3
Qualidade de Serviço

A qualidade de serviço pode ser definida como a habilidade configurada numa rede, que visa garantir e manter certos níveis de desempenho para cada aplicação de acordo com a demanda específica de cada usuário. Embora essa definição seja medida através do conceito de satisfação do usuário, a qualidade de serviço pode ser medida segundo alguns parâmetros de desempenho, tais como: vazão, atraso na entrega, jitter, probabilidade de perda de pacotes, ordem na entrega dos pacotes, dentre outros. Neste capítulo apresentam-se os conceitos de qualidade de serviço em redes de pacotes e mecanismos utilizados para implementação de uma rede com tais características.

A qualidade de serviço deve ser fim a fim, ou seja, o tráfego tem que ser tratado inicialmente na rede local (LAN) de origem, depois no próprio roteador, posteriormente nas conexões de longa distância (WAN) e roteadores intermediários, no roteador de destino, e finalmente na rede local de destino.

3.1. Fases de QoS

Pode-se dividir o ciclo de vida de um serviço com qualidade negociada em quatro fases: iniciação, requisição, provisão e gerência [13]. Essas fases devem funcionar conjuntamente e paralelamente, mas em diferentes escalas de tempo. O correto funcionamento depende de algoritmos assíncronos que gerenciem esses módulos. Na figura 16, pode-se visualizar estas quatro fases de qualidade de serviço.

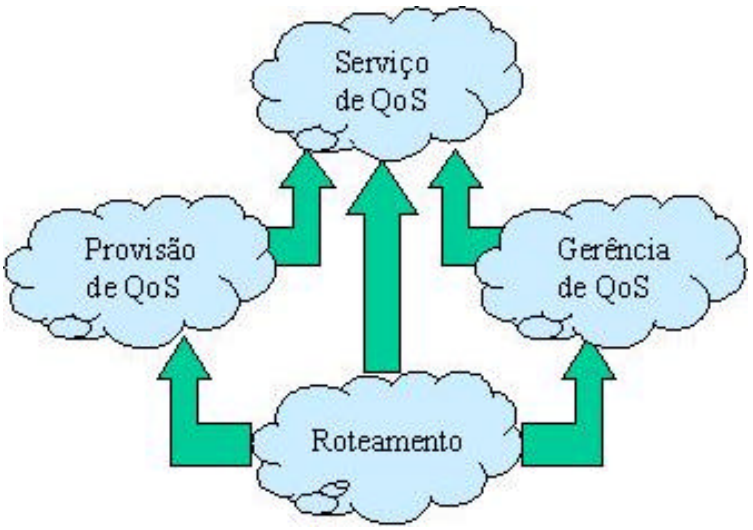

Figura 16 - Fases de Qualidade de serviço

Na fase de iniciação do sistema, as raízes das árvores de recursos virtuais são estabelecidas, bem como alguns de seus recursos virtuais com os respectivos escalonadores.

A garantia de QoS se faz reservando uma parcela de cada recurso a um determinado fluxo e que caberá ao escalonador do recurso compartilhá-lo conforme a reserva realizada. A cada recurso real de um provedor de serviços está associada uma árvore de recursos virtuais, conforme ilustrada na figura 17.

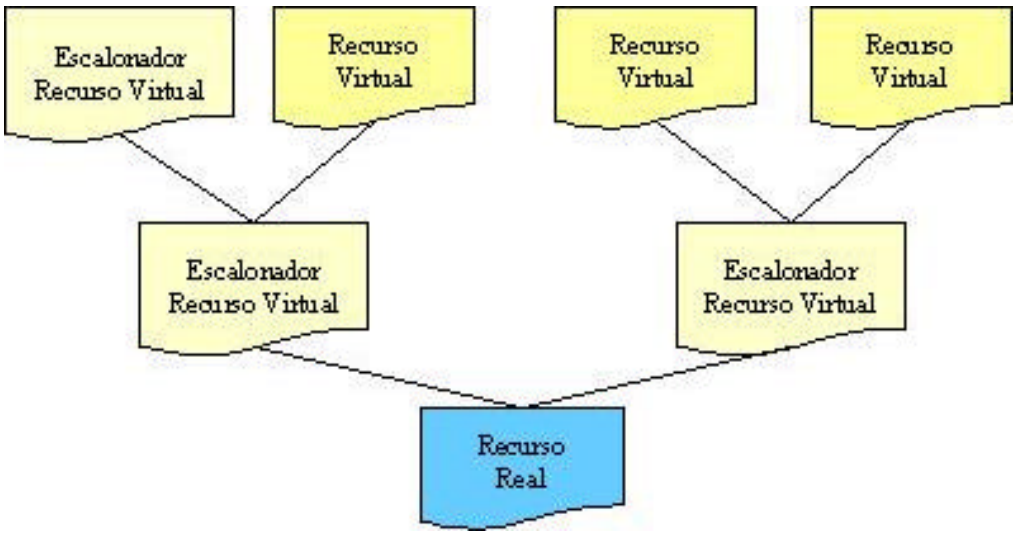

Figura 17 - Árvore de recursos virtuais

Dada uma requisição de QoS, um controle de admissão deve ser realizado no recurso para saber se existe a possibilidade de atender a QoS solicitada e, em caso afirmativo, fazer a reservada parcela do recurso, responsável pelo

atendimento. A essa parcela dá-se o nome de recurso virtual, pois é como se criasse um novo recurso, a partir do original, com a capacidade de atender apenas a solicitação. O escalonador de recursos é o responsável por garantir o compartilhamento correto deste recurso.

Parte da árvore de recursos virtuais é criada na fase de iniciação. As folhas da árvore serão criadas na fase de provisão, e poderão ser alteradas na fase de gerência. Para serem criadas, no entanto, o usuário do serviço precisa explicitar a qualidade desejada e o tipo de fluxo de dados que receberá esta qualidade. Com isso, recursos poderão ser reservados, criando novas folhas nas árvores.

Na fase de requisição de serviço o usuário deve especificar para o provedor qual a qualidade de serviço desejada para um determinado tipo de fluxo. Baseado nesta especificação começa todo um processo de negociação presente na fase de provisão.

Na fase de provisão, é realizada a divisão da responsabilidade pela provisão do serviço por todos os recursos participantes, para atingir a qualidade de serviço fim a fim. Uma vez realizada esta orquestração, uma parcela de cada recurso deve ser reservada.

A fase de gerência exige o monitoramento do sistema e do fluxo de entrada. Uma qualidade de serviço negociada pode ser degradada por violação da carga controlada por um usuário ou pelo congestionamento da rede. Em qualquer um dos dois casos, providências devem ser tomadas para trazer o sistema para um estado tal, que possa garantir os contratos de qualidade já negociados.

Alguns mecanismos de gerência são utilizados para manter a qualidade de serviço negociada: Policiamento de fluxo que verifica se a carga controlada pelo usuário está sendo mantida por ele; monitoramento que permite a cada nível do sistema inferir sobre a qualidade de serviço oferecida pelo nível inferior; a manutenção que compara a qualidade de serviço monitorada com a esperada e dispara mecanismos de ajuste; degradação que sinaliza perda de qualidade de serviço que os mecanismos de manutenção não conseguem recuperar e finalmente a sinalização que permite ao usuário especificar o intervalo de um ou mais parâmetros de qualidade de serviço, que devem ser monitorados e o usuário avisado (sinalizado).

3.2.Parâmetros de desempenho de QoS

Qualidade de serviço também pode ser definida como a medida do desempenho para uma rede que reflita sua disponibilidade e a qualidade do serviço de transmissão. A disponibilidade do serviço é um elemento crucial, ou seja, antes que todo o QoS possa ser executado com sucesso, a infra-estrutura da rede deve ser projetada para estar disponível de acordo com a demanda da rede implementada. As 5 (cinco) definições listadas a seguir são consideradas para as medições de desempenho do cenário proposto determinando a qualidade de transmissão da rede.

3.2.1.Disponibilidade

A fração de um intervalo de tempo em que há conectividade entre um ponto de ingresso e um ponto especificado de saída é definida como a disponibilidade da rede. A disponibilidade de serviço é definida como a fração de tempo em que o serviço está disponível entre um ponto especificado de ingresso e um ponto de saída com os limites definidos em um acordo de nível de serviço (SLA).

3.2.2.Perda de pacotes

É a medida comparativa entre os pacotes fielmente transmitidos e recebidos e o número total dos pacotes que foram transmitidos. A perda é expressa como a porcentagem dos pacotes que foram descartados. A perda é tipicamente uma função da disponibilidade. Em uma situação de rede altamente disponível, a perda (durante períodos de não congestionamento) seria essencialmente zero. Durante períodos de congestionamento, entretanto, os mecanismos de QoS determinariam quais pacotes seriam devidamente descartados. Sem QoS os pacotes são descartados indiscriminadamente em uma situação de buffer cheio.

3.2.3. Atraso (Delay)

Quantidade finita de tempo gasto por um pacote para alcançar o ponto de recepção após ser transmitido a partir de uma origem definida. É a soma de todos os intervalos de tempo, fixos e variáveis, gastos desde a origem até o destino. Inclui tempos de processamento, serialização, propagação, enfileiramento,

bufferização, entre outros. No exemplo da voz, o atraso é definido como o tempo gasto pelo som a partir da boca do emissor até atingir a orelha do ouvinte.

3.2.4.Jitter

É definido como a diferença no atraso fim a fim entre os pacotes. Por exemplo, se um pacote gasta 100 ms para atravessar a rede da origem até o destino e o pacote seguinte utiliza 125 ms para fazer o mesmo percurso, o jitter calculado é de 25 ms. No caso de voz, a existência de jitter afeta negativamente a qualidade da voz para o receptor. Cada estação de uma conversação de voz ou vídeo sobre IP possui um buffer dejitter. Este buffer é utilizado para atenuar a diferença de tempo de chegada entre os pacotes que contêm voz. Um buffer dejitter pode ser dinâmico ou adaptável e alguns Codec´s podem se ajustar para uma mudança de até 30 ms de variação no tempo de chegada dos pacotes.

3.2.5.Banda passante

É a largura de banda disponível ao usuário entre um ponto de presença de origem e um de destino.

3.3.Modelo IntServ x Modelo DiffServ

Os diferentes graus de sensibilidade dos tráfegos de dados, voz e vídeo, em relação ao desempenho da rede e, em particular, ao atraso de entrega dos pacotes, motivam o desenvolvimento de uma rede de pacotes que tenha a capacidade de dar garantias e tratamentos diferenciados aos diversos tipos de tráfego, ou seja, uma rede com uma qualidade de serviço acordada. Este desenvolvimento, atualmente em curso na IETF, deu origem a dois modelos estruturados conhecidos como Serviços Integrados (Intserv) e Serviços Diferenciados (Diffserv). O objetivo desse trabalho é caracterizar o modelo Diffserv, identificando as principais funções e estruturas de provisão de QoS presentes e informar os principais problemas encontrados no modelo IntServ, que justificam a não utilização do mesmo para a simulação de um core de rede IP/MPLS adequado aos padrões de QoS do UMTS.

3.3.1.Serviços Integrados

O modelo de serviços integrados foi o primeiro a ser especificado para o estabelecimento de classes de serviço com diferentes requisitos de QoS especificadas através de um contrato de serviço entre usuário e a rede. Essas classes se tornam necessárias devido ao desenvolvimento de novas aplicações de tempo real com uma grande sensibilidade ao atraso da rede; reserva de recursos para atendimento destes requisitos e processamento específico para cada fluxo de dados.

Para implementação deste modelo, a funcionalidade básica dos roteadores deve ser estendida, passando a incluir os seguintes componentes:

- Escalonador de Pacotes: mecanismo responsável por gerenciar o encaminhamento dos vários fluxos de dados. Esses escalonadores são implementados através de alguma disciplina de fila (apêndice A) no local onde os pacotes são enfileirados;
- Classificador: associa o pacote a sua específica classe no momento em que ele chega ao roteador. Pacotes pertencentes a uma mesma classe recebem tratamentos semelhantes no escalonador de pacotes;
- Controle de Admissão: Controla a entrada de um novo fluxo na rede. Para isto os recursos da rede devem ser provisionados de tal forma, que a entrada do novo fluxo não degrade o desempenho contratado dos demais;
- Protocolo de Reserva de Recursos: faz parte do controle de admissão sendo responsável pela sinalização de requisição de QoS aos roteadores pertencentes ao caminho de transmissão. O protocolo mais utilizado em fazer essas reservas de recursos no Intserv é o RSVP.

O modelo Intserv tem como uma das suas principais características, a alta granularidade na alocação de recursos, realizada por fluxos individuais. Na sua forma de implementação mais usual, o gerenciamento de recursos em provedores de serviços integrados é dinâmico, normalmente realizado com o uso do protocolo RSVP. A gerência dinâmica, aliada à alocação por fluxos, contribui para a alta carga de sinalização, dificultando a aplicação do modelo Intserv em provedores de serviços localizados próximos ao núcleo da infra-estrutura de rede. Neste trabalho o modelo Intserv não está sendo levado em consideração, pois atualmente, é

considerado como não sendo escalável para ser utilizado em core de redes de grande dimensões, como a rede “all-IP” UMTS. Em resumo, os principais problemas são:

- A quantidade de informações de estados cresce proporcionalmente ao número de fluxos que o roteador tem que tratar. Isso impõe uma sobrecarga nos roteadores, em termos de capacidade de armazenamento e processamento;
- Para cada fluxo deve haver sinalização em cada nó (sistema final ou roteador), ou seja, a quantidade de mensagens suficientes para a troca de informações de sinalização é muito grande;
- As exigências nos roteadores são bastante altas, ou seja, todos os nós têm que implementar RSVP, classificação, controle de admissão e escalonamento de pacotes.

3.3.2.Serviços Diferenciados

O modelo de serviços diferenciados foi proposto, pelo IETF, para processar agregados de fluxos ao invés de fluxos individuais, conforme proposto no modelo IntServ. Através desta agregação de fluxos em classes de serviço, o Diffserv permite que o número de estados a ser mantido nos roteadores, com a conseqüente carga de sinalização, seja reduzido. No modelo Diffserv, os fluxos de pacotes são processados individualmente ao entrar na rede.

Para controlar os parâmetros de qualidade de serviço propostos na solução são utilizados cinco definições para caracterizar as políticas de qualidade de serviço configuradas na rede de dados conforme detalhado na figura 18.

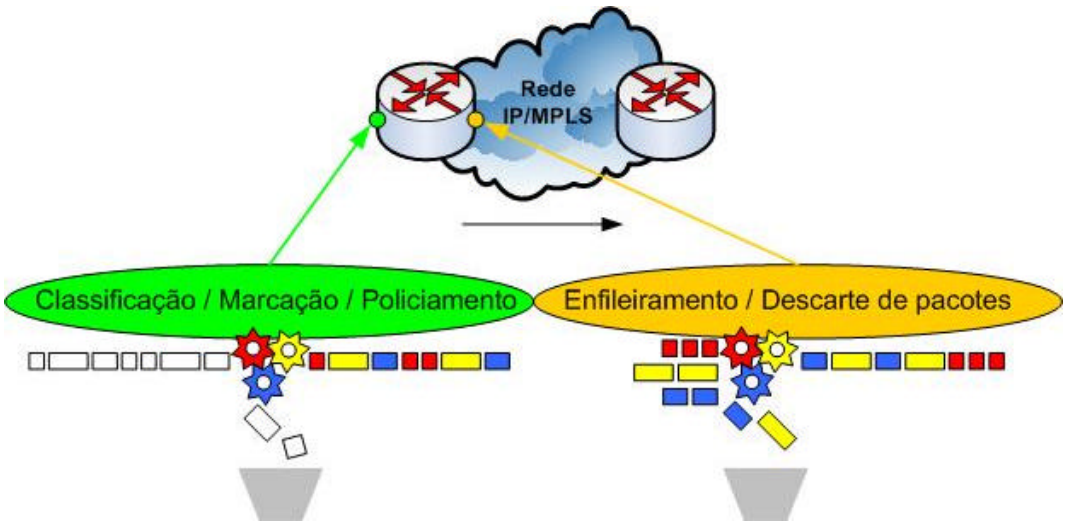

Figura 18 – Definições de QoS da rede

3.3.2.1.Classificação

Após a adequação de toda a infra-estrutura, as aplicações existentes são classificadas nos acessos do backbone. Esta marcação, bastante abrangente, será convertida em um modelo de quatro classes, definido no UMTS, para o backbone, onde serão aplicadas as políticas.

A primeira exigência da política de marcação é identificar o tipo de tráfego que requer o tratamento diferenciado, conforme já detalhado no item 3.1 das fases de QoS. A classificação pode ocorrer através de variado número de parâmetros, incluindo: endereço origem, endereço destino, protocolo, porta, IP Precedence, IP DSCP, dentre outros. Para os experimentos desse trabalho é utilizado o parâmetro de protocolo e porta para diferenciar os perfis de tráfego classificados. A classificação do tipo complexo (endereço, protocolo, etc.) deve ocorrer no ponto mais próximo da origem dos dados.

3.3.2.2. Marcação

No roteador de entrada, caso o pacote não tenha sido previamente marcado é enquadrado em uma determinada classe de agregados. A marcação deste pacote é feita no campo TOS do cabeçalho IP.

O modelo DiffServ utiliza 6 dos 8 bits alocados no campo TOS para transportar o DSCP que seleciona um PHB, ou seja, esse critério criado para

encaminhamento para uma determinada classe é denominado Per Hop Behaviour (PHB).

Embora receba o nome de “serviço diferenciado”, o IETF não tem a intenção de padronizar os serviços, especificando apenas PHB´s. Um PHB descreve como será realizado o encaminhamento de pacotes pertencentes a uma mesma classe de serviço em um roteador de borda.

A arquitetura DiffServ [9] usa o termo SLA para descrever o serviço contratado que o cliente deve receber. O SLA deve conter as regras que podem ser informadas aos clientes especificando o desempenho que o usuário deve esperar desses serviços bem como os acordos comercias.

Roteadores na borda da rede identificam pacotes baseados no precedente IP ou nos campos DSCP localizados no cabeçalho do datagrama IP versão 4. Os recursos de rede que suportam Diffserv usam os códigos DSCP no cabeçalho IP para selecionar um PHB para cada pacote que passa por esse roteador.

Os seis bits mais significativos do byte ToS são usados para definir o padrão DSCP, os dois menos significativos são definidos como ECN (Early Congestion Notification). O IP Precedence usa 3 bits, enquanto que o DSCP que é uma extensão do IP Precedence utiliza 6 bits para selecionar o PHB definido para pacote escoado nos elementos de rede. A figura 19 faz uma comparação entre o byte ToS e o campo Diffserv [12]:

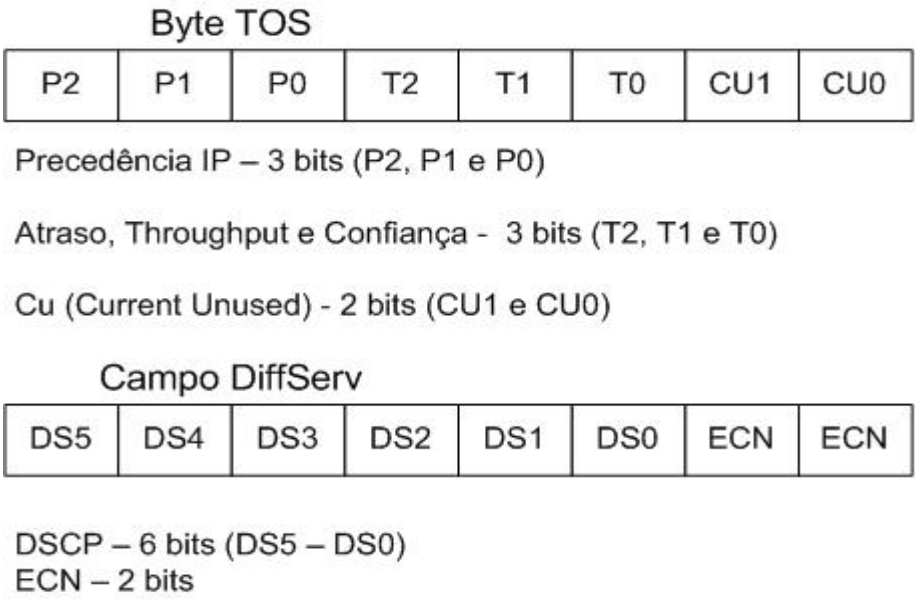

Figura 19 - Byte TOS x Campo DiffServ

A seguir pode-se verificar na tabela 1, a utilização dos bits no campo Diffserv:

XXX00000 bits 0, 1, 2 (DS5, DS4, DS3) são bits de precedência, onde:
111 – Controle de rede = Precedence 7
110 – Controle de Internetwork = Precedence 6
101 – CRITIC/ECP = Precedence 5
100 – Flash Override = Precedence 4
011 – Flash = Precedence 3
010 – Immediate = Precedence 2
001 – Prioridade = Precedence 1
000 – Routine = Precedence 0
000XXX00 bits 3, 4, 5 (DS2, DS1, DS0) são: atraso, throughput e bits Reliability.
Bit 3 = Delay [D] (0 = normal; 1 = baixo)
Bit 4 = Delay [D] (0 = normal; 1 = alto)
Bit 5 = Delay [D] (0 = normal; 1 = alto)
000000XX bits 6, 7 (ECN)

Tabela 1 - Utilização de bits no padrão DiffServ

Os 3 bits mais significativos dos bits de Precedência IP (P2, P1 e P0), já mencionados na figura 19, são reorganizados e renomeados dentro das categorias da tabela 2:

Nível de precedência	Descrição
7	“keep alive” dos protocolos de roteamento
6	Protocolos de roteamento IP
5	EF – tráfego de voz
4	Classe 4 – tráfego de vídeo conferência e vídeo streaming
3	Classe 3 – tráfego de sinalização de chamadas de voz e vídeo
2	Classe 2 – tráfego de dados
1	Classe 1 – tráfego de dados
0	Melhor esforço

Tabela 2 - Categorias DiffServ

O backbone MPLS configurado também é capaz de interpretar pacotes oriundos da rede MPLS utilizados para manter um indicador de QoS. São 3 (três) bits dentro do label MPLS onde por default, estes valores são copiados dos 3 bits mais significativos do campo TOS do header IP já mencionados anteriormente. Logo, permitem, até 8 valores.

Atualmente encontram-se padronizados dois PHB´s: Expedited Forwarding (EF) [10], e Assured Forwarding (AF) [11].

3.3.2.2.1.Assured Forwarding (AF)

O padrão Diffserv não especifica uma definição precisa de “baixa”, “média” e “alta” probabilidade de descarte dos pacotes envolvidos no tráfego da rede. A RFC 2597 [11] define quatro classes do tipo AF - AF1x, AF2x, AF3x e AF4x. Dentro de cada classe há três probabilidades de descarte representadas pela letra “x” da frase anterior. Dependendo da política de qualidade de serviço atribuída no backbone de dados, pacotes podem ser selecionados por um PHB baseado nas premissas de throughput, atraso, jitter, perda ou prioridade de acesso para os serviços de rede.

As classes 1 a 4 mencionadas na coluna do item Descrição da tabela 2 são classificadas dentro do padrão Diffserv como as classes do tipo AF. A tabela 3 ilustra o código DSCP para cada classe AF de acordo com a probabilidade definida. Os bits 0, 1 e 2 definem a classe; os bits 3 e 4 especificam o descarte de pacotes; bit 5 é sempre igual à zero.

	Classe 1	Classe 2	Classe 3	Classe 4
Descarte BAIXO	001010 AF11 DSCP 10	010010 AF21 DSCP 18	011010 AF31 DSCP 26	100010 AF41 DSCP 34
Descarte MÉDIO	001100 AF12 DSCP 12	010100 AF22 DSCP 20	011100 AF32 DSCP 28	100100 AF42 DSCP 36
Descarte ALTO	001110 AF13 DSCP 14	010110 AF23 DSCP 22	011110 AF33 DSCP 30	100110 AF43 DSCP 38

Tabela 3 - Código DSCP x padrão Diffserv

3.3.2.2.2.Expedited Forwarding (EF)

O PHB EF [10] deve ser implementado quando há a necessidade de escoar um perfil de tráfego que tenha características de: baixa perda, baixo atraso, pouca variação no atraso (baixo jitter) e largura de banda garantida; quando providos serviços fim a fim no domínio Diffserv. Os pacotes que não estiverem dentro do padrão configurado, são descartados imediatamente sem a possibilidade de uma remarcação e/ou reclassificação. O PHB EF é utilizado para identificar e encaminhar tráfego de aplicações multimídia como voz e vídeo em tempo real.

O Expedited forwarding (EF) é definido como uma classe única com encaminhamento expresso. A taxa com que os pacotes marcados com o PHB EF são escoados na interface de saída de um roteador deve ser ao menos configurada com uma taxa fixa que independa da carga oferecida pelos demais tráfegos marcados por classes diferentes da EF. Os DSCP´s binário e decimal equivalentes ao PHB EF são respectivamente 101110 e 46. A configuração dos perfis de tráfegos marcados com os DSCP´s citados pode ser realizada por diferentes mecanismos de escalonamento de filas. O mecanismo escolhido neste trabalho é o PQ, detalhado no apêndice A.

3.3.2.3. Policiamento e Shaping

"Policing" e "Shaping" são ferramentas utilizadas para identificar violações de quantidade de tráfego em uma interface, durante um intervalo de tempo (normalmente um segundo). A principal diferença entre elas é a maneira como estas ferramentas respondem às violações. A configuração de policiamento normalmente descarta o tráfego excedente e é utilizado para forçar que o tráfego recebido de um usuário não ultrapasse o valor contratado em cada intervalo de tempo. A configuração de "shaping" tipicamente atrasa o tráfego excessivo utilizando um buffer para armazenar os pacotes e modelar o fluxo quando a transmissão de uma determinada origem é maior do que o esperado. Normalmente é utilizado em interfaces de saída quando há uma configuração de policiamento controlando o volume de tráfego no link de entrada no roteador à frente. Tais buffer´s de enfileiramento são finitos e atuam de forma bastante semelhante a um líquido entrando em um container através de um funil. Se a água continuar entrando no funil muito mais rapidamente do que estiver saindo, eventualmente o

container vai ser sobrecarregado e o líquido irá derramar. Quando os buffer´s estão preenchidos, novos pacotes são descartados assim que encaminhados para o enfileiramento. Este mecanismo é conhecido como “tail-drop”.

3.3.2.4.Gerência de Congestionamento

Encontram-se nesta categoria as ferramentas que determinam como um frame ou pacote é encaminhado. Sempre que pacotes entram em um dispositivo mais rapidamente do que podem sair, pode ocorrer um ponto de congestionamento. Os dispositivos possuem buffer´s que podem ser configurados para permitir o encaminhamento de pacotes de alta prioridade mais cedo do que os de baixa prioridade. Os algoritmos de enfileiramento, detalhados no apêndice A, são ativados por interface apenas quando esta enfrenta congestionamento, e são desativados, quando volta ao funcionamento normal.

Sabe-se que o congestionamento provoca o transbordamento de buffer´s com conseqüente perda de pacotes [17]. Aumentando-se o tamanho do buffer diminui-se a perda de pacotes, mas se tende a aumentar o atraso. Observa-se, portanto, que a dimensão e o gerenciamento dos pacotes no buffer são fatores importantes em uma rede com QoS. São propostos dois mecanismos de gerenciamento ativos de fila, cuja ação específica depende do estado do buffer: RED e WRED.

3.3.2.4.1.RED (Random Early Detection)

Mecanismo proposto em 1993 por Floyd e Jacobson. O RED permite que o roteador descarte pacotes antes que ocorra uma saturação na fila. Consequentemente, uma resposta ao congestionamento ocorrerá mais cedo resultando num menor tamanho médio da fila. A utilização do RED é interessante por algumas razões: primeiro, o atraso na fila irá diminuir, tornando-se interessante no uso de aplicativos interativos; segundo, o descarte de pacotes não ocorrerá em surtos, pois os pacotes são descartados com uma probabilidade p_d detalhada abaixo:

$$p_d = \frac{p_b}{(1 - cont * p_b)} \tag{3.1}$$

onde *cont* é o número de pacotes desde o último marcado e,

$$p_b = \max_p * \frac{(avg - \min_{th})}{(\max_{th} - \min_{th})} \quad (3.2)$$

onde $\max_p$ é a probabilidade máxima que p_b pode assumir e *avg* é o tamanho médio da fila.

Como se pode verificar na figura 20, o RED possui um função com comportamentos distintos nos intervalos: [0, $\min_{th}$), [$\min_{th}$, $\max_{th}$), e [$\max_{th}$;+8). No primeiro intervalo, nenhum pacote é descartado, no segundo, pacotes são descartados com uma probabilidade p_d e no terceiro todos os pacotes são descartados.

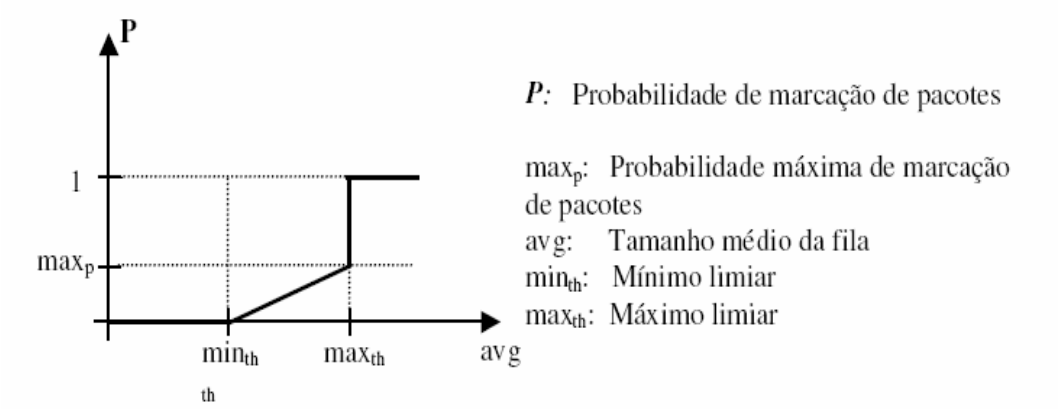

Figura 20 - Mecanismo RED

3.3.2.4.2.WRED (Weighted Random Early Detection)

WRED é uma implementação desenvolvida pelo fabricante Cisco que combina as funcionalidades do RED com a classificação por precedência IP. Essa implementação suporta até oito níveis de precedência, onde cada um desses níveis é configurado com distintos parâmetros RED conforme demonstrado na figura 21. Por ser uma extensão do RED, o gerenciamento ativo de fila WRED trabalha de forma semelhante a ele. Porém no WRED, os efeitos se estendem as i classes. Ou seja, se a fila média se encontrar antes do limiar $\min_{th}(i)$, nenhum pacote da classe i é descartado, se ela estiver no intervalo [$\min_{th}(i)$; $\max_{th}(i)$), os pacotes da classe i são descartados com probabilidade p-a(i) e se *avg* estiver acima de $\max_{th}(i)$, todos os pacotes da classe i são descartados.

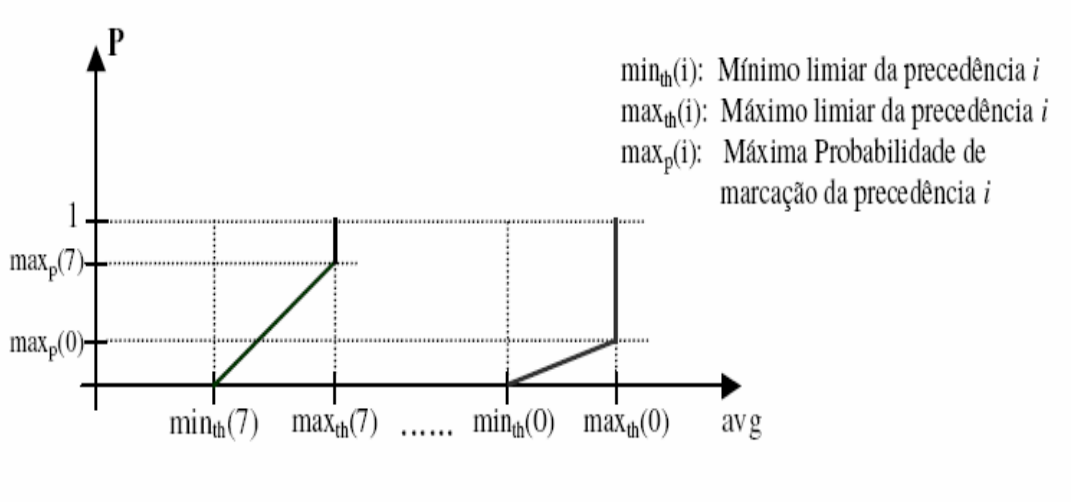

Figura 21 - Mecanismo WRED

3.4. Suporte de QoS em UMTS – conceitos e arquiteturas

Após explicitar neste capítulo os conceitos básicos envolvidos para a implementação de qualidade de serviço, relata-se nesta parte do trabalho, a especificação técnica, segundo [6], para a adequação de qualidade de serviço do padrão UMTS.

Os serviços de rede são considerados fim a fim numa rede UMTS quando há uma chamada concreta de um usuário para outro. Um serviço fim a fim deve ser implementado segundo a necessidade do usuário, que estará informando se a qualidade de serviço atribuída está satisfazendo às suas necessidades ou não.

O serviço de transporte inclui todos os aspectos para habilitar o provisionamento de uma qualidade de serviço contratada. Estes aspectos são: o controle de sinalização, transporte dos dados do usuário e a funcionalidade da gerência de qualidade de serviço. A arquitetura do nível de serviço de transporte do UMTS está detalhada na figura 22. A qualidade de serviço proposta neste trabalho está centrada no "Backbone Bearer Service", no entanto a origem dos pacotes é feita em máquinas que emulam tráfego equivalente ao tráfego de usuários (TE) sem considerar o atraso da interface aérea.

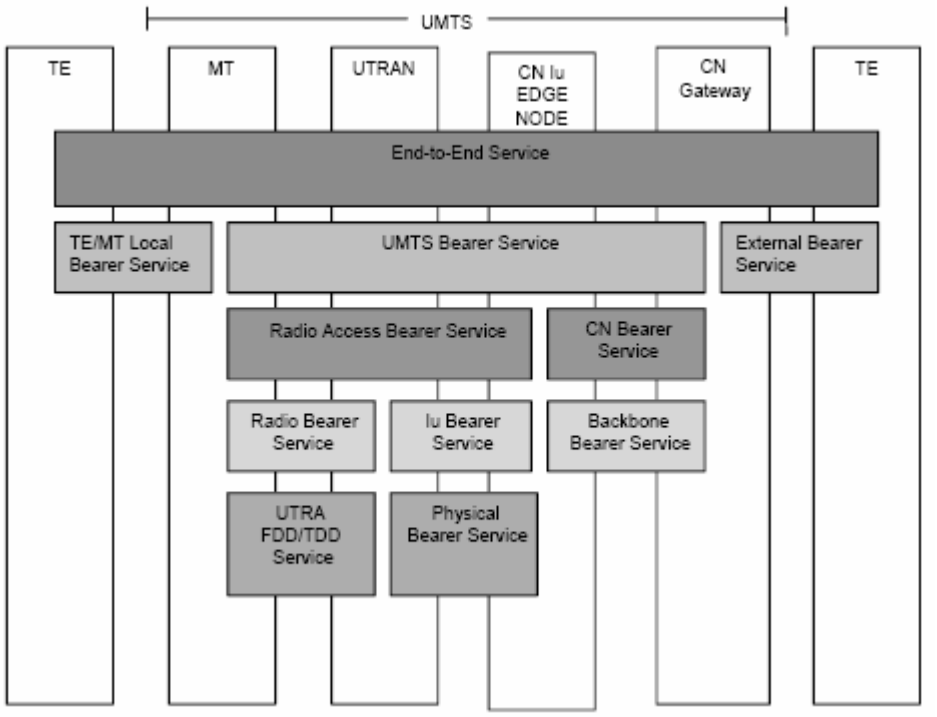

Figura 22 - Arquitetura QoS do UMTS

3.4.1.Requerimentos de QoS para o usuário final no core da rede de transporte

O serviço de transporte do core de rede do padrão UMTS conecta o nó UMTS CN Iu Edge com o Gateway CN para a rede externa. A regra desse serviço é para controlar e utilizar a rede do backbone com o objetivo de providenciar o serviço de transporte UMTS contratado. O pacote UMTS deve ser compatível com diferentes tipos de backbone de dados suportando a variedade de qualidades de serviços que podem ser utilizadas. O core de rede para o serviço de transporte não especifica uma configuração específica para transportar os serviços de uma rede UMTS. Neste trabalho está sendo utilizado um core de rede IP/MPLS como rede de transporte dos serviços vinculados ao padrão citado.

3.4.2. Classes de qualidade de serviço definidas no UMTS

Para a definição das classes de qualidade de serviço no padrão UMTS, as restrições e limitações da interface aérea devem ser levadas em consideração.

Os mecanismos de QoS providos para a rede celular proposta, devem ser robustos e capazes de atender as demandas específicas de um ambiente sem fio. São propostas quatro diferentes classes de QoS:

- Classe Conversational;
- Classe Streaming;
- Classe Interactive;
- Classe Background.

Obs.: O padrão das classes será mantido em inglês para o melhor entendimento deste trabalho.

O principal fator que distingue as quatro classes é como o senso de atraso no envio dos pacotes é percebido, isto é, a classe Conversational deve possuir o tráfego com menor atraso na rede, enquanto que a classe background deve classificar os pacotes que não precisam ter esse tipo de preocupação na marcação, ou seja , podem ter o maior atraso na rede.

As classes Conversational e Streaming devem ser usadas para carregar fluxos de tráfego que tenham o perfil de tempo real. O principal divisor entre estas duas classes é o quanto é sensível o atraso do tráfego especificado.

As classes Interactive e Background são usadas principalmente nas aplicações tradicionais de Internet, tais como WWW, e-mail, protocolos de gerência (Telnet, ssh, snmp), ftp dentre outros.

Conforme [2] as quatro diferentes classes são baseadas segundo o atraso individual, a taxa de bit imposta, a taxa de erro de bit e os requerimentos de prioridade atribuídos aos tráfegos.

3.4.2.1.Classe Conversational

O principal perfil de tráfego usuário deste esquema é a conversa de telefonia, ou seja, o tráfego de voz propriamente dito. Porém com o desenvolvimento da Internet e da multimídia, novas aplicações estão se enquadrando nessa classe. Como exemplo, pode ser citado o serviço de voz sobre IP e as ferramentas de vídeo conferência.

O perfil de conversa em tempo real é caracterizado pelo tempo de transferência dos pacotes que deve ser baixo, devido ao perfil desse tráfego. O máximo atraso na transferência é dado pela percepção de vídeo e áudio. Por essa razão, o limite para um atraso de transferência aceitável deve ser muito pequeno.

3.4.2.2.Classe Streaming

Quando o usuário está observando o vídeo em tempo real e ouvindo áudio em tempo real está classe pode ser aplicada. Para esta classe, uma porção da variação do atraso é tolerável devido ao buffer do nível de aplicação. Esta variação do atraso fim a fim deve ser limitada para preservar a variação de tempo entre os pontos de origem e destino da informação. Os fluxos de tráfego que são classificados como prioritários devido a sua importância crítica, são adequados a este perfil, por isso, no experimento, são configurados os protocolos de NGN.

3.4.2.3.Classe Interactive

Esta classe é aplicada nos serviços que requerem throughput seguro. Com o objetivo de assegurar melhores tempos de resposta para os pacotes que são classificados neste perfil de classe, uma maior prioridade é implementada se comparada com a classe background. Alguns exemplos incluem e-commerce, tráfego web interativo.

3.4.2.4.Classe Background

Esta classe comporta os serviços que podem ser configurados com características de melhor esforço, ou seja, podem ser citados tráfegos como download de e-mails e de arquivos, tráfegos com aplicativos destinados à gerência de rede, dentre outros. Esta classe tem a menor prioridade para escoamento de tráfego quando comparada as demais classes supracitadas.

3.4.3. Atributos de qualidade de serviço da rede de transporte UMTS

Para a definição dos atributos são suportados os serviços de transporte unidirecional e bidirecional. Para os serviços de transporte bidirecional deve ser possível ajustar separadamente para downlink e uplink os atributos de máxima taxa de bit, taxa de bit garantido e atraso na transferência.

Segue lista de atributos, especificada em [6], que detalha os parâmetros que devem ser utilizados para a classificação e medição de desempenho de uma rede UMTS com qualidade de serviço:

- Classe de Tráfego (conversational, streaming, interactive e background) é a classificação dada ao pacote de acordo com o tipo de aplicação utilizada;
- Taxa de bit máxima (kbps) é o número máximo de bits entregues pelo UMTS e para UMTS em um período de tempo, dividido pela duração do período;
- Taxa de bit garantida (kbps) é o número garantido de bits entregues pelo UMTS em um período de tempo, dividido pela duração do período;
- Ordem de entrega (sim/não) indica se o transportador UMTS deve providenciar a entrega da SDU na seqüência correta ou não;
- Tamanho máximo do SDU (octeto) que define o tamanho máximo permitido na rede;
- Informação sobre o formato do SDU (bits) que lista os possíveis tamanhos exatos que podem ser utilizados na rede;
- Taxa de erro do SDU é o parâmetro que indica a fração de SDU´s perdidos ou detectados como errados;
- Taxa de erro de bit residual indica a taxa de erro de bit não detectada nos SDU´s entregues ao destino. Se nenhuma detecção de erro é solicitada, a taxa de erro de bit residual indica a taxa de erro de bit nos SDU´s entregues;
- A entrega de SDU´s errados (sim/não/-) indica se os SDU´s detectados como errados devem ser entregues ou devem ser descartados;
- Atraso na entrega (ms) indica o atraso máximo para 95% das distribuições de atraso de todos os SDU´s entregues durante o período de vida de um serviço responsável pelo transporte dos pacotes. Atraso de um SDU é definido como o tempo que uma solicitação leva para transferir um SDU da origem ao destino;
- Prioridade do tráfego negociado especifica a importância relativa de todos os SDU´s que pertencem aos transportadores UMTS, quando comparados com os SDU´s de outros transportadores.

Alocação e retenção devido a prioridade especificam a importância relativa de vários transportadores UMTS quando comparados os parâmetros para alocação e retenção de um transportador UMTS. O atributo que prioriza a alocação/retenção é um atributo subscrito que não é negociado do terminal móvel

Na tabela 4 seguem sumarizados por classe de tráfego os atributos de transporte do padrão UMTS que devem ser considerados na classificação dos perfis de tráfego:

Classe de tráfego	Conversational	Streaming	Interactive	Background
Taxa de bit máxima	X	X	X	X
Ordem de entrega	X	X	X	X
Tamanho máximo do SDU	X	X	X	X
Informação sobre o formato do SDU	X	X		
Taxa de erro do SDU	X	X	X	X
Taxa de erro de bit residual	X	X	X	X
Entrega de SDU´s errados	X	X	X	X
Atraso na entrega	X	X		
Taxa de bit garantida	X	X		
Prioridade do tráfego negociado	X	X	X	
Alocação e retenção devido a prioridade	X	X	X	X

a 4 - Atributos de transporte do padrão UMTS por classe

A tabela 4 é bem útil ao longo do trabalho, pois é uma forma sumarizada de representar as características para a escolha adequada dos perfis de tráfego emulados no experimento. A partir desta tabela, pode-se recomendar que a classe Streaming seja a mais indicada para o escoamento dos perfis de tráfego dos protocolos NGN, pois os campos “atraso na entrega” e “taxa de bit garantida”, são extremamente importantes e representam o diferencial adequado para a escolha entre o perfil desta classe e o da classe Interactive.

3.4.4.Valores padronizados para os atributos das classes

Para os serviços de transporte, tanto da rede de acesso rádio, quanto da rede de dados propriamente dita, o UMTS lista alguns valores ou ranges de valores que especificam os atributos por classe. Assim como a tabela de atributos 3.4, este

trabalho dá um enfoque maior nos valores especificados para a rede de dados, sem levar em consideração a rede de acesso rádio, mesmo comparando e observando que a diferença entre as tabelas especificadas em [6] é muito pequena.

A tabela 5 [3] lista os valores especificados. O valor dos ranges reflete a capacidade da rede UMTS:

Características/Classes	**Conversational**	**Streaming**	**Interactive**	**Background**
Taxa máxima de bit (kbps)	< 2048	< 2048	< 2048 - cabeçalho	< 2048 - cabeçalho
Tamanho máximo do pacote (bytes)	= 1500 ou 1502	= 1500 ou 1502	= 1500 ou 1502	= 1500 ou 1502
Taxa de erro do pacote	10^{-2}, $7*10^{-3}$, 10^{-3}, 10^{-4}, 10^{-5}	10^{-1}, 10^{-2}, $7*10^{-3}$, 10^{-3}, 10^{-4}, 10^{-5}	10^{-3}, 10^{-4}, 10^{-6}	10^{-3}, 10^{-4}, 10^{-6}
Atraso de transferência (mseg.)	Valor máximo = 100	Valor máximo = 250	-	-
Taxa de bit garantida (kbps)	< 2048	< 2048	-	-

abela 5 - Valores dos atributos da rede de serviço de transporte UMTS

Na tabela 5 há uma observação para o tamanho máximo do pacote especificado em bytes. Caso o PDP seja do tipo PPP, o tamanho máximo do SDU é de 1502 bytes. Nos demais casos são 1500 bytes. A partir desta tabela, pode-se confirmar a utilização mais apropriada dos protocolos de NGN na classe Streaming, pois a outra classe Interactive que poderia adequar o perfil de tráfego destes protocolos de convergência, não possui um limite de atraso de transferência, só se preocupando com a prioridade dos tráfegos alocados a este perfil.