Scheduling for output queued switches: weighted fair queuing

November 18, 2002

- Queues $1 \ldots N$ sharing a server with capacity C
- Queue i has weight ϕ_i

  While queues i and j are nonempty, the link serves ϕ_i bits from queue i for each ϕ_j bits served from queue j (Intuition)

Assume fluid arrivals

Formal definition: Generalized Processor Sharing (GPS):

Notations: $Z_j(t), Z(t)$

$$Z_j(t) = \begin{cases} 1 & \text{if queue } j \text{ is nonempty} \\ 0 & \text{if queue } j \text{ is empty} \end{cases}$$

$$Z(t) = \begin{cases} 1 & \text{if at least one queue is nonempty} \\ 0 & \text{otherwise} \end{cases}$$

Service rate of queue j at time u is $C\phi_j Z_j(u) / \sum_{i=1}^N \phi_i Z_i(u)$

Cumulative service provided to queue j over the interval $[0,t]$ is $\int_0^t C\phi_j Z_j(u) / \sum_{i=1}^N \phi_i Z_i(u) du$

Average service rate applied to queue j over the interval $[0,t]$ is $\frac{\int_0^t \frac{C\phi_j Z_j(u)}{\sum_{i=1}^N \phi_i Z_i(u)} du}{\int_0^t Z_j(u)du}$

Since $Z_j(u) \leq 1$, the average service rate is lower bounded by

$$\phi_j \frac{\int_0^t \frac{CZ_j(u)}{\sum_{i=1}^N \phi_i} du}{\int_0^t Z_j(u)du}$$

$$\frac{\phi_j}{\sum_{i=1}^N \phi_i} C$$

Average service rate provided to queue j in any interval $[0,t]$ is at least $\frac{\phi_j}{\sum_{i=1}^N \phi_i} C$.

Minimum service curve $S_{\text{GPS}}(t) = \frac{\phi_j}{\sum_{i=1}^N \phi_i} Ct$

$$\textbf{Virtual time: } V(t) = \int_0^t \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du (\text{basic definition})$$

Properties of $V(t)$

- $V(0) = 0$, and $V(t) \geq 0$ for $t \geq 0$
- $V(t)$ is non-decreasing with t
- $V(t)$ is a piece-wise linear function whose slope increases with arrivals and decreases with departures.
- $V(t)$ does not change in the intervals when the system has no data to serve

$a_k(j)$ is the arrival instant of the kth packet in queue j

$L_k(j)$ is the amount of data in the kth packet arriving in queue j

$s_k(j)$ is the service initiation instant of the kth packet arriving in queue j

$d_k(j)$ is the departure instant of the kth packet arriving in queue j

If queue j is served by a dedicated server of capacity c_j, then

$$d_{k+1}^j = \max d_k^j, a_{k+1}^j + \frac{L_{k+1}^j}{c_j}$$

Service rate allocated to a queue changes in GPS

$$
\begin{aligned}
V(d_{k+1}^j) &= \int_0^{d_{k+1}^j} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du \\
&= \int_0^{s_{k+1}^j} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du + \\
& \quad \int_{s_{k+1}^j}^{d_{k+1}^j} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du
\end{aligned}
$$

Since $Z_j(u) = 1$ during (s_{k+1}^j, d_{k+1}^j),

$$\int_{s_{k+1}^j}^{d_{k+1}^j} \phi_j \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du = L_{k+1}^j$$

$$\int_{s_{k+1}^j}^{d_{k+1}^j} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du = L_{k+1}^j / \phi_j$$

Also $s_{k+1}^j = \max(d_k^j, a_{k+1}^j)$ Thus

$$
\begin{aligned}
V(d_{k+1}^j) &= \int_0^{\max(d_k^j, a_{k+1}^j)} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du + L_{k+1}^j / \phi_j \\
&= \max\left(\int_0^{d_k^j} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du, \int_0^{a_{k+1}^j} \frac{C}{\sum_{i=1}^N \phi_i Z_i(u)} I_{Z(u)>0} du \right) + L_{k+1}^j / \phi_j \\
&= \max\left(V(d_k^j), V(a_{k+1}^j) \right) + L_{k+1}^j / \phi_j \text{(recursion equation)}
\end{aligned}
$$

Simulation of GPS scheduler for fluid arrivals

1. The function $V(t)$ is simulated using the basic definition and updated at the arrivals and departure epochs using the recursion relation
2. Virtual finish time of a packet is the virtual time at its departure epoch. The packet is marked with this value when it arrives.
3. When the function $V(t)$ equals the virtual finish time of a packet, it is ejected from the system

Implementation of Weighted Fair Queueing (WFQ) a packetized GPS scheduler:

1. The function $V(t)$ is simulated as before.
2. The packet is marked with its virtual finish time when it arrives just as for GPS.
3. When a packet is selected for transmission, it is transmitted completely.
4. The packet with the smallest virtual finish time is the next packet selected.

Lemma 1 *The service order of PGPS between packets present at the scheduler at an instant is the same as the departure order for GPS among these packets.*

Proof: Consider two packets i and j present at time t in the scheduler.

PGPS selects i if $V(d_i) \leq V(d_j)$.

Since $V(t)$ is a non-decreasing function, $d_i \leq d_j$.

Let $\hat{d}_k$ denote the PGPS departure instant of packet k.

Theorem 1

$$\hat{d}_k \leq d_k + \frac{L_{\max}}{C} \textit{where } L_{\max} \textit{ is the maximum packet length}$$

Proof: Both GPS and PGPS are work-conserving.

Thus both serve the same set of packets each busy period, and the busy periods have the same duration in both.

Thus it is sufficient to show the above result for a single busy period.

Packet departure orders may differ.

Let $p_1, p_2, \ldots$ be packets in order of departure in PGPS.

Let the busy period start at t_0. Let $\hat{d}_k = t_0 + \sum_{i=1}^{k} L_i/C$

If none of the packets $p_1, p_2, \ldots p_{k-1}$ depart after d_k in GPS, $d_k \geq t_0 + \sum_{i=1}^{k} L_i/C = \hat{d}_k$

Let $d_1, \ldots d_m > d_k$ and $d_{m+1}, \ldots \ldots d_{k-1} \leq d_k$

Consider $\hat{s}_m$, the service initiation time for p_m in PGPS.

Clearly $p_{m+1}, \ldots, p_{k-1}$ are not present in the system at $\hat{s}_m$.

Thus $p_{m+1}, \ldots, p_{k-1}$ arrive after $\hat{s}_m$ and depart before d_k

$$\begin{aligned} d_k &\geq \hat{s}_m + \frac{\sum_{i=m+1}^{k} L_i}{C} \\ &= \hat{d}_k - \frac{L_m}{C} \end{aligned}$$

Let $c_j = \frac{\phi_j}{\sum_{i=1}^{N} \phi_i} C$ Then $S_{\mathrm{WFQ}}(t) = c_j C(t - \frac{L_{\max}}{C})^+$

For a peak rate controlled and leaky bucket controlled source with parameters(σ_j, ρ_j, R_j) the delay through a WFQ scheduler is bounded by $\frac{\sigma_j(R_j - c_j)}{c_j(R_j - \rho_j)} + L_{\max} C$

System Illustration:

System consists of a

1. coder which emits bytes at the peak rate of R bytes per second
2. leaky bucket shaper (σ, ρ) with an infinite source buffer
3. packetizer with maximum packet length $L_{\max}$
4. WFQ scheduler with minimum rate r and maximum packet length over all sessions $V_{\max}$

End to end delay of a bit which sees X_1 packets in the source buffer is upper bounded by

$$D = \frac{X_1}{\rho} + \frac{\sigma(R - r)}{r(R - \rho)} + \frac{L_{\max}}{\rho} + \frac{V_{\max}}{C}$$

Let

$$T' = \frac{\sigma(R - r)}{r(R - \rho)} + \frac{L_{\max}}{\rho} + \frac{V_{\max}}{C}$$

$\mathrm{Prob}(D > T) = \mathrm{Prob}(X_1 > (T - T')\rho)$

Computation of $V(t)$

It is possible to update $V(t)$ only at the arrival epochs

Between two arrivals N possible departures and changes in slope.

Thus complexity is $O(N)$

Severe computational bottleneck for large N.

Redress:

- Flow aggregation
- Simplifying the scheduling (SCFQ, STFQ)