Packets scheduling with a new concept of Work Progress Estimation (WPE)

Mudassir Tufail
INRIA/IRISA Rennes
Campus de Beaulieu, 35042, France.

Bernard Cousin
INRIA/IRISA Rennes
Campus de Beaulieu, 35042, France.

Abstract

Scheduling schemes are judged on three important properties: delay bound, fairness and complexity. Recently proposed queue service schemes [4, 5] though guarantee an optimal delay bound and an optimal fairness to a connection but are still delicate in implementation [3]. We propose a new scheduling scheme which works on two novel ideas: 1) work-based packet tags rather than service finish (or start) time-based tags, 2) session associated virtual time function which is not monotonic (unlike system associated time functions as proposed in earlier schemes) and hence is numerically bounded. The proposed scheme determines the session's relative work progress with reference to that in GPS model so is named as Work Progress Estimation (WPE) scheme. The simulation results, prove that the proposed WPE scheme provides an optimal delay bound, an optimal fairness and low implementation cost.

Key words: fair queuing, QoS, fairness, complexity, GPS model

1 Introduction

The Generalized Processor Sharing (GPS) model is based on an ideal fluid model where the packets are not transmitted one after another. A GPS model server serving N_t sessions is characterized by N_t positive real numbers, $\phi_1^*, \phi_2^*, \ldots, \phi_{N_t}^*$. The server operates at a fixed rate r and is work-conserving[1]. Let $W_{i,GPS}[t^*, t]$ be the amount of session i traffic served in the interval $[t^*, t]$, then a GPS model server is defined for which

$$\frac{W_{i,GPS}[t^*,t]}{W_{j,GPS}[t^*,t]} \geq \frac{\phi_i^*}{\phi_j^*} \qquad j = 1, 2, \ldots, N_t \qquad (1)$$

holds for any session i that is backlogged throughout the interval $[t^*, t]$.

[1] A server is said to be work-conserving if it does not stay idle unless all the sessions' queues are empty.

In GPS fluid flow model, the slope of the work progress curve, $W_{i,GPS}[t^*, t]$, for a session i is always greater than 0 at all instants during the interval $[t^*, t]$. On the other hand, in a real scheduling scheme s, the slope is, at times, equals to 0 even if the session i is backlogged for all instants. A backlogged session's service, under a scheme s, may lead or lag, in terms of its total number of bits transmitted until a given instant t, by its reference behavior under corresponding GPS model, refer figure 1. Notice that the s scheme server services the packets from session i during interval [1,2] and [4,6] whereas for intervals [0,1] and [2,4], there is no service offered to the session hence the work progress curve's slope equals to 0. We mean, by Work Progress Estimation (WPE), calculating the session's service lead/lag at a given instant and the queue service scheme, employing the WPE based packet stamp values, is hereby named as Work Progress Estimation fair queuing scheme.

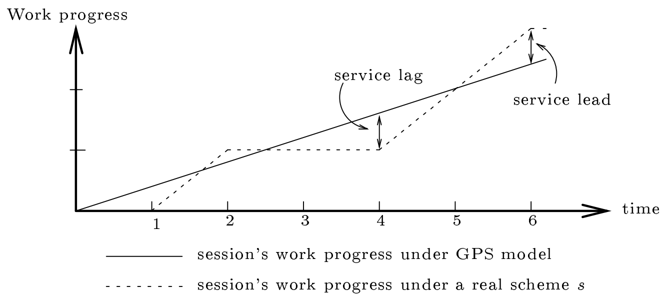

Figure 1: Session's work progress curve

1.1 Review of previous work

There are three important properties of a queue service scheme: tight delay bound, low complexity and fairness. The characteristics of some of the proposed scheduling schemes have been analyzed in [3] are shown in table 2. The schemes which do not define a

a_i^k	arrival time of the k^{th} packet on session i
$d_{i,s}^k$	departure time of the k^{th} packet on session i in the s scheme
$b_{i,s}^k$	service start time of the k^{th} packet on session i in the s scheme
$W_{i,s}[t_1, t_2]$	the amount of work received by the session i during the interval $[t_1, t_2]$ in the s scheme
L_i^k	size of the k^{th} packet on session i
$L_{i,max}$	the maximum packet size on session i
L_{max}	the maximum packet size
$B_s(\tau)$	the set of backlogged sessions at the time τ in the s scheme
r	link speed
ϕ_i^*	positive real number defining service share of session i
ϕ_i	proportional service share of session i
N_t	total number of sessions
N	number of backlogged sessions, N $\leq N_t$
$stamp_{i,s}^k$	the stamp value of a packet p_i^k in the s scheme
$v_{i,s}(\tau)$	the session's virtual time, at a given instant τ in the s scheme
$h_i(\tau)$	the sequence number of the packet at the head of session i's queue at a given instant τ
$\Delta_i^{k,T}(\tau)$	the service lag/lead of the packet p_i^k, at a given instant τ for the time interval T

Table 1: Notations used in this article

self computed virtual time function are marked with '-' in their corresponding entry in this table. WF^2Q+, though improved in complexity, is still delicate in implementation and has following complications: 1) sorting sessions with respect to stamp values of packets at the head of their respective queues along with their eligibility constraint leads to a delicate implementation of buffers [3], 2) computation of system's virtual time requires to know the minimum packet's service virtual start time, among the packets at respective queue heads waiting for service, and thus incurs a complexity[2] of O(log N). In most of the previously proposed schemes, packets are stamped with a time value indicating, virtually or really, its service start or finish time in GPS model and the stamp values do not get evolved with work progress of corresponding session in GPS model. Moreover, monotonic behavior of stamp values render them numerically boundless. In our proposed scheme, we stamp a session's packet with a value which measures how much the session's work, in real system, is lagging or leading by its work progress in corresponding GPS model without simulating it.

2 Work progress estimation

The $\Delta_i^{k,T}(t)$ determines the relative work progress, rendered to k^{th} packet of i^{th} session, in real scheduling scheme with reference to that in corresponding GPS model.

[2]All other queue service schemes, which define their own virtual time function, have complexity of O(1) for computing the virtual time at a given instant, refer table 2.

Definition 1 *We define $\Delta_i^{k,T}(t)$ for packet k of session i if served during interval T by a server of scheme s as:*

$$\Delta_i^{k,T}(t) = \int_t^{t+T} (W_{i,s}(t) - W_{i,GPS}(t))dt \quad \text{where } t = b_{i,s}^k \tag{2}$$

where T is an interval sufficient enough to transmit fully the packet k and is given by $T = \frac{L_i^k}{r}$.

scheme	delay	fairness	complexity	
			virtual t	over all
WFQ [1]	optimal	factor 2	-	O(N)
SCFQ [7]	f(N)	factor 2	O(1)	O(logN)
SFQ [6]	f(N)	factor ≤ 2	O(1)	O(logN)
FFQ [2]	optimal	factor ≥ 3	O(1)	O(logN)
LFVC [8]	optimal	factor 8	O(1)	O(logN)
WF^2Q [4]	optimal	optimal	-	O(N)
WF^2Q+ [5]	optimal	optimal	O(logN)	O(logN)

Table 2: Characteristics of queue service schemes

During interval $b_{i,GPS}^k \leq t \leq d_{i,GPS}^k$ (when a packet p_i^k is being served in GPS model) the slope of $\Delta_i^{k,T}(t)$ curve for a packet p_i^k is given by $-L_i^k\phi_i$. Thus the equation of the curve may be written as:

$$\Delta_i^{k,T}(t) = -L_i^k \phi_i t + \frac{(L_i^k)^2}{2r}(1-\phi_i) \quad \text{at } b_{i,GPS}^k \;\; t=0 \tag{3}$$

In above equation 3, we assume that system time t is initiated at the time when the packet p_i^k starts receiving service in corresponding GPS model. Simulating GPS model at high switching speed increases, significantly, the computational cost of the scheduling method. In order to render WPE scheduling scheme independent of continuous simulation of GPS model, we define in, the following, two important session associated variables:

Definition 2 $\overline{V}_i$ *(if positive) measures the time interval for which session i stays, eventually, out of competition for any further service by WPE scheme server. It is updated at each instant* $t = b_{i,WPE}^k$ *when a packet* p_i^k *is picked up for service in WPE scheme.*

Definition 3 $v_{i,WPE}(t)$ *measures the time instants by which the system time t lags or leads the instant* $b_{i,GPS}^{h_i(t)}$ *for a packet* $p_i^{h_i(t)}$. $v_{i,WPE}(t)$ *is session associated virtual time function in WPE scheme and may has positive or negative value.*

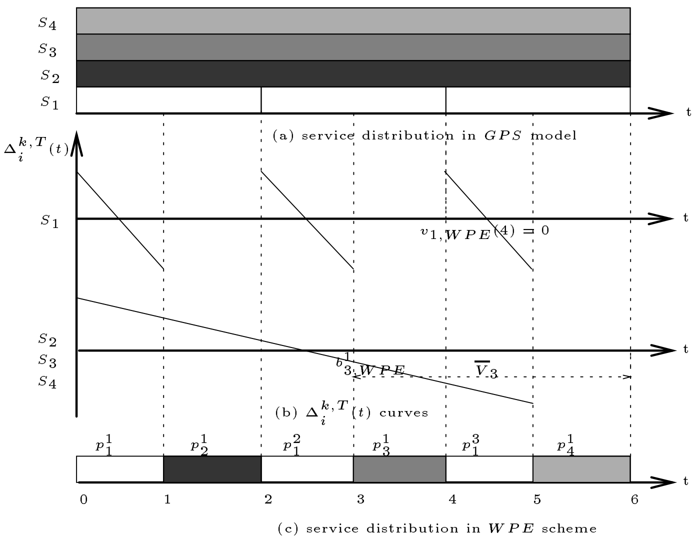

Figure 2: An example of *WPE* service order

2.1 WPE working principles

WPE scheme proceeds as follows at each time instant t when the server has to select a packet among those belonging to backlogged sessions.

1. If the session is newly backlogged or the packet p_i^k, at the head of the queue, is the first one to transmit then initialize $v_{i,WPE}(t)$ as:

$$v_{i,WPE}(t) = t - a_i^1 \tag{4}$$

else initialize $v_{i,WPE}(t)$ as:

$$v_{i,WPE}(t) = min(t - a_i^k, t - (b_{i,WPE}^{k-1} + \overline{V}_i)) \tag{5}$$

where $\overline{V}_i$ is updated at each instant when a packet from the session i is selected for service by WPE scheme server. Since the previously served packet for session i is p_i^{k-1} thus $\overline{V}_i$ is given by:

$$\overline{V}_i = \frac{L_i^{k-1}}{r_i} - v_{i,WPE}(b_{i,WPE}^{k-1}) \tag{6}$$

where $r_i = \frac{\phi_i^*}{\sum_{j \in B_{WPE}(b_{i,WPE}^{k-1})} \phi_j^*} r$.

2. At time t, calculate $\Delta_i^{k,T}(t)$ following equation 3 as:
if $v_{i,WPE}(t) \geq 0$

$$\Delta_i^{k,T}(t) = -L_i^k \phi_i v_{i,WPE}(t) + \frac{(L_i^k)^2}{2r}(1-\phi_i) \tag{7}$$

else

$$\Delta_i^{k,T}(t) = -L_i^k \phi_i v_{i,WPE}(t) + \frac{(L_i^k)^2}{2r}(1-\phi_i) + \frac{N_t (L_{max})^2}{2r_{min}} \tag{8}$$

3. Stamp the packets with normalized $\Delta_i^{k,T}(t)$ values as:

$$stamp_{i,WPE}^k \leftarrow \Delta_{i,norm}^{k,T}(t) = \frac{1}{\phi_i}(\Delta_i^{k,T}(t) + \frac{(L_i^k)^2}{2r}) \tag{9}$$

4. A packet having the minimum stamp value is selected for service at instant t.

2.2 Example

Consider four sessions of figure 2 which stay backlogged for all instants $0 \leq t < 6$ and all packets arrive at $t = 0$. Following WPE scheme, we calculate $v_{i,WPE}(t)$, $\Delta_i^{k,T}(t)$ and $stamp_{i,WPE}^k$ (or $\Delta_{i,norm}^{k,T}(t)$ at given instant t which, in fact, represents the starting time of a packet slot) and select a packet from the session having the minimum stamp value at that instant. Table 3 shows these variables values and the selected

t	$v_{i,WPE}(t)$				$\Delta_i^{h_i(t),T}(t)$				$stamp_{i,WPE}^{h_i(t)} \leftarrow \Delta_{i,norm}^{h_i(t),T}(t)$				sel-ect-ion
	session				session				session				
	1	2	3	4	1	2	3	4	1	2	3	4	
0	0	0	0	0	0.25	0.42	0.42	0.42	1.5	5.5	5.5	5.5	1
1	-1	1	1	1	2.75	0.25	0.25	0.25	26.5	4.5	4.5	4.5	2
2	0	-4	2	2	0.25	3.08	0.08	0.08	1.5	81.5	3.5	3.5	1
3	-1	-3	3	3	2.75	2.92	-0.08	-0.08	26.5	80.5	2.5	2.5	3
4	0	-2	-2	4	0.25	2.75	2.75	-0.25	1.5	79.5	79.5	1.5	1
5	-1	-1	-1	5	2.75	2.58	2.58	-0.42	26.5	78.5	78.5	0.5	4

Table 3: service distribution in WPE scheme

session at each instant t. After having been selected the session i for service at the given instant, we update its $\overline{V}_i$ value following the equation 6. Notice that we do not need to know the $\overline{V}_i$ value if the packet at the head of session i's queue be the first one. We evaluate 6 packet slots for 4 backlogged sessions following the WPE scheme and end up, eventually, with service distribution as shown in figure 2.c.

3 Proof

In the following we prove that the WPE scheme and the WF^2Q scheme produce the same service order for a given set of backlogged sessions provided that the implementation policy for tie-breaking situations is same for both the schemes. As per equation 9, we have:

$$\Delta_{i,norm}^{k,T}(t) = \frac{1}{\phi_i}(\Delta_i^{k,T}(t) + \frac{(L_i^k)^2}{2r}) \qquad (10)$$

Putting the $\Delta_i^{k,T}(t)$ value from equation 7 in above equation, we get:

$$\begin{aligned}\Delta_{i,norm}^{k,T}(t) &= \frac{1}{\phi_i}(-L_i^k\phi_i v_{i,WPE}(t) + \frac{(L_i^k)^2}{r} - \frac{(L_i^k)^2}{2r}\phi_i) \\ &= L_i^k(-v_{i,WPE}(t) + \frac{L_i^k}{r\phi_i} - \frac{L_i^k}{2r}) \qquad (11)\end{aligned}$$

where $\phi_i = \frac{\phi_i^*}{\sum_{j\in B_{WPE}(t)}\phi_j^*}$ with $j = 1,\dots N_t$. So if the set of backlogged sessions $B_{WPE}(t)$ at time t changes, the proportional service share ϕ_i of session i is also changed. It can be written that:

$$r_i = r\phi_i \qquad (12)$$

Substituting the r_i value in above equation 11,

$$\Delta_{i,norm}^{k,T}(t) = L_i^k(-v_{i,WPE}(t) + \frac{L_i^k}{r_i} - \frac{L_i^k}{2r}) \qquad (13)$$

According to the definition of $v_{i,WPE}(t)$ (equations 4 and 5), a session i has $v_{i,WPE}(t) = \frac{L_i^k}{r_i}$ at packet slot time t when its packet p_i^k departs (or finishes service) in GPS model. We can write the $v_{i,WPE}(t)$ values at all the precedent instants, back-spaced by packet slot intervals, as $(\frac{L_i^k}{r_i} - \frac{L_i^k}{r}), (\frac{L_i^k}{r_i} - \frac{2L_i^k}{r}), \dots$. For different $v_{i,WPE}(t)$ values, we may rewrite the above equation 13 as, provided that $v_{i,WPE}(t) \geq 0$:

$$\begin{aligned}\Delta_{i,norm}^{k,T}(t) &= -\frac{(L_i^k)^2}{2r} \qquad \text{if } v_{i,WPE}(t) = \frac{L_i^k}{r_i} \\ &= \frac{(L_i^k)^2}{2r} \qquad \text{if } v_{i,WPE}(t) = \frac{L_i^k}{r_i} - \frac{L_i^k}{r} \\ &= \frac{3(L_i^k)^2}{2r} \qquad \text{if } v_{i,WPE}(t) = \frac{L_i^k}{r_i} - \frac{2L_i^k}{r} \\ &= (\beta - \frac{1}{2})\frac{(L_i^k)^2}{r} \text{ if } v_{i,WPE}(t) = \frac{L_i^k}{r_i} - \frac{\beta L_i^k}{r}\end{aligned}$$

where $\beta \geq 0$. The above result indicates that $\Delta_{i,norm}^{k,T}(t)$ (which is actually the $stamp_{i,WPE}^k$ value) is function of β and L_i^k for all the sessions which have started their service in corresponding GPS model[3] at the given instant t. As the β value approaches zero (making consequently the $\Delta_{i,norm}^{k,T}(t)$ value lesser) $v_{i,WPE}(t)$ value approaches $\frac{L_i^k}{r_i}$ which is the $v_{i,WPE}(t)$ of the session i at time t when the packet departs in corresponding GPS model. Hence selecting a packet with minimum $\Delta_{i,norm}^{k,T}(t)$ value, at a given instant, in WPE scheme means looking for a packet, among eligible ones, that finishes its service first in corresponding GPS model, which is how the WF^2Q scheme works.
For the sessions which have $v_{i,WPE}(t) < 0$ at a given instant t are considered ineligible sessions in WF^2Q scheme. In WPE scheme, these sessions, however, participate in competition but are not selected. This is ensured by an addition of a constant

[3] Recall that this result is valid for sessions with $v_{i,WPE}(t) \geq 0$ and referred as eligible sessions in WF^2Q scheme.

$\frac{N_t(L_{max})^2}{2r_{min}}$, refer equation 8. As evident from equation 7 that for $v_{i,WPE}(t) \geq 0$, the $\Delta_i^{k,T}(t) < \frac{(L_i^k)^2}{2r}$ $\forall i$ whereas for $v_{i,WPE}(t) < 0$ the $\Delta_i^{k,T}(t)$ can never exceed $\sum_{i=1}^{N_t}$ where $_{i \in B_{WPE}(t)}$ $\frac{(L_{i,max})^2}{2r}$. Since $B_{WPE}(t)$ is a subset of all sessions at the server, thus addition of constant $\frac{N_t(L_{max})^2}{2r_{min}}$ ensures that a session with $v_{i,WPE}(t) < 0$ gets $\Delta_i^{k,T}(t)$ value sufficiently large that it does not have a the minimum $\Delta_{i,norm}^{k,T}(t)$ value at the instant t and thus is not selected by WPE scheme for service.

We conclude this section by the following two important remarks: 1) among the eligible sessions (as per WF^2Q scheme), the WPE and WF^2Q scheme select, at a given instant, the same session's packet for a given set of backlogged sessions, 2) the session(s) declared ineligible, at a given instant, as per WF^2Q scheme participate in competition but is(are) not selected for service at that instant in WPE scheme. Thus the service orders produced by WPE and WF^2Q schemes are exactly the same which means that WPE scheme has an optimal delay bound and an optimal fairness as those of WF^2Q scheme.

4 Simulation

Four queue service schemes: WF^2Q, WF^2Q+, SCFQ and WPE, are simulated in MatLab. These implementations are based on per session queue configuration[4].

4.1 Pseudocode

Considering the implementation in an ATM switch, the packet size becomes constant and may be written as $L_i^k = L$, where L means 53 bytes. Note that, that two operations: equations 7 + 9 or 8 + 9 according to the case, are merged in a single respective instruction in pseudocode.

$C = \frac{L^2}{2r} \leftarrow \frac{(L_i^k)^2}{2r}$

At packet arrival:

No operation except directing the packet towards the concerned queue.

Calculation of stamp values at time t:

for i=1 to N

  if $h_i(t) == 1$

    $v_{i,WPE} = t - a_i^1$

  else

    $v_{i,WPE} = min(t - a_i^{h_i(t)}, t - (b_{i,WPE}^{h_i(t)-1} + \overline{V}_i))$

  end

  if $v_{i,WPE} \geq 0$

    $stamp_{i,WPE} = -Lv_{i,WPE} + C(\frac{2}{\phi_i} - 1)$

  else

    $stamp_{i,WPE} = -Lv_{i,WPE} + C(\frac{2+N_t}{\phi_i} - 1)$

  end

end

Session selection:

Let the selected session be ss where $ss \in B_{WPE}(t)$

$stamp_{ss,WPE} = min(stamp_{1,WPE}, \ldots, stamp_{N,WPE})$

Session's packet departure:

$b_{ss,WPE}^{h_{ss}(t)} = t$

$\overline{V}_{ss} = \frac{L}{r_{ss}} - v_{ss,WPE}$

4.2 Results

The main objective behind the development of WPE scheme is to reduce the latency of queue service scheme which means the time taken by the scheme server to decide which session's packet to be served at the given instant t. This can only be achieved if the number of operations may be reduced. More are the number of operations in a queue service scheme more is the time taken by CPU to determine the packet to be served at a given time. In other words the latency of a queue service scheme reflects its implementation cost. The WF^2Q scheme requires a continuous simulation of GPS model in the background thus has large delay per packet which increases significantly if there

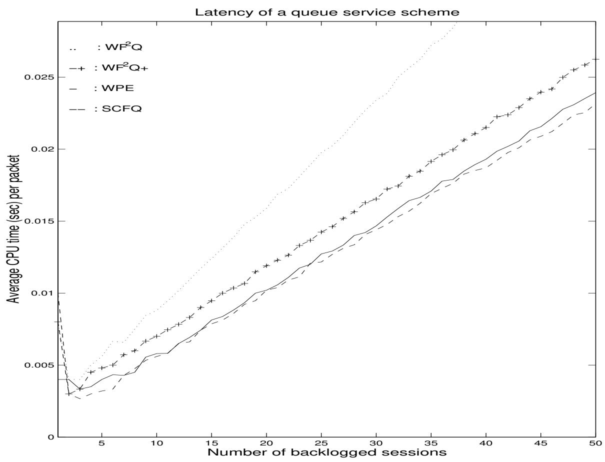

Figure 3: Comparative scheduling delays

[4]The working principles of these four scheduling scheme are independent of queues configuration which may be either one queue per session at an output port or one queue per output port.

are more sessions backlogged, refer figure 3. All other three schemes (WF^2Q+, WPE and SCFQ), contrarily to WF^2Q scheme, have close latency curve slopes, figure 3, which do not get increased abruptly when the number of backlogged sessions is large[5]. Notice that the latency curves of WPE and SCFQ schemes are very close and overlap each other at times. On the other hand, scheduling delay of WF^2Q+ scheme is remarkably greater than that of SCFQ scheme.

The scheduling delays per packet depends upon the number of backlogged sessions at a given instant. In the figure 3, the maximum number of backlogged sessions is limited to only fifty[6] and the latency curves of WPE and WF^2Q+ curves seem to progress with approximately same slope value which, in fact, is not the case. In the figure 4, the maximum number of backlogged sessions is relatively more significant and the schemes simulated are only WF^2Q+ and WPE as to avoid very large simulation times. The latency curves in figure 4 show that as the number of backlogged session increases the scheduling delays per packet incurred by two schemes: WPE and WF^2Q+, differ significantly.

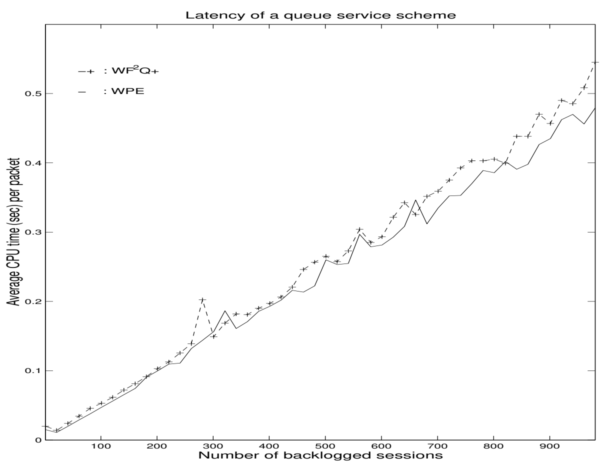

Figure 4: Comparative scheduling delays

5 Conclusion

We define a new concept of packet relative stamp values which are based upon the corresponding session's relative work progress with reference to that in GPS model. In this context we define $\Delta_i^{k,T}(t)$ which estimates the work progress. The packets stamp values are extracted from normalized value of $\Delta_i^{k,T}(t)$. These values, being relative, do not continue increasing with system time thus are finite. The finite nature of stamp values helps to eliminate the session's eligibility check from the WPE scheme and lowers its implementation cost. Another important issue is the constant background simulation of GPS model which incurs a high implementation cost. WPE scheme uses sessions associated virtual time function $v_{i,WPE}(t)$ which helps in determining the session's work progress in corresponding GPS model without simulating it. The $v_{i,WPE}(t)$ measure the time units by which the system time t lags or leads the instant $b_{i,GPS}^k$ for the packet p_i^k at the head of session i's queue. It may have positive or negative values and is numerically bounded. WPE scheme has been proved to guarantee an optimal delay bound and an optimal fairness to backlogged session with a lower implementation cost.

[5]This result confirms that the WF^2Q+ and SCFQ schemes are of the same complexity orders, refer table 2.

[6]Note that ATM has the a theoretical capacity of having as much as 2^{24} sessions per UNI (User Network Interface).

References

[1] Abhay K. Parekh. A Generalized Processor Sharing approach to flow control in integrated services networks. *Ph.D. dissertation, Massachusetts Institute of Technology, February 1992.*

[2] D. Stiliadis, A. Verma. Design and analysis of frame-based fair queuing: A new traffic scheduling algorithm for packet-switched networks. *Proc. ACM SIGMETRICS'96, Philadelphia, PA, 1996.*

[3] Franois Toutain. Gestion premtive et quitable de la qualit de service dans les rseaux de paquets intgration de services (Preemptive and fair QoS management in integrated services packet switched networks). *Ph.D. dissertation, University of Rennes 1, France, July 1997.*

[4] Jon C.R. Bennett, Hui Zhang, WF^2Q: Worst-case Fair Weighted Fair Queuing, *IEEE INFOCOM, Mar-96, pp. 120-128.*

[5] Jon C.R. Bennett, Hui Zhang, Hierarchical Packet Fair Queuing Algorithms, *SIGCOMM'96, Aug-96.*

[6] P. Goyal, H.M. Vin, H. Cheng. Start-time fair queuing: A scheduling algorithm for integrated services packet switching networks. *Proc. ACM SIGCOMM'96, Stanford, CA, August 1996.*

[7] S. Jamaloddin Golestani, A Self-Clocked Fair Queuing Scheme for Broadband Applications, *IEEE INFOCOM'94, Toronto, CA, June 1994, pp. 636-646.*

[8] S. Suri, G. Varghese and G. Chandranmenon. Leap forward virtual clock: An O(loglog N) fair queuing scheme with guaranteed delays and throughput fairness. *Proc. IEEE Infocom'97, Kobe, Japan, April 1997.*