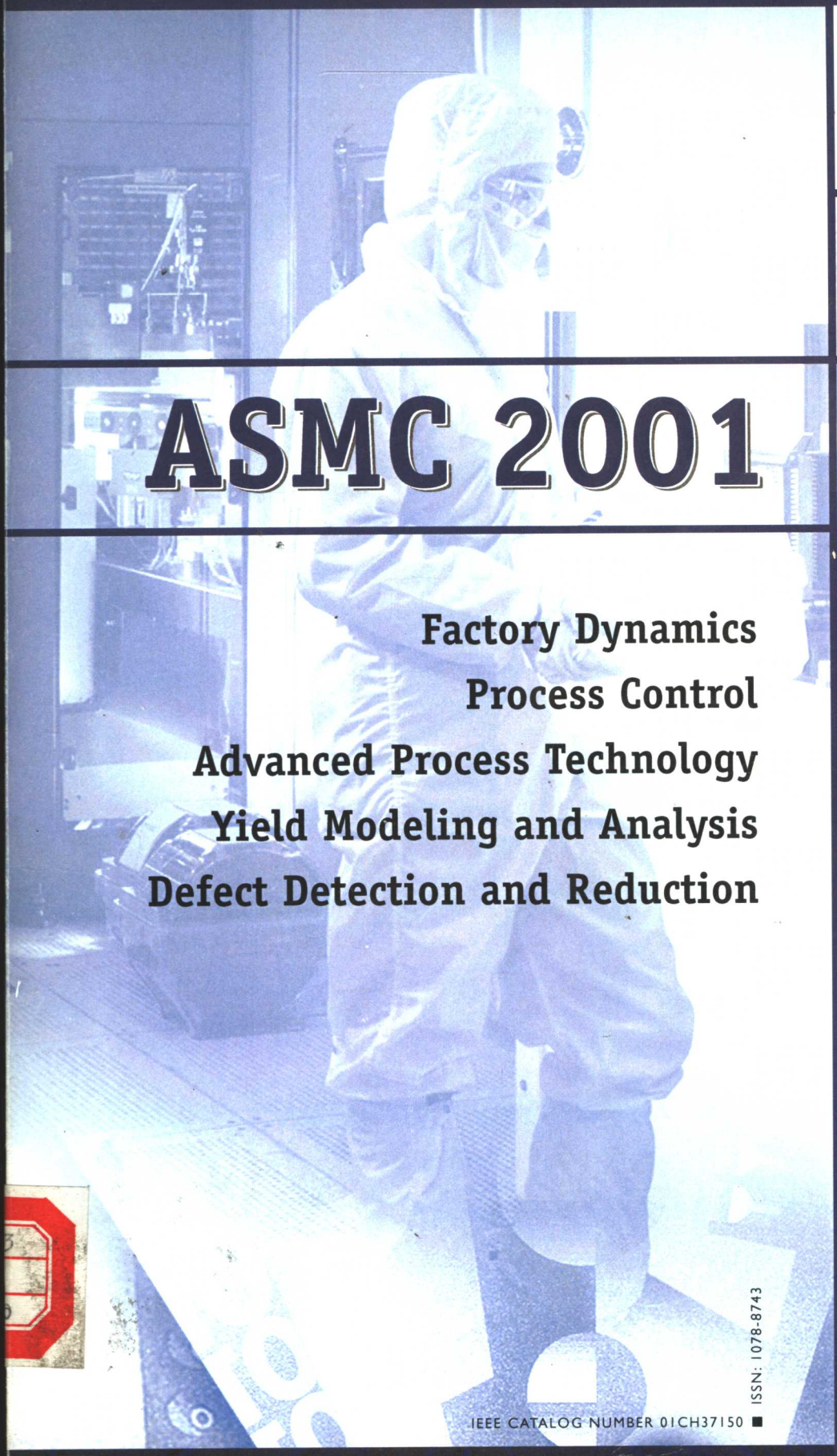
The 12th Annual IEEE/SEMI® Advanced Semiconductor Manufacturing Conference



Advanced Semiconductor Manufacturing Conference and Workshop

World Class Manufacturing

IEEE/SEMI

ASMC 2001

Factory Dynamics
Process Control
Advanced Process Technology
Yield Modeling and Analysis
Defect Detection and Reduction

April 23–24

International Congress Centre

Munich, Germany

Sponsored by:

ELECTRON DEVICES SOCIETY · CPMT · IEEE Networking the World · MICR

Produced by:

semi Semiconductor Equipment and Materials International

ISSN: 1078-8743

IEEE CATALOG NUMBER 01CH37150

ABOUT THE ASMC SPONSORS

Semiconductor Equipment and Materials International Semiconductor Equipment and Materials International (SEMI®) is an international trade association representing more than 2,400 member companies that develop, manufacture and supply the technology, equipment, materials and services used to manufacture semiconductors and flat panel displays. The equipment and materials industry provides critical technologies that enable the development and manufacture of advanced semiconductors and displays which in turn enable the more than $1 trillion global electronics industry.

Since its founding, SEMI has evolved into an international organization committed to free trade and open markets. SEMI maintains a presence in every significant semiconductor and FPD manufacturing region in the world, with eleven offices in North America, Europe, Russia, Japan and Asia-Pacific. SEMI's primary goal is to help its members achieve their common corporate goals by expanding global marketing opportunities, by improving access to customers and to industry, government and civic leaders and by promoting the value and importance of our industry. SEMI achieves this through sponsoring trade expositions, executive conferences and technical and business education programs; by facilitating the development of international standards; through advocacy of public policy issues; and by providing timely industry information and statistics that enable our members to better understand prevailing and future economic, business and technology trends. For up-to-date information, visit the SEMI website at www.semi.org.

IEEE

Networking the World™

The Electron Devices Society (EDS) and the Components, Packaging, & Manufacturing Technology (CPMT) Society are two of the 36 technical societies within the Institute of Electrical and Electronics Engineers, which, in turn, is the largest professional engineering organization in the world with over 350,000 members. It is transnational, with conferences and chapters in most countries. EDS and CPMT also sponsor a number of other related conferences:

- VLSI Chip Packaging Workshop
- Semiconductor Thermal and Temperature Management (SEMI-THERM) Symposium
- International Electron Devices Meeting (IEDM)
- International Symposium on Semiconductor Manufacturing (ISSM)
- Intersociety Conference on Thermal Phenomena in Electronic Systems (I-THERM)
- Electronic Components & Technology Conference (ECTC)
- International Electronic Manufacturing Technology (IEMT) Symposium

Copies of past proceedings of some of these conferences are available for purchase. In addition, EDS and CPMTS publish the Transactions on Semiconductor Manufacturing, the archival journal in this field. We invite you to consider membership for one of the IEEE Societies and to participate with us in furthering advancements in these fields. If you already belong to another professional society, you can affiliate with EDS or CPMT at reduced fees. Please refer to the back cover for additional information.

MICRO

MICRO magazine is the ultimate focused editorial and advertising vehicle for professionals concerned with defect reduction, process control, and yield enhancement strategies in the semiconductor, semiconductor equipment and materials, and related advanced microelectronics manufacturing industries. The magazine and its web site - www.micromagazine.com - offer a comprehensive, practical, and timely mix of technical articles and industry and product technology news coverage to a global audience. Since 1983, MICRO has been a leader in providing reliable, in-depth information to its readers in the semiconductor manufacturing community. MICRO magazine -- proud cosponsor of ASMC 2001 and member of SEMI.

0-7803-6555-0/01/$10.00©2001 IEEE

2001 ASMC ORGANIZING COMMITTEE

Our special appreciation to the following people who together volunteered countless hours to the organization of the IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop:

GENERAL CHAIR	**William Miller,** IBM Microelectronics
CONFERENCE CHAIR	**Patricia Gabella,** International SEMATECH (AMD Assignee)
TECHNICAL CHAIR	**Mart Graef,** Philips Research

Javier Bonal, Agere Systems
Javier Castaño, Agere Systems
Thomas de Paly, AMD Dresden
James Dougan, Motorola
John Goodman, Entegris, Inc.
Dave Gross, Advanced Micro Devices
Terry Leslie, Dominion Semiconductor
Thomas Kaufmann, Fraunhofer Institute IPA
Yuri Karzhavin, White Oak Semiconductor
Peter Kücher, Infineon Technologies SC300 GmbH & Co. KG
Sergio Malvar, Agere Systems
Scott R. McClure, IBM Microelectronics
Winfried Meier, Nikon Precision Europe GmbH
Thomas Piliszczuk, KLA-Tencor France
Linda Sattler, Consultant
Walter Schoenleber, Applied Materials GmbH
Volker Tegeder, Infineon Technologies AG
Larg Weiland, PDF Solutions

0-7803-6555-0/01/$10.00©2001 IEEE

Introduction to ASMC 2001

At present, the semiconductor manufacturing industry is transferring to 300 mm wafer fabs. As a consequence, tool utilization is becoming more important than ever. Hence, many of the contributions at the conference deal with factory dynamics and yield control. There is a tendency for process technologies to become more and more generic, whereas products are diversifying (for instance as systems-on-chip). Does this mean that process technology is becoming a "commodity"? This question was actually addressed during a panel discussion at ASMC 2000. Although opinions vary, most people will agree that timely availability of new technologies is a key factor for modern semiconductor fabs. The big challenge is now how to remain cost-effective in a rapidly changing industrial environment.

This state of affairs constitutes the setting for the 12th Advanced Semiconductor Manufacturing Conference – and the first one to be held in Europe, in an appropriate conjunction with Semicon Europa 2001.
The subtitle of the conference is "Advancing the Science and Technology of Semiconductor Manufacturing". This theme should inspire the participants to discuss ideas on possible improvements of the semiconductor manufacturing process. Undoubtedly, the oral presentations and the posters at ASMC 2001 will provide a well-balanced overview of the state-of-the-art in semiconductor manufacturing.

The various sessions of ASMC 2001 have been organized to reflect the following themes:

- *Factory Dynamics:* Cycle time management and cost control are the key subjects here. The presentations in this session discuss how these topics are related to fab methodology and process issues.
- *Process Control:* Process control quantifies process parameters in IC fabrication. This session focuses on metrics and methodologies to ensure consistent manufacturing.
- *Advanced Process Technology:* New materials and device concepts are being introduced at an unprecedented rate. The presentations in this session discuss the consequences of advanced process options for manufacturing methodologies.
- *Yield Modeling and Analysis:* Yield enhancement tools and methods are the focus of these two sessions, which include also inspection methodologies and SPC for yield analysis.
- *Defect Detection and Reduction:* Defect detection and reduction is essential for yield improvement. This session focuses on tools and methodologies for the identification and analysis of defect sources.

In addition to the oral presentations, a number of posters will be presented in a separate session. These cover a wide range of subjects, varying from process technology to factory control. They have in common that they all contribute to the improvement of semiconductor manufacturing methods.

I wish you a very useful and enjoyable conference!

Mart Graef
Technical Chair ASMC 2001
Strategic Program Manager, Philips Semiconductors

0-7803-6555-0/01/$10.00©2001 IEEE

Table of Contents

Poster Session

Yield Modeling

†Abstract

Yield Modeling and Analysis 2

Defect Detection and Reduction

Quantifying Operational Time Variability: the Missing Parameter for Cycle Time Reduction

J.H. Jacobs, L.F.P. Etman, J.E. Rooda
Eindhoven University of Technology
Systems Engineering Group
P.O. Box 513, 5300 MB Eindhoven
The Netherlands
{j.h.jacobs, l.f.p.etman, j.e.rooda}@tue.nl
http://se.wtb.tue.nl

E.J.J. van Campen
Philips Semiconductors MOS4YOU
Gerstweg 2
6534 AE Nijmegen
The Netherlands
edgar.van.campen@philips.com

Abstract

Operational time variability is one of the key parameters determining the average cycle time of lots. Many different sources of variability can be identified such as equipment breakdowns, setup, and operator availability. However, an appropriate measure to quantify variability is missing. Measures such as the Overall Equipment Efficiency (OEE) in semiconductor industry are entirely based on mean value analysis and do not include variances.

The main contribution of this paper is the development of a new algorithm that enables to estimate the mean effective process time t_e and the coefficient of variation c_e^2 of a multiple machine equipment family from real fab data. The algorithm formalizes the effective process time definitions as given by Hopp and Spearman [1], and Sattler [2]. The algorithm quantifies the claims of machine capacity by lots, which includes time losses due to down time, setup time, or other irregularities. The estimated t_e and c_e^2 values can be interpreted in accordance with the well-known $G/G/m$ queueing relations. A test example as well as an elaborate case from semiconductor industry show the potential of the new effective process time (EPT) algorithm for cycle time reduction programs.

Keywords

Factory dynamics, equipment modeling, data extraction, manufacturing line performance, capacity and cycle time losses

1. Introduction

Equipment in semiconductor manufacturing is subject to many sources of variability. An important source is formed by equipment downs, which occur due to the highly complex and technologically advanced semiconductor manufacturing processes [3]. Many other corrupting operational influences are also present, such as batching, rework, setup, and operator availability. All together, they introduce a substantial amount of variability in the interarrival and operational times of the lots during their flow through the fab.

Queue times are mainly influenced by variability and utilization. Utilization is usually high in semiconductor industry in order to maximize productivity and minimize costs. In combination with large variability, high utilization leads to large cycle times of lots. Especially, multi-process multi-product IC manufacturers are under high pressure nowadays to reduce cycle times and improve delivery performance. Without changing utilization this can only be achieved by reduction of operational time variability. Therefore, identification and reduction of the main sources of variability is a key action to improve upon the compromise between throughput and cycle time.

Unfortunately, in semiconductor industry no measures for operational time variability are used. The overall equipment efficiency (OEE) has been introduced by SEMI [4]. This measure is based on mean values with respect to availability and productivity. It includes for example mean time between failure and mean time to repair to characterize equipment downs, but it fails to include variances. Hopp and Spearman ([1], Section 8.4) show with a simple example that, besides the average capacity, the fluctuations of capacity in time should also be included to make the correct conclusion on how well an equipment is performing. Taking into account average capacity only, may lead to the wrong conclusion.

A suitable measure that quantifies the total process time variability is still missing. Such a measure would be highly valuable in variability reduction programs. Sturm et al. [5] observed that it is impossible to measure each individual source of variability. Instead, they measured cycle time distributions at equipment families, and used these in their simulation model. However, in these distributions the effects of utilization and

0-7803-6555-0/01/$10.00©2001 IEEE

variability are mixed up. Another approach is proposed by Hopp and Spearman [1]. They introduce the so-called effective process time, and describe it as: the time seen by lots from a logistical point of view. Basically, the effective process time includes all time losses due to down, setup, rework, and any other source of variability. A similar description is given by Sattler [2] who defined the effective process time as: all cycle time except waiting for another lot, which includes waiting for machine down time and operator availability and a variety of other activities.

Sattler [2] noticed that her definition of effective process time is difficult to measure. The same difficulty holds for the description given by Hopp and Spearman [1]. But the basic idea of effective process time to include time losses does give a starting point to compute effective process times of lots if a list of events is available with arrival and departure times of lots and track-in and track-out data of equipment. Since semiconductor industry is highly automated, this data is generally available. We propose a new method to actually compute effective process times from such a data set. In this way we are able to estimate the mean and variance of the effective process time of an equipment family. This gives the desired quantification of operational time variability. The approach is illustrated using real fab data of a Philips Semiconductors wafer fabrication facility.

2. Performance Measurement

Wafer fabs combine uncertain yields and unreliable equipment in a re-entrant process flow. In order to improve equipment productivity, SEMI [4] defined the overall equipment effectiveness (OEE). OEE separates equipment productivity into three basic corrective action categories: availability, performance, and quality. Availability efficiency is the fraction of time that the machine is in a condition to perform its intended function. Performance efficiency is the fraction of machine uptime that the machine is processing actual units at theoretically efficient rates. Finally, quality efficiency is the theoretical production time for effective units divided by the theoretical production time of actual units. Typically, OEE includes several sources of variability such as down times or time spend on monitoring. However, OEE is only based on mean values.

Besides mean process time, the average queue time of wafers in the queues of machine families is determined by utilization and variability. Hopp and Spearman [1] use the following approximation for the average queue time of wafers in $G/G/m$ queuing system, where m denotes the number of identical machines:

$$t_q = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^{(\sqrt{2(m+1)}-1)}}{m(1-u)}\right)t_e \qquad (1)$$

with the utilization defined as:

$$u = \frac{t_e}{t_a m} \qquad (2)$$

The first term of Equation (1) represents the variability and is the sum of the squared coefficients of variation of the interarrival times c_a^2 and the process times c_e^2. The squared coefficient of variation is defined as the quotient of variance and the mean squared. Thus, $c_a^2 = \sigma_a^2/t_a^2$, and $c_e^2 = \sigma_e^2/t_e^2$, where t_a and t_e are the mean interarrival time and mean process time, respectively. Hopp and Spearman [1] use in this respect the effective process time paradigm: t_e and c_e^2 include the effects of operational time losses due to e.g. machine downs, setup, rework, and other irregularities. Compared with the theoretical process time t_0 this typically means $t_e > t_0$ and $c_e^2 > c_0^2$. In accordance with [1] we call $c_e^2 < 0.5$ lowly variable, $0.5 < c_e^2 < 1.75$ moderately variable, and $c_e^2 > 1.75$ highly variable.

Equation (1) clearly identifies the contribution of utilization and variability. Cycle time increases linearly with the squared coefficients of variation of interarrival and effective process times, and increases nonlinearly with utilization. To reduce the mean waiting time there are two possible courses of action. The first is to reduce the loss of capacity due to irregularities. This gives a smaller mean effective process time t_e, which also means a lower utilization. This part is covered by performance measures such as OEE. If the mean capacity loss cannot be further reduced, the second action is to reduce the variation of the irregularities, giving a smaller variability term. The OEE fails to cover this term.

An important property is that variability propagates through the fab, which occurs for two reasons. First the arriving wafer flow at a machine is transformed into a departing flow. Second, the arrival pattern of a specific machine is determined by the departure patterns of its predecessors. In [1] the following linking approximation is used:

$$c_d^2 = 1 + (1-u^2)(c_a^2 - 1) + \frac{u^2}{\sqrt{m}}(c_e^2 - 1) \qquad (3)$$

This means that for low utilizations, the flow variability of the departing wafers equals the variability of the arriving flow, while for high utilizations, the flow variability of the departing wafers equals the effective process time variability. To be cost effective, wafer fabs operate at high machine utilizations. Thus, reducing operational time variability at one machine will posi-

0-7803-6555-0/01/$10.00©2001 IEEE

tively influence the arriving wafer flow of its successors.

Equation (1) implies that the mean EPT and the corresponding squared coefficient of variation are two fundamental process parameters with respect to cycle time performance. To use these parameters as performance measures, t_e and c_e^2 have to be determined from actual fab data. As semiconductor industry is highly automated, much data on all sorts of events is available. Unfortunately, the definitions given by Hopp and Spearman [1] and Sattler [2] cannot be used to measure the effective process time from this data set.

3. How to Measure EPT?

We have formalized the EPT definition and propose a new algorithm to actually compute EPTs from real-time fab data. With an equipment family or workstation we mean one or more machines that perform a similar operation and that share a single queue. The EPT definitions of Hopp and Spearman [1], and Sattler [2] include the theoretical process time as well as setup time, breakdown, operator availability, and all other operational times due to variability effects. For the cycle time of a lot it is of no importance whether the lot is waiting for an operator or for a machine that is being monitored. Generally stated, EPT is seen as the total amount of time a lot could have been, or actually was, processed on a machine. If a machine has only 50% of its regular capacity, for example because one of its chambers is being maintained, this should also be reflected in the EPT.

The new algorithm is developed such that it enables to calculate the EPT of a workstation from a list of events. This list of events consists of the arrival and departure times of the lots at a certain workstation, and the machine identification number the lot has been processed on. We start with investigating the EPT definition for a single machine workstation with first-in-first-out (FIFO) dispatching. This EPT definition is extended to include other dispatching policies as well. Finally, the EPT definition is generalized to a multiple machine workstation.

3.1 Single machine, FIFO dispatching

Consider a workstation with FIFO dispatching that consists of a single queue and a single machine. This single machine setup is used to formalize the conceptual idea that the EPT is the total amount of time a lot could have been or actually was processed on the machine. The event history with respect to arrival and departure can be visualized by a Gantt chart. An example is presented in Figure 1. The Gantt chart shows four lots which were processed in FIFO order on a single

time	lot	event
0	1	Arrival
1	2	A
2	1	Depart
3	3	A
5	2	D
6	3	D
7	4	A
9	4	D

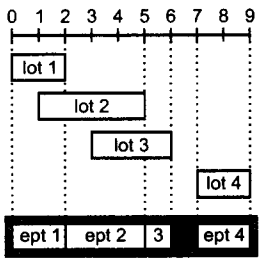


Figure 1: Single machine: FIFO dispatching

machine. The first lot arrived at $t = 0$ and departed at $t = 2$. The arrival and departure times of the other lots are depicted in the same way. Actual process times are not needed for determining the EPTs, and are therefore not depicted. The resulting EPTs are shown in the grey box at the bottom of Figure 1. The following paragraph explains how these EPTs have been determined.

Initially, no lots are present in the workstation, i.e. no lots are queued and no lot is in process. Since at $t = 0$ the first lot arrives at the workstation, this lot immediately claims capacity of the machine, independent of whether it is queued for a while or processed immediately. Before the first lot departs, a second lot arrives at $t = 1$. Since a FIFO dispatch policy is used, the first arrived lot still claims capacity of the machine. The second lot does not claim capacity of the machine, because the first lot already does. When the first lot has finished processing, and departs from the machine, it does not claim capacity of the machine anymore. The total amount of time this first lot has claimed capacity is called a realization of effective process time. From this point of time ($t = 2$), the second lot is the following lot that is processed next. Therefore, now this lot claims capacity of the machine until the lot departs. It does not make a difference if new lots are arriving, like the third lot at $t = 3$. The second EPT realization is then the time between the departure of the first lot and the second lot.

In general, for each lot that is processed next on the machine it holds that the total amount of time the lot was queued or processed between the departure of the previous lot and its own departure is a realization of effective process time. When enough individual EPTs have been determined, a complete EPT distribution arises. For the machine, an estimate of mean t_e and coefficient of variation c_e^2 can be calculated from the whole set of individual EPTs.

3.2 Single machine, general dispatching

If a single machine is considered, the lots in the system claim capacity of this particular machine. From

0-7803-6555-0/01/$10.00©2001 IEEE

the machine's point of view it does not matter which lot claims its capacity. If two lots are queued, capacity of the machine is claimed from the period the first lot has arrived in the queue. Whenever the second lot arrives, still capacity of the machine is claimed until one of both lots departs. Therefore we assume that the order in which the lots are processed does not have any influence on the EPT calculation. As a consequence, the EPT does not depend on the schedule, but only on the arrival and departure times of the lots. Thus, during the period that at least one lot is present in the family, the capacity of the machine is claimed until a lot departs from the family. The EPT calculation does not take into account *which* lot arrives or departs. This corresponds with observations that dispatch rules and overtaking by hot lots have hardly any effect on the average cycle time (see e.g. [6]).

In Figure 2 an event history and Gantt chart of a non-FIFO schedule is shown to illustrate how abovementioned rationale for general dispatching affects the EPT definition. Assume that lot 3 is a hot lot and, although lot 3 arrives after lot 2, it is processed before lot 2 is processed. Thus, when lot 1 leaves, initially, lot 2 claims capacity of the machine. After the arrival of lot 3, this hot lot claims capacity of the machine. But, for the machine it makes no difference which lot is processed next. During the complete time period between the departure of lot 1 and the departure of lot 3, a lot claims capacity of the machine and therefore the complete time period is a single EPT realization. It is of no concern whether this realization is based on the presence of a single or multiple lots. Notice that the Gantt chart of Figure 2 has equal arrival and departure times compared to Figure 1. Although the schedules produce the lots in different orders, the EPT calculation delivers equal EPT realizations.

Summarizing, the machine does not need to know which lot is claiming capacity. The EPT is the total amount of time a single lot or different lots are claiming capacity of the machine until a lot departs. Thus, whenever there is no lot in the workstation (queue empty and machine idle), capacity of the machine is not claimed. These time periods do not belong to EPT. But as soon as a new lot arrives, the next EPT realization will be the time between this arrival and the first next departure of any lot, not necessarily the first newly arrived. With lots present in the single machine workstation, the EPT is the time between two departures of two lots. This holds until no more lots are present in the family.

Algorithm SM – single machine workstation

An algorithm to calculate EPTs is proposed in Guarded Command Language in Figure 3. GCL is

time	lot	event
0	1	Arrival
1	2	A
2	1	Depart
3	3	A
5	3	D
6	2	D
7	4	A
9	4	D

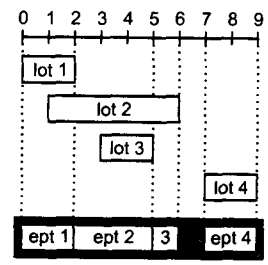


Figure 2: Single machine: general dispatching

summarized in the appendix and explained in detail in [7]. It is assumed that at the start of the algorithm the workstation is empty. The algorithm considers a single machine workstation and a general dispatch rule. First of all, the algorithm triggers for an event. When an event occurs, the algorithm reads (with the ? operator) the current time of the event τ and the type of the event ev. Each event ev can be an arrival of a lot (**A**) or a departure of a lot (**D**). The number of lots present in the family is denoted by n.

If a lot arrives, the workstation can be in two different states: (1) the family is empty, that means the number of lots n present in the system equals 0, or (2) the workstation is not empty, and thus $n > 0$. In the first case ($n = 0$), capacity of the machine has not been claimed until the lot arrives at time τ. From this point the capacity of the machine is claimed until a lot departs. Therefore, the start of EPT is set at $s := \tau$. In the second case, if a lot arrives and the family is not empty ($n > 0$), the start of EPT s has already been set by a lot arrived earlier. In that case, nothing has to be done. Finally, the number of lots present in the system has to be updated to the new value: $n := n + 1$.

If a lot departs, a realization of EPT can be calculated which equals the time between the actual time τ and the time the start of EPT was set: s. Thus, the realization of EPT equals $\tau - s$. This value is written (with the ! operator). Afterwards, the number of lots present in the system has to be updated: $n := n - 1$. Now again two different states can occur: (1) the workstation is empty, or (2) the workstation is not empty. If the workstation is empty, capacity of the machine is not claimed by a new lot and the start of the EPT is not set. Then s needs not to be set until a new lot arrives. If the workstation is not empty, capacity of the machine is immediately claimed by one of the lots still left in the system and s is set to the actual time.

In the practical case that lots are already present in the family when the algorithm is started, then the first departure of a lot cannot result in an EPT realization, since s has not been properly set. But, after this first departure, s can be properly set, and the algorithm can continue with all the other lots.

0-7803-6555-0/01/$10.00©2001 IEEE

```
[ n := 0
; *[ ?(τ, ev)
   : [ ev = A
       ⟶ [ n = 0 ⟶ s := τ
         [] n > 0 ⟶ skip
         ]
       ; n := n + 1
     [] ev = D
       ⟶ !τ − s
       ; n := n − 1
       ; [ n = 0 ⟶ skip
         [] n > 0 ⟶ s := τ
         ]
     ]
   ]
]
```

Figure 3: Algorithm SM – single machine

3.3 Multiple equipment

The EPT algorithm is generalized to cover multiple machines with general dispatching. The workstation now consists of a single queue which feeds a number of parallel machines. Again we follow the concept that a departure of a lot from a machine yields a new EPT realization. The EPT equals the total time capacity of this machine was claimed during this last departure. So, whenever a single lot is queued for or processed on a particular machine, capacity of this machine is claimed until a lot departs from this machine. This is similar to the single machine situation, and holds for each of the machines in the family. The number of machines for which capacity is claimed (m_e) should be equal to the minimum of the number of machines (m) and the number of lots (n) present in the family:

$$m_e = \min(m, n) \tag{4}$$

Usually if two lots are present in the workstation they will be processed by two different machines. Then it is clear that capacity of both machines is claimed ($m_e = 2$). But there are some other possibilities where it is less clear from which machine capacity is claimed. Imagine, for example, the following situation of a workstation with 2 machines. A lot is processed on the first machine and another lot is waiting to be processed on this first machine too. Capacity of this first machine can not be claimed twice, but according to Equation (4) capacity is still claimed for two machines. One could say that the second lot claims capacity of the second machine now, since it is the only available machine left. However, the actual value of the next EPT realization depends on whether or not a third lot will arrive before the first lot has finished and on which machine this third lot will be processed. This is explained in the following two paragraphs.

Consider the following scenario that no new lot arrives before the first one is finished, or that a new lot arrives that will also be processed on the first machine. In these situations, where the second machine stays idle, we follow Equation (4) and assume that the capacity of the second machine is claimed as long as at least two lots are present in the system. So, upon departure of a lot, we want the multiple machine EPT algorithm to compute the EPT realization according to the time this lot has claimed capacity of some machine, either the first one, or the second one, or both, as long as the claim is a continuous one. Consequently, an EPT realization in the multiple machine case cannot always be assigned to one particular machine.

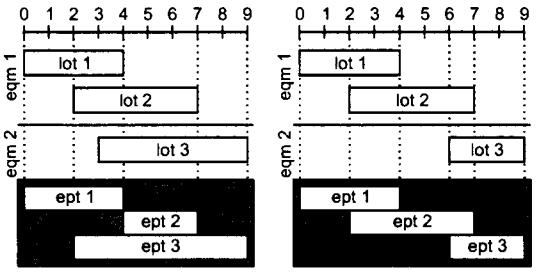


Figure 4: Multiple machines

We need the assumption of a continuous claim of capacity. This is explained by slightly changing the previous example. Consider again the two machines. One machine is processing a lot and capacity of this machine is claimed. Another lot is queued and will be processed in the future on the first machine too. As in the previous example, this second lot claims capacity of the second machine. Now imagine that the next lot that arrives will not be processed on the first machine, but will be processed on the second machine which has been idle so far. This third lot may arrive either before or after the first lot departs from the first machine, as shown in Figures 4(a) and 4(b), respectively. If it arrives before the first lot is finished but after the second lot has arrived (Figure 4(a)), then you might end up with a discontinuous claim of capacity and discontinuous EPTs if start and end times are not properly chosen: at time $t = 2$ lot 2 claims capacity from idle machine 2; then at $t = 3$ lot 3 arrives and takes over this claim of capacity of machine 2; for lot 2 the claim would end, and be resumed at $t = 4$ when machine 1 becomes available again. Instead, we describe it as: at time $t = 2$ a lot claims capacity from idle machine 2; at $t = 3$ another lot arrives and capacity of machine 2 is still being claimed; at $t = 4$ lot 1 departs which means the start of a new claim on machine 1; the respective EPTs run as long as the claims continue

0-7803-6555-0/01/$10.00©2001 IEEE

non interruptedly. The same consideration holds for the Gantt chart in Figure 4(b) with the difference that directly after the arrival of lot 2, this lot continuously claims capacity of the machines until it departs. In that case, the total time lot 2 is present in the family is a single EPT realization. Figure 4 illustrates the EPT realizations we obtain in this way at times $t = 7$ and $t = 9$ for both situations. Consequently, an EPT realization in the multiple machine situation cannot always be assigned to one particular lot. We already observed that for the single machine workstation.

Summarizing, for multiple machines capacity can be claimed in two different ways: (1) capacity is claimed by presence of a lot that will be processed on a machine. This means that it is clear which lot claims which machine, (2) capacity is claimed but it cannot be assigned to one specific lot or machine. In case of multiple idle machines it is even impossible to determine which machine is claimed and which not.

Algorithm MM – multiple machine workstation

The EPT algorithm for multiple machine workstations is presented in Figure 5. It is a generalization of Algorithm SM. Algorithm MM uses list ts, which is initially empty (a list is a vector of variable length). The algorithm uses ts to temporarily store all start values of new EPT realizations. From the time points in ts, capacity of machines is claimed. Besides list ts also tuple s is used to store start values of EPT realizations (a tuple is a vector of fixed length). The start values stored in s belong to capacity claims that can be assigned to specific machines. Thus, element $s.i$ is the start time that capacity of the i-th machine is claimed. From a certain time point when a value for $s.i$ can be set, its value is taken from list ts. In list ts all start values remain that cannot yet be assigned to a particular machine. This ensures continuous capacity claims.

A third variable is used in Algorithm MM: tuple nt. Element $nt.i$ of tuple nt equals the number of lots present in the family that is or will be dispatched into the i-th machine. The sum of $nt.i$ over all machines equals the number of lots n present in the workstation.

The algorithm again triggers on an event and determines besides the actual time τ and the event ev also the machine number i. In case of an arrival event, i is the number of the machine the lot that has been arrived will be processed on in the future. In case of a departure, i is the machine the lot was processed on.

At an arrival, Algorithm MM distinguishes four cases by combining two boolean expressions, $n \otimes m$ and $nt.i \otimes 0$, in which $\otimes$ denotes a relational operator:

- $n \otimes m$: If the number of lots is below the number of machines ($n < m$), a new EPT realization has to be started and the start time is added to the rear of list ts. If the number of present lots is larger than or equal to the number of machines in the family ($n \geq m$), capacity of all machines is claimed already. Therefore, no new EPT realization has to be started.
- $nt.i \otimes 0$: When a lot arrives for processing on machine i, an EPT start value can only be assigned to this specific machine if no other lot is already waiting for this machine or being processed on this machine. So, if $nt.i = 0$ is true, then $s.i$ is set to the head value of list ts, and the head of list ts is removed.

At a departure, the opposite holds:

- $n \otimes m$: a new EPT realization has to start directly only if $n \geq m$ to claim the machine that has become available. Then, the corresponding start value (τ) is added to list ts or set in $s.i$.
- $nt.i \otimes 0$: when a lot departs from machine i, a new EPT start value has to be assigned to this machine, only if other lots are waiting for this machine ($nt.i > 0$). This new EPT start value of $s.i$ depends on n: if $n < m$, $s.i$ becomes again the head of ts, but if $n \geq m$, other lots continuously claim capacity of all the other idle machines. Then $s.i$ can be set with the actual time τ, leaving ts in its original state.

```
[ n := 0; nt := ⟨0⟩^m; ts := []
; *[ ?⟨τ, ev, i⟩
   : [ ev = A
       ⟶ [ n < m ∧ nt.i = 0 ⟶ ts := ts ++ [τ]
                              ; s.i := hd(ts)
                              ; ts := tl(ts)
         [] n < m ∧ nt.i > 0 ⟶ ts := ts ++ [τ]
         [] n ≥ m ∧ nt.i = 0 ⟶ s.i := hd(ts)
                              ; ts := tl(ts)
         [] n ≥ m ∧ nt.i > 0 ⟶ skip
         ]
       ; n := n + 1; nt.i := nt.i + 1
     [] ev = D
       ⟶ !τ − s.i
       ; n := n − 1; nt.i := nt.i − 1
       ; [ n < m ∧ nt.i = 0 ⟶ skip
         [] n < m ∧ nt.i > 0 ⟶ s.i := hd(ts)
                              ; ts := tl(ts)
         [] n ≥ m ∧ nt.i = 0 ⟶ ts := ts ++ [τ]
         [] n ≥ m ∧ nt.i > 0 ⟶ s.i := τ
         ]
     ]
   ]
]
```

Figure 5: Algorithm MM – multiple machine

0-7803-6555-0/01/$10.00©2001 IEEE

Summarizing, if due to an event a new EPT realization has to be started, the corresponding time is added to the rear of list ts. If, however, it is possible to directly assign a new EPT start value to one specific machine, it is set to the oldest element of ts, and ts is updated.

4. Examples

This section describes a few test examples to validate the EPT definition and to show how the EPT can be interpreted. The validation is based on the following test setup. Consider an equipment family with multiple machines that suffer from one or more different sources of variability. A discrete-event simulation model of this family is implemented – using the χ language [7] – that explicitly includes each different source of variability in the model. Running the model gives an estimate of the expected cycle time CT as well as a list of generated arrival and departure events – for the test examples we carried out 10 simulation replications of 200,000 lots each. From this list of events we calculate the EPT realizations using algorithm MM, and estimate the mean effective process time t_e and coefficient of variation c_e^2. With these t_e and c_e^2 values a new gamma distribution is defined to replace the process time specifications of the machines in the original discrete-event model by the EPT based gamma distribution. Running this so-called metamodel gives another estimate of the expected cycle time CT^*, which can be compared with the CT of the original model. For the t_e and c_e^2 to be appropriate measures for capacity loss and variability, the CT and CT^* should behave similarly, that is, reducing e.g. the capacity variation in the initial model giving a lower CT value, should be represented by a reduced c_e^2 value with an accordingly decreased CT^* value in the metamodel. Ideally, in the $G/G/m$ queueing situation, CT and CT^* should have approximately equal values.

4.1 Unreliable machines

Consider a workstation with two identical unreliable machines, a single buffer, and Poisson arrival of lots. The mean theoretical process time of the machines is $t_0 = 0.8$ with a coefficient of variation of $c_0^2 = 0.25$. However, during processing a machine may break down and be temporarily unavailable for further processing (we assume that breaks do not occur when the machine is idle). Availability is a function of the mean process time to failure t_f and mean time to repair t_r and set in this example on a constant value of: $A = t_f/(t_f + t_r) = 0.8$. After repair, the machine is going up again and finishes the lot with the remaining process time. Both the failure times and repair times are exponentially distributed.

Table 1 shows the simulation results of the model with different settings for the failure and repair times and at various throughput levels. The throughput equals the arrival rate r_a and is the average number of lots that arrive per time unit, given by: $r_a = 1/t_a$, with t_a being the mean interarrival time. The t_e and c_e^2 values are obtained from Algorithm MM. Independent of the throughput the effective process time is estimated to be 1.0, which is correct since $t_e = t_0/A$. The squared coefficient of variation c_e^2 increases for increasing values of t_f and t_r. This reflects the previously mentioned statement of [1] that, for equal availability, machines with frequent but short outages are to be preferred to machines with infrequent but long outages. The estimated mean cycle time of the original model CT and the EPT based metamodel CT^* are approximately equal for all throughput levels and t_f and t_r values.

4.2 Unequal machines

In many practical cases, machines in an equipment family are often not (exactly) identical and may differ in the mean or variance of the process times. Let us therefore consider a workstation with two unequal machines. Both machines have gamma distributed process times with means $t_0(1)$ and $t_0(2)$, and squared coefficients of variation $c_0^2(1)$ and $c_0^2(2)$. Two cases are studied: (I) the machines have equal coefficients of variation $c_0^2(1) = c_0^2(2) = 0.5$, but different mean process times, (II) the machines have equal mean process times $t_0(1) = t_0(2) = 1.0$ but different coefficients of variation. The FIFO dispatch rule we use in the simulation study does not account for the difference in capacity or variability. If the two machines are both idle, they have an equal chance to start processing a lot.

The first case (I) considers a fast and a slow machine in parallel. We vary with different values for $t_0(1)$ and $t_0(2)$, but keep the mean capacity of the machines always at 2.0 lots/hour. Table 2 shows the results of three different settings for the process time of the fast ma-

Table 1: Unreliable machines

t_f/t_r	r_a	t_e	c_e^2	CT	CT^*
0.8/0.2	1.0	1.000	0.330	1.227	1.230
	1.4	1.000	0.331	1.642	1.653
	1.8	1.000	0.330	3.822	3.839
8.0/2.0	1.0	0.999	1.047	1.315	1.341
	1.4	1.000	1.049	1.968	1.984
	1.8	0.999	1.052	5.192	5.367
16.0/4.0	1.0	0.999	1.844	1.398	1.460
	1.4	1.000	1.844	2.266	2.254
	1.8	1.000	1.849	6.998	6.910

0-7803-6555-0/01/$10.00©2001 IEEE

Table 2: Unequal machines: a fast and a slow machine

$t_0(1)/t_0(2)$	r_a	t_e	c_e^2	CT	CT^*
0.9/1.125	1.0	1.004	0.518	1.261	1.267
	1.4	1.001	0.517	1.724	1.749
	1.8	1.000	0.520	4.232	4.209
0.75/1.5	1.0	1.038	0.685	1.302	1.373
	1.4	1.020	0.688	1.760	1.917
	1.8	1.006	0.687	4.277	4.831
0.60/3.0	1.0	1.172	1.636	1.471	1.969
	1.4	1.085	1.682	1.875	3.066
	1.8	1.023	1.698	4.272	8.827

Table 3: Unequal machines: machines with difference in variability of the process times

$c_0^2(1)/c_0^2(2)$	r_a	t_e	c_e^2	CT	CT^*
0.25/1.00	1.0	1.000	0.620	1.271	1.274
	1.4	1.000	0.624	1.783	1.789
	1.8	1.001	0.622	4.531	4.423
1.00/2.00	1.0	1.000	1.494	1.403	1.412
	1.4	1.000	1.495	2.180	2.210
	1.8	1.000	1.495	6.232	6.203

chine $t_0(1)$ and the slow machine $t_0(2)$. What we expect to happen is that for increasing difference in capacity of the machines, the variability of the effective process time of the workstation increases. However, the increase of cycle time CT is much less compared with the increase of c_e^2: the metamodel overestimates the effect of unequal capacity on the cycle time. This means that we should be careful with the interpretation of the c_e^2 if there is a large capacity difference in the machines of a workstation. Also a second effect can be observed; for larger capacity difference, the estimated mean effective process time tends to rise above the 1.0 value for decreasing throughput. This is due to the absence of a suitable dispatching rule in the model to account for the fact that the first machine is faster than the second. For low throughput we actually loose capacity due to unnecessary idle time of the fast machine. This is reflected in an increase of t_e.

The second case (II) considers a workstation with two machines with equal capacity rates set at 1 lot/hr, but with different variability in the process times. Table 3 shows the results for two settings of the variability, respectively machines with low and moderate variability, and machines with moderate and high variability. The estimated c_e^2 values equal the averages of the two SCVs of the process time of the machines. For both the original and the EPT based metamodel the estimated cycle times are approximately equal. The c_e^2 correctly represents the variability in the workstation.

5. Case Study

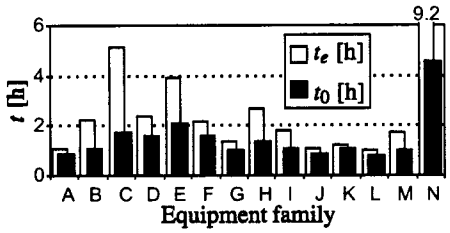
A case study from Philips Semiconductors is used to illustrate the potential of the EPT as a performance measure for cycle time reduction. The Philips wafer fabrication facility is a multi-product multi-process fab with more than 400 machines. Over 1.5 million track-in and track-out events were extracted from the manufacturing execution system (MES), covering a period of almost half a year in 1998. This data was converted to arrival and departure events, assuming that a track-out of a lot implies an arrival at the next process step neglecting transport time. Using this data, the mean effective process time t_e and the squared coefficient of variation c_e^2 were computed for fourteen single-lot equipment families (single-lot equipment process just one lot at at time).

The computed t_e and c_e^2 values are presented in Figures 6(a) and 6(b), and compared with the nominal process time t_0 and its variation c_0^2. The nominal process time only includes the time a lot has been actually in process. If the t_e of an equipment family is larger than t_0 this means that a considerable amount of capacity is lost due to irregularities. Similarly, the c_e^2 can be compared with c_0^2, revealing how much additional variability is effectively present. For most equipment families the natural variability c_0^2 is very small due to the highly automated and controlled processes in semiconductor manufacturing. Figure 6(b) shows that there is one equipment family, family N, with a c_0^2 larger than 1.0. This equipment family performs process capability measurements. Measurement time depends on the maturity of the process flow and can vary between 4 hours and 40 hours. For all other equipment families, the c_e^2 is significantly larger than c_0^2, showing how much variability is due to irregularities such as downs, setups, operator availability, etcetera.

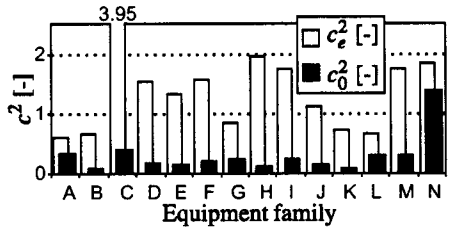
Figures 6(c), (d), and (e) show respectively the arrival coefficients of variation, the utilizations, and the cycle time factors for these equipment families. The cycle time factor is defined as the quotient of cycle time and nominal process time. It is an often used indicator in semiconductor industry to measure the contribution of waiting time to the cycle time. From Figure 6(e) we can observe that equipment family C, H, and N have a cycle time factor larger than 4.0, indicating a very bad cycle time performance. From Figures 6(a), (b), (c), and (d) we get more detailed information about this bad performance.

Equipment family C has a t_e value almost three times as large as t_0, an a c_e^2 of almost 4.0. The arrival coefficient of variation has a reasonable value of $c_a^2 = 1.3$. The combination of a high t_e and c_e^2 value causes the large cycle time, even though the utilization

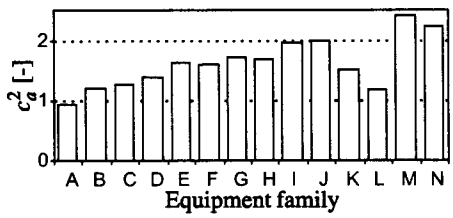
0-7803-6555-0/01/$10.00©2001 IEEE



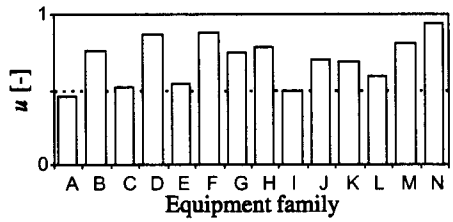
(a) Mean process times



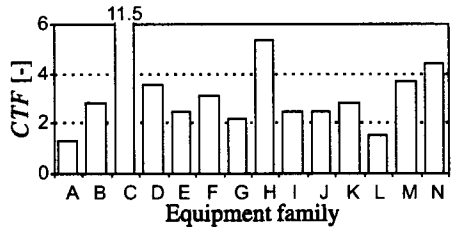
(b) Process time coefficients of variation



(c) Arrival coefficients of variation



(d) Utilization



(e) Cycle time factors

Figure 6: Case study MOS4YOU

is only 50%! So the problem is not a shortage of capacity, but the capacity loss and variability due to irregularities. Closer examination of family C should reveal the precise causes. The problem of family H mainly lies with high c_e^2 and c_a^2 values: variability in arrival and processing is the main cause. Finally, family N combines an extremely high t_e with an exceptionally high utilization. Reducing the capacity loss will reduce the t_e and u, and solve this problem. Even further improvement of family N can be obtained by reducing c_a^2 and c_e^2 which are in the high variability range.

6. Conclusion

A new algorithm has been developed that enables to estimate the mean effective process time and the corresponding squared coefficient of variation from data of the MES. The required data is arrival and departure times of lots at the workstations, and corresponding identification numbers of the machines the lots have been processed on. For our case study we were able to obtain this information from track-in and track-out data of the machines.

With this algorithm available, all four key parameters determining the cycle time of an equipment family can be measured: mean effective process time t_e, squared coefficient of variation of the effective process time c_e^2, the squared arrival coefficient of variation c_a^2, and the utilization u (see Equation (1) and (2)). The case shows how, using these four parameters, the main causes of large cycle times can be identified and which actions could be appropriate. Without estimates for t_e and c_e^2 this reasoning is not possible, and one can only rely on a cycle time based measure for each machine, such as the cycle time factor *CTF*, which however does not give a clue on the true cause of any large cycle time observed.

The EPT algorithm in this paper has been specifically developed for single-lot equipment families: only machines are considered that process one lot at a time. We also assume that the machines in an equipment family do not have large differences in processing times since this troubles the interpretation of the c_e^2. The latter assumption is usually not a serious restriction, but the single-lot assumption is: many equipment in semiconductor manufacturing process more than one lot at a time, such as litho machines or furnaces. A generalization of the EPT algorithm to include these types of machines is still being investigated. For the moment, EPT promises to be a very powerful tool in cycle time monitoring and improvement for semiconductor manufacturing.

0-7803-6555-0/01/$10.00©2001 IEEE

Acknowledgment

We would like to thank P.P. van Bakel for his contribution to the examples of Section 4.

The work of the first author is supported by the Dutch Technology Foundation STW, applied science division of NWO and the technology program of the Ministry of Economic Affairs of The Netherlands.

References

[1] J.W. Hopp and M.L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, Irwin/McGraw-Hill, New York, second edition, 2000.

[2] L. Sattler, "Using queueing curve approximations in a fab to determine productivity improvements," in *1996 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, Cambridge, MA, November 1996, pp. 140–145.

[3] R. Uzsoy, C.Y. Lee, and L.A. Martin-Vega, "A review of production planning and scheduling models in the semiconductor industry, part I: system characteristics, performance evaluation, and production planning," *IIE Transactions*, vol. 24, no. 4, pp. 47–61, 1992.

[4] V.A. Ames, J. Gililland, A. Konopka, and H. Barber, *Semiconductor manufacturing productivity; Overall Equipment Effectiveness (OEE) guidelines*, Technology transfer #950327443 A-GEN, Revision 1.0. Sematech, 1995.

[5] R. Sturm, T. Silvi, F. Frauenhoffer, and T. Treiber, "A simulation model for advanced release and order planning," *Future Fab International*, vol. 6, pp. 71–74, 1999.

[6] B. Ehteshami, R.G. Petrakian, and P.M. Shabe, "Trade-offs in cycle time management: hot lots," *IEEE Transactions on Semiconductor Manufacturing*, vol. 5, no. 2, pp. 101–106, 1992.

[7] J.E. Rooda, "Modelling of industrial systems," Lecture notes 4C530, Eindhoven University of Technology, 2000, http://se.wtb.tue.nl.

Appendix: Guarded command language

The two algorithms proposed in this paper are presented by statements from guarded command language [7]. This appendix gives an overview of the syntax rules for each of these statements. The following symbols are used to present the syntax rules:

x	variable
e, e'	expressions
b	boolean expression
l, l'	lists

Using these symbols the syntax rules for statement S become:

$S ::=$	skip	empty statement
	$\mid x := e$	assignment statement
	$\mid ?x$	receive statement
	$\mid !e$	send statement
	$\mid S; S$	catenation statement
	$\mid *[\ S\ ]$	repetition statement
	$\mid [\ G\ ]$	selection statement

Herein the syntax rules for selection statement G are:

$G ::=$	$b \longrightarrow S$	guarded command
	$\mid G\ [\ ]\ G$	selection

The skip statement is used when nothing has to be done. An assignment is used to set the value of an expression (such as: $n + 1$) to a variable. The receive and send statement wait for input and deliver output, respectively. Furthermore, a statement can consist of several statements (catenation). Repetition is used in this paper to endlessly repeat a set of statements. Selection is used to choose among several statements. In a selection statement several boolean expressions are evaluated (such as: $n < m$). For the boolean expressions that turn out to be true, one is chosen and the corresponding statements after the $\longrightarrow$ operator are executed.

Lists A list is a vector with variable length and denoted by squared brackets around the elements. A list with a single element can be constructed with $[e]$. A special list is the empty list: []. Concatenation ($l +\!\!+ l'$) is used to append one list l to another list l'. For a non-empty list, the first element is obtained by the function hd(l), the head of l. The list that appears after the removal of the first element is denoted by the function tl(l), the tail of l. This gives the following syntax rules L for lists:

$L ::=$	[]	empty list
	$\mid [e, \ldots, e]$	list constructor
	$\mid L +\!\!+ L$	concatenation
	$\mid$ hd(L)	first element of list
	$\mid$ tl(L)	tail of list

Tuples A tuple is a vector with fixed length and denoted by pointed brackets around the elements. The elements of a tuple can be found by an index. A dot is used as a projection operator to obtain the index in the tuple ($n.i$). To initialize a tuple, a constructor is used, such as $\langle 0 \rangle^m$, which constructs a tuple with m zero values. We have the following syntax rules T for tuples:

$T ::=$	$\langle e \rangle^{e'}$	tuple constructor
	$\mid T.e$	projection

0-7803-6555-0/01/$10.00©2001 IEEE