Qualidade de Serviço com Ganho de Multiplexação Estatística

Sibelius Lellis Vieira
Universidade Católica de Goiás, Departamento de Computação
Goiânia, GO, BRASIL, 74605-010
sibelius@ucg.br

Abstract

The Internet is still largely based on the best effort service, which does not provide enough support for multimedia applications with strict timing requirements, such as Voice over IP and videoconferencing. The network should provide to these applications bounds in the maximum delay and packet rate loss. In order to determine the required network characteristics to provide these services, we use a formal modeling of traffic and bandwidth service based on the network calculus. The network calculus provides a framework to identify the necessary resources to a given application, based on their traffic profile. The backlog and delay bounds can be evaluated given a representation of the service offered by the node and by the network as a whole. In general, the statistical analysis of the quality of service can provide a gain in the resource utilization when compared to the deterministic analysis. We try to identify and compare the deterministic and statistical calculus in this sense.

Keywords: Quality of Service, Network Calculus, Performance Evaluation.

Resumo

O serviço de melhor esforço, disnível em larga escala na Internet, não é suficiente para garantir um suporte adequado para aplicações com requisitos temporais rígidos, tais como Voz sobre IP and videoconferência. Este suporte tem como meta fornecer a estas aplicações garantias de atraso máximo e taxa máxima de perda de pacotes e deve ser estabelecido em termos de gerência de banda, controle de buffers e regulagem de tráfego. Neste trabalho, empregamos uma modelagem formal de controle de tráfego e serviço de banda baseada em cálculo de rede que tem como propriedade a identificação, a partir das características do tráfego, os recursos necessários para assegurar a qualidade das aplicações. O tamanho das filas e limites de atraso podem ser estimados a partir de uma representação do serviço oferecida pelo rede. Em geral, a especificação de qualidade em termos estatísticos pode fornecer um ganho na utilização dos recursos da rede em relação à qualidade determinística. Procuramos identificar e relacionar as vantagens e desvantagens do uso do cálculo de rede estatístico em relação ao cálculo determinístico.

Palavras chaves: Qualidade de Serviço, Cálculo de Rede, Análise de Desempenho de Redes.

1 INTRODUÇÃO

A transmissão de dados entre os elementos que compõe a Internet se baseia no serviço de melhor esforço, que é atualmente o único tipo de serviço oferecido em larga escala nesta rede. O objetivo deste serviço, como o nome sugere, é processar a comunicação dos pacotes de forma otimizada, através do uso racional dos recursos. Contudo, tal serviço não dá garantias específicas de desempenho ou de confiabilidade para as aplicações. Com o crescimento da Internet, este serviço se mostrou fundamental, pois permitiu que os elementos de rede tais como os roteadores se mantivessem simples, sendo amplamente disponibilizado. Além disto, tem propriedades importantes, tais como a escalabilidade e robustez, sendo adequado para grande parte das aplicações tradicionais. Entretanto, a Internet vem sendo utilizada cada vez mais por um contingente de aplicações com requisitos temporais e de consumo de taxa mais rígidos. As novas aplicações, tais como telefonia IP, video-conferência, *streams* de vídeo e áudio e outros necessitam de serviços mais sofisticados do que o serviço de melhor esforço pode oferecer. A telefonia IP, por exemplo, é uma das muitas aplicações que têm evidenciado a necessidade de diversas mudanças na infraestrutura da rede, pois experiências associadas a conversações telefonicas demonstram que um parâmetro tal como o atraso fim-a-fim deve ser limitado a 200ms. O tráfego destas aplicações possui requisitos drasticamente diferentes dos associados à rede de pacotes tradicional, tais como a limitação de atraso e a garantia de largura de banda mínima. Portanto,

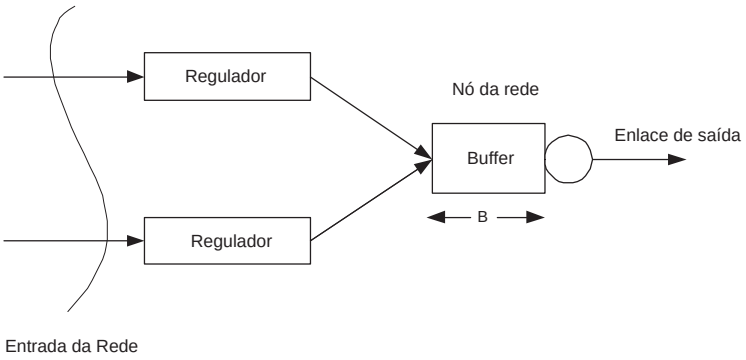

Figure 1: Um elemento de rede ou nó pode ser visto como um conjunto de portas de saída, cada qual com um buffer de tamanho finito. Em geral, múltiplos fluxos de entrada podem ser direcionados e multiplexados em uma única porta de saída. Fluxos entrantes na rede são regulados no primeiro nó através de um envelope de entrada A^*.

faz-se necessário analisar, modelar, desenvolver e implementar novos protocolos e serviços que permitam o suporte de aplicações com características de comunicação multimídia [11].

As novas aplicações requisitam uma rede tenha a capacidade de dar o suporte adequado a uma grande variedade de tráfego com requisitos de desempenho diferenciados e críticos. Esta infraestrutura de comunicação deve conter mecanismos de controle e gerenciamento de tráfego, assim como um controle do comportamento do tráfego submetido por diversas aplicações [12]. A análise de problemas de controle de tráfego em redes, escalonamento de pacotes e encaminhamento é extremamente importante para a garantia da qualidade de serviço (QoS), exemplificada anteriormente como o atraso fim-a-fim. De forma mais geral, esta QoS é especificada em termos de atraso máximo e perda de pacotes. O controle de tráfego assegura que os recursos da rede sejam divididos de forma eficiente, de tal forma a maximizar o uso destes recursos enquanto assegura o cumprimento da QoS para os usuários. Para assegurar o escalonamento de recursos da rede de forma a atender aos requisitos de QoS, é necessário também que cada fonte caracterize seu próprio tráfego. O gerenciamento de tráfego na rede é importante para assegurar que a qualidade de serviço negociada para as sessões seja mantida.

Neste contexto, os fluxos devem estabelecer contratos com a rede de maneira a limitar, em algum sentido, a quantidade de tráfego que os mesmos irão injetar na rede em determinado intervalo de tempo. O estabelecimento e manutenção destes contratos são pontos importantes para a garantia de fornecimento de serviço por parte da rede. Reguladores do tipo balde furado, por exemplo, são mecanismos convenientes para a definição e garantia de cumprimento dos contratos de tráfego [3]. As fontes que apresentam conformidade com a especificação de tráfego são ditas fontes reguladas por envelopes de tráfego.

É fundamental se determinar a quantidade de recursos necessária para garantir a QoS, expressada em termos de atrasos máximos ou perdas de pacotes. O atualmente denominado cálculo de rede oferece uma série de métodos para que esta quantidade seja estabelecida [4, 5]. Através deste formalismo, é possível calcular, usando como entrada as necessidades das aplicações e sua caracterização de tráfego, os recursos necessários para a garantia de limites nos valores de atraso e de perda de pacotes. O cálculo de rede determinístico provê uma abordagem formal elegante para análise de pior caso, que pode ser usada para derivar limites máximos em atrasos e tamanho da fila para uma grande variedade de mecanismos de escalonamento [2]. A alocação de recursos é feita através do conceito de curva de serviço, que aloca para cada conexão uma quantidade de recursos em pior caso. Uma vantagem do cálculo de rede determinístico é a sua capacidade de determinar os atrasos máximos e tamanho de filas em vários nós da rede.

Pode se argumentar, contudo, que a garantia absoluta de que nenhum pacote será perdido ou sofrerá um atraso maior do que o estabelecido é por demais conservadora para aplicações de mídia contínua, as quais podem tipicamente tolerar uma taxa pequena de perdas. De fato, os usuários não percebem qualquer degradação de qualidade quando a perda de pacotes é pequena e pouco frequente, especialmente se o receptor emprega técnicas de cancelamento de erros. Além do mais, esquemas que garantem a entrega total dos pacotes tipicamente tem uma baixa capacidade de transmitir tráfego com características de rajada. Posto de outra forma, as abordagens determinísticas que garantem que não há perda requerem uma superestimação dos recursos de rede necessários para garantir a QoS [1]. Portanto, a utilização de modelos determinísticos leva a um uso ineficiente dos recursos de rede.

Estas características do cálculo determinístico levantam questões importantes. A primeira é a possibilidade de desenvolver uma abordagem que provê garantias estatísticas de QoS em uma rede, ou seja, um limite na fração de tráfego que exceda um parâmetro de QoS tal como o atraso máximo [7]. A garantia estatística

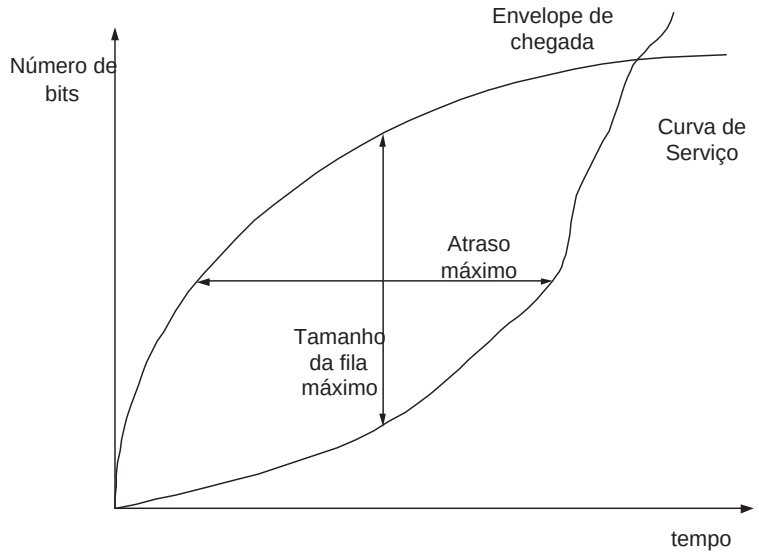

Figure 2: Illustração da relação entre envelope, curva de serviço, distância horizontal e distância vertical. A diferença horizontal indica o atraso máximo que pode ser associado ao fluxo e a diferença vertical, o tamanho máximo de ocupação de buffers.

de QoS em um contexto de rede é notoriamente difícil, pois os fluxos de tráfego normalmente perdem suas características estatísticas originais no processo de compartilhamento de filas de saída. A segunda questão diz respeito ao ganho advindo desta multiplexação estatística em relação à abordagem determinística [6].

Uma característica das redes de pacotes é a sua habilidade de explorar a multiplexação estatística de fontes de tráfego e portanto, alcançar um alto grau de utilização dos recursos disponíveis. Este ganho em termos de multiplexação estatística pode ser explorado tirando vantagem das propriedades estatísticas, da independência das fontes e dos diferentes requisitos de QoS das mesmas [10]. Os modelos estatísticos, baseados em um cálculo estatístico de rede, podem estabelecer garantias da forma probabilística, tais como a probabilidade de um valor de atraso exceder um limite pré-determinado é de $10^{-\epsilon}$, onde ϵ varia entre 3 e 9. Desta forma, uma pequena fração de tráfego pode violar as especificações de QoS. Estas abordagens estatísticas podem permitir um ganho substancial no uso dos recursos, evidenciado em uma relação que expressa que os recursos necessários para garantir estatisticamente N fluxos são bem menores do que os recursos para garantir deterministicamente estes mesmos fluxos.

Este trabalho procura investigar um importante aspecto do cálculo de rede determinístico e estatístico, qual seja o nível de utilização de recursos indicado pelo número de fluxos admissíveis em um nó. O esquema de gerenciamento de tráfego tem os seguintes componentes: cada fluxo deve ser regulado na entrada da rede por uma elemento que implemente um envelope de tráfego, todos os nós empregam um mecanismo de multiplexação de tráfego com memória e o controle de admissão é baseado na suposição de que o tráfego de uma fonte é independente do tráfego de outra na entrada da rede.

Na seção II, descrevemos as bases do cálculo determinístico e apresentamos os principais resultados que nos permitem obter os parâmetros de interesse dos fluxos, tais como limites de atraso, tamanho de filas e funções de limitação de tráfego. Na seção III, apresentamos a extensão do cálculo determinístico para obter garantias estatísticas e discutimos os resultados do cálculo estatístico e seus limites. Na seção IV, apresentamos uma análise de comparação de desempenho entre o cálculo determinístico e o cálculo estatístico para situações de fluxos fim-a-fim. Finalmente, na seção V apresentamos as conclusões.

2 CÁLCULO DETERMINÍSTICO

Nesta seção, apresentamos algumas definições e resultados que serão considerados e descritos coletivamente como "cálculo de rede". O princípio geral do cálculo de rede é fornecer uma abordagem formal para a obtenção de parâmetros importantes para caracterizar o desempenho de uma rede, tais como o atraso máximo ou a taxa de perda de pacotes. Através de uma coleção de resultados oriundos de uma álgebra conhecida como *min-plus*, podemos caracterizar sistemas de filas utilizados em redes de comunicação. Como exemplo, podemos utilizá-lo para entender os cálculos de atraso usados pelos servios garantidos propostos pelo IETF, como modelo comum para escalonadores, para identificar a influência de reguladores sobre o atraso, para calcular a banda efetiva necessária de um fluxo com garantias de atraso, entre outras aplicações.

Vamos considerar como exemplo funções não-decrescentes no tempo. A função $f(\ )$ é não-decrescente quando $f(s) \leq f(t)$ sempre que $s \leq t$. Para duas funções não-decrescentes f_1 e f_2, denominamos a operação de convolução *min-plus* entre f_1 e f_2 como $f_1 \otimes f_2$ correspondente à operação de convolução padrão e definida da seguinte forma:

$$f_1 \otimes f_2(t) = inf_{0 \leq u \leq t}\{f_1(u) + f_2(t-u)\} \quad \forall t \geq 0 \tag{1}$$

Seja um fluxo de dados descrito por sua função de chegada $A(t)$, que é igual ao número de bits do fluxo que chega em um nó da rede em um intervalo de tempo $[0, t]$. Portanto, a função $A(t)$ representa uma quantidade aleatória, cuja distribuição não é conhecida. Como o tráfego é cumulativo, podemos considerar $A(t)$ como função não-decrescente. Dada uma função não-decrescente A^*, dizemos que o fluxo é restrito por A^* se e somente se para todo $s \leq t$, temos que $A(t) - A(s) \leq A^*(t-s)$. Isto é equivalente a afirmar que $A(t) \leq A \otimes A^*(t)$ para qualquer $t \geq 0$.

A função A^* é chamada de curva de chegada ou envelope de fluxo. Por exemplo, um fluxo controlado por um balde furado tem uma curva de chegada na forma $A^*(t) = \sigma + \rho t$. A função A^* pode ser qualquer função não-negativa e não-decrescente, mas para definir uma restrição consistente deve também ser sub-aditiva, o que implica em $A^*(s+t) \leq A^*(s) + A^*(t)$ para qualquer $s, t \geq 0$. Na Figura 1, ilustramos uma situação em que cada fonte tem um envelope definido por um regulador. A utilidade básica do envelope do fluxo é garantir que os fluxos que entram na rede tem um tráfego máximo definido e controlado pelos reguladores que implementam este envelope.

Consideremos $\mathcal{S}$ um sistema visto na forma de uma caixa-preta, ou seja, este sistema é caracterizado apenas pela sua chegada e pela sua saída. Este sistema também pode ser representado graficamente pela ilustração da Figura 1. O sistema $\mathcal{S}$ recebe dados de entrada e os retransmite na saída após um atraso variável. Seja $A(t)$ a função de entrada, representando o tráfego de chegada e $D(t)$ a função de saída, esta última indicando o número de bits que deixam o sistema no intervalo de tempo $[0, t]$. A fila no tempo t tem um tamanho igual a $A(t) - D(t)$ e representa o número de bits dentro do sistema, supondo o sistema vazio no tempo 0. De forma similar, o atraso virtual no tempo t é dado por:

$$d(t) = inf\{T : T \geq 0 \ \ e \ \ A(t) \leq D(t+T)\} \quad \forall t \geq 0 \tag{2}$$

Este é o atraso que seria associado a um bit chegando no tempo t se todos os bits que foram recebidos antes deste forem servidos na ordem de chegada. Se a função de saída for contínua, então $D(t + d(t)) = A(t)$.

Os resultados do cálculo de rede fornecem regras computacionais para o limite de atrasos virtuais e tamanho de fila para sistemas arbitrários que representam as redes de dados. Dizemos que o sistema $\mathcal{S}$ oferece uma curva de serviço S se e somente se:

$$D(t) \geq A \otimes S(t) \quad \forall t \geq 0 \tag{3}$$

Na prática, isto é equivalente a afirmar que para todo $t \geq 0$, existe algum $t_0 \geq 0$, com $t_0 \leq t$ tal que $A(t) - D(t_0) \geq S(t - t_0)$. Por exemplo, em um sistema de escalonamento no qual uma parte da banda é reservada para um fluxo com taxa garantida r baseado em modelo de fluidos, o sistema oferece ao fluxo uma curva de serviço $S(t) = rt$. O conceito de curva de serviço pode ser relacionado ao de escalonamento, pois para cada tipo de escalonamento, podemos derivar uma curva de serviço representando este escalonamento. Porém, as curvas de serviço são mais gerais, e em alguns casos não representam disciplinas de escalonamento factíveis. Como exemplo de uma disciplina não-factível, temos a curva de serviço $S(t) = 0$ para $t < d$ e $S(t) = \infty$ para $t \geq d$.

Uma série de resultados foram derivados para o atraso e o tamanho de fila em um nó da rede, bem como para um caminho constituído de vários nós, dados a função de entrada e a curva de serviço [2]. O primeiro supõe a existência de um fluxo, com envelope A^* que atravessa um sistema que oferece uma curva de serviço S. O tamanho da fila $A(t) - D(t)$ para todo t é limitado por:

$$A(t) - D(t) \leq sup_{s \geq 0}\{A^*(s) - S(s)\} \quad \forall t \geq 0 \tag{4}$$

Se o sistema é um buffer simples em um nó, este tamanho pode ser interpretado como o tamanho instantâneo da fila. Por outro lado, se o sistema é mais complexo, tal como um conjunto de buffers em nós de um caminho, então este tamanho é o número de bits "em trânsito", supondo que podemos observar tanto a entrada quanto a saída simultaneamente. Este resultado indica que o tamanho da fila é limitado pela diferença vertical entre o envelope e a curva de serviço, conforme ilustrado na Figura 2.

Para descrevermos o próximo resultado, introduzimos uma nova operação, que pode ser interpretada como uma operação inversa à da convolução. Dizemos que a função $f_1 \oslash f_2(t) = sup_{u \geq 0}\{f(t+u) - f(u)\}$ é a deconvolução das funções f_1 e f_2. Supondo a existência de um fluxo com características dadas anteriormente,

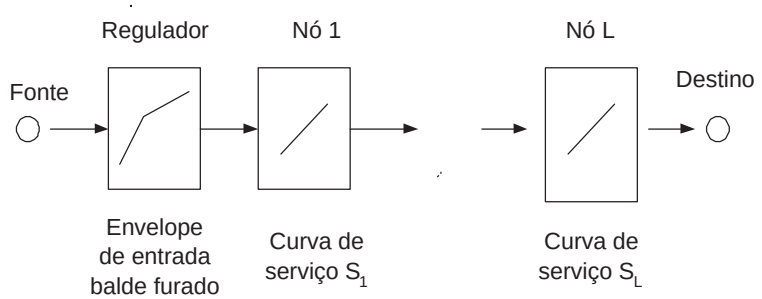

Figure 3: Um fluxo criado em uma fonte atravessa uma rede através de um caminho com N nós, cada qual oferecendo uma curva de serviço. Na entrada da rede, um regulador garante a conformidade com o envelope de entrada.

o fluxo de saída é restringido por um envelope dado por $A^* \oslash S(t)$. Portanto, dado um envelope de entrada e curvas de serviço, podemos obter envelopes para o tráfego em qualquer nó da rede.

Para duas funções não-decrescentes α e β, definimos a diferença horizontal entre as duas curvas como sendo $h(\alpha, \beta) = sup_{s \geq 0}(inf\{T : T \geq 0 \quad e \quad \alpha(s) \leq \beta(s+T)\})$. A definição pode ser entendida intuitivamente da seguinte forma: se α e β são contínuas e estritamente crescentes, então para todo t existe no máximo um número $d(t)$ tal que $\alpha(t) = \beta(t + d(t))$. Neste caso, $h(\alpha, \beta) = sup_t d(t)$. Em outras palavras, $h(\alpha, \beta)$ não é nada mais do que a fórmula usada para a computação do atraso máximo, que pode ser reescrita como $d(t) \leq h(\alpha, \beta)$ para qualquer t. O valor do atraso máximo pode ser representado como a diferença horizontal entre o envelope e a curva de serviço, conforme ilustrado na Figura 2. Desta forma, podemos computar o atraso máximo d como dado abaixo:

$$d = sup_{t \geq 0} inf\{T : T \geq 0 \quad e \quad A(t - T) \leq S(t)\} \tag{5}$$

Estes resultados podem ser aplicados em um caminho de rede envolvendo vários nós. Suponha a existência de dois sistemas $\mathcal{S}$ e $\mathcal{F}$ com curvas de serviço S e F, respectivamente. A concatenação destes dois sitemas oferece uma curva de serviço $T = S \otimes F$. Por exemplo, na Figura 3 ilustramos um fluxo com um regulador de entrada implementando um envelope de tráfego para este fluxo e a seguir, uma série de nós que fazem parte do caminho deste fluxo na rede. A curva de serviço de rede para este fluxo pode ser obtida como uma convolução das várias curvas de serviço individuais para o fluxo em cada um dos nós. Outras características do fluxo podem ser obtidas através do envelope de entrada e desta curva de serviço de rede, tais como o atraso fim-a-fim sofrido pelos pacotes deste fluxo.

Um regulador, com curva de regulagem γ é um dispositivo que força uma saída na forma de um envelope com função γ. O regulador atrasa o bit em um buffer sempre que o envio deste bit possa violar as restrições do tráfego de saída. Um balde furado é um regulador cuja função γ tem a forma $\gamma = \sigma + \rho t$, para $t > 0$ e $\gamma = 0$ para $t = 0$. O regulador tem duas propriedades importantes: em primeiro lugar, o regulador não aumenta o limite do atraso. Suponha que um fluxo com envelope de entrada A^* é direcionado para o sistema formado por $\mathcal{S}$ e $\mathcal{F}$ em sequencia. Suponha que um regulador $\gamma \leq A^*$ é inserido entre $\mathcal{S}$ e $\mathcal{F}$. O limite de retardo dado anteriormente é também válido para o sistema com o regulador. O regulador também conserva as propriedades originais do fluxo. Considere um fluxo, com envelope de entrada A^* direcionado para um regulador com curva γ. A saída do regulador ainda é restringida pela curva de entrada original A^*.

3 CÁLCULO ESTATÍSTICO

O cálculo estatístico extende o cálculo determinístico apresentado dentro de um contexto probabilístico no intuito de explorar o ganho de multiplexação estatística. Este ganho de multiplexação estatística pode ser observado concretamente através de mecanismos de controle de admissão, que indicam o número de fluxos aceitáveis em um sistema ou em algum nó. Um dos primeiros métodos que foram estudados extensivamente, o cálculo de banda efetiva, foi motivado por funções que apareceram na teoria de desvios grandes. Expressões para a banda efetiva de um fluxo foram derivados, associando de forma estatística a banda que um fluxo iria necessitar para garantia de sua QoS com uma probabilidade dada. Para esta computação, o utiliza-se uma expressão baseada na função de chegada $A(t)$. Neste trabaho, não investigaremos as relações entre o cálculo estatístico e o cálculo de banda efetiva. Contudo, estas relações tem sido analisadas, e os resultados indicam que o cálculo estatístico pode fornecer uma abordagem mais geral com garantias de desempenho similares ao cálculo de banda efetiva [8].

No formalismo de cálculo estatístico, consideramos as chegadas e partidas em um intervalo de tempo $[0, t]$ vistas como processos estocásticos que satisfazem certas propriedades, e as funções de chegada $A(t)$ e saída $D(t)$ representam estes processos. Para assegurar a validade do cálculo estatístico, precisamos estabelecer certas condições em relação ao processo de chegada:

(A1) Aditividade: Para qualquer $t_1 < t_2 < t_3$, temos que $A(t_1, t_3) = A(t_1, t_2) + A(t_2, t_3)$.

(A2) Sub-aditividade: O envelope $A^*(t)$ é subaditivo.

(A3) Estacionaridade: A função $A(t)$ é estacionária, ou seja, $Pr[A(t, t+\tau) \leq x] = Pr[A(t_1, t_1+\tau) \leq x]$, para qualquer $t, t_1 \geq 0$.

(A4) Independência: As chegadas A_i e A_j para dois fluxos diferentes são estocasticamente independentes.

Estas suposições são feitas apenas na entrada da rede, quando o tráfego está chegando no primeiro nó de seu caminho. Nenhuma suposição adicional é feita com relação ao tráfego dentro da rede. A subaditividade garante que o limite $lim_{\tau\to\infty} A^*(\tau)/\tau$ existe e é denotado por ρ. Desta forma, existe um limite superior para a taxa de chegada para $A(t)$ a longo prazo. A estacionaridade tem como características interessantes o fato de que valores esperados podem ser computados como médias a longo prazo. A independência dos fluxos permite explorar os ganhos com a multiplexação estatística.

3.1 Extensões ao cálculo determinístico

Vamos definir as funções do cálculo estatístico que generalizam os conceitos de envelope e curva de serviço apresentados para o cálculo determinístico. Para os fluxos de entrada, definimos o envelope efetivo $G^\epsilon(\tau)$ para um processo de chegada $A(t)$ representado o tráfego entrante em $[0, t]$ como sendo:

$$Pr[A(t+\tau) - A(t) \leq G^\epsilon(\tau)] \geq 1 - \epsilon \quad \forall t, \tau \geq 0 \tag{6}$$

De forma simplificada, o envelope efetivo representa um limite estacionário para os processos de chegada. De acordo com esta definição, é possível que o tráfego de chegada exceda a quantidade permitida pelo envelope efetivo com uma pequena probabilidade. Conforme será descrito posteriormente, é possível obter limites para a função de envelope efetivo a partir da função que representa o envelope de chegada. Estes limites são apertados, representando de forma acurada a função do envelope efetivo.

De forma análoga, podemos definir a curva de serviço efetiva como sendo uma medida da probabilidade do serviço disponível para o fluxo. A curva de serviço efetiva garante que uma grande quantidade de tráfego de entrada estará disponível na saída, através da função $D(t)$, como resultado do serviço efetivo. Dado um processo de chegada $A(t)$, a curva de serviço efetiva S^ϵ é uma função não-negativa que satisfaz para todo $t \geq 0$

$$Pr[D(t) \geq A \otimes S^\epsilon(t)] \geq 1 - \epsilon \tag{7}$$

A definição da curva de serviço efetiva impõe uma dificuldade para estabelecer uma curva de serviço global. Isto pode ser entendido se observarmos que, como a definição da curva de serviço efetiva não assegura que o valor de $D(t)$ não é menor do que o $inf_{s\geq t}\{A(t-s) + S^\epsilon(s)\}$, não é possível garantir uma expressão simples para a concatenação das curvas de serviço. Uma forma de diminuir a dificuldade no cálculo desta concatenação é adicionar uma suposição relativa ao escopo do $infimum$. Esta suposição se baseia na existência de um valor T tal que

$$Pr[D(t) \geq inf_{s\leq T}\{A(t-s) + S^\epsilon(s)\}] \geq 1 - \epsilon \quad \forall t \geq 0 \tag{8}$$

Portanto, T limita o escopo da convolução. Em geral, o valor de T está relacionado a um intervalo de tempo no qual a convolução faz sentido. Por exemplo, em uma porta de saída de um nó da rede, os intervalos de tempo importantes são aqueles em que existe tráfego requisitando serviço nesta porta, ou seja, os intervalos nos quais o buffer não está vazio. Os períodos nos quais o buffer está vazio tem solução trivial, pois não existe demanda de tráfego. O tamanho destes intervalos ocupados pode ser relacionado ao valor de T. Por exemplo, para curvas de serviço na forma $S(t) = kt$, o cálculo determinístico fornece um limite superior para o valor de T na dado como [8]:

$$T = inf\{\tau > 0 : A^*(\tau) \leq S(\tau)\} \tag{9}$$

O valor de T dado pela eq. (9) fornece um limite superior conservador, no sentido de que pode ser usado para o cálculo estatístico, que fornece limites mais otimistas [8].

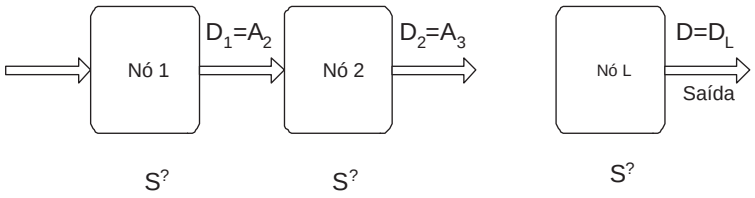

Figure 4: Nós do caminho de um fluxo com curva de serviço S^{ϵ}. A saída de um nó D_i corresponden a entrada do próximo nó A_{i+1}.

Os resultados apresentados no cálculo determinístico podem ser aplicados para o cálculo estatístico, de acordo com as proposição abaixo:

Suponha que G^{ϵ_g} é um envelope efetivo para as chegadas A em um nó e que S^{ϵ_s} seja uma curva de serviço efetiva satisfazendo a eq. (8) para algum $T < \infty$. Defina ϵ como $\epsilon = \epsilon_s + T\epsilon_g$. Temos que:

1 Envelope de saída: A função $G^{\epsilon_g} \oslash S^{\epsilon_s}$ é um envelope efetivo para o tráfego de saída do nó.

2 Limite de tamanho de fila: A função $G^{\epsilon_g} \oslash S\epsilon_s(0)$ é um limite probabilístico para o tamanho da fila, no sentido de que, para qualquer $t \geq 0$, temos $Pr[B(t) \leq G^{\epsilon_g} \oslash S^{\epsilon_s}(0)] \geq 1 - \epsilon$.

3 Se $d > 0$ satisfaz $sup_{\tau \leq T}\{G^{\epsilon_g}(\tau - d) - S^{\epsilon_s}(\tau)\} \leq 0$, então d é um limite de probabilístico do atraso, no sentido de que, para qualquer $t \geq 0$, temos $Pr[W(t) \leq d] \geq 1 - \epsilon$.

Podemos verificar que este teorema generaliza os resultados do cálculo determinístico, que ocorrem no limite de $\epsilon_g, \epsilon_s \to 0$.

O cálculo estatístico fornece também uma expressão probabilística para a convolução de curvas de serviço de vários elementos de rede dispostos em sequencia. Esta expressão fornece o serviço dado pela rede como um todo em um determinado caminho. Considere tal caminho de um fluxo como ilustrado na Figura 4 . Em cada nó, a chegada tem alocada uma curva de serviço, indicada como S^{i,ϵ_s} para o nó i. De forma similar à eq. (8), supomos que cada nó satisfaz a expressão $Pr[D^i(t) \geq inf_{\tau \leq T^i}\{A^i(t-\tau) + S^{i,\epsilon_s}(\tau)\}] \geq 1 - \epsilon_s$, para $T_i < \infty$, com $i = 1, \ldots, L$. Podemos obter uma curva de serviço efetiva da rede, $S^{tot,\epsilon}$ dada como:

$$S^{tot,\epsilon} = S^{1,\epsilon_s} \otimes S^{2,\epsilon_s} \otimes \ldots \otimes S^{L,\epsilon_s} \tag{10}$$

A violação de probabilidade é dada por $\epsilon = \epsilon_s \sum_{i=1}^{L}(1 + (i-1)T^i)$. Este valor de ϵ se degrada a medida que o fluxo atravessa os vários nós do caminho, tornando limitada a abordagem estatística baseada em envelope efetivo e curva de serviço efetiva.

3.2 Multiplexação Estatística

A partir da definição do envelope efetivo G^{ϵ} dada na eq. (6), podemos obter o valor do envelope efetivo baseado em funções de geração de momento das distribuições dos processos de chegada $A(t)$. Sabemos que a função de geração de momento da distribuição de $A(t)$ pode ser dada por $M(s,t) = E[e^{A(\tau,\tau+t)s}]$ [9]. A estacionaridade garante que as funções $M(s,t)$ não dependem de t. Através da utilização dos limite de Chernoff, que garante um limite superior para $P[Y \geq y]$, onde Y é uma variável aleatória e y é um valor possível para esta variável, através da expressão $P[Y \geq y] \leq e^{-sy}E[e^{A(\tau,\tau+t)s}]$ [1], temos que:

$$Pr[A \geq Nx] \leq [e^{-xs}(1 + \frac{\rho\tau}{A^*(\tau)}(e^{sA^*(\tau)} - 1)]^N \tag{11}$$

onde N é o número de fluxos agregados pelo mesmo envelope efetivo. Utilizando a definição de envelope efetivo, obtemos que o valor do envelope efetivo é dado por $G^{\epsilon}(t) = Nmin(x, A^*(\tau))$, onde x é o menor número satisfazendo

$$(\frac{\rho\tau}{x})^{(x/A^*(\tau))}(\frac{A^*(\tau) - \rho\tau}{A^*(\tau) - x})^{1-(x/A^*(\tau))} \leq \epsilon^{1/N} \tag{12}$$

Desta forma, podemos calcular o envelope efetivo de qualquer fluxos ou agregado de fluxos chegando em um nó. Para o próximo nó, podemos utilizar o resultado do Teorema 1 que fornece um envelope de saída efetivo dado por $G^{\epsilon_g} \oslash S^{\epsilon_s}$, onde S^{ϵ_s} é a curva de serviço efetiva do nó.

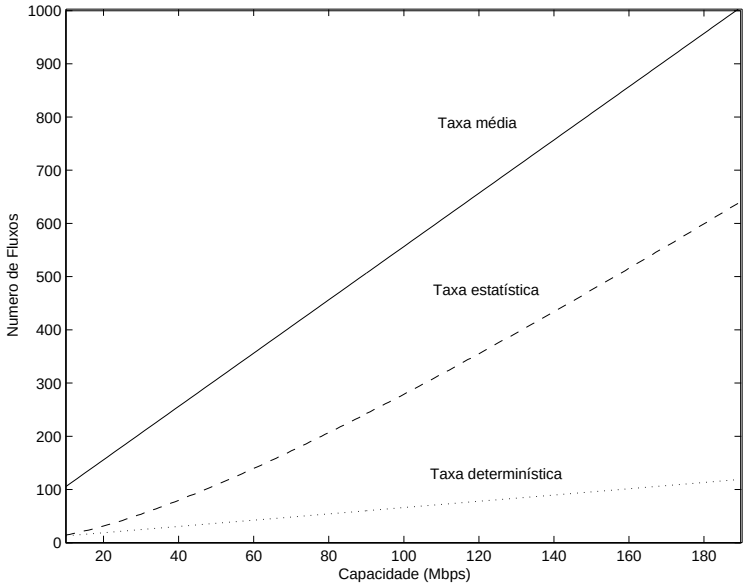

Figure 5: Número de fluxos em função da capacidade do enlace de saída para fluxos do tipo A.

4 AVALIAÇÃO NUMÉRICA

Nesta seção, apresentamos exemplos numéricos para a utilização das saídas dos nós, comparando os envelopes determinísticos com os envelopes efetivos, em uma rede com fluxos com caminhos envolvendo múltiplos nós. Vamos supor que os fluxos sejam individualmente controlados na entrada da rede por um envelope $A^*(\tau) = min\{P\tau, \sigma + \rho\tau\}$, onde P é a taxa de pico do fluxo, ρ é a taxa média e σ é a rajada de dados. Para os nossos experimentos, vamos supor a existência de dois tipos de fluxos: o primeiro, do tipo A, com grande variação entre a taxa média e taxa de pico e com alta quantidade de rajada, tendo como parâmetros $P_A = 2.0Mbps$, $\rho_A = 0.2Mbps$ e $\sigma_A = 0.1Mbit$; o segundo, com pequena variação entre a taxa média e a taxa de pico e com baixa quantidade de rajada, tendo como parâmetros $P_B = 0.2Mbps$, $\rho_B = 0.1Mbps$ e $\sigma_B = 0.01Mbit$.

No primeiro experimento, determinamos a utilização dos nós de saída em relação à capacidade do enlace de saída para o cálculo determinístico. Esta utilização é dada pelo número máximo de fluxos que podem ser acomodados em um enlace de saída sem que a QoS dos fluxos seja comprometida. Como parâmetros de QoS, vamos utilizar o atraso máximo d para os fluxos com um valor de $d = 10ms$. Vamos analisar o comportamento da rede separadamente para os fluxos do tipo A e B. A curva de serviço para o agregado de fluxos do tipo A ou B é dada por $S_R(t) = Ct$, onde C é a capacidade do enlace em Mbps e R representa o agregado. De acordo com os resultados do cálculo determinístico para a diferença horizontal vistos anteriormente, a curva de serviço para cada fluxo separadamente deve obeceder a relação indicada na eq. (5), $A^*(t-d) \leq S(t)$, onde $S^{tot}(t)$ é a curva de serviço de rede para cada fluxo, dada por $S^{tot}(t) = S^1 \otimes \ldots \otimes S^L(t)$, onde L é o número de nós no caminho dos fluxos. Supondo que cada fluxo tem uma curva de serviço por nó dada por $S^i(t) = ct$, onde c é a capacidade necessária para garantir o atraso máximo d, obtemos $S^{tot}(t) = ct$ como resultado das operações de convolução. Desta forma, podemos obter o valor da capacidade equivalente que cada fluxo requisita em cada nó para garantir o atraso máximo d.

Nas Figuras 5 e 6 ilustramos o número de fluxos que podem ser acomodados por enlace para o valor de atraso máximo requisitado. A capacidade máxima para fluxos do tipo A é de $c_A = 1.695$ e obecede a relação $P_A \geq c_a \geq \rho_A$. Para os fluxos do tipo B, $c_B = 0.181Mbps$ e também obedece uma relação equivalente. Na Figura 5, o número de fluxos é dado por uma relação linear na forma $N_A = C/c_A$ e na Figura 6, a mesma relação é dada por $N_B = C/c_B$. Ilustramos também o número de fluxos que poderiam ser acomodados se cada fluxo tivesse garantido apenas a capacidade média do envelope ρ.

No segundo experimento, determinamos a mesma utilização para a situação em que o atraso máximo é garantido estatisticamente. Para tal, calculamos o envelope efetivo dos fluxos do tipo A e B através da solução da eq. (12) para x de tal forma que $d > 0$ satisfaça a relação $sup_{\tau \leq T}\{G^{\epsilon_g}(\tau - d) - S^{\epsilon_s}(\tau)\} \leq 0$, onde $S^{\epsilon_s}(\tau) = Ct$ para o agregado de fluxos. Escolhemos para ϵ_s e ϵ_g o valor de -9. O valor de T_i é derivado da relação que fornece o um limite para o valor do período ocupado para um agregado de fluxos no caso determinístico, conforme apresentado através da eq. (8). Para a curva de serviço e envelope dados, $T_i = 82ms$ para os fluxos do tipo A e $T_i = 122ms$ para os fluxos do tipo B.

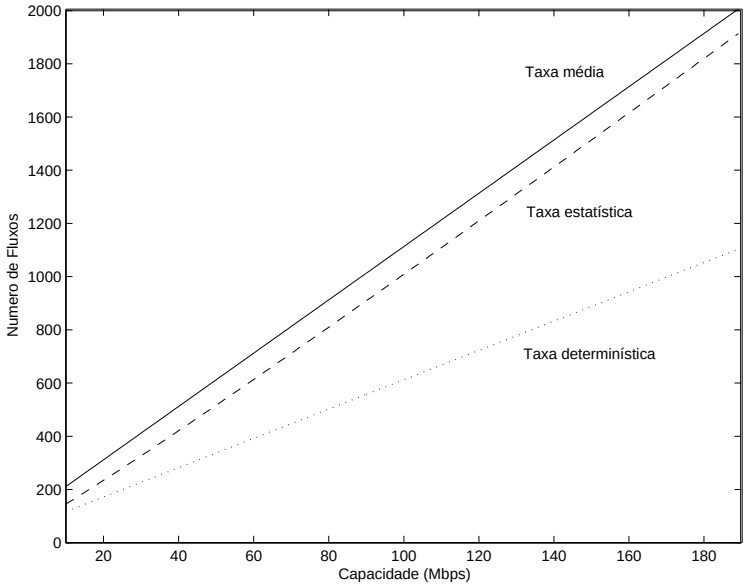

Figure 6: Número de fluxos em função da capacidade do enlace de saída para fluxos do tipo B

Table 1: Degradação do valor de ϵ para caminhos com 5 e 10 nós

Tipo de fluxo	L = 5	L = 10
A	8.10^{-6}	8.10^{-5}
B	10^{-5}	10^{-4}

Nas Figuras 5 e 6 ilustramos novamente o número de fluxos que podem ser acomodados no enlace de para uma garantia estatística do valor máximo de atraso. Podemos observar que o número de fluxos é significativamente maior do que o obtido para o caso determinístico. Na verdade, este número de aproxima do número de fluxos dado utilizando o valor da taxa média como sendo a capacidade reservada para cada fluxo. De forma intuitiva, a capacidade total dividida pelo número de fluxos indica a capacidade equivalente que cada fluxo deve reservar no nó de saída de forma a garantir seus requisitos de atraso máximo. Para a taxa média, esta capacidade reservada por fluxo garante apenas um valor finito para o atraso máximo que o tráfego do fluxo pode sofrer, mas não garante o atraso máximo requisitado. Para a taxa determinística, esta capacidade garante que o fluxo nunca sofre um atraso maior do que d e para a taxa estatística, esta capacidade garante que o fluxo pode sofrer um atraso maior do que d com uma probabilidade muito pequena. Podemos observar que para fluxos com menor rajada, a taxa estatística é mais próxima da taxa média do que para fluxos com maior rajada.

O problema da abordagem estatıustica surge em função da degradação do valor de ϵ. Embora os valores de ϵ_s e ϵ_g sejam muito pequenos, o valor de ϵ se degenera rapidamente quando os fluxos atravessam vários nós no caminho. Na tabela 1, ilustramos os valores de ϵ para $L = 5$ e $L = 10$.

5 CONCLUSÃO

Neste trabalho, procuramos comparar os resultados obtidos pelo cálculo de rede determinístico e estatístico, através do cálculo do número de fluxos que podem ser admitidos em um enlace com determinada capacidade em uma rede com múltiplos nós. Podemos observar a que o número de fluxos empregando a abordagem estatística é sensivelmente maior do que o número de fluxos com garantias determinísticas de atraso. Entretanto, a taxa de perda aumenta a medida que o fluxo atravessa os vários nós do caminho, a ponto do mesmo tornar-se inviável para aplicações com requisitos mais rigorosos em termos de perdas de pacote.

6 Agradecimentos

Agradecemos à PROPE/UCG e à CAPES pelo apoio recebido.

References

[1] R. Boorstyn, A. Burchard, J. Liebeherr and Oottamakorn, C, Statistical service assurances for traffic scheduling algorithms. *IEEE Journal on Selected Areas in Communications*, 18(12):2651–2664, 2000.

[2] J.-Y. Boudec and P. Thiran, *Network calculus*. Springer Verlag, Lecture Notes in Computer Science, LNCS 2050, 2001.

[3] C.S. Chang, *Performance guarantees in communications networks*, Springer, 2000.

[4] R. Cruz, A calculus for network delay, part I : Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–121, 1991.

[5] R. Cruz, A calculus for network delay, part II : Network analysis. *IEEE Transactions on Information Theory*, 37(1):121–141, 1991.

[6] E. Knightly and N. Shroff, Admission control for statistical QoS: Theory and practice. *IEEE Network*, 13(2):20-29, March 1999.

[7] C. Li, A. Burchard and J. Liebeherr, Statistical Network Calculus for Multiplexed Arrivals, Technical Report, University of Virginia, Computer Science Department 2003.

[8] J. Liebeherr, S. Patek and A. Burchard, Statistical Per-Flow Service Bounds in a Network with Aggregate Provisioning. *Proceedings of the INFOCOM*, 2003.

[9] R. Nelson, *Probability, Stochastic Processes, and Queueing Theory*, Springer-Verlag, 1995.

[10] M. Reisslein, K.Ross and S. Rajagopal, A framework for guaranteeing statistical qos. *IEEE Transactions on Networking*, 10(01):27–42, 2002.

[11] M. Schwartz, *Broadband Integrated Networks*. Prentice-Hall, 1996.

[12] H. Zhang, Service disciplines for guaranteed performance service in packet switching networks. *Proceedings of the IEEE*, 83:1374–1399, 1995.