University of Würzburg
Institute of Computer Science
Research Report Series

Two-Moment Analysis of Alternative Tool Models with Random Breakdowns

M. Mittler

Report No. 141 July 1996

Lehrstuhl für Informatik III, Universität Würzburg,
Am Hubland, D – 97074 Würzburg
Tel.: +49-931-8885518, Fax: +49-931-8884601
e-mail: mittler@informatik.uni-wuerzburg.de

Abstract: *We investigate a two-station, single-server queuing system with random breakdowns. Two classes of customers arrive at the system with finite buffers where one server is dedicated to only one customer class while the second server can process both types of customers. Interarrival and service times are exponentially distributed while failure and repair times follow the generalized exponential distribution. Motivated by the application of alternative tools in semiconductor manufacturing we consider two different control schemes that trigger processing of customers of the second class on the server that is usually dedicated to customers of the first class. We calculate the first two moments of customer flow times for both control schemes.*

Keywords: *alternative tools, semiconductor wafer fabrication, two-moment analysis, breakdowns, failure, repair*

1 Introduction

The fundamental characteristic of semiconductor manufacturing processes is that production planning and scheduling of these facilities and their performance evaluation are difficult.[1] The major reasons for this are reentrant product flows, multiple types of equipment, and complex manufacturing processes (Uzsoy et al. 1994). Besides this, a huge number of control and environmental factors aggravate the tasks of planning and scheduling, and performance evaluation (Bohn 1995). With respect to the material flow, among those factors are well-studied factors like batching and breakdowns, and factors to which only minor attention has been paid, like alternative tools (see de Ridder and Rodriguez (1994).

According to the APICS dictionary[2], the occurrence of alternative tools refers to a scenario in a manufacturing process where a particular operations can be carried out by various tools. Apart from a preferred tool by which the jobs are usually processed, jobs may follow a less preferred alternative routing. Nevertheless, each routing results to the same item but may require a different processing time.

In semiconductor manufacturing for example, tools processing stepper operations may be replaced by the latest models after a few years of operation. Nevertheless, the old equipment is usually not discarded. The new equipment may have a greater precision than the older one but may be slower. Hence, a shop floor manager will require particular sensitive processing steps to be performed by the new equipment while the other processing steps to be carried out on the old equipment. However, if the equipment with greater precision is idle or only few lots of wafers wait to be processed on it, it may also perform non-sensitive processing steps.

The contribution of this papers is as follows. Based on the description above we consider a two-station queuing system corresponding to a work center with two single-machine tools each of which is dedicated to a particular processing step. Additionally, one of the processing steps may alternatively be routed to the machine that is dedicated to the other processing step. Alternative processing is triggered according to two different control mechanisms. Both of them take into account status information on the buffer lengths in front of the tools and aim at reducing the flow time of the lots that can follow the alternative routing. The *handover* mechanism employs the number of lots that can alternatively be processed as control threshold, the buffer in front of the alternative tool being empty. On the other hand, the *takeover* control triggers alternative service if and only if the number of queued lots in front of the alternative tool is below a certain value.

[1] For detailed introductions to semiconductor wafer fabrication we refer to Gise and Blanchard (1986), Münch (1993), Sze (1983), Fordyce and Sullivan (1994), Kempf (1994), and Bohn (1995).

[2] Cited according to de Ridder and Rodriguez (1994).

Since semiconductor manufacturing equipment is commonly subject to unscheduled breakdowns and since it is known from queuing analysis (cf. Hopp and Spearman 1996) that breakdowns strongly affect cycle time (i.e., the time a lot needs to travel through the entire manufacturing system) we also consider random breakdowns. Machine breakdowns can be either time-dependent or operation-dependent. In both cases they are modelled by the time within a machine is to fail and the time needed to repair it after failure.

We assume that lot interarrival times and processing times are exponentially distributed. Processing times on alternative routes may be different. Further, we assume that the time to fail and the time to repair follow the generalized exponential distribution. Finally, the buffers in front of the machines are finite and the lots queue according to the FIFO sequencing rule.

For both control mechanisms, using a standard technique from the literature, we calculate not only the mean but also the variance of the lot flow time (cycle time) through the work center. We pay attention to the variance of flow time, too, since the larger the variance of the flow time, the more likely it is to exceed certain time requirements like due dates (cf. Hopp and Spearman 1996) or time bounds within subsequent processing steps have to be finished (cf. de Ridder and Rodriguez 1994).

2 Literature

Although there is an extensive amount of literature regarding the analysis of queuing systems with batch service and/or breakdowns only a few articles deal with alternative tools.

Balakrishnan (1989) derives a method to dispatch customers to machines to facilitate a given makespan in a model with j customer classes that are served by a number of special purpose and general purposes machines (servers). Arrival times are assumed to be known a priori and service times are deterministic. Special purpose machines are dedicated to a single customer class while general purpose machines can process any type of customer. Further, to each customer a due date is assigned. A customer may also be preempted. The type of preemptions is preemptive-resume.

Leachman and Carmon (1992) proposed procedures to evaluate capacity limitations of alternative machines types in corporate-level production planning models where the alternative machines have deterministic, identical or proportional, processing times across the tasks they have to perform. The aim of this approach it to determine optimal production rates of each product type over some planning horizon. As Leachman and Carmon (1992), Rohan (1992) also uses Linear Programming techniques to investigate the problem of how to use multiple resource by multiple task types given certain constraints on

flexibility and to predict machine utilizations.

To the author's knowledge, the work of Xu et al. (1992) is the only publication with an alternative tool control mechanism similar to the ones described in the previous section. They consider a system consisting of two multi-server queuing stations each of them being usually dedicated to a single class of customers. Interarrival times (possibly different means) and service times (identical means) are exponentially distributed for both types of customers. Given that a customer originates certain holding costs during its sojourn, Xu et al. (1992) show that the handover control mechanism is optimal to minimize the long-run average cost. Further, using the dynamic programming approach, they derive a procedure to optimize the threshold that triggers alternative service.

Nelson and Philips (1993) approximate the mean flow time in a shortest queue routing system with general interarrival and service times. A single class of customers arrives at the multiple-server queuing system where each server has its own queue. Upon arrival, a customer joins the shortest queue. Within individual queues, customers wait for service in order of their arrival (FIFO).

A class of models similar to the previous one cope with customer jockeying. Customer jockeying implies that customers do not only join the shortest queue upon their arrival. Additionally, if the difference of the longest queue and the shortest queue exceeds a certain threshold, commonly one, then the customer at the tail of the longest queue jockeys to the shortest queue.[3]

However, the aim of the jockeying models is different from our aim. The goal of the articles previously mentioned and the publication of Xu et al. (1992) is to minimize performance measures with respect system characteristics like holding costs or inventory. Instead of this, we consider performance characteristics with respect to individual customers, i.e., cycle times of individual lots. Additionally, jockeying models do not include unscheduled breakdowns. If a server in a jockeying model breaks down, however, the customer at the head of the corresponding queue has to wait either until the server becomes available or until another queue becomes empty. This behavior may contribute to an increased cycle time such that time bounds may be violated.

3 Basic Model with Alternative Processing

We consider basic queuing model incorporating alternative processing depicted in Figure 1. There are two classes of customers arriving to the system each of them representing a different processing step. Customer class 1 requires processing from server 1 (mean pro-

[3] See for example Elsayed and Bastani 1985, Kao and Lin 1990, Zhao and Grassmann 1995 among many others.

cessing time μ_{11}^{-1}), whereas customer class 2 can alternatively be processed by server 1 or server 2. The mean processing times of class-2 customers on different servers may be different, i.e., μ_{12}^{-1} on server 1 and μ_2^{-1} on server 2. Further, interarrival and service

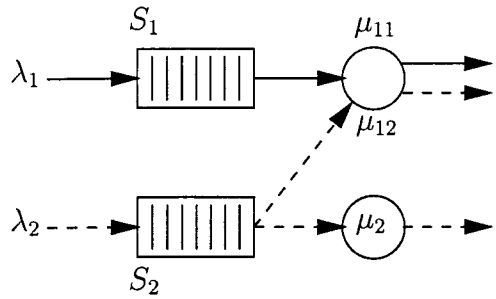

Figure 1 *Basic alternative-tools model*

times of customer class i are exponentially distributed with mean λ_i^{-1}. The buffers in front of both servers are finite where S_i denotes the capacity of buffer i. To complete the description of the basic model we refer to the unreliable server behavior. Both servers are subject to unscheduled breakdowns which can be operation- or time-dependent. The time to failure, F, and the time to repair, R, are assumed to follow a generalized exponential distribution to facilitate distributions with a coefficient of variation larger than 1 (Dallery 1994). The generalized exponential distribution function reads as follows:

$$F(t) \;=\; P[X \le t] \;=\; (1-q)\cdot 1 \;+\; q\cdot(1-e^{-\mu t}), \qquad t \ge 0, \tag{3.1}$$

The advantage of this distribution is that the resulting queuing models can be analyzed as if the failure and repair time would be exponentially distributed (Dallery 1994).

We applied two different control mechanisms to this model. These mechanisms differ from each other with respect to the scenario when alternative processing is triggered:

takeover: server 1 takes over customers from buffer 2 if the number of customers in buffer 1 is smaller than threshold A ($1 \le A \le S_1$);

handover: server 2 hands over customers to server 1 if the number of queued class-2 customers exceeds threshold B *and* if and only if queue 1 is empty at the same time ($0 \le B \le S_2$).

The aim of both control mechanisms is to reduce the first two moments of the sojourn time probability distribution function of class-2 customers. Our investigation represents a means how to adjust the thresholds A and B to achieve this goal.

The takeover model is motivated by a real application in a wafer fabrication plant (formerly owned by IBM) whereas the handover model is a simplified version of the

model investigated by Xu et al. (1992). The approaches of the two control mechanisms are different to some extend since in the takeover model, server 1 additionally delays customers of class 1 to process class-2 customers. On the other hand, the control mechanism of the handover model shortens the waiting time of class-2 customers server by dispatching them to server 1 also. Moreover, if A is set to 1, the takeover model approximately corresponds to two isolated finite-buffer $M/M/1$ systems since in this case, processing of class-2 customers on server 1 is only initiated if queue 1 is empty and server 1 is idle. Depending on the load, the differences among the performances measures of the takeover model and the two isolated queuing systems are more or less inconsiderable. On the other hand, if we set $B = S_2$ in the handover model, this model differs from two isolated $M/M/1$ systems with finite buffers only by that fact that a class-2 customer follows the alternative route only if server 1 is idle.

4 Analysis

4.1 General Approach

The stochastic behavior of a queuing system can be captured by a Markov process if interarrival and service times are exponentially distributed. In this case the stationary state probability vector $\boldsymbol{p} \triangleq (p_0, p_1, \ldots)$ can be calculated by solving the stationary equations

$$\boldsymbol{p} \cdot \boldsymbol{Q} = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{p} \cdot \boldsymbol{e}^T = 1, \tag{4.2}$$

where $\boldsymbol{Q}$ denotes the infinitesimal generator of the Markov process, $\boldsymbol{0} \triangleq (0, 0, \ldots)$, and $\boldsymbol{e} \triangleq (1, 1, \ldots)$. The stationary state probabilities immediately yield the mean cycle time by firstly calculating the mean number of customers in system and secondly applying Little's law.

The calculation of the variance of cycle time is more sophisticated and thus deserves further attention. From queuing analysis two different methods to calculate moments or distributions of cycle times are known: the exact method by Syski (1986) and Kühn (1972) and the approximate method of Melamed and Yadin (1984a, 1984b). We confine ourselves to the first method since the second one yields upper and lower bounds of the cycle time distribution only. We, however, aim at the calculation of moments. To make the paper selfcontained, we sketch the main ideas and the final equations of this approach and refer to Kühn (1972) for additional details. Though the original literature is devoted to moments of waiting times, the deliberations for cycle times are analogous.

Syski (1986) employed the *first-passage time* concept (also referred to as the concept of *taboo sets*) and obtained an ordinary differential equation for the Laplace-Transform

of the cycle time probability density function (PDF). The concept of taboo sets implies that the original state space is divided into two partitions, a regular set and a taboo set. Given that the tagged customer enters a regular state upon its arrival, one keeps track of its behavior until it reaches a state of the taboo set for the first time. At the instant a tagged customer enters the system its corresponding *sojourn process* is initiated. This process represents all customers (currently present in the system and possibly arriving in the future) that might have an impact on the sojourn time of the tagged customer. The sojourn process is finished when tagged customer leaves the system, i.e., when its service time has elapsed or when it is ruled out of the system. The generator matrix of the sojourn process can directly be derived from the generator matrix of the original state process by removing the state transitions that do not bias the tagged customer's sojourn time.

This approach yields the sojourn time (i.e., cycle time) PDF of a tagged customer given the initial state of the system upon its arrival. The entire cycle time PDF can then be obtained by unconditioning the previous result using the arrival probabilities (the probabilities the tagged customer to find a particular state upon its arrival). Fortunately, since we deal with a Markovian system, the arrival probabilities are exactly the same as the ordinary state probabilities because of the PASTA property (Wolff 1982).

The sojourn process is also a Markov process since its states are a subset of the states of the original Markov process. However, depending on the system, additional information on the system state has to be provided. Then, the corresponding state transitions have to be added to the sojourn process. Nevertheless, these transitions must not affect the Markovian property of the sojourn process.

Since the ordinary differential for the cycle time PDF obtained by Syski (1986) cannot be solved explicitly, Kühn (1972) confined himself to the calculation of moments of cycle time. Using the relation between a probability density function, its Laplace-Transform, and its moments, he showed that the cycle time moments, $\mathrm{E}[T_i^L]$, $L \geq 1$, conditioned on state i found upon arrival are given by the following recursive equation:

$$-\alpha_i \cdot \mathrm{E}[T_i^L] - \sum_{k \notin \mathcal{H}} \alpha_{ik} \cdot \mathrm{E}[T_k^L] = L \cdot \mathrm{E}[T_i^{L-1}], \qquad i \notin \mathcal{H}, \tag{4.3}$$

where $\mathrm{E}[T_k^0] = 1$. Here, the α_{ik}'s are the transition rates of the sojourn process and $\mathcal{H}$ represents the taboo set. Further, α_i is defined by

$$\alpha_i = -\sum_{j \neq i,\ j \notin \mathcal{H}} \alpha_{ij} - \beta_i, \qquad i \notin \mathcal{H}, \tag{4.4}$$

where β_i denotes the transition rate of the sojourn process to end from state i. Unconditioning Eqn. (4.3) with respect to the arrival probabilities p_i^a, any cycle time moment

can simply be calculated as follows:

$$\mathrm{E}[T^L] \;=\; \sum_{i \notin \mathcal{H}} p_i^a \cdot \mathrm{E}[T_i^L]. \tag{4.5}$$

We point out that although Eqns. (4.3) and (4.5) are not well-known, they are of the same fundamental importance for the sojourn process and the moments of sojourn times as Eqn. (4.2) for the original state process and the mean sojourn time.

Note that the method just described also facilitates the analysis of the basic alternative-tool models where the time to fail, F, and the time to repair, R, follow the generalized exponential distribution. Dallery (1994) has shown that if F and R follow this distribution, then the queuing system can be analyzed with the same complexity as if F and R would be exponentially distributed by simply adjusting their means, i.e., $\mathrm{E}[F]$ and $\mathrm{E}[R]$. The only required assumption is that failures have to be of the preemptive-resume type. This assumption is trivially fulfilled since, in case of exponentially distributed service times, there is no difference between failures of the preemptive-resume and the preemptive-repeat type in terms of the total processing time a customer experiences.

The application of Syksi and Kühn's moment calculation method requires the infinitesimal generator matrices has to be generated first. We did this directly by considering all state transitions caused by customer arrivals, service completions, failures, and repairs. Alternatively, one might also use a Queuing Petri Net (see Bause and Kritzinger 1996) to serve this purpose. Secondly, the arrival probabilities have to be determined. Although these are given by the ordinary state probabilities due to the PASTA property, care has to be taken since the sojourn process can be initiated from different states (see Section 4.2.2).

Admittedly, this approach represents a brute force method to calculate moments of cycle times. However, as van Dijk (1993, p. 38) points out, in general, solving the global balance equations (i.e., the first part of Eqn. (4.2)) does not yield explicit solutions such that one has to rely on numerical procedures or approximations. On the other hand, remembering that a queuing network has product form if its states posses the partial balance property (van Dijk 1993), one may try to replace the basic alternative processing queuing system by a product form queuing network to simplify the analysis. Unfortunately, this attempt fails since the partial balance property is violated by some of the system states.

The analysis of the two alternative-tools models is troublesome since for each of the models, the calculation of moments of cycle times has separately to be carried out for each customer class. Note that the analysis of a single class requires two generator matrices as well as the arrival probabilities need to be known. For these reasons, due to the lack of space, we do not report the generator matrices and the arrival probabilities

in detail. Instead, we confine ourselves to an example and illustrate the derivation of the moments of cycle time of class–2 customers in the handover model. The remaining derivations can be carried out in an analogous manner.

Finally, we point out that the variance of waiting time cannot simply be derived by subtracting the service time variance from the variance of cycle time since the control mechanism leads to a considerable correlation between waiting time and service time. Therefore, if one is also interested in the variance of waiting time, one has additionally to derive the waiting processes.

4.2 An Example: Class-2 Customers in the Handover Model

4.2.1 Mean Cycle Times

To capture the stochastic behavior of the handover model we need the following four-dimensional Markov process $\{X(t),\ t \in \mathbb{R}\}$,

$$X(t) \ = \ (Z_1(t), Q_1(t), Z_2(t), Q_2(t)), \tag{4.6}$$

where the individual random variables are defined as follows:

Z_1 **: state of server 1.** This random variable can take the following values:

U_0 : server 1 is idle;
U_i : service of a class-i customer, $i = 1, 2$;
D_0 : breakdown occurred in state U_0;
D_i : breakdown during service of a class-i customer, $i = 1, 2$.

State D_0 occurs only in systems with time-dependent failures since an idle server cannot break down by definition. On the other hand states D_1 and D_2 occur in systems with time-dependent and operation-dependent failures as well.

Z_2 **: state of server 2.** Since server 2 is dedicated to customers of class 2 only we have $Z_2 \in \{\mathrm{U}_0,\ \mathrm{U}_2,\ \mathrm{D}_0,\ \mathrm{D}_2\}$.

Q_i **: number of customers in queue** i**,** $i = 1, 2$**.** Note that, if we deal with time-dependent failures only, $Z_i = \mathrm{U}_0$ implies $Q_i = 0$ since an idle server instantaneously starts service upon the arrival of a customer.

To calculate mean values of certain performance measure requires to solve the stationary equations from Eqn. (4.2). This immediately yields server utilizations, blocking probabilities, and mean number of queued customers by summing the state probabilities appropriately. Then employing Little's law, the calculation of the mean cycle time is trivial. Since the method of Syski and Kühn facilitates the analysis of any moment of cycle time we are also able to calculate the mean cycle time by this method. Hence, we do not report these performance measures in detail.

4.2.2 Moments of Cycle Time

To obtain the sojourn process of class-2 customers, $\{\tilde{X}(t),\ t \in \mathbb{R}\}$, the original state space has to be extended since it is not sufficient to consider only the state of both servers and the number of customers currently waiting in both queues. The sojourn process of class-2 customers has additionally to account for the current position of the tagged customer (waiting in queue or receiving service) and the number of customers ahead of the tagged customer. Hence, we have

$$\tilde{X}(t) \ = \ (L_2(t), Z_1(t), Q_1(t), Z_2(t), V_2(t), Q_2(t)), \tag{4.7}$$

where the random variables incorporated in this equation are defined as follows:

L_2 : **residence of the tagged customer.** The tagged customer may either be waiting in queue 2 or may be receiving service by one of the two servers. Hence, random variable L_2 can take the following values:

W_2 : tagged customer waits in queue 2,
U_i : tagged customer receives service from server i, $i = 1, 2$.
D_i : tagged customer resides in server i which is broken down, $i = 1, 2$.

Z_1 : **state of server 1.** This random variable takes the same values as in the original state process $\{X(t),\ t \in \mathbb{R}\}$. Thus, we have $Z_1 \in \{\mathrm{U}_0, \mathrm{U}_1, \mathrm{U}_2, \mathrm{D}_0, \mathrm{D}_1, \mathrm{D}_2\}$.

Q_1 : **number of customers in queue 1.** The sojourn process has to account also for this number since service of class-2 customers is triggered on server 1 only if queue 1 is empty.

Z_2 : **state of server 2.** Server 2 cannot be idle during the sojourn time of the tagged customer. Hence, state U_0 does not appear and we have $Z_2 \in \{\mathrm{U}_2, \mathrm{D}_0, \mathrm{D}_2\}$.

V_2 : **number of customers in queue 2 prior to the tagged customer.** The tagged customer is only admitted to the system if queue 2 is not fully occupied upon its arrival, i.e., $V_2 \in \{0, \ldots, S_2 - 1\}$.

Q_2 : **number of customers in queue 2.** Since the tagged customer and V_2 customers ahead of it wait in queue 2, we have $Q_2 \geq V_2 + 1$.

It remains to derive the state probabilities upon the arrival of the tagged customer. Here, to simplify the notation, we use the abbreviation $k = (t_2, z_1, q_1, z_2, v_2, q_2)$ for any state $k \in \tilde{\mathcal{X}}$ where $\tilde{\mathcal{X}}$ denotes the set of states of the sojourn process. Further, we use the abbreviation $(\mathrm{D}_1, \ldots)$ to refer to a system state $(\mathrm{D}_1, z_1, q_1, z_2, v_2, q_2)$. The arrival probabilities following have to conditioned on the case, that the tagged customer is admitted to the system, i.e., not rejected.

The tagged customer receives immediately service if it finds server 2 idle upon its arrival. Hence, we have

$$p_0^a(\mathrm{U}_2,\dots) \;=\; \sum_{l\,=\,(z_1,q_1,\mathrm{U}_0,q_2)} \frac{p(l)}{1-p_2^b}, \tag{4.8}$$

where p_2^b is the blocking probability of class-2 customers.

Given the control parameter B set to 0, the tagged customer may immediately be processed by server 1 if server 2 is busy. To make this happen, server 1 has to be idle and queue 2 has to be empty. This situation does not occur when $B \neq 0$. Thus, the corresponding arrival probability is

$$p_0^a(\mathrm{U}_1,\dots) \;=\; \sum_{\substack{l\,=\,(\mathrm{U}_0,q_1,z_2,0)\\ B=0\ \wedge\ z_2\neq \mathrm{U}_0}} \frac{p(l)}{1-p_2^b}. \tag{4.9}$$

Since the tagged customer cannot occupy a server broken down upon beginning of its sojourn process we have

$$p_0^a(\mathrm{D}_i,\dots) \;=\; 0, \qquad i=1,2. \tag{4.10}$$

The sojourn process cannot be initiated with $q_2 \neq v_2+1$ since tagged customer is always the tail of the queue upon its arrival. Hence, the corresponding arrival probability is zero:

$$p_0^a(k) \;=\; 0, \qquad \forall k:\ l_2=\mathrm{W}_2\ \wedge\ q_2\neq v_2+1. \tag{4.11}$$

The initial state of the sojourn process can be entered from more than one predecessor states:

$$p_0^a(k) \;=\; \frac{p(\mathrm{U}_2,0,z_2,B-1)\;+\;p(\mathrm{U}_0,0,z_2,B)}{1-p_2^b}, \tag{4.12}$$
$$\forall k:\ l_2=\mathrm{W}_2\ \wedge\ q_2=v_2+1\ \wedge\ B>0\ \wedge\ z_1=\mathrm{U}_2\ \wedge\ q_1=0\ \wedge\ v_2=B-1.$$

The first term corresponds to the case where the length of queue 2 is incremented by one upon the arrival of the tagged customer, server 2 being busy and queue 1 being empty. Alternatively, server 1 can be idle upon the arrival of the tagged customer. If the length of queue 2 is equal to B at that time, the control mechanism triggers service of a class-2 customer on server 1. Hence, the number of customers in queue 2 is also B immediately after the arrival of the tagged customer.

The derivation of the remaining arrival probabilities is straightforward. The number of customers in queue 2 prior to the arrival of the tagged customer is simply the number of customers in front of the tagged customer just after its arrival. Hence,

$$p_0^a(k) = \frac{p(z_1, q_1, z_2, V_2)}{1 - p_2^b}, \qquad (4.13)$$
$$\forall k : \; l_2 = \mathrm{W}_2 \; \wedge \; q_2 = v_2 + 1 \; \wedge \; \neg(B \neq 0 \; \wedge \; z_1 = \mathrm{U}_2 \; \wedge \; q_1 = 0 \; \wedge \; q_2 = B).$$

5 Numerical Results

To compare the behavior of the two control mechanisms results are presented for the mean and coefficient of variation of sojourn time, the server utilizations, and the blocking probabilities. For this purpose, all service rates are set equal to 1, i.e., $\mu_{11} = \mu_{12} = \mu_2 = 1$. Further, customers arrive at buffers of size $S_1 = S_2 = 16$ with arrival rates to $\lambda_1 = 0.5$ and $\lambda_2 = 0.9$, respectively. The reason for the latter setting is that this investigation aims at studying the behavior of the control mechanisms in the high load regime. Server 1 experiences additional work load due to the alternative routing of class-2 customers. Hence, the arrival rate of class-1 customers is set to a relative low value to keep the corresponding blocking probability at a low level. Due to the finite capacity buffers in front of the servers, some customers get lost when the buffer is fully occupied upon their arrival.

Both servers are subject to operation-dependent breakdowns where the mean time between failures is 200 times larger than the mean service time. The time to repair is 20 times the mean service time. Therefore, the availability of the servers is approximately 90.91 %. The time between failures is exponentially distributed. Finally, the time to repair follows a GE distribution function with a coefficient of variation of 1.5.

In the following, all performance measures are plotted as functions of the control thresholds. To simplify the presentation of the results both control thresholds are assumed to be equal, i.e., $A = B = \theta$.

The impact of the control mechanisms on server utilization is depicted in Figure 2. As expected, the graphs indicate that the more customers of class 2 that take the alternative routing (i.e., increasing threshold in the takeover model and decreasing threshold in the handover model), the lower the utilization of server 2. The utilization of server 1 (server 2) approximately reaches its maximum value in the takeover (handover) model for the largest threshold. It is furthermore worth mentioning that a growing utilization leads to an increasing blocking probability.

Figure 3 shows the effect of the control schemes on blocking probabilities. The blocking probabilities stay on a moderate level for both control mechanisms and both cus-

tomer classes. Although the server utilizations differ significantly in the takeover model if threshold is set to 1, the corresponding blocking probabilities are approximately the same. Since the utilization of server 1 grows with increasing threshold A in the takeover model, the blocking probability of class-1 customers behaves the same. The utilization of server 2 and the blocking probability of class-2 customers behave exactly the opposite way. As far as the handover control mechanism is concerned, these considerations are analogous to the previous discussion with one exception. The lower threshold B is, the more customers of class 2 follow the alternative routing.

In Figure 4 mean sojourn times are plotted versus the control thresholds. First, the qualitative behavior of the graphs is not surprising and need no explanation. The quantitative behavior, however, is surprising for the following reasons. In the takeover model, if we compare the results for $\theta = 1$ and $\theta = 16$ the mean sojourn time $\mathrm{E}[T_2]$ of class-2 customers approximately decreases by 62.5 %. However, this reduction is at the expense of the mean cycle time of class-1 customers which increases by 250 % approximately. On the other hand, in the handover model, the difference between the two extremes $\theta = 0$ and $\theta = 15$ with respect to $\mathrm{E}[T_1]$ and $\mathrm{E}[T_2]$ is smaller. In this case, the mean cycle time of class-2 customers drops to 50 %, whereas the mean cycle time of class-1 customers grows by 115 %.

Figure 5 illustrates the effect of the control parameter θ on the cycle time coefficient of variation. The large magnitude of the CoVs is primarily due to server breakdowns and secondarily to the variability of arrival and service times. Furthermore, as the graphs indicate, if the mean cycle times stay on a moderate level (see customer class 1 in handover model and customer class 2 in the takeover model), the CoVs of cycle times are comparably large. Furthermore, the results for customer class 2 in handover model and customer class 1 in the takeover model exhibit that the cycle time CoVs decrease if the mean cycle times grow significantly.

This behavior can simply be explained by taking into account the definition of the coefficient of variation. First, the standard deviation of sojourn times seems to remain approximately the same for different control thresholds. This statement holds, in particular, for class-1 customers in the handover model and customers of class 2 in the takeover model (see Figures 4 and 5) since there are only minor changes in terms of the corresponding mean sojourn times and CoVs of sojourn times. The same conclusion can be applied to class-1 customers in the takeover model and customers of class 2 in the handover model. In these cases, the mean sojourn times increase tremendously, whereas the CoVs of sojourn times fall considerably with increasing threshold θ. Hence, the standard deviations of sojourn times remain on approximately the same level by the definition of the coefficient of variation.

As far as the mean and the coefficient of variation of cycle time of both customer

classes is concerned, it is furthermore worth noting that the handover control is always *better* for class-1 customers but is is also *worse* for class-2 customers. This behavior can be explained by the fact that, in the handover model, customers are routed to the alternative tool only if the buffer in front of it is empty.

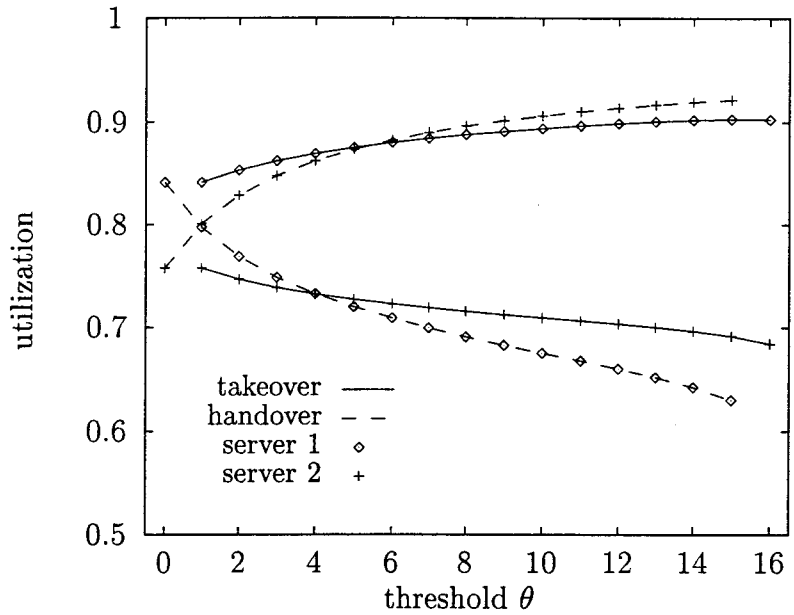

Figure 2 *Server utilization vs. control thresholds*

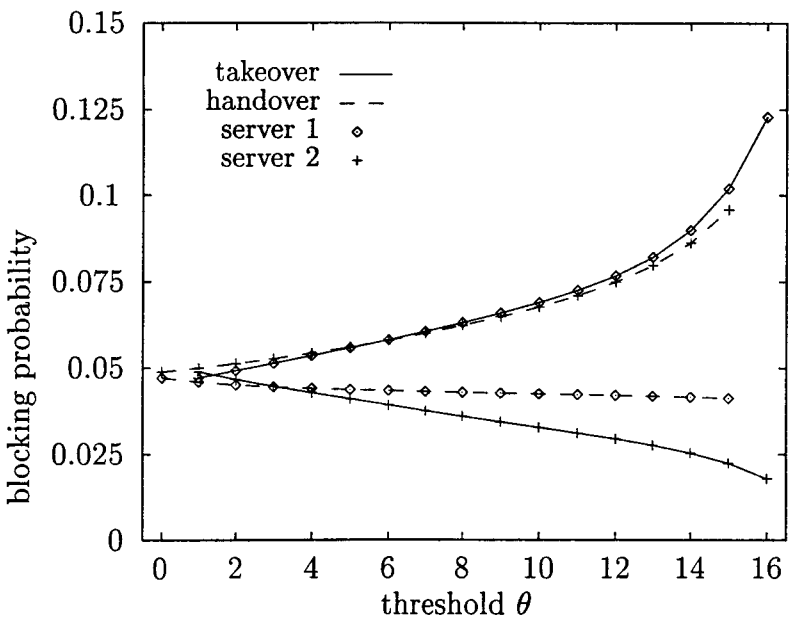

Figure 3 *Blocking probability vs. control thresholds*

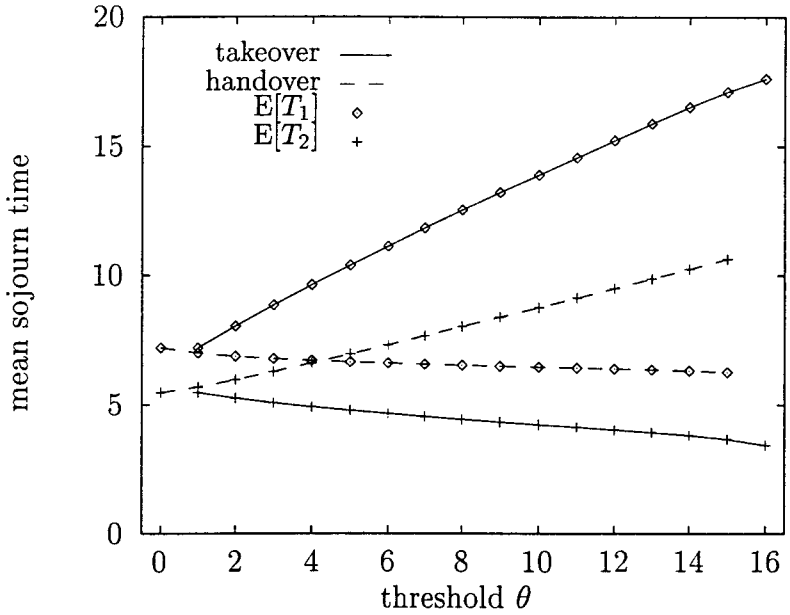

Figure 4 *Mean sojourn time vs. control thresholds*

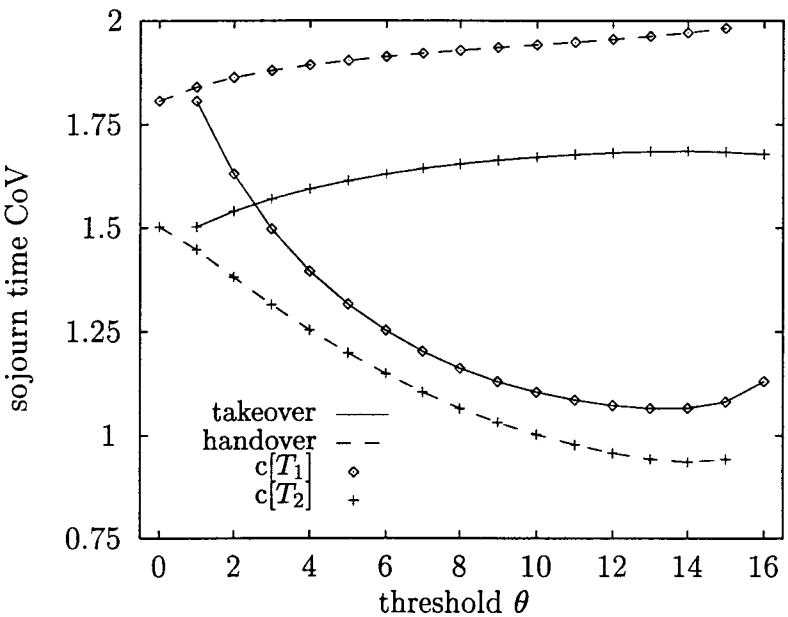

Figure 5 *Sojourn time CoV vs. control thresholds*

Additional experiments have shown the same results if the arrival rate of class-1 customers is increased up to the arrival rate of class-2 customers and also if there are no server breakdowns.

6 Conclusion

We investigated a basic queuing system with alternative processing. We applied two control mechanisms to this system to dispatch customers to an alternative tool. The results for an asymmetric system with different arrival rates exhibit that both control mechanisms achieve a reduction of the mean cycle time of the customers that can follow alternative routes (class-2 customers) only at the expense of a tremendous increase of the mean cycle time of the remaining customers (class-1 customers). As far as the cycle time coefficient of variation is concerned, the results are the better, the higher the mean cycle time is. Since in manufacturing practice the mean cycle time is of major importance, we suggest to set the control threshold to those values that enable the shortest mean cycle time. Finally, in terms of class-2 customers, the results of the takeover control are always inferior compared to the handover control. For class-1 customers, the takeover control outperforms the handover control. This is due to the fact that the alternative routing is only followed if the alternative tool is idle.

It remains for future research to consider alternative processing work centers consisting more than two tools. Although the method of Syski and Kühn theoretically facilitates the calculation of moments of cycle times for such systems too, the problem in practice will be to deal with the tremendous number of states of the corresponding stochastic processes. Furthermore, the manual derivation of these process and the arrival probabilities as well are prone to errors. Hence, higher level modeling techniques like Petri nets are needed to generate at least the generator matrices in an automatic way. Another interesting topic would be the queuing analysis of alternative tool models with non-exponentially distributed arrival and service times.

Acknowledgement The author would like to thank Dr. Ottmar Gihr and Dr. John Fowler for many discussions on this topic. The programming efforts of Horst Makitta are greatly appreciated.

References

Balakrishnan, A. (1989). Preemptive scheduling of hybrid parallel machines. *Operations Research 37*(2), 301–313.

Bause, F. and P. S. Kritzinger (1996). *Stochastic Petri Nets — An Introduction to the Theory.* Vieweg.

Bohn, R. E. (1995, January). Noise and learning in semiconductor manufacturing. *Management Science 41*(1), 31–42.

Dallery, Y. (1994). On modeling failure and repair times in stochastic models of manufacturing systems using generalized exponential distributions. *Queueing Systems 15*, 199–209.

de Ridder, L. and B. Rodriguez (1994). Measurement and improvement of manufacturing capacities (MIMAC): Definition of capacity loss factors. Technical report, Nimble NV, Maaltecentrum, Derbystraat 313, 9051 St-Denijs-Westrem (Gent), Belgium.

Elsayed, E. A. and A. Bastani (1985). General solutions of the jockeying problem. *European Journal of Operational Research 22*, 387–396.

Fordyce, K. and G. G. Sullivan (1994). Logistics management system (LMS): Integrating decision technologies for dispatch scheduling in semiconductor manufacturing. In M. Zweben and M. S. Fox (Eds.), *Intelligent Scheduling*, pp. 473–516. San Francisco, CA: Morgan Kaufmann.

Gise, P. and R. Blanchard (1986). *Modern semiconductor fabrication technology.* Englewood Cliffs: Prentice-Hall.

Hopp, W. J. and M. L. Spearman (1996). *Factory Physics: The Foundations of Manufacturing Management.* Irwin.

Kao, E. P. C. and C. Lin (1990). A matrix-geometric solution of the jockeying problem. *European Journal of Operational Research 44*, 67–74.

Kempf, K. G. (1994). Intelligently scheduling semiconductor wafer fabrication. In M. Zweben and M. S. Fox (Eds.), *Intelligent Scheduling*, pp. 517–544. San Francisco, CA: Morgan Kaufman.

Kühn, P. J. (1972). *Über die Berechnung der Wartezeiten in Vermittlungs- und Rechnersystemen.* Ph. D. thesis, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, Germany. In German.

Leachman, R. C. and T. F. Carmon (1992). On capacity modeling for production planning with alternative machine types. *IIE Transactions 24*(2), 62–72.

Melamed, B. and M. Yadin (1984a). Numerical computation of sojourn-time distributions in queueing networks. *Journal of the ACM 31*(4), 839–854.

Melamed, B. and M. Yadin (1984b, July–August). Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes. *Operations Research 32*(4), 926–944.

Münch, W. v. (1993). *Einführung in die Halbleitertechnologie.* Stuttgart: B. G. Teubner. (In German).

Nelson, R. D. and T. K. Philips (1993). An approximation for the mean response time for shortest queue routing with general interarrival and service times. *Performance Evaluation 17*, 123–139.

Rohan, D. (1992). Resource sharing in capacity analysis. In *1992 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 39–42.

Syski, R. (1986). *Introduction to Congestion Theory in Telephone Systems* (2nd ed.). Elsevier. Originally published 1960 by Oliver & Boyd, Edinburgh, Scotland.

Sze, S. M. (1983). *VLSI Technology.* New York, NY: McGraw-Hill.

Takagi, H. (1991). *Queueing Analysis, Volume 1: Vacation and Priority Systems.* Amsterdam: North-Holland.

Uzsoy, R., C.-Y. Lee, and L. Martin-Vega (1994). A review of production planning and scheduling models in the semiconductor industry — Part II: Shop floor control. *IIE Transactions on Scheduling and Logistics 26*, 44–55.

van Dijk, N. M. (1993). *Queueing Networks and Product Forms – A Systems Approach.* Chichester: Wiley.

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research 30*, 223–231.

Xu, S. H., R. Righter, and J. G. Shanthikumar (1992, November-December). Optimal dynamic assignment of customers to heterogeneous servers in parallel. *Operations Research 40*(6), 1126–1138.

Zhao, Y. and W. K. Grassmann (1995). Queueing analysis of a jockeying model. *Operations Research 43*(3), 520–529.

Preprint-Reihe
Institut für Informatik
Universität Würzburg

Verantwortlich: Die Vorstände des Institutes für Informatik.

[90] U. Hertrampf. *On Simple Closure Properties of #P.* Oktober 1994.

[91] H. Vollmer und K.W. Wagner. *Recursion Theoretic Characterizations of Complexity Classes of Counting Functions.* November 1994.

[92] U. Hinsberger und R. Kolla. *Optimal Technology Mapping for Single Output Cells.* November 1994.

[93] W. Nöth und R. Kolla. *Optimal Synthesis of Fanoutfree Functions.* November 1994.

[94] M. Mittler und R. Müller. *Sojourn Time Distribution of the Asymmetric $M/M/1//N$ - System with LCFS-PR Service.* November 1994.

[95] M. Ritter. *Performance Analysis of the Dual Cell Spacer in ATM Systems.* November 1994.

[96] M. Beaudry. *Recognition of Nonregular Languages by Finite Groupoids.* Dezember 1994.

[97] O. Rose und M. Ritter. *A New Approach for the Dimensioning of Policing Functions for MPEG-Video Sources in ATM-Systems.* Januar 1995.

[98] T. Dabs und J. Schoof. *A Graphical User Interface For Genetic Algorithms.* Februar 1995.

[99] M.R. Frater und O. Rose. *Cell Loss Analysis of Broadband Switching Systems Carrying VBR Video.* Februar 1995.

[100] U. Hertrampf, H. Vollmer und K.W. Wagner. *On the Power of Number-Theoretic Operations with Respect to Counting.* Januar 1995.

[101] O. Rose. *Statistical Properties of MPEG Video Traffic and their Impact on Traffic Modeling in ATM Systems.* Februar 1995.

[102] M. Mittler und R. Müller. *Moment Approximation in Product Form Queueing Networks.* Februar 1995.

[103] D. Roos und K.W. Wagner. *On the Power of Bio-Computers.* Februar 1995.

[104] N. Gerlich und M. Tangemann. *Towards a Channel Allocation Scheme for SDMA-based Mobile Communication Systems.* Februar 1995.

[105] A. Schömig und M. Kahnt. *Vergleich zweier Analysemethoden zur Leistungsbewertung von Kanban Systemen.* Februar 1995.

[106] M. Mittler, M. Purm und O. Gihr. *Set Management: Synchronization of Prefabricated Parts before Assembly.* März 1995.

[107] A. Schömig und M. Mittler. *Autocorrelation of Cycle Times in Semiconductor Manufacturing Systems.* März 1995.

[108] A. Schömig und M. Kahnt. *Performance Modelling of Pull Manufacturing Systems with Batch Servers and Assembly-like Structure.* März 1995.

[109] M. Mittler, N. Gerlich und A. Schömig. *Reducing the Variance of Cycle Times in Semiconductor Manufacturing Systems.* April 1995.

[110] A. Schömig und M. Kahnt. *A note on the Application of Marie's Method for Queueing Networks with Batch Servers.* April 1995.

[111] F. Puppe, M. Daniel und G. Seidel. *Qualifizierende Arbeitsgestaltung mit tutoriellen Expertensystemen für technische Diagnoseaufgaben.* April 1995.

[112] G. Buntrock, und G. Niemann. *Weak Growing Context-Sensitive Grammars.* Mai 1995.

[113] J. García and M. Ritter. *Determination of Traffic Parameters for VPs Carrying Delay-Sensitive Traffic.* Mai 1995.