MELHORES PRÁTICAS EM REDES COLABORATIVAS. ESTRATÉGIAS PARA ALTO DESEMPENHO E DISPONIBILIDADE EM EMPRESAS

Pedro Cardoso Damasceno

Orientador: Professor Ricardo Zelenovsky

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Elétrica da Faculdade de Tecnologia na Universidade de Brasília.

BRASÍLIA, 2018

UNIVERSIDADE DE BRASÍLIA – UNB
FACULDADE TECNOLOGIA

MELHORES PRÁTICAS EM REDES COLABORATIVAS. ESTRATÉGIAS PARA ALTO DESEMPENHO E DISPONIBILIDADE EM EMPRESAS

Soluções para assegurar alta disponibilidade e desempenho em redes VoIP. Abordagem com ênfase no modelo de implantação Cisco.

Damasceno, Pedro Cardoso

Orientador: Professor Ricardo Zelenovsky

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Elétrica da Faculdade de Tecnologia - UNIVERSIDADE DE BRASÍLIA – UNB

BRASÍLIA, 2018

AGRADECIMENTOS

Agradeço a todos que de alguma forma participaram desse trabalho. Provendo suporte, ideias, críticas e apoio.

Inicialmente gostaria de agradecer minha família por nunca duvidarem do meu potencial e investimento no meu desenvolvimento. Aos amigos de curso que tiveram importância singular na minha formação.

Ao professor Ricardo Zelenovsky por apoiar minhas ideias apresentadas todas as vezes e acreditar nos trabalhos propostos.

Ao professor Paulo Henrique Portela por me dar orientações e validar ideias propostas neste trabalho

À equipe da A.Telecom, por passar seu conhecimento e expertise sobre esse tipo de solução, investir na minha profissionalização e pela parceria criada durante meu período de trabalho.

RESUMO

A tecnologia em sistemas de comunicação sempre foi um tema com grandes inovações. Desde a invenção do telefone até os últimos avanços em tecnologias de telepresença, a área de comunicações tem se mostrado um agente de grande peso no desenvolvimento de grandes empresas.

Nas últimas décadas, uma das grandes evoluções dessa área tem sido a transição para o uso do IP na telefonia, se aproveitando do grande crescimento nos sistemas de redes de computadores e aprimoramento essenciais dos equipamentos e tecnologias (alta capacidade de *throughput*, PoE, protocolos de comunicação, plataformas de virtualização, integração dos sistemas, entre outros).

Embora esse desenvolvimento demonstre grande potencial e novas facilidades, surgem diversos obstáculos e peculiaridades para a implantação desses sistemas de forma que se integrem à rede de dados e garantam qualidade de serviço, disponibilidade e novas aplicabilidades dependendo das tecnologias, funcionalidades, topologias e equipamentos que uma empresa necessita e dispõe.

Considerando as especificidades que serão apresentadas ao se implantar esse tipo de sistema, o trabalho dispõe desenvolver e apreciar mecanismos para abordar a continuidade e qualidade dos serviços de VoIP e vídeo de uma empresa com arquitetura homogênea (equipamentos do mesmo fabricante) *multisite* e processamento de chamadas centralizado. O estudo propõe evitar e tratar de sobrecarga em links de rede e ações que mantenham o processamento de chamadas quando há falhas de conectividade, indisponibilidade, interrupção de serviços ou limitação de capacidade dos links.

Serão tratados temas como codecs e o efeito que provocam no consumo da banda, recursos de mídia e como aplicá-los para melhor aproveitamento para evitar consumo de banda em links WAN, QoS e sua imprescindível necessidade de se integrar aos sistemas VoIP. Serão descritos princípios de *design* e estratégias para emprego do QoS, gerenciamento de congestionamento de dados. Ao final, serão abordados métodos para garantir alta disponibilidade.

Será proposto um modelo para realizar as configurações das metodologias propostas e observar a diferença de comportamento após aplicadas as técnicas

abordadas. Será dada ênfase às tecnologias da Cisco, uma vez que são equipamentos amplamente empregados em empresas, com alta confiabilidade, desempenho, documentação e suporte. Procurou-se manter a rede homogênea, sem adicionar equipamentos de outros fabricantes.

ABSTRACT

Technology in communication systems has always been a theme with great innovations. From the invention of the telephone to the latest advances in telepresence technologies, the communications area has proven to be a major player in the development of large companies.

In the last decades, one of the great evolutions of this area has been the transition to the use of the IP in the telephony, taking advantage of the great growth in the systems of computer networks and essential improvement of the equipment and technologies (high capacity of throughput, PoE, protocols of communication, virtualization platforms, systems integration, among others).

Although this development demonstrates great potential and new facilities, several obstacles and peculiarities arise for the implementation of these systems, so that they integrate into the data network and guarantee service quality, availability and new applications depending on the technologies, functionalities, topologies and equipment that a company needs.

Considering the specificities that will be presented when this type of system is implemented, this work proposes the development and evaluation of mechanisms to address the continuity and quality of VoIP and video services of a company with homogeneous architecture (equipment of the same manufacturer), multisite and call processing centralized. The study proposes to avoid and deal with overhead in network links and actions that maintain the processing of calls when there are connectivity failures, unavailability, interruption of services or limitation of capacity of the links.

Topics such as codecs and the effect they cause on bandwidth consumption, media resources and how to apply them for better use to avoid bandwidth consumption in WAN links, QoS and their essential need to integrate with VoIP systems will be contemplated. It will be approached design methods and strategies for employment of QoS and management of data congestion. At the end, methods to ensure high availability will be addressed.

A model will be proposed to carry out the configurations of the proposed methodologies and to observe the difference in behavior after applying the techniques discussed. It will be placed emphasis on Cisco technologies as they are widely used

in enterprises, with high reliability, performance, documentation and support. The network was mantained homogeneous, without adding equipment from other manufacturers.

SIGLAS

VoIP	-	Voice Over IP
IP	-	Internet Protocol
PoE	-	Power Over Ethernet
PSTN	-	Public Switched Telephony Network
UC	-	Unified Communications
CUCM	-	Cisco Unified Call Manager
CME	-	Call Manager Express
CODEC	-	Coder Decoder
SIP	-	Session Initiation Protocol
QoS	-	Quality of Service
CAC	-	Call Admission Control
RMC	-	Rota de Menor Custo
AAR	-	Automatic Alternate Routing
TEHO	-	Tail End Hop Off
CUBE	-	Cisco Unified Border Element
SRST	-	Survivable Remote Site Telephony
MOH	-	Music On Hold
IOS	-	Internetwork Operating System
WAN	-	Wide Area Network
LAN	-	Local Area Network
MQC	-	Modular QoS CLI
DSP	-	Digital Signal Processor
CFUR	-	Call Forward Unregistered
DE	-	Discard Elegiable

ATM	-	Automated Teller Machine
CLP	-	Cell Loss Priority
MPLS	-	Multi Protocol Label Switching
IPP	-	IP precedence
PQ	-	Priority Queue
LLQ	-	Low Latency Queueing
RED	-	Random Early Detection
WRED	-	Weighted Random Early Detection
MPD	-	Mark probability denominator
CIR	-	Commited Information Rate
RTP	-	Real Time Protocol
SCCP	-	Skinny Call Control Protocol
CFNB	-	Call Forward No Bandwidth
DID	-	Direct Inward Dialing
SNR	-	Single Number Reach
SLA	-	Service Level Agreement
RTT	-	Round-Trip Time
VPN	-	Virtual Private Network
SONET	-	Synchronous Optical Network
IPSec	-	IP Security
LLQ/CBWFQ	-	Low Latency Queue / Class Based Weighted Fair Queue
ACL	-	Access Control List
CPU	-	Central Processing Unit
ITU	-	International Telecommunications Uion
DSP	-	Digital Signal Processor

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

Com avanços das tecnologias de transmissão de dados, a telefonia IP se tornou uma tecnologia de uso generalizado no meio empresarial. Esse sistema ficou popular devido às diversas vantagens que oferece, como o aproveitamento do investimento feito na rede de dados, escalabilidade do sistema, redução de despesas e custos de operação, maximização da produtividade e melhorias no serviço com clientes. De forma geral, apresentam um custo benefício evidente [1].

A tendência dos sistemas de comunicação nas empresas é tomar forma e evoluir para a convergência das redes tradicionais de telefonia, as quais eram providas separadamente, para uma infraestrutura única de transmissão em pacotes (voz, dados, imagens, vídeos, sons trafegando pela mesma infraestrutura) [1] conforme as figuras 1 e 2 apresentadas a seguir:

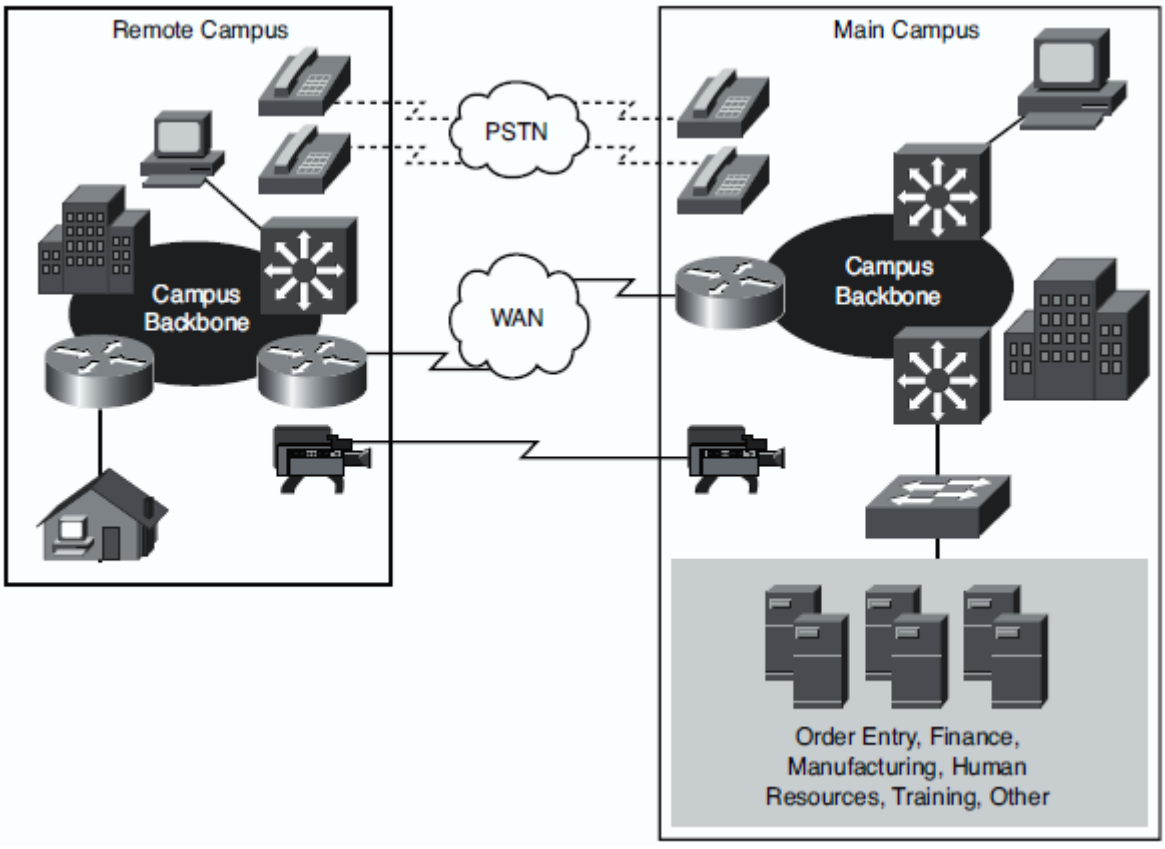


Figura 1 - Rede antes da convergência

Fonte: [2]

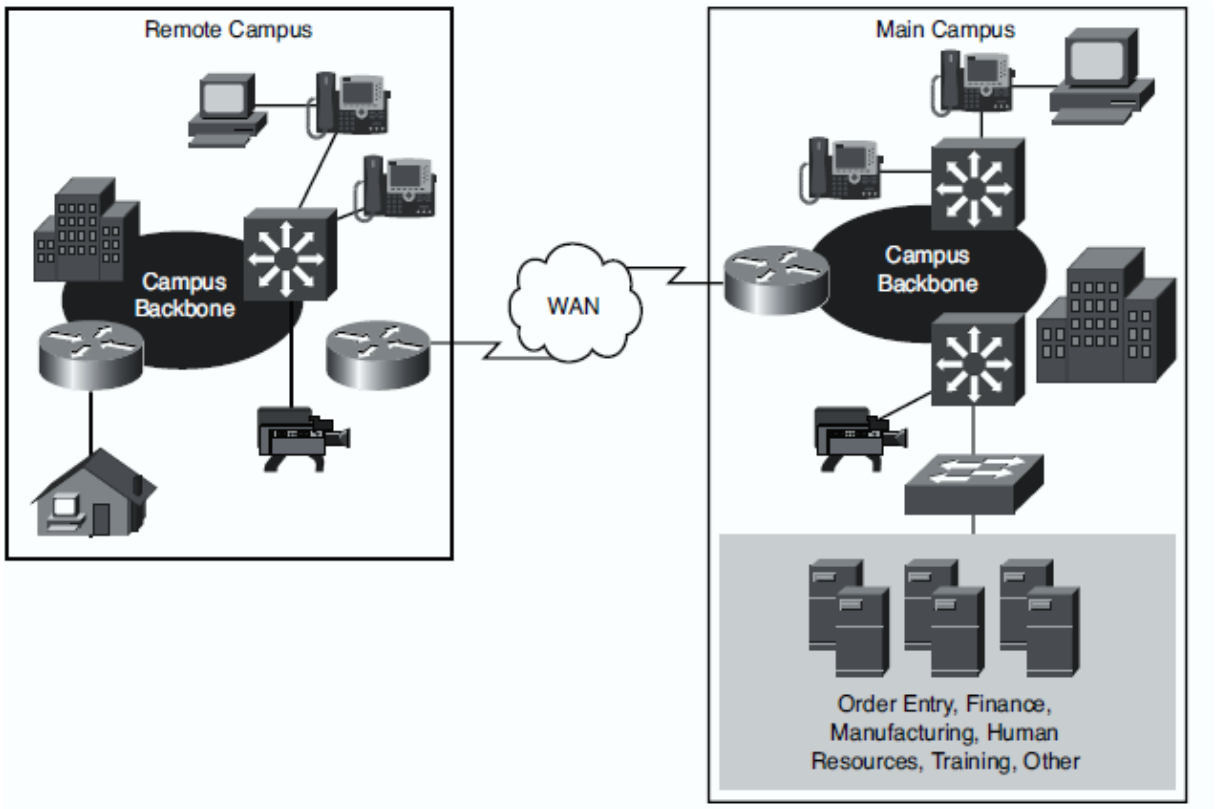


Figura 2 - Rede após convergência

Fonte: [2]

A principal diferença entre as redes convergentes e as redes tradicionais de comutação por circuitos está na estrutura de transmissão por pacotes utilizada no protocolo IP e adotada nessa nova solução. Para que possam trafegar nas novas redes, os sinais de voz precisam ser transformados em pacotes para serem transportados. Estes são adicionados ao tráfego de dados fluindo na rede e são tratadas as especificidades para abordar essa nova tecnologia. Essa função é realizada por *gateways* de voz, roteadores, centrais telefônicas e toda a infraestrutura física do ambiente.

Esse novo cenário apresentado ocasionou uma flexibilidade muito ampla na forma de implementar uma rede para satisfazer suas conveniências devido ao grande número de dispositivos e tecnologias que começam a ser integrados. Essa integração promove um novo conceito: Comunicações Unificadas.

Comunicações unificadas (UC) é um termo que descreve a integração de serviços de comunicação empresarial, como mensagens instantâneas (chat), informações de presença, voz (incluindo telefonia IP), recursos de mobilidade

(incluindo mobilidade de extensão e alcance único), áudio, web e videoconferência, e-mail, convergência fixo-móvel (FMC), compartilhamento de desktop, compartilhamento de dados (incluindo quadros interativos eletrônicos conectados à web), controle de chamadas e reconhecimento de fala com serviços de comunicação, como mensagens unificadas (correio de voz integrado, SMS, fax), integração com bancos de dados, mídia social, *bots*, entre outros. A UC não é necessariamente um único produto, mas um conjunto de produtos que fornece uma interface de usuário unificada consistente e uma experiência do usuário em vários dispositivos e tipos de mídia.

Em seu sentido mais amplo, a UC pode abranger todas as formas de comunicação que são trocadas através de uma rede.

Existem diversos padrões de implantação em uma rede de comunicação unificada, os quais devem ser abordados para servir às expectativas e funcionalidades que uma firma necessita. Em termos gerais, a arquitetura implementada é ajustada para que atenda às necessidades e topologias típicas e bem definidas das empresas.

Cada modelo possui seu grau de complexidade, entretanto observa-se que a maior parte dos problemas de design se manifesta na transição de um sistema *single-site* para um *multisite*, uma vez que envolve diversas considerações que não são aplicadas ao primeiro. Para fins de esclarecimento, o modelo que será realçado é o modelo de implantação de processamento de chamadas centralizado, o qual atende a empresas cuja área de cobertura operacional é baseada em vários sites ligados a um escritório central da matriz.

Esse trabalho visa apreciar a seguinte questão: Ao se adicionar mais de uma localidade ao cenário (conforme a figura 3 apresentada a seguir), diversas ressalvas e problemas com a implantação, garantia de serviço e qualidade são adicionados ao ambiente. Muitas vezes percebe-se que o serviço não está funcionando conforme o proposto devido à má gestão, configuração e dimensionamento da rede da empresa e não devido ao serviço entregue pelo provedor de serviço (ITSP). Essas ressalvas incluem disponibilidade, redundância, qualidade do serviço, questões de largura de banda, plano de discagem...

Serão investigados problemas associados à implantação de sistemas de voz e vídeo em mais de uma localidade e sugeridas metodologias e melhores práticas para

prover o bom funcionamento da tecnologia e dar amparo para as necessidades das empresas.

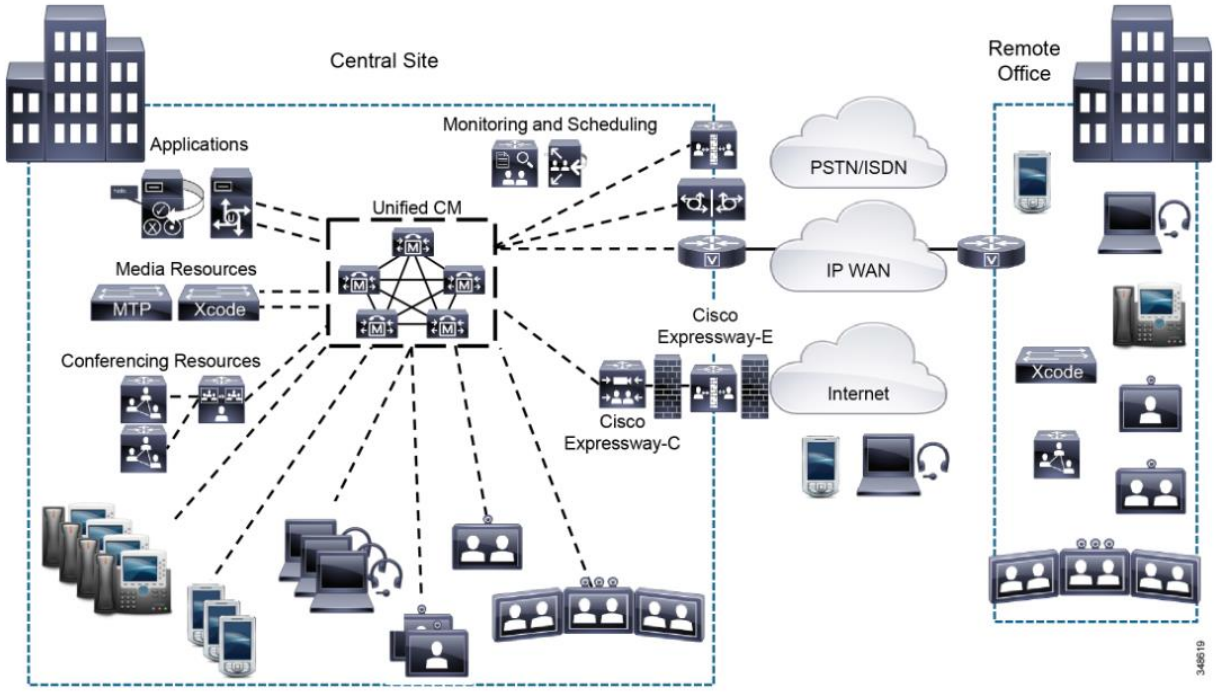


Figura 3 - Modelo de rede multisite com processamento de chamadas centralizado

Fonte: [3]

1.1. MOTIVAÇÃO E DEFINIÇÃO DO PROBLEMA

O desenvolvimento de uma empresa geralmente acarreta expansão e possivelmente adição de novas filiais ou localidades. Juntamente com esse crescimento, se adicionam diversas medidas de implementação para que o sistema de voz e dados da empresa acompanhem o desenvolvimento e entregue o serviço esperado.

Em ambientes pobremente projetados onde não foram aplicadas boas metodologias para a solução dos serviços de colaboração, o usuário final acaba por perceber diversos sintomas de problemas na rede: Degradação na qualidade de voz e vídeo (eco, ruído de fundo, voz metálica, cortes na fala, atraso), ausência de áudio em um ou ambos os sentidos, queda das ligações e conferências, indisponibilidade do serviço...

É fato que a implantação de um ambiente de UC *multisite* exige que sejam apreciados alguns aspectos e considerações de projeto únicos. Torna-se necessário um profundo entendimento das melhores práticas para criar uma organização e procedimentos que sustentem alocação de recursos de infraestrutura adequada, plano de discagem que permita escalabilidade e manutenção, largura de banda compatível com os serviços utilizados (não apenas para telefones IP, mas também para *endpoints* de vídeo), *design* e implementação de qualidade de serviço (QoS), arquitetura de rede de longa distância e da área local (LAN/WAN) altamente disponível, incluindo telefonia de local remoto com capacidade de sobrevivência (SRST).

Considerando as especificidades apresentadas ao se implantar esse tipo de sistema, o trabalho dispõe desenvolver e apreciar mecanismos para abordar a continuidade e qualidade dos serviços de VoIP e vídeo de uma empresa com arquitetura homogênea (equipamentos do mesmo fabricante) *multisite* e processamento de chamadas centralizado. O estudo propõe evitar e tratar de sobrecarga em links de rede e ações que mantenham o processamento de chamadas quando há falhas de conectividade, indisponibilidade, interrupção de serviços ou limitação de capacidade dos links.

Será proposto um modelo para realizar as configurações das metodologias propostas e observar a diferença de comportamento após aplicadas as técnicas abordadas. Será dada ênfase às tecnologias da Cisco, uma vez que são equipamentos amplamente empregados em empresas, com alta confiabilidade, desempenho, documentação e suporte. Procurou-se manter a rede homogênea, sem adicionar equipamentos de outros fabricantes.

É importante atentar e esclarecer que o trabalho não se destina a ensinar o leitor sobre configurações de equipamentos da Cisco, apesar de tratar aprofundadamente sobre esse meio, enfaticamente o âmbito de colaboração. Conjectura-se que o público alvo são profissionais que já têm conhecimento dessa tecnologia, porém verifica-se que existe uma lacuna de conhecimento técnico e prático ao se instaurar esse tipo de solução no ambiente corporativo.

Este trabalho é dedicado a técnicos e administradores de redes que já estão envolvidos no meio de colaboração e possuem conhecimento de configuração dos

equipamentos e tecnologias apresentados, mas que desejam refinar e aperfeiçoar o entendimento deste tipo de solução.

1.2. OBJETIVO DO TRABALHO

Esse trabalho possui os seguintes objetivos:

- Investigar criticamente elementos decisivos nos modelos de implantação do meio colaborativo;
- Apresentar os problemas mais comuns ao se implantar um sistema de voz em um ambiente com processamento de chamadas centralizado com solução *multisite*;
- Analisar e propor melhores práticas para se garantir alta disponibilidade de serviços, redundância, confiabilidade e tratamento de banda;
- Gerar conhecimento e habilidades nas soluções de implantação no ambiente de colaboração da Cisco para se diminuir a lacuna entre conhecimentos técnicos da tecnologia e implantações em ambientes reais;
- Implantar um sistema para simular a melhoria dos serviços ao se aplicar as técnicas apresentadas;

1.3. ORGANIZAÇÃO DO TRABALHO

A organização do trabalho será a seguinte:

O **capítulo 2** apresenta o referencial teórico:

Apresenta, de forma geral, os modelos de implementação e mecanismos de conectividade entre os sites e a PSTN e principais preocupações nesse tipo de solução.

Identifica os desafios relevantes e considerações em implantações *multisite* que utilizam soluções Cisco.

Descrição de mecanismos para gerenciar o consumo da banda, que é um dos recursos mais caros e preciosos nesse tipo de solução. Descreve opções de implementação do QoS e suas variantes.

Aborda também opções de redundância e disponibilidade para prover soluções de telefonia em *sites* remotos.

O **capítulo 3** propõe um sistema desenvolvido em laboratório com os objetivos principais de analisar e verificar os mecanismos abordados no trabalho para garantir desempenho sofisticado e alta disponibilidade dos recursos.

Por fim, a conclusão sobre o trabalho e aponta possíveis caminhos e desenvolvimentos tecnológicos que estão sendo desenvolvidos nesse meio. No final serão apresentadas as referências bibliográficas utilizadas para embasamento do trabalho e aprofundamento dos temas apresentados.

O **capítulo 4** expõe a conclusão do trabalho apresentando o que o trabalho se propôs a fazer, o que foi feito, resultados mais expressivos, e proposta de trabalhos futuros. Ao final são apresentadas informações mais específicas sobre a configuração dos ativos da rede, como os roteadores, switches e servidores, nos laboratórios propostos.

2. REFERENCIAL TEÓRICO

2.1. CONSIDERAÇÕES DA SOLUÇÃO MULTISITE

2.1.1. Modelos de implementação

Resumidamente, existem três modelos básicos de implantação para o sistema unificado de comunicações e colaboração:

Campus: Serviços de UC, terminais, gateways, controladores de borda, recursos de mídia e outros componentes associados estão todos localizados em uma única LAN de alta velocidade.

Centralizado: Serviços de UC estão localizados em um site ou centro de dados central do campus, mas os pontos de extremidade, gateways, recursos de mídia e outros componentes são distribuídos por vários sites remotos interconectados por uma WAN ativada por QoS.

Distribuído: Vários *campus* e / ou implantações centralizadas (Clusters) são interconectados por meio de uma plataforma de agregação de plano de discagem e tronco em uma WAN habilitada para QoS.

Há um número amplo de variações nesses três modelos básicos para se adequar a demandas específicas, mas as orientações apresentadas nesse trabalho se aplicam à maioria delas e são bases para elaborações mais complexas. [3]

2.1.2. Serviço Centralizados

Uma solução multisite pode ser implementada de diversas maneiras. Em muitos casos, os principais critérios que direcionam o design de cada serviço são a disponibilidade e a qualidade da rede IP entre esses sites.

Em uma organização a qual os sites de filiais da empresa estão geograficamente dispersos e interconectados em uma rede de longa distância, os

serviços de comunicação unificada (UC) podem ser implantados em um local central enquanto atendem aos terminais pelas conexões WAN.

A centralização dos serviços de comunicação unificada (UC) oferece vantagens de economia de escala tanto em despesas de capital quanto operacionais associadas à hospedagem, administração e operação de equipamentos em situações que a conectividade entre os sites oferece as seguintes características:

- Largura de banda suficiente para a carga de tráfego prevista, incluindo cargas de acesso de horário de pico;
- Alta disponibilidade, onde o provedor de serviços WAN adere a um acordo de nível de serviços (SLA) para manter e restaurar a conectividade imediatamente;
- Baixa latência, onde eventos locais no site remoto não serão prejudicados caso o *round-trip time* (RTT) para o site principal apresentar alguns atrasos nos tempos de resposta do sistema;

Além disso, quando um determinado serviço é implantado centralmente para atender aos pontos de extremidade em vários sites, há vantagens de transparência dos recursos oferecida pelo uso das mesmas faculdades de processamento para usuários de outras regiões.

Essas vantagens de transparência de recursos e economias de escala devem ser avaliadas em relação ao custo relativo de estabelecer e operar uma rede WAN configurada para acomodar as demandas do tráfego de comunicações unificadas.

Opções de conexão para a rede WAN IP incluem [4]:

- Linhas dedicadas (*Leased Lines*);
- *Frame Relay*;
- *Asynchronous Transfer Mode* (ATM);
- MPLS;
- VPN;
- IPSec VPN (V3PN);
- Satélite;
- SONET;

2.2. DESAFIOS A SEREM TRATADOS

O objetivo de qualquer implantação em vários sites é fornecer aos usuários finais experiência amigável com tecnologias de UC, mantendo disponibilidade máxima. O escopo e a complexidade do projeto entram em cena nesta última porção.

Diversos componentes se apresentam para prover a capacidade de manter a rede acessível e desimpedida:

- QoS através da WAN;
- Desafios para limitação de banda;
- Métodos de backup para disponibilidade;
- Modo de sobrevivência local (SRST);
- Considerações sobre localização de recursos de mídia;
- Escalabilidade;
- Soluções de mobilidade;
- Segurança;

2.2.1. Voz Sobre IP (VoIP)

Embora os pacotes que transportam tráfego de voz sejam tipicamente pequenos, eles não podem tolerar atraso ou variação nos atrasos à medida que atravessam a rede.

Os pacotes que transportam dados de transferência de arquivos são tipicamente grandes e podem sobreviver a atrasos e quedas. É possível retransmitir parte de um arquivo de dados descartado, mas não é possível retransmitir parte de uma conversa de voz. O fluxo de voz constante e de pequeno volume compete com os fluxos de dados em rajada. A menos que algum mecanismo medeie o fluxo geral, a qualidade de voz será seriamente comprometida em tempos de congestionamento de rede. O tráfego de voz crítico deve ter prioridade, uma vez que é muito sensível ao tempo. Não pode ter valores significativos de atrasado e não pode ser excessivamente descartado, ou a qualidade resultante de voz e vídeo sofrerá.

O tráfego de voz possui os seguintes requisitos adicionais unidirecionais:

(ITU-T G.114)

- Latência: recomendado 150ms *one-way*; porém, até 200ms é aceitável
- Jitter: 30ms ou menos;
- Perda de pacotes: 1% ou menos (CISCO);

Serão tratados os métodos para garantir essas cláusulas.

Caso esses requisitos não sejam atendidos, são apresentados alguns sintomas do problema de qualidade de voz. Algumas definições foram desenvolvidas e aplicadas a fim de categorizá-los [5]:

Ruído

Normalmente trata-se de qualquer ruído na linha ou em uma mensagem de correio de voz além do sinal de voz. Ruído normalmente deixa a conversa inteligível, mas ainda longe de ser excelente. Estática (Estática é uma distorção granular semelhante à má recepção no rádio. Causas comuns são interferência elétrica ou VAD.), zumbido e tons intermitentes são exemplos em que as partes chamadoras e chamadas podem se entender, mas com algum esforço. Alguns ruídos são tão graves que a voz se torna ininteligível.

Adulteração da voz

Normalmente é um problema que afeta a própria voz:

Voz com eco

O eco é o sintoma em que o sinal de voz é repetido na linha. Ele pode ser ouvido em qualquer uma das extremidades da chamada, em graus variados e com muitas combinações de atraso e perda dentro do sinal ecoado. É o sinal que vaza na extremidade oposta e retorna ao emissor (locutor). O falante ouve um eco de sua própria voz.

Causas comuns são:

- Perda insuficiente do sinal de eco.
- Os canceladores de eco no gateway adjacente ao híbrido da extremidade oposta não são ativados.
- Eco acústico causado pelo telefone do ouvinte.

Voz distorcida

Um sinal da voz distorcida é aquele em que o caráter real da voz é alterado para um grau significativo e, muitas vezes, tem uma qualidade que flutua. Em algumas ocasiões, a voz se torna ininteligível.

A distorção da voz é subdividida em algumas categorias como voz cortada, robótica, sintética.

Causas comuns são:

- Consecutivos pacotes perdidos ou excessivamente atrasados além dos limites do período de reprodução do *buffer dejitter*, de forma que a inserção preditiva de DSP não possa ser usada e o silêncio seja inserido.
- VAD

2.2.1.1. Latência

Latência, ou atraso, é o tempo que leva um pacote para percorrer uma rede de ponta a ponta. Em termos de telefonia, a latência é a medida do tempo que a voz do locutor leva para chegar ao ouvido do ouvinte. Grandes valores de latência não necessariamente degradam a qualidade do som de uma chamada telefônica, mas o resultado pode estar influenciando na sincronização da conversa de tal forma que há hesitações na interação dos interlocutores.

Geralmente, aceita-se que a latência de ponta a ponta deve ser inferior a 150 ms para chamadas telefônicas de qualidade. Para garantir que o orçamento de latência permaneça abaixo de 150 ms, é necessário levar em consideração as causas principais do atraso. Ao projetar uma rede multisserviço, o atraso total que um sinal ou pacote exibe é um somatório de todos os contribuidores de latência:

- **Atraso de pacote**

É o tempo que leva para os endpoints criarem os pacotes usados nos serviços de voz, ou seja, o tempo necessário para preencher um pacote com dados.

Geralmente, quanto maior o tamanho do pacote, maior o tempo necessário para preenche-lo. O atraso de empacotamento é regido pelo CODEC que está sendo usado. Esse problema também existe no lado de recebimento, porque o gateway de mídia deve remover e processar os dados no pacote. Se os pacotes forem mantidos pequenos, essa quantidade de atraso, em ambas as direções, é geralmente muito pequena. Essa operação não deve exceder 30ms.

- **Serialização**

Serialização refere-se ao tempo que leva para converter um quadro da Camada 2 em pulsos elétricos ou ópticos da Camada 1 para a mídia de transmissão. Portanto, o atraso de serialização é fixo e é uma função da taxa de linha (ou seja, a velocidade de clock do link). Esse atraso é inversamente proporcional à velocidade do link. Em outras palavras, quanto mais rápida a mídia, menor a latência. Esse valor é um pouco dependente da tecnologia de link usada e seu método de acesso (Multilink PPP, ATM, Frame Relay, MPLS).

A fórmula para serialização [6]:

$T = F/ L;$

onde:

T = Atraso de serialização [s];

F =Tamanho do frame [bits];

L = Velocidade do link [bps];

Por exemplo, um circuito T1 (1,544 Mbps) requer cerca de 8 ms para serializar um quadro Ethernet de 1500 bytes no fio, enquanto são necessários 187,5 mili-segundos para colocar o mesmo quadro em um circuito de 64 Kbps. Embora esse atraso seja inevitável (independentemente da taxa empregada), usar interfaces com alta taxa de transmissão reduz a latência geral. A Figura 4 ilustra o atraso para velocidades e tamanhos de frames distintos.

Entretanto, em alguns locais remotos, adição de banda não é uma opção, logo o atraso por serialização será significativo e sensível aos pacotes de voz. Como pode-se perceber na Figura 1Figura 4 a seguir, quanto maior o pacote que deve ser enviado, maior o atraso. Para esses casos, uma opção é utilizar a técnica de LFI ou Intercalação de Fragmentação de Links como mostrado naFigura 5. O LFI ameniza a questão do atraso de serialização cortando pacotes grandes em pedaços menores antes de serem enviados. Isso permite que o roteador mova o tráfego de VoIP crítico entre as partes fragmentadas do tráfego de dados, conforme Figura 5. Esse processo garante que os pacotes de voz tenham um atraso variável mais consistente e diminua significativamente o jitter de voz, entretanto, aumenta a sobrecarga nos links devido ao aumento no número de cabeçalhos criados para cada fragmento.

É importante notar que a partir de uma certa velocidade de canal, essa técnica não é mais interessante porque o *tradeoff* do aumento de sobrecarga começa a ser mais prejudicial ao sistema do que o ganho de diminuição do atraso de serialização.

Atraso = Tamanho do Frame (bits) / Banda do link (bps)

	1 Byte	64 Bytes	128 Bytes	256 Bytes	512 Bytes	1024 Bytes	1500 Bytes
56 kbps	143 us	9 ms	18 ms	36 ms	72 ms	144 ms	214 ms
64 kbps	125 us	8 ms	16 ms	32 ms	64 ms	126 ms	187 ms
128 kbps	62.5 us	4 ms	8 ms	16 ms	32 ms	64 ms	93 ms
256 kbps	31 us	2 ms	4 ms	8 ms	16 ms	32 ms	46 ms
512 kbps	15.5 us	1 ms	2 ms	4 ms	8 ms	16 ms	32 ms
768 kbps	10 us	640 us	1.28 ms	2.56 ms	5.12 ms	10.24 ms	15 ms
1536 kbps	5 us	320 us	640 us	1.28 ms	2.56 ms	5.12 ms	7.5 ms

Figura 4 - Atraso de Serialização para Diversos Tamanhos de Frames em Links de baixa Velocidade

Fonte: [7]

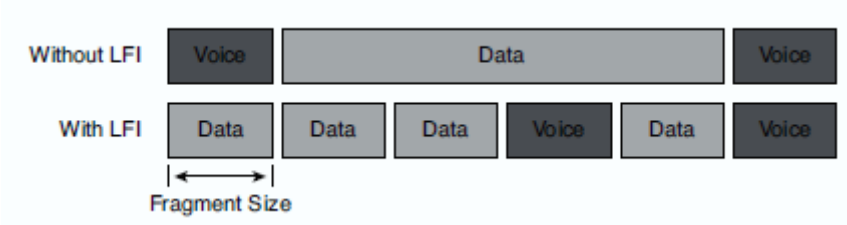


Figura 5 - Fragmentação e Interleaving

Fonte: [2]

- **Atraso de propagação**

Este tipo de atraso está relacionado à transferência física de sinais e é um componente fixo, sendo definido em função da distância física que os sinais precisam percorrer entre o ponto de origem e o ponto final de recebimento. O atraso de propagação depende, é claro, da tecnologia usada e também da distância ao longo da qual o sinal é transferido. Por exemplo, no caso da distância avaliada em dezenas de quilômetros, o impacto desse atraso é insignificante [6]. O impacto deste tipo de atraso é significativo no caso de redes de transporte (por exemplo, caminhos transoceânicos ou transcontinentais, etc.), onde linhas de longa distância são usadas. Neste tipo de redes, a fibra ótica é usada principalmente como meio. A transmissão da luz pode ser descrita como transmissão de ondas eletromagnéticas no ambiente. A taxa de transmissão depende do índice de refração desse ambiente e pode ser calculada pela seguinte fórmula [6]:

$v = c/\eta = 2.0\ 10^8$ [m/s];

onde:

$c = 3*10^8$ [m/s] – velocidade da luz no vácuo;

η = índice de refração para o vidro de silício com comprimento de onda λ = 1.33μm;

v = velocidade da luz se espalhando em fibra ótica [m/s];

Então obtem-se o valor de propagação pela seguinte fórmula:

$T = L/v$;

Onde:

T = atraso de propagação [ms];

L = comprimento [m];

O valor final do atraso de propagação pode ser em arrendondado em torno de 4,38 microssegundos por km. Este valor é irrelevante em comparação com outros valores de atraso nas transmissões de curta distância. Somente no caso de transmissões muito longas, o valor do atraso de propagação pode ser calculado em dezenas de milissegundos. Para estimar o atraso de propagação, uma estimativa popular de 10 microssegundos / milha ou 6 microssegundos / km é normalmente utilizada [8].

- **Atraso de fila**

O atraso de fila é uma função do congestionamento dos componentes da rede está congestionado e, em caso afirmativo, quais políticas de agendamento foram aplicadas para resolver eventos de congestionamento.

Os aplicativos em tempo real geralmente são mais sensíveis a jitter do que a latência como um todo porque os pacotes precisam ser recebidos em buffers de compensação de jitter antes de serem executados. Se um pacote não for recebido dentro da janela permitida pelo buffer de compensação de jitter,ele será perdido e poderá afetar a qualidade geral da voz ou da video-chamada.

Dado que a maioria dos fatores que contribuem para a latência da rede são fixos, atenção especial deve ser dada ao atraso na fila, porque esse é o único fator de latência diretamente sob o controle do administrador da rede, por meio de suas políticas de enfileiramento. Portanto, compreender o sistema de enfileiramento dos ativos de rede, incluindo a operação Tx-Ring (Buffer) e LLQ / CBWFQ (termos que serão abordados mais a frente), serve para ajudar os administradores de rede a otimizar essas políticas críticas.

Esse assunto será abordado com mais profundidade e tratado junto com as técnicas de QoS apresentadas mais à frente.

- **Atraso de processamento/envio**

É o tempo que um dispositivo de rede leva para armazenar em buffer um pacote e tomar decisão de encaminhamento. Incluído nessa decisão pode ser a interface para encaminhar o pacote, seja para descartar ou para enviar o pacote para uma lista de controle de acesso (ACL) ou política de segurança.

Esse atraso depende de fatores como a velocidade e utilização da CPU, arquitetura do roteador, e as facilidades configuradas nas interfaces. [9]

Em alguns cenários, atender a meta de latência pode simplesmente não ser possível devido às distâncias envolvidas e à relativa objetividade de seus respectivos caminhos de transmissão. Nesses cenários, se a meta de latência unidirecional G.114 de 150 ms não puder ser atendida, a ITU mostra que a qualidade da comunicação em tempo real não começa degradar significativamente até que a latência unidirecional exceda 200 ms [1], conforme ilustrado na Figura 6, da ITU G.114, onde se relaciona qualidade da fala em tempo real versus atraso absoluto. Para cada valor é apresentado um nivelamento de satisfação dos usuários:

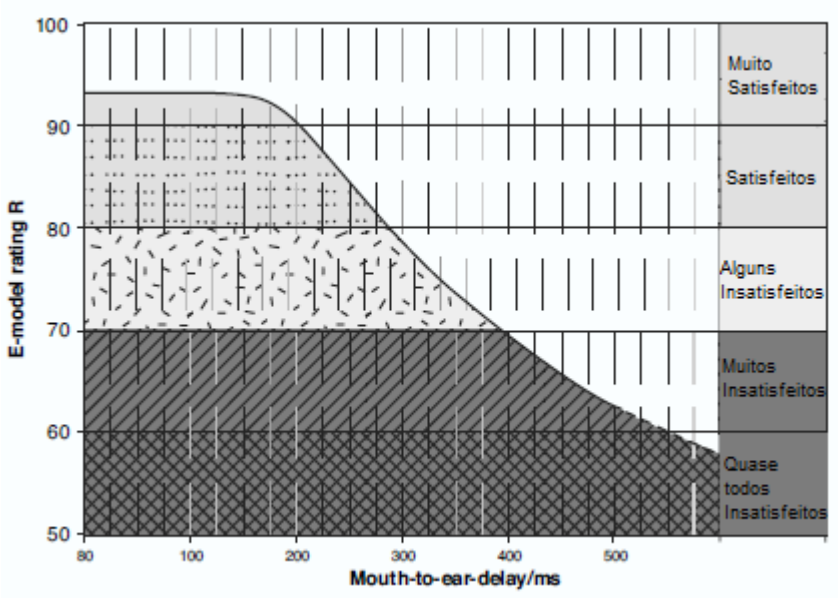


Figura 6 - ITU G.114 Gráfico de qualidade de voz em tempo real vs latência

Fonte: [10]

É importante ressaltar que o contexto até agora tem sido para circuitos WAN através de caminhos terrestres; para circuitos satélites, a latência esperada pode estar na faixa de 250 a 900 ms. Por exemplo, os sinais transmitidos por satélites

geoestacionários precisarão ser enviados para uma altitude de aproximadamente 36.000 km acima do nível do mar (do equador) para o espaço e depois de volta para a Terra novamente. Não há nada que se possa fazer para diminuir a latência em tal cenário, pois não se consegue aumentar a velocidade da luz ou das ondas de rádio. É necessário instruir a base de usuários para que as expectativas de desempenho realistas sejam definidas. [11]

2.2.1.2. Jitter

O jitter é a diferença entre o atraso unidirecional de pacotes selecionados. [12]

De outra forma, o jitter é a medida do tempo entre o momento em que um pacote deve chegar e quando efetivamente chega. Por exemplo, com uma taxa de transmissão de pacotes constante a cada 20 ms, espera-se que cada pacote chegue ao destino exatamente a cada 20 ms (jitter nulo). Esta situação nem sempre é o caso.

A maior culpada do jitter é a latência variável, ou seja, no enfileiramento causados por alterações de cargas no tráfego de rede. Outro fator são os pacotes que às vezes podem seguir caminhos diferentes, os quais possuem atrasos de propagação diferentes para chegar ao destino. Às vezes, é necessário eliminar o impacto da variação do atraso usando um buffer de compensação de jitter [6]. Esse buffer pode ser considerado como uma memória que armazena o fluxo dos pacotes de modo que a forma de onda de voz reconstruída não seja afetada pelo jitter. O mais importante é que o tamanho dessa memória é limitado pela variação de atraso e, por outro lado, esse tipo de jitter forma uma parte adicional (dinâmica) de atraso.

Os buffers de reprodução podem minimizar os efeitos do jitter, mas não podem eliminar valores muito altos de jitter. Embora seja esperado um pouco de jitter, valores elevados podem causar problemas de qualidade de voz, pois o gateway de mídia pode descartar pacotes que chegam fora de ordem.

2.2.1.3. Perda de pacotes

A perda de pacotes ocorre por vários motivos e, em alguns casos, é inevitável. Muitas vezes, a quantidade de tráfego que uma rede vai transportar é subestimada. Durante o congestionamento da rede, os roteadores e switches podem estourar seus buffers de filas e ser forçados a descartar pacotes. A perda de pacotes para aplicativos em tempo não real, como navegadores da Web e transferências de arquivos, é indesejável, mas não é crítica.

Para pacotes de voz, as restrições são mais rigorosas, embora algumas perdas possam ser toleradas. Uma certa quantidade de desperdiço de pacotes para serviços de voz pode ser aceitável, desde que essa perda seja distribuída por uma grande quantidade de usuários. Desde que a quantidade de perda de pacotes seja inferior a 5% para o número total de chamadas (ITU-T), a qualidade geralmente não é afetada negativamente. Geralmente é melhor descartar um pacote, em vez de aumentar a latência de todos os pacotes entregues armazenando-os em buffer.

2.2.2. BANDA

Cada site em uma implantação multisite geralmente é interconectado por uma WAN IP. Largura de banda em links WAN é limitada e relativamente cara. O objetivo é usar a largura de banda disponível da maneira mais eficiente possível. O tráfego desnecessário deve ser removido dos links IP WAN por meio da filtragem de conteúdo, firewalls e listas de controle de acesso (ACLs). Como a largura de banda disponível na WAN pode se tornar escassa, qualquer período de congestionamento pode resultar em degradação do serviço, a menos que a QoS seja implantada em toda a rede.

Os fluxos de voz RTP produzidos pelos terminais de telefones IP são um tamanho de pacote constante e previsível. Eles são pequenos em tamanho, mas enviados a uma taxa de frequência muito alta (ou seja, um grande número de pacotes de tamanho pequeno atravessando o fio ou o link de rede). Em locais com problemas de banda, links WAN de baixa velocidade, os fluxos de voz podem ser considerados

um desperdício se as técnicas para atenuar a utilização da banda não forem empregadas.

Algumas dessas técnicas para conservação WAN incluem [13]:

- Utilização de codecs de baixo consumo;
- Utilização de Compressão de Cabeçalho RTP;
- Implementação de anunciadores locais ou desabilitar anunciadores remotos;
- Implementação de conferências locais;
- Implementação de *Media Termination Points* (MTP) locais;
- Implementação de Transcoders;
- CONSIDERAÇÕES DE RECURSOS DE MÍDIA;
- Limitação do número de chamadas utilizando CAC ou RSVP

Os codecs são um dos fatores mais decisivos e expressivos no consumo da largura de banda disponível no link para o envio dos dados [14].

2.2.2.1. Codecs

A voz, é um sistema essencialmente analógico. Dependendo do tipo de sistema pelo qual esta irá trafegar, deve ser convertida para um formato digital. Após a digitalização, a representação digital da voz ou vídeo é transmitida para o receptor e depois convertida de volta para analógico, a fim de ser entendida pelo humano na outra extremidade. Na maior parte, esse processo de conversão analógico-digital e digital-analógico é tratado por um codec ou codificador-decodificador. Grande parte do trabalho feito com codecs foi em um esforço para melhorar a qualidade das ligações e reduzir a quantidade de largura de banda consumida pelo fluxo de voz através do uso de compressão.

Existem muitas técnicas diferentes usadas para lidar com esses fluxos de áudio e vídeo. A maioria dos codecs populares usados hoje é padronizada nas recomendações da ITU-T (International Telecommunications Union-Telecom), embora existam outros. Grande parte do trabalho feito com codecs foi em um esforço para reduzir a quantidade de largura de banda consumida pelo fluxo de voz através

do uso de compressão [15]. As implantações de VoIP exigem que os mesmos processos de conversão ocorram, embora nem sempre tenham as preocupações de largura de banda de topologias mais tradicionais.

Um codec tem a tarefa de transformar a onda analógica em uma série de amostras e, em seguida, fornecer um valor binário para essa amostra [16]. Isso é feito para que os dados de voz possam ser transferidos através das partes digitais da rede. No receptor, essas amostras devem ser desmontadas e convertidas novamente em uma forma de onda analógica que possa ser interpretada corretamente.

Quatro etapas são necessárias para transformar um sinal analógico em um sinal digital comprimido. As etapas ocorrem na seguinte ordem [17]:

1. Amostrar o sinal de voz analógico.
2. Quantizar a amostra.
3. Codificar a amostra digital.
4. Comprimir a amostra codificada.

Os codecs de áudio e vídeo mais populares são padronizados na série ITU-T [16] (G.711, G.722, G.726, G.729, entre outros):

Antes de aprofundar os detalhes dos codecs, é importante esclarecer alguns termos [18]:

Taxa de Bits do Codec (Kbps)

Com base no codec, esse é o número de bits por segundo que precisam ser transmitidos para fornecer uma chamada de voz. (taxa de bits de codec = tamanho de amostra de codec / intervalo de amostra de codec).

Tamanho da amostra de codec (bytes)

Com base no codec, esse é o número de bytes capturados pelo DSP (*Digital Signal Processor*) em cada intervalo de amostragem do codec. Por exemplo, o codificador G.711 utiliza *payloads* de 160 *bytes* enquanto o G.729 usa 20 *bytes*. A diferença está na codificação realizada pleo G.729.

Intervalo de amostra de codec (ms)

Este é o intervalo de amostragem no qual o codec opera. Por exemplo, os codificadores G711 e G.729 operam em intervalos de amostragem de 20 ms.

2.2.2.1.1. G.711

Esse codec define codificação de voz de 64Kbps (*payload* RTP). Isso é feito por amostragem do fluxo de voz 8000 vezes por segundo (125µs) e atribuindo 8 bits por amostra. Não se deve esquecer, no entanto, que isso é apenas o *payload* do pacote. Deve-se considerar ainda a sobrecarga do cabeçalho das outras camadas. A formação dos pacotes em cabeçalhos IP/UDP é realizada a cada 20ms.

Esse codec fornece o melhor desempenho em geral nas configurações tradicionais e de VoIP. No entanto, também consome a maior quantidade de largura de banda de rede. Isso normalmente não é uma preocupação para a *LANs*, mas é um fator a ser considerado e analisado quando se trata do estresse que pode suscitar em uma conexão WAN.

2.2.2.1.2. G.729

O G.729 ITU realiza a amostragem do sinal de voz na mesma taxa do G.711 de 8.000 amostras por segundo, de acordo com o teorema da taxa de Nyquist. Também como G.711, a taxa de bits é fixada em 8 por amostra. A principal diferença

entre G.711 e as variações de G.729 tem a ver com a compressão. Usando o que é conhecido como predição de revestimento conjugativo - estrutura algébrica - código – excitado (CS-ACELP), os codecs G.729 usam métodos de amostragem alternativos e expressões algébricas como um livro de códigos para prever a representação numérica real [17] .

O G.729 consome uma carga útil de 8Kbps mais a sobrecarga de empacotamento. A formação de pacotes é realizada igual ao G.711, 50 vezes por segundo (a cada 20ms) e com o *palyload* de 20 *bytes* (50*20*8 = 8kbps). O encapsulamento da voz digitalizada em um cabeçalho RTP, UDP, IP e camada 2 é extremamente alto em comparação com o tamanho dos dados de voz em si [14].

Essa sobrecarga presente independentemente do codec consiste em um cabeçalho RTP de 12 bytes, um cabeçalho UDP de 8 bytes e um cabeçalho IP de 20 bytes. Isso adiciona um overhead de 16kbps (40[Bytes] * 8[bits/Byte] / 20[ms]). Ainda assim, essa métrica só considerou o encapsulamento até a camada IP. O real encapsulamento da camada de enlace varia com base na tecnologia WAN, e por isso não foi considerado. Por exemplo, o tamanho de um cabeçalho de um GRE ou VPN IPsec através de um transporte de camada 2 é muito maior que um PPP.

Dessa forma, uma ligação utilizando o G.711 já consome uma banda de 80kbps por ligação. E a relação do cabeçalho para o *payload* é 1:4. Já o G.729 possui uma razão de 2:1, com um cabeçalho maior que a própria carga útil.

Historicamente, a escolha de qual codec usar foi baseada em termos de qualidade e custo [19]. Se dinheiro e largura de banda não fossem impedimentos, as organizações provavelmente usariam o G.711 exclusivamente.

No entanto, a conectividade externa gera despesa, e muitos dos codecs e algoritmos de compactação já foram desenvolvidos para obter uma qualidade semelhante à do G.711, usando menos recursos. Chamadas que viajam através de links que têm muito menos largura de banda, portanto, apresentam a seleção de codecs com um impacto evidente. Por exemplo, uma organização conectada ao mundo externo por meio de um link T1 é limitada a 1,544 Mbps. Usando o G.711, cada chamada consumiria 1:24 dessa capacidade. Um codec como o G.729 usa menos da metade disso. É claro que, em algum momento, aumentar a compactação para economizar largura de banda começa a afetar a qualidade das chamadas [20]

Tabela 1	***G711***	***G729***
Payload	160 B	20 B
Codec BW	64 kbps	8 kbps
Cabeçalho L3/L4:		
RTP	12	12
UDP	8	8
IP	20	20
Payload Total Cabeçalho:	40	40
Total antes do L2	200 B	60 B
BW por ligação (sem L2)	**80 kbps**	**24 kbps**

Frame Relay Overhead

Frame Relay	6 B	6 B
Total L2	206 B	66 B
BW por ligação	82.4 kbps	26.4 kbps

Multilink PPP Overhead

Frame Relay	6 B	6 B
Total L2	206 B	66 B
BW por ligação	82.4 kbps	26.4 kbps

Ethernet Overhead

Ethernet	18 B	18 B
Total L2	218 B	78 B
BW por ligação	87.2 kbps	31.2 kbps

cRTP Aplicado

cRTP (Comp. Cabeçalho)	40 -> 4	40 -> 4
Antes do L2	65.6 kbps	9.6 kbps
BW por ligação	68 kbps	12 kbps
Ganho com cRTP %	18%	60%

Casos especiais: Talvez seja necessário encapsular ou criptografar o tráfego de voz na rede. Se este for o caso, deve-se incluir a sobrecarga para qualquer protocolo que estiver sendo usado. Aqui está um pequeno resumo dos bytes adicionados por algumas das técnicas de criptografia mais populares que estão sendo usadas atualmente [21]:

IPSec: 50 - 57 bytes

Protocolo GRE de encapsulamento: 4 - 20 bytes (Cisco usa 8 bytes)

Marcação MPLS: 4 bytes por tag (pode ser mais de uma tag presente).

2.2.2.2. Compressão de Cabeçalho

Quando se usa compactação de cabeçalho RTP (cRTP), o cabeçalho IP, UDP e RTP podem ser compactados para 2 ou 4 bytes (dependendo se a soma de verificação UDP é preservada), comparado aos 40 bytes que é exigido por esses cabeçalhos se o cRTP não for utilizado.

O cRTP é ativado por link/nó, nas duas extremidades de uma interligação WAN ponto-a-ponto. Deve ser usado seletivamente em um link WAN lento, normalmente com menos de 768 kbps. Não é necessário ser ativado de ponta a ponta em todos os links WAN mais rápidos. A Figura 7 exemplifica a utilização do cRTP em um link WAN Frame-Relay, juntamente com a alteração do codec G711 para G729 [22].

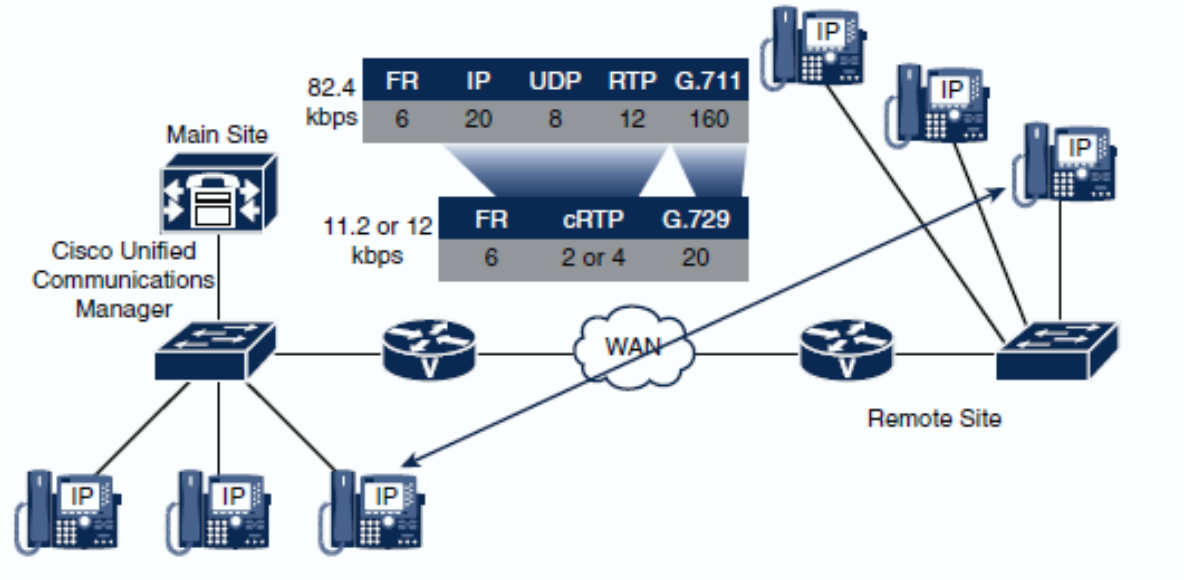


Figura 7 - Codecs de baixo consumo de banda e compressão de cabeçalho RT

Fonte: [19]

2.2.3. CONSIDERAÇÕES DE RECURSOS DE MÍDIA

2.2.3.1. Transcoders

Um transcodificador é necessário somente quando os dois pontos de extremidade de uma chamada não podem encontrar um codec comum permitido pela configuração da região. Por exemplo, os telefones IP remotos (que suportam G.711 e G.729) não podem usar o G.711 pela WAN IP, e o sistema de correio de voz da sede e a ponte de conferência de software não suportam o G.729. A central telefônica detecta esse problema com base em suas configurações de região e a negociação de capacidade executada durante a sinalização de configuração de chamada identifica a necessidade de um transcodificador.

Neste exemplo recorrente, um sistema de correio de voz que suporta apenas o G.711 é implantado no site principal. Um servidor de voz está fornecendo uma ponte de conferência de software que também suporta apenas o G.711. Se os telefones IP remotos forem configurados para usar o G.729 na WAN para conservar a largura de banda da WAN, eles não poderão ingressar em conferências ou acessar o sistema de

correio de voz. Para permitir que esses telefones IP usem o G.729 e acessem os serviços somente do G.711, um transcodificador de hardware é implantado no site principal no gateway usando recursos de DSP. As recomendações descrevem a alocação de transcodificadores centralmente no ambiente e pontes de conferência em locais remotos [19].

Os telefones IP remotos agora enviam fluxos de voz G.729 para o transcodificador pela WAN IP, o que economiza largura de banda. O transcodificador altera o fluxo para G.711 e o transmite para o sistema de bridge de conferência ou de correio de voz, permitindo que a conexão de áudio funcione.

Práticas Principais para Projeto de Transcodificador:

Ao projetar transcodificadores para permitir que dispositivos G.711 somente se comuniquem com telefones IP remotos usando G.729, prossegue-se com as seguintes etapas:

Etapa 1: Implementar um recurso de transcodificação baseado em hardware. Como os servidores de voz geralmente não suportam recursos de transcodificação de software, a única opção é usar um recurso de transcodificação baseado em hardware, configurando primeiro o transcodificador no gateway (roteador) e, em seguida, vincular o transcodificador à central.

Etapa 2: Implementar regiões de forma que apenas o G.729 seja permitido na WAN IP, e um transcodificador seja usado no qual o telefone IP tenha acesso por meio de um *pool* de dispositivos ou codificado como uma configuração para o telefone em si. Para fazer isso, todos os telefones IP e dispositivos G.711, como sistemas de correio de voz de terceiros ou pontes de conferência de software localizados no site da sede, são colocados em uma região (como região sede) e telefones IP remotos colocado em outra região (como a região da Filial).

2.2.3.2. MTP

Um MTP conecta dois fluxos de mídia e permite que eles sejam configurados ou desconfigurados de forma independente [23].

Um MTP pode ser usado como uma instância de conversão entre fluxos de áudio incompatíveis, para sincronizar o clock, ou para ativar serviços suplementares para dispositivos que não suportam alguma opção de recursos de algum protocolo.

Os pontos de terminação de mídia local (MTPs) podem ser necessários em situações em que é preciso converter G.711uLaw para G.711aLaw (como codecs) ou em situações em que é a transição do H.323 *slow-start* para H.323 *fast-start* é necessária.

Mais recentemente, com os troncos SIP se evidenciando nas soluçoes do mercado, para fornecer suporte do SIP *early-offer* para o SIP *delay-offer* pode ser exigido um MTP. Uma situação típica para a utilização do MTP é quando a operadora contratada estabelece que a forma de operação das ligações deve utilizar o SIP *early-offer*, porém alguns dos dispositivos do cliente não é capaz de suportar esse tipo de formato (equipamentos antigos por exemplo). Nesse caso, é necessário que seja configurado o MTP para realizar a transição e retificar as informações para o formato adequado.

São utilizados também para gerar tons DTMF quando integrados com um *endpoint* [24]. De forma geral, dispositivos VoIP antigos não suportam dígitos DTMF tradicionais por padrão [25]. Pode ser necessário permitir que os *endpoints* IP usem DTMF para se comunicar com serviços não baseados em VoIP. O relé DTMF pode ser usado para facilitar essa conversão. Existem vários métodos para configurar o relé DTMF. Todos eles exigem o uso de DSPs para transportar adequadamente o tom DTMF descompactado em uma rede IP.

Os MTPs locais são implantados usando uma mistura de configurações de software do servidor de voz ou DSPs baseados em hardware em locais remotos. Os MTPs implantados localmente impedem que o tráfego atravesse a WAN IP e use MTPs centralizados.

2.2.3.3. Anunciador

Um anunciador é uma função de software do aplicativo de streaming de mídia de voz IP que fornece a capacidade de transmitir mensagens faladas ou vários tons de progresso de chamada do sistema para um usuário [19]. Ele é capaz de enviar vários fluxos RTP (*Real-Time Transport Protocol*) unidirecionais para dispositivos telefones e *softphones* IP ou gateways e usa protocolos de sinalização para estabelecer o fluxo RTP. Os anúncios podem ser personalizados apropriadamente para substituir os tons de reordenamento.

Resumindo, o anunciador é uma maneira fácil para a central reproduzir mensagens de erro em uma voz humana, fornecendo ao usuário final uma mensagem de erro ou uma mensagem informativa, em vez de um simples tom de inválido. Exemplos disso são mensagens de para números não alocados, chamadas rejeitadas, números alterados, formato de número inválido e nível de precedência excedido.

Se os anúncios não devem ser enviados por um link WAN IP saturado, deve ser feita a configuração para que os telefones remotos não tenham acesso ao recurso de mídia do anunciador, que pode ser implementado em um design como o mostrado na Figura 8.

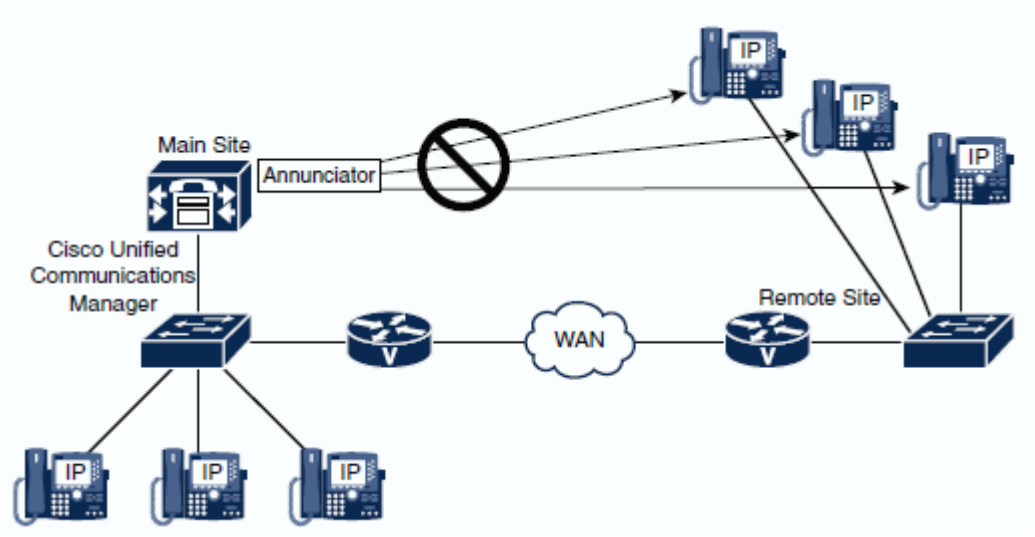


Figura 8- Desabilitar Anunciador Remoto

Fonte: [19]

2.2.3.4. Conferências

Em uma chamada padrão entre dois terminais o fluxo RTP e RTCP flui somente entre os dois *endpoints* [16]. Quando se inicia uma conferência, o fluxo de dados não segue mais esse padrão. Aparece nesse cenário um elemento que gerencia o fluxo RTP de cada terminal (recebe o fluxo de cada participante da conferência e entrega o fluxo apropriado a cada aparelho). Esse gerenciador é a ponte de conferência, podendo ser baseada em software e hardware [19].

Quando recursos de mídia estão envolvidos no fluxo de chamadas, não há transmissão direta de áudio de telefone IP para telefone IP. Os fluxos de tráfego de portadores RTP são enviados dos telefones IP para o recurso de ponte de conferência que mistura o áudio. O recurso de conferência mistura os fluxos de áudio e envia de volta um fluxo de áudio exclusivo para os telefones IP. O fluxo de áudio deve subtrair o fluxo de áudio da pessoa que recebe o fluxo de áudio para que nenhum eco seja ouvido. Alguns dispositivos de conferência, devido a limitações de processamento, misturam apenas os três alto-falantes mais altos. As mensagens de sinalização (tráfego de controle) são trocadas entre os telefones IP, o servidor de voz e o recurso de conferência (se estiver usando um recurso de hardware). O número de conferências individuais e o número máximo de participantes por conferência variam de acordo com o recurso em uso.

Conferências Modo Misto

A ponte de conferência baseada em software, implementada como um serviço do servidor de voz, suporta apenas conferências usando um único codec de áudio pois o servidor não realiza *transcoding* [26].

Algumas pontes de conferência de hardware podem suportar vários tipos de fluxo de baixa taxa de bits, como G.729, G.723 e iLBC. Uma conferência de modo misto é uma conferência na qual vários codecs de áudio são usados para diferentes fluxos de áudio [19]. Uma ponte de conferência de modo misto tem o fardo adicional de transcodificar os fluxos de portadora RTP. Conferências de modo misto limitam o número de participantes da conferência e conferências ativas com base nos recursos

do hardware. Existem várias famílias de *bridge* de conferência de hardware que devem ser investigadas.

A escalabilidade da conferência de software é limitada pela plataforma do servidor em que está sendo executada. Os recursos de conferência do servidor são limitados por padrão, pois é assumido que o servidor de voz estará executando tarefas de processamento de chamadas enquanto fornece recursos de conferência.

Conferências Locais

A implementação de uma ponte de conferência local é uma técnica em que os telefones IP enviam seus fluxos RTP para um recurso de conferência local configurado em um gateway. Essa técnica economiza largura de banda evitando fluxos RTP que atravessam a WAN destinada a uma ponte de conferência centralizada. Pontes de conferência baseadas em hardware são obtidas pela configuração de módulos de dados de voz de pacote (PVDMs) no roteador. Se houver um Assinante local na LAN ou no site de filial, as conferências G.711 baseadas em software poderão ser obtidas usando o Assinante local como um recurso de conferência. É importante observar que os PVDMs e as conferências baseadas em hardware podem suportar vários codecs mistos, enquanto a conferência baseada em software pode suportar o G.711 apenas por padrão.

Conforme mostrado na Figura 9, se uma ponte de conferência local for implantada no site remoto usando os DSPs ou PVDMs do gateway do site remoto, ela manterá os fluxos de voz fora da WAN IP para conferências nas quais todos os membros estão fisicamente localizados no site remoto. A mesma solução pode ser implementada para MTPs.

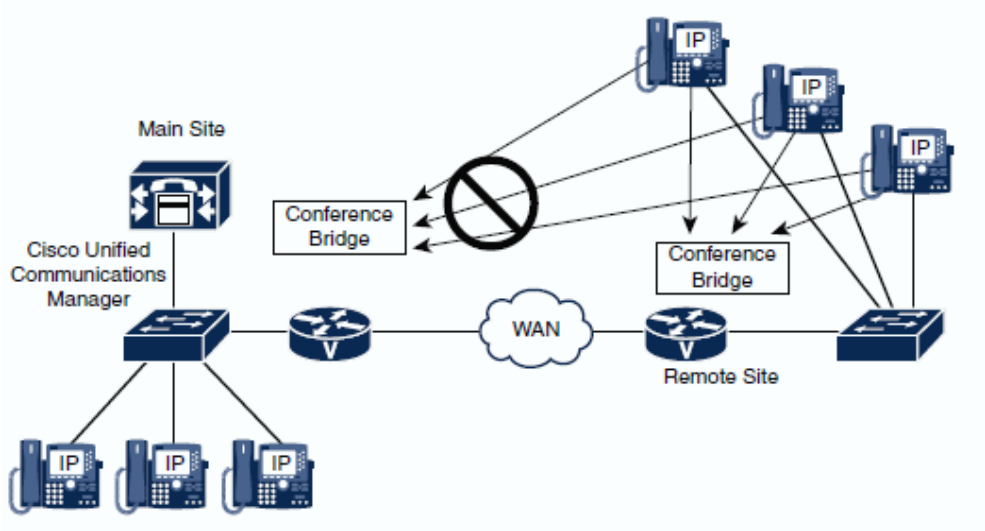


Figura 9- Conferência Local

Fonte: [19]

2.2.3.5. MoH

Para evitar a ineficiência do consumo de banda por pacotes de voz, todos os fluxos de voz desnecessários devem ser mantidos longe da WAN IP. Um ótimo exemplo disso são os recursos de mídia, em particular a música em espera (MoH), as pontes de conferência (CFB) e os anunciadores. Cada um dos tipos de recursos requer largura de banda adicional na WAN IP. Esses recursos de mídia podem ser otimizados de forma que não precisem atravessar a WAN IP o tempo todo, economizando largura de banda. Pode-se conseguir essa otimização colocando recursos de mídia local nos locais remotos, quando aplicável.

Multicast MoH a partir da *flash* de roteadores da filial é uma técnica para implantação em vários locais que usam processamento centralizado de chamadas junto com um servidor MoH centralizado que transmite o MoH de transmissão múltipla (*multicast*). O roteador é configurado para reproduzir o mesmo arquivo MoH e falsifica o endereço IP *multicast* usado pelo processo centralizado do servidor MoH. O gateway remoto é configurado para MoH de transmissão múltipla e envia continuamente esse fluxo, independentemente de o roteador estar no modo SRST (modo de *fallback*). Nem a central telefônica nem os telefones IP remotos estão cientes de que o gateway está

envolvido. Para eles, parece que um fluxo MOH de transmissão múltipla foi gerado pelo servidor MOH e recebido pelos telefones IP remotos.

Esse mecanismo funciona da seguinte maneira: O gateway “falsifica” o endereço IP e a porta MoH *multicast* da central, fornecendo *multicast* ao telefone IP na filial. Para conseguir isso, os telefones IP remotos são configurados para usar o servidor de voz MoH centralizado como sua origem MoH. O servidor MoH é configurado para *multicast* MoH (obrigatório), e o valor *max-hops* na configuração do servidor MOH é definido como 1 para as origens de áudio afetadas. O parâmetro *max-hops* especifica o valor de tempo de vida (TTL) usado no cabeçalho IP dos pacotes RTP. O servidor MoH e o gateway localizado no site remoto precisam usar o mesmo endereço *multicast* e o número da porta para seus fluxos. Dessa forma, pacotes MoH gerados pelo servidor no site central são descartados pelo roteador da camada de agregação do site central porque o TTL foi excedido. Como consequência, os pacotes MoH não cruzam a WAN IP. O gateway gera permanentemente um fluxo MoH de *multicast* com um endereço IP *multicast* e um número de porta idênticos, de forma que os telefones IP simplesmente escutem esse fluxo, pois parece que ele vem do servidor MoH.

Ao usar o MoH de *multicast* a partir da flash de um roteador de filial, o G.711 deve ser habilitado entre o servidor MoH e os telefones IP remotos. Isso é necessário porque o recurso SRST MoH do ramo suporta apenas o codec G.711. Portanto, o fluxo que é configurado pela central nas mensagens de sinalização também tem que ser G.711. Como os pacotes não são enviados pela WAN, a configuração do codec de alta largura de banda G.711 não é um problema, desde que seja habilitada apenas para o MoH. Todos os outros fluxos de áudio (como chamadas entre telefones) enviados pela WAN devem usar o codec G.729 de baixa largura de banda. Uma melhor prática recomendada envolve a criação de uma região MoH de *multicast* no servidor e o estabelecimento de sua relação com todas as outras regiões para o G.711.

Às vezes, o MOH multicast do flash do roteador do site remoto não pode ser usado. Por exemplo, talvez o roteador do site remoto não suporte o recurso ou as políticas de segurança da empresa sejam um fator limitante. Nesse caso, considerar as seguintes alternativas [19]:

- Uso de MOH de multicast na WAN versus unicast na WAN: Quando se usa o MoH de multicast na WAN de IP, o número de fluxos de MOH necessários pode ser reduzido significativamente. Assim, é necessária menos largura de banda em comparação com vários fluxos de MoH de unicast. A rede IP, no entanto, deve oferecer suporte ao roteamento multicast para o caminho do servidor MoH para os telefones IP remotos.
- Usando o G.729 para o MoH para sites remotos: Se o MoH multicast não for uma opção (por exemplo, porque o roteamento multicast não pode ser habilitado na rede), ainda pode-se reduzir a largura de banda consumida pelo MoH unicast. Se for alterado o codec que é usado para os fluxos de MOH para G.729 e, potencialmente, habilitar cRTP na WAN de IP, cada fluxo de MoH individual exigirá menos largura de banda e, portanto, reduzirá a carga no link de WAN. As economias de largura de banda são idênticas àquelas alcançadas ao usar G.729 e cRTP para fluxos de áudio padrão, o que foi discutido anteriormente.

2.2.4. QoS CONCEITO BÁSICO

A qualidade de serviço (QoS) refere-se à capacidade da rede de fornecer um serviço melhor ou especial a um conjunto de usuários, aplicativos e tráfego selecionados às custas de outros. O objetivo fundamental é gerenciar a contenção de recursos de rede para maximizar a experiência do usuário final de uma sessão - qualquer tipo de sessão. Respeitando que nem todos os pacotes de rede são iguais, eles não devem ser tratados igualmente. Os recursos de QoS implementam um sistema de “gerenciamento de arbitrariedade e desigualdade” na rede. [27]

Algumas sessões recebem prioridade sobre outras; Sessões sensíveis a atraso contornam filas de pacotes que mantêm sessões menos sensíveis; quando os buffers de filas estouram, os pacotes são descartados primeiro nas sessões capazes se recuperarem da perda ou naquelas que podem ser eliminadas com um impacto mínimo. Para liberar espaço para os pacotes pertencentes a sessões de alto impacto na empresa que não podem tolerar perdas sem afetar a experiência do usuário final, outras sessões são gerenciadas (isto é, pacotes são seletivamente atrasados ou

descartados quando a contenção surge) com base nas decisões de política de QoS implementadas na rede.

Isso é exatamente o que o tráfego de voz precisa ao cruzar a rede: serviço melhor ou “especial” que o tráfego de dados típico, como navegação na Web, transferências por FTP, tráfego de e-mail e assim por diante. O tráfego de voz precisa disso não tanto devido aos requisitos de largura de banda (o VoIP usa muito pouca largura de banda em comparação com a maioria dos aplicativos de dados), mas sim os requisitos de atraso. Ao contrário dos dados, o tempo que leva um pacote de voz para ir de uma ponta à outra da rede é crítico. Não há grande prejuízo se um pacote de dados cruzando a rede sofrer um atraso, uma transferência de arquivos pode levar mais alguns segundos para ser concluída ou uma página da Web levará um tempo a mais para ser carregada. Do ponto de vista de um usuário, isso não é grave. No entanto, se o tráfego de voz cruzando a rede sofrer atrasos, as conversas começarão a se sobrepor (uma pessoa começa a falar ao mesmo tempo que a outra); a conversa se quebra; e, em alguns casos extremos, a chamada de voz cai. A chave aqui é que se ouvem os problemas de rede em tempo real, pois isso atrapalha a conversa telefônica; as pessoas notam, e isso é inaceitável porque estamos acostumados com a fala soando natural e gera um impacto importante nos negócios.

Para combater esses problemas, é necessário garantir não apenas que haja largura de banda disponível para o tráfego VoIP, mas que o tráfego VoIP tenha prioridade absoluta no envio dos pacotes. Isso significa que, se ocorrer um gargalo na rede e um roteador tiver que enfileirar o tráfego antes de ser enviado, o roteador moverá o tráfego de voz em espera à frente do tráfego de dados e dará prioridade de transmissão aos pacotes de voz. Conseguir isso é o trabalho do QoS.

O QoS não é uma ferramenta em si, mas uma categoria de muitas ferramentas que visam dar controle total sobre o tráfego que cruza a rede. Como e quando se usa cada uma das ferramentas de QoS depende dos requisitos de rede do seu tráfego e das características (como largura de banda, atraso e assim por diante) da rede que suporta o tráfego.

As ferramentas de eficiência de link evitam que fluxos de pacotes de grande porte (como transferências de arquivos) afetem fluxos de pacotes de tamanho pequeno gravemente degradantes, como voz. A seguir estão as principais práticas associadas à implementação de QoS [2] e ilustrados na Figura 10:

- Identificar vários tipos de tráfego (voz, vídeo, sinalização, missão crítica, planejamento de recursos empresariais, dados e assim por diante).
- Dividir o tráfego em classes (tráfego em tempo real de voz e vídeo, tráfego de missão crítica, sinalização, tráfego menos importante e assim por diante).
- Criar uma política de QoS por classe, o que normalmente é feito em um link WAN de agregação ou porta de saída conectando uma rede de alta velocidade a uma rede de baixa velocidade. Essa política de QoS aplica um tratamento diferenciado às várias classes de tráfego criadas e identificadas.

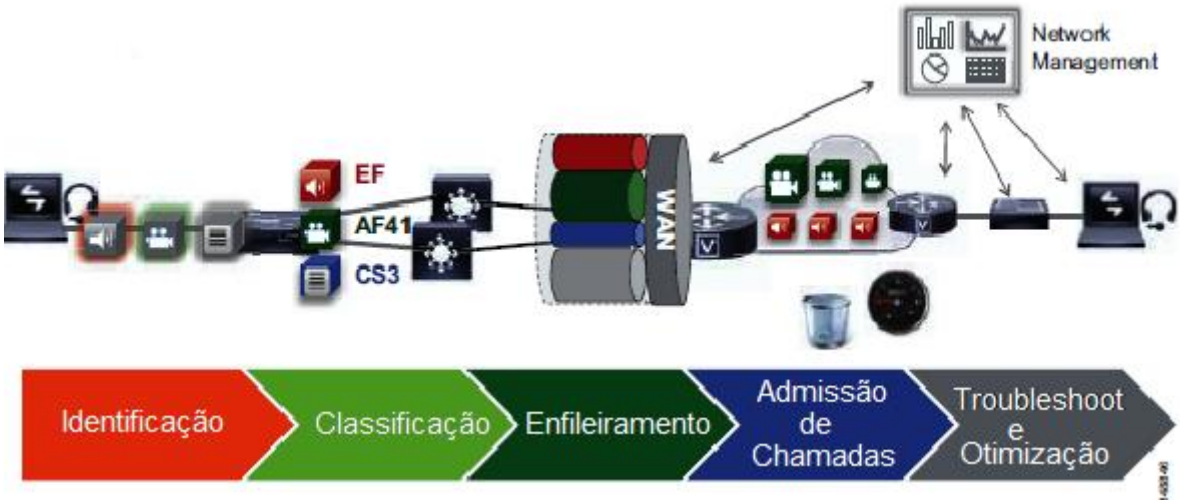


Figura 10 - Elementos da Arquitetura QoS para Colaboração

Fonte: *[15]*

2.2.4.1. CLASSIFICAÇÃO E MARCAÇÃO

2.2.4.1.1. Overview

O cabeçalho IP é definido no RFC 791, incluindo um campo de 1 byte chamado byte do tipo de serviço (ToS). O byte ToS foi planejado para ser usado como um campo para marcar um pacote para tratamento com ferramentas de QoS.

Uma série de RFCs chamados coletivamente de Serviços Diferenciados (DiffServ) surgiu mais tarde. O DiffServ precisava de mais de 3 bits para marcar os pacotes, então o DiffServ padronizou uma redefinição do byte ToS. O próprio byte ToS

foi renomeado para o campo Differentiated Services (DS) e o IP Precedence (IPP) foi substituído por um campo de 6 bits [28], conforme ilustrado na Figura 11.

Várias RFCs DiffServ sugerem um conjunto de valores para usar no campo DSCP e um significado implícito para essas configurações. Por exemplo, o RFC 3246 define um DSCP do decimal 46, com um nome Expedited Forwarding (EF – Recomendado que seja usado para marcar pacotes de voz). De acordo com essa RFC, os pacotes marcados como EF devem receber preferência de enfileiramento para que experimentem latência mínima, mas os pacotes devem ser controlados para evitar que eles assumam um link e impeçam que outros tipos de tráfego saiam de uma interface durante períodos em que o tráfego de alta prioridade atinge ou excede a largura de banda da interface. Essas configurações sugeridas e o comportamento associado de QoS recomendado ao usar cada configuração são chamados de Comportamento por pulsos (PHB) por DiffServ [29].

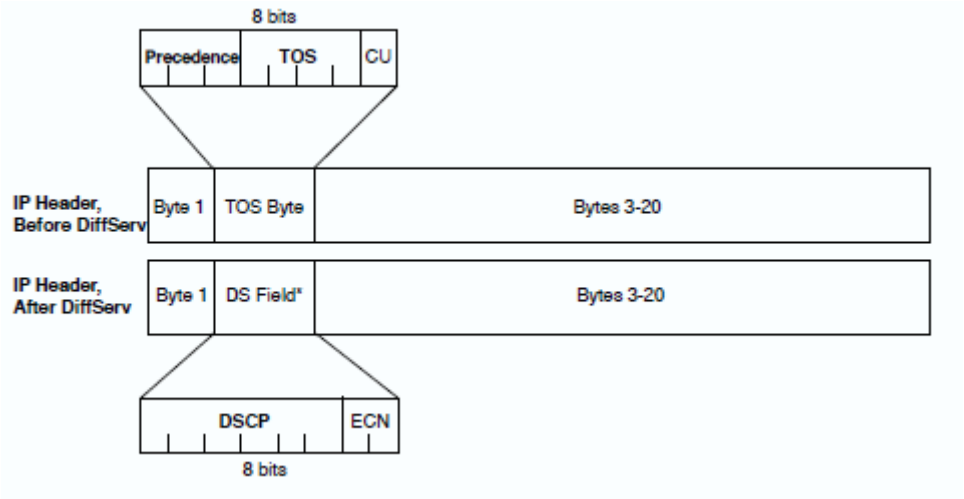


Figura 11- Comparação Byte ToS e Campo DS

Fonte: [28]

2.2.4.1.2. Marcação de Cabeçalho - Camada 2

Os outros campos de marcação possíveis residem dentro dos cabeçalhos da Camada 2, o que significa que os cabeçalhos são descartados quando encaminhados por um processo da Camada 3. Assim, esse não pode ser usado para transportar marcações de QoS além do *hop* atual.

ETHERNET

Ethernet suporta um campo de marcação QoS de 3 bits, mas o campo só existe quando o cabeçalho Ethernet inclui um cabeçalho de entroncamento 802.1Q ou ISL. Embora essas duas formas de cabeçalhos implementem métodos e marcações com nomes diferentes, geralmente a maioria dos comandos do IOS e administradores referem-se a esses campos como CoS, independentemente do tipo de entroncamento.

WAN

O Frame Relay e o ATM suportam um único bit que pode ser configurado para fins de QoS, mas esses bits únicos são destinados a um uso muito estrito relacionado à probabilidade de queda.

Nomeados como o bit de Elegibilidade de Descarte de Frame Relay (DE) e o bit de Prioridade de Perda de Célula ATM (CLP), esses bits podem ser definidos por um roteador ou por um comutador ATM ou de Frame Relay. Os recursos de roteador e queda de comutador podem ser configurados para permitir a queda mais agressiva de quadros e células que tenham o conjunto de bits DE ou CLP, respectivamente.

O MPLS define um campo de 3 bits chamado bit MPLS Experimental (EXP) que é destinado à marcação geral de QoS. Freqüentemente, as ferramentas C&M são usadas na borda de redes MPLS para remapear valores DSCP ou IPP para valores MPLS Experimental bit para fornecer QoS dentro da rede MPLS [28].

2.2.4.1.3. QoS Pré-Classificação

Com o tráfego não criptografado e não encapsulado, os roteadores podem corresponder e marcar valores de QoS e executar ações de entrada e saída com base em marcações, inspecionando os cabeçalhos de IP. No entanto, o que acontece se o tráfego for criptografado? Se encapsularmos o tráfego dentro de um túnel VPN, os cabeçalhos originais e o conteúdo do pacote não estarão disponíveis para inspeção. A única coisa com que se pode trabalhar é o byte ToS do pacote original, que é copiado automaticamente para o cabeçalho do túnel (no modo de transporte IPsec, no modo de encapsulamento e nos túneis GRE) quando o pacote é encapsulado.

O problema que surge deste comportamento inerente do encapsulamento de túneis é a incapacidade de um roteador tomar ações de QoS de saída com base em tráfego criptografado. Para atenuar essa limitação, surge um recurso chamado pré-classificação de QoS. Esse recurso pode ser ativado nos roteadores de ponto de extremidade de VPN para permitir que o roteador tome decisões de QoS de saída com base no tráfego original, antes do encapsulamento, em vez de usar apenas o cabeçalho do túnel. A pré-classificação de QoS funciona mantendo o tráfego original e não criptografado na memória até que as ações de QoS da saída sejam realizadas.

2.2.4.2. GERENCIAMENTO DE CONGESTIONAMENTO

2.2.4.2.1. CBWFQ e LLQ

O CBWFQ e LLQ são apresentados porque são os principais mecanismos atuais de enfileiramento baseados em roteador recomendados para redes de mídia avançada em que diferentes tipos de tráfego compartilham a mesma mídia de transmissão [30]. Os métodos de enfileiramento herdados não serão abordados.

Os atuais e muito mais recentes mecanismos de enfileiramento recomendados e adequados para redes multimídia procuraram combinar os melhores recursos dos algoritmos legados e, ao mesmo tempo, minimizar suas desvantagens. O tráfego sensível ao *delay* em tempo real requer dois atributos de um algoritmo de

enfileiramento: uma garantia absoluta de largura de banda e uma garantia de atraso. Outros tráfegos não precisam ser reprimidos na presença de tráfego em tempo real. Os algoritmos atuais de enfileiramento recomendados são os seguintes:

CBWFQ: Um algoritmo de enfileiramento híbrido combinando uma garantia de largura de banda com a integridade dinâmica a outros fluxos dentro de uma classe de tráfego. Ele não fornece garantia de latência e, como tal, é adequado apenas para gerenciamento de tráfego de dados.

LLQ: Esse método adiciona um recurso de prioridade estrita ao CBWFQ e, portanto, é adequado para misturas de tráfego em tempo real e não-tempo real. Ele fornece garantias de latência e largura de banda.

O CBWFQ permite a criação de filas, relacionadas a classes de tráfego. Cada fila é atendida com base na largura de banda atribuída a essa classe. O CBWFQ é configurado usando a palavra-chave de largura de banda em um mapa de políticas. Com o CBWFQ, uma largura de banda mínima é explicitamente definida e aplicada. A largura de banda pode ser especificada em termos absolutos ou percentuais. É feita uma garantia mínima de largura de banda, a qual é associada à classe. Em segundo lugar, todos os outros tráfegos (que se enquadram na classe *default*) são agrupados na fila de espera.

No caso de congestionamento, o Tx-Ring (mecanismo de enfileiramento padrão) para a interface é preenchido e empurra os pacotes de volta para as filas CBWFQ (se configuradas). Cada classe CBWFQ é atribuída a sua própria fila. As filas do CBWFQ também podem ter um gerenciador de fila justa aplicado (usando a palavra-chave *fair-queue* em um mapa de políticas) para gerenciar fluxos múltiplos que competem por uma única fila. Além disso, cada fila do CBWFQ é atendida em um modo *round-robin* ponderado (WRR) com base na largura de banda atribuída a cada classe. O planejador CBWFQ então encaminha os pacotes para o Tx-Ring.

O LLQ é essencialmente o CBWFQ combinado com um único PQ estrito. O tráfego atribuído à fila recebe prioridade absoluta, usando a palavra-chave *priority*, é atendido até sua largura de banda atribuída antes que outras filas do CBWFQ sejam atendidas. Todo o tráfego em tempo real deve ser configurado para ser atendido pela fila de prioridade. Diversas classes de tráfego em tempo real podem ser definidas e garantias de largura de banda separadas são fornecidas a cada uma delas, mas uma

única fila de prioridades programa todo o tráfego combinado. Assim como no CBWFQ, pode-se configurar o LLQ com alocações de largura de banda absoluta ou baseada em porcentagem.

2.2.4.2.2. Sincronização Global

A sincronização global TCP em redes de computadores pode acontecer com fluxos TCP / IP durante períodos de congestionamento, pois cada remetente reduzirá sua taxa de transmissão ao mesmo tempo em que ocorre a perda de pacotes.

Os roteadores normalmente têm filas de pacotes em *buffer* para permitir que eles mantenham os pacotes quando a rede está ocupada, em vez de descartá-los.

Como os roteadores têm recursos limitados, o tamanho dessas filas também é limitado. A técnica mais simples para limitar o tamanho da fila é conhecida como *tail drop*. A fila tem permissão para preencher seu tamanho máximo e, em seguida, todos os novos pacotes são simplesmente descartados, até que haja espaço na fila novamente.

Isso causa problemas quando usado em roteadores que manipulam vários fluxos TCP, especialmente quando há tráfego intermitente. Enquanto a rede está estável, a fila está constantemente cheia e não há problemas, exceto que a fila cheia resulta em alta latência. No entanto, a introdução de um surto súbito de tráfego pode fazer com que um grande número de fluxos já estabelecidos e estáveis percam pacotes simultaneamente.

O TCP tem recuperação automática de pacotes descartados, que ele interpreta como congestionamento na rede (o que geralmente é correto). O remetente reduz sua taxa de envio por um determinado período de tempo e, em seguida, tenta descobrir se a rede não está mais congestionada, aumentando a taxa novamente. Isso é conhecido como o algoritmo de início lento. [31]

Quase todos os remetentes usarão o mesmo atraso de tempo antes de aumentar suas taxas. Quando esses atrasos expirarem, ao mesmo tempo, todos os remetentes enviarão pacotes adicionais, a fila do roteador voltará a estourar, mais

pacotes serão descartados, todos os remetentes recuarão por um atraso fixo... e assim sucessivamente;

Esse padrão de cada remetente diminuindo e aumentando as taxas de transmissão ao mesmo tempo que outros remetentes é chamado de "sincronização global" e leva ao uso ineficiente da largura de banda, devido ao grande número de pacotes descartados, que devem ser retransmitidos e os remetentes têm uma taxa de envio reduzida, em comparação com o estado estável, enquanto são recuados, após cada perda.

Este problema tem sido objeto de muita pesquisa [28]. O consenso parece ser que o algoritmo de *tail-drop* é a principal causa do problema, e outros algoritmos de gerenciamento de tamanho de fila, como Random Early Detection (RED) e Weighted RED, reduzirão a probabilidade de sincronização global, bem como diminuirão o tamanho das filas em face de carga pesada e tráfego em rajadas [32].

2.2.4.2.3. WRED

O *tail drop* pode ter um efeito geral negativo no tráfego da rede, particularmente no tráfego TCP. Quando os pacotes são perdidos, por qualquer motivo, os remetentes TCP diminuem sua taxa de envio de dados devido ao *windowing*. Quando o *tail drop* ocorre, vários pacotes são perdidos e as conexões TCP diminuem ainda mais. Além disso, a maioria das redes envia uma porcentagem muito maior de tráfego TCP do que o tráfego UDP, o que significa que a carga geral da rede tende a diminuir depois que vários pacotes são descartados. [11]

A taxa de transferência geral pode ser melhorada descartando-se alguns pacotes à medida que a fila começa a ser preenchida, em vez de esperar pelo impacto maior do *tail drop*. Foi desenvolvido o mecanismo de detecção antecipada aleatória ponderada (WRED) com o objetivo específico de monitorar o tamanho da fila e descartar uma porcentagem dos pacotes na fila para melhorar o desempenho geral da rede. Conforme a fila fica mais longa, o WRED começa a descartar mais pacotes, esperando que uma pequena redução na carga oferecida a seguir seja suficiente para impedir o preenchimento da fila.

O WRED usa várias configurações numéricas ao tomar suas decisões. Primeiro, o WRED usa a profundidade da fila média medida ao decidir se uma fila foi preenchida o suficiente para começar a descartar pacotes. Em seguida, o WRED compara a profundidade média com um limite de fila mínimo e máximo, executando diferentes ações de descarte, dependendo do resultado.

Quando a profundidade média da fila é muito baixa ou muito alta, as ações são um pouco óbvias. Quando a profundidade média aumenta acima do limite máximo, o WRED descarta todos os novos pacotes. Caso a profundidade média da fila esteja entre os dois limites, o WRED descarta uma porcentagem de pacotes. A porcentagem cresce linearmente à medida que a profundidade média da fila aumenta do limite mínimo até o máximo, conforme mostrado na Figura 12 - WRED Lógica de Descarte (que mostra uma configuração de exemplo).

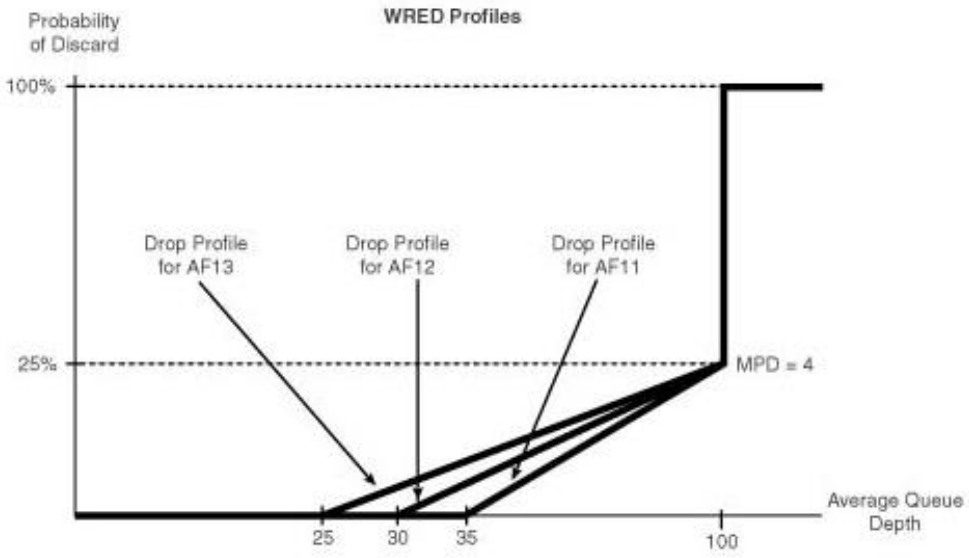


Figura 12 - WRED Lógica de Descarte

Fonte: [28]

Conforme apresentado na figura acima, fator do WRED que afeta sua lógica é o denominador de probabilidade de marca (MPD), a partir do qual a porcentagem máxima de 25% é atingida. O sistema calcula a porcentagem de descarte usada no limite máximo com base na fórmula 1/MPD. Na figura, um MPD de 4 gera um valor calculado de 1/4, o que significa que a taxa de descarte aumenta de 0% a 25%, à medida que a profundidade média da fila cresce do limite mínimo até o máximo. Além

disso, quando o WRED descarta pacotes, ele escolhe aleatoriamente os pacotes a serem descartados [25].

2.2.4.2.4. POLICE & SHAPING

Os mecanismos de *shaping e policing* são mecanismos de condicionamento de tráfego usados em uma rede para controlar as taxas de tráfego. Ambos os mecanismos usam a classificação para identificar e diferenciar o tráfego. Ambos medem a taxa de tráfego e comparam essa taxa com a política de traffic shaping ou traffic policing configurada [27], conforme demonstrado na Figura 13. A diferença entre políticas de shaping e policing pode ser descrita em termos de sua implementação.

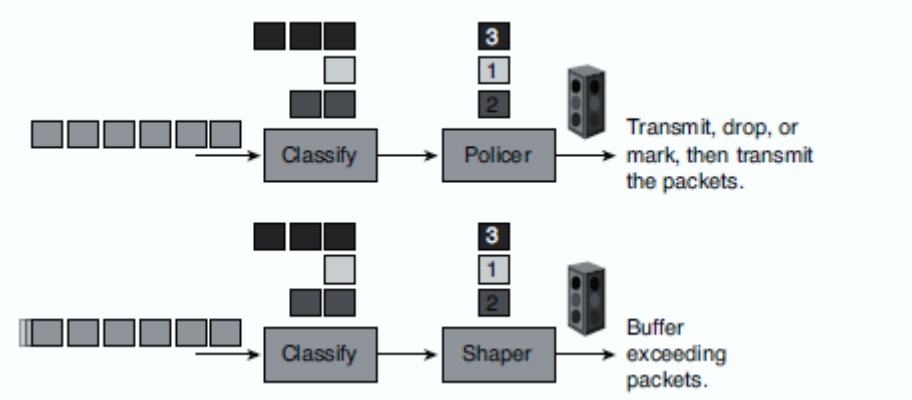


Figura 13- Police & Shaping classificação e políticas

Fonte: [2]

O *shaping* reduz o tráfego excessivo para que este fique dentro de uma taxa desejada. Com essa técnica, as intermitências de tráfego são suavizadas ao enfileirar o excesso de tráfego para produzir um fluxo de dados mais estável. Reduzir explosões de tráfego ajuda a reduzir o congestionamento na rede.

O *policing* descarta o excesso de tráfego para manter o fluxo dentro dos limites de taxa especificados. O *policiing* não introduz nenhum atraso no tráfego que esteja em conformidade com as políticas de tráfego. O policiamento de tráfego pode causar mais retransmissões TCP, porque o tráfego além dos limites especificados é simplesmente descartado.

O *shaping* resolve alguns tipos gerais de problemas que podem ocorrer em redes multiacesso [25]: Se um provedor de serviços descartar intencionalmente qualquer tráfego em um circuito quando a taxa de tráfego exceder a taxa de informação comprometida (CIR), faz sentido que o roteador não envie tráfego mais rápido que o CIR. Assim o cliente mantém controle local da regulamentação do tráfego. O segundo tipo de problema é o bloqueio de saída.

O bloqueio de saída é o segundo problema para o qual o *shaping* fornece algum alívio. Este fenômeno ocorre quando redes WAN com diversos nós e largura de banda assimétricas em cada link são usados ao longo do caminho do tráfego. Se o *shaping* não for usado ocorre um enfileiramento muito alto nos pontos mais lentos, causando atraso e estouro de *buffer*, e consequentemente quedas.

Já o *policing* é mais utilizado para satisfazer algum destes requisitos [27]:

Limitar a taxa de acesso em uma interface quando uma infraestrutura física de alta velocidade é usada no transporte. A limitação de taxa é frequentemente usada pelos provedores de serviços para oferecer acesso aos clientes.

Remarcar o excesso de tráfego com uma prioridade mais baixa na Camada 2 ou na Camada 3, antes de enviar o excesso de tráfego.

A Figura 14 ilustra a diferença de tratamento entre o *policing* e *shaping*, evidenciando que o *policing* apenas descarta os pacotes em excesso e o *shaping* tenta promover a adequação à taxa de transmissão configurada

.

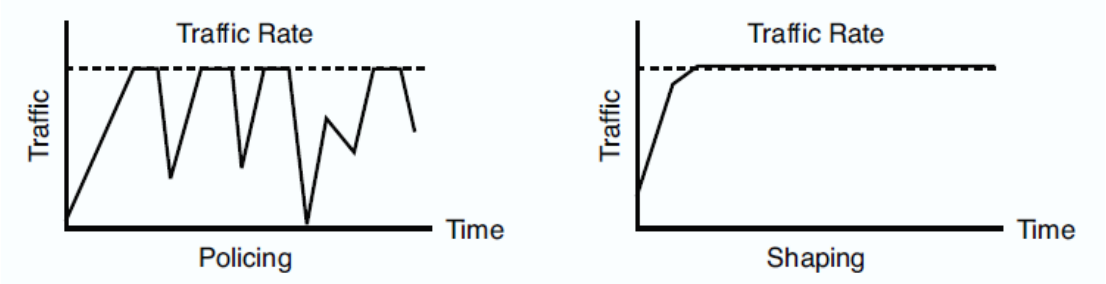


Figura 14- Police & Shaping Comparação

Fonte: [2]

2.2.5. PRINCÍPIOS DE DESIGN E ESTRATÉGIAS QoS

Princípios de Design

Tendo revisado os principais conjuntos de ferramentas de QoS e os requisitos de negócios e aplicativos, será possível reunir todos esses recursos para formar uma estratégia de QoS de ponta a ponta para uma organização.

Geralmente há mais de uma solução para quaisquer desafios de QoS, especialmente com o rico conjunto de ferramentas de QoS à disposição, havendo mais de uma ferramenta para resolver um determinado problema. Algumas ferramentas podem ser cruas, mas eficazes; outras podem ser elegantes e eficientes. Por exemplo, foi apresentada uma visão geral de mecanismos que poderiam atiginr o resultado de marcar e controlar o fluxo de um pacote como o CBWFQ, LLQ, WRED, *shaping* e *policing*. Todos podem alcançar o resultado de controlar o fluxo de um pacote, mas o fazem de maneira diferente e, portanto, essas ferramentas são melhor usadas em diferentes contextos de projeto.

Dessa forma, para combinar a ferramenta de QoS certa com o desafio certo de QoS, é útil apontar algumas das práticas recomendadas de design de QoS: [29] [33]

QoS Hardware versus Software

- Sempre ativar as políticas de QoS no hardware - em vez de software - quando houver uma opção.

Classificação e Marcação

- Classificar e marcar os aplicativos o mais próximo possível de suas origens, validando viabilidade técnica e administrativa.
- Usar a marcação DSCP sempre que possível.
- Seguir as marcações com base em padrões PHB e DSCP para garantir a interoperabilidade e expansão futura.

Policiamento e Markdown

- Ativar o *policing* o mais próximo possível de sua fonte.

- Sempre que possível, marcar de acordo com as regras baseadas em padrões.

Enfileiramento e Eliminação de Melhores Práticas

- Ativar as políticas de enfileiramento em todos os nós com potencial de congestionamento.
- Sempre que possível, atribuir cada classe de aplicativo à sua própria fila dedicada.
- Usar apenas plataformas / provedores de serviços que ofereçam no mínimo quatro comportamentos de enfileiramento baseados em padrões:
    - EF RFC 3246
    - AF RFC 2597
    - DF RFC 2474
    - LE RFC 3662

Recomendações da Fila EF: A Regra 33% LLQ

- Limitar a quantidade de enfileiramento de prioridade restrita a 33% da capacidade de largura de banda do link.
- Gerenciar o tráfego de prioridade estrita com um mecanismo de controle de admissão (CAC).
- Não habilitar o WRED nesta fila.

Recomendações da Fila AF

- Provisionar alocações de largura de banda garantida de acordo com os requisitos da aplicação.
- Ativar o WRED baseado em DSCP nessas filas.

Recomendações da Fila DF

- Prover pelo menos 25% da largura de banda do link para a classe padrão *Best Effort*.
- Ativar o WRED na classe padrão.

Recomendações da Fila da Classe Scavenger

- Atribuir uma largura de banda mínima à fila da classe Scavenger.
- O WRED não é necessário na fila da classe Scavenger.

Estratégias de Design

Tendo analisado esses princípios de design de QoS de melhores práticas, deve-se montá-los em uma estratégia de ponta a ponta para um determinado negócio ou organização. Obviamente, nunca haverá uma solução de tamanho único, porque os objetivos e as restrições de negócios variam. Portanto, três modelos genéricos são apresentados como exemplos: [27]

- Estratégia de QoS do modelo de 4 classes (Figura 15);
- Estratégia de QoS do modelo de 8 classes (Figura 16);
- Estratégia de QoS de 12 classes (Figura 17);

Esses modelos possuem a estratégia de 4 classes como base e mantêm essa estrutura, adicionando mais classes e especificações conforme a quantidade de programas, aplicativos e necessidade de um negócio se mostram presentes. Será apresentado o modelo de 4 classes com mais granularidade, enquanto que os outros dois modelos serão apenas ilustrados, uma vez que seguem a mesma linha de raciocínio.

4 Classes:

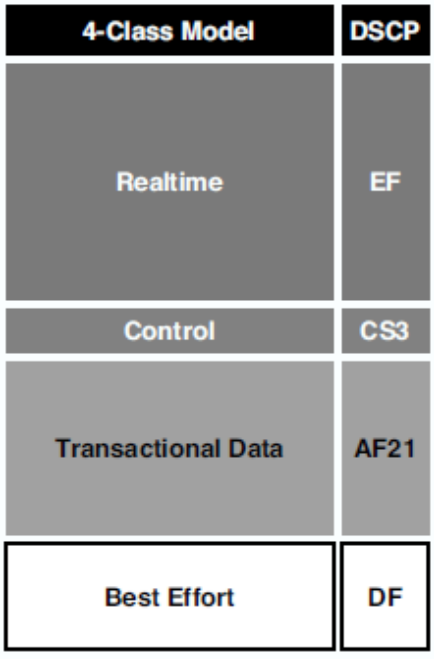
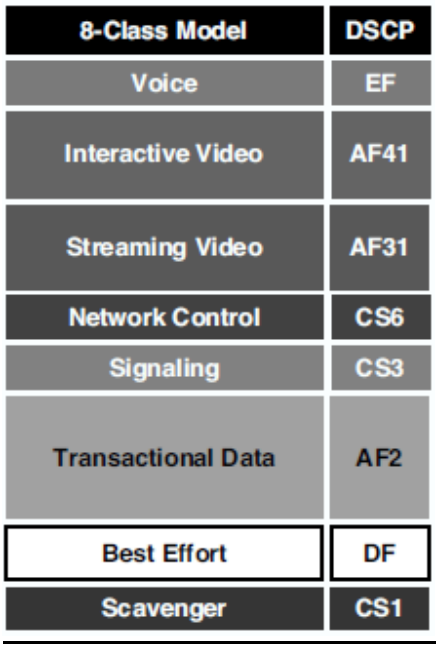


Figura 15 - Estrutura do Modelo QoS de 4 Classes

O modelo de QoS de quatro classes representa uma estratégia básica de QoS de ponta a ponta. Quando as empresas começaram a implantar a telefonia IP, elas precisavam de um modelo de três classes (uma para voz, sinalização e melhor esforço /padrão). O modelo de quatro classes insere apenas uma classe adicional a esses requisitos mínimos para telefonia IP, conforme decidido pelos objetivos de negócios de QoS: A classe adicional pode ser uma classe AF para aplicativos de dados transacionais (conforme mostrado no exemplo a seguir) ou pode ser uma classe AF para tráfego de vídeo/multimídia/aplicativo específico, ou pode até ser uma classe Scavenger. Neste exemplo de modelo QoS de quatro classes, as classes, marcações e tratamentos recomendados são os seguintes:

- Voz: Marcada EF e tratada com um EF PHB, mas limitada a 33 por cento da largura de banda de prioridade estrita.
- Controle: Marcado CS3 e provisionado com uma alocação de largura de banda garantida de 7 por cento. Embora essa classe seja destinada principalmente ao serviço de tráfego de sinalização, outro tráfego de controle também pode ser atendido dentro dessa classe.
- Dados transacionais: Marcado AF21 e tratado com um AF, provisionado com alocação de largura de banda garantida de 35%, com WRED baseado em DSCP ativado.
- Melhor esforço: Marcado DF e provisionado com uma alocação de largura de banda garantida de 25 por cento, com WRED habilitado.

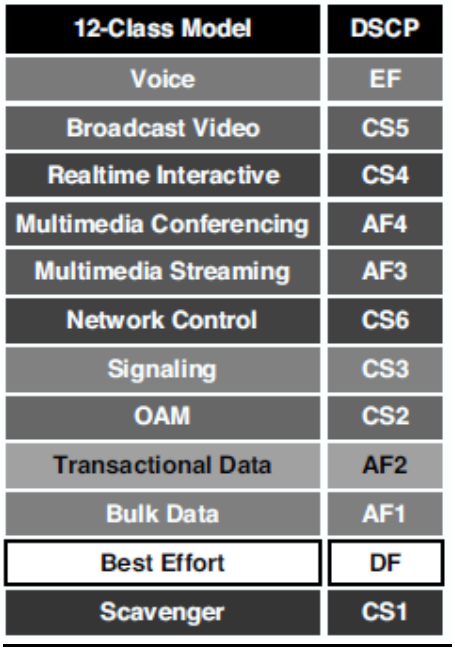
8 Classes



8-Class Model	DSCP
Voice	EF
Interactive Video	AF41
Streaming Video	AF31
Network Control	CS6
Signaling	CS3
Transactional Data	AF2
Best Effort	DF
Scavenger	CS1

Figura 16- Estrutura do Modelo QoS de 8 Classes

12 Classes



12-Class Model	DSCP
Voice	EF
Broadcast Video	CS5
Realtime Interactive	CS4
Multimedia Conferencing	AF4
Multimedia Streaming	AF3
Network Control	CS6
Signaling	CS3
OAM	CS2
Transactional Data	AF2
Bulk Data	AF1
Best Effort	DF
Scavenger	CS1

Figura *17- Estrutura do Modelo QoS de 12 Classes*

Como podemos ver na Figura 18, as classes 8 e 12 são derivadas da classe 4, porém com mais granularidade, caso o cliente tenha na sua rede equipamentos e dados de streaming de vídeo, multimídia, entre outros presentes na rede.

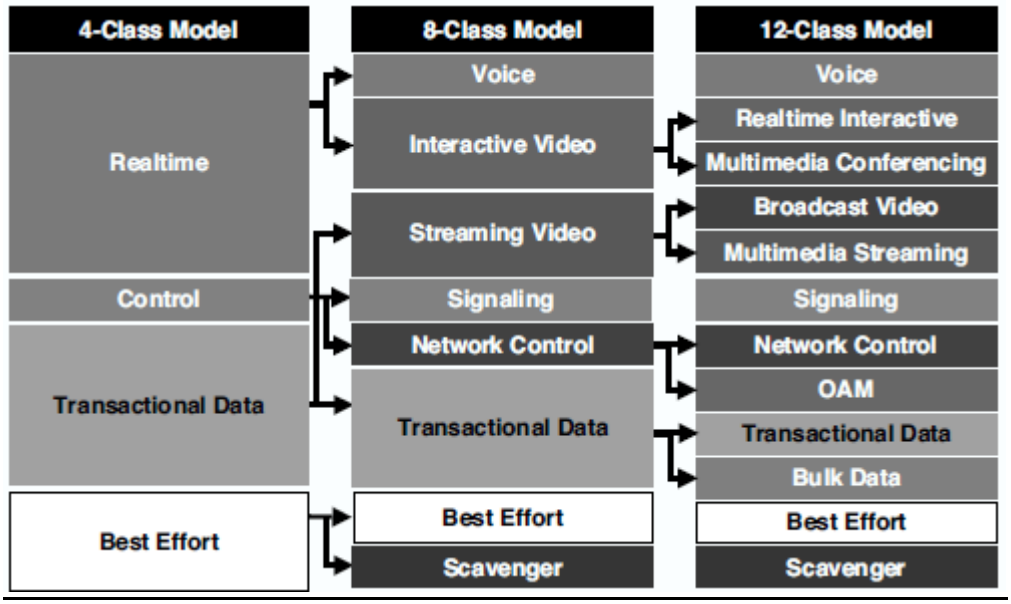


Figura 18- Comparação dos 3 modelos de Classe

2.2.6. CAC - LIMITAÇÃO DO NÚMERO DE CHAMADAS

Quando se implementa mecanismos de QoS atraves do link WAN, se garante uma quantidade de banda para as aplicações de voz e, normalmente, se estabelece um limite dessa banda para que as aplicações de voz não consumam toda a capacidade da conexão, garantindo assim que outros serviços pussuam um fluxo adequado para que sejam atendidos devidamente, e vice-versa. [34]

Entretanto, esse limite garante apenas que as aplicações de voz não vão afetar nem serão afetadas por outros tipos de serviços. É necessário um mecanismo que proteja os serviços de voz deles mesmos. Por exemplo, considerando que existe um link WAN sendo utilizado e que este tenha um limite de 128 kbps para o fluxo de voz. Esse limite garante boa qualidade para 1 ligação utilizando o codec G.711. Caso sejam feitas mais ligações atravessando esse canal, será necessário mais que a capacidade permitida e todas as chamadas começarão a apresentar problemas. Uma vez que um fluxo de voz não possui preferência sobre outro, todos serão tratados pelo modelo de melhor esforço, prejudicando todas as chamadas.

Nesse contexto surge o CAC, *Call Admission Control*, um mecanismo para assegurar que caso seja feita uma ligação e esta acarrete na extrapolação do limite de ligações que preservem a qualidade do funcionamento dos serviços, esta seja

negada de atravessar o link WAN. O funcionamento padrão desse tipo de ferramenta é simplesmente bloquear a tentativa da chamada, o que, na maioria dos casos não é o comportamento desejado.

Resumindo, o objetivo do CAC é evitar a assinatura excessiva do link WAN quando as chamadas de voz / vídeo são colocadas de um site físico / lógico para outro. Em outras palavras, há um limite para quantas chamadas são permitidas em toda a rede em um link WAN. As chamadas de voz / vídeo são admitidas na rede somente enquanto a rede puder garantir qualidade de serviço (QoS) suficiente. [35]

2.2.6.1. AAR

Como consequência da implementação do CAC, as chamadas além de um limite específico de largura de banda ou característica de link serão descartadas para manter a qualidade das chamadas existentes.

Isso se traduz diretamente em usuários finais que perdem ligações ou não conseguem realizá-las nos horários de pico. Esse resultado pode ser indesejável, especialmente quando há chamadas críticas a serem feitas. É possível redirecionar automaticamente as chamadas através da PSTN ou de outras redes quando as chamadas são bloqueadas devido à largura de banda insuficiente. Esse mecanismo é conhecido como roteamento alternativo automatizado (AAR). Usando AAR, o chamador não precisa desligar e rediscar o número chamado. O recurso AAR permite que a central estabeleça um caminho alternativo para a mídia de voz quando o caminho preferencial (IP) entre dois pontos finais intracluster ficar sem largura de banda disponível, conforme determinado pelo mecanismo de localidade CAC.

2.2.7. ALTA DISPONIBILIDADE

Outro tópico para discussão é alta disponibilidade em implantações de vários sites. Deve-se garantir que o processamento de chamadas e os recursos permaneçam intactos durante a falha ou congestionamento da WAN. A alta disponibilidade pode ser obtida de várias maneiras: [19]

- Backup de PSTN: Utilizando a PSTN como backup para chamadas entre sites internas e rotas de menor custo. Os telefones IP que normalmente transmitem RTP e a sinalização pela WAN para chamadas de site a site podem usar o PSTN como um caminho alternativo no caso de a WAN ficar inativa. Uma observação importante é que a manipulação de dígitos precisa ocorrer para expandir o número discado para um número PSTN totalmente qualificado.
- Retirada de telefones IP com SRST: Os telefones IP que usam SIP ou SCCP devem registrar-se em um dispositivo de processamento de chamadas para que os telefones funcionem. Os telefones IP que se registram na WAN IP podem ter um gateway local SRST configurado como backup para um servidor de voz em sua configuração. Quando a conexão com o servidor primário é perdida, eles podem se registrar novamente com o gateway SRST local.
- Desvio de chamadas não registradas (CFUR): Esta é uma configuração de encaminhamento de chamadas de telefones IP que se torna efetiva quando o telefone IP não está registrado. Essa configuração pode ser usada para encaminhar chamadas para o correio de voz ou, talvez, para um local alternativo, se o telefone não estiver registrado na central telefônica.
- Roteamento alternativo automatizado (AAR) e Redirecionamento de chamadas sem largura de banda (CFNB): o AAR permite que as chamadas sejam reencaminhadas pelo PSTN quando as chamadas na WAN IP não são admitidas pelo CAC. Essas duas técnicas podem ser usadas para redirecionar chamadas para destinos alternados no caso de a WAN estar com excesso de assinaturas.
- Soluções de mobilidade: quando usuários ou dispositivos se movimentam entre sites, eles podem perder recursos ou ter uma configuração abaixo do ideal por causa de uma alteração em sua localização física real. A mobilidade de extensão e o recurso de mobilidade de dispositivo podem resolver esses

problemas. Além disso, o recurso de mobilidade colaborativa permite a integração de telefones celulares e telefones de escritório em casa, permitindo a acessibilidade em qualquer dispositivo por meio de um único número (de escritório).

2.2.7.1. Roteamento Otimizado

O uso de uma WAN IP permite economizar no custo de chamadas PSTN interurbanas ou internacionais em um ambiente multisite. Há duas maneiras de economizar custos em chamadas PSTN interurbanas ou internacionais em uma implantação multisite:

- Desvio de chamadas: as chamadas entre sites que usam a WAN IP em vez da PSTN são chamadas de pedágio. O PSTN é usado somente quando as chamadas pela WAN IP não são possíveis (devido a uma falha na WAN ou porque a chamada não é admitida pelo CAC). Um exemplo é a discagem entre telefones IP em dois sites; a chamada atravessa a WAN IP (para os protocolos RTP e de sinalização) em vez de atravessar a PSTN.

- TEHO ou RMC: TEHO amplia o conceito de desvio de pedágio, usando também a WAN IP para chamadas para destinos remotos na PSTN. Com o TEHO, a WAN IP é usada o máximo possível e a quebra da PSTN ocorre no gateway que está mais próximo do destino PSTN discado. A quebra do PSTN local é usada como backup no caso de uma falha na WAN IP ou no CAC. Um exemplo é discando um número de longa distância (DDD ou DDI). Desde que se tenha uma IP WAN conectando ambos os sites, a chamada pode fluir pela WAN IP e sair de um gateway de voz como uma chamada local, economizando assim custos caros da PSTN.

Ao usar a WAN IP para alcançar destinos PSTN remotos ou números internos em um site diferente, é importante considerar os caminhos de backup. Quando a WAN IP está inativa ou quando não há largura de banda suficiente disponível para uma chamada de voz adicional, as chamadas devem ser roteadas pelos gateways PSTN locais como um caminho de backup.

No exemplo mostrado na Figura 19 uma chamada de São Paulo (filial) para Brasília (matriz) seria roteada conforme mostrado nas etapas a seguir:

Etapa 1. Um usuário de São Paulo disca 0 61 xxxx-xxxx, um número de telefone PSTN localizado em Brasília.

Etapa 2. A chamada é sinalizada para o cluster CUCM em Brasília.

Etapa 3. O CUCM em Brasília direciona a chamada para o gateway Brasília, que irrompe na PSTN com uma chamada local para a PSTN de Brasília.

Se a WAN não estivesse disponível por algum motivo antes da chamada, o gateway de São Paulo teria que ser configurado adequadamente para rotear a chamada com a manipulação de dígitos apropriada através da PSTN, a um custo de pedágio potencialmente mais alto para o telefone de Brasília. O contrário também é válido.

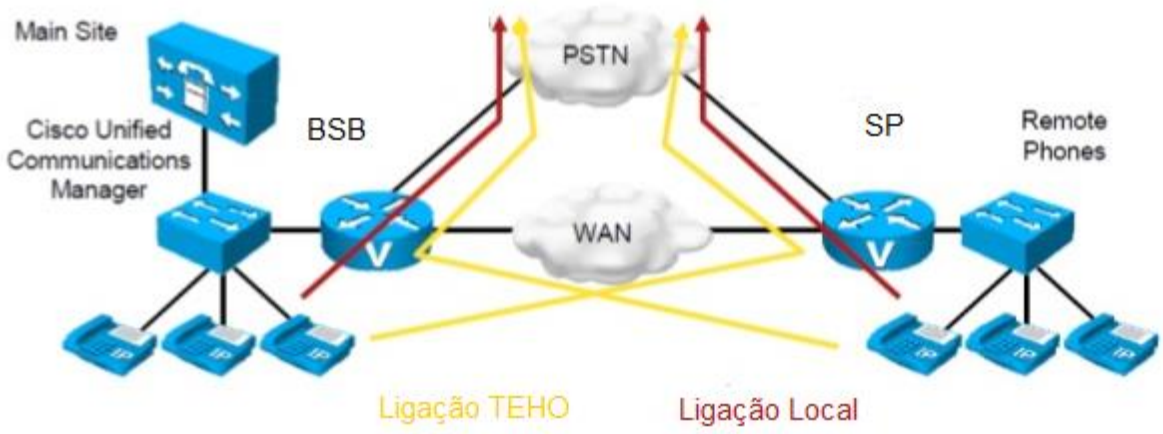


Figura 19 - Exemplo TEHO/RMC

Vale ressaltar que em casos que a WAN interconecta sites em países diferentes, e empresas com grande volume de ligações, esse mecanismo se mostra muito vantajoso e com uma economia expressiva [3].

2.2.7.2. Backup PSTN

As chamadas para um site remoto ou filial no mesmo cluster são configuradas para usar a WAN IP primeiro. No caso de uma falha na WAN ou em tempos de congestionamento da WAN, o AAR pode ser implementado para usar a PSTN como uma opção de backup. O resultado final é um custo operacional reduzido com desvio de pedágio na WAN e entrega bem-sucedida das mesmas chamadas pela PSTN, mas potencialmente a um custo operacional mais alto se a WAN falhar.

É importante lembrar que, quando ocorre *failover* para a PSTN, a expansão do número deve ocorrer, o que significa que a extensão discada de quatro dígitos original precisa ser expandida para um número PSTN totalmente qualificado. Tudo isso é feito sem intervenção do usuário final; o usuário final disca quatro dígitos e não tem conhecimento do reencaminhamento através da PSTN e da expansão do número.

2.2.7.3. Fallback para telefones IP (CISCO)

O fallback para telefones IP é fornecido pelo recurso SRST Cisco IOS e melhora a disponibilidade de telefones IP remotos. Quando os telefones IP da Cisco perdem contato com todos os servidores CUCM, eles se registram no roteador local Cisco Unified SRST para manter o recurso de processamento de chamadas necessário para fazer e receber chamadas. Quando a conexão WAN entre um roteador e o CUCM falha ou quando a conectividade com o CUCM é perdida por algum motivo, os telefones IP Cisco Unified no site remoto ficam inutilizáveis durante a falha. O Cisco Unified SRST supera esse problema e garante que os telefones IP unificados da Cisco ofereçam serviços contínuos (embora mínimos) fornecendo suporte de manipulação de chamadas para telefones IP unificados da Cisco diretamente do roteador Cisco Unified SRST

Um link WAN conecta os telefones IP em um site remoto ao CUCM em um site central por meio do protocolo SIP ou SCCP para sinalização. Se o link WAN falhar, o Cisco Unified SRST ativará o gateway local para fornecer serviços de processamento de chamadas para telefones IP. Os telefones IP registram-se no gateway (que é

listado como um servidor CUCM de backup ou referência SRST na configuração de grupo do servidor dos telefones IP). A ordem de configuração real é uma referência SRST adicionada ao CUCM contendo o endereço IP do roteador da filial. Além disso, um grupo de servidores CUCM é criado contendo até três servidores CUCM para o telefone IP tentar se registrar. Cada um dos componentes anteriores é carregado no pool de dispositivos para uma filial ou site remoto específico, que por sua vez é aplicado ao telefone IP. O telefone IP enviará uma mensagem keepalive para seu servidor CUCM primário e para todos os backups. No caso de todos os servidores CUCM estarem indisponíveis (por exemplo, devido a uma interrupção da WAN), o telefone IP tentará se registrar com seu roteador local usando a referência SRST.

Quando o link WAN ou a conexão com o CUCM primário é restaurado, o tratamento de chamadas é revertido para o CUCM primário.

Um item adicional para é que existem mais de uma formas de SRST:

- SCCP SRST ou SRST tradicional para dispositivos baseados em SCCP.
- CME SRST, que permite a configuração avançada no modo SRST. O CME SRST adiciona recursos, como os pilotos e grupos de captura a um telefone registrado no SRST.

2.2.7.4. Alcance de telefones do site remoto durante a falha WAN (CISCO)

Conforme apresentado, os telefones IP localizados em locais remotos podem usar um gateway SRST como um backup para o CUCM em caso de falha de WAN IP. O gateway pode usar seu plano de discagem local para rotear chamadas destinadas aos telefones IP no site principal por meio do PSTN.

Entretanto, como as chamadas entre sites devem ser roteadas do site principal para o remoto enquanto a WAN IP está inativa? O problema neste caso é que o CUCM não considera nenhuma outra entrada em seu plano de discagem se um número discado combina um diretório configurado mas não registrado. Portanto, se os usuários no site principal discarem extensões internas durante a interrupção de IP WAN, suas chamadas falharão (ou irão para o correio de voz). Para permitir que

telefones IP remotos sejam acessados a partir dos telefones IP no site da sede, configure o CFUR (call forward unregistered) para os telefones do local remoto. O CFUR deve ser configurado com o número PSTN do gateway do site remoto para que as chamadas internas para telefones IP remotos sejam encaminhadas para o número PSTN apropriado.

2.2.8. MOBILIDADE

Ao se adicionar uma ou mais localidades participando do mesmo sistema colaborativo, surge de forma mais expressiva o conceito de mobilidade. Esta seção fornece uma visão geral, apenas apresentando as soluções de mobilidade que resolvem problemas resultantes de usuários de roaming, dispositivos de roaming e usuários com vários telefones (telefone comercial, telefone celular, telefone residencial e assim por diante).

Quando usuários ou dispositivos circulam entre sites, surgem problemas que podem ser resolvidos por soluções de mobilidade:

Mobilidade do dispositivo: resolve problemas causados por dispositivos móveis, incluindo configurações inválidas de dispositivos, como *regions*, referências SRST, grupos AAR, permissões de chamadas e assim por diante. Um exemplo de *roaming* de dispositivo é quando um funcionário de escritório se muda para um prédio ou local diferente. Se esse usuário mover seu telefone físico para um local diferente, o telefone manterá a configuração original do site. O recurso de mobilidade do dispositivo permite que as configurações do dispositivo que dependem da localização física do dispositivo sejam sobrescritas automaticamente se o dispositivo aparecer em um local físico diferente.

Mobilidade de extensão: soluciona problemas resultantes de usuários de roaming que usam telefones IP compartilhados localizados em outros escritórios. Normalmente, é possível ver isso em um *call-center* ou ambiente de "*hot desk*", em que vários usuários entram em um escritório, sentam-se em um cubículo ou espaço

aberto e fazem login em seu PC e telefone. Se eles não puderem efetuar *login* no telefone e o ramal permanente e as configurações forem aplicadas, poderão ocorrer problemas como diretório incorreto, falta de assinaturas do serviço de telefone IP, permissões de ligação e assim por diante. A mobilidade de extensão permite que os usuários façam *login* nos telefones e substituam a configuração do telefone IP pela configuração do telefone IP do usuário conectado.

Mobilidade unificada: resolve os problemas de ter vários telefones e, consequentemente, vários números de telefone, como um telefone do escritório, um telefone celular, um telefone de casa e assim por diante. Essa solução permite que os usuários sejam alcançados por um único número, independentemente do telefone que é realmente usado. Às vezes, isso é chamado de alcance de número único (SNR). Com a mobilidade unificada, quando um usuário recebe uma chamada em seu ramal de discagem direta interna (DID), várias chamadas simultâneas de saída podem ser feitas para vários dispositivos em uma tentativa de "localizar" o usuário. Quando vários dispositivos tocam simultaneamente, um normalmente tem uma taxa de sucesso maior ao conectar a chamada.

3. SIMULAÇÃO

Nessa porção do trabalho, foi montado um cenário para que fossem analisadas as diferenças de desempenho quando acatadas as orientações indicadas para validar as principais problemáticas abordadas neste projeto.

Inicialmente, essa sessão apresenta o ambiente configurado para a simulação em um nível mais abrangente apenas para entender a disposição e o papel de cada componente.

Em seguida, foi exposto mais a fundo como cada um dos componentes se interligam, a motivação e justificativa para a escolha dos mecanismos ou técnicas para os testes.

Foi apresentada a proposta de cada teste, juntamente com o que se esperava validar em cada um. Logo após, já realizada a ambientação e entendendo a proposição, foram discutidas as especificidades, granularidades de cada proposta e os resultados obtidos para validar os procedimentos tratados.

3.1. AMBIENTE

O contexto proposto foi simplificado para que a análise fosse clara e permitisse destacar as situações relevantes para o trabalho.

A simulação consiste em representar uma empresa que possui uma matriz em uma localidade (Brasília) e uma filial estabelecida em outra região (São Paulo) conectadas via um link WAN para comunicação entre os dois sites. Cada um desses sites terá sua própria conexão à rede pública de telefonia (PSTN) via protocolo SIP conforme a Figura 20 apresenta.

Os testes irão abordar mecanismos de garantir o bom funcionamento das comunicações que fluem através deste link WAN em diversas situações propostas.

Serão conectados dois telefones analógicos ao roteador da PSTN, simulando que cada um pertence a um estado brasileiro (Telefone 1 simula um número proveniente de Brasília e o Telefone 2 simula um número proveniente de São Paulo).

Estão dispostos na matriz, o servidor de voz Call Manager da Cisco, o qual é responsável por registrar e enviar as configurações dos aparelhos telefônicos (telefones IP e *softphones*) e roteamento inicial das chamadas. Neste site, o computador utilizado é um Windows 7, o qual foi empregado para o monitoramento das chamadas e do link WAN por meio das ferramentas adequadas apresentadas na próxima sessão. Além disso, foi utilizado um programa para simular um servidor e estabelecer comunicação com o computador utilizado na filial.

A filial contém um roteador com capacidades de rotear chamadas e estabelecer troncos de comunicação, um telefone IP da Cisco e dois computadores conectados (Windows 10 e Ubuntu). O Windows 10 tem como contribuição simular um cliente e estabelecer comunicação com o Windows 7 localizado na matriz para gerar um fluxo de pacotes que atravessam o link WAN entre os sites a uma taxa definida. O Ubuntu será usado para captura de tráfego para análise dos dados.

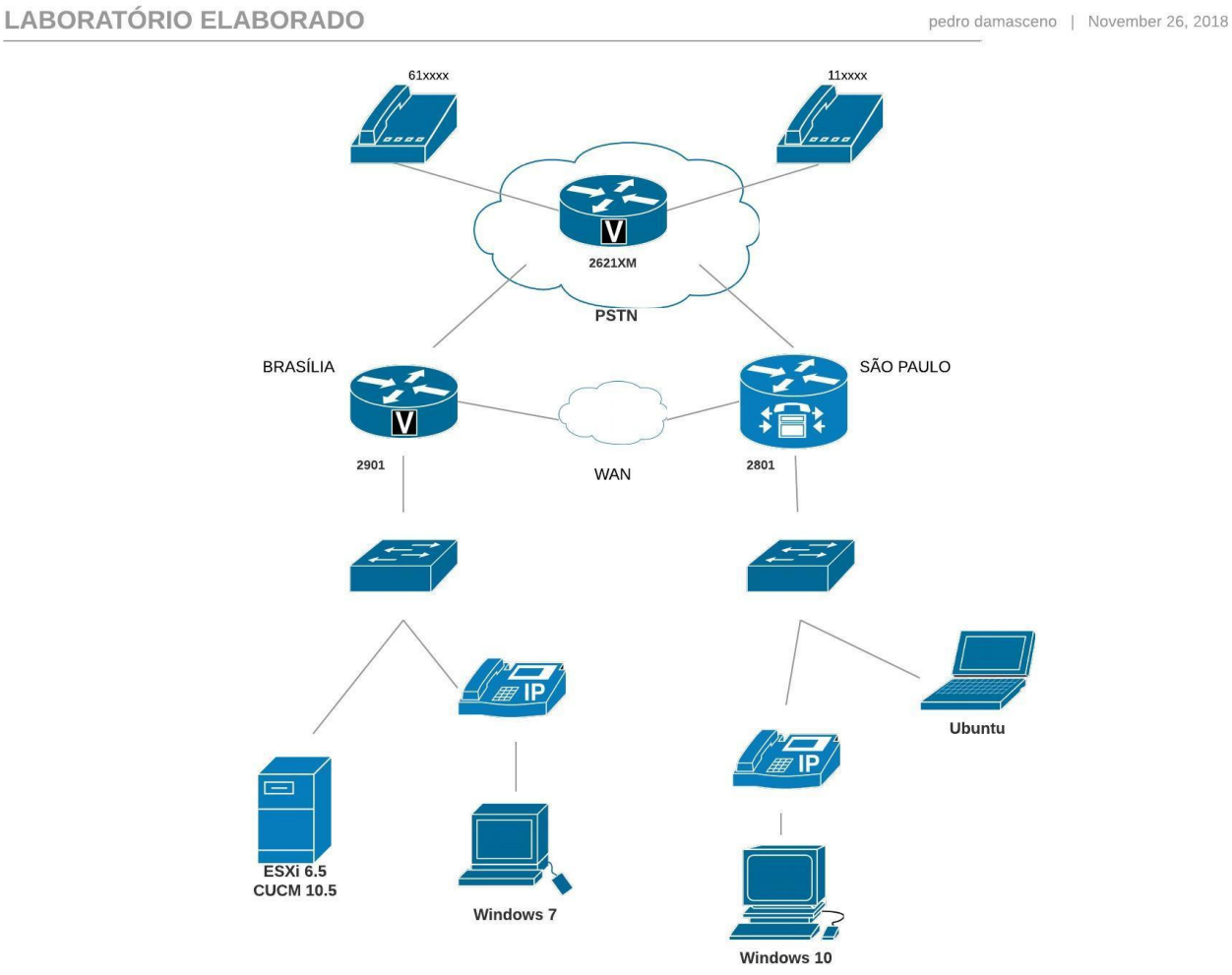


Figura 20 - Topologia Proposta

3.2. AMBIENTE TÉCNICO

As especificações de infraestrutura dos equipamentos utilizados foram as seguintes:

- Roteador 2901, 2801 e 2621XM da Cisco;
- Switches 2960 da Cisco;
- Servidor de rede com hypervisor ESXi 6.5 da VMWare
- CUCM versão 10.5;
- Telefones IP 7911 da Cisco;
- Computadores (Windows 10, Windows 7 e Ubuntu);
- Telefone analógico;
- Servidor PRTG para captura dos dados SNMP e NetFlow dos ativos de rede;
- *Softphones* (3CX, Eikiga, CIPC, X-Lite, MicroSIP);

Os roteadores 2901 e 2801 foram conectados por um link serial simulando uma conexão WAN utilizando o PPP como protocolo de camada 2. O ideal seria utilizar links WAN MPLS, uma vez que este é o método que vem ganhando espaço e sendo mais utilizado, entretanto, devido à falta de equipamentos, a utilização do PPP foi suficiente para a realização dos testes e não afetou o desempenho nem a validade de cada um.

As conexões à PSTN foram simuladas conectando os roteadores de cada site ao 2621XM por meio de troncos SIP (outra opção seria conectar utilizando troncos analógicos ou digitais R2/ISDN, entretanto estes vêm sendo gradativamente substituídos por troncos SIP). O roteador 2621XM foi conectado a dois telefones analógicos que simularam números públicos provenientes de cada localidade. Para isso, foram configuradas as portas de cada um para que estes possuam o número identificador (ANI) adequado e as *Dial-Peers* de forma que ao reconhecer o DDD de cada localidade, a chamada fosse encaminhada para o telefone encarregado.

Foi utilizado o hypervisor esxi6.5 da VMWare, que é disponibilizado por 60 dias como versão de teste, encarregado de proporcionar o ambiente para instalação do servidor de voz CUCM. Foi utilizada o fabricante VMWare pois esta possui parceria

com a Cisco e provê o ambiente já com os requisitos necessários para a instalação da central de voz.

O CallManager foi utilizado como servidor TFTP para configuração base dos telefones IP (este modelo é utilizado em redes de menor porte. Dependendo do número de dispositivos que forem atendidos, é necessária a instalação de um servidor TFTP separado).

Foram configuradas as r*egions* de BSB e SP para o tratamento dos codecs entre cada site durante a fase de destes. Os números dos telefones utilizaram os ramais 4XXX e não foram tratados mecanismos de sobreposição e escalabilidade dos ramais, uma vez que não é escopo dos testes, entretanto, é válido atentar para esta característica.

Foram configurados troncos SIP entre o CUCM e o roteador 2901 e o 2801 para roteamento das chamadas dependendo do número discado. A partir do encaminhamento dos dados das chamadas para os roteadores, estes devem conter suas próprias configurações e especificações de perfis de tradução para adição e remoção das máscaras e prefixos dos ramais.

O software de monitoramento de rede utilizado foi o PRTG, já que possui versão gratuita de até 100 sensores e é capaz de monitorar e classificar condições do sistema, como uso de largura de banda ou tempo de atividade, coletar estatísticas de *hosts* diversos como switches, roteadores, servidores e outros dispositivos e aplicativos e é capaz de interpretar dados fornecidos pelo Netflow, uma ferramenta de monitoramento do fluxo de tráfego. O NetFlow realiza a coleta de estatísticas sobre o tráfego que atravessa um determinado dispositivo nas interfaces habilitadas com esse recurso (vide figura). Essa ferramenta é totalmente transparente para os dispositivos e aplicações da rede, não requerendo nenhuma intervenção de configuração entre os vários elementos monitorados - a não ser habilitá-lo na interface monitorada.

Para simular a sobrecarga do link WAN acima da capacidade projetada, foi efetuado o estresse do link WAN por meio de um gerador de tráfego (packETH1.6). A banda máxima permitida nesse link foi limitada utilizando QoS nos roteadores (*policing*). Para cada teste o limite foi ajustado de forma que facilite a observação dos resultados.

O trafego de pacotes nas interfaces e links foi identificado por meio de configuração do Netflow e SNMP nos roteadores e sensores no PRTG.

A captura de tráfego de rede foi feita por meio de configuração de sessões de monitoramento no switch e utilizando o Wireshark no Ubuntu.

3.3. PROPOSTAS DOS TESTES

Esta sessão de testes apresenta quais testes foram realizados e qual o objetivo que se desejou alcançar com cada um.

Foram realizadas 5 simulações, que trataram de forma geral os seguintes fatores:

Validação do consumo de banda de cada dos codecs e forma de diminuí-lo. Demonstração de *transcoding* para casos em que os equipamentos não suportam um codec específico.

Validação do atraso de serialização e forma de contornar esse empecilho.

Validação da necessidade do QoS aplicado ao ambiente de colaboração.

Validação de mecanismos de alta disponibilidade e backup caso ocorra falha no link WAN e este se torne momentaneamente indisponível.

TESTE 1: Foi feita a verificação da ocupação de banda no link para ligações utilizando diferentes tipos de codecs atravessando a WAN. Inicialmente foi utilizado o G711, seguido do G729 e G729 com cRTP.

Dessa forma demonstrou-se empiricamente que a qualidade das ligações é satisfatoriamente conservada enquanto o consumo de banda para cada ligação é diminuído. Essa técnica se mostrou muito importante em conexões que possuem capacidade de banda limitada. Além disso, foi demonstrado como realizar o *transcoding* para garantir que serviços que só aceitam o codec G.711 consigam atravessar a WAN utilizando o G.729 e recodifica-los para o G.711 após a travessia do link.

TESTE 2: Esse teste teve como objetivo validar a importância do LFI em links que possuem taxa de transmissão lenta e verificar que, mesmo sem a sobrecarga desse link, a qualidade da ligação é afetada ao iniciar a transmissão de dados devido ao atraso de serialização.

Foi realizada a limitação do link para um *clock* baixo:128kbps e feita limitação do link para essa mesma taxa de transmissão. Para ser testado o impacto da serialização foi realizada uma ligação que atravessa o link WAN e injetado um tom de voz de 1000 Hz para ser observado como esse tom começa a ser deteriorado quando pacotes começam a trafegar pelo canal, mesmo que a taxa de transmissão dos pacotes não estivesse sobrecarregando o *link*. Logo após foi feita a configuração do LFI e repetido o teste para ratificar a eficácia dessa técnica.

Essa técnica demonstrou que *links* que possuem o *clock* lento, sofrem uma perda de qualidade significativa nas ligações quando outros tipos de tráfego começam a fluir por eles, mesmo que este não esteja sobrecarregado.

TESTE 3: Esse teste se propôs demonstrar a degradação da qualidade das chamadas de voz de um site para o outro caso não houvesse nenhum mecanismo que garantisse prioridade do trafego de voz sobre outros.

Nesse cenário, o link WAN foi limitado a 1Mbps e em seguida foram feitas ligações de um site para o outro pelo link WAN utilizando o codec G711. Foi validada a qualidade das ligações.

Em seguida, foi feita a sobrecarga no link WAN utilizando o Iperf e logo após, realizou-se uma ligação de um site para o outro. Foi avaliada a qualidade da ligação que atravessa um link sem nenhum tipo de prioridade sobre outros tipos de dados, ou seja, um modelo de envio de dados de melhor esforço.

TESTE 4: Ao fim da realização deste teste, validou-se a eficiência, importância e apresentou-se a forma de se configurar e mapear os dados para se aplicar corretamente as técnicas de QoS.

Esse teste repetiu as configurações utilizadas no TESTE3 adicionando as técnicas de CBWFQ e LLQ para o tráfego RTP nos roteadores e switches. Foi feita a mesma sobrecarga do link WAN (com a mesma limitação de banda), em seguida foram realizadas as ligações para avaliar a qualidade das mesmas. Isso confirmou a importância e efetividade do design e mecanismos QoS.

TESTE 5: Esta sessão visou abordar e validar mecanismos para manter a alta disponibilidade dos serviços no caso de quedas do link WAN, como o mecanismo de encaminhamento de ligações para dispositivos que perderam conexão e, por isso, foram cancelados os registros com a central telefônica. Foi validado que os telefones no site da filial consiguissem realizar chamadas mesmo não sendo capaz de se comunicar com o servidor de voz.

Caso nenhuma configuração adicional fosse realizada e uma falha momentânea no link WAN da empresa ocorresse, os serviços de voz das filiais seriam paralisados. Para realizar esse teste foi feita a desconexão do link entre os roteadores 2901 e 2801, para simular a queda do link, após a implantação do SRST e configuração do CUCM para criação de rotas alternativas das ligações (juntamente com o CFUR), garantindo relutância à falha em casos de perda de conexão entre os sites.

Esses cinco testes finalizam o laboratório do projeto e ratificam os artifícios adotados para garantir alta disponibilidade e garantia de qualidade dos serviços.

3.4. RESULTADOS DOS TESTES

TESTE 1:

Foram configuradas *regions* para cada localidade e criadas as relações de codecs para cada uma. Essas regiões são aplicadas às *device pools* que estão presentes na configuração de cada telefone, para que estes possuam as configurações e relações dos codecs referentes à sua respectiva localidade.

Na primeira e segunda sessão deste teste foi adotado o G729 entre as *regions*.

Na terceira parte do teste a relação da região BSB - SP foi adotado o G711.

As *regions* estabelecem a relação dos codecs que são utilizados entre elas. Cada dispositivo é vinculado a uma *region*. Quando a chamada é realizada, o número discado é ligado a um padrão de roteamento de chamada ou a outro dispositivo que, por sua vez, já é relacionado a outra *region*. O codec utilizado na ligação é determinado dessa forma.

Verificou-se durante os testes que os *softphones* utilizados que não são da Cisco não são capazes de realizar chamadas com o codec G.729. Dessa forma todas as chamadas eram anuladas. Para ser dada a continuidade desse teste foi necessária a configuração de *transcoding* entre os escritórios. A transcodificação é obtida usando módulos de dados de voz e pacote (PVDM) e processadores de sinal digital (DSPs), já que o CUCM não é capaz de transcodificar ligações.

Essa técnica se faz necessária em casos que os dispositivos exigem que um codec específico seja utilizado, como o Unity (correio de voz) e o Contact Center Express (UCCX), os quais só aceitam fluxos do G.711, dependendo de como estão instalados. Em uma arquitetura centralizada de processamento de chamadas, esses aplicativos geralmente estão localizados em uma sede ou data center.

Foram realizadas 4 chamadas entre os sites utilizando cada um dos codecs. Para validação do resultado teórico, foram capturados os dados de entrada e saída das interfaces dos roteadores via NetFlow.

G729:

Banda teórica consumida: 4*26.4 = 105.6kbps;

G729 com cRTP:

Banda teórica consumida: 4*12 = 48kbps;

G711:

Banda teórica consumida: 4* 82.4 = 329.6kbps:

Conforme esperado, os resultados apresentados na Figura 21 foram válidos com os teóricos. As primeiras ligações consumiram em torno de 100kbps, segundo a imagem. Em seguida, utilizando a compressão, foi verificado o consumo ligeiramente abaixo de 50kbps. Ao se utilizar o codec G.711, o consumo do link, para as mesmas 4 ligações, subiu para cerca de 330kbps.

Pode-se perceber que, para uma mesma quantidade de ligações, o consumo da capacidade do link é evidentemente economizado ao se utilizar codecs diferentes. A diferença de consumo se provou notória conforme o codec aplicado.

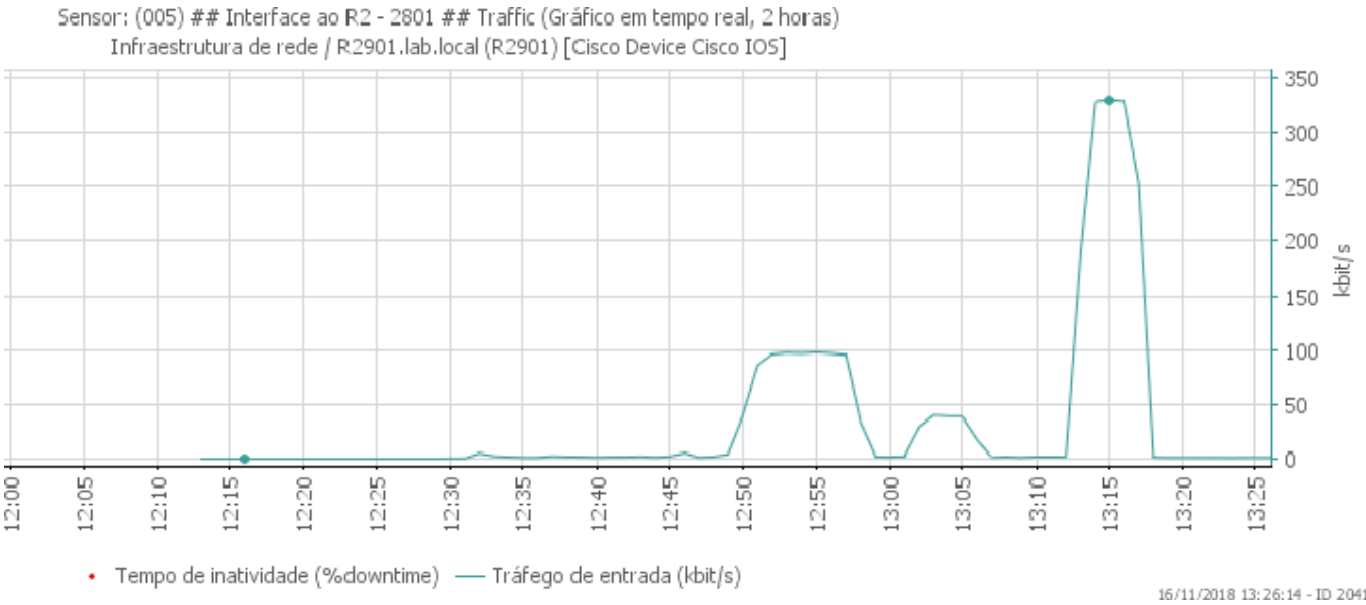


Figura 21- Consumo de 4 ligações G729, cG729 e G711

Em instituições que necessitam otimizar a demanda de banda essa é uma solução válida e fundamental. Se mostra extremamente oportuna essa técnica

especialmente em empresas com um volume muito alto de ligações ou que o link WAN possui baixa capacidade.

É necessário notar que é válido um estudo de viabilidade financeira nesses casos: Para realizar a transcodificação é preciso adquirir PVDMs e DSPs, que são materiais de custo elevado. É valida a análise de custo entre a aquisição dos recursos necessários para utilização de transcodificação ou aumento da banda dos links para se adequar a necessidade do negócio.

A configuração do cRTP e do *transcoding* entre os links está disposta no Apêndice Sessão 0 - cRTP

TESTE 2:

Nesse teste, desmonstrou-se a sensibilidade dos pacotes de voz a sobrecargas de dados caso não seja tomada nenhuma providência em relação a prioridade do tráfego.

Para isso, o link WAN foi limitado à uma taxa de transmissão de 1Mbps, validado pela Figura 22, por meio da técnica de *rate-limit/policing* (para simular como o tráfego excedido seria tratado pela operadora WAN) configurado para transmitir pacotes enquanto essa taxa não fosse alcançada e descartando os pacotes que a excedam.

Utilizou-se o PRTG para análise e confirmação da banda ocupada no canal WAN por meio do protocolo SNMP para gerenciamento do roteador 2901. Essa ferramenta não só monitora o tráfego de pacotes nas interfaces dos dispositivos da infraestrutura de redes, como também faz o agrupamento deles em fluxo (por conexão), o que permitiu caracterizar o perfil de operação da rede.

Os dados do tráfego total de saída do roteador foram configurados para serem enviados ao IP e porta adequados do PRTG, o qual estava com sensores já configurados para receber as informações.

Empregou-se o Iperf para geração de tráfego. Essa ferramenta funciona com um modelo cliente-servidor para envio e teste de recepção dos dados. O Windows 10 funcionou como servidor, recebendo o fluxo de dados gerado pelo cliente, o Windows 7. Foram transmitidos pacotes UDP na porta 5201 a taxas de 1 a 3Mbps, como mostrado na Figura 23 no campo *Bandwidth*, para gerar uma fila razoável e perceber a degradação ocasionada.

Utilizou-se o Wireshark para a captura de pacotes dos dados de voz da ligação para analisá-los e observar o comportamento na situação gerada.

Os comandos utilizados para geração do tráfego foram:

Windows 10: iperf3.exe -s
Windows 7: iperf3.exe -c <ip_destino> -p 5201 -t 240 -b [1000000 – 3000000] -u
Switch 2 (2960):

monitor session 1 source interface Fa0/4

monitor session 1 destination interface Fa0/20

Roteador 2901:

int serial 0/0/0

rate-limit output 1000000 240000 240000 conform-action transmit exceed-action drop

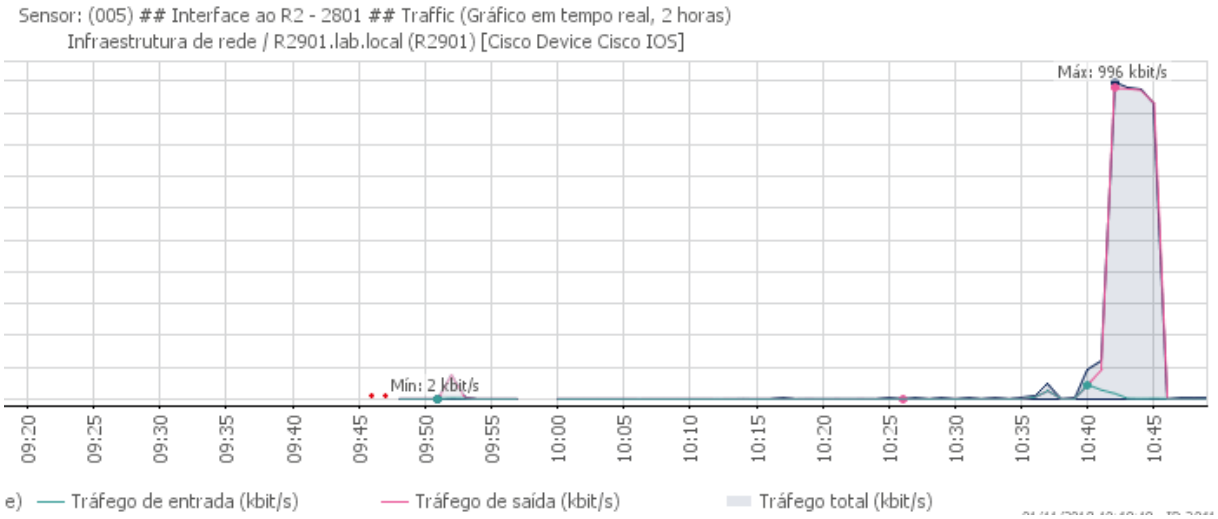


Figura 22 - R2901 Taxa de transmissão interface serial0/0/0 limitada a 1Mbps

```
C:\Users\Cliente\Desktop\iperf-3.1.3-win64>iperf3.exe -c 10.2.40.11 -p 5201 -t 2
40 -b 3000000 -u
Connecting to host 10.2.40.11, port 5201
[  4] local 10.1.40.13 port 51552 connected to 10.2.40.11 port 5201
[ ID] Interval           Transfer     Bandwidth       Total Datagrams
[  4]   0.00-1.00   sec   480 KBytes  3.91 Mbits/sec  60
[  4]   1.00-2.01   sec   232 KBytes  1.89 Mbits/sec  29
[  4]   2.01-3.01   sec   360 KBytes  2.95 Mbits/sec  45
[  4]   3.01-4.01   sec   368 KBytes  3.01 Mbits/sec  46
[  4]   4.01-5.01   sec   368 KBytes  3.01 Mbits/sec  46
[  4]   5.01-6.01   sec   368 KBytes  3.01 Mbits/sec  46
[  4]   6.01-7.01   sec   360 KBytes  2.95 Mbits/sec  45
[  4]   7.01-8.01   sec   368 KBytes  3.01 Mbits/sec  46
[  4]   8.01-9.01   sec   368 KBytes  3.01 Mbits/sec  46
[  4]   9.01-10.01  sec   368 KBytes  3.01 Mbits/sec  46
[  4]  10.01-11.01  sec   368 KBytes  3.01 Mbits/sec  46
```

Figura 23 - Iperf Transmissão 3Mbps (10primeiros segundos)

Foram feitas ligações que utilizassem o link WAN para observar o comportamento destas antes da sobrecarga apenas para validação das configurações básicas do ambiente, uma vez que não havia tráfego pelo link. Conforme observado na Figura 24, confirmou-se que a ligação teve um bom desempenho devido ao atraso (*delta*) e *jitter*, os quais estão dentro dos padrões estabelecidos pela ITU. Por meio dos pacotes capturados pelo Wireshark é possível reproduzir o áudio capturado.

Packet	Sequence	Delta (ms)	Jitter (ms)	Skew	Bandwidth	Marker	Status
336	334	20.01	0.01	0.08	80.00		✓
338	335	20.02	0.01	0.06	80.00		✓
340	336	19.98	0.01	0.08	80.00		✓
342	337	19.98	0.01	0.10	81.60		✓
344	338	20.02	0.02	0.08	81.60		✓
346	339	19.98	0.02	0.10	81.60		✓
348	340	20.02	0.02	0.07	81.60		✓
350	341	19.98	0.02	0.09	81.60		✓

Figura 24 - Dados Inicio Ligação (antes da sobrecarga)

Durante a ligação, iniciou-se o sobre fluxo de dados no link. Imediatamente percebeu-se a queda na qualidade da chamada, independentemente da sobrecarga aplicada. Intuitivamente é possível deduzir que durante a sobrecarga de 3Mbps a degradação foi mais considerável, porém, para manter a consistência com o teste de QoS a seguir, será apresentado sempre o pior caso.

Como pode ser observado na Figura 23, foi gerado um trafego com taxa de transmissão de 3Mbps, porém só foi permitida a transmissão de 1Mbps pela interface serial do roteador, causando enfileiramento, atraso e consequente perda de diversos pacotes (tanto de voz quanto de dados, uma vez que eles possuem a mesma prioridade).

A análise feita pelo Wireshark nos mostra com melhor granularidade os dados capturados da chamada e as condições de qualidade:

Pelas Figura 25 e Figura 26, foi possível notar a grande degradação decorrente da sobrecarga do link. Foi observada uma perda de 25,8% dos pacotes (um valor notoriamente mais elevado do que a norma exige), valores de atraso muito mais altos que os obtidos na primeira fase desse teste (Max Delta 743,063 ms) e valores de jitter oscilando no limite estipulado de 30ms, evidenciado pela Figura 27.

Wireshark · RTP Streams · teste1_sobrecarga3Mbps

Source Address	Source Port	Destination Address	Destination Port	SSRC	Payload	Packets	Lost	Max Delta (ms)
1.1.1.1	16398	10.2.30.12	16754	0x1dcd0303	g711A	870	303 (25.8%)	743.063

Figura 25 - Wireshark Stream Sobrecarga 3Mbps

Packet	Sequence	Delta (ms)	Jitter (ms)	Skew	Bandwidth	Marker	Status
1139	747	96.08	29.10	-82.77	92.80		Wrong
1140	749	0.83	29.73	-43.59	94.40		Wrong
1141	751	0.82	30.32	-4.41	96.00		Wrong
1148	753	107.86	32.67	-72.27	91.20		Wrong
1149	755	0.78	33.08	-33.05	92.80		Wrong
1150	756	0.82	32.21	-13.88	94.40		✓
1154	758	63.07	31.64	-36.95	83.20		Wrong

Figura 26 - Dados durante sobrecarga 3Mbps

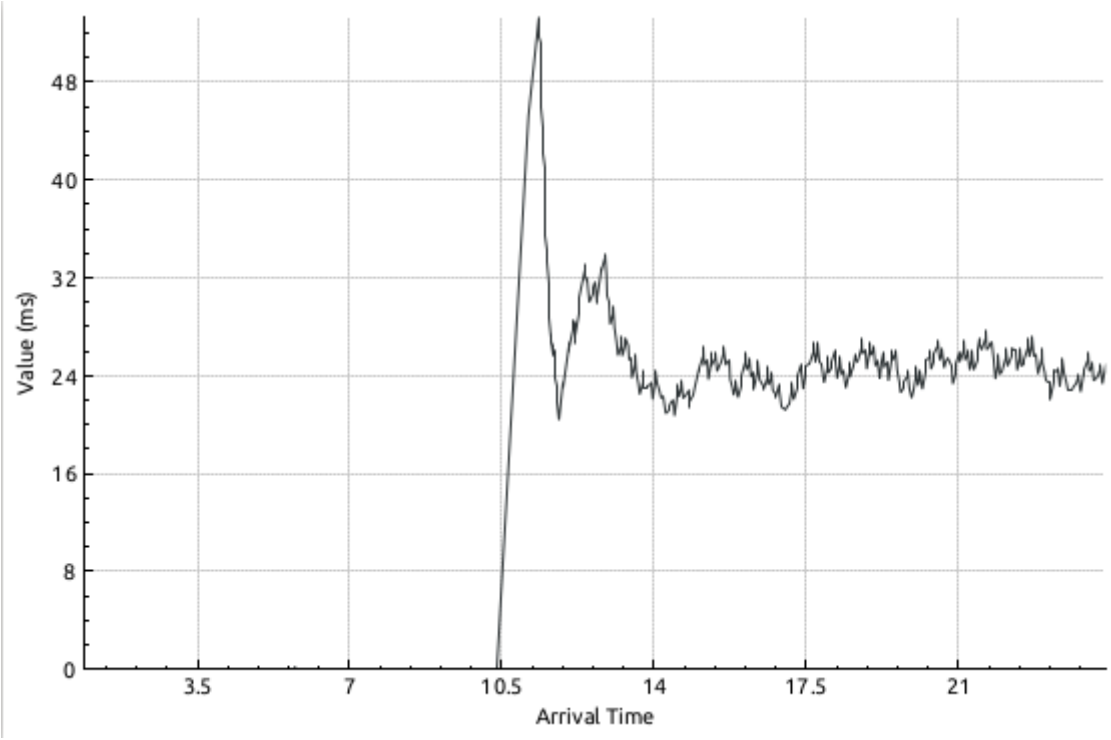


Figura 27 - Jitter Sem QoS

Percebeu-se que a qualidade da ligação caiu drasticamente e foram captados pacotes com latência de até 750ms e jitter até 53.35ms, valores muito acima do recomendado.

TESTE 3:

Nesse teste, desmonstrou-se a validação da importância e indispensabilidade do emprego das técnicas de QoS ao se implementar uma solução de VoIP para garantir resiliência a sobrecargas de transmissão de dados.

Foram utilizadas as configurações já provenientes do TESTE 3 da limitação da WAN a 1Mbps, comunicação com exportação de dados NetFlow ao servidor PRTG e utilização do Iperf para sobrecarga do link. A sobrecarga realizada foi apenas a de 3Mbps, uma vez que se for provada a eficácia desse mecanismo para essa taxa de transmissão, está garantido que funcionará para qualquer taxa abaixo.

O primeiro passo dado foi o de classificar os dados RTP de voz que fluíam na rede. Para isso foi necessário adicionar marcações de camada 2 (COS) nos pacotes provenientes dos *softphones* que não eram Cisco por meio de *access-lists*, uma vez que estes não enviavam as marcações DSCP nativamente e por isso, não estavam recebendo prioridade no tráfego.

Logo após foi realizada a configuração no roteador para a classificação do tráfego com campo QoS de camada 2 marcado, provenientes dos *softphones* de terceiros e com campo DSCP proveniente dos telefones Cisco.

Após classificação em um mesmo grupo (VoIP), foi criada a política de prioridade e configurado para que todos os pacotes tivessem a marcação DSCP no campo adequado. Essa política foi então aplicada à interface serial adequada.

O QoS implementado forneceu formatos para o comportamento dos dados RTP. Os pacotes de voz foram limitados à uma banda de 200Kbps (2 ligações utilizando o codec G711).

Inicialmente foi verificada uma ligação antes de ser feita a sobrecarga para garantir o funcionamento padrão da rede (Figura 28).

Novamente foi feito o congestionamento do link com uma taxa de 3Mbps e analisados os dados da ligação.

Os resultados obtidos:

Packet	Sequence	Delta (ms)	Jitter (ms)	Skew	Bandwidth	Marker	Status
149	2156	19.57	0.10	0.15	81.60		✓
151	2157	20.01	0.09	0.14	81.60		✓
153	2158	20.03	0.09	0.11	81.60		✓
155	2159	19.98	0.08	0.13	81.60		✓
157	2160	20.05	0.08	0.08	81.60		✓
159	2161	20.20	0.09	-0.11	81.60		✓

Figura 28 - Dados Inicio Ligação (antes da sobrecarga)

Conforme, a Figura 29, percebeu-se que mesmo com a taxa de transmissão muito acima do permitido pela configuração do *policing* todo tráfego de voz recebeu a prioridade necessária e não sofreu perda alguma (de 25,8% no teste 2 para 0%), um resultado muito significativo e relevante.

Com a Figura 30 e Figura 31, foi possível verificar que os níveis de atraso e *jitter* foram mantidos dentro dos termos aceitáveis (atraso máximo de 32,18ms e *jitter* máximo em torno de 4,2ms).

Wireshark · RTP Streams · teste2_sobrecarga3Mbps

Source Address	Source Port	Destination Address	Destination Port	SSRC	Payload	Packets	Lost	Max Delta (ms)
10.1.30.12	18820	10.2.40.11	40048	0xc7171740	g711U	1281	0 (0.0%)	32.184

Figura 29 - Wireshark Stream com QoS

Packet	Sequence	Delta (ms)	Jitter (ms)	Skew	Bandwidth	Marker	Status
2640	2953	25.44	3.42	-5.15	80.00		✓
2646	2954	25.31	3.54	-10.46	80.00		✓
2651	2955	19.20	3.37	-9.66	80.00		✓
2652	2956	9.94	3.79	0.39	81.60		✓
2654	2957	20.05	3.55	0.34	81.60		✓
2661	2958	31.39	4.04	-11.04	80.00		✓

Figura 30 - Dados durante sobrecarga 3Mbps

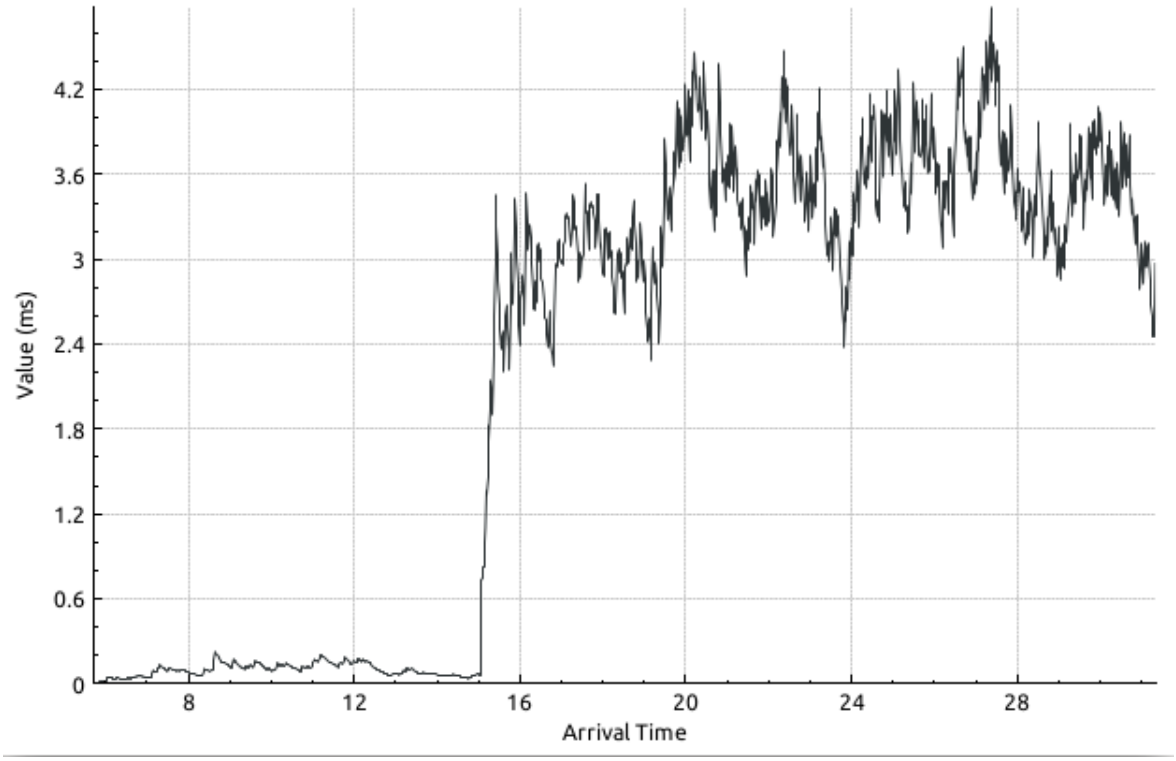


Figura 31 - Jitter QoS

Observou-se que quando foi utilizado os mecanismos de QoS, a ligação ocupou a primeira posição na fila e são sofreu atrasos significativos, nem *jitter* expressivo para prejudicar a comunicação. Além disso, um fator expressivo foi que não houve perda de pacotes durante a comunicação, o que a tornou muito mais satisfatória. Foi comprovada a extrema importância das práticas de QoS por toda a rede para identificar os tráfegos que devem possuir prioridade de transmissão. Esses fatores se mostram muito mais desafiadores e significativos quando adicionamos esses dispositivos em uma rede *Wireless*.

TESTE 4:

Existem áreas de difícil acesso e que a possibilidade de comunicação se torna complicada. Verifica-se constantemente que nesses casos, o emprego de links de alta capacidade é algo custoso. São exemplos de algumas mineradoras, fazendas, áreas remotas, locais com ADSL distantes da central, clientes com contratos antigos de baixa velocidade...

Nestes casos, ao se aproveitar do sistema de comunicação centralizado da matriz são necessários alguns ajustes para que o funcionamento se torne adequado e aceitável.

Neste teste, foi demonstrada a necessidade da fragmentação dos pacotes nos links de baixa velocidade. Foi realizada a limitação do link WAN para um *clock* baixo de 128kbps, para simular um dos casos exemplificados, e testar o impacto da serialização.

Foi feita uma ligação da porta analógica da PSTN para um telefone IP localizado na filial. Foram realizados ajustes no roteamento das chamadas pelos roteadores (via *dial-peers*) redirecionando a ligação para que trafegue pelo link WAN. A escolha dessa metodologia é devido à uma função que só é possível a partir de uma ligação feita por uma porta analógica, que é a injeção de um tom sonoro na ligação. Por isso a ligação não foi iniciada de um telefone IP já configurado na matriz.

Por meio desse tom, se torna mais evidente a presença do atraso de serialização.

Foi injetado um tom de voz pela porta analógica de 1000 Hz (*voice-port x inject-tone network 1000*) e observado como esse tom começa a ser deteriorado quando pacotes de 1500B começam a trafegar pelo canal, sem sobrecarga no link. Logo após foi feita a configuração do LFI para ratificarmos a eficácia dessa técnica. O codec utilizado nessa ligação é o G729.

Inicialmente foi realizada a ligação para verificar o comportamento padrão desta quando não há nenhum tipo de empecilho para dificultar o fluxo. Foi executada a injeção dos tons de voz e foi observado que nenhum pacote foi perdido e não houve

delay entre eles. Conforme a Figura 32 apresenta, não houve perda de pacotes (*loss* 0.0%) e o tom ouvido foi contínuo e suave.

Source Address	Source Port	Destination Address	Destination Port	SSRC	Payload	Packets	Lost	Max Delta (ms)	Max Jitter
3.3.3.3	17840	172.16.23.2	17728	0xd2f0303	g711A	733	0 (0.0%)	20.353	0.103

Figura 32 – Pacotes de voz em link de 128kbps

Logo após foi feita uma injeção de pacotes a uma taxa de 64kbps para iniciar um fluxo de dados que não consumissem toda a banda disponível no link e realizada a ligação.

Foi notado que o tom injetado, mesmo com QoS habilitado e recebendo precedência sobre qualquer tráfego, ficou ligeiramente picotado e que ao tentar realizar uma conversa, esta estava seriamente degradado. Percebeu-se, segundo a Figura 33, que ocorreu uma perda pequena de pacotes (3,6%, valor acima do estipulado pela ITU) e apesar de o atraso e o *jitter* estarem dentro dos limites aceitos pelas recomendações.

Esse fenômeno ocorre porque os pacotes de dados, mesmo tendo prioridade sob qualquer tráfego, sofrem atraso quando pacotes grandes atravessam o link, devido ao tempo necessário para o envio destes

Source Address	Source Port	Destination Address	Destination Port	SSRC	Payload	Packets	Lost	Max Delta (ms)	Max Jitter
1.1.1.1	16422	10.2.40.12	16420	0x2c70303	g711A	803	30 (3.6%)	40.179	0.489

Figura 33 - Fluxo de 64kbps junto com chamada sem LFI

A seguir, foi feita a configuração nas interfaces para que fosse adotada a fragmentação dos pacotes antes destes serem enviados e realizada sobrecarga de 256kbps para demonstrar a ação do LLQ QoS e do LFI juntos.

Observou-se por meio da Figura 34 que o áudio, mesmo com a sobrecarga do link, não apresentou perdas e não foi verificada nenhuma degradação durante a conversa:

Source Address	Source Port	Destination Address	Destination Port	SSRC	Payload	Packets	Lost	Max Delta (ms)	Max Jitter	Mean Jitter
1.1.1.1	16458	10.2.40.12	16408	0x32f0303	g711A	1858	0 (0.0%)	30.278	4.785	1.392

Figura 34 - Aplicado LFI e LLQ durante a sobrecarga de 256kbps

Percebeu-se que a ação do LFI em links com baixa velocidade é claramente impactante para aperfeiçoamento do desempenho das chamadas.

A configuração está disposta no Apêndice Sessão - LFI

TESTE 5:

Nesse teste demonstro-se a necessidade da configuração de um método de se manter a operação do sistema caso ocorra uma falha no link WAN.

Para isso, foi feita a configuração do SRST no roteador 2801 da filial para fornecer informações suficientes para que os telefones remotos não fiquem indisponíveis. Outra opção seria a configuração do SRST CME, a qual provê mais funções aos telefones, como funções de *parking*, *hunt groups*, grupos de captura, conferência, entretanto, devido a esse acréscimo de funcionalidades, suporta um número reduzido de telefones.

Foi realizada a configuração no CUCM para que fosse possível redirecionar as chamadas pela PSTN (com expansão da ANI) caso os telefones não apareçam como registrados no sistema.

Logo após foi realizada a desconexão do link WAN, como pode ser verificado nas Figura 35 e 36, onde se percebe que a taxa de transferência de dados se reduziu a 0 kbps. Testou-se a disponibilidade do sistema.

Como se percebe na Figura 37, o telefone em modo SRST está registrado, porém perde algumas funções como a de *hold* e apresenta uma mensagem ao usuário informando que o telefone não está completamente funcional (SRST MODE na figura).

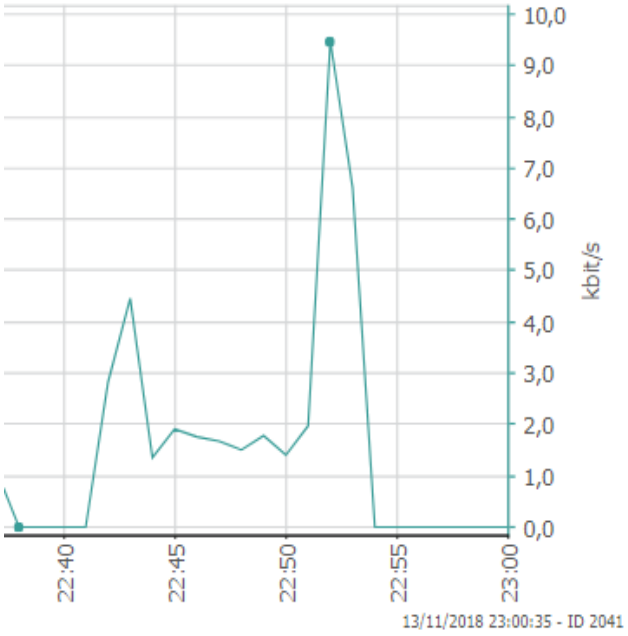


Figura 35- R2901 Interface Serial 0/0/0

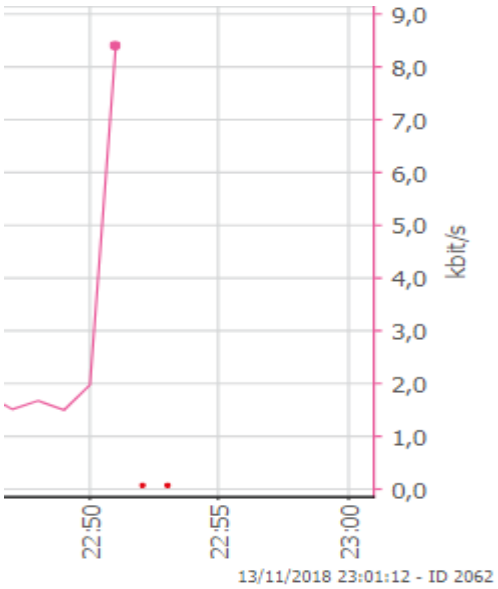


Figura 36 - R2801 Interface Serial 0/3/0

Figura 37- IP Softphone SRST Mode

As configurações do SRST bem como as *dial-peers* estarão dispostas no Apêndice Sessão 890 - SRST

4. CONCLUSÃO

As organizações estão cada vez mais olhando para o VoIP como uma alternativa atraente para a tradicional PSTN. No entanto, o êxito na implantação do VoIP não é trivial em contextos específicos e rigorosos, por isso é importante que uma organização considere todas as funcionalidades que serão exigidas de sua rede e esteja ciente dos possíveis contratempos associados à implantação deste tipo de solução.

Assim como as empresas adotam vários protocolos e tecnologias para suas redes de dados, elas devem definir os parâmetros para seus requisitos de VoIP, dependendo das imposições comerciais e técnicas que se adequem ao modelo e necessidade do negócio.

A grande variedade de protocolos VoIP, mecanismos de implantação, granularidade de configuração e amplitude de soluçoes, promove uma flexibilidade que torna os sistemas de voz baseados em VoIP muito atraentes devido ao valor que agregam ao empreendimento.

Visando garantir que as tecnologias e equipe escolhidas sejam capazes de resolver as dificuldades e obstáculos associados a esta tecnologia, esse trabalho se propôs a investigar elementos decisivos nos modelos de implantação do meio colaborativo e apresentar quais são os fatores críticos mais comuns ao se implantar um sistema de voz em um ambiente com processamento de chamadas centralizado e solução multisite.

Para isso, foi desenvolvido um estudo sobre os componentes que afetam o tráfego dos dados VoIP e técnicas e mecanismos para abordar a continuidade e qualidade dos servicos colaborativos de uma empresa. O estudo se destinou a evitar sobrecarga links de rede, priorização de dados em tempo real e ações que mantenham o processamento de chamadas ativo quando ocorrem falhas de conectividade, indisponibilidade ou interrupcao de servicos.

Foram apresentadas as considerações em ambiente *multisite*, desafios a serem tratados como: latência, jitter, perda de dados, como os *codecs* afetam a banda, o efeito da utilização de recursos de mídia centralizados e distribuídos em cada localidade utilizando *hardware*. Foi feito um estudo aprofundado sobre QoS e como

este influencia significativamente esse tipo de aplicação. Foram apresentados princípios de *design* e estratégias ao utilizar o QoS. Tratou-se, também, sobre alta disponibilidade, roteamento otimizado de chamadas, e mecanismos de *backup* em casos de falha no link WAN.

No final, foi proposto um modelo para realizar as configurações e simular um ambiente VoIP, observar o comportamento em diferentes cenários e validar os resultados e diferenças quando aplicadas as técnicas propostas. Confirmou-se, nesses testes, a indispensabilidade e conveniência desses mecanismos para a consolidação desse tipo de aplicação.

Sabendo-se que a tecnologia sempre está evoluindo e que o sistema colaborativo está se desenvolvendo e ampliando rapidamente, propõe-se a continuidade desse estudo para situações em que a infraestrutura do ambiente estará disposta na núvem. Sugere-se observar também o efeito da integração desses sistemas com mídias sociais, *contact-centers*, bancos de dados, sistemas de e-mail, as necessidades específicas para vídeo, entre outros. Dessa forma, se desenvolve a visão da composição e incorporação dos sistemas VoIP a outras soluções em diversos tipos de ambientes e *designs*.

Finaliza-se este trabalho firmando a necessidade do profundo entendimento das minúcias desse tipo de solução para a expansão de redes colaborativas escalonáveis e confiáveis que suportem os requisitos de redes de próxima geração.

Figura 38- Laboratório Montado

BIBLIOGRAFIA

1. STEFAN BRUNNER, A. A. A. **Understanding VoIP Networks**. Sunnyvale: Juniper Networks, 2004.
2. WALLACE, K. **Implementing Cisco Unified Communications Voice over IP and QoS**. Indianapolis: Cisco Press, 2011.
3. COLLABORATION Deployment Models. **Cisco Collaboration System 10.x Solution Reference Network Designs (SRND)**, 2015. Disponivel em: <https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab10/collab10/models.html>. Acesso em: 15 outubro 2018.
4. CISCO Collaboration System 10.x Solution Reference Network Designs (SRND). **cisco.com**, 2015. Disponivel em: <https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab10/collab10/models.html>. Acesso em: 18 out. 2018.
5. RECONHECENDO e categorizando os sintomas de problemas com a qualidade de voz. **cisco.com**, 2017. Disponivel em: <https://www.cisco.com/c/pt_br/support/docs/voice/voice-quality/30141-symptoms.html>. Acesso em: 28 novembro 2018.
6. BAKYT KYRBASHOV, I. B. M. K. V. J. Evaluation and Investigation of the Delay in VoIP. **RADIOENGINEERING**, Slovakia, v. 20, Junho 2011. ISSN 2.
7. CISCO. VoIP over PPP Links with Quality of Service (LLQ / IP RTP Priority, LFI, cRTP). **cisco.com**, 2006. Disponivel em: <https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/7111-voip-mlppp.html>. Acesso em: 11 nov. 2018.
8. UNDERSTANDING Delay in Packet Voice Networks. **cisco.com**, 2006. Disponivel em: <https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/5125-delay-details.html>. Acesso em: 18 novembro 2018.
9. H.SCHULZRINNE, S. C. RTP: A Transport Protocol for Real-Time Applications. **IETF**, 2003. Disponivel em: <https://tools.ietf.org/html/rfc3550>. Acesso em: 23 out. 2018.
10. SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS (G.114). **INTERNATIONAL TELECOMMUNICATION UNION**, maio 2003. 20.
11. NARBIK KOCHARIANS, P. P. **CCIE Routing and Switching v5.0**. 5. ed. Indianapolis: Cisco Press, v. 1, 2015. 957 p. Acesso em: 2018.

12 C. DEMICHELIS, T. L. P. C. RFC 3393. **tools.ietf.org**, 2002. Disponivel em: . <https://tools.ietf.org/html/rfc3393#section-1.2>. Acesso em: 20 nov. 2018.

13 OLSEN, C. **Implementing Cisco Collaboraton Applications**. Indianapolis: Cisco Press, . 2015.

14 VOIP - Per Call Bandwidth Consumption. **cisco.com**, 2016. Disponivel em: . <https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/7934-bwidth-consume.html>. Acesso em: 22 out. 2018.

15 BANDWIDTH Management. In: _____ **Cisco Collaboration System 11.x SRND**. [S.l.]: . [s.n.], 2017. Cap. 13, p. 124.

16 HARTPENCE, B. **Packet Guide to Voice over IP**. 1. ed. Sebastopol: O'Reilly, 2013. .

17 FROEHLICH, A. Voice Study Guide. In: FROEHLICH, A. **Voice Study Guide**. . Indianapolis: Wiley, 2010. Cap. 2, p. 47-52.

18 VOICE Over IP - Per Call Bandwidth Consumption. **cisco.com**, 2016. Disponivel em: . <https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/7934-bwidth-consume.html>. Acesso em: 22 novembro 2018.

19 WILLIAM ALEXANDER HANNAH, A. B. **Implementing Cisco IP telephony and Video,** . **Part 2**. Indianapolis: Cisco Press, 2016.

20 ASHARSIDD. codec-bandwidth-calculation-g711g729. **VoiceOnBits.com**, 2010. . Disponivel em: <https://voiceonbits.com/2010/08/15/codec-bandwidth-calculation-g711g729/>. Acesso em: 03 nov. 2018.

21 VOICE Over IP - Per Call Bandwidth Consumption. **Cisco**, 2016. Disponivel em: . <https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/7934-bwidth-consume.html>. Acesso em: 15 outubro 2018.

22 S. CASNER, V. J. Compressing IP/UDP/RTP Headers for Low-Speed Serial Links RFC . 2508. **IETF.org**, 1999. Disponivel em: <https://www.ietf.org/rfc/rfc2508.txt?number=2508>. Acesso em: 11 nov. 2018.

23 FINKE, J. **Implementing Cisco Unified Communications Manager Part 1**. Indianapolis: . Cisco Press, 2011.

24 H. SCHULZRINNE, C. U. S. P. RFC2833. **ietf.org**, 2000. Disponivel em: . <https://www.ietf.org/rfc/rfc2833.txt>. Acesso em: 20 nov. 2018.

25 WALLACE, K. Kevin Wallace Training, LLC. **KWtrain**. Disponivel em: . <https://www.kwtrain.com/>. Acesso em: 11 nov. 2018.

26 SALVATORE COLLORA, E. L. A. S. **Cisco CallManager Best Practices**. Indianapolis: . Cisco Press, 2004.

27 TIM SZIGETI, C. H. **End-to-End QoS Network Design**. Indianapolis: Cisco Press, 2013. .

28 NARBIK KOCHARIANS, T. V. **CCIE Routing and Switching v5.0**. 5. ed. Indianapolis: . Cisco Press, v. 2, 2015. 846 p. Acesso em: 2018.

29 A. CHARNY, C. S. J. C. R. B. M. K. B. T. J. Y. L. B. E. W. C. S. D. . N. N. D. S. L. T. RFC . 3246 - An Expedited Forwarding PHB (Per-Hop Behavior). **IETF.org**, 2002. Disponivel em: <https://tools.ietf.org/html/rfc3246>. Acesso em: 07 nov. 2018.

30 CISCO. Cisco IOS Quality of Service Solutions Configuration Guide, Release 12.2. . **cisco.com**, 2014. Disponivel em: <https://www.cisco.com/c/en/us/td/docs/ios/12_2/qos/configuration/guide/fqos_c/qcfcar.html>. Acesso em: 1 set. 2018.

31 FILHO, J. E. M. **Análise de Trafego em Redes TCP/IP**. Sao Paulo: novatec, 2017. .

32 BOGART, K. INE. **INE Training**, 2016. Disponivel em: <https://ine.com/>. Acesso em: 10 . nov. 2018.

33 J. BABIARZ, K. C. N. N. C. S. F. B. RFC 4594 - Configuration Guidelines for DiffServ . Service Classes. **IETF.org**, 2006. Disponivel em: <https://tools.ietf.org/html/rfc4594>. Acesso em: 07 nov. 2018.

34 BEHL, A. **CCIE Collaboration Quick Reference**. Indianapolis: Cisco Press, 2014. .

35 VOIP Call Admission Control. **Cisco**, 2001. Disponivel em: . <https://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/voip_solutions/CAC.html>. Acesso em: 15 outubro 2018.

36 SIÉCOLA, P. C. VoIPFIX: Uma ferramenta para análise e detecção de falhas em sistemas . de telefonia IP, 10 fevereiro 2011. 106.

37 PINOTTI, F. L. Simulação e Emulação de Tráfego Multimídia em Redes IP, agosto 2011. .

38 DEUS, M. A. D. Estratégias de Gerenciamento de Banda IP/MPLS para o Transporte . Eficiente de Serviços Integrados, set. 2007. 144.

39 CHRISTINA HATTINGH, D. S. Configuring SIP Trunks for PSTN Access, 21 julho 2010. . 46.

40 CHRISTINA HATTING, D. S. **SIP Trunking**. Indianapolis: Cisco Press, 2010.
.

41 SIP-BASED Trunk Managed Voice Services Solution Design and Implementation Guide.
. San Jose, CA: Cisco Press, 2009.

42 CISCO'S IP Telephony Solution. **cisco.com**, 2002. Acesso em: 18 out. 2018.
.

43 QUALITY of Service for Voice over IP. **cisco.com**, 2011. Disponivel em:
. <https://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/qos_solutions/QoSVoIP/QoSVoIP.html>. Acesso em: 1 nov. 2018.

APÊNDICE

Estruturação de IPs da Rede

REDES

Matriz		
	Voz:	10.1.30.x
	Dados:	10.1.40.x
	Gerenciamento:	10.1.254.x
Filial		
	Voz:	10.2.30.x
	Dados:	10.2.40.x
	Gerenciamento:	10.2.254.x
R2901:		
Interface s0/0/0:		172.16.12.1
Interface g0/1:		172.16.13.1
Interface g0/0.30		10.1.30.1
	g0/0.40	10.1.40.1
	g0/0.254	10.1.254.1
Interface loopback0		1.1.1.1
R2801:		
Interface s0/3/0:		172.16.12.2
Interface g0/1:		172.16.13.2
Interface g0/0.30		10.2.30.1
	g0/0.40	10.2.40.1
	g0/0.254	10.2.254.1
Interface loopback0:		2.2.2.2
R2621XM:		
Interface f0/0:		172.16.12.3
Interface f0/1:		172.16.13.3
Interface loopback0:		3.3.3.3
SERVIDOR ESXi:		10.1.254.10
CUCM:		10.1.40.10

COMPUTADORES e SOFTPHONES: 10.x.40.x

TELEFONES: 10.x.30.x

SRST

CUCM:

Criada referência a SRST no ip 2.2.2.2 porta 2000

Adicionada a referência no Device Pool de cada filial

R2801:

```
call-manager-fallback
  ip source-address x.x.x.x port 2000
  max-ephones 6
  max-dn 12
  secondary-dialtone 0
  transfer-system full-consult
  time-format 24
  date-format dd-mm-yy
  system message SRST Mode
```

LFI

Iperf3.exe -c <ip> -p 5201 -b 128000 -t 180 -4 -u (CLIENTE)

Iperf3.exe -s (SERVIDOR)

R2901

```
interface Serial0/0/0
  description ## Interface conectada no R2 - 2801 ##
  bandwidth 128
  no ip address
  encapsulation ppp
  ppp multilink
  ppp multilink group 1
  clock rate 128000
!
class-map match-all VOIP-RTP
  match protocol rtp
```

```
  match ip dscp ef
!
policy-map Qos-Policy
  class VOIP-RTP
    priority 80
    set dscp ef
  class class-default
    fair-queue
!
interface Multilink1
  ip address 172.16.12.1 255.255.255.0
  ppp multilink
  ppp multilink group 1
  ppp multilink interleave
  ppp multilink fragment delay 10
  service-policy output Qos-Policy
```

R2801

```
interface Serial0/3/0
  description ## Interface conectada no R1 - 2901 ##
  bandwidth 128
  no ip address
  encapsulation ppp
  ppp multilink
  ppp multilink group 1
!
interface Multilink1
  ip address 172.16.12.2 255.255.255.0
  ppp multilink
  ppp multilink interleave
  ppp multilink group 1
  ppp multilink fragment delay 10
  service-policy output Qos-Policy
!
class-map match-any ID_VOIP
  match protocol rtp
  match ip dscp ef
class-map match-all DADOS
  match access-group 140
policy-map Qos-Policy
```

```
 class ID_VOIP
  priority 80
  set dscp ef
 class DADOS
 class class-default
  fair-queue
```

R2621XM

```
test voice port 1/0/1 inject-tone network 1000
```

cRTP

```
int serial x/x/x

ip rtp header-compression

! comando necessário em ambas as interfaces

! outra opção é configurar por meio de QoS (vantagem de comprimir apenas pacotes desejados)

!

policy-map Qos-Policy-compression
 class ID_VOIP
  priority 80
  set dscp ef
  compress header ip rtp
```

TRANSCODING

R2901

```
voice-card 0
 dsp services dspfarm
sccp local Loopback0
sccp ccm 10.1.40.10 identifier 1 priority 1 version 4.1
sccp
!
sccp ccm group 1
 bind interface Loopback0
 associate ccm 1 priority 1
```

```
 associate profile 1 register transcode
!
dspfarm profile 1 transcode
 description ## TRANSCODE PROFILE 1 ##
 codec g729abr8
 codec g729ar8
 codec g711alaw
 codec g711ulaw
 codec g729r8
 codec g729br8
 maximum sessions 4
 associate application SCCP
```

R2801

```
voice-card 0
 dsp services dspfarm
sccp local Loopback0
sccp ccm 10.1.40.10 identifier 2 priority 1 version 4.1
sccp
!
sccp ccm group 2
 bind interface Loopback0
 associate ccm 2 priority 2
 associate profile 10 register transcodefilial
!
dspfarm profile 10 transcode
 description ## TRANSCODE PROFILE 10 ##
 codec g711alaw
 codec g711ulaw
 codec g729r8
 codec g729br8
 maximum sessions 4
 associate application SCCP
```

CONFIGURAÇÃO ROTEADORES E SWITCHES

R2901

```
Current configuration : 9420 bytes
```

```
!
Last configuration change at 11:09:08 BSB Fri Nov 16 2018 by fuso
!
version 15.7
hostname R2901
!
no aaa new-model
clock timezone BSB -4 0
!
ip dhcp excluded-address 10.1.30.0 10.1.30.10
ip dhcp excluded-address 10.1.40.0 10.1.40.10
!
ip dhcp pool VOZ
 network 10.1.30.0 255.255.255.0
 default-router 10.1.30.1
 option 150 ip 10.1.40.10
 lease 0 1
!
ip dhcp pool DADOS
 network 10.1.40.0 255.255.255.0
 default-router 10.1.40.1
!
ip dhcp pool MGMT
 network 10.1.254.0 255.255.255.0
 default-router 10.1.254.1
!
no ip domain lookup
ip domain name lab.local
ip name-server 8.8.8.8
ip cef
!
multilink bundle-name authenticated
!
template mon
!
crypto pki trustpoint TP-self-signed-876476126
 enrollment selfsigned
 subject-name cn=IOS-Self-Signed-Certificate-876476126
 revocation-check none
 rsakeypair TP-self-signed-876476126
```

```
!
!
crypto pki certificate chain TP-self-signed-876476126
voice-card 0
 dsp services dspfarm
!
voice service voip
 ip address trusted list
  ipv4 10.1.40.10
  ipv4 3.3.3.3
  ipv4 2.2.2.2
 allow-connections sip to sip
 fax protocol t38 version 0 ls-redundancy 0 hs-redundancy 0 fallback none
 sip
!
!
voice class uri 1 sip
 host 10.1.40.10
voice class codec 1
 codec preference 1 g711alaw
 codec preference 2 g711ulaw
 codec preference 3 g729r8
!
voice class codec 2
 codec preference 1 g729r8
 codec preference 2 g711alaw
 codec preference 3 g711ulaw
!
voice class h323 1
  h225 timeout tcp establish 2
  h225 timeout setup 2
  call preserve
!
voice class e164-pattern-map 2
  e164 4...
!
voice class e164-pattern-map 1
  e164 00300T
  e164 00800T
  e164 09090T
```

```
  e164 33024...$
  e164 000T
  e164 010...$
  e164 010..$
  e164 011[9].....$
  e164 061[9].....$
  e164 090T
  e164 01[389].$
  e164 19.$
  e164 0[0][^0].[2-6].....$
  e164 0[0][^0].[7].....$
  e164 0[0][1-9].[9].....$
  e164 0[0][1-3]...[7].....$
  e164 0[0][1-9]...[9].....$
  e164 0[7].....$
  e164 0[9].....$
  e164 0[2-6].....$
 !
!
voice class server-group 1
 ipv4 10.1.40.10
!
voice translation-rule 1
 rule 1 /^00800\(.*\)/ /0800\1/
 rule 2 /^00300\(.*\)/ /0300\1/
 rule 3 /^000\(.*\)/ /0021\1/
 rule 4 /^00\(.*\)/ /0\1/
 rule 5 /^0\(.*\)/ /\1/
!
voice translation-rule 5555
 rule 1 /5555\(....\)/ /\1/
!
!
voice translation-profile REMOVE-0
 translate called 1
!
voice translation-profile RETIRA-MASCARA-LFI
 translate called 5555
!
hw-module pvdm 0/0
```

```
!
class-map match-any ID_VOIP
 match access-group 100
 match protocol rtp
 match access-group 130
 match ip dscp ef
class-map match-all DADOS
 match access-group 140
class-map match-all VOIP-RTP
 match access-group 110
!
policy-map Qos-Policy
 class VOIP-RTP
  priority 80
  set dscp ef
 class class-default
  fair-queue
!
interface Loopback0
 ip address 1.1.1.1 255.255.255.0
!
interface Multilink1
 ip address 172.16.12.1 255.255.255.0
 shutdown
 ppp multilink
 ppp multilink interleave
 ppp multilink group 1
 ppp multilink fragment delay 10
 service-policy output Qos-Policy
!
interface Embedded-Service-Engine0/0
 no ip address
 shutdown
!
interface GigabitEthernet0/0
 description ## Interface LAN ##
 no ip address
 ip virtual-reassembly in
 duplex auto
 speed auto
```

```
!
interface GigabitEthernet0/0.1
 description Interface Nativa
 encapsulation dot1Q 1 native
!
interface GigabitEthernet0/0.30
 description Interface Voz
 encapsulation dot1Q 30
 ip address 10.1.30.1 255.255.255.0
!
interface GigabitEthernet0/0.40
 description Interface Dados
 encapsulation dot1Q 40
 ip address 10.1.40.1 255.255.255.0
 h323-gateway voip interface
 h323-gateway voip bind srcaddr 10.1.40.1
!
interface GigabitEthernet0/0.254
 description Interface MGMT
 encapsulation dot1Q 254
 ip address 10.1.254.1 255.255.255.0
!
interface GigabitEthernet0/1
 description ## Interface R3 - 2621XM ##
 ip address 172.16.13.1 255.255.255.0
 duplex auto
 speed auto
!
interface Serial0/0/0
 description ## Interface ao R2 - 2801 ##
 bandwidth 2000
 ip address 172.16.12.1 255.255.255.0
 encapsulation ppp
 clock rate 2000000
ip flow egress
service-policy output Qos-Policy
!
ip forward-protocol nd
!
ip http server
```

```
ip http authentication local
ip http secure-server
!
no ip ftp passive
ip route 2.2.2.2 255.255.255.255 172.16.12.2
ip route 3.3.3.3 255.255.255.255 GigabitEthernet0/1
ip route 10.2.0.0 255.255.0.0 172.16.12.2
!
snmp-server community LAB_LOCAL1 RW
access-list 100 permit udp any range 16384 32767
access-list 100 permit udp any range 16384 32767 any
access-list 110 permit ip any precedence critical
access-list 110 permit ip any dscp ef
access-list 130 permit ip 10.1.30.0 0.0.0.255 any
access-list 140 permit ip 10.1.40.0 0.0.0.255 any
!
control-plane
!
mgcp behavior rsip-range tgcp-only
mgcp behavior comedia-role none
mgcp behavior comedia-check-media-src disable
mgcp behavior comedia-sdp-force disable
!
mgcp profile default
!
sccp local Loopback0
sccp ccm 10.1.40.10 identifier 1 priority 1 version 4.1
sccp
!
sccp ccm group 1
 bind interface Loopback0
 associate ccm 1 priority 1
 associate profile 1 register transcode
!
dspfarm profile 1 transcode
 description ## TRANSCODE PROFILE 1 ##
 codec g729abr8
 codec g729ar8
 codec g711alaw
 codec g711ulaw
```

```
 codec g729r8
 codec g729br8
 maximum sessions 4
 associate application SCCP
!
dial-peer voice 1 voip
 description ## Entrada PSTN ##
 session protocol sipv2
 incoming called-number .
 codec g711alaw
 no vad
!
dial-peer voice 2 voip
 description ## Saida Geral PSTN ##
 translation-profile outgoing REMOVE-0
 session protocol sipv2
 session target ipv4:3.3.3.3
 destination e164-pattern-map 1
 voice-class sip bind control source-interface Loopback0
 voice-class sip bind media source-interface Loopback0
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
 no vad
!
dial-peer voice 3 voip
 description ## Entrada CUCM para CUBE ##
 session protocol sipv2
 incoming called-number .
 codec g711alaw
 no vad
!
dial-peer voice 4 voip
 description ## Saida CUBE para CUCM ##
 session protocol sipv2
 session target ipv4:10.1.40.10
 destination e164-pattern-map 2
 voice-class sip bind control source-interface Loopback0
 voice-class sip bind media source-interface Loopback0
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
```

```
 no vad
!
dial-peer voice 55553605 voip
 description ## TESTE LFI ##
 destination-pattern 55553605
 session protocol sipv2
 session target ipv4:2.2.2.2
 voice-class sip bind control source-interface Loopback0
 voice-class sip bind media source-interface Loopback0
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
 no vad
!
ip flow-export destination 10.1.40.11 9996
ip flow-export version 9
!
ntp master
ntp server 173.231.187.61
!
end
```

R2801

```
Current configuration : 7084 bytes
!
! Last configuration change at 11:09:59 BSB Fri Nov 16 2018 by fuso
version 15.1
!
hostname R2801
!
no aaa new-model
!
clock timezone BSB -4 0
!
ip dhcp excluded-address 10.2.30.0 10.2.30.10
ip dhcp excluded-address 10.2.40.0 10.2.40.10
!
ip dhcp pool VOZ
 network 10.2.30.0 255.255.255.0
```

```
 default-router 10.2.30.1
 option 150 ip 10.1.40.10
 lease 0 1
!
ip dhcp pool DADOS
 network 10.2.40.0 255.255.255.0
 default-router 10.2.40.1
!
ip dhcp pool MGMT
 network 10.2.254.0 255.255.255.0
 default-router 10.2.254.1
!
ip cef
no ip domain lookup
ip domain name lab.local
ip name-server 8.8.8.8
!
multilink bundle-name authenticated
!
voice service voip
 ip address trusted list
  ipv4 1.1.1.1
  ipv4 3.3.3.3
  ipv4 10.1.40.10
 allow-connections sip to sip
 sip
!
voice class codec 1
 codec preference 1 g711alaw
 codec preference 2 g711ulaw
 codec preference 3 g729r8
!
voice class codec 2
 codec preference 1 g729r8
 codec preference 2 g711alaw
 codec preference 3 g711ulaw
!
voice translation-rule 1
 rule 1 /^00800\(.*\)/ /0800\1/
 rule 2 /^00300\(.*\)/ /0300\1/
```

```
 rule 3 /^000\(.*\)/ /0021\1/
 rule 4 /^00\(.*\)/ /0\1/
 rule 5 /^0\(.*\)/ /\1/
!
!
voice translation-profile REMOVE-0
 translate called 1
!
voice-card 0
 dsp services dspfarm
!
class-map match-any ID_VOIP
 match access-group 100
 match protocol rtp
 match ip dscp ef
class-map match-all DADOS
 match access-group 140
!
policy-map Qos-Policy
 class ID_VOIP
  priority 80
  set dscp ef
 class DADOS
 class class-default
  fair-queue
!
interface Loopback0
 ip address 2.2.2.2 255.255.255.0
!
interface Multilink1
 ip address 172.16.12.2 255.255.255.0
 shutdown
 ppp multilink
 ppp multilink interleave
 ppp multilink group 1
 ppp multilink fragment delay 10
 service-policy output Qos-Policy
!
interface FastEthernet0/0
 description ## Ligada ao R3 - 2621XM ##
```

```
 ip address 172.16.23.2 255.255.255.0
 duplex auto
 speed auto
!
interface FastEthernet0/1
 description ## Interface LAN ##
 no ip address
 duplex auto
 speed auto
!
interface FastEthernet0/1.1
 description Interface Nativa
 encapsulation dot1Q 1 native
!
interface FastEthernet0/1.30
 description Interface Voz
 encapsulation dot1Q 30
 ip address 10.2.30.1 255.255.255.0
!
interface FastEthernet0/1.40
 description Interface Dados
 encapsulation dot1Q 40
 ip address 10.2.40.1 255.255.255.0
!
interface FastEthernet0/1.254
 description Interface MGMT
 encapsulation dot1Q 254
 ip address 10.2.254.1 255.255.255.0
!
interface Serial0/3/0
 description ## Interface ao R1 - 2901 ##
 bandwidth 2000
 ip address 172.16.12.2 255.255.255.0
 encapsulation ppp
 service-policy output Qos-Policy
!
ip forward-protocol nd
ip http server
ip http authentication local
ip http secure-server
```

```
!
!
ip route 1.1.1.1 255.255.255.255 172.16.12.1
ip route 3.3.3.3 255.255.255.255 FastEthernet0/0
ip route 10.1.0.0 255.255.0.0 172.16.12.1
!
access-list 100 permit udp any range 16384 32767
access-list 130 permit ip 10.2.30.0 0.0.0.255 any
access-list 140 permit ip 10.2.40.0 0.0.0.255 any
!
snmp-server community LAB_LOCAL1 RW
!
control-plane
!
voice-port 0/2/0
 cptone BR
 station-id name TESTE-LFI
 station-id number 33653605
 caller-id enable
!
voice-port 0/2/1
!
mgcp profile default
!
sccp local Loopback0
sccp ccm 10.1.40.10 identifier 2 priority 1 version 4.1
sccp
!
sccp ccm group 2
 bind interface Loopback0
 associate ccm 2 priority 2
 associate profile 10 register transcodefilial
!
dspfarm profile 10 transcode
 description ## TRANSCODE PROFILE 10 ##
 codec g711alaw
 codec g711ulaw
 codec g729r8
 codec g729br8
 maximum sessions 4
```

```
 associate application SCCP
!
dial-peer voice 1 voip
 description ## Entrada ##
 session protocol sipv2
 incoming called-number .
 codec g711alaw
 no vad
!
dial-peer voice 2 voip
 description ## Saida Geral PSTN ##
 translation-profile outgoing REMOVE-0
 destination-pattern 011[9].....$
 session protocol sipv2
 session target ipv4:3.3.3.3
 voice-class sip bind control source-interface Loopback0
 voice-class sip bind media source-interface Loopback0
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
 no vad
!
dial-peer voice 4 voip
 description ## SAIDA PARA CUCM ##
 destination-pattern 4...
 session protocol sipv2
 session target ipv4:1.1.1.1
 voice-class sip bind control source-interface Loopback0
 voice-class sip bind media source-interface Loopback0
 no vad
!
gateway
 timer receive-rtp 600
!
call-manager-fallback
 secondary-dialtone 0
 max-conferences 4 gain -6
 transfer-system full-consult
 ip source-address 2.2.2.2 port 2000
 max-ephones 6
 max-dn 12
```

```
 system message primary SRST MODE
 system message secondary SRST MODE
 keepalive 20
 time-format 24
 date-format dd-mm-yy
!
ntp server 1.1.1.1
```

R2621XM

```
Current configuration : 5844 bytes
!
version 12.4
!
hostname R2621XM
!
no aaa new-model
clock timezone BSB -4
ip cef
!
no ip domain lookup
ip domain name lab.local
ip name-server 8.8.8.8
!
multilink bundle-name authenticated
!
voice service voip
 allow-connections sip to sip
 sip
  bind control source-interface Loopback0
  bind media source-interface Loopback0
!
voice translation-rule 1
 rule 1 /3316\(....\)/ /\1/
!
voice translation-rule 2
 rule 1 /3302\(....\)/ /\1/
!
voice translation-profile RETIRA-MASCARA-FILIAL
```

```
 translate called 2
!
voice translation-profile RETIRA-MASCARA-HQ
 translate called 1
!
interface Loopback0
 ip address 3.3.3.3 255.255.255.0
!
interface FastEthernet0/0
 description ## Ligado ao R2 ##
 ip address 172.16.23.3 255.255.255.0
 duplex auto
 speed auto
!
interface FastEthernet0/1
 description ## Interface conectada ao R1 ##
 ip address 172.16.13.3 255.255.255.0
 duplex auto
 speed auto
!
ip route 1.1.1.1 255.255.255.255 FastEthernet0/1
ip route 2.2.2.2 255.255.255.255 FastEthernet0/0
ip route 10.1.0.0 255.255.0.0 FastEthernet0/1
ip route 10.2.0.0 255.255.0.0 FastEthernet0/0
ip route 172.16.12.0 255.255.255.0 FastEthernet0/1
!
ip http server
ip http authentication local
ip http secure-server
!
snmp-server community LAB_LOCAL1 RW
!
voice-port 1/0/0
 cptone BR
 station-id name SP - 011
 station-id number 01134294000
 caller-id enable
!
voice-port 1/0/1
 cptone BR
```

```
 description TELEFONE BRASILIA 061
 station-id name PSTN BSB 61
 station-id number 061999849514
 caller-id enable
!
dial-peer voice 1 voip
 description ## Entrada R1 ##
 session protocol sipv2
 incoming called-number .
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
 no vad
!
dial-peer voice 2 voip
 description ## Saida R1 ##
 translation-profile outgoing RETIRA-MASCARA-HQ
 preference 1
 destination-pattern 33164...
 session protocol sipv2
 session target ipv4:1.1.1.1
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
 no vad
!
dial-peer voice 4 voip
 description ## Saida R2 ##
 translation-profile outgoing RETIRA-MASCARA-FILIAL
 destination-pattern 3302....
 session protocol sipv2
 session target ipv4:2.2.2.2
 dtmf-relay rtp-nte sip-kpml
 codec g711alaw
 no vad
!
dial-peer voice 30 pots
 description ## Entrada PSTN ##
 incoming called-number .
!
dial-peer voice 40 pots
 port 1/0/1
```

```
!
dial-peer voice 100 pots
 description ## LIGACAO EMERGENCIA ##
 destination-pattern 1..$
 port 1/0/1
!
dial-peer voice 102 pots
 description ## LIGACAO SERVICOS ##
 destination-pattern 0[38]00T
 port 1/0/1
!
dial-peer voice 104 pots
 description ## LIGACAO A COBRAR ##
 destination-pattern 9090T
 port 1/0/1
!
dial-peer voice 108 pots
 description ## LICAGAO LOCAL MOVEL ##
 destination-pattern [9].....$
 port 1/0/1
!
dial-peer voice 110 pots
 description ## LICAGAO INTERURBANA FIXO ##
 destination-pattern [0]..[2-6].....$
 port 1/0/1
!
dial-peer voice 112 pots
 description ## LICAGAO INTERNACIONAL ##
 destination-pattern 00T
 port 1/0/1
!
dial-peer voice 55553605 voip
 description ## TESTE LFI ##
 destination-pattern 55553605
 session protocol sipv2
 session target ipv4:1.1.1.1
 codec g711alaw
 no vad
!
dial-peer voice 61 pots
```

```
 description ## CHAMADA PARA PSTN BSB ##
 destination-pattern 61+
 port 1/0/1
!
dial-peer voice 11 pots
 description ## CHAMADA PARA PSTN Sao Paulo ##
 destination-pattern 11+
 port 1/0/0
!
ntp server 1.1.1.1
```

SW1 / SW2

```
Current configuration : 5825 bytes
!
version 12.2
hostname SW1_2960
!
no aaa new-model
!
no ip domain-lookup
ip domain-name lab.local
!
mls qos
!
vlan internal allocation policy ascending
!
class-map match-any ID_VOIP
  match access-group 111
!
policy-map QOS_VOIP
  class ID_VOIP
   set dscp ef
  class class-default
!
interface FastEthernet0/1-24
 switchport access vlan 40
 switchport mode access
 switchport voice vlan 30
```

```
 mls qos trust device cisco-phone
 mls qos trust dscp
 spanning-tree portfast
!
interface GigabitEthernet0/1
 description ## Servidor ESXi ##
 switchport mode trunk
!
interface GigabitEthernet0/2
 description ## Rtr 2901 ##
 switchport mode trunk
!
interface Vlan1
 shutdown
!
interface Vlan254
 ip address 10.1.254.2 255.255.255.0
 no ip route-cache
!
ip http server
access-list 30 permit 10.1.30.0 0.0.0.255
access-list 111 permit udp any range 16384 32767
access-list 111 remark RECEBENDO CHAMADA SIP
access-list 111 permit udp any range 16384 32767 any
access-list 111 remark FAZENDO CHAMADA SIP
snmp-server community LAB_LOCAL1 RW
```