*"Oh, well. It still works,*
*It's just not work-conserving!".*

— The Art of Chutzpah[†]

# F

# Proofs for Chapter 6

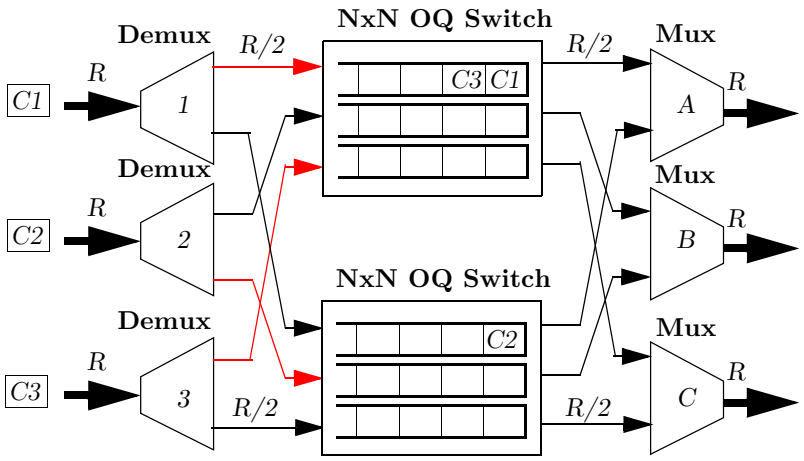
## F.1 Proof of Theorem 6.1

**Theorem 6.1.** *A PPS without speedup is not work-conserving.*

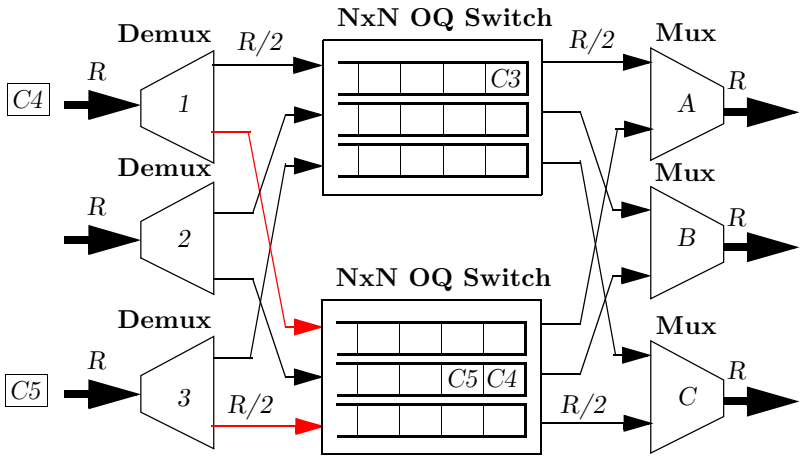
*Proof.* (By counter-example). Consider the PPS in Figure F.1 with three ports and two layers ($N = 3$ and $k = 2$). The external lines operate at rate $R$, and the internal lines at rate $R/2$.

Assume that the switch is empty at time $t = 1$, and that three cells arrive, one to each input port, and all destined to output port $A$. If all the input ports choose the same layer, then the PPS is non-work-conserving. If not, then at least two of these inputs will choose the same layer and the other input will choose a different layer. Without loss of generality, let inputs 1 and 3 both choose layer 1 and send cells $C1$ and $C3$ to layer 1 in the first time slot. This is shown in Figure F.1(a). Also, let input port 2 send cell $C2$ to layer 2. These cells are shown in the output queues of the internal switches and await departure. Now we create an adversarial traffic pattern. In the second time slot, the adversary picks the input ports that sent cells to the same layer in the first time slot. These two ports are made to receive cells destined to output port $B$. As shown in the figure, cells $C4$ and $C5$ arrive at input ports 1 and 3, and they both must be sent to layer 2; this is because the internal line rate between

[†]HPNG Group Meeting, Stanford University, California, Jan 1999.




**(a) Time Slot 1 (C1, C2, C3 arrive for output A)**




**(b) Time Slot 2 (C4, C5 arrive for output B)**

**Figure F.1:** *A* $3 \times 3$ *PPS with an arrival pattern that makes it non-work-conserving. The notation* $Ci : A, m$ *denotes a cell numbered* $i$*, destined to output port* $A$*, and sent to layer* $m$*.*

the demultiplexor and each layer is only $R/2$, limiting a cell to be sent over this link only once every other time slot. Now the problem becomes apparent: cells $C4$ and $C5$ are in the same layer, and they are the only cells in the system destined for output port $B$ at time slot 2. These two cells cannot be sent back-to-back in consecutive time slots, because the link between the layer and the multiplexor operates only at rate $R/2$. So, cell $C4$ will be sent, followed by an idle time slot at output port $B$, and the system is no longer work-conserving. And so, trivially, a PPS without speedup cannot emulate an FCFS-OQ switch. ❒

## F.2 Proof of Theorem 6.5

In what follows, we will use $T$ to denote time in units of time slots. We will also use $t$ to denote time, and use it only when necessary. Recall that if the external line rate is $R$ and cells are of fixed size $P$, then each cell takes $P/R$ units of time to arrive, and $t = TP/R$. Before we prove the main theorem, we will need the following results.

**Lemma F.1.** *The number of cells $D(i,l,t)$ that demultiplexor $i$ queues to FIFO $Q(i,l)$ in time $T$ slots, is bounded by*

$$D(i,l,T) \leqslant T \quad \text{if} \quad T \leqslant N$$
$$D(i,l,T) < \frac{T}{k} + n \quad \text{if} \quad T > N.$$

*Proof.* Since the demultiplexor dispatches cells in a round robin manner for every output, for every $k$ cells received by a demultiplexor for a specific output, exactly one cell is sent to each layer. We can write $S(i,T) = \sum_{j=1}^{N} \overline{S}(i,j,T)$, where $\overline{S}(i,j,T)$ is the sum of the number of cells sent by the demultiplexor $i$ to output $j$ in any time interval of $T$ time slots, and $S(i,T)$ is the sum of the number of cells sent by the

demultiplexor to all outputs in that time interval $T$. Let $T > N$. Then we have,

$$D(i,l,T) \leqslant \sum_{j=1}^{N} \left\lceil \frac{\overline{S}(i,j,T)}{k} \right\rceil \leqslant \left\lceil \sum_{j=1}^{N} \frac{\overline{S}(i,j,T)}{k} \right\rceil + N + 1 = \left\lceil \frac{S'(i,T)}{k} \right\rceil + N - 1 \leqslant \left\lceil \frac{T}{k} \right\rceil + N - 1 < \frac{T}{k} + N \quad \text{(F.1)}$$

since $S(i,T)$ is bounded by $T$. The proof for $T \leqslant N$ is obvious. ❒

We are now ready to determine the size of the co-ordination buffer in the demultiplexor.

**Theorem F.1.** *(Sufficiency) A PPS with independent demultiplexors and no speedup can send cells from each input to each output in a round robin order with a co-ordination buffer at the demultiplexor of size $Nk$ cells.*

*Proof.* A cell of size $P$ corresponds to $P/R$ units of time, allowing us to re-write Lemma F.1 as $D(i,l,t) \leqslant Rt/Pk + N$ (where $t$ is in units of time). Thus the number of cells written into each demultiplexor FIFO is bounded by $Rt/Pk + N$ cells over all time intervals of length $t$. This can be represented as a leaky bucket source with an average rate $\rho = R/Pk$ cells per unit time and a bucket size $\sigma = N$ cells for each FIFO. Each FIFO is serviced deterministically at rate $\mu = R/Pk$ cells per unit time. Hence, by the definition of a leaky bucket source [219], a FIFO buffer of length $N$ will not overflow. ❒

It now remains for us to determine the size of the co-ordination buffers in the multiplexor. This proceeds in an identical fashion.

**Lemma F.2.** *The number of cells $D'(j,l,T)$ that multiplexor $j$ delivers to the external line from FIFO $Q'(j,l)$[1] in a time interval of $T$ time slots, is bounded by*

$$D'(i,l,T) \leqslant T \quad \text{if} \quad T \leqslant N$$
$$D'(i,l,T) < \frac{T}{k} + n \quad \text{if} \quad T > N.$$

[1]FIFO $Q'(j,l)$ holds cells at multiplexor $j$ arriving from layer $l$.

*Proof.* Cells destined to multiplexor $j$ from a demultiplexor $i$ are arranged in a round robin manner, which means that for every $k$ cells received by a multiplexor from a specific input, exactly one cell is read from each layer. Define,

$$S'(j,T) = \sum_{j=1}^{N} \overline{S'}(i,j,T), \tag{F.2}$$

where $\overline{S'}(i,j,T)$ is the sum of the number of cells from demultiplexor $i$ that were delivered to the external line by multiplexor $j$ in time interval $T$, and $\overline{S'}(i,T)$ is the sum of the number of cells from all the demultiplexors that were delivered to the external line by the multiplexor in time interval $T$. Let $T > N$. Then we have,

$$D'(i,l,T) \leqslant \sum_{j=1}^{N} \left\lceil \frac{\overline{S'}(i,j,T)}{k} \right\rceil \leqslant \left\lceil \sum_{j=1}^{N} \frac{\overline{S'}(i,j,T)}{k} \right\rceil + N + 1 = \left\lceil \frac{S(i,T)}{k} \right\rceil + N - 1 \leqslant \left\lceil \frac{T}{k} \right\rceil + N - 1 < \frac{T}{k} + N \tag{F.3}$$

since $S'(i,T)$ is bounded by $T$. The proof for $T \leqslant N$ is obvious. ❒

Finally, we can determine the size of the co-ordination buffers at the multiplexor.

**Theorem F.2.** *(Sufficiency) A PPS with independent multiplexors and no speedup can receive cells for each output in a round robin order with a co-ordination buffer of size $Nk$ cells.*

*Proof.* The proof is almost identical to Theorem F.1. From Lemma F.2, we can bound the rate at which cells in a multiplexor FIFO need to be delivered to the external line by $Rt/Pk + N$ cells over all time intervals of length $t$. Cells are sent from each layer to the multiplexor FIFO at fixed rate $\mu = R/Pk$ cells per unit time. We can see as a result of the *delay equalization* step in Section 6.8.2 that the demultiplexor and multiplexor systems are exactly symmetrical. Hence, if each FIFO is of length $N$ cells, the FIFO will not overflow. ❒

Now that we know the size of the buffers at the input demultiplexor and the output multiplexor – both of which are serviced at a deterministic rate – we can bound the relative queueing delay with respect to an FCFS-OQ switch.

**Theorem 6.5.** *(Sufficiency) A PPS with independent demultiplexors and multiplexors and no speedup, with each multiplexor and demultiplexor containing a co-ordination buffer of size $Nk$ cells, can emulate an FCFS-OQ switch with a relative queuing delay bound of $2N$ internal time slots.*

*Proof.* We consider the path of a cell in the PPS. The cell may potentially face a queuing delay as follows:

1. The cell may be queued at the FIFO of the demultiplexor before it is sent to its center stage switch. From Theorem F.1, we know that this delay is bounded by $N$ internal time slots.
2. The cell first undergoes delay equalization in the center stage switches and is sent to the output queues of the center stage switches. It then awaits service in the output queue of a center stage switch.
3. The cell may then face a variable delay when it is read from the center stage switches. From Theorem F.2, this is bounded by $N$ internal time slots.

Thus the additional queuing delay, *i.e.*, the relative queueing delay faced by a cell in the PPS, is no more than $N + N = 2N$ internal time slots.

Note that in the proof described above, it is assumed that the multiplexor is aware of the cells that have arrived to the center stage switches. It issues the reads in the correct FIFO order from the center stage switches, *after* they have undergone delay equalization. This critical detail was left out in the original version of the paper [36], leading to some confusion with subsequent work done by the authors in [128]. On discussion with the authors [220], we concurred that our results are in agreement. Their detailed proof appears in [128]. ❐