

# ESTIMATION OF TOKEN BUCKET PARAMETERS FOR VIDEOCONFERENCING SYSTEMS IN CORPORATE NETWORKS

J.Glasmann, M. Czermin, A. Riedl

Lehrstuhl für Kommunikationsnetze, Technische Universität München,

Arcisstr. 21, 80333 München, Germany

Tel.: +49 89 28 92 35 05, Fax: +49 89 28 96 35 05, E-mail: [Josef.Glasmann@ei.tum.de](mailto:Josef.Glasmann@ei.tum.de)

**Abstract:** *In this paper we examine bandwidth requirements of videoconferencing systems in corporate networks. Based on extensive measurements with different H.323 applications under various conditions, important statistical traffic properties are determined. In particular, we estimate token bucket parameters and calculate effective bit rates according to the Guaranteed Service approach specified by the IETF in the Integrated Services Architecture. Finally, we evaluate the theoretical results through simulations using the obtained traces as traffic input.*

**KEYWORDS:** *H.323, IntServ, Guaranteed Service, QoS, Effective Bandwidth*

## 1. INTRODUCTION

As the Internet evolves into a universal network platform, which is being used for all kinds of applications, quality of service issues become increasingly important. Many applications such as the videoconferencing systems that we consider in our paper impose stringent real-time requirements on the network, thus, calling for special packet handling mechanisms. As a consequence, two major service architectures have been specified by the IETF. The Differentiated Services initiative (DiffServ) [1] favors soft Quality of Service (QoS) guarantees through service discrimination with only a few traffic classes, while the Integrated Service Architecture (IntServ) [2] also proposes per-flow resource reservation and packet handling with hard QoS guarantees. For both approaches to work, it is necessary to know about the traffic, which has to be transferred over the network. Especially for IntServ a thorough understanding of individual flow characteristics is vital, so the usually tight QoS requirements can be met.

Characterization in the given context refers to the description of statistical traffic properties according to a certain traffic model. This traffic model might be used for issues of network planning (e.g., topology design, capacity assignment, or scheduling strategies) or Call Admission Control (CAC) and resource reservation.

Usually, the desired quality of service (delay, loss rate) is given, and the necessary resources ("effective bandwidths") have to be calculated. In section 2 we introduce the Guaranteed Service model, which is the basis for our characterization efforts.

In recent literature, traffic is often characterized by packet or bit rates only, summing up the number of packets or bytes observed in certain time intervals. However, for the investigation of bandwidth requirements of high speed data sources and delay distributions for packets of individual flows, the packet interarrival times and the packet lengths are significant and, therefore, have to be taken into account.

In order to determine the bandwidth requirement of H.323 videoconferencing systems [3], extensive traffic measurements with applications from PictureTel (LiveLan), Vcon (ARMADA Escort 25Pro), and Microsoft (Netmeeting v. 2.11) were carried out. The applications run on PCs (Pentium II) with Windows'95/Windows'98 as operating system. While using the conferencing systems over an Ethernet Local Area Network (LAN), we traced the generated traffic. The header information of successfully transferred Ethernet packets together with the respective timestamp was stored. The entire set of measurements covers approximately 200 individual measurements with varying program settings and picture dynamics.

The selected applications differ in the employed codecs, attainable bit rates, and packetization behavior. In general, the task scheduling of the operating system and networking effects (CSMA/CD protocol) have also great impact on the packet flow on the LAN. The latter can be eliminated by using switched LANs and full duplex mode. On a main processor normally more than one task (e.g. system/application processes) run at the same time. Without any task prioritization of the sending process of real-time packets, the packet arrival process on the Ethernet is biased and characteristics of the codec are hard to observe on packet level. Windows'95 for example supports no task prioritization while Windows'98 shows some improvements.

The videoconferencing systems LiveLan and ARMADA Escort implement the H.261 video codec standardized by the ITU [4]. Both conferencing systems provide quite comfortable user interfaces, which offer several program adjustments. For example, it is possible to adapt the video data rate to different transmission speeds in a range from 64 kbps (only audio) up to 768 kbps. The set bit rates include audio and video data generated by the applications and are not exceeded. The total of the signaling traffic (RTP/RTCP, H.245, H.225.0) [3,4] amounts to less than 1 kbps and, thus, is negligibly small.

For audio transmission both applications employ constant bit rate codecs. While PictureTel relies on the G.711 encoding scheme [5] with 64 kbps data rate, Vcon uses G.728 [8] with 16 kbps.

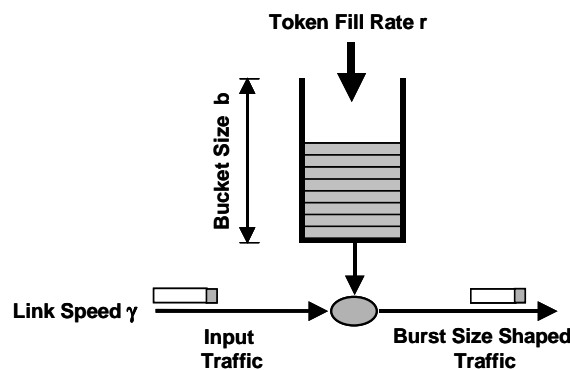
Netmeeting uses a proprietary codec for video and a G.723.1 implementation (6.3 kbps) [7] for audio, with "silent suppression" as a default setting. The adjustment possibilities that Netmeeting offers do not refer to data rates but allow a more

intuitive setting of parameters. The user can select the picture size by choosing "small", "medium", or "large", and vary the degree of quality between "faster" and "higher quality". Over timescales of several seconds the bit rates can diverge strongly from their average value.

## 2. GUARANTEED SERVICE (INTSERV)

The Integrated Services Architecture (RFC 1632) [1] specified by the IETF defines two service classes: Controlled Load Service [10] and Guaranteed Service [11]. For applications with strict real time requirements the Guaranteed Service provides hard upper bounds of the end-to-end network delay. Before real-time data are transmitted, the required network resources (buffer, bandwidth) are reserved for every flow using a separate signaling protocol. Since every data flow in the network uses its dedicated resources, it can be considered as an isolated stream, which is independent of all other traffic streams.

The characterization of a source according to IntServ's Guaranteed Service is based on the Token Bucket principle. The Token Bucket is a policing unit in the network. It monitors the traffic that is generated by a single source and if necessary limits with traffic by dropping individual packets. Figure 1 depicts the function of the Simple Token Bucket filter (STB). A bucket holds up to  $b$  tokens. For every byte sent by the source, one token is taken out while new tokens are put into the bucket with rate  $r$ . At times when the bucket is empty, arriving packets are dropped. Thus, if the bucket is full, a maximum burst of  $b$  bytes can pass the token bucket without being affected. However, in the long run, the average data rate cannot be greater than  $r$ .



**Figure 1: Simple Token Bucket Filter**

The Complex Token Bucket (CTB) is a more advanced type of token bucket filter. In addition to STB's functionality, it has the capability of limiting the peak rate  $p$  of a source. Even if the bucket is full, the source cannot necessarily send packet bursts with link speed.

Every traffic source can now be characterized by the given token bucket parameters  $(r, b)$  or  $(r, b, p)$  for STB and CTB, respectively. On the basis of these worst case traffic parameters an effective bit rate  $R$  and the required buffer space

[11,12] can be calculated for a desired upper delay bound. This bit rate  $R$  has to be reserved along the data path in the network. When using a work conserving scheduling discipline like weighted fair queuing (WFQ) [14] in the network nodes the flow can normally achieve a higher bandwidth than  $R$ . Work-conserving scheduling disciplines share unused link capacities among all active flows in a fair manner.

The determination of the effective bit rate is based on a worst case calculation of the maximum delays. The method is introduced in the Guaranteed Service specification (RFC 2212) [11] for the Complex Token Bucket and in [12] also for the Simple Token Bucket. The concept behind the CTB-formula is shortly explained in the following section.

It is assumed that the token bucket  $b$  is filled at the beginning of the transmission. The source starts with sending the maximum burst allowed by the token bucket filter and the bucket is emptied completely with the peak rate  $p$ . It is further assumed that one link along the path is overloaded. On this link only the reserved rate  $R$  is available for the data flow. Therefore, the maximum queuing delay is determined by the fact that the maximum possible burst arrives at this link and is shaped by the reserved rate  $R$ .

In addition to the maximum queuing delay, the end-to-end delay includes further components, which are caused by the scheduling procedure and by the propagation delay in the network. These delays occur on every link along the data path. In case of WFQ for example, the delay due to scheduling procedures consists of a maximum transmission time of a packet currently served ( $L_{max}/\gamma$ ) and the scheduling interval ( $L_{max}/R$ ) with  $L_{max}$  being the maximum packet size. The maximum transmission time of a packet is the maximum packet length divided by the link capacity. The Scheduling Interval is the time, during which a rate-based scheduler has to serve an incoming packet. It corresponds to the transmission time of a packet served with the reserved rate  $R$  according to the Fluid Flow Model [14,15]. Link rates are usually substantially larger than reserved rates  $R$ . Therefore, transmission times are negligible small compared to Scheduling-Intervals.

The total end-to-end delay in a network consisting of  $K$  hops can be given by following equations:

$$D_{e2e} \leq \begin{cases} \left[ \frac{(b-L)(p-R)}{R(p-r)} + \frac{L}{R} + \sum_{j=1}^K \left[ \frac{C_j}{R} + D_j \right] \right] & , \text{if } p > R \\ \left[ \frac{L}{R} + \sum_{j=1}^K \left[ \frac{C_j}{R} + D_j \right] \right] & , \text{if } p \leq R \end{cases} \quad D_{e2e} \leq \frac{b}{R} + \sum_{j=1}^K \left[ \frac{C_j}{R} + D_j \right]$$

**Equation 1: Maximum Delay  
(Complex Token Bucket)**

**Equation 2: Maximum Delay  
(Simple Token Bucket)**

If the effective bandwidth is larger than the peak rate, no queuing-delay occurs, and there are no bursts to shape. The delay consists only of Scheduling-Intervals, transmission times and propagation delay. When using WFQ as scheduling mechanism, the link parameters  $C$  and  $D$  are [15]:

$$C_j = L$$

$$D_j = \frac{L_{\max, j}}{\gamma_j} + T_{prop, j}$$

**Equation 3: Link Parameters**

Here,  $L = L_{\max}$  stands for the maximum packet length of the IP data flows including the Ethernet headers.  $T_{prop, j}$  is the propagation delay and  $\gamma_j$  is the capacity of link  $j$ . Setting the packet length for all data sources equal to the maximum possible value in Ethernet networks (1536 bytes) results in effective bit rates, which are too pessimistic, especially when sources have only lower bandwidth requirements (e.g. audio sources) and the hop count is high. In these cases the Scheduling-Interval, which is significantly influenced by the maximum packet length multiplied by the hop count, becomes the dominating component of the end-to-end delay. As a consequence, the token bucket parameters  $r$ ,  $b$ , and  $p$  are not enough to describe the traffic characteristics with respect to the bandwidth requirement in a sufficient way.

We can solve Equations 1 and 2 for the effective bit rate  $R$  and obtain:

$$R = \begin{cases} \frac{(b + LK)p - (K + 1)Lr}{b - L + (p - r) \left[ D_{e2e} - \sum_{j=1}^K \frac{L_{\max, j}}{\gamma_j} + T_{prop, j} \right]}, & \text{if } p > R \\ \frac{\sum_{j=1}^{K+1} L}{D_{e2e} - \sum_{j=1}^K \frac{L_{\max, j}}{\gamma_j} + T_{prop, j}}, & \text{if } p < R \end{cases} \quad R = \max \left[ r, \frac{b + (K - 1)L}{D_{e2e} - \sum_{j=1}^K \left[ \frac{L_{\max, j}}{\gamma_j} + T_{prop} \right]} \right]$$

**Equation 4: Effective Bandwidth  
(Complex Token Bucket)**

**Equation 5: Effective Bandwidth  
(Simple Token Bucket)**

### 3. ESTIMATING TOKEN BUCKET PARAMETERS

There is no standardized procedure to determine the Token Bucket parameters on the basis of packet traces. The bucket filling rate must be larger or equal to the mean transmission rate of the source and lower than the calculated effective bit rate.

The formula for the calculation of the effective bit rate according to the CBT-method (Equation 4) shows for bucket filling rates smaller than 150 kbps and bucket sizes smaller than 10 kbyte a strong dependency on the bucket size.

However, the bucket size depends on the source traffic and the token filling rate. Therefore, we have to make a tradeoff between setting the token filling rate and minimizing the token bucket size. The peak rate has substantially smaller influence on the effective bit rate than the bucket size.

The audio sources of all applications employ codecs with constant upper limits in their bandwidth requirements  $r_{set,au}$ . The audio data rates are independent of other application settings and can be used as the token filling rate. The mean transmission rate of the video traces  $r_{mean,vi}$  depends on the selected settings and the movement behavior of the users. The bit rate  $r_{mean,vi}$  usually lies approximately 10-15% below the bitrate-setting  $r_{set,vi}$  and achieves these rates only with very high movement proportions in the picture. In case of LiveLan and Vcon applications, the rates  $r_{set,vi}$  are determined on the basis of the bit rate settings  $r_{set,all}$ , which represent the sum of all data streams (audio and video) generated by the application. The total bit rate sets an upper limit for the video codec algorithm and can be used as the token filling rate for video sources. For short periods of time (about 100ms) the video bit rate may exceed the rate limit  $r_{set,vi}$ . In these cases the application normally reduces - after some delay (about 100ms) - the quality (resolution, quantization) of the picture.

Netmeeting does not offer such a setting possibility for the video codec. There is not always a direct causal connection between the application settings like picture size and quality and the video bit rate. The token filling rate can only be estimated on basis of several measurement series by calculating the mean bit rates of the traces, taking the highest ones, and rounding them up.

The rates  $r_{Link}$  (Table 1) correspond to the bandwidth requirements on the link. These bit rates take into account the packet-overhead (including Ethernet header) and are higher than the settings of the codec.

For all traces, the bucket size  $b$  is determined by the maximum number of tokens that are required in the bucket so no packets are dropped. The traces show high variance regarding to their bucket sizes. Normally, our proposal represents the maximum  $b$  value of all traces with comparable settings. If using the mean rate  $r_{mean,vi}$  of the individual traces as the bucket filling rate and  $b(r_{mean})$  as bucket size for video sources, the effective bit rates become unrealistically high. We achieve much smaller bucket sizes and more realistic effective bit rates by using  $r_{set,vi}$  as the token filling rate. Only a few heavy bursts in the LiveLan and Vcon traces consisting of 5-8 packets each are responsible for the high bucket sizes. In some special cases we tolerate packet losses smaller than 0.1% (belonging to a single exceptionally high burst) to reduce the bucket size considerably.

The peak rates are calculated by dividing the packet lengths by the appropriate packet interarrival times for all packets of a trace. Usually the highest peak rate is chosen as the peak rate of the trace. Sometimes the 99.9% quantile of all sample peak rates is taken in order to eliminate the disturbing influence of measuring errors. In the following, the effective bit rates are calculated according to Equation 4 for a network consisting of 5 hops, a link capacity of 100 Mbps, WFQ

schedulers, and a delay criterion of at most 100 ms. The maximum end-to-end delay contains only the network delay neglecting coding or packaging delays in the terminals.

The following two tables show the estimated token bucket parameters for our video and audio sources with different adjustments. The right column contains the effective bit rates calculated with the STB and CTB method.

As we expected, the CTB method results in bandwidth requirements that are for the same type of traffic 5-20% lower than the ones computed with the STB method. The smaller the peak rates are the bigger is the difference between the two methods. Another remarkable thing is that especially for sources with low mean rates the bandwidth requirements are very high. LiveLan and Vcon audio sources have bandwidth demands higher than their peak rate.

Video-Application	Settings		Token Bucket			Packet Length	Effective Bit rate Calculation	
	$r_{set,all}$ [Kbps]	$r_{link,vi}$ [Kbps]	$r$ [Kbps]	$b(r)$ Kbyte]	$p$ Mbps]	$L_{max}$ [Kbyte]	$r_{v,STB}$ [Kbps]	$r_{v,CTB}$ [Kbps]
LiveLan	768	740	750	7.0	9.5	1.52	1 200	1 160
	384	338	350	7.0	2.0	1.52	1 200	1 120
	174	117	120	3.2	0.3	1.52	875	741
Vcon	384	390	400	8.0	9.9	1.30	1 280	1 230
	128	119	120	3.2	6.2	1.47	875	859
	64	52	52	2.7	3.0	1.46	835	809
Netmeeting	P: large Q: high	~130	200	7.0	9.9	1.43	1 180	1 130
	P: large Q: medium	~30	70	5.0	9.9	1.42	1 020	989
	P: medium Q: medium	~20	35	2.5	9.9	1.42	815	808
	P: medium Q: medium	~15	25	2.0	4.5	1.42	774	768
	P: medium Q: fast							

**Table 1: Token Bucket Parameters (Video), WIN'95**

Audio-Application	Settings		Token Bucket			Packet Length	Effective Bit rate Calculation	
	$r_{codec,au}$ [Kbps]	$r_{link,au}$ [Kbps]	$r$ [Kbps]	$b(r)$ Kbyte]	$p$ [Kbps]	$L_{max}$ [Kbyte]	$r_{v,STB}$ [Kbps]	$r_{v,CTB}$ [Kbps]
LiveLan	64	72.25	73	1.16	130	578	325	279
Vcon	16	23.04	24	0.44	45	216	122	104
Netmeeting	6.3	23.63	24	0.45	400	90	101	93

**Table 2: Token Bucket Parameters (Audio), WIN'95**

Further, we notice that the peak rates for Netmeeting video sources are very high although the mean bit rates are very low and the gaps in between consecutive video frames are very large. But Netmeeting video frames normally consist of several

packets, which are transmitted in the form of short bursts. This behavior is responsible for the high peak rates.

## 4. SIMULATION RESULTS

To evaluate the calculated effective bandwidth values, simulations with the network simulator ns from Berkeley were carried out. The sample topology consists of 5 nodes, including one sink and one source. In this scenario each packet of a trace is fed into a FIFO queue per individual hop, which supports a minimal (guaranteed) link rate of  $R_{sim}$ . This guarantees that no packet - between the source and sink - exceeds the calculated bandwidth. In this simulation no packet loss is permitted and all links are configured with the same parameters ( $T_{prop} = 2.5\mu s$ ,  $\gamma_{Link} = R_{sim}$ ). In addition to that, the source traffic is policed by a token bucket filter. With these conditions it is possible to take into account the worst-case time parameters such as queuing delay, scheduling interval, and propagation delay per link. The only additional parameter, which is not considered for the maximum delay, is the transmission time of each packet at every hop. With an average link capacity of 100 Mbps and 5 hops, as mentioned before, the maximum total transmission time sums up to only 0.62 ms which is negligible. The same is true for the propagation delay since we consider only scenarios with small dimensions. Examination of the different delay components, i.e., comparing the queuing delay and the scheduling interval with all other delay components, validates these assumptions.

Video-Application	Settings			Effective Bit rate			Simulation		
	$\Gamma_{link,vi}$ [Kbps]	$\Gamma_{v,STB}$ [Kbps]	$\Gamma_{v,CTB}$ [Kbps]	$R_{v,sim}$ [Kbps]	Queuing Delay [ms]	Scheduling-Interval [ms]			
LiveLan	740	1 200	1 160	930	48	52			
	338	1 200	1 120	850	43	57			
	117	875	741	600	20	80			
Vcon	390	1 280	1 230	900	45	55			
	119	875	859	600	20	80			
	52	835	809	590	25	75			
Netmeeting	~130	1 180	1 130	900	45	55			
	~30	1 020	989	750	46	54			
	~20	815	808	650	26	74			
	~15	774	768	650	24	76			

Table 3: Video Sources, WIN'95

Audio-Application	Settings			Effective Bit rate			Simulation		
	$\Gamma_{link,au}$ [Kbps]	$\Gamma_{au,STB}$ [Kbps]	$\Gamma_{au,CTB}$ [Kbps]	$R_{au,sim}$ [Kbps]	Queuing Delay [ms]	Scheduling-Interval [ms]			
LiveLan	72.25	325	279	225	20	80			
Vcon	23.04	122	104	85	20	80			

<b>Netmeeting</b>	23.63	101	93	65	55	45
-------------------	-------	-----	----	----	----	----

**Table 4: Audio Sources, WIN'95**

The results of the simulations (Table 3 and 4) show that the impact of the maximum scheduling intervals on the calculated effective bandwidth, regarding sources with high rates, is equal to the queuing-delay. For sources with low rates the scheduling interval dominates the cumulative delay.

Detailed examination of Table 3 reveals that the bandwidth values inferred from the simulations are 20-40% below the calculated values. In this context it should be noted that the simulated values are the rates, which are needed in reality. This difference is a consequence of the pessimistic assessment of the maximum bursts according to the token bucket model. The simulations have shown that there is a small possibility that the whole bucket will be depleted with peak rate.

To improve this model we examine the dependency between the token fill rate and the respective token bucket size. An increase of the token fill rate reduces the necessary token bucket size, which causes a smaller calculated effective bandwidth. The results of this modification of the model shows that the possible reduction of the token bucket size is very small and the calculated effective bandwidth, according to the complex token bucket formula, increases for higher token fill rates. In addition to that, we analyze how the packet loss probability depends on a step-by-step reduction of the bucket size. Here we observe some significant distinctions between the two operating systems. When using Windows'95 the step-by-step reduction of the token bucket size causes a great enhancement of the packet loss (>1%) already for very small steps. The gained reduction of the effective bandwidth is very small in comparison to the measured packet loss. When using Windows'98 on the other side the bucket size for high-speed video traffic can be approximately halved while the packet losses stayed tolerable small (< 0.05 %). In the case of for example 384 kbps video the effective bitrate (CTB) can be scaled down to 755 kbps (Vcon) and 828 kbps (LiveLan).

## 5. CONCLUSION

In this paper, we examine the applicability of effective bandwidth formulas given in IntServ's Guaranteed Service specification to H.323 video conferencing traffic. Based on packet traces generated by several applications with various settings, the necessary parameters are determined and the effective bandwidth values are computed according to these formulas. Alternatively, the effective bit rates of the individual streams are investigated by simulation where the packet traces are used as input. By comparing the results, it is possible to evaluate the accuracy of the formulas when applied to the given scenarios.

As the Guaranteed Service model is designed to support hard QoS guarantees, it has to estimate the bandwidth requirements of real-time traffic very conservatively.

Comparing the theoretic results with our simulation data, we can observe that for every scenario the computed bandwidth is indeed larger than the actual required one, thus, allowing the Guaranteed Service approach to provide the requested QoS. However, as simulation results also show, in many cases a lot less bandwidth would suffice to achieve the same QoS. Especially for video sources the calculations are too pessimistic.

In view of the high bandwidth requirements, it is questionable whether such QoS guarantees justify the costs for additional infrastructure and network components. For certain traffic streams, the effective bandwidth can be greatly reduced, if one only allows a few packet losses. This indicates that often only a few bursts lead to the unreasonably high bandwidth demands.

## REFERENCES

[1]	S.Blake, D.Black, M.Carlson, E.Davies, Z. Wang, W. Weiss: An Architecture for Differentiated Services, IETF RFC 2475, December 1998.
[2]	Braden, Clark, S. Shenker: Integrated Services in the Internet Architecture: an Overview, IETF RFC 1633, June 1994.
[3]	ITU-T Rec. H.323: Visual Telephone Systems an Terminal Equipment for Local Area Networks which Provide a Non-Guaranteed Quality of Service,1996.
[4]	ITU-T Rec. H.261: Video CODEC for Audiovisual Services at p x 64 kbit/s, 1993.
[5]	ITU-T Rec. H.225.0: Media Stream Packetization and Synchronization for Visual Telephone Systems on Non-Guaranteed Quality of Service LANs, 1996.
[6]	ITU-T Rec. H.245: Control of Communications between Visual Telephone Systems and Terminal Equipment, 1996.
[7]	ITU-T Rec. G.711, Pulse Code Modulation (PCM) of voice frequencies
[8]	ITU-T Rec. G.728, Speech coding at 16 kbit/s, 1992
[9]	ITU-T Rec. G.723.1, Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, March 1996
[10]	J. Wroclawski: Specification of the Controlled-Load Network Element Service, IETF RFC 2211, September 1997
[11]	S. Shenker, C. Partridge, R. Guerin: Specification of Guaranteed Quality of Service, IETF RFC 2212, September 1997
[12]	S. Verma, R.K. Pamkaj und A. Leon-Garcia: Call Admission and Resource Reservation for Guaranteed QoS Services in Internet, Computer Communications 21(1998), p. 362-373.
[13]	L.Georgiadis, R.Guerin, V.Peris, R.Rajan: Efficient Support of Delay and Rate Guarantees in an Internet, ACM SIGCOMM '96
[14]	H. Zhang: Service Disciplines for Guaranteed Performance Service in Packet Switching Networks. Proc. IEEE. S. 1373-1396, October 1995.
[15]	A.K. Parekh and R.G. Gallager: A generalized processor sharing approach to flow control in integrated services networks: The single node case. IEEE/ACM Trans. on Networking, 2(2): 137-150, April 1994