*Proceedings of the 2011 Winter Simulation Conference*
*S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds.*

# EFFECTIVE WIP DEPENDENT LOT RELEASE POLICIES : A DISCRETE EVENT SIMULATION APPROACH

Raha Akhavan-Tabatabaei
Carlos F. Ruiz Salazar

Universidad de los Andes
Cra 1 A No. 18ª-10
Bogotá, Colombia

## ABSTRACT

In this paper we explore a lot release policy for wafer fabs that is based on the WIP threshold of the bottleneck station. Our results show that this policy is effective in cycle time improvement while keeping the same level of throughput compared with a case where no policy is applied. The application of this policy is practical and needs less considerations compared to policies that aim at keeping the WIP constant throughout the fab.

## 1 INTRODUCTION

Cycle time reduction is a critical issue in semiconductor manufacturing systems (SMS) since it is closely related with the lead time which is crucial to customer satisfaction. In addition, constantly declining prices of semiconductor products dictate selling the newly introduced products to the market as fast as possible in order to enjoy higher profit margins. These factors render the cycle time a critical performance measure in SMS.

In the literature regarding the performance of SMS and specifically the fabrication facilities (fabs) where the chips are manufactured there has been a wide range of studies regarding the factors that contribute to long cycle times. Many of such publications suggest that reducing the variability in arrival process of lots to the toolsets (workstations) and the variability in the service process of each tool can significantly improve the cycle time. For a detailed survey of literature on this topic see (Shanthikumar, Ding and Zang 2007).

To improve the overall fab cycle time some have proposed scheduling and sequencing methods. These methods are usually based on optimization schemes and assuming that the demand and due dates are not subject to substantial changes. The solutions and recommendations of such models are usually used in long-range planning over coarse units of time such as weeks and months (Fowler, Hogg and Mason 2002). This renders them less useful for hour-by-hour plans and during the course of a single shift.

In contrast dynamic flow control models based on heuristic methods aim at satisfying the need of fab managers for dynamic decision making and responding to current conditions. Such models include dispatching rules that select the initiation time of a particular lot into the production line. Policies such as CONWIP (Hopp and Spearman 1991), CONLOAD (Rose 1999), CONWIP with dynamic load changes (Rose 2001) and optimal lot sizing (Wang and Wang 2007) are examples of decision making schemes. CONWIP, CONLOAD and similar policies aim at keeping the WIP level of the whole fab or a particular toolset constant.

In this paper we propose a lot release policy that is based on the variability in the WIP level of the bottleneck toolset in a fab. We examine the effectiveness of this policy on the Mini-Fab model proposed by Intel (2011). The Mini-Fab provides a simple example of a fab with three toolsets and six distinct operations. The process flow of this model creates reentrance in the toolsets which is a common and important characteristic of SMS fabs.

The WIP dependent lot release policy that is under consideration chooses the arrival rate to the first operation of the Mini-Fab according to the WIP level of the bottleneck toolset (WL). This policy chooses a threshold W and adjusts the arrival rate of raw wafers to the Mini-Fab based on the observations on WL. When the WL raises above W the arrival rate is dropped to zero and when it is decreased below W the arrival rate is adjusted to an accelerated rate in order to compensate the starvation period.

We build a discrete event simulation model of the Mini-Fab and test the WIP dependent lot release policy in that framework with various setting for the service time distributions. We show that this policy is effective for cycle time improvement while keeping the throughput constant or even in some cases resulting in higher throughput of the fab.

The rest of this paper is designed as follows. Section 2 describes the details of the Mini-Fab and shows the application of the WIP dependent lot release policy on a number of cases that vary in their service time distributions. Section 3 proposes a procedure based on trial and error to find the best values for the WIP threshold (W) in different cases. Section 4 discusses conclusions and direction for future research while Section 5 is dedicated to bibliography.

## 2 MINIFAB MODEL

Kempf (Intel 2011) presents a five-machine six-step Mini-Fab model that represents the existing complexities in a semiconductor manufacturing fab. This model includes three workstations (toolsets) of diffusion, lithography and implantation with corresponding reentrant flows. Figure 1 shows the layout of the Mini-Fab.

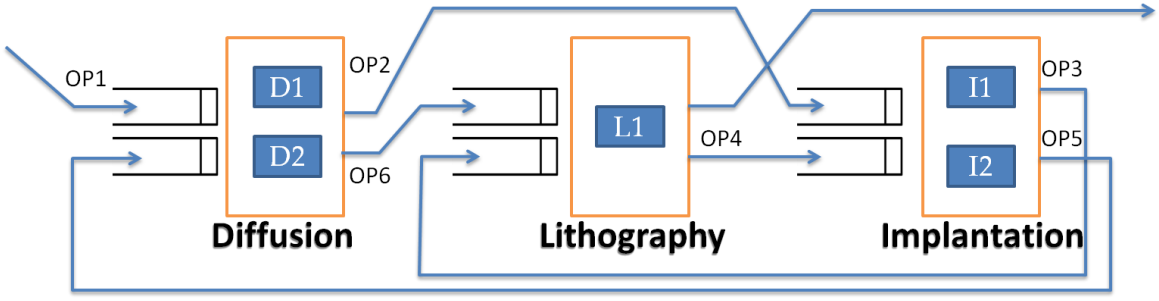


Figure 1: Process flow in Mini-Fab

In this paper we use the basic layout and process flow of the Mini-Fab to experiment the proposed WIP dependent lot release policy. We assume a single product type environment with distinct processing times on each machine to represent the variability between the generations of the same tool within a toolset, a condition that is often observed in real fabs. The individual tools within the toolsets are subject to failure with distinct time to fail and time to repair distributions as well.

The parameters of these distributions are chosen such that the Lithography toolset becomes the bottleneck, a consistent fact with high volume manufacturing fabs. In the following subsections we present a number of scenarios that are different in their distributions of service time in each toolset and examine the application of the proposed policy.

### 2.1 Initial Mini-Fab Model

We first assume that the inter-arrival time distribution to the first operation (OP1) is exponential with rate 2.6 lots per hour and all the service related distributions are also exponential. The parameters of these distributions are chosen such that becomes the bottleneck with 93.4% utilization.

Now we apply the following lot release policy on the initial model. When the WIP level at the lithography toolset (WL) passes the threshold W we drop the arrival rate to zero, otherwise we increase the ar-

rival rate to a recommended value. We find this recommended value such that the long run average arrival rate of the case with policy is equal to 2.6 lots per hour, the same as the initial model with no rule.

For W ranging from 1 to 9 we perform a series of simulation runs with the described characteristics, applying the lot release policy. Each simulation run is replicated 150 times and the replication run length is 300 hours. We compare the results with the case where the policy is not applied.

The mean cycle time and the 95% confidence intervals around the mean are shown in Figure 2. As we observe in this figure for all the values of W ranging from 1 to 9 the case with the lot release policy outperforms the no policy case. In Figure 2, we also include the cycle time behavior of a CONWIP policy, whose WIP value is equal to the average of each Policy Case. This means that we first measure the average WIP of each policy case and round it to the nearest integer to define the CONWIP value. Table 1 shows this in more detail. From this graphical comparison shown in Figure 2, we observe that although CONWIP outperforms our release policy in some of the cases, the differences are not significant at the optimal value of W which is 7. However, in terms of total output CONWIP policy is better in all cases than our policy and the No Policy Case. However, in practice applying a CONWIP policy might be more challenging than our proposed lot release policy. Table 2 shows a summary of results of the comparisons between the No Policy Case, the Policy Cases and the CONWIP case.

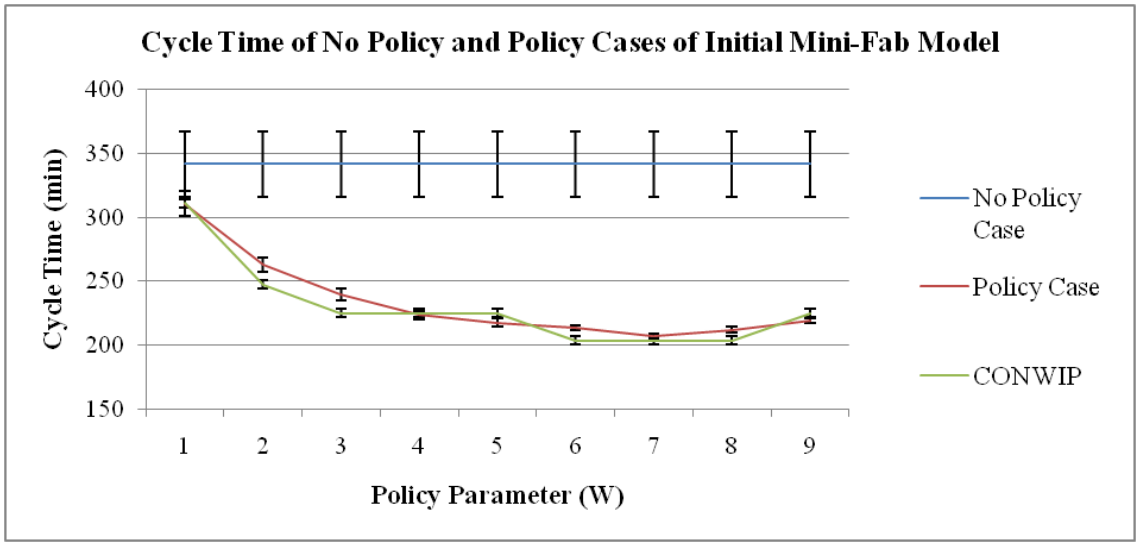


Figure 2: Comparison of Cycle Time in initial Mini-Fab model with and without lot release policy.

Table 1: Definition of CONWIP Values for each Policy Case

| Policy Parameter (W) | WIP Average Level | CONWIP Value |
|---|---|---|
| 1 | 13.62 ± 0.41 | 14 |
| 2 | 11.33 ± 0.24 | 11 |
| 3 | 10.46 ± 0.25 | 10 |
| 4 | 9.82 ± 0.17 | 10 |
| 5 | 9.50 ± 0.12 | 9 |
| 6 | 9.23 ± 0.09 | 9 |
| 7 | 8.99 ± 0.08 | 9 |
| 8 | 9.26 ± 0.09 | 9 |
| 9 | 9.66 ± 0.10 | 10 |

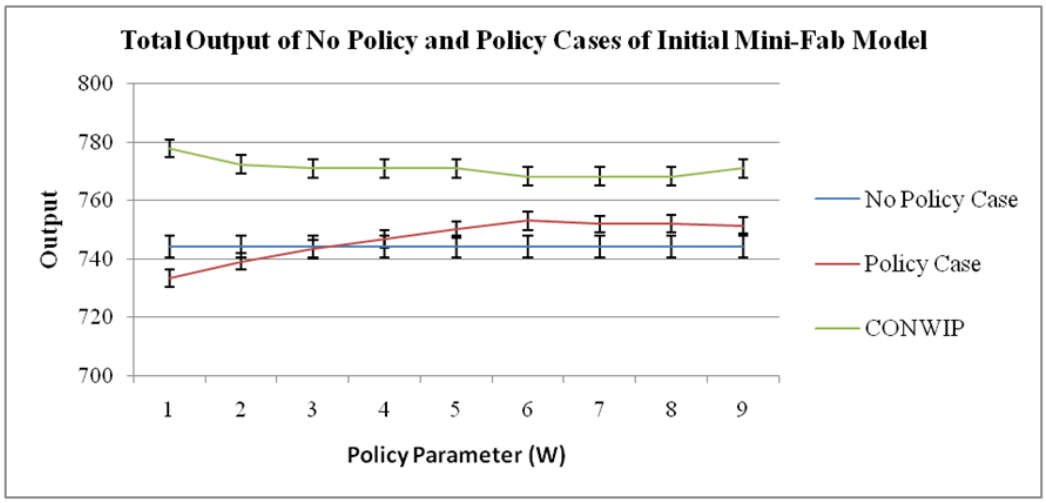


Figure 3:Comparison of Total Output in initial Mini-Fab model with and without relase policy

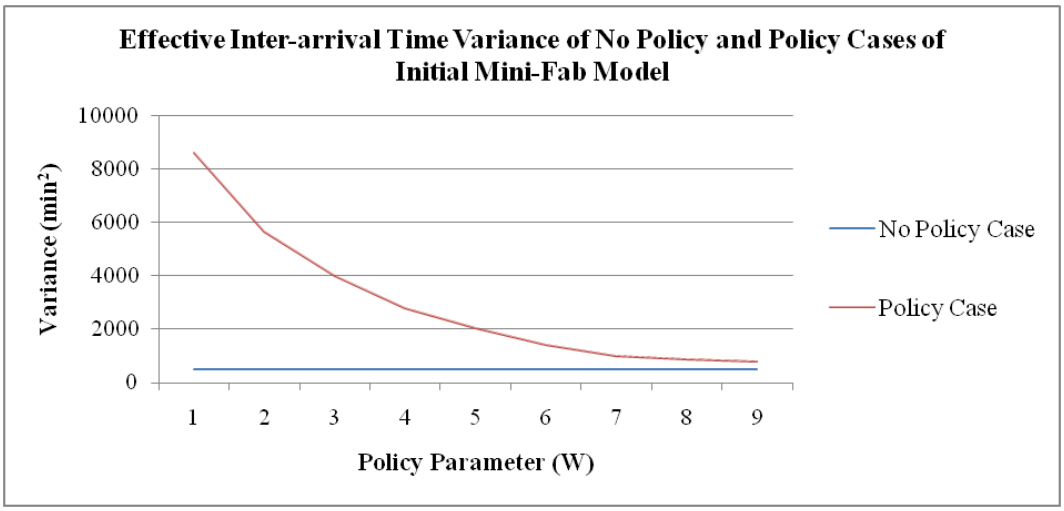


Figure 4: Comparison of Effective Inter-arrival Time Variance in initial Mini-Fab model with and without release policy

Table 2: Percentage difference of performance measures between No Policy and Policy Cases in initial Mini-Fab model

| Policy Parameter (W) | Cycle Time % Improvement | | Total Output % Improvement | | Effective Inter-arrival Variance % Difference |
|---|---|---|---|---|---|
| | W/ Release Policy | W/ CONWIP Policy | W/ Release Policy | W/ CONWIP Policy | |
| 1 | 9.10% | 8.98% | 1.45% | 4.52% | 1563.40% |
| 2 | 23.05% | 27.62% | 0.69% | 3.79% | 993.33% |
| 3 | 29.98% | 34.14% | 0.11% | 3.61% | 671.88% |
| 4 | 34.56% | 34.14% | -0.34% | 3.61% | 439.04% |
| 5 | 36.29% | 34.14% | -0.79% | 3.61% | 290.66% |
| 6 | 37.48% | 40.40% | -1.19% | 3.23% | 176.47% |
| 7 | 39.33% | 40.40% | -1.04% | 3.23% | 96.12% |
| 8 | 37.90% | 40.40% | -1.05% | 3.23% | 70.27% |
| 9 | 35.87% | 34.14% | -0.98% | 3.61% | 52.86% |

## 2.2 Mini-Fab Model with Non-exponential Distributions

In this part we model the Mini-Fab of Figure 1 with non-exponential arrival and service distributions. The distribution of the inter-arrival times for the case with no policy is Gamma(8,3) with the first entry being the scale parameter. For the cases where we apply the policy we use the parameters of the gamma distribution according to Table 3. Similar to Section 2.1 these parameters are chosen based on the selected value of W (policy parameter) in an attempt to keep the long-run average number of arriving lots close to the case with no policy.

Table 3: Policy Case Parameters

| WIP Threshold (W) | Gamma Parameters (Scale, Shape) | | Effective Inter-arrival Mean (minutes) | Effective Inter-arrival Variance (minutes$^2$) |
|---|---|---|---|---|
| 1 | 2.00 | 2.75 | 24.11 ± 0.08 | 4399.45 |
| 2 | 3.00 | 3.00 | 23.90 ± 0.08 | 2500.09 |
| 3 | 4.00 | 3.50 | 23.93 ± 0.08 | 1035.19 |
| 4 | 4.75 | 4.00 | 24.08 ± 0.07 | 430.81 |
| 5 | 6.50 | 3.20 | 23.97 ± 0.07 | 349.40 |
| 6 | 6.75 | 3.25 | 23.93 ± 0.07 | 276.40 |
| 7 | 7.00 | 3.25 | 23.99 ± 0.08 | 237.73 |
| 8 | 7.00 | 3.30 | 23.92 ± 0.07 | 213.43 |
| 9 | 7.00 | 3.35 | 23.98 ± 0.07 | 197.36 |

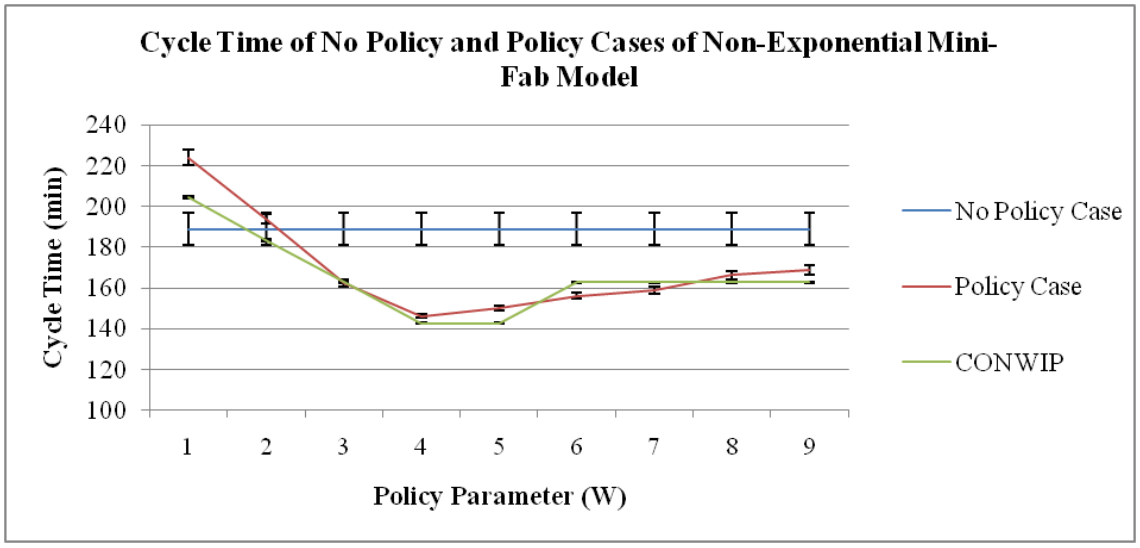


Figure 5: Comparison of Cycle Time in non-exponential Mini-Fab model with and without release policy

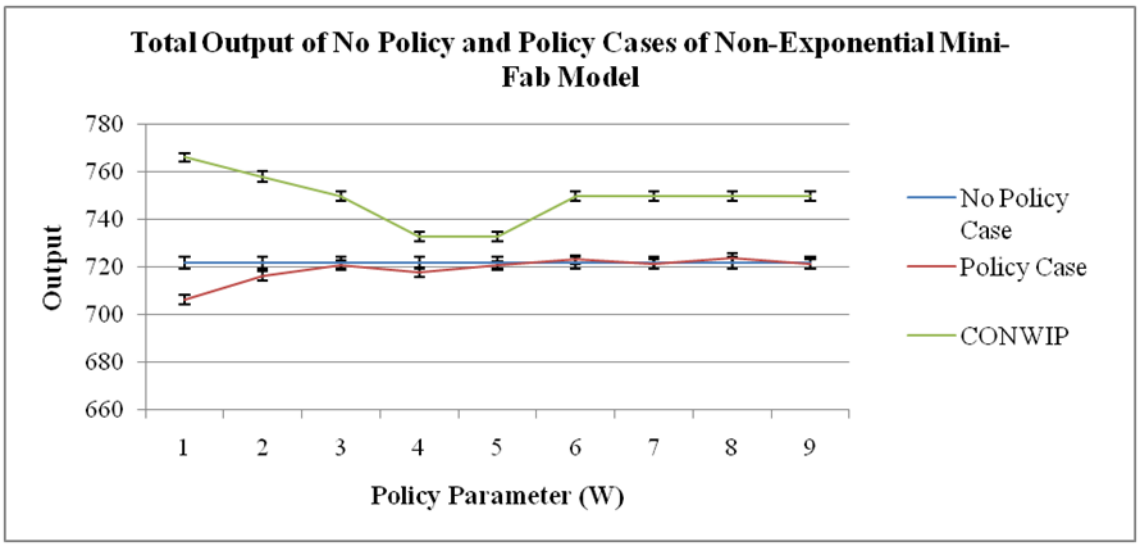


Figure 6: Comparison of Total Output in non-exponential Mini-Fab model with and without release policy

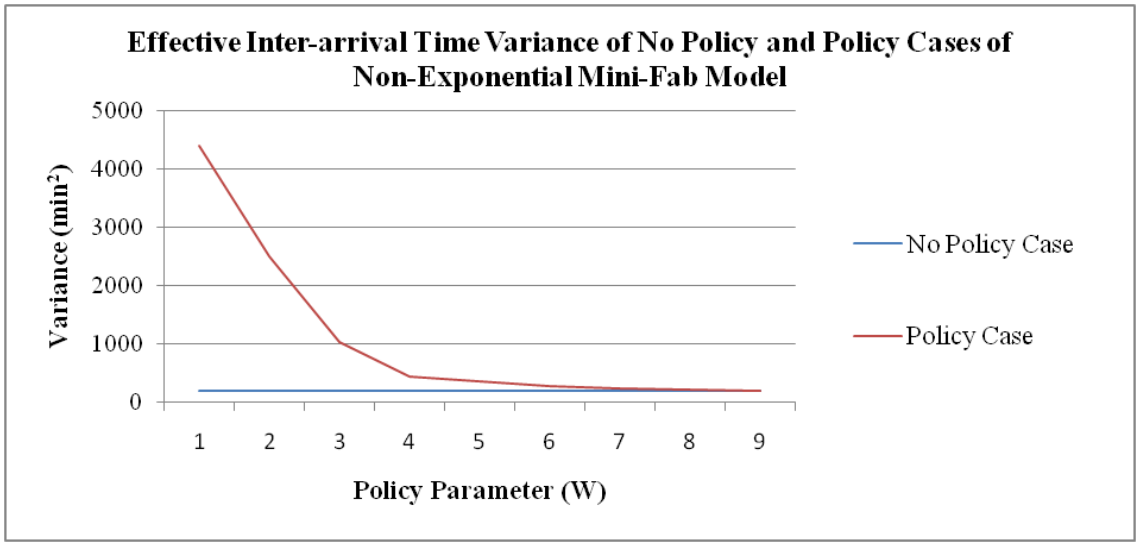


Figure 7: Comparison of Effective Inter-arrival Time Variance in non-exponential Mini-Fab model with and without release policy

Table 4: Percentage difference of performance measures between No Policy and Policy Cases in non-exponential Mini-Fab model

| Policy Parameter (W) | Cycle Time % Improvement | | Total Output % Improvement | | Effective Inter-arrival Variance % Difference |
|---|---|---|---|---|---|
| | W/ Release Policy | W/ CONWIP Policy | W/ Release Policy | W/ CONWIP Policy | |
| 1 | -18.58% | -8.28% | -2.13% | 6.15% | 2208.35% |
| 2 | -2.61% | 2.72% | -0.75% | 5.03% | 1211.78% |
| 3 | 14.04% | 13.83% | -0.12% | 3.91% | 443.15% |
| 4 | 22.57% | 24.41% | -0.54% | 1.53% | 126.04% |
| 5 | 20.40% | 24.41% | -0.13% | 1.53% | 83.33% |
| 6 | 17.36% | 13.83% | 0.20% | 3.91% | 45.02% |
| 7 | 15.85% | 13.83% | -0.04% | 3.91% | 24.74% |
| 8 | 11.91% | 13.83% | 0.31% | 3.91% | 11.98% |
| 9 | 10.71% | 13.83% | -0.03% | 3.91% | 3.55% |

As shown in the results of the exponential and non-exponential Mini-Fab models we observe that our policy achieves similar cycle time reductions as a CONWIP policy. However, differences between both policies are evident when analyzing the fab total output. Through a series of simulation experiments we have discovered that one reason for this result is that the CONWIP policy shows a slightly higher inter-arrival rate of lots than our policy. Consequently, due to Little's Law the CONWIP policy shows a slightly better cycle time and throughput rate when compared against our policy. This means that with our policy we lose some output when changing the original arrival rate to zero each time the WIP at the bottleneck is higher than the policy parameter.

## 3 DETERMINING THE APPROPRIATE WIP LEVEL

As it is observed in Section 2 applying the WIP dependent lot release policy can be an effective tool for cycle time improvement and often for throughput improvement as well. However, the degree of improvement in the cycle time compared with no policy case depends on the policy parameter. In this section we present a procedure to find the appropriate policy parameter W, that yields the minimum cycle time.

As observed in Section 2 this policy can affect the cycle time and throughput performance at the same time and for lower levels of WIP it can be damaging to both performance metrics. In the proposed procedure here we attempt to choose W while taking into account both metrics of throughput and cycle time.

The procedure begins with running the simulation model for no policy case with W=0 and recording the resulting cycle time and throughput. Then this step is repeated through incrementing the value of W by one and increasing the value of arrival rate in such a way that the effective arrival rate of cases with policy remain very close to that of the no policy case. The iterations stop when no significant improvement in the cycle time is observed.

Here we present a step by step description of this procedure:

1. Run replications of the simulation model with no policy. Save the effective inter-arrival time mean ($\lambda$) and variance ($\sigma^2$), cycle time (CT) and total throughput (T).
2. Set W = 1, CTI *(cycle time improvement)* = 0, BCT *(best cycle time)* = 0, BW *(best WIP)* = 0.
3. While CTI $\geq$ 0 Do:
   Without changing the type of inter-arrival time distribution choose its parameters such that the effective mean inter-arrival time of the No Policy and Policy cases become statistically equal.
   Repeat step 1 using the following lot release policy: no job enters the system while WL > W
   - 3.1. Save $\lambda_{eff}$ and $\sigma_{eff}^2$ (effective inter-arrival time mean and variance), $CT_w$ and $T_w$ (cycle time and throughput for the designated level of W).
   - 3.2. If BCT = 0 then
     - 3.2.1. Set BCT = $CT_w$
     - 3.2.2. Set BW = W
     - 3.2.3. Set CTI = 0
   - 3.3. Else
     - 3.3.1. Update CTI = $CT_{W-1} - CT_W$
   - 3.4. End If
   - 3.5. For an arbitrary value $\gamma$, if CTI > 0 and $|T-T_W| / T < \gamma$, then
     - 3.5.1. Set BCT = $CT_w$
     - 3.5.2. Set BW = W
   - 3.6. End If
   - 3.7. Set W = W + 1
4. End While
5. For an arbitrary value of $\varepsilon$ if $|BCT-CT| < \varepsilon$ then BW is the optimal WIP level to apply release policy. However, total throughput values should be compared so that throughput rate is not significantly affected by cycle time reduction. If BCT $\geq$ CT then, the release policy is not effective to reduce cycle time in the system. Note that *Best Cycle Time* (BCT) is only updated when the cycle time is improved with respect to the previous iteration and besides if throughput or total output is not negatively affected more than $\gamma$.

We now describe the algorithm in more detail. Step 1 refers to choosing a base case or No Policy Case which will act as a benchmark for comparing all Policy Cases. Step 2 initializes some variables in order to compare Policy Cases and to keep track of the best solution found so far. Step 3 builds Policy Cases and makes comparisons between them. Step 3.1 intends to choose appropriate parameters for each

value of *W* such that long-run inter-arrival time mean is the same between each Policy Case and the No Policy Case. Since this choice of the parameters is not evident it may be found through trial and error.

In Table 5 we give some insight of how to choose these parameters for each value of W (from 1 to 9) for high utilization levels of Lithography toolset (≥ 85%). Applying this procedure on a number of cases including those presented in Section 2 shows that the difference between the mean and variance of the inter-arrival times in the no policy case and the cases with policy are within a range that is presented in Table 5 for various values of W.

For example, for W = 2 a good guess of the inter-arrival distribution parameters would be one such that the mean of the Policy Case is about 30% of that of the No Policy Case. The general idea is to start with a low mean inter-arrival time (for W = 1) and low variance and as *W* increases increase both. This implies that as W increases the corresponding Policy Case approaches to No Policy Case in its mean and variance of inter-arrival times. This result is consistent with intuition since the probability of having more than W jobs in the Lithography station is lower as W increases.

Table 5: Policy Cases parameters choice insight

| WIP Threshold W | Mean % Factor | Variance % Factor |
|---|---|---|
| 1 | 15% - 25% | 10% - 25% |
| 2 | 25% - 35% | 15% - 30% |
| 3 | 40% - 55% | 30% - 40% |
| 4 | 55% - 75% | 40% - 50% |
| 5 | 60% - 80% | 50% - 65% |
| 6 | 70% - 90% | 60% - 70% |
| 7 | 80% - 95% | 65% - 80% |
| 8 | 90% - 97% | 65% - 90% |
| 9 | 95% - 99% | 70% - 99% |

The rest of lines in step 3 compare the Policy Cases in terms of the average cycle time they render. The algorithm is set to stop as soon as a Policy Case performs worse than the previous one (W-1). This is the result of observations in figures of Section 2 where it is shown that the cycle time of the Policy Cases is a convex function of W. Therefore $CT_{W-1} > CT_W$ indicates that a global minimum is achieved.

## 4 CONCLUDING REMARKS

In this paper we discuss the application of a WIP dependent lot release policy that is based on the WIP threshold of the bottleneck station in a miniature semiconductor manufacturing fab. We first show that applying such a policy can be effective in cycle time improvement of the fab. This policy only considers a static bottleneck which is identified as the most utilized station in the long-run. This means that we do not consider shifting bottlenecks when applying this policy. Later we discuss the procedure to find the effective WIP threshold to improve the cycle time without the loss in throughput level at the same time.

We also compare this policy with a policy where the WIP is kept constant throughout the Mini-Fab. The results show that in some cases the constant WIP policy can be more effective in cycle time improvement. Our results also show that total output performance measure is also improved with the constant WIP policy and this may be due to the slight loss of output when suspending the release process of lots with the proposed policy. However, in practice keeping the WIP constant across the fab may come across as a much harder task than controlling the WIP level of the bottleneck.

The choice of WIP threshold is also done through an iterative procedure and through keeping the overall effective arrival rate constant. This guarantees no loss in throughput and renders the policy cases comparable with the case where no policy is applied. In order to find the effective arrival rates some insight is given based on experiments.

Finally, this model can serve as a basis to give insight to the line managers on how to choose lot release policies. To extend this work we plan to work on a similar model for opportune maintenance and show that adding variability in time to repair if done in sync with the traffic of the bottleneck station can render lower cycle time as well.

## 5 REFERENCES

Fowler, J.W., G.L. Hogg, S.J. Mason. 2002. "Workload Control in the Semiconductor Industry." *Production Planning & Control Vol. 13 No. 7*: 568-578.

Hopp, W.J., M.L. Spearman. 1991. "Throughput of a Constant Work in Process Manufacturing Line Subject to Failures". *International Journal of Production Research* Vol. 29 No. 3: 635-655.

Hopp W.J., M.L. Spearman. *Factory Physics*. Singapore: McGraw Hill, 2008.

Intel. "Intel Five-Machine Six Step Mini-Fab Description: Dr. Karl Kempf". Accessed April 2011. http://aar.faculty.asu.edu/research/intel/papers/fabspec.html.

Rose, O. 1999. "CONLOAD: A New Lot Release Rule for Semiconductor Wafer Fabs." In *Proceedings of the 1999 Winter Simulation Conference*, edited by P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, 850-855. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Rose, O. 2001. "CONWIP-like Lot Release for a Wafer Fabrication Facility with Dynamic Load Changes." *Proceedings of the 2001 International Conference on Semiconductor Operational Modeling and Simulation*: 41-45.

Shanthikumar J.G., S. Ding, M.T. Zang. 2007. "Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems." *IEEE Transactions on Automation Science and Engineering* Vol. 4 No. 4: 512-522.

Wang, C., C. Wang. 2007. "A Simulated Model for Cycle Time Reduction by Acquiring Optimal Lot Size in Semiconductor Manufacturing." *International Journal of Advanced Manufacturing Technology* Vol. 34: 1008-1015.

**Raha Akhavan-Tabatabaei** is an assistant professor at the Department of Industrial Engineering, Los Andes University in Bogota, Colombia. She has received her master's and Ph.D. degrees in operations research and industrial engineering from Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, USA. (e-mail: r.akhavan@uniandes.edu.co).

**Carlos F. Ruiz Salazar** is a former undergraduate student at the Department of Industrial Engineering, Los Andes University in Bogota, Colombia. He is currently a master student in the Operations Research and Management Science Program at Tilburg University, The Netherlands.