5.9 Norton’s Theorem for Closed Queueing Networks

Norton’s Theorem for the analysis of closed queueing networks draws its inspiration from its analogy with Norton's Theorem in the analysis of electrical circuits where a complex circuit is compactly represented as a current source with parallel impedance driving the load impedance. The current source and its parallel impedance then represents the rest of the circuit other than the load impedance whose effects are required to be studied, i.e. the voltage across the load and the current through it.

Norton’s Theorem for closed networks performs an essentially similar function. It is basically a technique to reduce a closed queueing network with K FCFS exponential service queues and M jobs/customers circulating in the network so that the performance of one of the queues (any queue in the network) or the performance of a sub-network of the queues may be easily studied. Following this method, a smaller equivalent network may be obtained by replacing all queues except those in a designated sub-network by a single *Flow Equivalent Server (FES)*. The authors (Chandy, Herzog and Woo) of this method [CHW75] show that for certain types of system parameters, the behaviour of the equivalent network will be exactly the same as that of the original network - this is also known as the Chandy-Herzog-Woo Theorem and is illustrated next. This approach may also be extended as an approximation to a closed network of FCFS queues with general service times. It can also be extended to networks with other service disciplines and may also be used for networks that have several classes of customers. In this section we will, however, limit our discussion to the simple, closed queueing networks with a single class of jobs, which have exponentially distributed service times and probabilistic routing.

In order to illustrate the application of Norton’s Theorem to such a closed queueing network consider the example network of Figure 5.13 where we have identified the queue Q_i as the queue whose performance needs to be studied. (Note that we could have also considered a sub-network of queues to be studied instead of just one queue Q_i.)

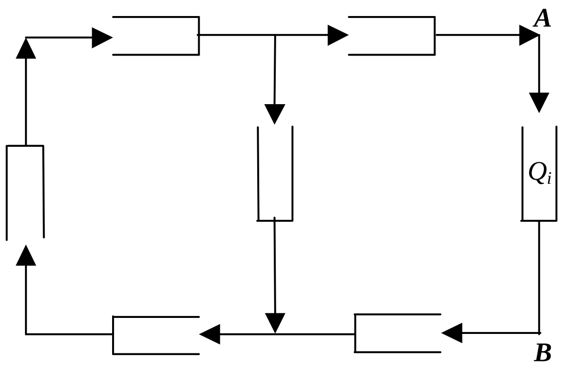

Figure 5.13. Original Queueing Network Before Reduction

Consider the situation where we want to characterize the queue Q_i in different ways and see how this affects the performance of the queue and the throughput of the overall network. This, for example, may be done by changing the service rates at Q_i. Obtaining the queueing statistics at Q_i for different characterizations of it will be simplified if the rest of the closed queueing network can be given a more compact representation - preferably as a single queue. The technique of Norton Reduction may be applied to this network to obtain such a compact representation of the rest of the queueing network other than the queue Q_i. The final objective of this objective of this reduction will be to obtain the network given in Figure 5.14 where the sub-network other than Q_i is replaced by a *single queue with a state dependent service rate m(j)* where *j* is the number of jobs in that queue. This queue is also referred to as a *Flow Equivalent Server (FES)* representation of the equivalent queue replacing the rest of the network (other than Q_i). The reason for calling it a flow equivalent server is because it represents that sub-network of queues by a single queue, which is equivalent in terms of the overall flow from that sub-network. It should be noted, however, that the actual nature of the equivalence will depend on the number of jobs that one considers the closed network to have - the equivalent service rate *m(j)* of the flow equivalent server will vary depending on the number of jobs circulating in the network.

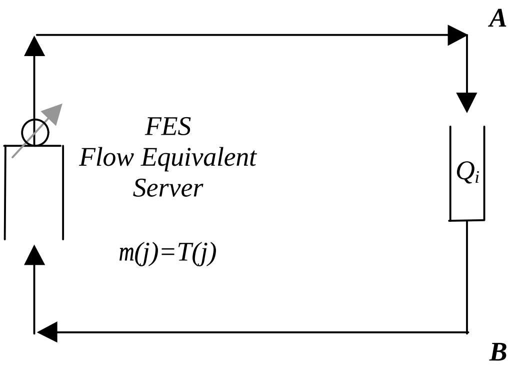

Figure 5.14. Equivalent Network with Flow Equivalent Server

The flow rate *m(j)* of the FES may be calculated as *T(j)* where this is obtained by shorting A and B and evaluating *T(j)* as the throughput between these points when there are *j* jobs circulating in that modified network. This is exactly analogous to way the "short-circuit" current is evaluated in applying Nortons' Theorem to an electronic circuit where the strength of the equivalent current source is taken to be the value of this short-circuit current. This is illustrated in more detail in Figure 5.15. As shown in the figure, the flow rate of the FES is calculated by removing the queue Q_i of interest and connecting its two end points A and B directly, i.e. by short-circuiting the queue. This may also be done by making the service time of Q_i to be zero. The flow between A and B will then be the network's throughput and is calculated as *T(j)* when there are *j* jobs circulating in this modified network.

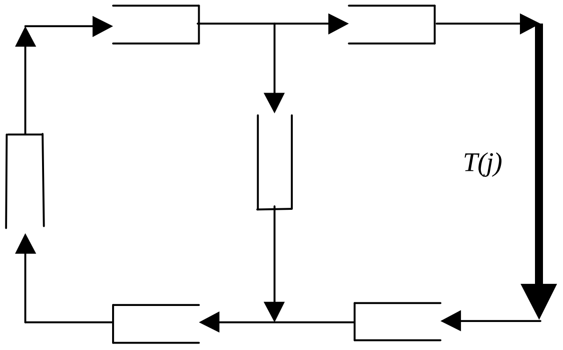

Figure 5.15. Network to Obtain the Flow Rate *T(j)* of the FES with *j* Jobs in the Network

In general, the portion of the network that can be isolated is not limited to a single queue and can be a network of queues. Following usual terminology, such a network of queue is called the *designated network* and the remaining network is called the *aggregate*. The aggregate network is reduced to a single *FEC* and the performance of the designated network is studied for variations in the parameters of the designated network. This hierarchical decomposition produces exact results for a large class of networks - i.e. the ones typical referred to as *BCMP Networks* [BCM75] where local balance conditions hold. The procedure is summarized below.

1. Select the *designated sub-network of queues* from the original network that is to be studied. The remaining network is the *aggregate network* to be reduced to a single FES.

2. Generate a queueing network in which the service times at all the queues in the designated network is set to zero, i.e. this is the process of shorting them. Note that the designated network should be selected such that the throughput through all the shorts in the new network should be identical.

3. Solve the above network using any of the known techniques. Solve this for all possible values of the network population, i.e. *j=1,...., M*. The throughput through the short for different populations correspond to the service rate of the FES with that number of jobs in the queue, i.e. *T(j)*.

4. The service rates of the equivalent FES are now available for different values of the number of jobs *j* circulating in the network. We now consider the equivalent network with the designated network and the FES where the FES replaces the aggregate network. The results for the designated network in this equivalent network will be the same as those in the original network.

Aggregation produces exact results for queueing networks with a product-form solution. However, this approach may be computationally more expensive than using the usual techniques (MVA or Convolution) if the objective is to solve for only one set of parameters for the designated network.

The real advantage of aggregation may actually lie in solving *non-product form queueing networks*. In this case, the usual strategy is to put the components that do not have the product form in the designated network. A FES is then obtained for the aggregate, which contains only those queues that would have a product form solution. The equivalent network may now be solved using either approximate non-product form solution techniques or by simulation. Since solving an actual non-product form network is computationally very expensive, the reduced number of queues in the equivalent model would reduce the computation time. However, in this case, aggregation is only an approximation. This is because the FES cannot exactly model the behaviour of the aggregate as the information on the location of the customers in the aggregate is discarded. However, in many cases of practical interest, the approximation provided by this approach gives acceptable results.