja, o recurso está ocioso e conseqüentemente os pacotes de voz chegam sem perdas, pois a quantidade de enviados corresponde à quantidade de pacotes recebidos. Os resultados apresentados na Figura 5.14 mostram uma variação significativa do *throughput* de voz em função do TTPD.

A avaliação do *throughput,* considerando o TTPD variando no intervalo 0,1 ms e 1,5 ms, mostra que o número de pacotes de voz por segundos diminui consideravelmente a medida que se aumenta o tráfego de dados. O *throughput* de voz, quando o TTPD é 0,1 ms, corresponde a 20 pps. Da mesma forma, *throughput* de voz é 77.4 pps quando o TTPD é igual a 1,5 ms (Figura 5.14). Em outras palavras, a quantidade de pacotes recebidos pelo cliente B diminui quando se aumenta o tráfego de dados.

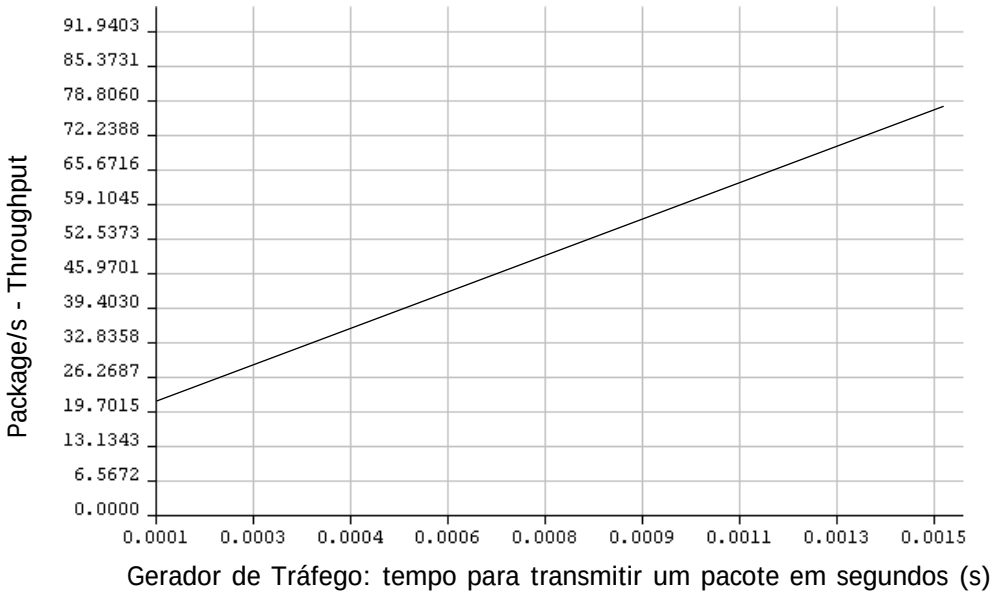


Figura 5.14 Throughput de voz recebido pelo cliente B (entre 0.0001 e 0.0015s)

Se a quantidade de pacotes de voz recebidos pelo cliente B diminui quando se aumenta o tráfego de dados, então a probabilidade de gargalo (PGARG) aumenta. Se a PGARG aumenta, a comunicação de voz tem impactos de desempenho (degradação na qualidade da voz) por perdas ou atrasos dos pacotes de voz. A PGARG calculada é 85% quando TTPD é 0,1 ms, e 39%, para o tempo de 1,5 ms (Figura 5.15).

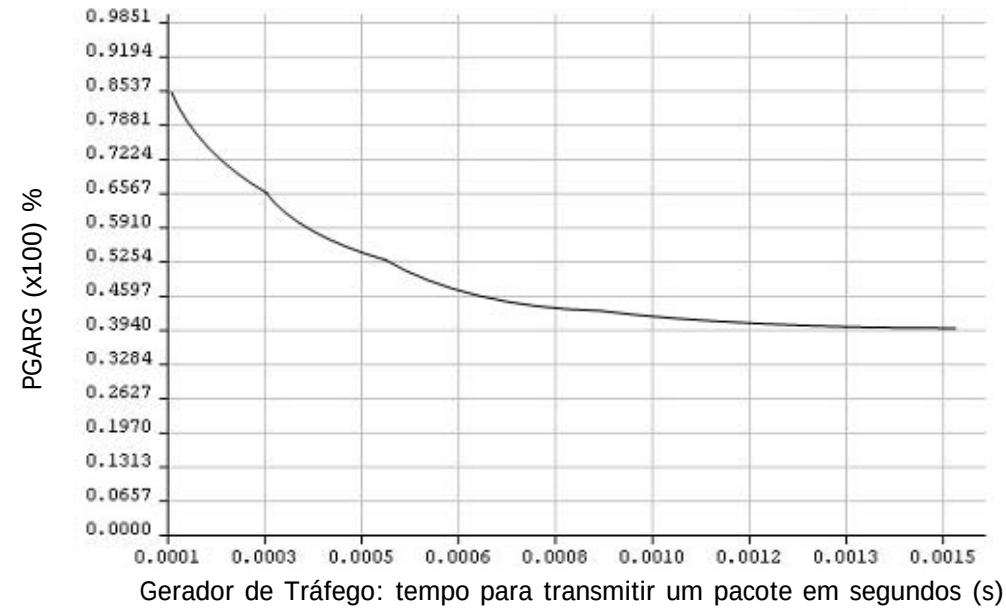


Figura 5.15 PGARG sobre a variação de TTPD

A Figura 5.16 mostra a percentagem de utilização do sistema sobre a variação da carga de tráfego.

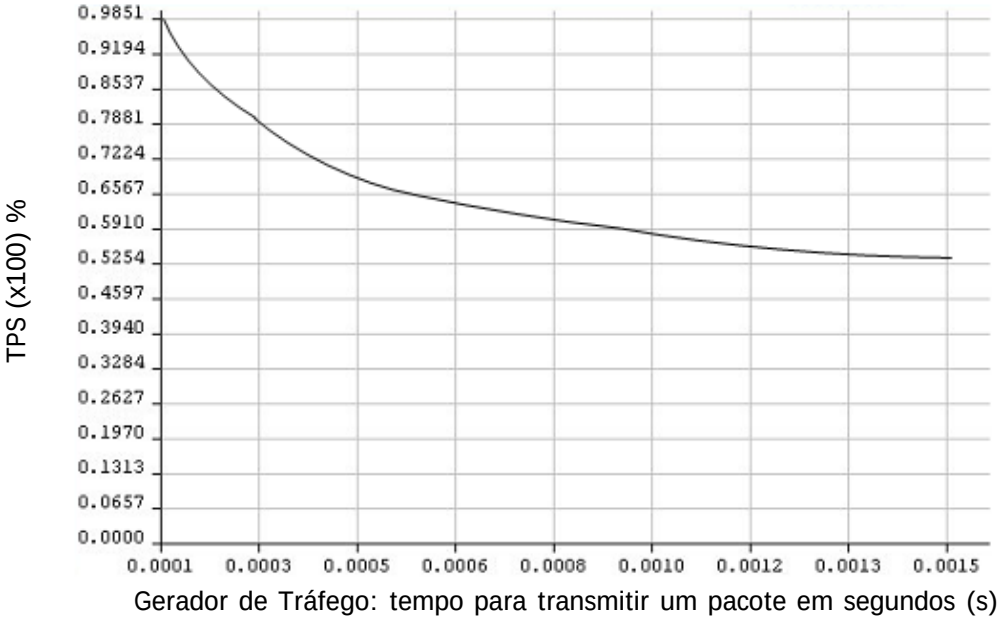


Figura 5.16 Utilização do Sistema(TPS) sobre a variação do Gerador de Tráfego

Pode-se constatar no gráfico da Figura 5.16 que o sistema está sendo bastante requisitado (98%) quando o tempo de transmissão no gerador de tráfego é de 1 ms. No tempo de 1,5 ms (gerador de tráfego) o sistema trabalha com relativa disponibilidade (53 % de utilização).

Capítulo 6 - Estudo de Caso

Uma característica comum do uso das aplicações de VoIP é sua utilização em um meio compartilhado. Em outras palavras é utilizar os recursos de VoIP com outras aplicações simultaneamente.

Avaliar a comunicação de voz com outras aplicações exige o conhecimento das características de desempenho de todos os componentes envolvidos. É preciso medir o desempenho dos componentes em função do tráfego para se conhecer a largura de banda utilizada. Conhecendo-se o nível de utilização dos recursos associado comunicação de voz, é necessário identificar os limites suportados por todos os componentes para os demais tipos de tráfego de forma que a qualidade do tráfego de voz não seja afetada. A Figura 6.1 mostra o contexto deste estudo de caso.

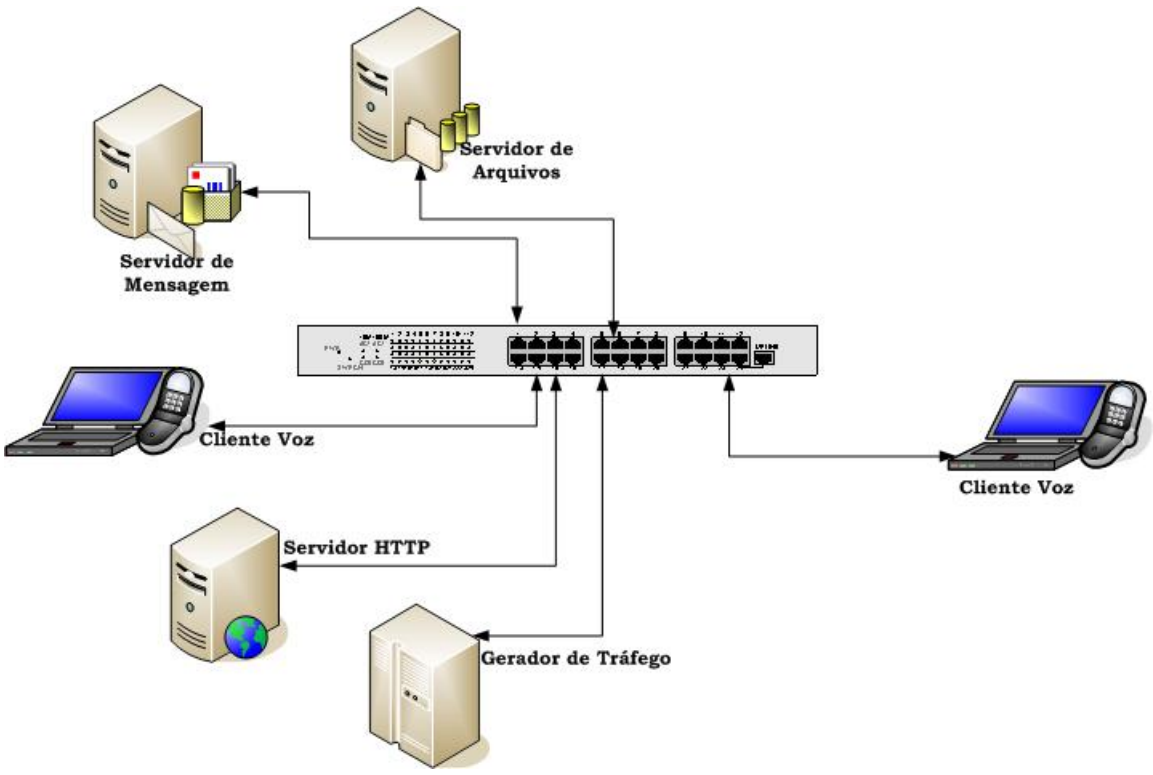


Figura 6.1 Estudo de Caso

A Figura 6.1 demonstra um cliente de voz comunicando-se com outro cliente de voz, com um servidor de aplicações Web (HTTP), em acesso a arquivos do servidor de arquivos e consultando o correio eletrônico. Além desse tráfego, coloca-se um gerador de tráfego para simular situações nas quais temos um tráfego adicional. Esse tráfego pode ser associado alguma aplicação específica, atualização de determinado *software* ou até mesmo tráfego indevido (tráfego de aplicações não necessárias).

No estudo de caso são detalhados os recursos de transmissão dos componentes envolvidos e, se baseando na metodologia usada neste trabalho e no modelo de desempenho, cria-se alguns cenários de interesse para se avaliar o desempenho das aplicações de VoIP nessa infra-estrutura.

Os recursos (desktops, laptops, servidores e *switches*) utilizados neste experimento foram isolados, de forma que se tivesse um ambiente controlado, para avaliação. Nenhum outro tipo de tráfego foi gerado no período em que os experimentos foram realizados. As computadores usados neste estudo de caso foram Ahtlon XP 2500+ com 1 Gb de memória RAM.

Os valores mensurados associados às métricas de interesse foram obtidos para cada componente de forma isolada. A coleta dos dados de cada componente foi obtida por um período de 5 minutos. Observou-se que não era necessário um período maior de medição, pois, neste caso, a taxa de transmissão (pacotes por segundos), não se altera em um período de coleta maior.

O cenário apresentado na Figura 6.2 demonstra o ambiente lógico de avaliação deste estudo de caso. Os quadrados simbolizam os pacotes de voz e os círculos simbolizam os demais tipos de pacote. Este estudo de caso visa responder perguntas como: qual é o impacto na voz, se o cliente faz o *download* de diversos arquivos? Qual é o tráfego e o impacto na comunicação de voz se o cliente consulta sua caixa de mensagens? Qual é o impacto de uma determinada aplicação web na comunicação de VoIP? Qual é o impacto na comunicação de voz, quando se insere uma nova aplicação? Como o

administrador de rede pode configurar o sistema para que se evite problemas de desempenho na comunicação de voz?

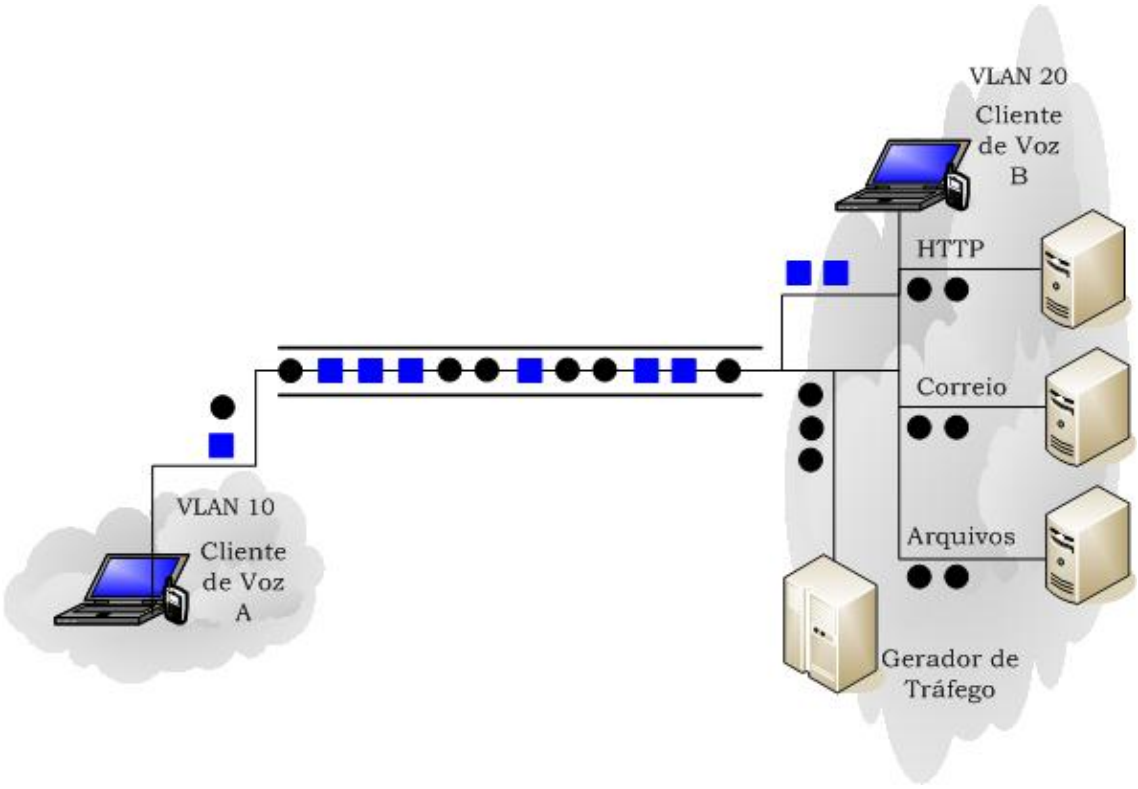


Figura 6.2 Cenário de Avaliação

A Figura 6.2 nos mostra os clientes de voz em redes isoladas. O cliente A pertence à rede 182.127.63.0/24 (chamada de vlan 10) e o cliente B pertence à rede 182.127.62.0/24 (chamada de vlan 20). O gerador de tráfego, o servidor de aplicações, o servidor de arquivos e o servidor de mensagens pertencem a vlan 20. Duas interfaces de roteamento foram configuradas para cada rede individualmente.

O tráfego referente ao cliente de voz foi medido anteriormente. As explicações referentes às medições e processos de coleta estão descritos no Capítulo 5. O codificador e decodificador (codec) utilizado pela aplicação de voz foi o G.711 e a ferramenta usada pelo cliente de voz foi o *MyPhone*.

O tempo de envio dos pacotes de voz medido corresponde a 74 pps (1 pacote para cada 1.35 x 10^{-2} s) e o tamanho médio dos pacotes de voz 284 bytes. Para padronizar o tamanho e o tempo de transmissão dos pacotes, a unidade de tráfego foi uniformizada por uma unidade básica (pacote de voz e tempo de transmissão do pacote de voz). A ferramenta utilizada pelo gerador de tráfego foi a mesma usada no modelo de desempenho (Capítulo 5), TfGen. O tempo médio de transmissão do gerador de tráfego não foi definido porque depende do cenário que será analisado.

O servidor de aplicação HTTP específico deste estudo transmite pacotes a uma taxa constante. Porém, há períodos de silêncio relativo ao tráfego gerado por essa aplicação. Essa característica acontece porque a aplicação não realiza transações com o cliente de forma constante. Por exemplo, se o cliente está lendo ou digitando alguma informação na aplicação, não existe um tráfego considerável gerado por esta aplicação. O período médio para se iniciar um período de ociosidade pode ser avaliado, assim como o seu impacto no desempenho do sistema. Neste estudo, considerou-se 0.5s. De maneira similar, neste estudo, considerou-se o período de silêncio igual 0.5s. O tamanho médio dos pacotes de transmissão enviados pelo servidor de aplicação é de 1050.8 bytes. Para encontrar o tempo médio de transmissão, divide-se o tamanho do pacote HTTP (1050.8 bytes) pelo tamanho do pacote de voz (284) e o resultado (3.70) é multiplicado pelo tempo de envio dos pacotes de voz (3.70 * 1.35 x 10^{-2}). Portanto, o tempo médio de envio do pacote http é de 0.05s.

Nesse estudo de caso o cliente tem acesso ao correio eletrônico. Portanto, existe um tráfego constante devido a verificação de mensagens entre o cliente e o servidor de mensagens. O tamanho médio do pacote que é transmitido é de 1262 bytes e o tempo médio de transmissão de 0.06s.

O servidor de arquivos que o cliente acessa possui diversos arquivos armazenados disponíveis para *download*. Inicialmente o servidor de arquivos está limitado para permitir o download de apenas 4 arquivos por vez. O servidor possui 50 arquivos com tamanho médio de 2103 bytes. O tempo médio de transmissão é de 0.1s para cada arquivo.

Para o *switch*, o tempo médio de transmissão (TMT) foi o mesmo, pois se utiliza o mesmo equipamento. Portanto, o coeficiente de variação (0.107252553), a média (0.000576948 s) e o desvio padrão (0.0000618791755 s) foram os mesmos. Da mesma forma, segundo o método de *moment matching,* o inverso do coeficiente de variação sugere que se utilize uma distribuição Hipo-exponencial como a distribuição expolinomial que aproxima os dados mensurados. Os valores de μ_1, μ_2 e γ que representam os valores médios associados às fases exponenciais e o número de fases são os mesmos do modelo de desempenho (Capítulo 5).

$$\mu_1 = 0.00056864$$
$$\mu_2 = 0.0000083077$$
$$\gamma = 86$$

A estimativa do tamanho do buffer também não foi alterada, pois se utiliza o mesmo *switch*. O valor estimado permanece 28.

Após as medições realizadas, o modelo refinado (que será avaliado) é obtido a partir do modelo abstrato. O modelo abstrato está representado na Figura 6.3.

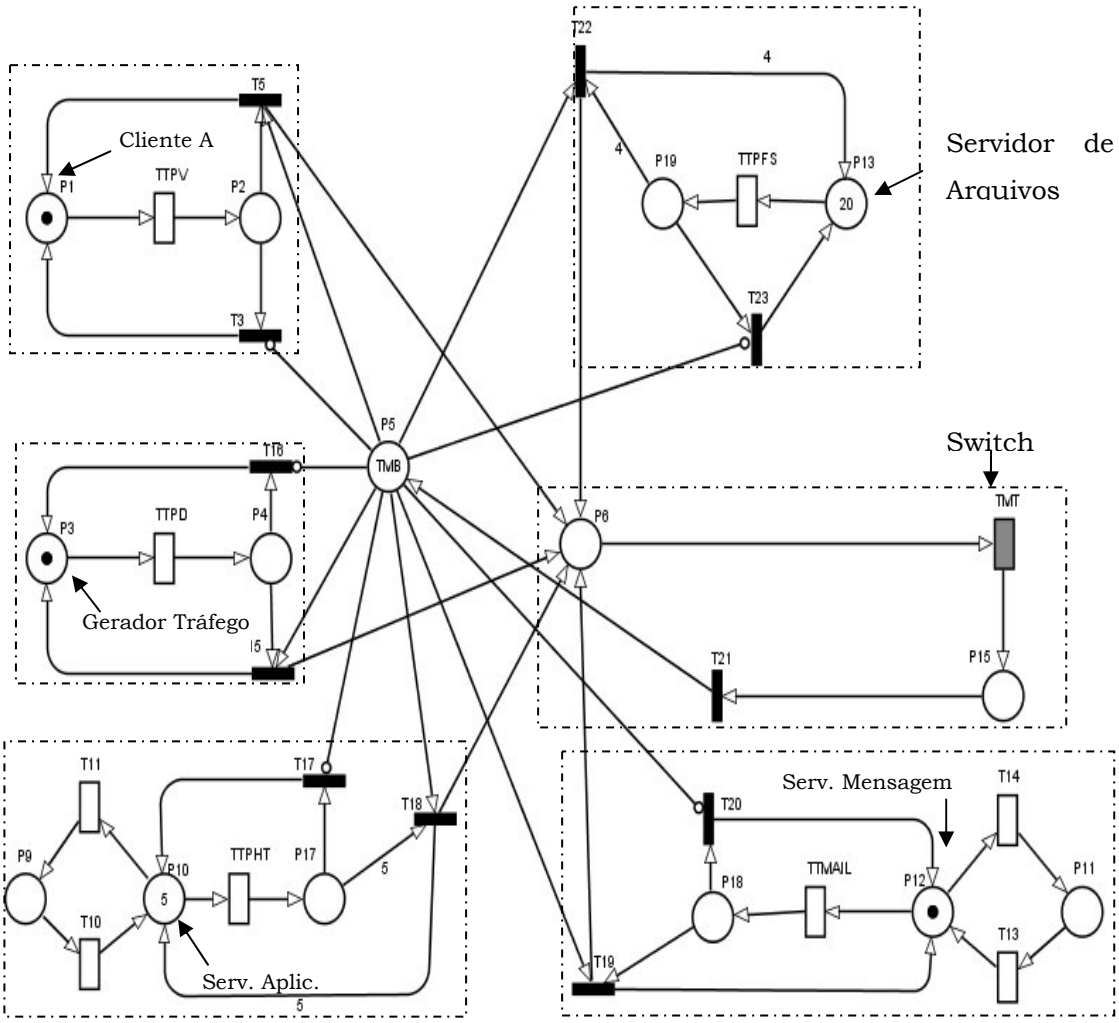


Figura 6.3 O modelo abstrato

A Figura 6.3 mostra o modelo abstrato com todos os componentes definidos no cenário de avaliação (Figura 6.2). Desmembrando esse modelo abstrato, cada componente do sistema é detalhado abaixo.

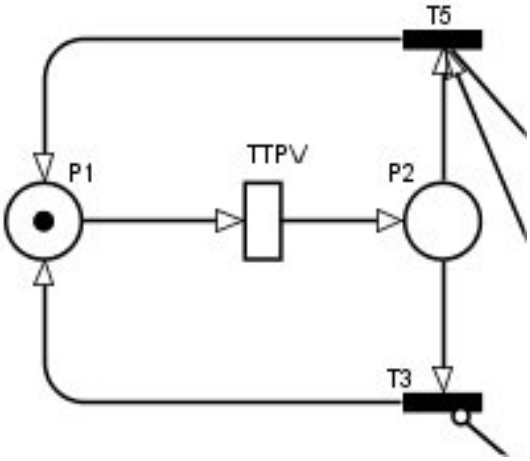


Figura 6.4 Modelo do Cliente de Voz

O cliente de voz é representado pela rede da Figura 6.4. O tempo de transmissão do pacote de voz é atribuído à transição TTPV. A marcação no lugar P1 representa um pacote de voz.

A Figura 6.5 apresenta a rede que representa o gerador de tráfego. A transição TTPD representa a transmissão do pacote de dados. O tempo para transmissão de um pacote de dados é atribuído a esta transição. A marcação no lugar P3 representa o pacote de dados.

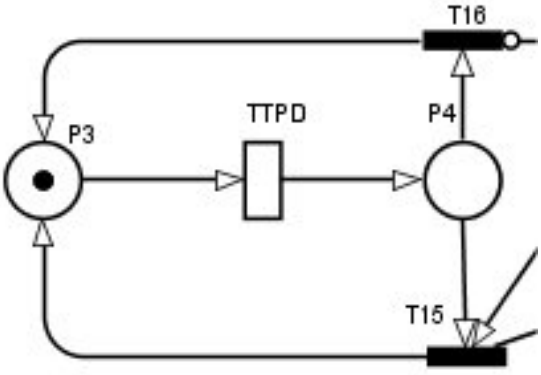


Figura 6.5 Modelo do Gerador de Tráfego

A Figura 6.6 mostra o modelo do servidor de arquivos. O número de marcas no lugar P13 representa a quantidade de arquivos no servidor. A transição TTPFS representa o envio dos pacotes e a multiplicidade dos arcos saindo do lugar P19 e da transição T22 representam a quantidade de pacotes

transmitidos simultaneamente. O tempo de transmissão dos pacotes é atribuído à transição TTPFS.

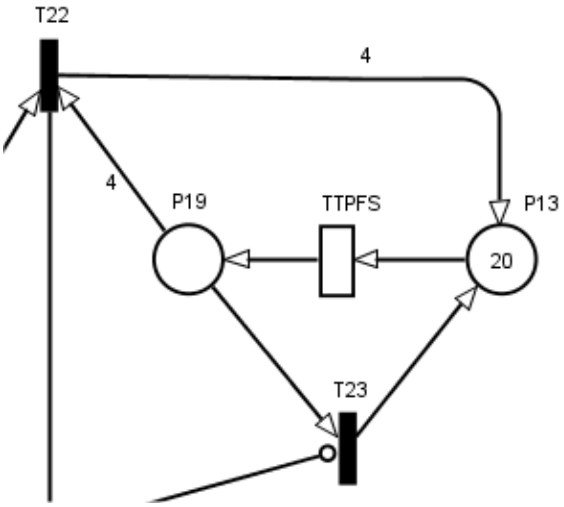


Figura 6.6 Modelo do Servidor de Arquivos

A Figura 6.7 apresenta a rede que representa o servidor de aplicações. O tempo de transmissão dos pacotes do servidor de aplicação é atribuído à transição TTPHT, que representa a transmissão de pacotes. À transição T11 associa-se o tempo médio em que ocorrem os períodos de silêncio. A duração dos períodos de silêncio é atribuída à transição T10.

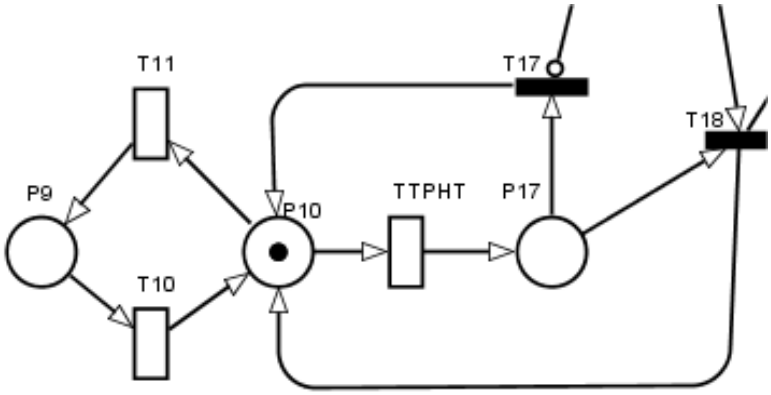


Figura 6.7 Modelo do Servidor de Aplicações

A Figura 6.8 mostra o modelo que representa o servidor de mensagens. O tempo médio de requisição é associado à transição TTMAIL.

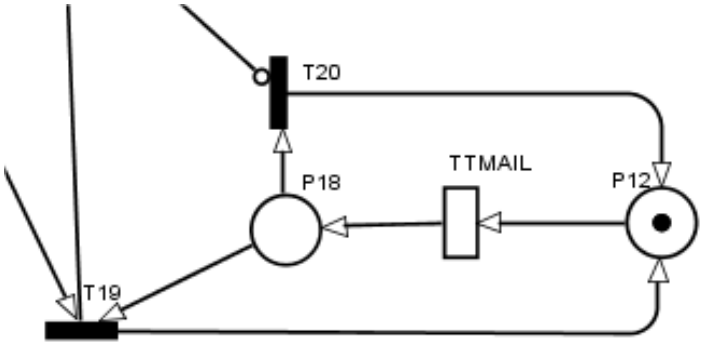
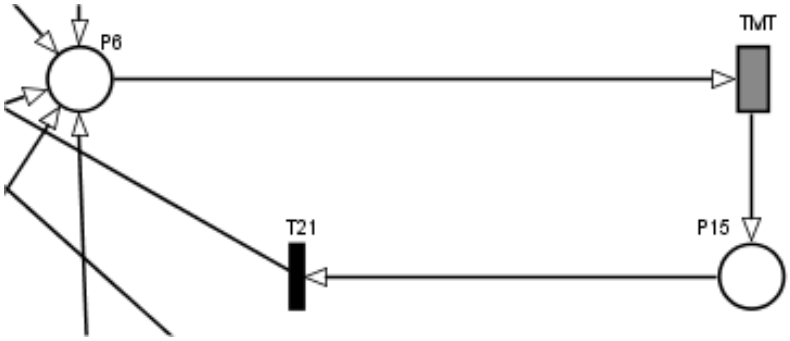
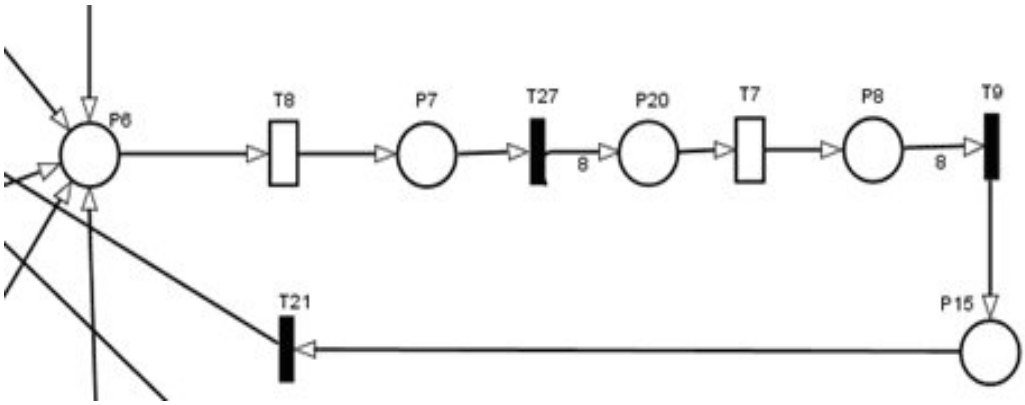


Figura 6.8 Modelo do Servidor de Mensagens

A transição TMT representa a distribuição expolinomial. As estatísticas para se encontrar a média e o desvio padrão foram obtidas no modelo de desempenho (Capítulo 5) e através da técnica de *moment marching* definiu-se a distribuição hipo-exponencial como a mais adequada para representar a transição. A Figura 6.9 mostra a representação da transição no modelo abstrato e o respectivo refinamento desta transição através de uma sub-rede que representa uma distribuição hipo-exponencial.



(a)



(b)

Figura 6.9 Refinamento da transição TMT - Tempo Médio de Transmissão (*Switch*)

Os valores obtidos de μ_1 e μ_2 são atribuídos às transições T8 e T7 da Figura 6.9(b). A multiplicidade (γ) dos arcos que conectam a transição T27 e do lugar P8 e o lugar P8 e a transição T9 é representada na Figura 6.9(b). A semântica de serviço das transições T7 e T8 é SSS (*single server semantics*)[10] .

As propriedades comportamentais e estruturais são analisadas após a geração do modelo abstrato. Com base nessa análise e da execução do *token game,* na qual se identificam eventuais propriedades indesejáveis, o avaliador pode reformular o modelo de acordo com as propriedades que se julgue importante e redefinir as expressões que representam as métricas de desempenho escolhendo o tipo de técnica adequada para executar a avaliação (análise numérica ou simulação estocástica).

O modelo gerado neste estudo de caso tem as seguintes propriedades comportamentais: o modelo é limitado, é livre de *deadlock,* reversível e também conservativo. Com relação às propriedades estruturais o modelo é coberto por invariantes de lugar, ou seja, o modelo é estruturalmente limitado (espaço de estados finito) e consistente (propriedade necessária para *liveness*).

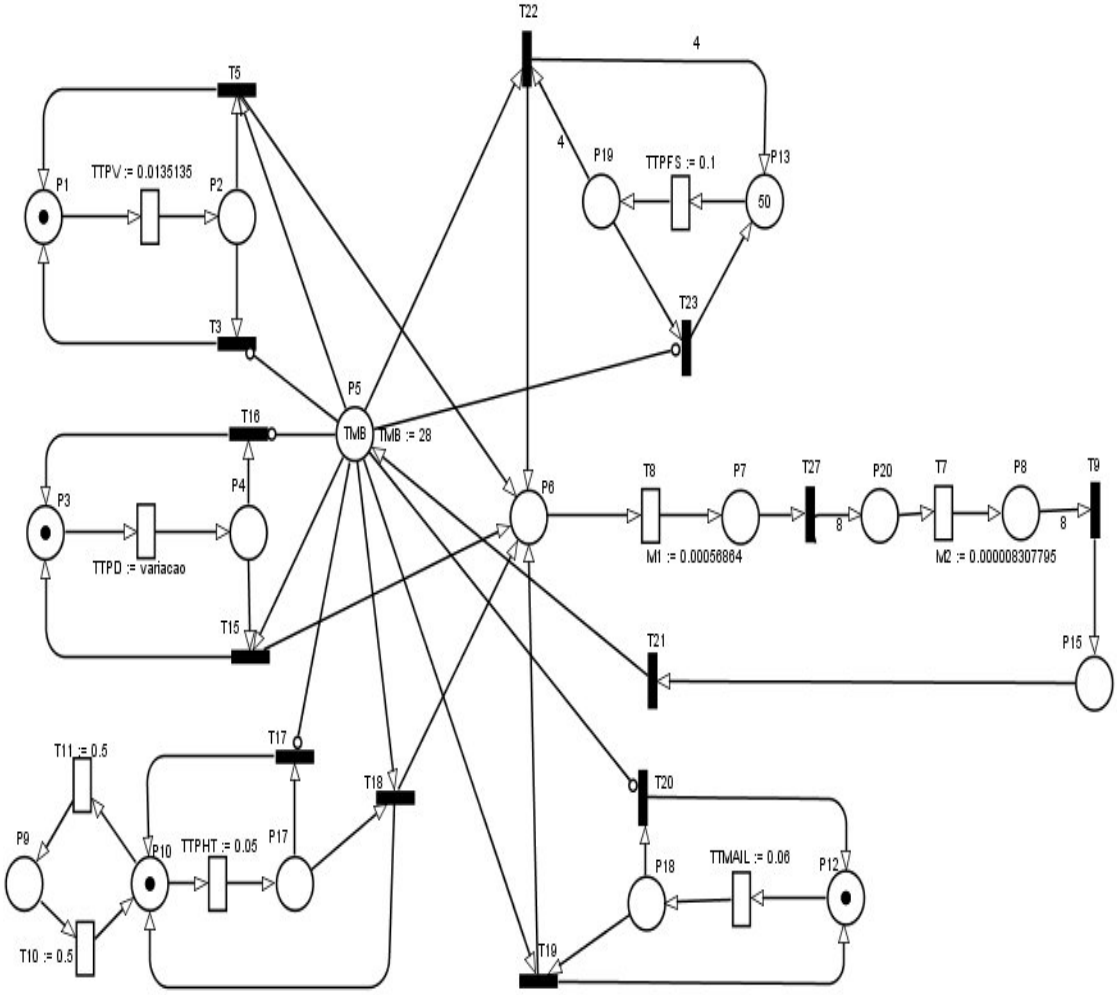


Figura 6.10 Modelo Refinado

A Tabela 6.1 mostra as características das transições temporizadas. Todas as transições temporizadas, com exceção da transição T10 e T11, têm semântica *single-server*. As transições T10 e T11 têm *infinite-server semantics*.

Tabela 6.1 Características das Transições Temporizadas

Transição	Tempo Associado as Transições (s)	Semântica
TTPV	0.0135135	*single-server*
TTPD	*****	*single-server*
TTPHT	0.05	*single-server*
TTMAIL	0.06	*single-server*
TTPFS	0.1	*single-server*
T7	0.00056864	*single-server*
T8	0.0000083077	*single-server*
T10	0.5	*infinite-server*
T11	0.5	*infinite-server*

Os pesos e as prioridades das transições imediatas do modelo refinado não foram alterados, ou seja, todas as transições imediatas permanecem com os mesmos pesos e prioridades dos atribuídos no modelo abstrato. Após a obtenção do modelo refinado, é necessário validá-lo. Os resultados da validação podem ser vistos no Capítulo 5.

Diversos cenários de avaliação são criados para avaliar o desempenho da aplicação de VoIP. As seguintes métricas são utilizadas para analisar o modelo:

- *Throughput* (essa métrica representa o *throughput* total do sistema).

$$TP = \frac{\mathrm{P}(\#\mathrm{P20} > 0)}{\mathrm{M2}}$$

- A utilização dos recursos (essa métrica representa a percentagem de utilização do sistema).

TPS = 1-P{#P6<=0}

- Probabilidade de Gargalo (a probabilidade de gargalo varia dependendo da disponibilidade do sistema em transmitir o pacote de voz).

  PGARG = P{#P1=1AND#P5=0AND#P3=1}

Baseado no modelo refinado e nos parâmetros inseridos foram definidos alguns cenários relevantes para avaliação.

Cenário 1

O primeiro cenário concerne a avaliação do desempenho da aplicação de voz variando-se o tráfego de dados (realizado através do gerador de tráfego). O tempo de transmissão de todos os outros componentes não foram alterados, ou seja, as medidas relativas a todos os componentes são inseridas no modelo e se variou apenas o tempo associado à transição temporizada correspondente ao gerador de tráfego.

Análise estacionária foi conduzida variando-se o tempo da transição TTPD, que corresponde ao tempo de transmissão dos pacotes do gerador de tráfego. Os limites do tempo associados à transição TTPD são 0,120 ms e 12,0 ms. Utiliza-se esse intervalo, pois se verifica um acréscimo significativo na probabilidade de gargalo (Figura 6.11). A Figura 6.11 mostra a probabilidade de gargalo em função da variação no tempo de transmissão dos pacotes do gerador.

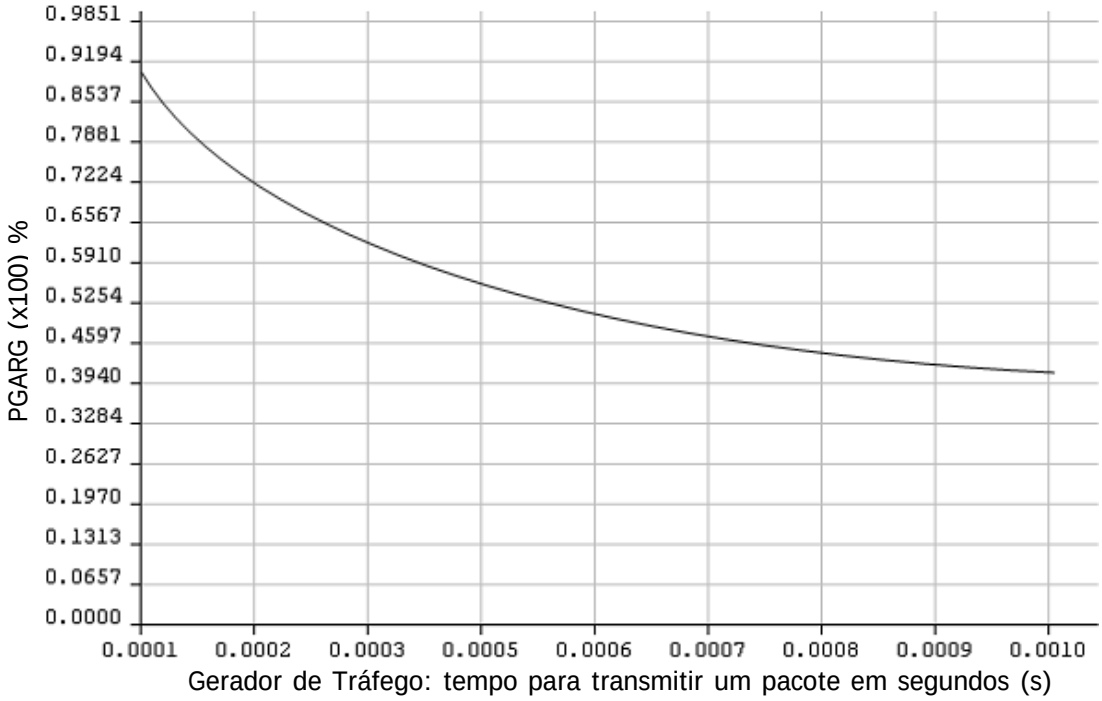


Figura 6.11 Probabilidade de gargalo sobre TTPD

Pode-se verificar através do gráfico da Figura 6.11 que a probabilidade de gargalo diminui quando o tráfego de dados diminui. Neste cenário a PGARG atinge 89% quando o TTPD é de 0,1 ms e 42% para o tempo de 1,0 ms. Considerando o modelo de desempenho (Capítulo 5), observa-se uma probabilidade de gargalo muito alta para o tempo de 0,1 ms, na qual se pode afirmar a degradação na qualidade da voz. No entanto, para o tempo de 1,0 ms, a PGARG é aceitável para que a comunicação de voz não tenha impactos de desempenho.

A Figura 6.12 mostra a percentagem de utilização do sistema sobre a variação da carga de tráfego. Pode-se observar no gráfico da Figura 6.12 que o sistema está sendo bastante utilizado, chegando a 98,5% no tempo de 0,120 ms.

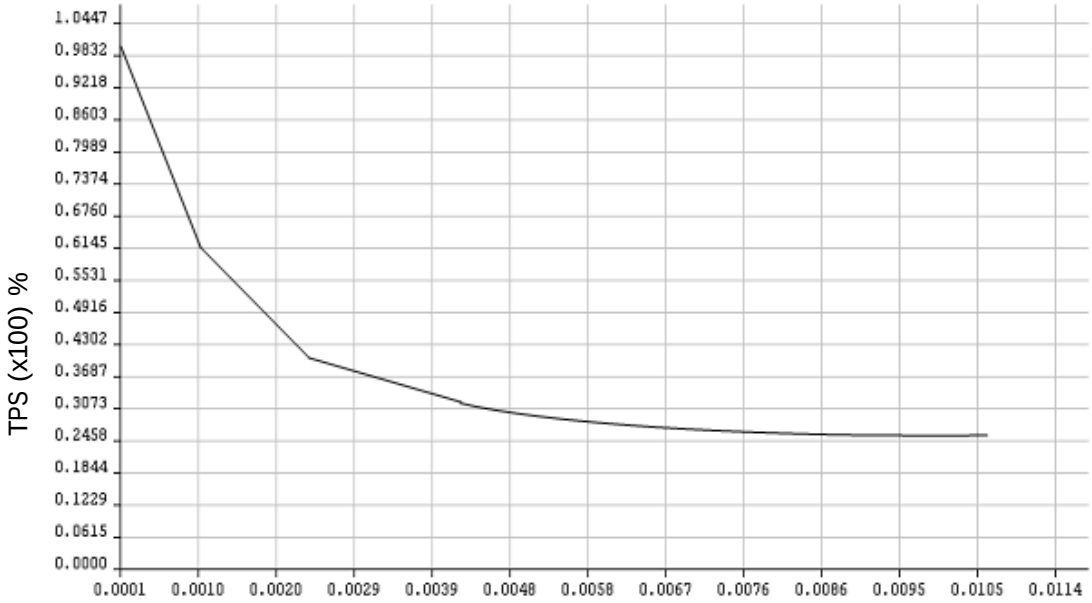


Figura 6.12 Utilização máxima dos recursos sobre TTPD

A utilização do sistema fica abaixo de 25% para TTPD igual a 0,0102 s, o que demonstra um baixo nível de utilização do sistema, quando se consideram estas condições. Embora não permaneça constante, o nível de utilização tem apenas uma variação suave quando o tempo de transmissão dos dados supera 0,0040s.

A avaliação de cenários permite que os administradores de rede configurem os sistemas de forma que se evite degradação na comunicação de VoIP. Os administradores podem definir alarmes que informem quando a probabilidade de gargalo ultrapassa limites estabelecidos ou quando o *throughput* estiver próximo de um determinado limiar. Neste estudo, sugere-se 50% como um limite associado a probabilidade de gargalo, pois foi verificado no Capítulo 5 que com essa probabilidade os pacotes de voz enviados são recebidos.

Cenário 2

Este cenário corresponde a avaliação do desempenho da aplicação de voz em função do número de arquivos transmitidos pelo servidor de arquivos,

ou seja, avalia-se os efeitos sobre o desempenho quando se aumenta o número de arquivos que são obtidos (*download* de arquivos) do servidor de arquivos e se avalia a probabilidade de degradação da comunicação de voz.

Para avaliação deste cenário, o tempo de transmissão de todos os outros componentes não foram alterados, ou seja, as medidas relativas a todos os componentes são inseridas no modelo e se variou apenas o lugar associado ao número de arquivos do servidor.

Neste cenário, consideraram-se as seguintes marcações do lugar P13: 1, 5, 10, 20, 30, 40, 50, 60 e 70, que correspondem ao número de arquivos armazenados no servidor disponível para *download*. O cliente inicia um *download* de arquivos e conseqüentemente gera tráfego. Com o aumento da quantidade de arquivos adquiridos a probabilidade de gargalo tende a aumentar.

A avaliação (análise estacionária) mostra que independentemente do número de arquivos armazenados no servidor, a probabilidade de gargalo é zero. Isso ocorre porque o número de arquivos que podem ser adquiridos simultaneamente foi limitado a 4 arquivos. O nível de utilização chega a 45% demonstrando que a máquina não está sobrecarregada.

Retirando a limitação do número de arquivos que podem ser adquiridos, o resultado da avaliação demonstra que a probabilidade de gargalo e o nível de utilização dos recursos são alterados com número de arquivos transmitidos pelo servidor, conforme mostra a Tabela 6.2.

Tabela 6.2: Variação do Número de Arquivos

Quantidade de Arquivos	PGARG %	TPS %
1	0	26.94
05	0	50.07
10	0	78.48
20	25.15	99.99
30	47.36	100
40	59.48	100
50	67.07	100
60	72.26	100
70	76.04	100

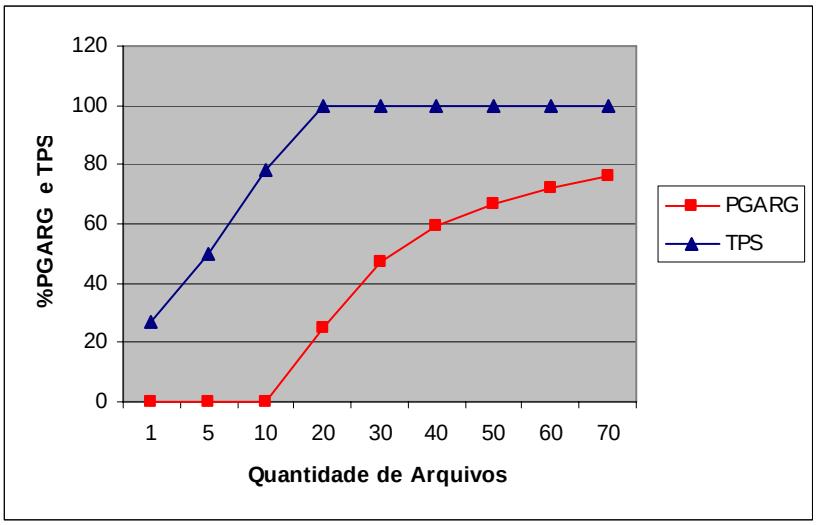


Figura 6.13 % PGARG e TPS em função da Quantidade de Arquivos

Observando a Figura 6.13, pode-se verificar que a probabilidade de gargalo aumenta consideravelmente quando a quantidade de arquivos adquiridos é maior do que 10. A análise também mostra que a partir de 20 arquivos o percentual de utilização chega a 100%**,** mas a probabilidade de gargalo é de 25%, em função do *buffer* (fila de transmissão) do *switch*. Portanto, os administradores de rede podem prevenir a degradação na comunicação de voz pela limitação do número de arquivos transmitidos. Esta limitação também permite um melhor gerenciamento do tráfego de rede.

Cenário 3

O cenário 3 observa o desempenho da aplicação de voz dado o aumento do tráfego de mensagens. Avalia-se uma situação que acontece geralmente quando o cliente recebe uma demanda muito grande de mensagens.

Neste cenário todas as outras transições do modelo permaneceram com seus tempos de transmissão inalterados. Para aumentar o tráfego referente à quantidade de mensagens que são enviadas, gradualmente altera-se o número de marcas no lugar P12. Realizando uma análise estacionária obtêm-se os seguintes resultados:

Tabela 6.3: Variação do número de Mensagens

Quantidade de Mensagens	PGARG %	TPS %
1	0.0	23.06783
05	0.0	60.97714
10	7.52808	99.03835
20	49.86885	100
30	66.01141	100
40	74.29227	100
50	79.32864	100
60	82.71494	100
70	85.14795	100

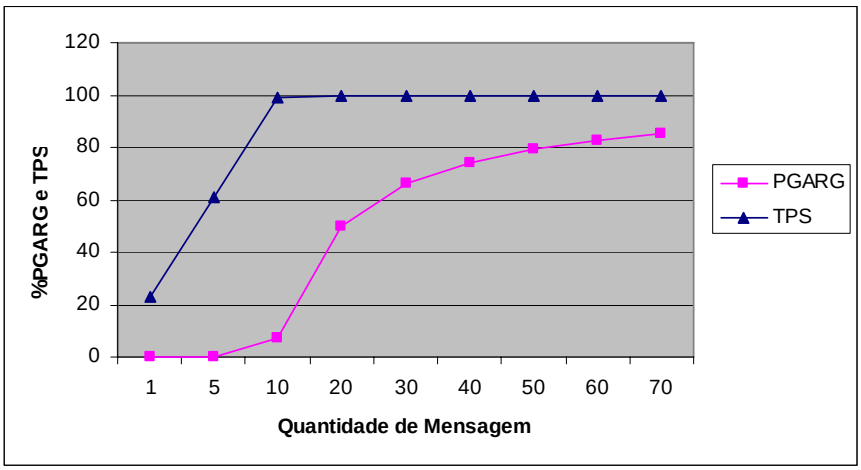


Figura 6.14 % (PGARG and TPS) em função da Quantidade de Mensagens

Pode-se observar neste cenário que quando a percentagem de utilização máxima é de 100%, o servidor de mensagens está enviando um total de 10 ou mais mensagens para o cliente (Figura 6.14). A partir de 10 mensagens também é verificado um aumento na probabilidade de gargalo, que chega a 80% quando o cliente recebe um total de 50 mensagens.

Este resultado permite configurar o servidor de mensagens para enviar uma mensagem de alarme quando o total de mensagens enviadas para o cliente excede 50, de forma que se evite a degradação da aplicação de voz. Neste caso o servidor de mensagens limitaria a quantidade de mensagens enviadas para o cliente, colocando-as na fila para serem enviadas posteriormente. Se esta fosse uma situação inusitada como uma infinidade de mensagens (vários *SPAM)* o cliente estaria protegido.

Cenário 4

O cenário 4 avalia o impacto da aplicação de voz em função da variação do tráfego do servidor de aplicações. O tempo de transmissão de todos os outros componentes não foram alterados, ou seja, as medidas relativas a todos os componentes são inseridas no modelo e se variou apenas o tempo associado à transição temporizada correspondente ao servidor de aplicações.

O resultado da análise estacionária (Figura 6.15) apresenta a probabilidade de gargalo em função da variação no tempo de transmissão. Os limites associados aos tempos de transmissão foram 0.00005s e 0.005s (associado à transição TTPHT).

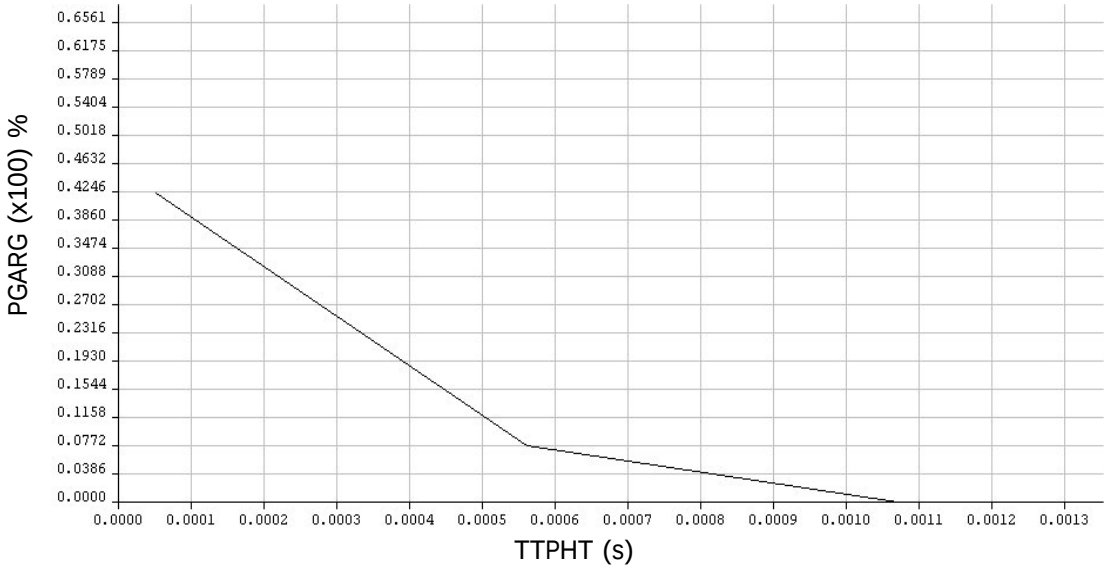


Figura 6.15 *PGARG* em função de TTPHT

A PGARG atinge 42% no tempo de 0,05 ms (Figura 6.15). Portanto, mesmo com alto tráfego no servidor de aplicações, a qualidade da comunicação de voz não é afetada.

A Figura 6.16 mostra a percentagem de utilização dos recursos considerando a variação do TTPHT.

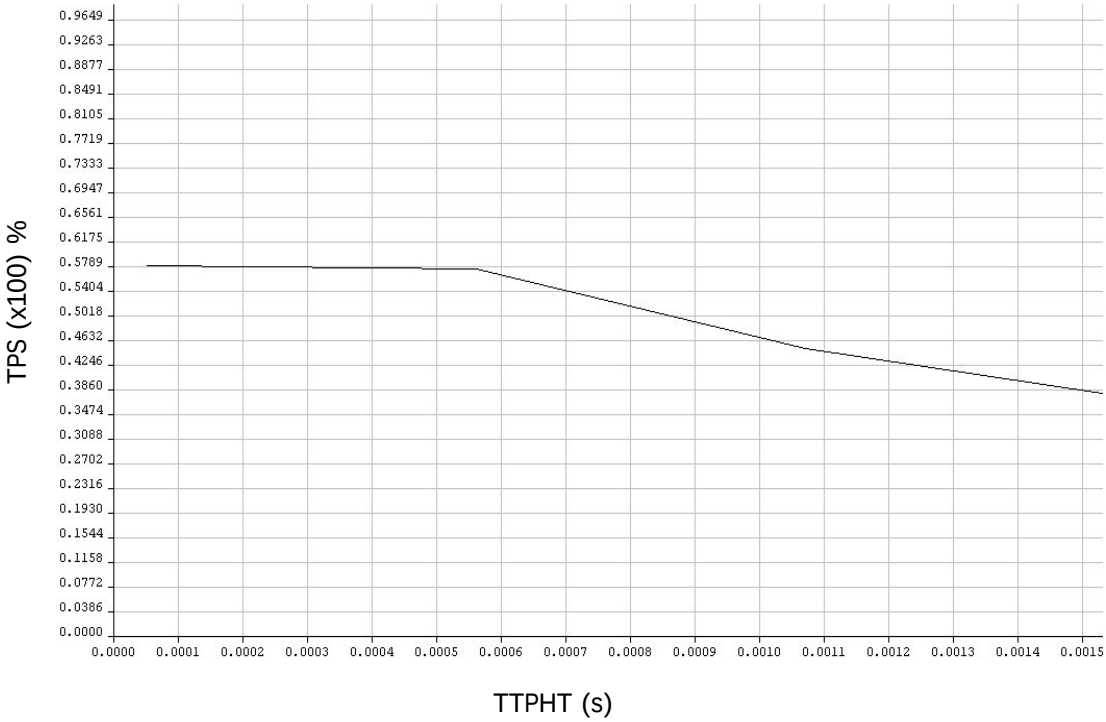


Figura 6.16 TPS em função de TTPHT

Os resultados apresentados na Figura 6.16 mostram que a aplicação de voz não sofre degradação de desempenho quando o tráfego do servidor de aplicação está em 0,05 ms. O gráfico mostra que mesmo com o tempo de transmissão de 0.0001 s, o percentual de utilização não chega 60%.

Os resultados mostram que o servidor de aplicações não tem grande impacto na comunicação de voz. A probabilidade de gargalo é de apenas 11% no tempo de 0.5 ms (Figura 6.15), mas o administrador de redes deve levar em consideração a utilização relativa a este tráfego porque em conjunto com outras aplicações o desempenho da aplicação de voz pode sofrer degradação.

Capítulo 7 - Conclusões

Este capítulo apresenta as considerações finais e as principais contribuições resultantes do desenvolvimento deste trabalho. Por fim, são expostas as principais sugestões para trabalhos futuros que estendem o trabalho apresentado.

Este trabalho apresentou um modelo para avaliação de desempenho de VoIP. A solução utilizada para avaliação de desempenho é baseada em modelos estocásticos utilizando distribuições poli-exponenciais.

O trabalho avaliou o desempenho da comunicação de VoIP utilizando o codec G.711 sobre diversas condições de tráfego. As métricas inseridas no modelo de avaliação correspondem às medições feitas em experimentos e cenários reais. As análises foram realizadas utilizando o software TimeNET que nos permite criar modelos GSPN.

Os diversos componentes utilizados neste trabalho tiveram características próprias de comportamento de tráfego. Em outras palavras, no estudo de caso, as aplicações Web, de correio e de arquivos tiveram características específicas de tráfego mensuradas.

O modelo proposto foi validado através da avaliação de um conjunto de cenários significativos e um conjunto de métricas. As avaliações foram realizadas por meio de métodos numéricos de análise ou através de simulação. A simulação foi adotada como opção nas situações em que as dimensões do espaço de estado inviabilizava a análise numérica. Contudo, alternativas de representação foram consideradas para redução do espaço de estados gerado pelo modelo GSPN. A estratégia utilizada para redução do número de estados consiste, basicamente, em representar cuidadosamente apenas os componentes do sistema sobre os quais se deseja a avaliação de uma métrica em particular. Sempre que possível, optou-se por representar os componentes ou conjunto de componentes do sistema através de modelos de alto-nível, ou seja, se não se tem interesse na avaliação de métricas “internas” de um

determinado componente ou sub-sistema, esta parte do sistema é representada por um único componente do modelo. Portanto, foi necessário definir de forma adequada a granularidade e a proporcionalidade dos valores obtidos. Por exemplo, verificou-se, em algumas situações, que a multiplicidade dos arcos que conectam alguns lugares a algumas transições (e vice-versa) não alterava o resultado numérico da avaliação, contudo comprometia fortemente as dimensões do espaço de estado, bem como o tempo de resposta da avaliação. Técnicas de redução também foram adotadas para que o número de estados gerados pelo modelo de alto-nível (GSPN) viabilizasse a avaliação do sistema.

Os resultados obtidos da análise do modelo GSPN se aproximaram de forma significativa dos resultados obtidos através da medição, o que permitiu a avaliação tanto do processo de medição adotado quanto o modelo proposto.
Um importante aspecto deste trabalho foi o desenvolvimento da metodologia adotada para avaliar o desempenho de ambiente de VoIP. Os modelos concebidos, dados os resultados obtidos no processo de validação, mostraram-se adequados para representação das características de desempenho desta classe de sistema.

A avaliação baseada em modelos possibilita a análise de desempenho de sistemas sem a necessidade de interferência no sistema real. Avaliar um sistema em operação pode não ser uma tarefa de fácil implementação ou ainda pode prejudicar sua operação. O uso de modelos como mecanismo de avaliação também torna possível a avaliação de cenários complexos cuja implementação real pode ser inviável devido aos custos ou mesmo em função da indisponibilidade temporária dos serviços, o que pode provocar redução de receita.

Como sugestão para trabalhos futuros, considera-se a adoção do modelo para avaliar o desempenho do tráfego de voz sobre com diferentes codecs; a avaliação aplicações de VoIP sobre redes com priorização de tráfego;
avaliar o desempenho de aplicações de VoIP tais como Gtalk e Skype; avaliar o desempenho de VoIP entre a rede 3G e a rede ADSL através do modelo GSPN proposto.

Referências Bibliográficas

[1] Andrew S. Tanenbaum, “Computer Networks”, 4th Ed, Prentice-Hall, 2003.

[2] Bolch G., Greiner S., Meer H., and Trivedi K. S., (1998): Queuing Networks and Markov Chains Modelling and Performance Evaluation with Computer Science Applications. John Wiley and Sons.

[3] Boudewijn R. Haverkort, Henrik C. Bohnenkamp, Connie U. Smith, Computer Performance Evaluation: Modelling Techniques and Tools : 11th International Conference, TOOLS 2000, Schaumburg, IL, USA, March 27-31, 2000.

[4] Carl A. Petri, Kommunikation mit Automaten. PhD, University of Bonn, West Germany, 1962.

[5] Catherine B., Gianluca I. and Chistophe D. (2002) “Impact of link failures on VoIP performance”, International Workshop on Network and Operating System Support for Digital Audio and Video, Miami, Florida, USA.

[6] Chandra, S. & Mathur, M., "Multimedia Standards - H.323 for Audio Visual Conferencing", 2000. Disponível em http://www-personal.ksu.edu/~mohit/multimedia.html.

[7] C. Huitema. Real Time Control Protocol (RTCP) attribute in Session Description Protocol (SDP). IETF RFC 3605 October 2003.

[8] David D. Clark and Wenjia Fang, “Explicit allocation of best-effort packet delivery service”, IEEE/ACM Trans. Netw., 1998, Piscataway, NJ, USA.

[9] DataBeam Corporation, "A Primer on the H.323 Series Standard", 1998. Disponível em http://www.lotus.com/products/sametime/sametime.nsf/standards . Acessado em 02 de Janeiro de 2008.

[10] Desrochers, A. A. and Al-Jaar, Robert Y (1995). “Applications of petri nets in manufacturing systems”. IEEE Press, Piscataway, NJ.

[11] Fouad A. Tobagi, Athina P. Markopoulou, Mansour J.Karam, “Is the Internet ready for VoIP.”, url= "citeseer.ist.psu.edu/575418.html", 2002.

[12] Hamdi, M., Verscheure, O., Hubaux J., “Voice Service Interworking for PSTN and IP Network”, Institute for Compute Communications and Applicatins – ICA (EPFL), 1999.

[13] Hoene, B. Rathke, A. Wolisz, C. (2003) “On the Importance of a VoIP Packet ”, In Proc. of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems, Berlin.

[14] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. IETF RFC 3550, July 2003.

[15] H. Schulzrinne, and J. Rosenberg, “Signaling for Internet Telephony”, IEEE September 1998.

[16] http://www.itu.int/net/home/index.aspx, acessado em 15 de janeiro de 2008.

[17] http://www.tiaonline.org/, acessado em 15 de dezembro de 2007.

[18] http://www.its.bldrdoc.gov/fs-1037/fs-1037c.htm, acessado em 01 de dezembro de 2007.

[19] HUREL, J-L et al. Mobile Network Evolution: From 3G Onwards. Alcatel Telecommunications Review – 4 Quarter 2003/1 Quarter 2004.

[20] Klepec, B.; Kos, A. (2001) “Performance of VoIP applications in a simple differentiated services network architecture” EUROCONapos; Trends in Communications, International Conference.

[21] ITU-T Recommendation H.323, Packet-Based Multimedia Communications Systems, Novembro 2000.

[22] ITU-T Recommendation G.711, Pulse Code Modulation (PCM) of Voice Frequencies, Novembro 1998.

[23] ITU-T Recommendation G.726, 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM), Dezembro 1990.

[24] ITU-T Recommendation G.722, 7 kHz audio-coding within 64 kbit/s, Novembro 1988.

[25] ITU-T Recommendation G.723.1, Speech Coders: Dual Rate Speech Coder for Multimedia Communitations Transmitting at 5.3 and 6.3 kbit/s, Março 1996.

[26] ITU-T Recommendation G.728, Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction, Setembro 1992.

[27] ITU-T Recommendation G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-code-Excited Linear-Prediction (CS-ACELP), Março 1996.

[28] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler. SIP: Session Initiation Protocol. RFC3261. June 2002.

[29] Jayant, N. S. and Noll, P. "Digital Coding Waveforms – Principles and Applications to Speech and Vídeo Englewood Cliffs", Prentice Hall 1984.

[30] Jeong-Soo Han, Seong-Jin and Jin-Wook, "Study of delay patterns of weighted voice traffic of end-to-end users on the VoIP Network", INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT Int. J. Network Mgmt 2002; 12: 271 – 280.

[31] Kostas, T. J. et al. Real Time Voice Over Packet-Switch Networks, IEEE Network, Vol. 12, n. 1, pp. 18-27, Janeiro 1998.

[32] Lathi, B. P., "Modern Digital and Analog Communication Systems", 3rd ed, Oxford University Press, 1998.

[33] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli and G. Franceschinis (1994), "Modelling With Generalised Stochastic Petri Nets", Università degli studi di Torino, Dipartimento di Informatica.

[34] M. Ajmone Marsan, G. Balbo and G. Conte. A class of generalized stochastic Petri nets for the performance analysis of multiprocessor systems. ACM Transactions on Computer Systems, May 1984.

[35] M. Carmo, J. Sá Silva, E. Monteiro, P. Simões, M. Curado and F. Boavida, "Avaliação da QoS em Redes Ethernet", Universidade de Coimbra, CISUC – Dep. Eng. Informática Polo II, 3030 Coimbra, PORTUGAL.

[36] Mehta, C. P., and Udani, S., Overview of Voice over IP, Technical Report MS-CIS-01-31, University of Pennsylvania, fevereiro 2001.

[37] Michel Mouly, Marie-Bernadette Pautet, “The GSM System for Mobile Communications”, Telecom Publishing (June 1992).

[38] P.M. Fiorini: Voice over IP for Enterprise Networks: Performance Implications & Traffic Models. Computer Measurement Group (CMG), 2000.

[39] P. Maciel, R. Lins, and P. Cunha. Introdução às Redes de Petri e Aplicações. X Escola de Computação Campinas-SP, julho de 1996.

[40] Performance Analysis, First EEF/Euro Summer School on Trends in Computer. Scienc Berg em Dal, The Netherlands, July 3-7, 2000 Revised Lectures.

[41] Rajavelsamy R, Venkateswar Jeedigunta, Balaji Holur, Manoj Choudhary and Osok Song, “Performance Evaluation of VoIP over 3G-WLAN Interworking System”, Volume: 4, On page(s): 2312- 2317 Vol. 4, March 2005, Samsung India Software Operations, India.

[42] Raj Jain, Art Of Computer Systems Performance Analysis, 1991.

[43] Sheldon M. Ross, “Introductory Statistics”, Academic Press, Hardcover, 2nd Bk&CD edition, March 2005.

[44] Qingguo Shen , “Performance of VoIP over GPRS”, Institute of Communications Engineering, PLAUST; Advanced Information Networking and Applications, 2003. AINA 2003. 17th International Conference, March 2003.

[45] Schulz, T., Voice over IP – White Paper, Eicon Technology Corporation, February 2000.

[46] S. Kent and R. Atkinson, ”Security Architecture for the Internet Protocol”, IETF RFC 2401, November 1998.

[47] Stallings, W., Christianson, L., Brown, K., “Data and Computer Communications and Computer”, Prentice Hall, 2004.

[48] Tadao Murata. Petri Nets: Properties, analysis and applications. In Proceedings of IEEE, 1989.

[49] Tadao Murata. “Modeling and analysis of concurrent systems,” in Handbook of Software Engineering (C. R. Vick and C. V. Ramamoorthy, eds.), ch.3, New York: Van Nostrand Reinhold, 1984.

[50] Toga, J., Elgebaly, H. Demystifying Multimedia Conferencing Over the Internet Using the H.323 Set of Standards, Intel Architecture Labs, Intel Corporation, 1998. Disponível em http://www.intel.com.tw/technology/itj/q21998/articles/art4.htm .

[51] Wenyu Jiang and Henning Schulzrinne, (2002) “Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss” Department of Computer Science, Columbia University, New York, USA.

Dissertação de Mestrado apresentada por **Antonio Ricardo Pereira Cavalcanti** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título **"Avaliação de Desempenho de VoIP Através de Modelos Estocásticos Utilizando Distribuições Poli-exponenciais"**, orientada pelo **Prof. Paulo Romero Martins Maciel** e aprovada pela Banca Examinadora formada pelos professores:

Prof. Paulo Romero Martins Maciel
Centro de Informática / UFPE

Prof. Ricardo José de Paiva Britto Salgueiro
Departamento de Computação e Estatística

Prof. Ricardo Massa Ferreira Lima
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 27 de agosto de 2008.

Prof. FRANCISCO DE ASSIS TENÓRIO DE CARVALHO
Coordenador da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Cavalcanti, Antonio Ricardo Pereira
Avaliação de desempenho de VoIP através de modelos estocásticos utilizando distribuições poli-exponenciais / Antonio Ricardo Pereira Cavalcanti - Recife : O Autor, 2008.
115 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, 2008.

Inclui bibliografia.

1. Redes de computadores. 2. Avaliação de desempenho. 3. Modelos estocásticos. 4. Redes de Petri. I. Título.

004.6 CDD (22. ed.) MEI2009-052