Uma política de escalonamento justa para redes Ethernet com elementos ópticos passivos

Flávio de M. Pereira[1], **Nelson L. S. da Fonseca**[1,2] **e Dalton S. Arantes**[3]

[1]Optical Internet Laboratory
http://www.ic.unicamp.br/~nfonseca/oil

[2]Instituto de Computação
Universidade Estadual de Campinas
Cx. Postal 6.176 — 13084-971 Campinas/SP Brasil

[3]Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas
Cx. Postal 6.101 — 13083-970 Campinas/SP Brasil

flaviomelopereira@yahoo.com.br, nfonseca@ic.unicamp.br,

dalton@decom.fee.unicamp.br

Resumo. *Neste artigo, propõe-se uma nova política de escalonamento para os fluxos de subida em redes com elementos ópticos passivos do tipo Ethernet. Essa política, denominada Compartilhamento Proporcional com Reserva de Carga, oferece garantias individuais de banda e de retardo a cada fluxo da rede. Além disso, ela é capaz de redistribuir proporcionalmente a capacidade ociosa da rede entre os fluxos ativos segundo a sua prioridade.*

Abstract. *We propose a novel discipline for scheduling upstream flows in Ethernet Passive Optical Networks, called Proportional Sharing with Load Reservation, which provides bandwidth and delay guarantees on a per flow basis. Moreover, it redistributes the unused bandwidth among active flows in proportion to their priority level.*

1. Introdução

As redes com elementos ópticos passivos do tipo Ethernet (EPONs) são redes ópticas ponto-multiponto que transportam dados em quadros Ethernet IEEE 802.3 e que têm em seu interior apenas elementos ópticos passivos unidirecionais tais como combinadores, divisores e acopladores ópticos. As transmissões dos fluxos de subida e de descida se fazem por meio de comprimentos de onda independentes, transportadas por fibras monomodo [Kramer et al. 2001, Takeuti 2005].

As redes EPONs são vistas como uma alternativa interessante para redes de acesso pois permitem levar a fibra óptica diretamente aos usuários finais, aumentando a banda disponível e reduzindo os custos de implantação e de manutenção [Kramer et al. 2001,

Durante a realização deste trabalho, F.M.Pereira estava na Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas.

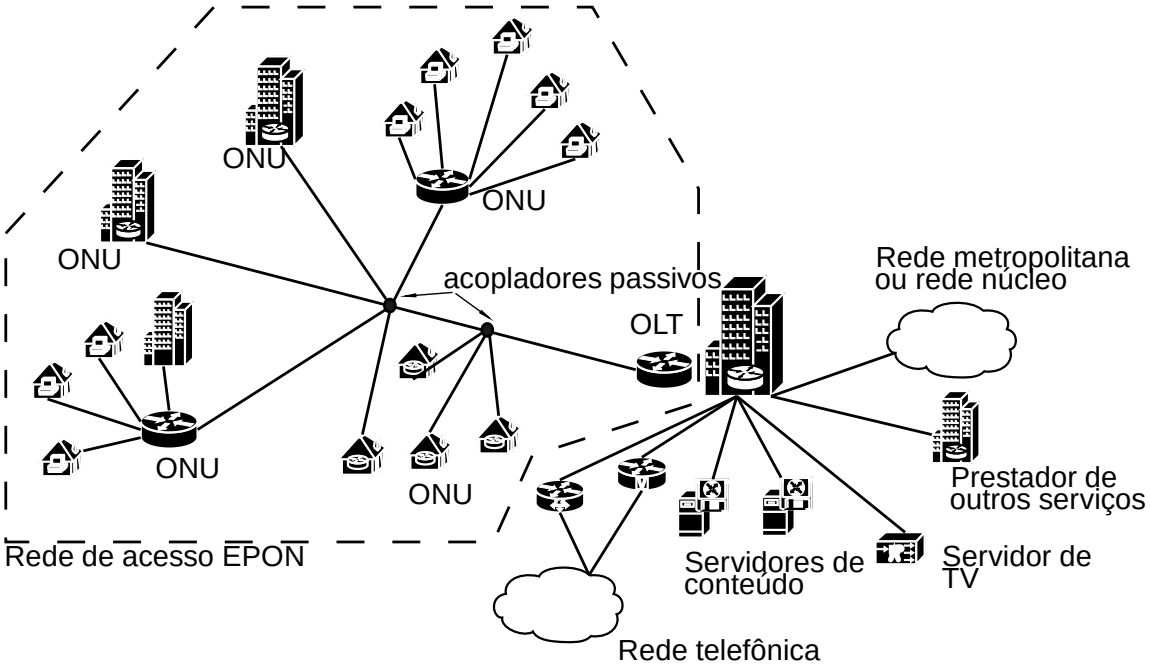

Figura 1. Exemplo de rede de acesso EPON.

Pereira 2006]. Em uma rede de acesso EPON, os usuários se conectam por meio de unidades de rede óptica (*Optical Network Units*, ONUs) exclusivas ou compartilhadas, que são conectadas a um terminal de linha óptica (*Optical Line Terminal*, OLT) disposto junto ao provedor de serviços. Um exemplo de rede de acesso EPON é apresentado na Figura 1.

As transmissões em redes EPON são feitas segundo a especificação da norma IEE 802.3. No fluxo de descida, a OLT transmite os quadros para as ONUs por difusão e cada ONU seleciona os quadros que lhe são destinados com base em campos de endereçamento lógico específicos. Nos fluxos de subida, a rede emprega um protocolo de *polling* denominado Protocolo de Controle Multiponto (*Multipoint Control Protocol*, MPCP) para controlar o acesso das ONUs ao meio compartilhado. A norma não estabelece, contudo, qual seria a política de escalonamento requerida para arbitrar este acesso nem para regular a concorrência dos fluxos de cada ONU pelos recursos da rede.

Políticas de escalonamento para redes EPON são hoje um tema de intensa pesquisa [Kramer et al. 2002, McGarry et al. 2004, Naser and Mouftah 2006]. Dados os elevados valores de largura de banda e de tempo de propagação envolvidos, as políticas quadro-a-quadro tradicionais não podem ser implementadas de maneira eficiente em uma rede EPON. Por isso, usualmente consideram-se políticas baseadas em janelas que distribuem as tarefas de escalonamento entre a OLT e as ONUs. Há, então, duas etapas que são denominadas escalonamento inter-ONU e intra-ONU. No escalonamento inter-ONU, a OLT atribui às ONUs janelas de tempo cuja duração é calculada com base nas demandas que elas informam ao final da sua última transmissão. Durante a sua janela de tempo, a ONU ocupa sozinha o canal compartilhado e pode então transmitir os quadros dos seus fluxos. A divisão da capacidade da janela de tempo entre os fluxos é realizada pela própria ONU por meio do escalonamento intra-ONU.

A maioria das propostas na literatura pressupõe que os escalonamentos intra-ONU e inter-ONU são realizados de maneira independente. Neste caso, entretanto, as ONUs não compartilham qualquer informação sobre a divisão da capacidade das suas janelas entre seus fluxos, o que pode levar a uma divisão desbalanceada dos recursos entre fluxos de ONUs diferentes. Com isso, a prioridade dos fluxos não seria respeitada e os usuários seriam tratados de maneira injusta [Kramer et al. 2004].

Para contornar esse problema, Kramer et al. [Kramer et al. 2004] propuseram uma política de escalonamento descentralizada denominada Enfileiramento Justo com Curvas de Serviço (*Fair Queueing with Service Envelopes*, FQSE). A política FQSE pressupõe uma relação hierárquica entre os escalonamentos inter-ONU e intra-ONU, tal que a OLT forneça às ONUs um indicador que lhes permita determinar o volume de tráfego de cada fluxo que se pode transmitir tal que os recursos da rede sejam divididos entre os fluxos de maneira balanceada e, portanto, justa.

Neste artigo, propõe-se uma nova política denominada Compartilhamento Proporcional com Reserva de Carga (*Proportional Sharing with Load Reservation*, PSLR), que também assegura a distribuição justa dos recursos entre os fluxos de diferentes ONUs. Assim como a política FQSE, a política PSLR permite que a rede reserve uma certa quantidade de recursos, na forma de taxa de serviço mínima, a cada um dos fluxos e atribua-lhes uma dada prioridade na concorrência pelos recursos ociosos. A diferença entre as duas políticas está na forma como a OLT e as ONUs regulam essa concorrência, o que leva à política PSLR a apresentar uma complexidade computacional inferior àquela imposta pela política FQSE. Além dessa vantagem, a política PSLR apresenta limitantes analíticos para a justiça relativa entre fluxos e para o retardo introduzido nos fluxos. Tais resultados ainda não foram obtidos para a política FQSE.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 apresenta a notação e o conceito de justiça adotado neste trabalho. A Seção 3 introduz a política PSLR. São apresentados limitantes para a justiça relativa e para o retardo introduzido nos fluxos pela política PSLR. A Seção 4 ilustra o funcionamento da política PSLR por meio da simulação de uma rede EPON típica. Finalmente, a Seção 5 apresenta algumas conclusões deste trabalho.

2. Aspectos preliminares

2.1. Notação e descrição do sistema

Neste trabalho, considera-se que o enlace de subida de uma rede EPON tem capacidade r e é compartilhado por N fluxos, que estão agrupados em J ONUs. Dentro das ONUs, cada fluxo tem uma fila exclusiva para armazenar os seus quadros em espera. O i-ésimo fluxo da j-ésima ONU é identificado pelo sub-índice ij.

Supõe-se que a OLT emprega um mecanismo de *polling* para atribuir janelas de tempo para que cada uma das ONUs transmita uma certa quantidade de quadros dos seus fluxos. No n-ésimo ciclo de *polling*, a duração (em bits) da janela de tempo atribuída pela OLT à j-ésima ONU é representada por $C_j[n]$. Durante essa janela, a ONU transmite $c_{ij}[n]$ bits de cada fluxo, quantidade esta que inclui os bits de preâmbulo e os bits entre quadros. O volume de tráfego (em bits) do ij-ésimo fluxo que esteja em espera dentro da ONU no instante em que se inicia a sua n-ésima janela para transmissão é dado por

$Q_{ij}[n]$. Considera-se que o tempo de passagem entre dois fluxos de uma mesma ONU é desprezível.

A duração total do n-ésimo ciclo de polling é então dada por $b[n] = hr + \sum_{ij} c_{ij}[n]$, sendo $h = \sum_j h_j$, em que a constante h_j representa o maior tempo de passagem entre a j-ésima ONU e a seguinte, que inclui o tempo de processamento necessário para o chaveamento.

2.2. Critério de justiça

Neste trabalho, supõe-se que uma política de escalonamento é justa se ela é capaz de reservar uma taxa mínima de serviço de ρ_{ij} a cada fluxo ij e, ao mesmo tempo, de compartilhar o restante da capacidade da rede entre os fluxos com tráfego em espera segundo sua prioridade. Seja $\mathcal{B}(t)$ o conjunto de fluxos que têm tráfego em espera no instante t. Uma política justa ideal é aquela que destina a cada fluxo $ij \in \mathcal{B}(t)$ uma taxa de

$$s_{ij}(t) = \rho_{ij} + \frac{w_{ij}}{\sum_{lm \in \mathcal{B}(t)} w_{lm}} \left[r - \sum_{lm \in \mathcal{B}(t)} \rho_{lm} - \sum_{hk \notin \mathcal{B}(t)} s_{hk}(t) \right], \tag{1}$$

em que w_{ij} é o fator de ponderação que representa a prioridade do fluxo ij e r é a capacidade da rede.

Como foi dito na Seção 1, as redes EPON normalmente requerem políticas baseadas em janelas em que as tarefas de escalonamento sejam distribuídas entre a OLT e as ONUs. Neste caso, é difícil obter um algoritmo de escalonamento que efetivamente se aproxime de (1) ao longo dos ciclos. Alternativamente, contudo, é possível propor um critério de justiça menos rigoroso e mais adequado ao caso de políticas distribuídas baseadas em janelas. Suponha que a duração dos ciclos de *polling* seja constante e dada por B bits. Suponha ainda que a rede destine a cada fluxo ij uma janela de transmissão cujo tamanho seja dado por

$$c_{ij}[n] = \begin{cases} \tilde{Q}_{ij}[n], & ij \notin \mathcal{B}[n] \\ \dfrac{\rho_{ij} B}{r} + w_{ij}\xi[n], & ij \in \mathcal{B}[n], \end{cases} \tag{2}$$

em que $\tilde{Q}_{ij}[n] = Q_{ij}[n-1] - c_{ij}[n-1]$ é o tráfego em espera do fluxo ij no final da $(n-1)$-ésima janela de transmissão da ONU j. Note que $\tilde{Q}_{ij}[n]$ é o volume de tráfego em espera que a ONU j utiliza para requerer recursos da OLT no final desse ciclo. O conjunto $\mathcal{B}[n]$ é o conjunto de fluxos que já teriam algum tráfego em espera ao final do n-ésimo ciclo mesmo que nenhum novo quadro chegue entre este ciclo e o anterior. Matematicamente, este conjunto é dado por $\mathcal{B}[n] = \left\{ ij \mid c_{ij}[n] < \tilde{Q}_{ij}[n] \right\}$. A variável $\xi[n]$ representa o volume adicional de serviço (por unidade de peso) que deve ser destinado a um fluxo $ij \in \mathcal{B}[n]$ no n-ésimo ciclo, que é dado por

$$\xi[n] = \frac{1}{\sum_{lm \in \mathcal{B}[n]} w_{lm}} \left\{ B - hr - \sum_{lm \in \mathcal{B}[n]} \frac{\rho_{lm} B}{r} - \sum_{hk \notin \mathcal{B}[n]} \tilde{Q}_{hk}[n] \right\}. \tag{3}$$

Seja agora $\mathcal{B}[\cdot]$ o conjunto de fluxos que têm tráfego em espera durante todos os ciclos de $n+1$ a $n+q$. O total de serviço a ser destinado a um fluxo $ij \in \mathcal{B}[\cdot]$ neste intervalo é então dado por

$$\begin{aligned} S_{ij}[n+1;n+q] &= \sum_{k=n+1}^{n+q} c_{ij}[k] \\ &= q\left(\frac{\rho_{ij}B}{r}\right) + \frac{w_{ij}}{\sum\limits_{lm\in\mathcal{B}[\cdot]} w_{lm}} \left\{ q\left(B - hr - \frac{B}{r}\sum_{lm\in\mathcal{B}[\cdot]}\rho_{lm}\right) - \right. \\ &\quad \left. \sum_{hk\notin\mathcal{B}[\cdot]}\sum_{k=n+1}^{n+q}\tilde{Q}_{hk}[n] \right\}. \end{aligned} \tag{4}$$

A taxa média de serviço destinada ao fluxo ij naquele intervalo é dada por

$$\begin{aligned} s_{ij}[\cdot] &= \left(\frac{r}{qB}\right) S_{ij}[n+1;n+q] \\ &= \rho_{ij} + \frac{w_{ij}}{\sum\limits_{lm\in\mathcal{B}[\cdot]} w_{lm}} \left\{ r\frac{B-hr^2}{B} - \sum_{lm\in\mathcal{B}[\cdot]}\rho_{lm} - \frac{r}{qB}\sum_{hk\notin\mathcal{B}[\cdot]}\sum_{k=n+1}^{n+q}\tilde{Q}_{hk}[n] \right\} \end{aligned} \tag{5}$$

que é parecida com a taxa destinada pela política ideal, dada por (1). Assim, neste trabalho consideramos que uma política é justa se (4) vale para todo fluxo $ij \in \mathcal{B}[\cdot]$.

2.3. A política FQSE

A política FQSE visa a dividir os recursos da rede entre os fluxos de maneira justa e independente da ONU a que pertençam. Esta política permite que a rede reserve uma certa quantidade de recursos, na forma de taxa de serviço mínima, a cada um dos fluxos e atribua-lhes uma dada prioridade na concorrência pelos recursos ociosos.

Para atingir esse objetivo, a política FQSE utiliza funções denominadas curvas de serviço, que expressam a quantidade de tráfego de cada fluxo que a rede deve transmitir em um dado ciclo. Tais funções são dadas em termos de um índice não-negativo denominado Parâmetro de Satisfação (PS), que por sua vez representa o quanto a rede é capaz de atender às demandas dos fluxos [Kramer et al. 2004].

Na política FQSE, cada ciclo de *polling* é dividido em uma fase de requisição e uma fase de permissão. Na fase de requisição, a OLT consulta as ONUs para obter a curva de serviço agregada de seus fluxos, que é representada por um conjunto de pontos que são os vértices da sua aproximação por uma função linear por partes. A OLT agrega as curvas de serviço das ONUs para calcular o índice PS da rede para o ciclo em questão, com o que também calcula a duração da janela de transmissão de cada ONU. Na fase de permissão, a OLT atribui as janelas de transmissão por meio de mensagens de permissão, que carregam o índice PS calculado. As ONUs, por sua vez, utilizam esse índice para dividir a capacidade da sua janela de transmissão entre os seus fluxos.

Em [Kramer et al. 2004], estuda-se a complexidade computacional da política FQSE. À parte do custo de se calcular o índice PS, a política FQSE tem complexidade computacional de $\mathcal{O}(n \log n)$ tanto em termos do número de ONUs como do número de fluxos por ONU. Além disso, essa política pode apresentar algum grau de injustiça na divisão de recursos devido a mecanismos utilizados para prevenir bloqueios por quadros que não caibam na janela de transmissão [Kramer et al. 2004, Sect. IV.A] e para aumentar sua utilização [Kramer et al. 2004, Sect. IV.B]. Considerando-se apenas o primeiro mecanismo, os resultados indicam que a diferença entre o volume de tráfego transmitido em um ciclo para fluxos de igual prioridade e igual reserva de taxa mínima é inferior a $2(L-1)$, em que L é o tamanho máximo de um quadro Ethernet.

3. Compartilhamento proporcional com reserva de carga

Apesar da simplicidade de (2)–(5), não é trivial elaborar um algoritmo computacional que atinja essa solução. Isto porque não há como se determinar a priori quais são os fluxos que compõem o conjunto $\mathcal{B}[n]$.

Lapsley e Low [Low and Lapsley 1999] estudaram algoritmos de controle de fluxo no tempo contínuo que envolvem problemas de otimização semelhantes àqueles de que se originam (2) e (3). Esses autores propõem a solução do problema por meio do algoritmo adaptativo baseado no gradiente projetado, demonstrando a sua convergência sempre que o conjunto de alocações factíveis for fechado e, nele, a função utilidade for contínua, $\cap$-convexa e monotonicamente crescente [Low and Lapsley 1999]. Dado que (2) e (3) advêm de um problema de otimização com função objetivo logaritmica, que satisfaz estas propriedades, pode-se adotar a mesma estratégia para elaborar um algoritmo de escalonamento para redes EPON.

Neste algoritmo, denominado Compartilhamento Proporcional com Reserva de Carga (*Proportional Sharing with Load Reservation*, PSLR), cada ONU envia à OLT a demanda total dos seus fluxos e a soma dos fatores de ponderação dos fluxos que têm tráfego em espera no final da sua janela de transmissão. Esses valores são utilizados pela OLT para obter a demanda global da rede e corrigir o tamanho das janelas de transmissão de cada ONU do próximo ciclo. Assim, cada ONU recebe uma janela que corresponde à sua taxa mínima adicionada da parcela da capacidade ociosa que deve ser destinada aos seus fluxos. Para que a ONU possa realizar internamente a divisão justa desses recursos, a OLT envia junto com a mensagem de permissão um indicador que expressa o tamanho da janela por unidade de ponderação que foi utilizado para estimar a parcela adicional de banda a ser destinada a cada fluxo com demanda reprimida. Este indicador é atualizado a cada ciclo com base justamente nas informações enviadas pela ONU, tal que a alocação de banda aos fluxos convirja iterativamente à alocação justa.

3.1. Algoritmo computacional

O algoritmo, que é ilustrado na Figura 2, é dividido em uma fase de inicialização e uma fase operacional. Na fase de inicialização, a OLT prepara as ONUs para as transmissões e mede o seu tempo de ida e volta (*round-trip time*, RTT). Na fase operacional, a OLT atribui janelas de tempo para que cada ONU transmita os seus quadros.

Durante a fase de inicialização do algoritmo, a OLT envia a cada ONU uma mensagem inicial para prepará-la para transmissão e para medir o seu RTT. Esta mensagem

pode ser encapsulada em uma mensagem GATE do protocolo MPCP ou ainda transmitida diretamente no enlace utilizando seqüências de escape, como proposto por Kramer et al. para a política IPACT [Kramer et al. 2002]. Quando uma ONU recebe aquela mensagem, ela marca os primeiros quadros em espera de cada fluxo ij até um total de $\rho_{ij}B/r$ bits. Os quadros marcados serão transmitidos no primeiro ciclo de *polling* da fase operacional. Vale notar que a ONU não marca os quadros que excederiam o dado limite de bits. Para cada fluxo ij, o total de bits marcados durante a fase de inicialização é representado por $c_{ij}[1]$.

Após marcar os quadros, a ONU envia à OLT uma mensagem de requisição contendo os seguintes campos:

- a identidade lógica da ONU j;
- a demanda total da ONU para o primeiro ciclo de *polling* da fase operacional, que é dado por $C_j[1] = \sum_i c_{ij}[1]$;
- o total de pesos dos fluxos que ainda terão tráfego a transmitir após o primeiro ciclo de *polling* da fase operacional, dado por $W_j[1] = \sum_i w_{ij}\mathbb{I}\left\{\tilde{Q}_{ij}[2] > 0\right\}$, em que $\mathbb{I}\{z\}$ é uma função indicadora que vale 1 quando a condição z é verdadeira e 0 caso contrário.

A mensagem de requisição pode ser encapsulada em uma mensagem REPORT do protocolo MPCP ou enviada diretamente pelo enlace utilizando seqüências de escape. Como mostra a Figura 2, a OLT espera a requisição de cada ONU antes de enviar a mensagem de inicialização para a ONU seguinte. Isto é necessário pois a OLT ainda desconhece o RTT de cada ONU e, por isso, não pode antecipar o envio das mensagens de modo a minimizar os tempos de passagem entre ONUs consecutivas.

Ao receber a mensagem de requisição da última ONU, a OLT inicia a fase operacional do algoritmo enviando uma mensagem de permissão para que a primeira ONU possa transmitir. A OLT calcula ainda o instante de tempo em que deve enviar as mensagens de permissão das demais ONUs para o ciclo corrente. Note que isto só é possível pois a OLT conhece tanto o RTT como a demanda total $C_j[1]$ de cada uma delas. Assim como ocorre no algoritmo IPACT, esse cálculo é feito de modo a antecipar as mensagens e, assim, minimizar o tempo de passagem entre ONUs consecutivas [Kramer et al. 2002].

Seja t_0 o instante de tempo em que a OLT envia a mensagem de permissão da primeira ONU para o n-ésimo ciclo de *polling* . O instante de tempo em que a OLT deve enviar a permissão da ONU j nesse ciclo é dado por

$$t_j = \max \begin{cases} t_0 \\ t_{j-1} + r^{-1}C_{j-1}[n] + G - R_j \end{cases} \tag{6}$$

em que R_j é o tempo de ida e volta da ONU j e G é o intervalo mínimo de tempo necessário para preparar a camada física para as transmissões e para compensar pequenas variações no tempo de ida e volta [Kramer et al. 2002, Kramer et al. 2001, Kramer 2005]. Note que, ao contrário do que ocorre no algoritmo IPACT, a OLT não envia permissões para um novo ciclo de *polling* antes que todas as requisições tenham sido recebidas. Isto é necessário porque a OLT precisa da demanda de todas as ONUs para calcular o tamanho de janela a que corresponde a parcela justa de recursos de cada ONU.

As mensagens de permissão que a OLT envia às ONUs contêm os seguintes campos:

- a identidade lógica da ONU, j;
- o serviço adicional por unidade de peso a ser destinado a cada fluxo, $\xi[n+1]$; e
- a duração, em bits, do ciclo de *polling* anterior, $b[n-1]$.

Assim como as demais mensagens de controle, a mensagem de permissão pode ser encapsulada em uma mensagem GATE do protocolo MPCP ou enviada diretamente pelo enlace utilizando seqüências de escape. Como condição inicial, supõe-se que $b[0] = B$ e $\xi[1] = 0$. Estes valores estão de acordo com o total de bits de cada fluxo que foi marcado na fase de inicialização do algoritmo. Para $n > 1$, $b[n]$ é dado pela diferença entre o valor de t_0 no ciclo de *polling* atual e o seu valor no ciclo anterior, multiplicada pela taxa r. O cálculo do valor de $\xi[n+1]$ é discutido mais adiante.

Quando a ONU j recebe a mensagem de permissão para o n-ésimo ciclo, ela calcula a quantidade de tráfego de cada fluxo ij que será transmitida durante o ciclo seguinte. Caso a quantidade calculada resulte na fragmentação de um quadro, a ONU pressupõe a sua transmissão integral e utiliza um contador para registrar quantidade extra de bits, que é descontada nos ciclos seguintes para que, em média, garanta-se a alocação justa dos recursos. Note que este mecanismo é similar ao que é utilizado pela política *Deficit Round Robin* [Shreedhar and Varghese 1995].

A quantidade de bits que cada fluxo ij pode transmitir durante o $(n+1)$-ésimo ciclo de *polling* é então dada por

$$c_{ij}^{e}[n+1] = \left[\frac{\rho_{ij} b[n-1]}{r} + w_{ij}\xi[n+1] - d_{ij}[n]\right]^{+}, \tag{7}$$

em que $0 \leq d_{ij}[n] < L$ é o contador do fluxo ij e $[z]^{+} = \max(0; z)$. A ONU então marca os quadros a serem transmitidos durante o $(n+1)$-ésimo ciclo, incluindo o que seria fragmentado se exatos $c_{ij}^{e}[n+1]$ bits fossem transmitidos. O total de tráfego que a ONU j marca para o fluxo ij é dado por $c_{ij}[n+1]$.

Após marcar os quadros de todos os fluxos, a ONU atualiza os contadores dos fluxos fazendo

$$d_{ij}[n+1] = \left[c_{ij}[n+1] + d_{ij}[n] - \frac{\rho_{ij} b[n-1]}{r} - w_{ij}\xi[n]\right]^{+}. \tag{8}$$

Supõe-se que $d_{ij}[n+1] = 0$ sempre que não houver quadros desmarcados do fluxo ij no instante em que os contadores são atualizados. Finalmente, a ONU envia à OLT os quadros que já haviam sido marcados na fase de inicialização (se for o primeiro ciclo de *polling*) ou durante o $(n-1)$-ésimo ciclo da fase operacional. A ONU termina sua transmissão enviando uma nova mensagem de requisição à OLT contendo os seguintes campos:

- a identidade lógica da ONU, j;
- a demanda total da ONU para o próximo ciclo de *polling*, dada por $C_j[n+1] = \sum_i c_{ij}[n+1]$;
- o total de pesos dos fluxos ij que ainda terão tráfego a transmitir após o ciclo seguinte, i.e., $W_j[n] = \sum_i w_{ij}\mathbb{I}\left\{c_{ij}[n+1] < \tilde{Q}_{ij}[n+1]\right\}$.

Após receber a requisição da última ONU, a OLT atualiza o valor de $\xi[n]$ por meio de

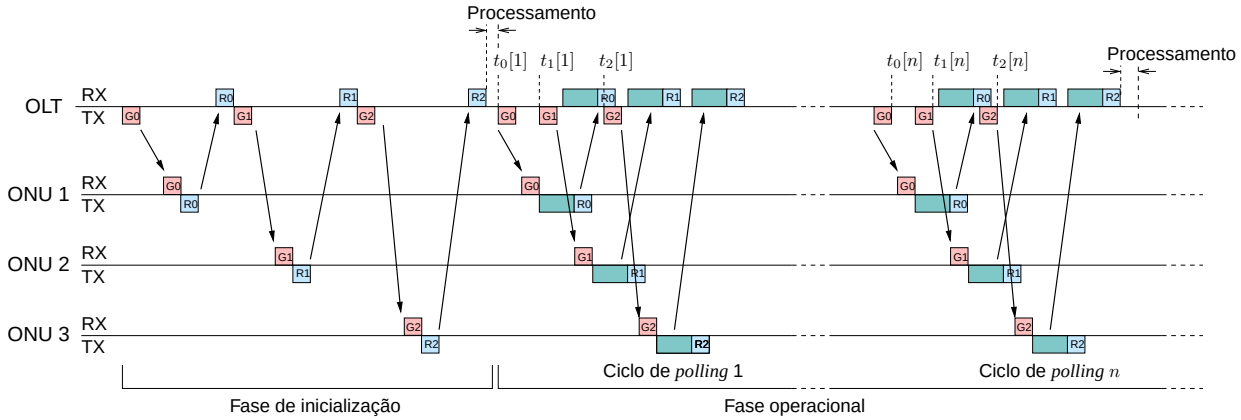

Figura 2. Operação do algoritmo PSLR.

$$\tilde{\xi}[n+1] = \begin{cases} \tilde{\xi}[n], & \sum_j W_j[n] = 0, \\ \left[\tilde{\xi}[n] + \eta\,(B - b[n])\right]^+, & \sum_j W_j[n] > 0 \end{cases} \tag{9}$$

$$\xi[n+1] = \frac{1}{\sum_j W_j[n]}\tilde{\xi}[n+1], \tag{10}$$

em que o parâmetro $\eta > 0$ corresponde ao passo do algoritmo. Finalmente, a OLT calcula a duração (em bits) do ciclo de *polling* e utiliza (6) para calcular o instante de tempo em que deve enviar uma nova mensagem de permissão a cada ONU. O processo de *polling* então se reinicia.

Em [Pereira 2006, Teorema B.1], demonstra-se que o algoritmo da política PSLR converge para a alocação justa (4) e (5) se o parâmetro η em (9) satisfizer a desigualdade $\eta < 1 + \frac{1}{r}\sum_{ij\in\mathcal{B}[\cdot]}\rho_{ij}$. Dado que o conjunto $\mathcal{B}[\cdot]$ varia durante a operação da rede, supõe-se que $\eta < 1$ é capaz de assegurar a convergência sob qualquer condição.

É fácil ver que a complexidade do algoritmo apresentado é $\mathcal{O}(n)$ tanto em relação ao número de ONUs como em relação ao número de fluxos por ONU [Pereira 2006, Teorema B.2]. Por isso, a política PSLR se mostra bastante atrativa em termos de escalabilidade. Vale notar ainda a sua simplicidade, dado que o algoritmo requer apenas operações aritméticas triviais para atingir a alocação justa de recursos aos fluxos. Ambas as características constituem importantes vantagens em relação à política FQSE, que tem complexidade $\mathcal{O}(n \log n)$ e que requer a execução de um complexo algoritmo de aproximação para obter as curvas de serviço de cada ONU.

3.2. Limitante de justiça

Para determinar o quão justa é a divisão de recursos promovida por uma política de escalonamento real, é necessário definir alguma medida que reflita a discrepância entre o serviço que ela oferece aos fluxos e o que deveria ser oferecido no caso ideal. Na literatura, uma das métricas mais utilizadas é o limitante de justiça relativa (*Relative Fairness Bound*, RFB) [Zhou 2003]. No caso de uma política que reserve uma banda mínima para os fluxos, o RFB deve ser calculado de modo a desconsiderar o serviço correspondente à

banda reservada aos fluxos ij e lm, já que esta parcela não advém da divisão dos recursos ociosos. Assim, o presente trabalho redefine o RFB como

$$\mathrm{RFB} = \max_{ij,lm \in \mathcal{B}(\tau;t)} \max_{\tau;t} \left| \frac{S_{ij}(\tau;t) - \phi\rho_{ij}}{w_{ij}} - \frac{S_{lm}(\tau;t) - \phi\rho_{lm}}{w_{lm}} \right|, \tag{11}$$

em que ρ_{ij} e ρ_{lm} representam a banda reservada pelos fluxos ij e lm, respectivamente, e $\phi \geq 0$ é o maior valor para o qual as frações são ambas positivas.

No caso da política PSLR, o limitante de justiça relativo satisfaz a desigualdade [Pereira 2006, Teorema B.4]

$$\mathrm{RFB} \;\leq\; \frac{1}{\min_{ij} w_{ij}} \max_{ij} \left[\rho_{ij} \frac{B + 2N(L-1)}{r - \sum_{lm} \rho_{lm}} + B - h + 2(L-1) \right]. \tag{12}$$

A demonstração deste resultado é omitida por restrições de espaço. A partir de (12), é fácil perceber que a diferença entre o serviço total oferecido a um dado fluxo e o que é oferecido a outro é sempre limitada, mesmo que os fluxos pertençam a ONUs diferentes.

Cabe observar ainda que é possível obter um limitante para o RFB menor do que (12) se o intervalo de interesse $(\tau;t)$ compreender n ciclos de *polling* inteiros. Conforme é demonstrado em [Pereira 2006, Teorema B.5], $\mathrm{RFB}_{\mathrm{CC}} \leq 2(L-1) / \min_{ij} w_{ij}$, ou seja, a diferença entre os serviços totais oferecidos a dois fluxos é sempre inferior ao tamanho de dois pacotes, ponderados pelo menor peso atribuído pela rede aos fluxos. Se forem considerados fluxos de igual prioridade e taxa mínima, o valor de $\mathrm{RFB}_{\mathrm{CC}}$ indica que a diferença de serviço entre dois fluxos sob a política PSLR é sempre inferior a $2(L-1)$, valor este similar ao obtido no caso da política FQSE quando se considera apenas o mecanismo para prevenir bloqueios por quadros maiores do que a janela de transmissão.

3.3. Latência e atraso

Stiliadis e Varma [Stiliadis and Varma 1998] definiram uma classe de políticas de escalonamento denominada servidores latência-taxa para a qual podem ser obtidos limitantes determinísticos para o atraso e para a latência de fluxos individuais. Uma política pertence à classe dos servidores latência-taxa com parâmetros $(\rho_{ij};\theta_{ij})$ se o serviço oferecido a qualquer fluxo ij que esteja latente em todo o intervalo $[\tau;t]$ satisfaz a relação $S_{ij}(\tau;t) \geq [\rho_{ij}(t-\tau-\theta_{ij})]^+$, em que ρ_{ij} é a taxa de serviço de longo prazo proporcionada pela política ao fluxo ij e θ_{ij} é o seu parâmetro de latência temporal. Este parâmetro representa o maior tempo necessário para que o escalonamento passe a atender o fluxo ij continuamente à taxa ρ_{ij}.

Para uma política que pertença à classe dos servidores latência-taxa, é possível obter limitantes de pior caso para a latência e para o atraso dos fluxos se o tráfego for policiado pelo algoritmo do balde furado. Tais limitantes são dados por $Q_{ij}(t) \leq \sigma_{ij} + \rho_{ij}\theta_{ij}$ e $D_{ij}(t) \leq \sigma_{ij}/\rho_{ij} + \theta_{ij}$ [Stiliadis and Varma 1998].

É possível demonstrar que a política PSLR pertence à classe dos servidores latência-taxa [Pereira 2006, Teorema B.7]. Nesse caso, o valor de θ_{ij} é dado pelo seguinte teorema:

Teorema 1. *Seja ρ_{ij} a banda reservada a um fluxo ij pela política PSLR. Neste caso, o escalonamento pode ser representado por um servidor latência-taxa $(\rho_{ij}; \theta_{ij})$ em que*

$$\theta_{ij} \leq \frac{1}{r}\left[3B - h + 4N(L-1) + \frac{B + 2N(L-1)}{r - \sum_{kp}\rho_{kp}}\left(r + 2\sum_{lm \neq ij}\rho_{lm}\right)\right]. \qquad (13)$$

Por restrições de espaço, omite-se a prova detalhada deste teorema, que é apresentada em [Pereira 2006, Teorema B.7]. Destaca-se, contudo, que o resultado desse teorema indicam que os limitantes de $Q_{ij}(t)$ e $D_{ij}(t)$ apresentados em [Stiliadis and Varma 1998] podem ser utilizados para caracterizar o máximo tamanho de fila e o máximo atraso de cada fluxo na rede de acesso EPON sob política PSLR.

4. Exemplo numérico

Nesta seção, ilustra-se a operação da política PSLR por meio de um exemplo numérico. Considera-se uma rede de acesso EPON com 16 ONUs, cada uma contendo 48 filas independentes com capacidade de 500kb. A distância entre a OLT e as ONUs é uniformemente distribuída entre 1km e 20km, de modo que o tempo de propagação entre os elementos de rede varia de 5μs a 100μs. Para o algoritmo PSLR , supõe-se que a duração desejada para os ciclos de polling (B) seja de 2Mb e que o intervalo mínimo entre as transmissões de ONUs consecutivas (G) seja de 1μs. Estes valores são típicos em simulação de redes EPON. Além disso, supõe-se que o passo do algoritmo PSLR (η) seja igual a 0,1.

Cada fila é utilizada por uma fonte de taxa de bit variável com taxa média de 5Mb/s. Os enlaces que conectam as filas e as fontes têm capacidade de 100Mb/s. Os parâmetros de serviço requeridos por cada fonte e uma descrição sucinta do correspondente perfil de serviço são apresentados na Tabela 1. De modo a verificar o efeito de variações de carga das ONUs na operação do algoritmo PSLR, supõe-se que os fluxos estejam ativos em diferentes intervalos de tempo $[t_{on}; t_{off}]$ que também são apresentados na Tabela 1.

Para cada fonte, o tráfego foi obtido por meio do gerador dado em [Kramer], que agrega 256 subfluxos on/off independentes com períodos de silêncio e de atividade paretianos com $\gamma = 1,3$. O tamanho dos quadros segue a distribuição obtida em uma rede de acesso real por Sala e Gummalla [Sala and Gummalla 2001]. Assim, o tráfego obtido corresponde ao que seria produzido por uma aplicação multimídia real que tenha tráfego com dependência de longa duração e $H = 0,85$.

Os resultados de simulação ora apresentados foram obtidos por meio do simulador Omnet++, em que foram implementadas classes C++ para as ONUs, a OLT e o enlace óptico [Varga]. A taxa destinada pela política PSLR a alguns dos fluxos da ONU 1 e da ONU 9 são apresentados nas Figuras 3 e 4. Estes valores correspondem a taxa média de bits transmitidos em intervalos de tempo não-sobrepostos de 0.1s. Os resultados obtidos para outros fluxos são omitidos por restrições de espaço.

Note que, para cada intervalo $[0; 10\text{s}]$, $[10\text{s}; 20\text{s}]$ e $[20\text{s}; 30\text{s}]$ em que não há variação na carga das ONUs, os valores de taxa obtidos por simulação são próximos aos que seriam obtidos aplicando (5) aos fluxos ativos. Além disso, os mesmos valores são obtidos para fluxos correspondentes tanto na ONU 1 como na ONU 9, mesmo quando o conjunto de fluxos ativos em cada uma dessas ONUs é diferente. Isso indica que o algo-

Tabela 1. Parâmetro de serviço dos fluxos.

Fluxo	ρ_i	w_i	ONU 1...8		ONU 9...16		Descrição
	[Mb/s]		t_{on}[s]	t_{off}[s]	t_{on}[s]	t_{off}[s]	
1...6	1	0	0	20	0	30	CBR Básico
7...12	1	1	0	20	0	30	VBR Básico de baixa prioridade
13...18	1	2	10	20	10	30	VBR Básico de alta prioridade
19...24	2	0	0	20	0	30	CBR Premium
25...30	2	1	10	30	10	30	VBR Premium de baixa prioridade
31...36	2	2	0	30	0	30	VBR Premium de alta prioridade
37...42	0	1	0	30	0	30	Melhor esforço de baixa prioridade
43...48	0	2	0	30	0	30	Melhor esforço de alta prioridade

ritmo satisfaz o critério de justiça dado por (5) mesmo quando se consideram fluxos de ONUs diferentes.

É possível ainda ver nas Figuras 3 e 4 que o algoritmo converge rapidamente para a nova alocação justa de taxas quando a carga da ONU varia em $t = 10\text{s}$. Para $t = 20\text{s}$, a resposta é mais lenta pois a rede precisa primeiro transmitir todo o tráfego em espera dos fluxos que se tornaram inativos antes de atingir a nova alocação justa.

A Figura 5(a) apresenta a duração dos ciclos de *polling* medidos durante a simulação. Note que o algoritmo rapidamente converge para a duração desejada, mesmo quando há variação na carga das ONUs. A vazão da rede é apresentada na Figura 5(b), que mostra que mais de 98% da capacidade da rede é utilizada na transmissão dos quadros. Este ótimo resultado se deve principalmente à antecipação do envio das mensagens de permissão às ONUs, o que reduz o tempo ocioso entre as janelas de transmissão das ONUs. Colaboram também para isso o baixo custo computacional do algoritmo, que reduz o intervalo ocioso entre os ciclos de *polling*,e o reduzido tamanho das mensagens de controle.

Os resultados de simulação indicam que a política PSLR é efetivamente satisfaz o critério de justiça definido na Seção 2. Além disso, ela é capaz de garantir taxa mínima aos fluxos e de distribuir a capacidade ociosa entre eles de maneira eficiente, independentemente da ONU a que pertençam.

5. Conclusões

Neste artigo, foi proposta uma nova política de escalonamento para redes EPON denominada Compartilhamento Proporcional com Reserva de Carga (PSLR). Esta política apresenta significativas vantagens em relação a outras disciplinas já apresentadas na literatura. As mais importantes são o oferecimento de garantias de desempenho fluxo-a-fluxo, a capacidade de prover uma distribuição justa dos recursos da rede entre os fluxos e o baixo custo computacional. Na política PSLR, cada fluxo pode estabelecer seu próprio contrato de serviço com a rede e receber a banda mínima contratada e uma parcela justa dos recursos ociosos independentemente da ONU em que se origina. A política FQSE proposta em [Kramer et al. 2004] atinge resultados similares, mas a custa de uma maior complexidade computacional. Ademais, a política FQSE não dispõe de expressões analíticas para limitantes de justiça e de retardo dos fluxos. Cabe ressaltar ainda que, assim como a política FQSE, a política PSLR pode também ser empregada em outras redes ponto-multiponto, dado que ela não requer qualquer característica particular das redes EPON

para a sua operação.

Agradecimentos

Os autores agradecem à FAPESP pelo apoio financeiro a este projeto (Processos 01/14379-4, 03/08264-5 e 03/08277-0).

Referências

Kramer, G. Generator of self-similar traffic (version 3).

Kramer, G. (2005). *Ethernet Passive Optical Networks*. McGraw-Hill.

Kramer, G., Banerjee, A., Singhal, N., Mukherjee, B., Dixit, S., and Ye, Y. (2004). Fair Queuing with Service Envelopes (FQSE): A cousin-fair hierarchical scheduler for subscriber access networks. *IEEE J. Select. Areas Commun.*, 22(8):1497–1513.

Kramer, G., Mukherjee, B., and Pesavento, G. (2001). Ethernet PON (ePON): Design and analysis of an optical access network. *Photonic Network Communications*, 3(3):307–319.

Kramer, G., Mukherjee, B., and Pesavento, G. (2002). Interleaved Polling with Adaptive Cycle Time (IPACT): A dynamic bandwidth distribution scheme in an optical access network. *Photonic Network Communications*, 4(1):89–107.

Low, S. and Lapsley, D. (1999). Optimization flow control — I: basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7(6):861–874.

McGarry, M. P., Maier, M., and Reisslen, M. (2004). Ethernet PONs : a survey of dynamic bandwidth allocation DBA algorithms. *IEEE Commun. Mag.*, 42(8):s8–s15.

Naser, H. and Mouftah, H. T. (2006). A joint-ONU interval-based dynamic scheduling algorithm for ethernet passive optical networks. *IEEE/ACM Trans. Networking*, 14(4):889–899.

Pereira, F. M. (2006). *Modelagem, policiamento e escalonamento de tráfego em redes Ethernet PON*. Tese de doutorado, Universidade Estadual de Campinas.

Sala, D. and Gummalla, A. (2001). PON functional requirements: Services and performance.

Shreedhar, M. and Varghese, G. (1995). Efficient fair queueing using Deficit Round Robin. In *Proc. SIGCOMM*, pages 231–242.

Stiliadis, D. and Varma, A. (1998). Latency-Rate servers: a general model for analysis of traffic scheduling algorithms. *IEEE/ACM Transactions on Networking*, 6(5):611–624.

Takeuti, P. (2005). Projeto e dimensionamento de redes ópticas passivas (PONs). Dissertação de mestrado, Universidade de São Paulo.

Varga, A. The Omnet++ simulator (version 3.0).

Zhou, Y. (2003). *Resource allocation in computer networks: fundamental principles and practical strategies*. PhD thesis, Drexel University.

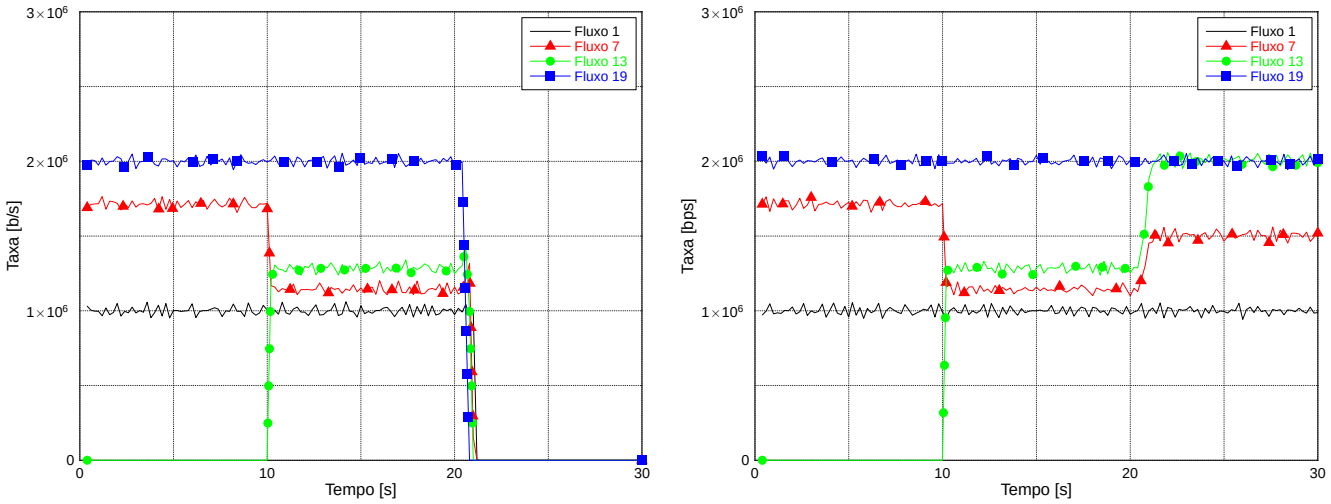

Figura 3. Resultados para os fluxos 1, 7, 13 e 19 das ONU 1 e 9.

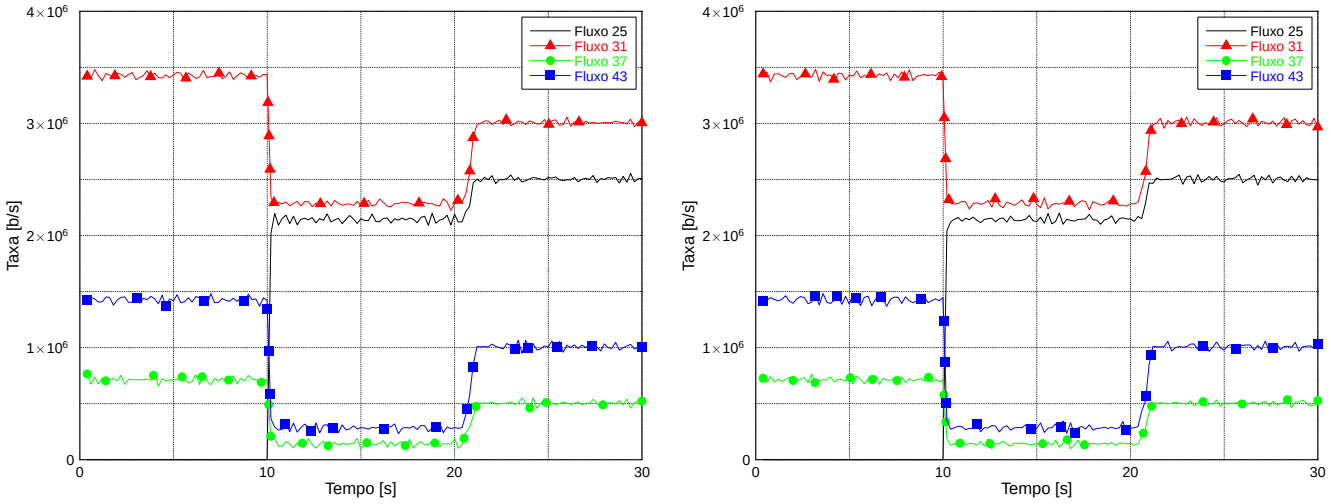

Figura 4. Resultados para os fluxos 25, 31, 37 e 43 das ONU 1 e 9.

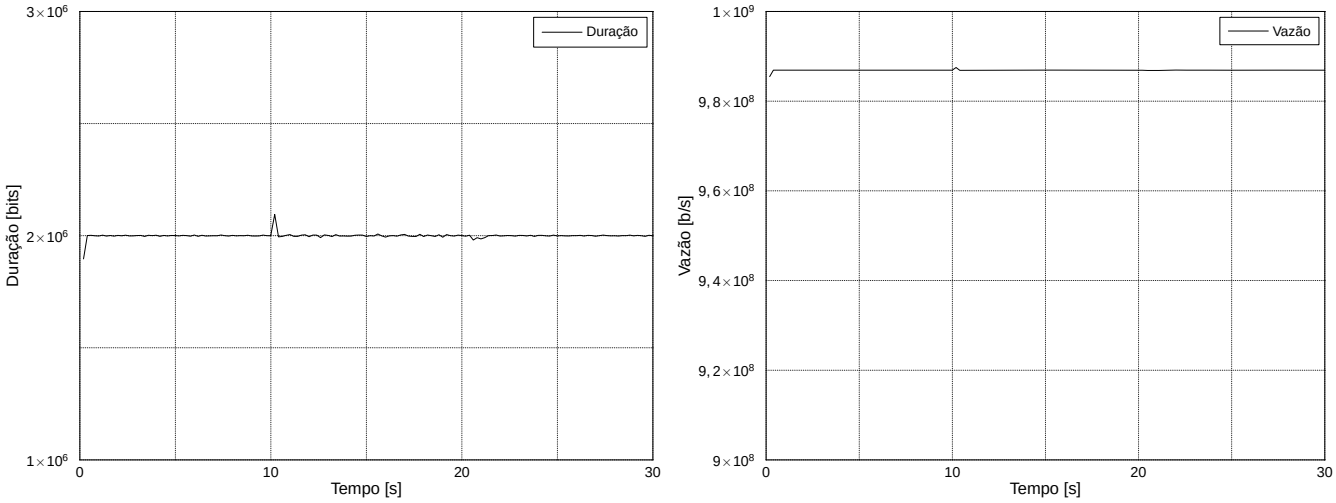

Figura 5. Operação da política PSLR. (a) Duração dos ciclos. (b) Utilização da rede.