6 Addendum to: General Networks with Batch Movements and Batch Services

6.6 Closed networks of generalized S-queues with unreliable servers

S-queues with rather general service regime are described in Section 6.4 as models for discrete-time systems in which customers can both arrive and be served in batches. As indicated, these queues can be connected into networks, which we shall sketch here for the case of closed networks.

The model encompasses the case of servers that break down and are repaired. Similar to Section 5, when a node breaks down the customers present there are frozen and no new arrivals are admitted to that node.

So with the possibility of having unreliable nodes we are faced to define rerouting strategies (protocols) for customers to get round inactive nodes. We shall discuss several alternative rerouting regimes.

It is shown that the equilibrium distribution for these networks has product form and is insensitive against varying the rerouting schedule in the described class of schedules. Moreover, we compute some availability and performance measures.

This Section is based on work with Kersten Tippner [Tip06].

We consider a discrete-time closed queueing network with $J$ nodes and $K$ customers. The nodes are unreliable, i. e. they can break down in the course of a time slot and then have to be repaired beginning with the next time slot. Thus, the description of the system state has to contain information about which nodes are working and which nodes are currently being repaired. Moreover, we examine the joint queue-length process for the $J$ nodes of the network. The states of the network are thus given in the form

$$x = (n, \bar{I}) = (((n_1, \ldots, n_J), (\bar{I})): n_i \in \mathbb{N}, i = 1, \ldots, J; \sum n_i = K; \bar{I} \subseteq J = \{1, \ldots, J\}),$$

$n_i$ being the number of customers at node $i$, $\bar{I}$ being the set of nodes currently under repair. Changes in the state of the system occur due to

a) breakdowns of active nodes and/or repairs of inactive nodes, and

b) departures of customers from nodes and their arrival to other nodes.

Breakdowns and repairs are assumed to occur independently from the queue-lengths at the various nodes of

\[\text{wr6ad.tex}\]
the network. If, at the beginning of a time slot, the nodes in $I \subseteq \bar{J}$ are inactive, the probability that, by the end of the same time slot, the set of inactive nodes is $L \subseteq \bar{J}$, is $\gamma(I, L)$, where

$$\gamma = (\gamma(I, L) : I, L \subseteq \{1, \ldots, J\})$$

is a reversible transition matrix.

Customers whose service times end leave their respective nodes at the end of the respective time slot and are then immediately routed to the next node. We denote the joint queue-length process by

$$X = ((X_t(1), \ldots, X_t(J)) : t \in \mathbb{Z})$$

and

$$D = ((D_t(1), \ldots, D_t(J)) : t \in \mathbb{Z})$$

are the sequences of departure and arrival vectors, respectively. By $A$, we denote the set of all possible departure and arrival vectors, which can be summarized as "transfer vectors".

The joint queue-length process is defined pathwise by

$$X_{t+1} = X_t - D_t + A_t;$$

its state space is $S(K, J) := \{(n_1, \ldots, n_J) : \sum_{i=1}^J n_i = K\}$.

If $\bar{I} = \emptyset$; i.e. if all nodes are active, customers are served at all nodes. The nodes are assumed to be working independently from each other and to be generalized S-queues, that is,

a) customers both arrive at the nodes and depart from them in batches, and

b) $D_t(i), i \in \bar{J}$, does not depend on the past of the system but via $X_t(i)$.

Moreover, we assume that the departure probability $q_i(n_i, a_i) := P(D_t(i) = a_i | X_t(i) = n_i)$ is given by

$$q_i(n_i, a_i) = \frac{\Psi_i(n_i - a_i)}{(a_i)!\Phi_i(n_i)}.$$  \hspace{1cm} (1)

Setting

$$\Phi_i(n_i) := \left(\frac{n_i}{a_i}\right)^{a_i} \left(1 - p_i\right)^{n_i - a_i}$$

and

$$\Psi_i(n_i) := \frac{1}{(n_i)!} \left(\frac{1 - p_i}{p_i}\right)^{n_i},$$

it is obvious that $q_i(n_i, a_i)$ fulfills (1).

Example 6.1 Infinite server network

All customers are served at the same time as though each of them were the only customer in the network. The service of a customer at node $i$ terminates in the current time slot with probability $p_i$; with probability $(1 - p_i)$, service continues for at least one more time slot.

$$\Rightarrow q_i(n_i, a_i) = \left(\frac{n_i}{a_i}\right)^{a_i} \left(1 - p_i\right)^{n_i - a_i}.$$  \hspace{1cm} (2)
A customer having left node $i, i \in J$ is routed to node $j, j \in \bar{J}$ with jump probability $r(i, j)$, independent of the routing of all other customers. The routing probability $r(a, a') := P(A_t = a' \mid D_t = a)$ is thus given by

$$r(a, a') = \sum_{\tilde{a}} \prod_{i=1}^{J} a_i \prod_{j=1}^{J} \frac{r(i, j)^{a_{i,j}}}{(a_{i,j})!}$$

where the summation is over all $\tilde{a} := ((a_{1,1}, \ldots, a_{1,J}), (a_{2,1}, \ldots, a_{2,J}), \ldots, (a_{J,1}, \ldots, a_{J,J}))$ such that $\sum_{i=1}^{J} a_{i,j} = a_i$ and $\sum_{i=1}^{J} a_{i,j} = a'_i$.

If, however, $\bar{I} \neq \emptyset$, that is, if at least one node is inactive, either all nodes immediately stop serving customers ("stalling", see section 6.7.3), or just the nodes in $\bar{I}$ interrupt services and reject the arrival of new customers so that one of the re-routing strategies described in section 6.7 has to be applied; the active nodes continue their services. In either case, though, all customers being served by a node that breaks down have to stay there until the node is repaired and their service time is terminated.

6.7 Re-routing strategies

Given that the nodes in $\bar{I} \subseteq \bar{J}$ are inactive, $A_k := \{a \in A : a \text{ consists of } k \text{ customers}\}$ can be divided into two disjunct subsets, the permitted transfer vectors ($=: A_k^P = \{a \in A_k : a_i = 0 \forall i \in \bar{I}\}$) and the prohibited ones ($=: A_k^\bar{P} = \{a \in A_k : \exists i \in \bar{I} \text{ such that } a_i \neq 0\}$). Departure vectors are (by construction) always permitted; an arrival vector, though, may be either permitted or prohibited.

The re-routing strategies described below are then applied to prohibited arrival vectors.

For more details and proofs see [SD03]. The framework there is continuous time queueing network theory, but the routing chain in these models are discrete as they are here.

6.7.1 Re-routing at customer level

As customers are routed independently, we can determine for each of them whether they are making a permitted or a prohibited jump, i. e. whether they are jumping to an active or an inactive node. Customers making prohibited jumps are then re-routed according to RS-RD ("Repeated Service – Random Destination") or skipping rules.

RS-RD (Repeated Service – Random Destination) Each customer making a prohibited jump is sent back to the node they have just left; there they served once more. After having terminated the additional service time, they jump to a node in the network according to the jump matrix $r$, whenever they select an up-node. Otherwise the procedure is repeated until the jump is successfully performed. In the context of blocking networks this protocol is called transmission blocking. The jump matrix $r_{\bar{I}}$ for the system with inactive nodes is thus given by

$$\tilde{r}_{\bar{I}}(i, j) = \begin{cases} r(i, j), & i, j \in J \setminus \bar{I}, i \neq j \\ r(i, i) + \sum_{k \in \bar{I}} r(i, k), & j = i. \end{cases}$$
Skipping  Customer selecting a prohibited jump just make a virtual jump to the node of their choice; having arrived there, they immediately jump to the next node according to the jump matrix \( r \) as though they had just left the respective inactive node. A customer has to jump until they reach an active node. In that case, the jump matrix for the system with inactive nodes is given by

\[
\tilde{r}_I(i,j) = r(i,j) + \sum_{k \in I} r(i,k) \tilde{r}_I(k,j) \quad i, j \in \bar{J} \setminus \bar{I}
\]

and similarly for \( i \in \bar{I}, j \in \bar{J} \setminus \bar{I} \).

6.7.2 Re-routing at transfer-vector level

Instead of re-routing individual customers, though, one can also determine whether an arrival vector is permitted or prohibited and then, if the vector is prohibited, re-transform it according to some specified global RS-RD or global skipping rules.

Global RS-RD  If a departure vector is transformed into a prohibited arrival vector, the transformation is invalidated. That is, all customers, independently of whether they would have made a permitted or a prohibited jump, are sent back to the nodes they have just left; there they are served once again. The routing matrix for the system with inactive nodes \( \tilde{r}_I \) is then given by

\[
\tilde{r}_I(a,a') = \begin{cases} 
  r(a,a'), & a, a' \in A_k, a \neq a' \\
  r(a,a) + \sum_{a'' \in A_k} r(a,a'') \tilde{r}_I(a'',a'), & a' = a
\end{cases}
\]

Global Skipping  If a departure vector \( a \) is transformed into a prohibited arrival vector \( a' \), the vector \( a' \) has to be re-transformed into a vector \( a'' \) etc. until a permitted transfer vector \( \hat{a} \) is reached. In that case, the routing matrix \( \tilde{r}_I \) for the system with inactive nodes is given by

\[
\tilde{r}_I(a,a') = r(a,a') + \sum_{a'' \in A_k} r(a,a'') \tilde{r}_I(a'',a'), \quad a, a' \in A_k
\]

and similarly for \( a \in A_k, a' \in A_k \).

6.7.3 Stalling

Stalling, as opposed to the more theoretical re-routing strategies describes above, is a technique already used in practice, for instance in quality control. It implies that, as long as there are nodes under repair in the network, not a single customer is served. All nodes immediately interrupt service. The inactive nodes are repaired; it is, however, possible that further breakdowns occur during a such "idle" period. Services are only continued when all nodes are in active status again.

6.8 Equilibrium distribution for networks of generalized S-queues with unreliable servers

Theorem 6.2  In closed networks of generalized S-queues with departure probabilities in the form (2) and with unreliable servers, the equilibrium distribution for the network process is independent of the strategy applied and is
\[
\pi((n_1, \ldots, n_J), \overline{I}) = \tilde{K}_{K,J}^{-1} \tilde{\pi}(\overline{I}) \prod_{i=1}^{J} y(i)^{n_i} \Phi_i(n_i),
\]

(2)

\(\tilde{K}_{K,J}^{-1}\) being the norming constant, \(y(\cdot)\) solving the balance equations

\[
y(i) = \sum_{i=0}^{J} y(i) r(i, j)
\]

and \(\tilde{\pi}(\cdot)\) being the probability solution of the balance equations for the breakdown - repair process.

\[
\tilde{\pi}(\overline{I}) = \sum_{L \subseteq \overline{J}} \tilde{\pi}(L) \gamma(\overline{J}, \overline{I}), \quad L \subseteq \overline{J}.
\]

Proof: (2) fulfills the global balance equations of the system. This can be seen by direct computation, either with the help of detailed balance equations or with the test by time reversibility.

\section*{Availability and performance measures}

Knowing the equilibrium distribution \(\pi\) for the network process, we can compute different performance and availability measures:

\begin{theorem}
\begin{enumerate}
\item The stationary joint point availability of subnetwork \(\overline{K} \subseteq \overline{J}\) at time \(t \geq 0\) is

\[
\text{Av}(\overline{K})(t) = 1 - \sum_{\overline{I} \supseteq \overline{K}} \gamma(\overline{I})
\]

\item The mean queue lengths \(E(X_i), i \in \{1, \ldots, J\}\), are the same as in the according network without breakdowns and repairs.
\end{enumerate}
\end{theorem}

\begin{theorem}
In a network with unreliable servers working with a re-routing strategy, the throughput at node \(j\), \(TH(j)\), is

\[
TH(j) = \hat{TH}(j) \cdot \text{Av}(\{j\})(t),
\]

\(\hat{TH}(j)\) being the throughput at node \(j\) in the according network without breakdowns and repairs.

If, however, stalling is applied, the throughput is

\[
TH(j) = \hat{TH}(j) \cdot \tilde{\pi}(\emptyset).
\]
\end{theorem}

\section*{References}
