A DISCRETE TIME MODEL OF A CAR ASSEMBLY SYSTEM

K. Furmans

University of Karlsruhe, Material Handling Systems
Hertzstr. 16, 76187 Karlsruhe, Germany
e-mail: kai.furmans@mach.uni-karlsruhe.de

Abstract

The performance evaluation of assembly lines under stochastic influence (changing model mix and varying processing times) is usually done by time consuming simulation experiments. In this paper a discrete time queueing model is presented that is a basis of an approximate performance analysis of car assembly systems. The model elements are described and approximate calculation methods for the waiting time and interdeparture time distributions at each work station are derived.

Problem Characterization

Part of simultaneous engineering is to link the development of the product and its manufacturing system. The benefit of such an approach is that the product designers get quickly information about the consequences of product design decisions on the manufacturing process. In a design-to-cost environment, the feedback should allow an estimation of the costs or savings involved by an alteration of the product and its production process. In the early stage of the development process there usually are many solutions to be quickly considered and evaluated to make sure that no possible solution is neglected.

The design of cars and their assembly systems is usually done in several steps. The layout of the assembly system is determining the work force needed and therewith has a relevant impact on the production cost of a car. The evaluation of such a layout is a complex task. Cars are usually assembled on a mixed-model line, where several elementary work steps (i.e. inserting a screw) are combined in a so called task (i.e. fixing the exhaust pipe). These tasks have to be assigned to work stations, where at least one worker is performing the required work steps for each car. The number of work steps as well as the time for performing a work step may vary from car to car according to the level of special equipment to be installed. This results in several possible processing times for each task. Due to the model mix, the processing times will also vary from car to car. The processing times for the tasks that are assigned to a work station can be described by statistical distributions.

Due to spatial restrictions, a worker usually starts his task at a car after the preceding task is finished. This leads to statistically distributed interarrival times of cars at his work station depending on the interrelease distribution of the preceding step. In order to prevent blocking, buffers between work stations are necessary.

In the assembly systems considered in this paper (as can be found at the Mercedes-Benz plant in Rastatt, Germany) the workers are moved along the line on platforms which are carrying the car they are working on. The platforms move with uniform speed. Instead of physically buffering the cars, the relative position of the workers is changing when they start a new job. This position depends on the time the preceding job is finished and the time the workers have finished the last item they have been working on.

When optimizing the assembly processes the goal usually is to minimize worker idle time by allocating an equal amount of work to each work station while observing the precedence restrictions of the elementary work steps. This could also be done regarding the expected amount of work (for typical algorithms see for instance Jackson [Jack56], Buxey [Buxe74] and Pinto et al. [PiDK81]). An undesirable effect by utilizing these optimization methods could be the introduction of more unproductive work due to increased floating of the workers, because each work station is assigned to a specific segment of the line. Here the required parts for the work station are stored or delivered. The relative position of the worker to the assigned segment is changing due to the statistical characteristic of the interarrival and the processing times. This in turn might increase the time to pick up the required parts because of longer distances to the segment where the parts are stored.

In the planning stage of an assembly system it is desirable to evaluate the above mentioned effects for different work station allocations and different assembly sequences. This could be done with the queueing model, subsequently presented in this paper.

Description of Model Elements

Assumptions

It is assumed, that the distributions of the processing times at the work stations are independent of each other. If for instance a car is equipped with air-conditioning it might be fitted with a sunroof with the same probability as those cars without air-conditioning.

The processing times are determined by the amount of time needed for fitting a specified part, weighted with the probability, that such a part is requested, i.e. if 75% of all cars are equipped with a sunroof and fitting a sunroof takes 10 minutes then the processing time 10 minutes occurs with probability 0.75 (if all other possibilities require a processing time unequal to 10 minutes) at the workstation for sunroofs. All necessary parts are always present, so that no additional time for material handling is required.

The variability of the processing time introduced by manual operation has not been modeled in the studied case. It is possible to include non time-dependent variation of the processing times by multiplying the known processing times with the distribution of the processing time variation. Breakdowns and temporary interruptions can be modeled in the same manner.

Blocking occurs in the sense, that a worker starts working on an assembly step only when the preceding step has been finished.

Queueing Systems

The assembly line can be modeled as a queueing network, consisting of M queueing systems, representing the work stations. The stochastic demand for processing resources is modeled by the queueing systems. Each work station i ($i = 1, ..., M$) is characterized by the stochastic variables $\{T_{a,i}\}$ for the interarrival time and $\{T_{s,i}\}$ for the service time (processing time). It is assumed, that the arrival and the service processes are renewal processes. The increment between subsequent values for all discrete time variables shall be t_{inc}. Then the density functions of $t_{a,i}$ and $t_{s,i}$ at work station i are defined as:

$$P\left(T_{a,i} = t_{a,i} = j \cdot t_{inc}\right) = : a_{i,j} \quad j = 0, \dots j_{a,max}$$
$$P\left(T_{s,i} = t_{s,i} = j \cdot t_{inc}\right) = : s_{i,j} \quad j = 0, \dots j_{s,max}$$

$$a_i = \left(a_{i,0}, a_{i,1}, \dots, a_{i,j_{a,max}}\right)^T; \quad s_i = \left(s_{i,0}, s_{i,1}, \dots, s_{i,j_{s,max}}\right)^T \tag{1}$$

with $j_{a,max}$, $j_{s,max}$ defining the range of possible values for j.

The arrival rate λ_i and the service rate μ_i at work station i are computed as follows:

$$\lambda_i = \frac{1}{E\left(t_{a,i}\right)} \quad E\left(t_{a,i}\right) = t_{inc} \sum_{j=0}^{j_{a,max}} j\, a_{i,j} \quad \text{and} \quad \mu = \frac{1}{E\left(t_{s,i}\right)} \quad E\left(t_{s,i}\right) = t_{inc} \sum_{j=0}^{j_{s,max}} s_{i,j} \tag{2}$$

The assembly line is moving with a constant speed v. The length of the associated segments of the work stations is denoted by l_i. If a task is performed by exactly one work station, to achieve steady state it is necessary that:

$$E(t_{s,i}) < l_i / v \tag{3}$$

The density functions of the stochastic variables $\{T_{w,i}\}$ of the waiting time distribution as well as $\{T_{r,i}\}$ for the idle time distribution at work station i are defined as

$$P\left(T_{w,i} = t_{w,i} = j \cdot t_{inc}\right) = w_{i,j} \quad j = 0,\ldots j_{w,max}$$

$$P\left(T_{i,i} = t_{i,i} = j \cdot t_{inc}\right) = i_{i,j} \quad j = 0,\ldots j_{i,max}$$

(4)

with $j_{w,max}$, $j_{i,max}$ defining the range of possible values for j in this case.

The intervals between two successive releases of cars from the system are characterized by the interdeparture time $\{T_{d,i}\}$ with density function $f(t_{d,i})$, thus describing the flow of cars leaving the queueing-system:

$$P\left(T_{d,i} = t_{d,i} = j \cdot t_{inc}\right) = d_{i,j} \quad j = 0,\ldots j_{d,max}$$

(5)

The distribution of the sojourn time $f(t_{q,i})$ is approximately available from the convolution of the waiting and the service time distribution. The vectors d, w, q and i are defined similarly to (1).

Several methods have been developed to compute the density functions $f(t_w)$, $f(t_i)$, and $f(t_d)$ for discrete time queueing systems, by solving Lindley's equations or by analyzing a spectral representation of the interarrival and service time distributions (see Kleinrock [Klei75] pp. 283ff, Haßlinger and Rieger [HaRi91]). The proposed approximation uses a computing time efficient algorithm that was developed by Grassmann and Jain ([GrJa88], [GrJa89]), based on a Wiener-Hopf-Factorization of the underlying random walk. It is yielding a numerical approximation of the density functions $f(t_w)$, $f(t_i)$ and $f(t_d)$.

Theoretically, t_d may become infinitely large. Therefore a finite vector representation of the distribution $f(t_d)$ without any compensation leads to a defective distribution, thus an approximation of the true interdeparture time is required.

In order to link the queueing systems further model elements, called nodes, are introduced, which split the flow of cars and join several streams to one. To simplify notation, in the subsequent sections the index of the node itself is omitted.

Deterministic Split Node

If the expected service time $E(t_s)$ for an operation is larger than l/v, the workload has to be distributed on n parallel work stations. The workers are taking turns in working on the cars, the first one working on car number $\{1, n+1, 2n+1, \ldots\}$ the second on car number $\{2, n+2, 2n+2, \ldots\}$. This is modeled by a *deterministic split* which splits a common input stream into several parallel queueing systems in the above described manner.

The vector d describing the density function of the departure process in each of the output streams is computed by a $n\text{-}1$-fold convolution of the density functions of the input process:

$$d = \underset{1}{a} \otimes \underset{2}{a} \otimes \ldots \otimes \underset{n-1}{a}$$

(6)

The deterministic split node can be used in conjunction with n immediately succeeding single-server queueing systems to model the behavior of a multiple-server queueing system with round-robin service discipline.

Stochastic Split Node

The flow of cars through the assembly system may split stochastically, for instance if some equipment (i.e. sunroof) is not installed in all cars. This is modeled by a *stochastic split* node, which transforms one input stream in two or more output streams. The assignment of cars to one of the output streams is done according to the assigned probability p_h that indicates the probability of an arriving customer exiting at stream h. The input stream is characterized by its density function $f(t_a)$, each output stream h by its interdeparture time density function $f(t_{d,h})$.

In contrast to the deterministic split, the number of arrivals at the node that lead to a subsequent departure at stream h is subject to the stochastic assignment process of cars to a specific stream.

Without loss of generality we assume, that the previous customer has been directed to stream h. Then, with probability p_h the next customer will take the same direction h. Therefore with probability p_h the interdeparture time at stream h is determined by the interarrival distribution. With probability $(1-p_h)\,p_h$ the second succeeding customers will leave at exit h. Thus with probability $(1-p_h)\,p_h$ the interdeparture process is the result of the convolution of $f(t_a)$. Extending this pattern, the interdeparture time density function of stream h can be computed by (7).

$$
d = \begin{array}{l}
\left(1 - p_h\right)^0 p_h\, a\ + \\[2mm]
\left(1 - p_h\right)^1 p_h\, a \otimes a\ + \\[2mm]
\left(1 - p_h\right)^2 p_h\, a \otimes a \otimes a\ + \\[2mm]
\vdots
\end{array}
\tag{7}
$$

It is obvious, that d is a vector of infinite dimension, therefore any limited evaluation of the values of d will lead to a defective distribution. To make $f(t_d)$ a proper distribution the last component $d_{j,max}$ is set to:

$$
d_{j,max} = 1 - \sum_{j=0}^{j_{d,max}\,-\,1} d_j
\tag{8}
$$

As a consequence the expected value $E(t_d)$ of the interdeparture time will be underestimated, leading to a higher utilization of the downstream queueing systems. The subsequent dimensioning of the work stations therefore will be conservative.

Merge Streams Node

A number of independent streams can be merged by a *merge* node yielding a single output stream. Its departure process will be approximated by a renewal process, whose interdeparture time distribution is computed by determining the minimum time to the next arrival in one of the n input streams (see Haßlinger and Rieger [HaRi91]). It is sufficient to treat the case of two incoming streams with indexes h and i, because all other cases could be handled by starting with two streams and adding the subsequent streams to the previous result.

The distribution of the time to the first arrival from stream h, $t_{r,h}$ is different from all subsequent interarrival times, because in general the first interval did not start with a renewal event of the associated stream.

The limiting density function of $t_{r,h}$, the time to the first arrival (also called excess), will be defined as:

$$P\left(T_{r,h} = t_{r,h} = j \cdot t_{inc}\right) = r_{i,j} \qquad \forall j = 0,\ldots j_{h,max}$$ (9)

The actual values of the distribution function are obtained by the following procedure (see Feller pp. 334 - 335 [Fell68]). To simplify notation we are introducing $\alpha_{h,k}$, the probability of an interarrival interval from stream h being longer than $k\,t_{inc}$:

$$\alpha_{h,k} = \sum_{n=k+1}^{j_{a,max}} a_{h,n}$$ (10)

The time that has passed since the last arrival of a customer in stream h is called the age of the arrival process. The limiting distribution of the age of the arrival process h will be defined by $\beta_{h,k}$, the probability of seeing the arrival process of stream h at an age of k. It is calculated by:

$$\beta_{h,k} = \alpha_{h,k}\lambda$$ (11)

In the continuous distribution case, the limiting distribution of age is identical to that of excess e.g., the time to the next arrival in the stream h (see also [Fell68]). In the discrete time case, we have to take into account the probability of an arrival event at a specific time. The probability of an excess of length $j \cdot t_{inc}$ equals:

$$r_{h,j} = \begin{cases} \beta_{j-1} & \text{for } j = 1,2,\ldots \\ 0 & \text{for } j = 0 \end{cases}$$ (12)

The excess of the merged stream is the minimum of the time to the next arrival in all incoming streams. The probability r_j^{sum} of the time to the next arrival in the combined stream being of duration $j \cdot t_{inc}$ is given by:

$$r_j^{sum} = r_{h,j} \sum_{k=j+1}^{j_{i,max}} r_{i,k} + r_{i,j} \sum_{k=j}^{j_{h,max}} r_{h,k}$$ (13)

Now, reversing the steps defined with formulas (10) - (12) we get the probabilities β_j of the age being $j \cdot t_{inc}$:

$$\beta_j^{sum} = r_{j+1}^{sum}$$ (14)

The throughput of a merge node λ^{sum} is the sum of the arrival rates λ_i from all streams i. With λ^{sum} the probabilities d_j of the interdeparture times are derived from:

$$\alpha_j^{sum} = \frac{\beta_j^{sum}}{\lambda^{sum}}$$ (15)

$$d_j = \begin{cases} \alpha_j - \alpha_{j+1} & \text{for } j = 0,\ldots,j_{max}-1 \\ \alpha_{j,max} & \text{for } j = j_{d,max} \end{cases}$$ (16)

The dimension $j_{d,max}$ of the resulting vector d, is given by $\min\{j_{h,max}, j_{i,max}\}$.
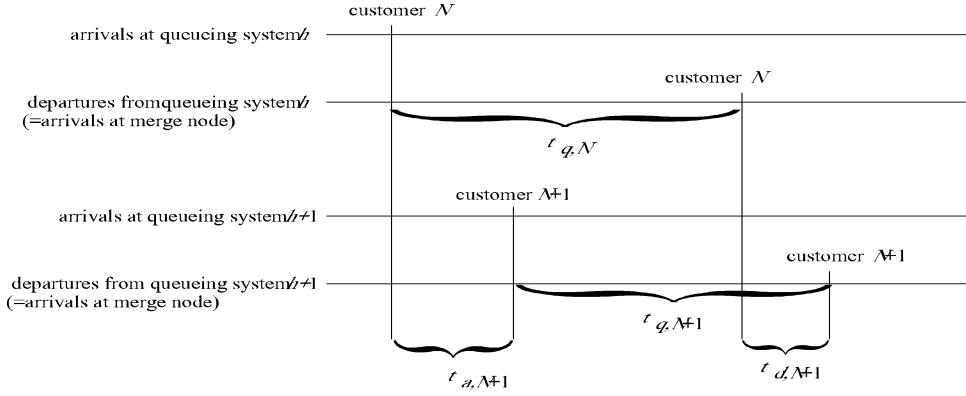
Merge Dependent Streams Node

Figure 1: Interarrival processes of two dependent streams

The independence of the incoming streams at a merge node is not in all cases a reasonable assumption. This is especially the case for a merge node that succeeds a round robin service station. The round robin service is modeled by a deterministic split node with subsequent queueing systems. If the sequence of the jobs in the outgoing stream is the same as in the incoming stream, or if only pairwise exchanges in sequence occur, the interdeparture time distribution can be approximated by observing two subsequent departures. In such a case the interdeparture times are determined by the difference of the interarrival time between two successive cars (see figure 1) N and $N+1$. The interdeparture time between the N-th and the $N+1$-st car is determined by:

$$t_{d, N+1} = t_{a,N+1} + t_{q,N} - t_{q,N+1} \tag{17}$$

The distribution of $\{T_a + T_q\}$ can be approximately computed from the convolution of the respective vectors a and q and will be denoted by w. The elements d_j of vector d, the probabilities of an interarrival event of length $j \, t_{inc}$ are accordingly computed from:

$$d_j = \sum_{k = j + 1}^{j_{q,max}} w_k \cdot q_{k - j} + \sum_{k = 0}^{j} w_k \cdot q_{k + j} \qquad \text{for } j > 0 \tag{18}$$

$$d_j = \sum_{k = 0}^{j_{q,max}} w_k \cdot q_{k - j} \qquad \text{for } j = 0 \tag{19}$$

Fork Stream Node

Although the cars are conveyed linear along the line, the interdependencies between tasks and work stations are not strictly linear. The dependencies between tasks are described by the precedence relations. The sequence of tasks has to be organized in a way that precedence relations are not violated. Within the limits described by the precedence relations some tasks may be performed simultaneously, leading to parallel processing of different tasks (i.e. fitting right and left side doors to the car simultaneously). This may lead to two (or more) parallel streams of tasks assigned to the left and right side of the line. A *fork node* multiplies one input stream into n output streams, yielding for instance the input distribution of tasks for the right and the left side of the line.

The elements d_{ij} of the interdeparture time distribution for all output streams $i= 1,...,n$ therefore will be the same as those of the interarrival distribution.

$$d_{i,j} = a_j \quad \forall \; i = 1,\dots, n, \quad j = 0, \dots, j_{a,max} \tag{20}$$

Join Streams Node

The *join node* is used to model the synchronization that is necessary to assure that two parallel sequences of tasks having originated at the same fork node are both finished. As an example consider an assembly line with work stations on the left and right hand side of the line with a robot at the end of the line, that can start his operation on a car only when both sequences of tasks are finished. The length $\{T_q\}$ of a sequence of k tasks can be computed by a convolution of the sojourn time distributions of the queues visited during the sequence.

The vectors describing the density of two sojourn time distributions for the sequences i and j of common origin shall be denoted by q_i and q_h. The associated interarrival times from the sequences are a_i and a_h. We approximate the point-process of the interdeparture stream by a renewal, whose distribution is constructed by examining five cases:

A. The interdeparture time is equal to the interarrival time of stream i if the sojourn time of two successive customers in sequence i was longer than that of the respective customer in sequence h. This case contributes $\delta_{A,j}$ to the total value of d_j.

B. The sojourn time of the second to the last customer in sequence h was longer than that of the

$$\tag{21}$$

$$\delta_{A,j} = a_{i,j} \left[\sum_{l=1}^{j_{q,i,max}} q_{i,l} \sum_{m=0}^{l-1} q_{h,m} \right]^2$$

respective customer in sequence i, while the opposite is true for the last customer. The interdeparture time in this case is determined by the excess of the arrival stream of sequence i:

$$\delta_{B,j} = r_{i,j} \sum_{l=1}^{j_{q,i,max}} q_{i,l} \sum_{m=0}^{l-1} q_{h,m} \cdot \sum_{l=1}^{j_{q,h,max}} q_{h,l} \sum_{m=0}^{l-1} q_{i,m} \tag{22}$$

The cases C and D are symmetric to A and B respectively, therefore the expressions (21) and (22) apply as well with i and j exchanged accordingly.

E. In case E the sojourn times in both streams are equal, the interdeparture time will be determined by the maximum of the excesses in both arriving streams:

$$\delta_{E,j} = r_{i,j} \sum_{l=0}^{j-1} r_{h,l} + r_{h,j} \sum_{l=0}^{j-1} r_{i,l} \tag{23}$$

Finally the values d_j are obtained by:

$$d_j = \delta_{A,j} + \delta_{B,j} + \delta_{C,j} + \delta_{D,j} + \delta_{E,j} \tag{24}$$

Example

To check the applicability of the chosen approach, an assembly line for cars at the Mercedes-Benz plant in Rastatt, Germany has been modeled with the above described elements. A part of the resulting network is shown in figure 2. The whole network consists of 11 different tasks that are performed at 39 work stations.
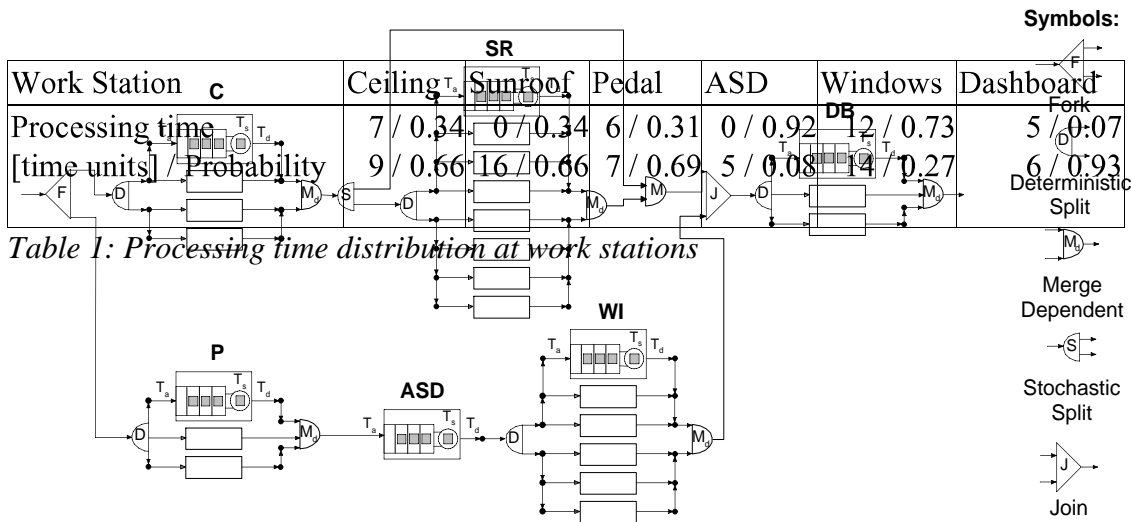
Work Station	Ceiling	Sunroof	Pedal	ASD	Windows	Dashboard
Processing time [time units] / Probability	7 / 0.34	0 / 0.34	6 / 0.31	0 / 0.92	12 / 0.73	5 / 0.07
	9 / 0.66	16 / 0.66	7 / 0.69	5 / 0.08	14 / 0.27	6 / 0.93

Table 1: Processing time distribution at work stations

Figure 2: Partial view of the queueing network model

The numerical evaluation of the model takes a few seconds on a PC or a workstation. The results of the waiting time distributions are compared to the results from a simulation experiment that was performed for the same model. The model was implemented on a DEC-Alpha 3000/300 in SIMPLE++. The simulation run covered 485.000 finished cars, requiring approximately 96 hours of CPU time.

Waiting Time	Ceiling QM	Ceiling Sim.	Sunroof QM	Sunroof Sim.	Pedal QM	Pedal Sim.	ASD QM	ASD Sim.	Windows QM	Windows Sim.	Dashboard QM	Dashboard Sim.
0	120661.7	120698	60391.5	60689	158293.2	158342	436032.9	438218	62001.2	62163	131425.4	133805
1	728.3	681	121.4	5	3398.9	3436	12139.0	10828	5173.5	5242	13595.7	14021
2	0.0	5	60.7	0	0.0	78	18451.3	18009	4122.6	4042	9549.4	7916
3	0.0	0					12624.6	12644	3071.8	3100	4855.6	4624
4							3884.5	3750	2182.6	2172	1456.7	1276
5							1456.7	1259	1535.9	1461	647.4	185
6							485.6	584	970.0	952	323.7	22
7							0.0	155	646.7	591		
8							0.0	69	404.2	415		
9							0.0	24	242.5	245		
10							0.0	2	161.7	172		
11							0.0	1	80.8	107		
Test Value	1.57E-02		1.86E-38		2.99E-12		8.41E-97		2.86E-02		1.39E-208	

Table 2: Comparison of expected number of waiting time observations (QM based on queueing model) and number of observations in simulation experiment (Sim).

The processing times at the work stations and their probabilities are shown in table 1. The stochastic split directs with a probability of 0.66 bodys to the sunroof assembly. The external interarrival time takes the value 2 [time units] with probability 0.3, and 3 with probability 0.7. In table 2 the expected number of observations for a specific waiting time and the number of observations of a specific waiting time are shown for all work stations. The approximation results of the queueing network model are very close to the simulation results. The test values for the X^2-Test are shown in the last line. Although the applicability of the X^2-Test depends on the independence of observations, which is not given in a simulation environment, the results show a low probability of having to discard the assumption of identical distributions of waiting time. Based on the calculated waiting times, other important performance measures, like sojourn time for the whole line or certain segments can be computed. These in turn are needed to calculate the distribution of the position of the workers in relation to their assigned segments.

The results before have been compared with the hands-on experience of the employees at the assembly line. It was found, that the results reflected their daily experience very well.

Bibliography

[Buxe74]   G. M. Buxey, *Assembly Line Balancing with Multiple Stations,* Management Science, Vol. 20, No. 6, pp. 1010-1021, 1974

[Fell68]   William Feller, *An Introduction to Probability Theory and its Applications*, Wiley, New York, Third Edition, 1968

[GrJa88]   Joti J. Jain, Winfried K. Grassman, *Numerical Solution for the Departure Process from the GI|G|1 Queue*, Computing Operations Research, Vol. 15, No. 3, pp. 293-296, 1988

[GrJa89]   Joti J. Jain, Winfried K. Grassman, *Numerical Solutions of the Waiting Time Distribution and Idle Time Distribution of the Arithmetic GI|G|1 Queue*, Operations Research, Vol. 37, No. 1, 1989

[HaRi91]   Gerhard Haßlinger, Erik S. Rieger: *Modellierung offener Warteschlangennetzwerke durch Erneuerungsprozesse im diskreten Zeitbereich,* in: Lehmann, Lehmann (Eds.): *Messung, Modellierung und Bewertung von Rechensystemen,* Informatik Fachberichte, 286, Springer-Verlag, Berlin, 1991

[Jack56]   James R. Jackson: *A Computing Procedure for a Line Balancing Problem,* Management Science, Vol. 2, No. 3, pp. 261-271, 1956

[Klei75]   Leonrad Kleinrock: *Queueing Systems, Volume 1: Theory,* John Wiley, New York, 1975

[PiDK81]   Peter A. Pinto, David G. Dannenbring, Basheer M. Khumawala: *Branch and Bound and heuristic procedures for assembly line balancing with paralleling of stations,* International Journal of Productions Research, Vol. 19, No. 5, pp. 565-576, 1981

[Wolf89]   Ronald W. Wolff: *Stochastic Modeling and the Theory of Queues*, Prentice Hall, Englewood Cliffs, N.J., 1989

Acknowledgements