2014 IEEE International Conference on
Automation Science and Engineering (CASE)
Taipei, Taiwan, August 18-22, 2014

Target Setting with Consideration of Target-induced Operation Variability for Performance Improvement of Semiconductor Fabrication

Yu-Ting Kao, Shi-Chung Chang, and Chun-Ming Chang, *Member, IEEE*

***Abstract*— Production target setting is common in practice to guide operations such as machine allocation and lot dispatching to achieve master production schedule (MPS). As targets affect operations and hence wafer flows, wafer flow estimation under given daily production targets is a basis of adjusting targets and machine allocation. This paper presents an innovative design of target setting algorithm (TaSIV) that develops the target-tracking service model by characterizing target-induced mean and variability and designs a hybrid flow time approximation by exploiting transient tendon queue analysis between two stages to set targets for improving production performance. The design first adopts a Bernoulli trial with proportional-to-target probability to model the target-tracking machine allocation and FIFO dispatching and then characterize the target-induced variability (TIV). To capture the effect of TIV on wafer flows, the design then approximates the time for the last wafer in initial WIP of a stage to finish processing at the next stage, named two-stage penetration time approximation, APT-2, by using Markov chain analysis of tandem queues with given initial number of wafers. By Integrating APT-2 into a recursive algorithm, SOPEA, the design estimates penetration time of multiple stages and wafer flows in a fab. Finally, our design integrates APT-2/SOPEA into a fixed-point iteration between wafer flow estimation and capacity allocation for target setting with consideration of TIV. Over a mini-Fab example and given targets generated by TaSIV, simulation of proportional-to-target machine allocation and FIFO dispatching demonstrates that targets generated by TaSIV reduces over-optimism and close to actual moves by 30.7% of bottleneck machine groups as compared to a mean-value based scheme frequently adopted by practitioners of fab operation management. TaSIV also leads to reductions of 1.2% in mean cycle time and 15.4% in cycle time variance at 1.1% throughput increase.**

***Keywords*—Daily wafer flow estimation, daily target setting, machine allocation variability**

I. Introduction

Production target setting is a short-interval production management function that determines for each product type the target number of wafers processed (moves) at each stage during one day. Given the delivery requirements from customers, wafers are released to the production line, which are processed in a serial flow by different tools. In a fab, the fabrication process of each type of wafers may require more than three hundred fabrication steps. For managerial purpose, every few consecutive steps are represented by one key step, which is defined as a stage. The whole process involves tens of delicate and expensive equipment groups. The process flow of a type of wafers through the fab is highly reentrant because the wafers make multiple visits to equipment groups as successive circuit layers are added onto the wafer. This reentrant feature poses a unique challenge to production, since wafers of different part types as well as wafers of the same type but at different layers of fabrication may compete for the finite capacity of an equipment group. The production target setting is equivalent to properly allocate the tool capacity to different products and stages. In practice, the targets are one of the key review items in the daily production meeting of a fab, which provides direction and serves as a guideline for machine scheduling and lot dispatching [2].

In the literature, there has been much research on the target setting problem [1-3]. Intel has adopted linear programming based optimization engine to convert fab requirements into targets for the function areas. Integrated with scheduling and dispatching for its 300-mm high-volume on the lithography area, the target setting reportedly achieves a significant improvement in output and cycle time [2]. Chang *et al.* [3] designed a daily target setting system (TSS), which allocates the capacity to products as well as stages under master production schedule in a horizon of one day. As the targets are the inputs for scheduling and dispatching, how the target is set affects the overall productivity, making the target problem important. Field applications demonstrated that quality DTS leads to over 20% increase in daily moves and more than 8% decrease of wafers-in-process (WIPs) of a foundry fab case [3] as shown in Figure 1.

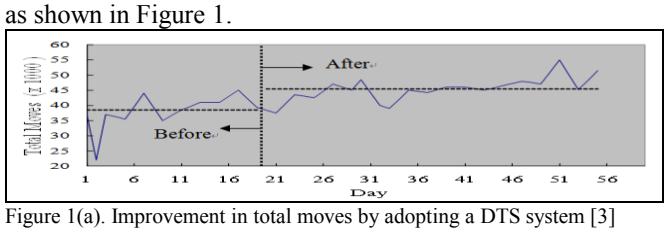

Figure 1(a). Improvement in total moves by adopting a DTS system [3]

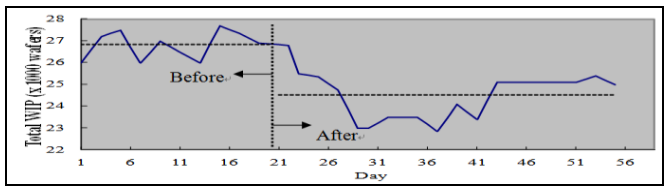

Figure 1(b). Reduction in total WIPs over time by adopting a DTS system [3]

Despite the successful results, all the above target setting schemes are mean-value based approaches without considering variability of manufacturing. There are two types of variability in production: (1) operation variability and (2) manufacturing variability. Manufacturing variability comes

* This work was supported in part by the National Science Council, Taiwan, ROC, under grants NSC 100-2218-E-002-027-MY3, NSC 102-2221-E-002-206 and NSC 102-2219-E-002 -012.

Y.-T. Kao is with the Graduate Institute of Industrial Engineering, National Taiwan University, Taipei, 10617 Taiwan, ROC (e-mail: f97546010@ntu.edu.tw).

S.-C. Chang[1] and C.-M.Chang[2] are with the Department of Electrical Engineering, National Taiwan University, Taipei, 10617 Taiwan, ROC (1. corresponding author; e-mail: scchangee@ntu.edu.tw; 2, e-mail: r02921003@ntu.edu.tw).

from equipment and fab facility and this paper is focused on operation variability injected by man-made operation decisions such as target setting, machine allocation and dispatching. Many studies have demonstrated that operation variability reduction leads to significant waste reduction and fab performance improvements. Li et al. [7] reported that, by reducing operation variability through proper scheduling, a fab accomplished 50% cycle time reduction and 35% throughput increase.

For the purpose to improve the production performance such as cycle time reduction and throughput increase this paper proposes an innovative design of target setting algorithm (TaSIV). The algorithm includes variability of machine allocation induced by tracking targets over time into wafer flow estimation to set targets. The remainder of this paper is organized as follows. Section II addresses the need for considering variability in target setting and its challenges. In Sections III, modeling of service variability caused by target-tracking machine allocation is developed. Section IV proposes an algorithm incorporating the service variability into production target setting. Simulation study over a mini-Fab case is performed in Section V. Finally, Section VI concludes the paper.

II. Need for Consideration of Variability in Target

This section describes the existing mean value-based target setting approach, TSS, developed by [3] and the needs to include variability into target setting. In spite of successful implementation of [3], in practice, it is observed that the targets obtained from TSS are much higher than actual moves, thus requiring manual adjustments. This is because the estimated flow times are much shorter than actual ones. Our in depth analyses of and comparison with fab data show that the actual machine capacity allocation over time differs significantly from TSS, which assumes all the machine groups are independent to each other. Lacking of the consideration of interactions between different machine groups such as machine allocations and dispatching leads to over-optimistic target setting. Machine allocation variability is therefore what we want to incorporate to make more reasonable targets than mean value-based approach.

A. Mean value-based target setting (TSS)

Consider a fab, which fabricates various types of circuitry on semiconductor wafers. Given the demanded moves, capacity estimates and available WIP distribution as inputs, TSS adopts an iterative, proportional resource allocation approach so that up-to-date cumulative production demands are first considered for equipment capacity allocation and then the maximization of machine utilization. The capacity of an equipment group is proportionally allocated to the needs from MPS or available WIPs among individual production stages using the same machine group.

In one hand, WIP availability at individual stages affects capacity allocation and target setting to the stages; on the other hand, target setting affects the number of wafers flowing into the stage from its upstream. WIP flow estimation and machine allocation/target setting are difficult chicken-egg problems. The reentrant property further magnifies the effects since not only different product types but the different stages of one product requiring the same machine group compete the limited capacity as well. The target setting problem can therefore be viewed as a fixed-point iteration problem [7, 3], which iterates between capacity allocation and wafer flow estimation until a fixed-point capacity allocation is achieved.

Chang et. al [3] designed a deterministic queueing analysis-based flow estimation method, named Stage of Penetration Estimation Algorithm (SOPEA). The method estimates the penetration time between stage *j* and *j'*, which is the time the initial wafers of stage *j* need to finish processing and flow into any downstream stage *j'*. It captures daily machine capacity availability, initial WIP distribution, and the coupling between target setting/tool capacity allocation and wafer flow estimation. By using SOPEA to set daily targets with WIP flow estimation, there was a further gain of 5% total moves in a foundry fab case [3]. Moreover, the percentage of stages that have less than 10% difference between the scheduled and the actually achieved targets, i.e., the target hit rate, was increased by about 7%.

B. Deficiency due to target-induced machine allocation

Although the improvements by integrating the flow-in estimation function into target setting, manual adjustments have been needed to achieve such good results. Figure 2 demonstrates a representative difference between the estimated flow for target setting and actual achievements in a memory fab case. The horizontal axis shows the stage index, while the vertical axis shows the achievable flow quantity percentage. The estimation, blue line, is up to 40% higher than the actually achieved flows, the red line, for some products at certain stages.

The phenomenon, where estimated flow-in is higher than actual flow-in, is because the estimated penetration times are much shorter than actual ones. TSS assumes all the machine groups are independent to each other. However, as many stages share the same machine capacity, the machines are allocated to different stages dynamically, which generates the service variability of each stage and arrival variability to the downstream stage. In Kingman's equation [4], the cycle time performance can be estimated by means of a "two-moment" approximation, which makes use of only the mean and standard deviation of the inter-arrival and service time distributions. The cycle time and WIP approximation are the function of arrival and service variability, which is affected by machine allocation. Dynamic machine allocation within different stages thus leads to variation in processing rates of individual stages.

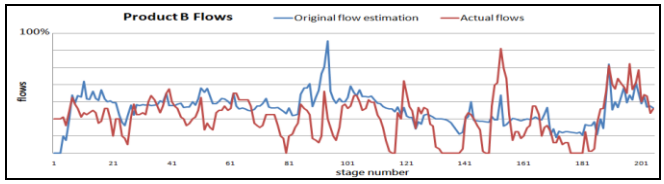

Figure 2. Comparison of SOPEA estimated and actual flows

III. Modeling Service Variability Caused by Target-Tracking Machine Allocation

To incorporate variability into target setting, this section addresses how we model the variability due to machine

allocation and dispatching in one stage. Consider a stage with a given daily target, corresponding machine capacity allocation and initial WIP of the scheduling day, e.g., 7:00am. Our key idea is modeling a supervisor's decision of target-tracking machine allocation and lot dispatching by Bernoulli trial-based proportional-to-target machine allocation and first-in-first-out (FIFO) lot dispatching. Such probabilistic machine allocation directly relates machine service variability of the stage to the target. Analysis of the model then yield analytic descriptions of service time distribution and how much variability is induced to track the given stage target. Details of modeling are described as follows.

A. Bernoulli trial-based model of target tracking allocation

In tracking the production targets of individual stages, at a decision point of machine allocation, say every half-hour or when need emerges, a supervisor first reviews WIP and machine availability and the workload of remaining targets to achieve among individual stages competing for one tool group. The supervisor then allocates machines to stages for the next period of time, the higher the workload the higher the priority. The allocation decision is dynamic, and an important part of allocation logic in the common practice is time division and proportional to targets. Once a machine is allocated to a stage, it processes lots in the WIP of the stage according to fab dispatching rules. Supervisors' dynamic machine allocation and dispatching based on WIP and machine availability to track targets incurs service variability to the production. To capture such variability, the first step is to model the supervisor's dynamic machine allocation to individual stages under given targets. For simplicity of presentation, assume a single type of wafer flow and FIFO as the dispatching rule.

Let us first define some notations and assumptions.

Notations

J : total number of stages;

j : stage index, $j=1,\ldots,J$;

M: total number of machine groups;

m: machine group index, $m=1,\ldots,M$;

S_m: the set of stage that require group m for processing; i.e., $S_m \equiv \{ j \mid$ stage j requires processing by machine group $m\}$, where $S_m \cap S_{m'} = \varnothing$ for $m \neq m'$ and $\cup S_m = \{1, 2, \ldots, J\}$;

C_m: estimated capacity of group m in the unit of machine-time per day;

W_j : amount of wafer lots at stage j in the beginning of the scheduling day (e.g. 7:00 am);

T_j : stage-j target of the day of in terms of lots;

τ_j : mean processing time of each lot at stage j; if stage j deals with batching process, τ_j is equivalent as the time processing a batch divided by number of maximal lots in a batch.

Figure 3 depicts a queueing model of machine allocation and dispatching for target tracking by supervisors, where the probabilistic switching models the allocation decisions at the completion of a processing lot, i.e., the time when a machine in group m finishes processing. The allocation decision is modeled as a Bernoulli trial at each decision instance; there is a probability p_j to allocate the machine to process stage j if it has WIPs available for processing and probability $1\text{-}p_j$ of not allocating. Stages without available WIPs at the time of decision will not be considered for allocation.

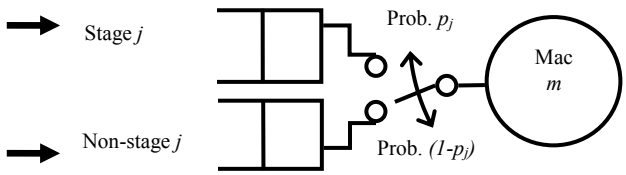

Figure 3. Probabilistic switching model of machine allocation

The probability, p_j, is proportional to T_j of stage j, which is defined by stage target T_j multiplied by mean processing time τ_j and divided by the total capacity C_m:

$$p_{j \in S_m} \equiv \frac{T_j \cdot \tau_j}{C_m}. \quad (1)$$

It models how often an available machine is allocated to processing stage j during a time slot and is proportional to the target and the corresponding workload. A target fulfilled stage will not get proportional-to-target machine allocation until all the stages competing for the same machine group achieve their respective targets.

B. Derivation of service time distribution

Let R_j be defined as the inter-departure time of two successive lots of stage j assuming WIP is always available. As the machine doesn't idle if there is WIP available for processing, R_j can also be interpreted as the service time of a lot at stage j. The target-tracking machine allocation can be modeled as Bernoulli trials with probability p_j. When a machine finishes processing a lot of stage j, the machine will be allocated to stage j again with allocation probability p_j and non-stage j with probability $(1\text{-}p_j)$. Note that τ_j is the average raw processing of stage j and $\tau_{(-j)}$ is the average raw processing time of non-stage j using the same tool group as stage j. R_j is therefore the processing time of stage j plus sum of the number of allocation decisions which processes non-stage j. The probability density function of R_j under a given machine allocation probability p_j is

$$f_{R_j}(r_j) = A\lambda_j e^{-\lambda_j r} - B\lambda_{(-j)} p_j e^{-\lambda_{(-j)} p_j r}, \quad (2)$$

$$\text{where } A = \frac{p_j(\tau_j - \tau_{(-j)})}{p_j\tau_j - \tau_{(-j)}} \text{ and } B = \frac{(1-p_j)\cdot\tau_{(-j)}}{p_j\tau_j - \tau_{(-j)}}.$$

Sketch of Proof (stage index omitted)

(1) Approximate the raw processing times of the stage and the processing time of the stages using the same tool group excepting the stage (non-stage) as exponentially distributed random variables.

(2) Model the number of machines allocated to the stage at a decision point as a geometric random variable with p.

(3) Show that R is the raw processing time of the stage plus sum of non-stage processing times before the allocation again to stage.

(4) Apply moment generating function analysis technique to the random sum of independent variables and obtain Eq. (2)

**Remark: Note that the emphasis here is on modeling target-tracking and proportional machine allocation. The approximation of exponential distributed raw processing time is adopted for characterizing stage service time under given targets, which can be transferred to any other distribution by the procedure of proof.*

The mean service time of a lot at stage j is then

$$E[R_j] = \tau_j + \tau_{(-j)} \frac{(1-p_j)}{p_j}. \quad (3)$$

It can be observed from Eq. (2) that, given the processing time of stage-j, i.e., τ_j, the p.d.f. of R_j depends on p_j, which is proportional to target T_j, and the difference between $\tau_{(-j)}$ and τ_j. By plotting the p.d.f. in Eq, (2), we observe its shape similar to an exponential distribution. So we adopt an exponential distributed random variable Y_j, with mean value $[\tau_j p_j + \tau_{(-j)}(1-p_j)]/p_j$ to approximate the service time of stage-j lots. The role of the target-tracking machine allocation probability is clear.

IV. Target Setting Algorithm Incorporating Service Variability of Machine Allocation (TASIV)

This section first designs a penetration time and flow-in estimation algorithm based on Bernoulli trial-based model of target tracking machine allocation with the aim to provide more realistic flow-in estimation for target setting. Input items are the allocating probability and service time distribution from previous section, and the output item is the flow-in quantity to each stage in one day. Production targets then could be set more reasonably with the variability considered in flow-in estimation.

The algorithmic design first adopts an innovative two-stage tandem-queue analysis based on the service time distribution described in previous section to estimate the mean time of penetrating two tandem stages (APT-2). The innovation lies in stochastic analysis of how the given initial WIPs flow in a two-stage tandem-queue. The mean two-stage penetration time not simply is a sum of mean processing times of WIPs by individual stages but also involves the service variability and interactions between the two stages. After characterizing the first-order effect of variability on penetration time in the two-stage approximation, we then integrate APT-2 into the recursive approximation of SOPEA [3] for multiple-stage penetration time estimation. The estimated daily flow-ins then follows. Details of the algorithm are described as follows.

A. Approximation of mean penetration time for two-stage (APT-2)

First the interaction of machine allocation between two stages is considered. Define penetration time

$PT_{jj'} \equiv$ the expected time needed for the last lot in initial WIP of stage j to finish processing at stage j'.

The penetration time for two-stage, $PT_{j(j+1)}$ depends on relative WIP distributions and service time distributions between the two stages. Figure 4 illustrates our idea of analysis, where we separate $PT_{j(j+1)}$ into two parts. The first part is PT_{jj} under given W_j and τ_j, and the second part is the expected cycle time of the last wafer lot of W_j at stage $j+1$. The second part is equivalent to the expected queue length at stage $j+1$ at the arrival of W_j-th lot from stage j multiplies the expected service time of stage $j+1$. When one wafer of stage j finishes processing, the queue length of stage $j+1$ is probabilistic determined by relative service time between stage j and stage $j+1$.

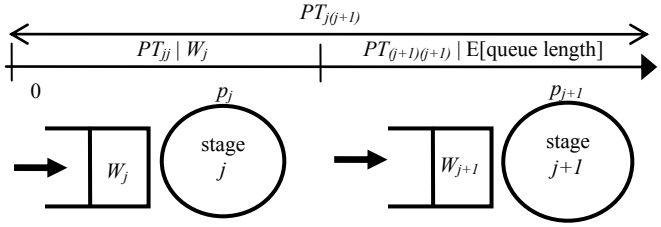

Figure 4. Illustration of two-stage penetration time $P_{j(j+1)}$

To approximate the expected queue length at stage $j+1$ at the arrival of W_j-th wafer lot from stage j, first we denote $(L_{j+1})_{W_j}$ = the number of lots present in stage $j+1$ at time PT_{jj}, which is a random variable and $1 \le (L_{j+1})_{W_j} \le W_j + W_{j+1}$.

For $W_j \ge 1$, $W_{j+1} \ge 1$, and $1 \le (L_{j+1})_{W_j} \le W_j + W_{j+1}$, define the probability that, when the W_j-th wafer finishes processing stage j and arrives the buffer of stage $j+1$, there are i lots in the buffer of stage $j+1$ (in queue or in service) given W_{j+1} lots are individually present at stage $j+1$ at the beginning of day:

$$P_{W_{j+1}}(W_j, i) = P\{(L_{j+1})_{W_j} = i \mid W_{j+1} \text{ already present at time 0}\}. \quad (4)$$

Now we derive how machine allocation variability affects the expected queue length. Let Y_j and Y_{j+1} represent the service time of stage j and service time of stage $j+1$, respectively. The utilization of stage $j+1$, ρ_{j+1}, is $E[Y_j]$ dividing by $E[Y_{j+1}]$.

Let α_{j+1} = Prob. $(Y_j < Y_{j+1}) = \dfrac{\rho_{j+1}}{\rho_{j+1}+1}$ and $\beta_{j+1} = 1 - \alpha_{j+1}$.

The parameter α_{j+1} models the probability that the service time of stage $j+1$ is longer than the service time of stage j *(relative service time)*. The α_{j+1} increases when p_j decreases or p_{j+1} increases, which could be explained intuitively that the probability of longer queue length in stage $j+1$ increases.

As the two-stage penetration time is determined by WIP quantity of stage $j+1$, we innovatively integrate the analysis of [5] here to approximate the queue length of downstream stage $j+1$, which makes extensive use of the memorylessness of the service times of both stages. At the time the last lot of initial W_j finishes processing at stage j, PT_{jj}, the queue length of stage $j+1$ is probabilistic distributed instead of deterministic due to the impact of the machine allocation variability. Here, total machines are aggregated as single machine. M/M/1 queue is therefore be adopted, which begins with the condition that W_{j+1} customers are already present in the system at time zero, and the W_j-th "new" customer arrives at time PT_{jj}, for $W_j \ge 1$.

The algorithm for obtaining the $P_{W_{j+1}}(W_j, i)$ is given below.

Given: α_{j+1}, β_{j+1}, W_j, and W_j

Step 1: **Calculation of the probability of the queue length at stage $j+1$ of i lots when W_j-th lot arrives, $i \in$ [1 to $W_j + W_{j+1}$]**

if $W_j = 0$,

there is no arrival from stage j, therefore the probability of being W_{j+1} is equal to 1 and 0 of others.

$P_{W_{j+1}}(0, i) = 1, \quad \text{if } i = W_{j+1}$

$P_{W_{j+1}}(0,i) = 0, \quad \text{if } i \neq W_{j+1}$

elseif W_j = 1; when the W_j-th lot arrives,

no lot finishes processing of stage $j+1$;

$P_{W_{j+1}}(1,i) = \alpha_{j+1}, \quad \text{if } i = W_{j+1}+1$

all lots of W_{j+1} have finished processing ;

$P_{W_{j+1}}(1,i) = (\beta_{j+1})^{W_{j+1}}, \quad \text{if } i = 1$

i lots are in the queue.

$P_{W_{j+1}}(1,i) = \beta_{j+1} P_{W_{j+1}}(1,i+1)$,

in the order $i = W_{j+1}$, W_{j+1}-1, ..., 2.

else W_j > 1; when the W_j-th lot arrives,

no lot finishes processing of stage $j+1$;

$P_{W_{j+1}}(W_j,i) = (\alpha_{j+1})^{W_j}, \quad \text{if } i = W_{j+1}+W_j$

all lots of W_{j+1} have finished processing;

$P_{W_{j+1}}(W_j,i) = (\beta_{j+1}/\alpha_{j+1}) P_{W_{j+1}}(W_j,2), \quad \text{if } i = 1.$

i lots are in the queue.

$P_{W_{j+1}}(W_j,i) = \alpha_{j+1} P_{W_{j+1}}(W_j-1,i-1) + \beta_{j+1} P_{W_{j+1}}(W_j,i+1)$,

in the order $i = W_{j+1}+W_j$-1, $W_{j+1}+W_j$ -2, ..., 2.

Step 2: **Calculation of the expected queue length of stage $j+1$ when the W_j-th wafer arrives by summing of the number of the wafers waiting in the buffer of stage $j+1$ multiplying the probability of the condition from 1 to $W_j + W_{j+1}$.**

$$E[(L_{j+1})_{W_j}] = \sum_{i=1}^{W_j+W_{j+1}} i \cdot P_{W_{j+1}}(W_j,i) \qquad (5)$$

Step 3: **Calculation of the expected penetration time for two-stage**

$$PT_{j(j+1)} = E[R_j] \cdot W_j + E[R_{j+1}] \cdot E[(L_{j+1})_{W_j}] \qquad (6)$$

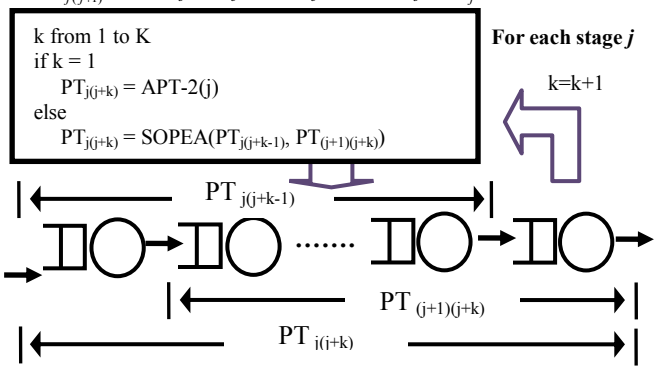

Figure 5. Recursion in SOPEA

B. Fab-wide time approximation (APT-2/SOPEA)

The target-induced variability effect is considered in the two-stage flow time approximation by capturing the increase of queue size of the downstream stage, shown in Figure 5. The variability-included approximation replaces the original deterministic approximation used in SOPEA. SOPEA is a recursive algorithm for computing PT_{jk} by using $PT_{j(k-1)}$ and $PT_{(j+1)k}$. By using the $PT_{j(j+1)}$ and $PT_{(j-1)j}$ calculated from the previous discussion, the three stage flow time estimation, $PT_{(j-1)(j+1)}$ can then be obtained, which implicitly includes the variability effect. And according to this recursive structure, the flow time estimation of multi-stages can be calculated.

V. Performance Improvement via Target Considering Variability

Simulation studies in this section compare and analyze between TaSIV and a mean-value based scheme, TSS. Closeness of targets to actual moves and their differences in throughputs and cycle times are the performance indices. Fab and stage cycle times as well as throughputs from the simulator, which tracks the target setting from TaSIV and TSS respectively, are collected to evaluate the performance.

The simulation studies are over a 3-station (machine group) and six-step re-entrant semiconductor fabrication mini-fab as depicted in Figure 6 [1, 6]. This mini-fab is exemplary but exhibits all of the essential features for fab scheduling. [1] Table 1 lists the parameters where t and j are station and stage indices respectively. P_j is the processing time of stage j, B_j is the batch size of stage j, C_t indicates the number of machines in each station, and R_t is the maximum throughput (capacity) per shift of each machine group. Note that full-batch policy is required in Station 1. Each lot has to wait in the buffer until the number of the lots waiting exceeds B_1. The R_3 is the smallest among all, representing Station 3 as the capacity bottleneck (CA-BN). Although the R_1 is slightly higher, the batching requirements make the lot in the buffer waits until a full batch is formed, resulting large arrival variability to the downstream stage. Station 1 is therefore viewed as a cycle time bottleneck (CT-BN). Station 2 is a non-bottleneck (non-BN). Simulation results on CA-BN and CT-BN will be analyzed respectively.

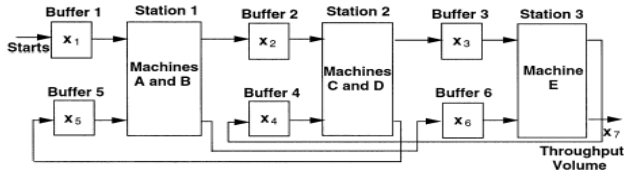

Figure 6. A five-machine six-step mini-Fab [1]

Table 1. Specification of the mini-Fab

t	j	P_j	B_j	C_t	R_t
1	1	225	3	2	6.95
	5	255	3		
2	2	30	1	2	9.97
	4	50	1		
3	3	55	1	1	6.38
	6	10	1		

To compare TaSIV and TSS, the simulator includes the two algorithms for setting targets and implements target-driven machine allocation. The simulation starts from the empty production line. The time horizon simulated is 20 days. The daily output demand from MPS is 18 lots to balance the production line. Of the beginning of each day, targets by stages are set by TaSIV or TSS. Each stage therefore has an allocation probability from Eq. (1), which guides the target-driven dispatching. We collect the moves of stage j in simulating day as actual moves. Closeness, difference between actual moves and the targets, and the cycle time performance are the indices to evaluate the two algorithms.

A. Closeness of targets to actual moves

Define closeness of each stage as

$$C_j = \frac{1}{RD} \sum_{r=1}^{R} \sum_{d=1}^{D} | T_{jdr} - M_{jdr} |, \quad \forall j = 1,...,J \ , \qquad (7)$$

where T_{jdr} and M_{jdr} are the target and actual moves of stage j in day d of r-th replication. Closeness quantifies the daily lot differences between planned targets and actual moves in average. Small C_j indicates that the simulated shop-floor

execution results match the target setting, which corresponds to reasonable targets.

Table 3. Closeness of TSS and TaSIV

	Station 3 (CA-BN)		Station 1 (CT-BN)		Station 2 (Non-BN)	
	S3	S6	S1	S5	S2	S4
TaSIV (lot/day)	0.89	3.43	0.11	1.24	0.16	1.37
TSS (lot/day)	0.92	4.52	0.12	1.79	0.18	1.21
Relative Error	3.3%	24.1%	8.3%	30.7%	11.1%	13.2%

Results of closeness are in Table 3. It can be observed that targets set by TSS deviate from actual moves more than that obtained from TaSIV. The main differences lie on stage 6 and 5, which are the stage of bottleneck station. It is supposed that TSS is over-optimistic as the estimated wafer flow by mean value is higher than actual. Take one replication for demonstration illustrated in Figure 7. The horizontal axis is the simulated day, and the vertical axis indicates the difference between target and actual move in each day. The targets obtained from TSS are plotted as blue line, while that from TaSIV are shown as red line. TaSIV reduces over-optimism and is closer to actual moves achieved than TSS.

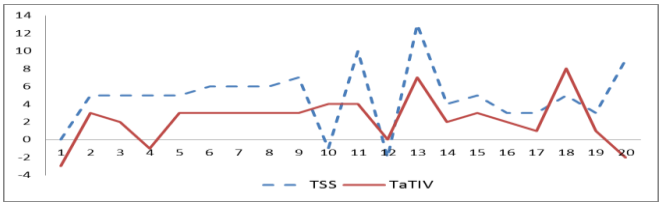

Figure 7. Closeness of TSS and TaSIV of one replication

B. *Performance improvements*

The fab cycle time performance and total throughputs after 20 days are shown in Table 4. As the 90% confidence interval of throughput performance, which tracks the targets from TaSIV, is higher than that from TSS, it could be stated that TaSIV leads to better throughput performance. With the same confidence level, the cycle time performance of TaSIV is also better than TSS. The relative error (R.E.) is defined as difference between the mean value of TaSIV and that of TSS divided by the mean value of TSS. In average, TaSIV leads to 1.1% higher throughput and 1.2% lower average cycle time than TSS. There is also a 15.4% reduction in variance of cycle time by considering TIV.

The stage cycle times are also shown in Table 5 to analyze the improvements between stages. It could be seen that the reductions of cycle time by TaSIV come mainly from the stages of CT-BN while the cycle times of Non-BN stages are almost the same. Note that of the stages of CA-BN, the mean cycle time of TaSIV increases compare with TSS. It is because the higher throughput level by TaSIV under the same production horizon leads to the increase of utilization on CA-BN, which thus prolongs the waiting times before processing. In summary, TaSIV leads to reductions in 1.2% of mean cycle time and 15.4% of variance under capacity allocation than TSS as the flow estimation is more reasonable after including TIV into the flow-in estimation.

VI. CONCLUSIONS

In this paper, an innovative design of target setting algorithm (TaSIV) that includes variability of machine allocation induced by tracking targets over time into wafer flow estimation for better target setting has been developed and tested over a mini-Fab case. A probabilistic proportional-to-target machine allocation model is proposed to capture target-induced variability. Two-stage penetration time approximation (APT-2) captures variability in order to estimate the expected delay needed for daily initial wafers at one stage to finish processing at the downstream stage. Simulation study shows that APT-2 captures both arrival and service variability caused by machine allocation. After the integration of APT-2 into mean-value based scheme, SOPEA, the first order effect of variability captured in two-stage penetration times propagates through the recursive calculation to approximate fab-wide flows. Over a mini-Fab example, it also demonstrates that TaSIV reduces over-optimism and close to actual moves by 30.7% of bottleneck machine groups as compared to a mean-value based scheme. TaSIV also leads to reductions of 1.2% in mean cycle time and 15.4% in cycle time variance at 1.1% throughput increase.

Table 4. Fab performance under given targets

	Throughput (lots) [90% C.I]	Mean Cycle Time (minutes) [90% C.I]	Cycle Time Variance [90% C.I]
TaSIV	348.5 [347.1, 350.0]	1495.6 [1486.2, 1505.1]	61100.4 [54208.0, 67992.8]
TSS	344.8 [343.4, 346.4]	1513.8 [1504.7, 1522.9]	72272.7 [64726.0, 79819.3]
R.E.	1.1%	- 1.2%	- 15.4%

Table 5. Mean value of stage cycle time (minutes)

		TaSIV	TSS
		Mean Stage Cycle Time [90%C.I.]	Mean Stage Cycle Time [90%C.I.]
Station 3 (CA-BN)	S3	187.3 [185.1, 189.4]	183.7 [180.8, 186.7]
	S6	127.3 [125.6, 129.0]	130.2 [127.5, 132.5]
Station 1 (CT-BN)	S1	650.1 [642.2, 658.1]	662.5 [650.2, 674.8]
	S5	418.8 [413.8,423.8]	425.0 [422.1, 427.9]
Station 2 (Non-BN)	S2	54.0 [53.1, 55.0]	54.8 [54.0, 55.7]
	S4	58.1 [57.5, 58.7]	57.8 [57.4, 58.2]

REFERENCES

[1] F.D. Vargas-Villamil, D.E. Rivera, K.G. Kempf, "A hierarchical approach to production control of reentrant semiconductor manufacturing lines", *IEEE Transactions on Control Systems Technology*, Vol. 11, No. 4, pp. 578 – 587, 2003.

[2] N. Govind, E. W. Bullock, L. He, B. Iyer, M. Krishna, and C. S. Lockwood, "Operations Management in Automated Semiconductor Manufacturing With Integrated Targeting, Near Real-Time Scheduling, and Dispatching," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 21, No.3, pp. 363 – 370, 2008.

[3] S. C. Chang, " Demand-Driven, Iterative Capacity Allocation and Cycle Time Estimation for Re-entrant Lines," in *Proceedings of 38th IEEE Conference on Decision and Control,* Phoenix, AZ, Dec., 7-10, 1999, pp.2270~2275.

[4] J. F. C. Kingman, "The Single Server Queue in Heavy Traffic," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 57, No. 04, October 1961, pp. 902 – 904.

[5] W. D. Kelton and A. M. Law, "The Transient Behavior of the M/M/s Queue, with Implications for Steady-State Simulation," *Operations Research*, Vol. 33, No. 2, pp. 378-396, 1985.

[6] M. K. El Adl, A. A. Rodriguez, K. S. Tsakalis, "Hierarchical Modeling and Control of Re-entrant Semiconductor Manufacturing Facilities," in *Proceedings of the 35th Conference on Decision and Control*, Kobe, Japan, Dec., 1996.

[7] R. Burden, J. Faires, Numerical Analysis, 5th ed., PWS-KENT Publishing Co., Boston, 1993.

[8] W. J. Hopp and M. L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, 2nd ed. London, U.K.: Irwin McGraw-Hill, 2001.