Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Implementation of a Fluctuation Smoothing Production Control Policy in IBM's 200mm Wafer Fab

James R. Morrison, *Member, IEEE*, Brian Campbell, Elizabeth Dews, and John LaFreniere

***Abstract*—Efficient operation of IBM Vermont's 200mm semiconductor wafer fabrication facility is essential to achieve the objective of transforming the site into a world class foundry manufacturer. To that end, we develop a fluctuation smoothing for the variation of cycle time (FSVCT) production control policy capable of allowing for a diversity of cycle time commitments. The policy directs which lot should next receive processing when a tool becomes available. In the absence of a validated cycle time model for the fabricator, we obtain estimates of achievable cycle time performance in the presence of business constraints on the cycle times of certain lots via a consequence of Little's law. The mean and variation of cycle time implications for IBM Vermont's facility are highlighted.**

I. Introduction

Semiconductor wafer manufacturing features a process flow which distinguishes it from traditional flow shops and job shops. During production, a lot of wafers (a lot typically consists of up to twenty five wafers) may return to a collection of equivalent tools, termed a tool group, multiple times. As lots may reenter a given tool group, semiconductor wafer manufacturing is said to possess a reentrant process flow. IBM's 200mm semiconductor wafer fabrication facility in Burlington Vermont USA offers a menu of chips from approximately thirty technologies each of which may require four hundred stages of processing. The facility features products with a highly reentrant process flow, operates over two hundred distinct tool groups, and can produce on the order of one thousand wafers per day. A simple example of reentrant structure for a single process flow is depicted in Fig. 1.

Efficient operation of a semiconductor manufacturing plant is critical to achieve profitability, maintain customer satisfaction and to help offset the substantial costs of construction, which have spiraled to $4 billion US for modern 300mm semiconductor wafer manufacturing facilities. One facet of manufacturing operations is the control of the production of work in process (WIP) within the facility. There are two primary decisions regarding WIP to be managed. First, when should a lot be released into the production facility? Second, which lot should next be processed (given that a tool is available)? We do not consider the question of whether an idle tool should remain idle in the presence of available WIP and assume that an idle tool will select a lot from among those in its queue and begin processing (though, for example, idling can be a part of an optimal control policy in batch servers).

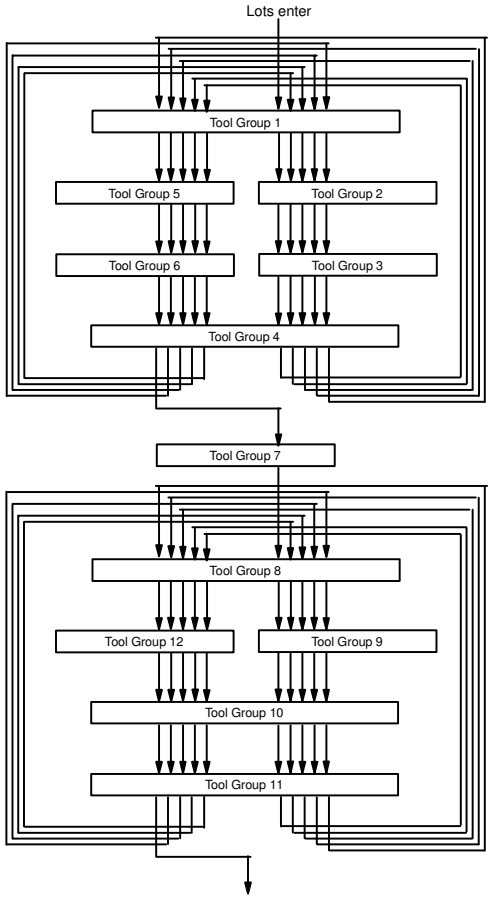

Fig. 1. *A reentrant line with a single process flow.*

An important metric for assessing the operating performance of a manufacturing facility is termed cycle time or turn-around time and is the total time for a lot of wafers to exit the fabricator once it has been released into production (including process time and non-process times such as queue time and travel time between tool groups). For a fabricator such as IBM's 200mm semiconductor wafer fabrication

Manuscript received March 07, 2005. This work was fully supported by the IBM Corporation.

Please address all correspondence to the first author.

James R. Morrison, Ph.D. is with the IBM Corporation, Essex Junction, VT 05452 USA and adjunct to the University of Vermont Department of Electrical and Computer Engineering, Burlington, VT 05405 USA (phone: 802-769-2196; fax: 802-769-9452; e-mail: jrmorr@us.ibm.com).

Brian Campbell is with the IBM Corporation, Essex Junction, VT 05452 USA (phone: 802-769-9600; e-mail: campbelb@us.ibm.com).

Elizabeth Dews is with the IBM Corporation, Essex Junction, VT 05452 USA (phone: 802-769-8515; e-mail: edews@us.ibm.com).

John LaFreniere is with the IBM Corporation, Essex Junction, VT 05452 USA (phone: 802-769-8736; e-mail: jlafreni@us.ibm.com).

0-7803-9568-9/05/$20.00 ©2005 IEEE

facility, cycle times typically vary from forty to sixty days depending on the product under construction and its relative priority to the business and consist of about twenty days of actual process time. The value of cycle time may be quantified in various ways, see [1] and [2] for examples. Through production control the mean and variance of cycle time may be improved. Customer satisfaction and on-time delivery are closely related to the mean and variance of cycle time.

As shown in [3], the release policy, which dictates when a lot should be released into production, can have a significant impact on cycle time performance. We do not address our release policy here.

Our focus is on determining which lot should next receive processing when a tool becomes available. In simulation studies and reports of semiconductor industry results, improved production control decisions have been demonstrated to decrease mean fabricator cycle time by as much as 25% over baseline performance while maintaining an equivalent throughput rate [1], [2], [3], and [4]. Many methods for the control of WIP have been proposed over the years and there has been growing interest recently in the use of controls derived from the study of asymptotic behavior of multi-class queueing networks, see, as a sample, [5] and [6]. We employ the fluctuation smoothing for the variation of cycle time (FSVCT) methodology detailed in [7] and tested in the simulation studies of [1] and [2]. The FSVCT policy can capture many of the features that one would consider as desirable in a fabwide scheduling policy in addition to being relatively simple to implement (in comparison to algorithms which employ mathematical programming techniques to deduce the production control decisions).

The paper is organized as follows. Section II reviews the FSVCT policy and develops extensions deemed necessary to the successful operation of the burgeoning foundry business model at IBM Vermont's 200mm semiconductor wafer fabrication facility. As the FSVCT policy requires estimates of fabricator cycle times, we address the measurement of fabricator cycle times and the prediction of achievable cycle times (in the presence of business constraints) in Section III. Section IV highlights the results obtained from the implementation of the FSVCT methodology at IBM Vermont's 200mm fabricator. Concluding remarks are presented in Section V.

II. Review of the FSVCT Policy and Extensions

The fluctuation smoothing for the variation of cycle time (FSVCT) production control policy assigns to each lot a number termed slack. When a tool becomes available to accept a lot into production, it selects the lot with the least slack (from among those lots that it can process) and begins production. For tool groups such as batch processing tools, which may process more than one lot at a time (e.g., furnaces and chemical cleans), and training tools, for which the processing of lots with similar process requirements in sequence can improve tool group throughput (e.g., ion implant and photolithography), the slack values can help guide batching and training decisions.

A. *The basic fluctuation smoothing for the variation of cycle time (FSVCT) policy*

In a production line with a single product flow, for a lot l the FSVCT slack $s(l)$ is defined as

$$s(l) := -[\text{Now} - \alpha(l)] - \rho(\sigma(l)),$$

where Now is the present time at which the tool is to select a lot, $\alpha(l)$ is the release time (also termed the arrival time) of the lot into the fabricator, $\sigma(l)$ in $\{1, 2, \ldots, P\}$ is the stage of processing at which the lot l resides (e.g., the 115th stage of processing on the P = 247 stage processing route) and $\rho(\sigma)$ is an estimate of the expected remaining cycle time (not only raw process time) from stage σ to the end of the production line. Since the expression $[\text{Now} - \alpha(l)]$ is the time that lot l has been in the fabricator, the FSVCT slack defined is thus the negative of the expected cycle time for lot l if it continues at the expected pace.

With this slack value, the FSVCT policy chooses that lot with the least slack, or equivalently the greatest expected total cycle time. Under this production control policy, lots which are projected to be ahead of the expected total cycle time will slow, allowing lots which are behind to accelerate their progression through the line. Hence, the FSVCT policy selects lots in an attempt to drive all lots in the fabricator to the same cycle time. If one attempts to drive all lots to the same expected total cycle time, the variation of total cycle time should be reduced (hence the name of the policy). A reduction in the variation of cycle time should result in a reduction in the mean cycle time as suggested by the Pollaczek-Khintchin formula, see [8]. Studies in [1], [2], and [7] of the application of the FSVCT policy to models of wafer fabrication facilities have demonstrated that the mean and variance of cycle time are substantially improved over baseline policies such as FIFO, critical ratio (CR) and other common dispatching rules.

B. *Multiple products and multiple cycle time targets*

IBM Vermont's 200mm semiconductor wafer fabrication facility has hundreds of distinct product flows which may, for convenience, be aggregated into around thirty major technology groupings. Each of the technology groupings may have a different total process time and a different anticipated total mean cycle time. Within each technology grouping, there is often a distinction between lots which must proceed at an aggressive pace through the line and those which must perforce accept a higher mean cycle time.

To incorporate the fact that diverse cycle times will be wanted, we extend the basic FSVCT policy. Let there be G lot arrival processes. Assume that all lots from an arrival process have the same expected total cycle time, the same stages of production and the same expected remaining cycle time from each stage of production. Thus, for lots from

arrival process g, the stages of production and the expected remaining time until the lots exit the line from each stage of production are common. For a lot l from arrival process g, let $\sigma(l)$ be the stage of processing at which the lot presently resides and let $\rho_g(\sigma)$ denote the expected remaining cycle time for a lot from arrival process g at stage σ. Let $\rho_g(1)$ denote the total expected cycle time for a lot from arrival process g to complete processing, that is the expected time from entering the first stage of processing to exiting the line.

Several approaches could be employed to generalize the basic FSVCT policy to allow for diverse cycle times. Here we define the multiple cycle time FSVCT slack for a lot l from arrival process g as

$$s(l) := -[\text{Now} - \alpha(l) - \rho_g(\sigma(l))]\frac{CT_{NOR}}{\rho_g(1)},$$

where $CT_{NOR} > 0$ is an arbitrary normalization constant to which all total expected cycle times are scaled, so that lots from different arrival processes may be compared. The multiple cycle time FSVCT policy then recommends that the lot with the least slack receive processing from the next available tool capable of catering to its current stage.

Example 1. *Consider lots l_1 and l_2 from arrival process g_1 with $\alpha(l_1)$ = -10 days and $\alpha(l_2)$ = -21 and lot l_3 from arrival process g_2 with $\alpha(l_3)$ = -31. Suppose that $\rho_{g1}(\sigma(l_1))$ = 40 days, $\rho_{g1}(\sigma(l_2))$ = 30 days and $\rho_{g2}(\sigma(l_3))$ = 50 days. Let $\rho_{g1}(1)$ = 50 days and $\rho_{g2}(1)$ = 80 days. Suppose that Now is day 0 and our normalization cycle time CT_{NOR} = 70 days.*

The slack for each lot may be calculated as

*$s(l_1)$ = [-10 - 0 - 40] * [70 / 50] = - 70.0 days*
*$s(l_2)$ = [-21 - 0 - 30] * [70 / 50] = - 71.4 days*
*$s(l_3)$ = [-31 - 0 - 50] * [70 / 80] = - 70.9 days.*

The multiple cycle time FSVCT policy recommends lot l_2 for processing as it has the greatest scaled expected total cycle time (l_3 is the second choice). This choice will ensure that the total cycle time of l_2 becomes closer to that of l_1 and l_3 after scaling to the normalization cycle time.

C. Adjusting cycle time targets during production

Particularly at the end of the year and the end of each quarter, business conditions may necessitate an adjustment to the expected cycle times for a subset of the lots in production. For reasons such as the need to satisfy quantities of deliveries to some customers or to extract the most profit from the line before the end of a target time, business objectives may dictate that certain groups of lots must increase the rate at which they proceed through production and other groups of lots will by necessity suffer the consequences (resulting in an increase their cycle times).

If the multiple cycle time FSVCT policy is provided with expected cycle times reflective of the new rates at which we want the groups of lots to proceed, and if these cycle times are achievable, then the multiple cycle time FSVCT policy may be used to reprioritize the lots to attempt to attain the desired cycle time changes. Lots released into production following a change in cycle time targets should proceed through the line at near their new cycle time targets (assuming that this is in fact possible). However, for all lots presently in production, the arrival date must be adjusted.

Many approaches could be used to adjust the arrival dates to achieve the objective of reducing the variation of the new cycle times. Two are suggested here. For lots from an arrival process g, let $\rho_g(1)$ be the expected total cycle time prior to adjusting the objectives and $\rho'_g(1)$ be the new expected total cycle time. If one chooses to forgive the lateness (the deviation from the expected cycle time) of a lot l from arrival process g presently in the line, let

$$\alpha'(l) = Now - [\rho'_g(1) - \rho'_g(\sigma(l))].$$

To adjust the arrival dates without forgiveness for lateness, let

$$\alpha'(l) = Now + [\alpha(l) - Now] * [\rho'_g(1) / \rho_g(1)].$$

Implicit in the latter approach is the assumption that for all lots from arrival process g the new estimated remaining cycle times $\rho'_g(\sigma)$ are a constant multiple of the original $\rho_g(\sigma)$, that is $\rho'_g(\sigma) = K_g * \rho_g(\sigma)$, for all σ and each g.

III. Predicting Achievable Cycle Times and Incorporating Business Objectives

In the definition of the basic and multiple cycle time FSVCT policies, expected cycle times play an important part. If one has available a model for the fabricator this can be used to provide estimates for the cycle times to be used with the FSVCT policy. The estimates should be reflective of the cycle times that will be achieved on average when the system is operated under the FSVCT policy. The cycle times can be difficult to predict prior to implementation without use of a model. In the absence of a validated simulation model for IBM Vermont's 200mm wafer fabrication facility, we obtain cycle time predictions via a corollary of Little's law.

A. Aggregate rate of completion of stages of production

For arrival process g, denote the long term average throughput rate as λ_g, the expected number of lots in production as N_g and the expected total cycle time from release to completion of production as W_g. Little's law then states, assuming the averages exist, that $N_g = \lambda_g W_g$ (see, for example, [8]). Suppose that the stages of production for a lot from arrival process g are enumerated as 1, 2, …, P_g where P_g denotes the number of stages of production required in the fabrication of a lot of semiconductor wafers from arrival process g (one can readily generalize the production path to incorporate probabilistic routing). The aggregate rate at which stages of production are completed must equal the

throughput rate times the number of stages of production. This fact is summarized in the subsequent lemma.

Let $D_g(\sigma,T)$ be defined as the number of departures from stage σ (in $\{1, 2, \ldots, P_g\}$) of lots from arrival process g in the time period [0,T) divided by T.

Lemma 1. If the limits exist,

$$\lim_{T\to\infty} \sum_{i=1}^{P_g} D_g(i,T) = \lambda_g P_g .$$

It is convenient to define the rate at which stages of production are completed for lots from arrival process g as

$$\Lambda_g := \lim_{T\to\infty} \sum_{i=1}^{P_g} D_g(i,T).$$

The following corollary is an immediate consequence of Little's law and enables one to relate the rate that stages of production are completed within the fabricator to the expected cycle time for lots.

Corollary 1. If the limits exist,

$$\Lambda_g = P_g \frac{N_g}{W_g}.$$

Example 2. *Consider a manufacturing system in which lots from arrival process g require 300 stages of production (P_g = 300). Assume that the expected number of lots N_g in the fabricator is 500 lots and that the rate at which stages of production are completed Λ_g is equal to 3000 stages/day. Corollary 1 can be employed to deduce that the expected cycle time W_g for a lot from arrival process g is 50 days (N_gP_g/Λ_g). From Lemma 1, the throughput rate λ_g is 10 lots/day (Λ_g/P_g).*

B. Prediction of achievable cycle times

The multiple cycle time FSVCT policy requires reasonable predictions of the expected cycle times of lots. By assuming that the equality of Corollary 1 holds for collections of arrival processes, we estimate expected cycle times from the historical rate of completion of stages of production and the number of lots presently in the system. Our approach allows us to account for complications such as changes in desired cycle time targets, insufficient fabricator capacity (to meet all cycle time targets) and the absence of a validated fabricator model. We rely on Corollary 1 rather than Little's law directly because the rate at which stages of production are completed is subject to less variation than the fabricator throughput and is thus deemed more reliable for use in cycle time estimation.

Assumption 1. Let Λ_g^H denote the historical rate at which stages of production are completed for lots from arrival process g. Let n_g denote the number of lots from process g in the system (at a given time). The expected cycle time for lots from arrival process g is given as $CT_g(> 0)$. We assume that the set of G arrival processes $\{g_1, g_2, \ldots, g_G\}$ can be partitioned into C sets with c denoting an element of the partition such that

$$\Lambda^c := \sum_{g\in c} \Lambda_g^H = \sum_{g\in c} \frac{P_g n_g}{CT_g}.$$

Example 3 demonstrates the ease with which estimates of achievable cycle time targets may be obtained by application of Assumption 1. Naturally, the resulting cycle time estimates must be reasonable to be considered achievable (e.g., normalized cycle time cannot be less than 1).

Example 3. *Let there be two arrival processes g_1 and g_2 which comprise an element c of the partition satisfying Assumption 1. Suppose that the historical rate of completion of stages of production Λ^c for lots from both g_1 and g_2 is 3000 stages/day. Suppose the number of lots are given as n_{g1} = 100 lots and n_{g2} = 400 lots. Let $P_{g1} = P_{g2}$ = 300 stages. Assuming that CT_{g1} = 40 days, lots from g_1 will complete stages of production at a rate of approximately 750 stages/day, leaving an allocation of $\Lambda^c - P_{g1}n_{g1}/CT_{g1}$ = 2250 stages/day for lots from g_2. Thus, we estimate the expected cycle time for lots from g_2 as $CT_{g2} = P_{g2}n_{g2}/2250$ = (300)(400)/2250 = 53.3 days.*

Given cycle time targets CT^T_g for lots from each arrival processes g, we can invoke Assumption 1 to determine achievable cycle time values for use in the multiple cycle time FSVCT policy. In the event that not all cycle time targets can be met (e.g., the fabricator is running at 40 days average cycle time and all lots are targeted at 30 days cycle time), we partition the arrival processes into two classes. Let R denote the set of those arrival processes which are *restricted* to achieve (if possible) their target cycle time. Let U denote the set of those arrival processes which are *unrestricted* in that one is less concerned that they achieve their target (though one hopes to achieve or come near to the target).

Suppose that, for each collection c of arrival processes which is an element of the partition satisfying Assumption 1, Λ^c is sufficient to support the cycle time targets for the lots from the restricted arrival processes, that is

$$\Lambda^c > \sum_{g\in c\cap R} P_g n_g / CT_g^T .$$

We then determine the achievable total cycle time for use in the FSVCT policy as $\rho_g(1) = CT^T_g$ for each g in R (i.e., we use the given total cycle time target). For lots from an unrestricted arrival process (in the partition element c), we determine the scaling factor ε^U_c by which lots can be slowed to ensure that the equality of Assumption 1 holds as

$$\varepsilon_c^U = \frac{\sum_{g \in c \cap U} P_g n_g / CT_g^T}{\Lambda_c - \sum_{g \in c \cap R} P_g n_g / CT_g^T}.$$

We then determine the achievable total cycle time targets for use in the FSVCT policy as $\rho_g(1) = \varepsilon^U_c * CT^T_g$ for unrestricted g in partition element c.

Example 4. *Consider the system of Example 3, with CT^T_{g1} = 20 days and CT^T_{g2} = 50 days. $P_{g1} = P_{g2}$ = 300 stages. Let g_1 be a restricted arrival process and g_2 be unrestricted. Since we seek to provide lots from arrival process g_1 with the accommodations they seek, we expect the 100 lots from arrival process g_1 (n_{g1} = 100 lots) to consume $P_{g1}n_{g1}/CT^T_{g1}$ = (300)(100)/20 = 1500 stages/day. The remaining production available for lots from the unrestricted process is Λ^c – 1500 = 1500 stages/day. Invoking Assumption 1, if we write the achievable total cycle time for lots from the unrestricted process as $\rho_g(1) = \varepsilon^U_c * CT^T_g$, we may solve Λ_c – 1500 = $P_{g2}n_{g2}/(\varepsilon^U_c * CT^T_{g2})$ to obtain ε^U_c = 1.6. We thus relax our expectations for lots from the unrestricted process and use a total cycle time of $\rho_g(1) = \varepsilon^U_c * CT^T_g$ = 80 days in the FSVCT policy.*

Note that when the inequality of $\Lambda^c > \Sigma_{\{g \text{ in } c \text{ and } R\}} P_g n_g / CT^T_g$ is very near equality, the resulting ε^U_c may be quite large and one may not wish to target the restricted objectives as closely. The introduction of an ε^R_c, with $\rho_g(1) = \varepsilon^R_c * CT^T_g$ for restricted g, can serve to reduce the rate at which stages of production are completed for lots from restricted arrival processes.

Of course the expected *remaining* cycle times from each stage to the end of the line used in the settings for the multiple cycle time FSVCT policy must also be achievable and the initial targets may be scaled by the same proportion as the total expected cycle time targets. One could also consider the application of our cycle time estimation approach at each tool group.

The FSVCT policy should improve the overall fabricator cycle time once implemented (assuming a less well performing policy prior to FSVCT) and hence the target cycle times should be later updated to account for the improvements. However, as imposing business constraints on groups of lots hinders the ability of the FSVCT policy to control all lots to the same cycle time (where appropriate), one expects less overall fabricator cycle time improvement than had one adaptively learned the FSVCT policy as in [1] and [2].

IV. Implementation Results

A multiple cycle time FSVCT policy was successfully implemented April 2005 in IBM's 200mm semiconductor wafer fabricator and guides the order in which lots receive production resources. A second release of the software, incorporating additional control features and flexibility, was installed September 2005. The ability to reduce the variation of cycle times and simultaneously exert careful control over the rates at which lots from different arrival processes flow through the fabricator was considered to be a significant increase in capability beyond control methodologies previously implemented within the fabricator.

Before and after the implementation, a steady decrease in the number of lots in the fabricator coupled with a reduction in cycle times was experienced as a consequence of reduced releases into the facility. Thus, in the absence of a validated fabricator model, reduced loading, shifts in mix and volume, changes in staffing, varying tool availabilities (a function of loading and spending) and other efficiency activities obscure the clarity of our results.

The variation of cycle time behavior, however, appears correlated with the implementation of the FSVCT policy (though we must still extrapolate from system behavior before implementation to reach this conclusion). Figure 2 provides variation of cycle time data as a function of the average cycle time for lots exiting the fabricator. In the months following implementation, a substantial drop in cycle time variation was observed. The extent to which the reduction in variation is due to reduced fabricator loading is uncertain (and without a model it is difficult to remove the influence of loading).

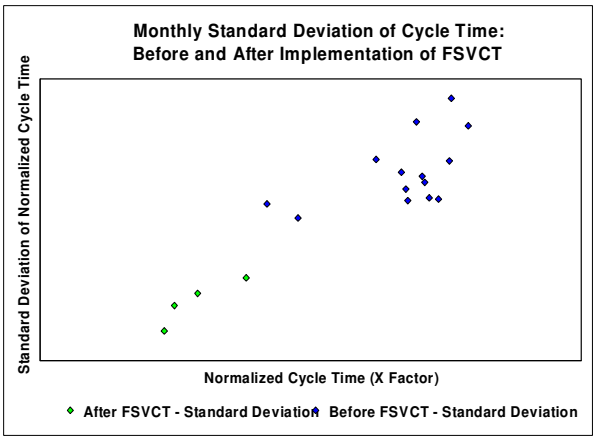

Fig. 2. *A reduction in the standard deviation of cycle time is experienced following implementation of the multiple cycle time FSVCT policy and a reduction in loading.*

Figure 3 provides the throughput rate within the fabricator (rate at which stages of production are completed for the entire fabricator) as a function of the total number of wafers in the fabricator. There does not appear to be a difference in the general trend of the performance. This is not surprising as at the time of the implementation of the multiple cycle time FSVCT production control policy the proportion of fabricator cycle time due to queueing for service was around 30% (so that there was little opportunity to reduce queueing cycle time or roughly equivalently, to improve the throughput at constant WIP).

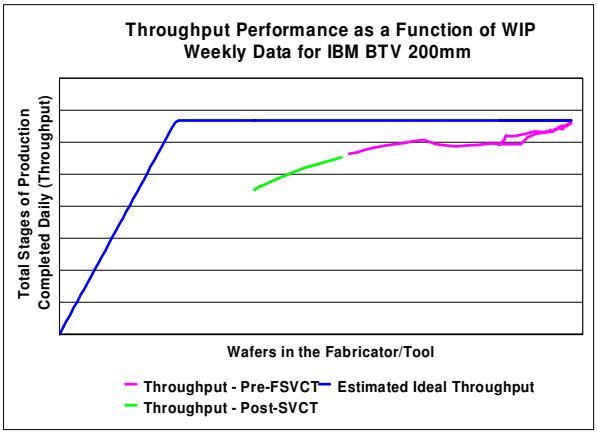

Fig. 3. *Average throughput does not significantly appear to improve beyond previous performance (in a low/moderate WIP state).*

V. Concluding Remarks

An important element of efficient semiconductor wafer fabrication facility operation is the implementation of a production control policy. Focusing our attention on the question of which lot should next be processed from among those currently in the WIP, we develop extensions of the fluctuation smoothing for the variation of cycle time (FSVCT) policy of [7].

To enable the policy to incorporate important business constraints we ensure the ability of the FSVCT policy to drive different subsets of lots to different cycle time targets while maintaining the variation reduction properties within each subset of WIP. We thereby expect an overall cycle time improvement, as variation of cycle time will have been reduced for each subset of WIP (and by the Pollaczek-Khintchin formula, see [8], we also expect a mean cycle time reduction).

With the objective of deducing achievable cycle time inputs for the FSVCT policy and in the absence of a validated wafer fabricator model capable of predicting cycle time behavior, we develop an approximate cycle time estimation procedure based in Little's law. The result is that we are able to obtain estimates for how well we can track the business cycle time targets and assess the consequences of favoritism.

Our software implementation of the multiple cycle time FSVCT control policy was successfully installed in IBM's 200mm semiconductor wafer fabricator in April 2005 and has been guiding production decisions since installation. In addition to the variation reduction properties, the ability to control the rate at which lots from different arrival processes flow through the facility was considered a substantial increase in functionality. The impact of the multiple cycle time FSVCT policy on the IBM Vermont 200mm semiconductor wafer fabricator was studied. The variation of cycle time decreased dramatically following implementation. However, a concurrent reduction in fabricator loading (reduced WIP and release levels) obscures the extent to which our control is responsible for the change. The throughput behavior as a function of WIP level did not appear to be changed by our control in the low/moderate loading regime in which the fabricator was operating.

Directions for future work include the consideration of state dependent settings for the expected remaining cycle times which incorporate tool states (up or down) and WIP levels. Alternative approaches to implementing the multiple cycle time feature of the policy could be investigated.

References

[1] J. R. Morrison, M. Janakiram and P. R. Kumar, "A comparative study of scheduling policies at Motorola," Proceedings of the International Conference on Semiconductor Manufacturing Operational Modeling and Simulation (SMOMS), San Francisco, CA, pp. 51-56, January 1999.

[2] P. Gupta, P. R. Kumar, D. Anderson, T. Ivanova and E. Reitman, "On cycle-time performance improvement under varying mixes by choice of scheduling policies," Proceedings of the International Confereence on Semiconductor Manufacturing Operational Modeling and Simulation (SMOMS), San Francisco, CA, pp. 9-12, January 1999.

[3] L. M. Wein, "Scheduling semiconductor wafer fabrication," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 1, No. 3, pp. 115-130, 1988.

[4] R. M. Dabbas and J. W. Fowler, "A new scheduling approach using combined dispatching criteria in wafer fabs," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 16, No. 3, pp. 501-510, 2003.

[5] M. Chen, R. Dubrawski and S. P. Meyn, "Management of demand-driven production systems," *IEEE Transactions on Automatic Control*, Vol. 49, No. 5, pp. 686-698, May 2004.

[6] J. M. Harrison and A. Zeevi, "Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime," *Operations Research*, Vol. 52, pp. 243-257, 2004.

[7] S. C. H. Lu, D. Ramaswamy and P. R. Kumar, "Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 7, No. 3, pp. 374-388, August 1994.

[8] L. Kleinrock, Queueing Theory, Volume 1: Theory, John Wiley - Interscience, New York, N.Y., 1975.