Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

A QUEUEING NETWORK BASED SYSTEM TO MODEL CAPACITY AND CYCLE TIME FOR SEMICONDUCTOR FABRICATION

Horst Zisgen

Industrial Software Solutions
IBM Systems and Technology Group
Mainz, 55131, GERMANY

Ingo Meents

Architecture and Development of Open Systems 2
IBM Systems and Technology Group
Mainz, 55131, GERMANY

Benjamin R. Wheeler

300mm Industrial Engineering
IBM Microelectronic Division
Hopewell Junction, NY 12533-6683, U.S.A.

Thomas Hanschke

Institute of Mathematics
Clausthal University of Technology
Clausthal-Zellerfeld, 38678, GERMANY

ABSTRACT

In today's competitive semiconductor business environment, wafer manufacturers are facing continuous pressure to accurately predict cycle time and tool utilization, gauge the impact of changes in capacity available, assess the impact of changes in product mix and quantity, and determine action plans to improve operational performance. Discrete Event Simulation (DES) is a widely used approach to perform such an analysis. However, DES has some inherent shortcomings for these planning tasks. Analytical models, like queueing networks, have much shorter response times and additional advantages compared to DES. But due to the complexity of semiconductor manufacturing systems (SMS) queueing models were not able to model all the peculiarities of those. This paper provides an overview of the main features of the IBM Enterprise Production planning and Optimization System (EPOS), a queueing network based system, which closes this gap. EPOS has been in use in the 300mm fabrication of IBM in Fishkill for more than 2 years and has turned out to be an invaluable tool to analyze the trade-offs of cycle time and capacity within this complex environment.

1 INTRODUCTION

In today's competitive semiconductor business environment, wafer manufacturers are facing continuous pressure to accurately predict cycle time and tool utilization, gauge the impact of changes in capacity available, assess the impact of changes in product mix and quantity, and determine action plans to improve operational performance. It is a cutting edge industry where pricing premiums are placed on the latest technology and being first to market often leads to more market share and higher revenue. Furthermore, the competitive market allows the customer to choose a supplier not only based on price and quality but also on lead time. Reduced cycle time is also highly desirable in order to enable faster yield learning, resulting in an increase in good chips per wafer earlier in a development cycle. Thus responsiveness is a key to success. And, if possible, this has to be considered already as early as possible in the fab design.

The biggest challenge in gauging and improving operational performance is the nature of the semiconductor manufacturing process itself – characterized by complex reentrant flows, large variations in raw process times, differences in batch sizes for tool sets, cascading tool sets, and a rich set of complexity in the tools themselves. This results in a significant portion of lead time being non-productive queue time (waiting to be serviced) and a well known trade-off between reducing cycle time and increasing equipment utilization sometimes known as the operating curve.

These challenging characteristics have limited the ability of the modeling community over the past 40 years to provide fab planners modeling software that meets their requirements for the speed of execution, ease of use, and accuracy to answer the key business question about gauging and improving operation performance.

Fab level discrete event simulations (DES) have been tried but model maintenance requirements, runtime constraints, and the limitation of only being able to investigate one scenario at a time have limited any real application

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

to very high level aggregated runs with limited accuracy. Apart from the run time issue, there are other inherent disadvantages of DES that limit its applicability to capacity planning of semiconductor manufacturing systems (SMS). By DES one can only arrive at meaningful results for the highest utilized tool as bottleneck. Further bottlenecks become visible after the most utilized tool is resolved and additional capacity is added to modeled resources. Furthermore, the DES run is not able to calculate the percentage of overload. This can only be figured out in a laborious and time consuming iteration of simulation runs. But especially in the case of the tactical capacity planning regarding the demand for a year in the future and new technologies the capacity planner needs that information and wants to perform what-if analysis quickly.

Optimization methods have been successful in handling cascading for tool planning, like in IBM's 200mm fab in Burlington, Vermont (Bermon and Hood 1999), but are unable to determine cycle time.

From this perspective analytical models based on queueing theory are the appropriate choice. They are able to provide the capacity analysis results including bottleneck overload percentage etc. and integrate this with the corresponding cycle time calculations. As pointed out in (Shanthikumar et al. 2007) analytical modeling systems using various queueing models (solutions) have proved insufficient of adequately capturing the complexity of SMSs in the past.

In this paper the EPOS system of IBM is presented which closes this gap and allows for this type of modeling. EPOS is a queueing network based simulation system for tactical and operational production planning and production management which can be integrated with the fab MES to capture routes, tools, raw process times, rework rates, and WIP. The maintenance of input planning parameters is supported by a continuous statistical monitoring feedback loop within the fab reporting system.

EPOS is based on advanced queueing theory algorithms developed by IBM and is currently implemented and running in IBM's 300mm fab in East Fishkill, NY. These algorithms take into account typical manufacturing characteristics of semiconductor fabs, like batch processing, process and downtime variability, rework, sampling and scrap rates, and varying product lot sizes. They use open queueing networks in order to build the models. These networks include an approximation of $G^X/G(b,b)/c$ which is based on an application of the renewal counting process and on diffusion approximation. The accuracy of the approximation has been shown - besides the practical experience - by a comprehensive comparison of the analytical outcome with the corresponding discrete event simulation results.

The rest of this paper is structured as follows: Section 2 describes the queueing model characteristics and how they meet the requirements of SMSs. The WIP movement prediction based on fluid models is discussed in the third section. In section 4 we present the implementation of the queueing model as system into the IT landscape of the 300mm line, the established business process and the results gained in IBM's fab in East Fishkill. Finally section 5 provides the conclusion and remarks on future work.

2 QUEUEING MODEL CHARACTERISTICS

2.1 General Model Characteristics

In SMSs batch processing is a widely spread, especially for tools like furnaces and wet cleaning equipment. This special way of performing an operation requires special queueing models taking into account that wafers to be collected up to a certain number (called batch size). Wafers are not only performed in batches but they are also moved as lots, so called Front Opening Unified Pods (FOUPs), between subsequent operations. The filling degree of the FOUPs does not need to be equal. It depends on the product group and other parameters. Thus the arrival stream of wafers at an equipment could be composed of FOUPs carrying a different number of wafers. These FOUPs could be grouped into batches which could be bigger, equal or smaller than the incoming lot sizes. From a modeling perspective this means that the arrival stream has to be modeled as bulks with an arbitrarily distributed size. The service process on the other hand has to be modeled as batch process too. For a general introduction into queueing systems with bulk arrivals and batch service refer to (Chaudhry and Templeton 1983). Thus the key is to model the way of batch processing and batch creation of different incoming lots appropriately. There are some approximations for batch processing in the literature, like (Bitran and Tirupati 1989) and (Chiamsiri and Leonard 1981). The later one provides a model with bulk arrivals and batch service for the $(1,b)$-rule based on a diffusion approximation. But in this model a wafer can join a running process as long as the process batch is not filled up to its maximum, which is usually not the case on the real shop floor. Bitran and Tirupati on the other hand focus only on the departures at batch processing work stations. In EPOS we have implemented a server queue with generally distributed interarrival times I^X of arriving bulks of an arbitrarily distributed size X. The service process is modeled with generally distributed service times S taking place with a fixed batch size b, listed as $G^X/G(b,b)$-queue. Let be $\lambda^X = 1/E(I^X)$ and $\mu = 1/E(S)$.

For a $G^X/G(b,b)/1$-queue Zisgen has developed in (Zisgen 1999) a diffusion approximation yielding to the following formula for the average number of wafers $E(Q)$ waiting in queue

$$E(Q) = \rho \left[\frac{\tilde{\rho}\hat{\rho}}{1-\hat{\rho}} + \tilde{\rho}b - \frac{E(X)}{b}\frac{\tilde{\rho}\hat{\rho}}{1-\hat{\rho}} + \tilde{\rho}b - \frac{b-1}{2} \right]$$

where $\hat{\rho} = e^{\gamma}$ with

$$\gamma = \frac{2\beta}{\alpha} = \frac{2(\lambda^X E(X) - \mu b)}{(Var(X) + E(X)^2 C^2(I^X))\lambda + b^2 C^2(S)\mu}$$

and

$$\tilde{\rho} = \frac{1}{\gamma}\frac{2\Lambda}{\gamma\alpha} R(\hat{\rho}^{b-1} - \hat{\rho})$$

and R is the probability of having less than b wafers in the system and Λ the time needed to fill the system up to b wafers. Furthermore $C^2(S)$ is the squared coefficient of variation of the service time S and $C^2(I^X)$ the squared coefficient of variation of the interarrival time I^X.

In the multiple server queue Hanschke (Hanschke 2006) has shown an approximation for the average number of wafers in $G^X/G(b,b)/c$ queue by modifying the Allan Cuneen approximation by a modulation of the batch arrival stream at the $G^X/G(b,b)/c$ queue applying renewal theory. Let I^b be the arrival rate of a complete batch at the server and $\lambda^b = 1/E(I^b)$. This yields to

$$E(Q) = \frac{\rho^2 b(C^2(I^b) + C^2(S))}{2(1-\rho)} ,$$

where $\rho = \lambda^b/\mu$ and

$$C^2(I^b) = E(X)(C^2(X) + C^2(I))/b .$$

The fixed batch size has the shortcoming that in the reality on the shop floor batches are not necessarily filled up to their maximum. In some case technical constraints might yield to a lower filling level, e.g. sometimes the number of wafers within the FOUPs locked at the load ports of a Furnace is less than the maximum batch size of the furnace. Or sometimes one has to wait unreasonably long to fill up a batch to its maximum since the products to be grouped together have only low volumes or due to some other incidents on the shop floor causing that lots that would fit into the batching pattern will not show up soon. Therefore EPOS adjusts the fixed batch size b to a so called effective batch size b_{eff} which is calculated dynamically on the basis of a maximum waiting time for batches to be filled up or due to other tool parameters, like the number of buffer slots (refer to 2.5) and FOUP filling degrees per product. This adjusted effective batch size is then used as b in the queueing formula and thus the capacity as well as the lead time calculation is adjusted.

2.2 Multiple Server Queues and Chamber Tools

In general one-of-a-kind equipments are the exception on the shop floor of a SMS. In the opposite, usually equivalent or similar equipments are grouped together in a sector. These equipments interact as alternates. Alternate tools require a queueing model based on multiple-server queues. Therefore Conners et.al. provide in (Conners et al. 1996) a model based on tool groups assuming to consist only of identical tools. But often within a tool group not all tools are qualified to perform the same operation which requires a more flexible approach in order to get the correct utilizations and queues. Therefore in EPOS tools are modeled as $G^X/G(b,b)/1$ queues if their set of operations is unique or as $G^X/G(b,b)/c$ server server queues, where c is the number of identical servers at the multi-server queue. In the later case the concept of batch processing with bulk arrivals was extended to the multiple-server case by exchanging ρ by Erlang's loss formula

$$P_c = \frac{\frac{(c\rho)^c}{c!}\frac{1}{1-\rho}}{\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!} + \left(\frac{(c\rho)^c}{c!}\frac{1}{1-\rho}\right)}$$

as an approximation of the probability that all server are busy.

This replacement yields to

$$E(Q) = \frac{\rho P_c b(C^2(I^b) + C^2(S))}{2(1-\rho)} .$$

Some process equipment consists of multiple chambers. These chambers are either processing the wafers in parallel and are acting more or less as equivalent servers within a tool. Or these chambers differ from each other and have different operations assigned to them. In the later case EPOS handles each chamber as an independent server and is considered as an individual tool. In the case that the chambers are identical there are specific tool types in EPOS used to cover these specialities (refer to section 2.5).

2.3 Multi Product- Multi-Class Open Queueing network

The major goal of the development of EPOS was the integration of capacity planning and cycle time planning for the entire fab and not only for isolated equipments.

In order to cover that goal the fabrication is considered as an open queueing network. A decomposition approach, enhanced on the basis of (Pujolle and Ai 1986) and (Gelenbe and Pujolle 1987), is used to determine the queuelengths for each equipment in the network. Cycle time as an additional performance measurement is derived by Little's rule (refer to (Little 1961)). Let be S the service time at a work station and I the interarrival time. The needed traffic rates and the traffic variability are calculated on the basis of the following approximation of the interdeparture times

$$D^b = \begin{cases} S & \text{with probability } \rho \\ I^* + S & \text{with probability } 1-\rho \end{cases}$$

where I^* is the time required to allocate a batch of size b, which yields the mean interdeparture time

$$E(D^b) = E(I^*)$$

and the squared coefficient of variation of the interdeparture times

$$C^2(D^b) = \rho^2 C^2(S) + (1-\rho)\rho + (1-\rho)C^2(I^*) \; .$$

Splitting this departure process and superpositioning the interarrival process per workcenter yields to a linear system of equations for the mean interarrival times per FOUP $E(I^X)$ and to a linear system of equations for the corresponding squared coefficient of variation $C^2(I^X)$. For details refer to (Hanschke and Zisgen 2005).

By this the queueing network in EPOS models the traffic of FOUPs and not that of individual wafer movement.

2.4 Load Balancing and Routing Probabilities

Unlike other approaches, like those in Conners et. al. (Conners et al. 1996), EPOS is modeling individual tools instead of tool groups. The main difference is that in the case of tool groups the assumption is that all tools within a tool group are enabled to perform exactly the same set of operations.

But usually all the equipments on the shop floor have their own set of operations they are capable to perform, e.g. due to different process qualifications or customer requirements. This impacts directly the capacity and other performance indicators of these equipments. Therefore in EPOS each tool can have an individual set of operations assigned to. Sometimes even individual chambers in a tool have a distinct set of operations. In this case a chamber is modeled as a separate tool. But this flexibility raises up the question of how to appoint the routing probabilities which are a pre-requisite in order to calculate the queue-lengths per equipment in the queueing network. Obviously these routing probabilities have a significant impact on the performance and have to be chosen appropriately. This is done in EPOS by a linear program which minimizes the sum of mutual utilization differences by choosing the appropriate routing probabilities with respect to the given tool dedications. Thus the load of tools within a class is most balanced, whereby a class of tools is defined as a set of tools which share at least on process step with at least one other tool in the class. (Kramer and Meents 2001). This approach is justified by the goal of the operations management of the fab to avoid utilization peaks for certain tools and thus keeping the lead time and the x-factor for the overall fab as low as possible.

2.4.1 Main Flow and Rework Steps

The routing is defined as a sequence of operations which can be performed at one or several alternative tools. In the case that rework is needed a lot (FOUP) is allowed to branch off the main flow and to proceed with a sequence of rework operations before it re-enters the main flow at a well-defined point. A FOUP can be sent to rework multiple times. The rework rates are gathered from the history tables in the MES (refer to section 4.1).

2.4.2 Scrap

Srap can occur in different types, like lot scrap, wafer scrap and bad dies marked in the wafer map. Bad dies do not have any impact on cycle time since the wafer has to processed anyhow. All other scrap rates are converted to lot scrap, allowing product lot sizes to be treated as fixed values. By doing this, different products can have different FOUP sizes in EPOS, but lot sizes do not decrease over the length of the product route. Obviously this is the most seldom scrap event but by this approach the lot sizes can be supposed to be fixed. But different products can have different FOUP sizes in EPOS. The scrap rates are also gathered from the MES (refer to section 4.1).

2.5 Tool Type Modeling

Having defined the traffic equations and the general way of handling the batch processing above the next step is to decompose the network into single tools and to find appropriate tool models for each tool type usually found on a semiconductor shop floor. Because of the variety of process equipment on the shop floor of a SMS, like furnace, lithography equipment, wet cleaner, etch tools, etc., one can not lump all these types together and model them with one specific server queue type. Therefore in EPOS we have defined a set of tool types from a queueing modeling point of view. Currently there are four general tool types defined and implemented which are capable to model all the equipment installed on the shop floor in the fab in East Fishkill, NY. The tool types are

1. Lot based equipment
2. Pipeline equipment
3. Batching equipment
4. Metrology equipment

Based on the specific tool type characteristics which have to be taken into account the mean process cycle time

$E(S)$ of the corresponding queue server is adjusted. The process time in EPOS is split into a mean fixed time $E(F)$ per batch and a mean variable time $E(V)$ per wafer. The fixed time could be the time which is needed to setup the tool or the chamber into the right state done up-front the process. The variable time is the process time needed per wafer. Thus EPOS can consider dependencies caused by the different filling degree of FOUPs or batch sizes. The mean process time per batch is computed as

$$E(S) = E(F) + E(V) \cdot b \,,$$

where b is the number of wafers in the processed batch.

2.5.1 Lot based equipment

Lot based tools are the most common ones on the shop floor. These tools perform all the wafers of an arriving FOUP at a time. They can have κ equal chambers which can process batches independently at the same speed. In the case that the number of wafers in an arriving FOUP L, is greater than $b \cdot \kappa$ the FOUP has to be split and processed in $\lceil L/(b \cdot \kappa) \rceil$ runs. Vice versa if $L < b \cdot \kappa$ lots can be mixed in the tool in the manner that wafers of a FOUP arriving at a busy but not fully loaded equipment can get started. The variable process time V_t is adjusted accordingly.

2.5.2 Pipeline equipment

Pipeline tools process wafers in a sequence of different process steps, e.g. a sequence of several chemical baths the wafers have to stay in for a different period of time. A subsequent wafer can be released into the tool while the first one is still in the tool but not before a trigger time t has passed. The trigger time is usually the cycle time of the longest process step. Pipeline tools are modeled as tandem queues. The first queue gets the trigger time assigned as cycle time. Since there is no queue between the two servers of the tandem the second queue is modeled as an infinite server with its cycle time being the sum of the remaining process step's process time. In order to get full FOUPs departing the tandem queue the second server has the batch size of the FOUP.

2.5.3 Batching equipment

Batching equipments, like furnace tools, are usually batch tools where the batch size is a multiple of a lot or FOUP size. Unfortunately batches can not always be filled up to the maximum batch size due to special constraints. E.g. the number of internal buffer slots could constrain the batch size. Furthermore in cases of low volume products it might happen that one has to wait a long time until enough FOUPs carrying that product have shown up to fill up a batch to its maximum. Therefore an effective batch size b_{eff} is calculated based on the following parameters

- Number of buffer slots p
- Maximum waiting time to fill up a batch

2.5.4 Metrology equipment

Metrology tools often work on a sample basis. There are two kinds of samples, lot samples and wafer samples. In the first case lots get picked for sampling with a sample rate s_r and in the later case a sample number s_n of wafers out of one lot is chosen for the measurement.

2.6 Machine outages

Especially in semiconductor fabrication machine outages can not be neglected. In EPOS those outages get incorporated into the mean service time at a tool via the availability of that tool in an analogous manner to (Gaver 1962). The availability is defined as $R = MTBF/(E(D)+MTBF)$, where $E(D)$ is the mean down time and MTBF the mean time between failures of the tool. Supposing that the time between to outages is exponentially distributed with parameter $\omega = 1/MTBF$. This yields to the mean completion time of the tool $E(C) = E(S)/R$. Accordingly the squared coefficient of the completion time is

$$C^2(C) = C^2(S) + \frac{R(1-R)E(D)(1+C^2(D))}{E(S)} \,.$$

3 WIP MOVEMENT PREDICTION BY FLUID MODELS

The queueing networks used in EPOS are based on the assumption that the system is in a stationary state. This assumption has been proofed reasonable in the daily business process for mid and long term planning. However, these models have their shortcomings in the context of short term forecasts taking into account the current state of the shop floor. In the operational business of running a semiconductor fab the planners have to answer questions like how the WIP will move downstream based on actual WIP positions in the fab or how fast WIP bubbles can be worked off. In order to tackle these operational planning issue a fluid model based algorithm was developed. By choosing a fluid model based approach the existing queueing model can be used as basis for the fluid model.

The fluid model is on the same level of granularity as the queueing model or even more granular. E.g. tool dedications on chamber level taken into account. By their nature operational questions require short response times. In order to keep the runtime short the scheduling of the fluids is rule-based instead of using time consuming linear

programming, like in (Conners et al. 1994). Furthermore, in practice it is often more than difficult to get the appropriate cost to parameterize the objective function for the linear programs. The rules implemented allow prioritization on EC-Level (or job class) as well as thresholds for specific throughputs per product or for WIP levels in general. To setup the fluid model the current WIP positions are fetched from the Manufacturing Execution System (MES) and fed into EPOS. This snap shot includes also prioritization information of each lot and some additional information for selecting certain lots. This queried information can be adopted or changed manually by the EPOS user interface, e.g. to perform what/if-scenarios. The fluid model approach is a pure deterministic approach. Therefore it is used to conduct short term analysis only. But since the runtime of the model is pretty fast and since the tight integration of EPOS into the IT systems allows to create the model in a short period of time fluid model runs can be set up quickly whenever the conditions on the shop floor have changed significantly. Thus the lack of randomness in the model can be overcome by repeating iterations. For details refer to (Meents and Zisgen 2004).

4 IMPLEMENTATION

4.1 IT Infrastructure

In addition to the challenges involved in mathematically modeling a complex SMS, there is also the non-trivial challenge of handling the massive amount of data required to feed the model. It is of course important to initialize the model by populating it with data from the manufacturing line, but it is also crucial to update model parameters to reflect process and equipment changes to ensure accuracy. This can only be achieved by embedding the mathematic modeling system into the overall IT landscape of the manufacturing facility. With tight integration between the simulation model and the fab MES (SiView in IBM 300mm), a large amount of data, such as process times, tool assignments, etc., can be automatically generated and maintained. Still, the engineering community may prefer to review and manually enter some modeling parameters. Also, for expected performance improvements or capital strategies, model adjustments for future time periods may be required. For this manual data editing, the EPOS Graphic User Interface can be used to conveniently navigate to the desired parameters and make individual record or mass updates. In the EPOS installation at IBM's 300mm fab in East Fishkill, NY, approximately 80% of the nearly 250,000 records in the model are automatically generated and maintained, while the remaining 20% are manually updated. Simulation results are available as dynamic web reports in the IBM Intranet and can be used by this easily for analysis, meetings, or management reviews.

Figure 1 shows the data flow for EPOS implemented in IBM's 300mm fab.

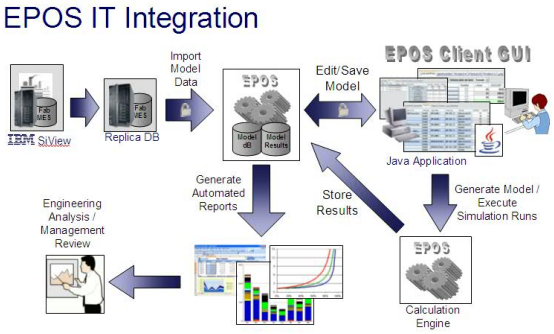

Figure 1: Planning process data flow

4.2 Inegration into the business process of production planning

Along with integration into the IT environment, it is also essential to have the modeling system and analysis results be integrated into fab operations and business processes. For management of model inputs, the Java-based GUI for EPOS allows the user community to edit data in a password protected, secure environment. Each user has a unique ID and permission settings, allowing easy change tracking and access control. On the model output side, all simulation results are output to a queryable database, and many standard dynamic web reports are available for analysis and reporting. Figure 2 represents a flow chart illustrating the processes around parameter control, simulation modeling, and results analysis.

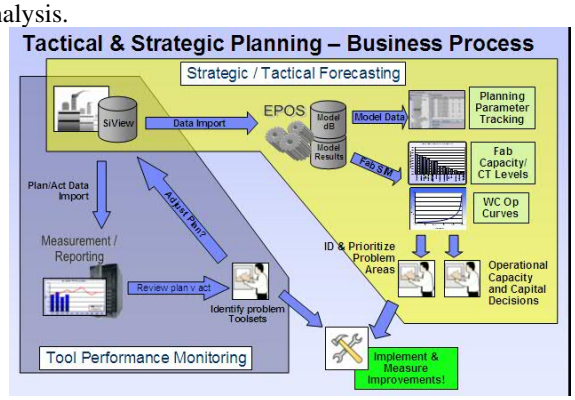

Figure 2: EPOS capacity planning process

4.3 Observed Benefits

At IBM's 300mm wafer fab, the EPOS system has completely replaced legacy spreadsheet-based models and been used as the exclusive fab planning system for more than 2

years. The implementation has brought about a paradigm shift in the general understanding of the trade-offs between capacity utilization and fab cycle time and armed the planning community with analysis capabilities that were previously not possible. Today, capital investment strategies are made not only by focusing on tools which are heavily utilized, but also considering second tier tools which may have a large contribution to overall cycle time due to high variability or multiple mask levels. In many cases, equipment which ordinarily would have been purchased due to utilization forecasts that are higher than an arbitrary threshold, have been supplanted with the purchase of tools which were less utilized, but had a higher contribution to overall fab cycle time, and were less expensive. This shift in capital strategy to focusing on impact to cycle time using EPOS rather than only the most highly utilized tools has already saved multi millions of dollars in capital costs. In addition to capacity planning, EPOS is used to easily and quickly forecast fab WIP levels and product lead times for a given product mix and volume. This is particularly important in an environment with high product differentiation and sensitivity to mix changes. Today, with a given volume plan, the East Fishkill model has been able to consistently predict average fab WIP and x-factor to within 10 percent.

Because the model is analytic in nature, unlike a DES, optimization techniques can also be applied to quickly guide decisions concerning fab loading strategies. For instance, given product margin data, the system can be used to determine the most profitable mix and volume which can be fed back through the MRP system or supply chain organization. Another advantage of the analytic model is the ability to automatically generate an empirical plot of wafer starts vs. product cycle time, or fab operating curve, with just one simulation run. This capability can be used both tactically and strategically for determining the maximum fab loading, while not exceeding a desired product cycle time. The graph in Figure 3 shows an operating curve for the entire fab created by EPOS. This type of trade-off analysis is commonly used by fab management in determining factory operating policies over different time horizons.

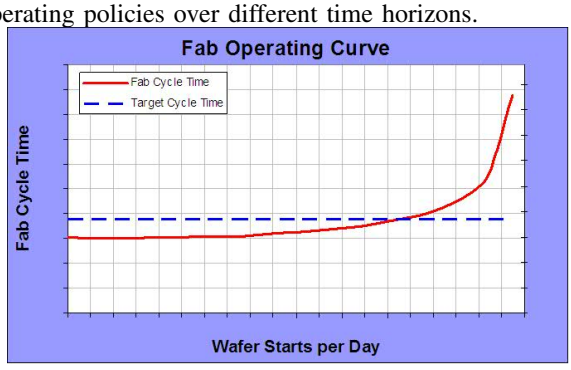

Figure 3: Fab operating curve by EPOS

5 CONCLUSION

The queueing models of EPOS are capable to model the key characteristics of real-world SMS. These models have improved the accuracy of queueing analysis in a semiconductor environment. With that queueing models become applicable in this state space. EPOS is in invaluable tool in IBM's 300mm fab to model its capability and quickly analyze trade-offs between cycle time and capacity. Its success enlarges the applicability of queueing theory for performance analysis to a more universal class of manufacturing lines.

REFERENCES

Bermon, S., and S. Hood. 1999. Capacity optimization planning system (CAPS). *Interfaces* 29 (5): 31–50.

Bitran, G., and D. Tirupati. 1989. Approximations for product departures from a single-server station with batch processing in multi-product queues. *Management Science* 35:851–878.

Chaudhry, M., and J. Templeton. 1983. *A first course in bulk queues*. New York: John Wiley & Sons.

Chiamsiri, S., and M. Leonard. 1981. A diffiusion approximation for bulk queues. *Management Science* 27:1188–1199.

Conners, D., G. Feigin, and D. Yao. 1994. Scheduling semiconductor lines using a fluid model. *IEEE Transactions on Robotics and Automation* 10 (2): 88–98.

Conners, D., G. Feigin, and D. Yao. 1996. A queueing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing* 9 (3): 412–427.

Gaver, P. 1962. A waiting line with interrupted service, including priorities. *J.Roy.Stat.Soc.Ser.B* 24:73–90.

Gelenbe, E., and G. Pujolle. 1987. *Introduction to queueing networks*. New York: John Wiley & Sons.

Hanschke, T. 2006. Approximations for the mean queue length for the $GI^X/G(b,b)/c$ queue. *Operations Research Letters* 34:205–213.

Hanschke, T., and H. Zisgen. 2005. Queueing networks with batch service. IBM Internal Technical Report.

Kramer, M., and I. Meents. 2001. *Integrated simulation - optimization strategies and logistical process control for production planning based on collaboratively maintained queueing models*. Ph.D. thesis, Clausthal University of Technology, Clausthal-Zellerfeld, Germany.

Little, J. 1961. A proof for the queueing formulae $L = \lambda W$. *Operations Research* 9:383–387.

Meents, I., and H. Zisgen. 2004. Simulation of production processes by means of continuous fluid models. IBM Internal Technical Report.

Pujolle, G., and W. Ai. 1986. A solution for multiserver and multiclass open queueing networks. *INFOR* 24:221–230.

Shanthikumar, J., S. Ding, and M. Zhang. 2007. Queueing theory for semiconductor manufacturing systems: A survey and open problems. *IEEE Transactions on Automation Science and Engineering* 4 (4): 513–532.

Zisgen, H. 1999. *Warteschlangennetzwerke mit Gruppenbedienung*. Ph.D. thesis, Clausthal University of Technology, Clausthal-Zellerfeld, Germany.

AUTHOR BIOGRAPHIES

HORST ZISGEN is the team leader of the SCM Industrial Software development team of IBM's Systems and Technology Group in Germany, Mainz. He is a member of the German Operations Research Society (GOR). Additionally he is an adjunct associate professor at Clausthal University of Technology. He received a diploma degree in mathematics and a Ph.D. degree in mathematics from Clausthal University of Technology, Germany. His research focus is mathematical modeling of material flows, especially in semiconductor, and his email address for these proceedings is <hozisgen@de.ibm.com>.

INGO MEENTS works as an accredited IT architect in research and development for IBM Germany. Starting his work in the field of Operations Research with simulation, modeling, and optimization he is currently working on the software design and development for emerging architectures. His current focus is on HPC algorithms for the heterogeneous multi-core processor 'Cell Broadband Engine'. He received a diploma degree in computer science and a Ph.D. degree in computer science from Clausthal University of Technology, Germany. He is a member of the GOR (Gesellschaft für Operations Research, Germany), and his email address for these proceedings is <meents@de.ibm.com>.

BENJAMIN R. WHEELER is an Industrial Engineer at IBM's 300mm wafer fabrication facility in East Fishkill, NY. He holds a B.S. from the Rochester Institute of Technology in Industrial Engineering. His professional interests lie in the application of simulation and network queueing models for fab decision support. He is also active in continuous improvement and lean manufacturing initiatives in IBM Microelectronics. His email address for these proceedings is <brw@us.ibm.com>.

THOMAS HANSCHKE holds the Chair for Stochastic Models in Engineering Science at the Institute of Mathematics of Clausthal University of Technology, Germany. He received an M.S. degree in mathematics and a Ph.D. degree in mathematics from the Karlsruhe University, Germany. Actually he is vice president of Clausthal University of Technology, Germany. Before joining the University Thomas Hanschke was a Senior Technical Staff Member at IBM. His research focuses on performance analysis of complex manufacturing facilities and air traffic systems. His web address is <www.math.tu-clausthal.de> and his email address for these proceedings is <hanschke@math.tu-clausthal.de>.