Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

OPTIMIZING STARTS FOR CAPACITY, VELOCITY, AND OUTPUT DURING THE RAMP-UP PERIOD OF A SEMICONDUCTOR FAB

Adar A Kalir

Fab/Sort Manufacturing Division
Intel Corporation
2 Ha'Avatz ST
Qiriat-Gat 82109, ISRAEL

Kosta Rozen

Fab/Sort Manufacturing Division
Intel Corporation
2 Ha'Avatz ST
Qiriat-Gat 82109, ISRAEL

ABSTRACT

In this paper, we address the question of how to maximize output over the ramp-up period of a semiconductor fab. We postulate that by keeping a margin between the constraint capacity and the ramp starts cumulative output can be increased over a desired planning horizon (e.g. first year of production). A mathematical framework of the problem is offered and then used to prove our hypothesis. We demonstrate the benefit via a numerical example and a set of simulation experiments.

1 INTRODUCTION

In the semi-conductor industry, factories ramp their production capacity in parallel to tool (machine) installations in order to maximize total output over time; and to put the expensive equipment into immediate use while it begins to depreciate. Traditionally, factory starts are set to the factory's ramping capacity, which is determined by the pace of installations of the most expensive tool (typically a lithography tool). Capacity, cycle time (CT) and fab constraint utilization are inter-related per the well-known operating curve (Fayed and Dunnigan 2007), illustrated in Figure 1 (next page). The operating curve, which emerges from basic queuing theory, suggests that as production volume increases towards the factory capacity (determined by its actual constraints), CT increases non-linearly. This CT behavior gives rise to the following concept: delay starts to create a capacity buffer to push more wafers to the market faster during ramp. This seemingly counter-intuitive principle of reducing starts to increase output earlier during ramp only works if by delaying starts a commensurate lower CT is attained. This problem can be stated as follows: What should the optimal starts levels (and subsequently the utilization of the fab constraint) be during ramp, in order to maximize the cumulative output over a desired planning horizon? The purpose of this paper is to formulate this problem analytically, and offer fundamental insights for practitioners.

2 RELATED LITERATURE

The ramp-up of factories in semiconductor manufacturing is a complex period, characterized by a combination of dynamically changing production capacity, relatively low yield, and high demand. Hence, more output and faster cycle time during ramp are essential - and provide the following important benefits (Terwiesch and Bohn 2001; Haller et al. 2003):

- Faster time-to-market for new products resulting in market leadership;
- Increased profitably (as product prices are highest in the beginning of their life cycle, and competition is scarce during that time);
- Faster learning speed by shortening info-turns, resulting in faster yield improvement; and
- Faster equipment qualification by shortening info-turns.

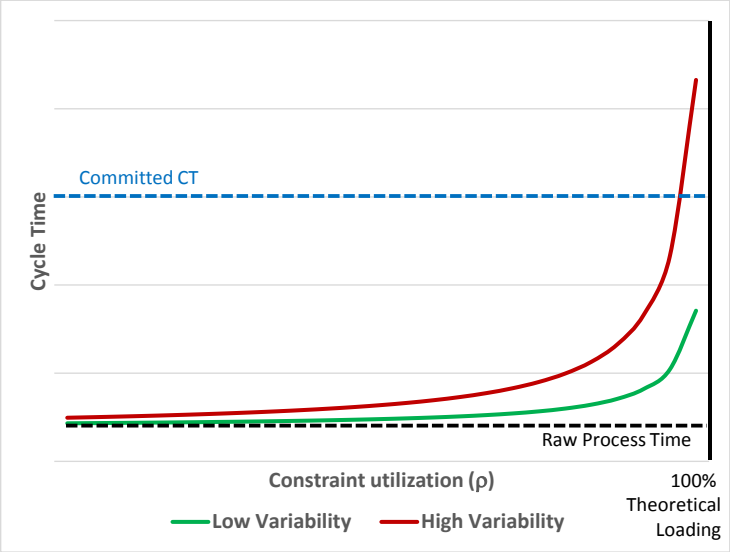

Figure 1: The Operating Curve (based on G/G/m queuing model).

Leachman and Hodges (1996) show that semiconductor manufacturing is characterized by immature processes and immature processing equipment with relatively short lives, resulting in low yields during ramps, and that short cycle time and fast learning speed is critical for long-term success. Terwiesch and Bohn (2001) analyze major trade-off that occurs during ramp between learning speed (which requires experiments that consume production capacity) and production output (which competes for the same capacity). A different trade-off, between increasing production output and improving yield, is discussed in Benfer (1993) and Tirkel et al. (2016). The latter evaluate the trade-off between cost and yield, based on amount of in-line inspection (measurement of WIP wafers). Grewal et al. (1998) describe the development and application of an integrated static capacity and dynamic simulation analysis methodology for defining equipment purchasing plan with the goal of reducing CT during ramp while minimizing capital equipment expenditures by quantifying CT and capital trade-offs for different toolsets. However, no work to date has looked at trade-offs between installed capacity, CT, and output during ramp.

3 FORMULATION OF THE PROBLEM

In a typical ramp, executed in fixed ramp steps (i.e. every τ weeks), starts are increased by a volume of v wafers, until the end of the ramp is reached, at weekly starts of nv. The question we pose is: considering a planning horizon of T, and given the fixed ramp steps, what would be the optimal starts scheme in order to maximize total output? Is it truly the one of nearly matching starts with the installed capacity for a utilization of 99% or maybe this is a case of "less is more"; and by introducing less starts more output can be achieved over the ramp (and post ramp) planning horizon?

The notation used in this paper is summarized in Table 1. Let us denote the expected cycle time of an average wafer during the ramp by CT. Then, as shown in Figure 2, initial output of the fab would begin to accumulate after time CT. Therefore, if the planning horizon for which we maximize the cumulative output is T, the total duration for which output is accumulated is T-CT.

Table 1: Notation and Abbreviations.

n	Number of ramp steps, $k = 1,\ldots,n$.
τ	Number of weeks at each ramp step k.
v	Volume by which each ramp step is increased (fixed)

T	Planning horizon
CT	Mean Cycle Time
ρ	Planned Utilization (or occupancy rate) of the fab's constraint tool-set (station).
A	Planned factor's constraint availability
RPT	Raw Process Time of a single unit.
EPT	Effective Process Time. $EPT = RPT/A$
C_{AR}	Coefficient of variation of the inter-arrival times
C_{EPT}	Coefficient of variation of the effective process (service) time
V	Variability term of the CT equation
U	Utilization term of the CT equation

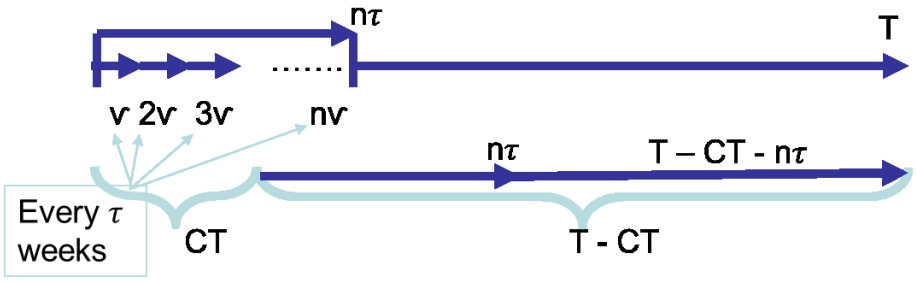

Figure 2: Relationships between ramp starts, *CT*, and *T*.

In order to develop a mathematical model for this problem, note that shifting starts by a certain duration (e.g. by 4 weeks) is just one way to enforce a gap between the starts level and the fab's constraint capacity. A more generic way would be to consider starts that are a portion of the installed capacity at the constraint, which would thenbe equal to the fab's constraint utilization, denoted by ρ. In that case, starts at any point would be $k(\rho v)$ instead of kv, where k is the k[th] step in the ramp. Knowing what the fab constraint would be, we can utilize the famous P-K equation to estimate the *CT* (Hopp and Spearman 2000). The equation is given by (1).

$$CT \cong \left(\frac{C_{AR}^2+C_{EPT}^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)EPT + EPT \quad = V \cdot U \cdot EPT + EPT \tag{1}$$

Equation (1) is actually an approximation for a G/G/1 queuing system. It states that *CT* is approximately equal to the multiplication of three terms: a variability term, a utilization term, and a an effective process time term, respectively. Note that ρ in this expression represents the utilization of the fab constraint tool-set, and it is the same ρ which was introduced earlier to denote the starts entering the fab at each of the ramp steps. Next, we differentiate between two cases:

- Case 1: When T-CT < nτ (ramp period)
- Case 2: When T-CT ≥ nτ (peak, or post ramp, period)

In case 1, for the cumulative output, we have:

$$Output = \sum_{k=1}^{\frac{T-CT}{\tau}} \tau \cdot k \cdot (\rho v) \tag{2}$$

The cumulative output function in Equation (2) is calculated as follows. There is a period of length *CT* in which starts are introduced into the fab but no output is yet attained (and therefore cumulative output is zero during this period). Then, output begins to accumulate but since only a period of *T*-*CT* is considered for this output (which is shorter than the entire ramp period, nτ), the output function sums-up only over the proportion of starts till the cutoff time *T*, which occurs before the end of the ramp. Re-arranging and simplifying (2), we get:

$$Output = \tau \cdot (\rho v) \cdot \sum_{k=1}^{\frac{T-CT}{\tau}} k = \\ = \frac{\tau \cdot (\rho v)}{2} \cdot \left(\frac{T-CT}{\tau}\right)\left(\frac{T-CT}{\tau}+1\right) \cong \frac{\tau \cdot (\rho v)}{2} \cdot \left(\frac{T-CT}{\tau}\right)^2 \quad (3)$$

And when *T*-*CT*= $n\tau$, or alternately: $\frac{T-CT}{\tau} = n$, then the equation becomes:

$$Output = \frac{\tau \cdot (\rho v)}{2} \cdot n^2 \quad (4)$$

In case 2, when *T*-*CT* $\geq n\tau$, we get:

$$Output = \frac{\tau \cdot (\rho v)}{2} \cdot n^2 + (T - CT - n\tau) \cdot (\rho n v) \quad (5)$$

Equation (5) implies that post ramp, fab loading is still below its capacity, which is not a realistic case. More realistically, loading should increase to maximal capacity once the ramp is complete. This is captured in (6) by omitting ρ in the second term:

$$Output = \frac{\tau \cdot (\rho v)}{2} \cdot n^2 + (T - CT - n\tau) \cdot (n v) \quad (6)$$

An important note to make on Equations (2)-(6) is that yield loss is ignored in the output function. It has been implicitly assumed that the impact of yield loss is the same across all *CT* scenarios (i.e. that yield is *CT* neutral). Practically, at least in recent years, *CT* is presumably found to be positively correlated with higher yield (Tirkel et al. 2009).

The basic premise of our proposed model is that when substituting Equation (1) within Equation (6) for the *CT* term, a strictly concave function emerges for the output, which suggests a global unique maxima for a specific value of ρ. The proof of this can be found in Kalir and Rozen (2018).

4 NUMERICAL EXAMPLE

To illustrate the concavity nature of the output function, consider the numerical example depicted in Table 2. In this example, a ramp of 1K wafers every 4 weeks, to a peak of 10K is analyzed. The planning horizon is a full year (52 weeks). The variability coefficient is constant at 0.12. This value is based on actual fab data. As to why the coefficient is held constant through the ramp, the reader is referred to Kim and Uzsoy (2008), Manda et al. (2016) where offsetting effects are discussed, that may keep the coefficient fairly constant during ramp.

Based on this data, Figure 5 plots the output function against ρ. Note that the cumulative output has a global maxima for ρ = 0.80., i.e. when starts into the fab are introduced at a ratio of 80% of the fab constraint capacity. Also, for values of ρ within ±5%, the output sensitivity is relatively low and merely reduces by 2%. However, further deviation from the optima causes a rapid deterioration in output.

Table 2: Parameter values for numerical example.

Parameter	Unit	Value
v	wafers	1000
τ	weeks	4
n		10
T	weeks	52
V (Variability coefficient)		0.12
EPT (approx. ~RPT)	weeks	8

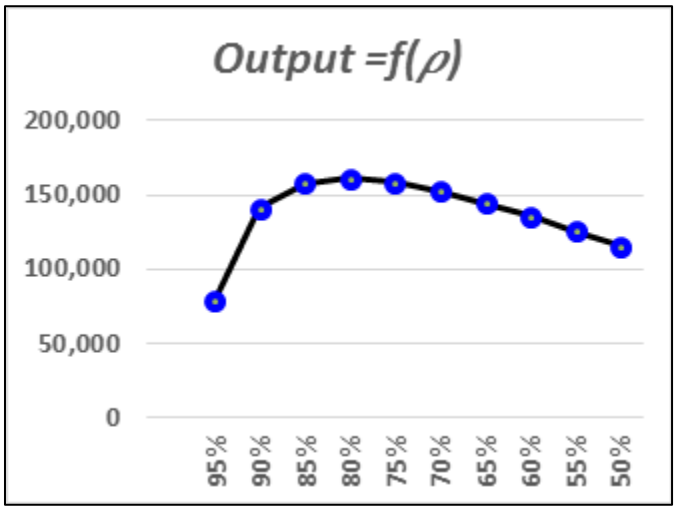

Figure 3: Concave output function (numerical example).

5 SIMULATION EXPERIMENTS

Next, we outline a small-scale simulation model that has been developed in order to validate the model from the previous section. Arena simulation software (version 15.00), was utilized and experiments were executed on a Core i5-4300U CPU at 1.9 GHz. Each experiment was replicated 200 times and simulated for a total of 15 equal periods: 10 ramp step periods and 5 stabilization periods (post end of ramp). As this is a transient simulation, the fab started without WIP and no warm-up period was introduced. The simulation model consists of three toolsets, with the following characteristics:

- Toolset A is a constraint ($\rho = 97\%$, remains constant throughout the ramp).
- Toolsets B and C are near-constraints ($\rho = 90\%$).
- The number of tools is scaled linearly as ramp progresses:
    - Toolset A begins with 2 tools at ramp start (1,000 wafer starts per week, or WSPW) and eventually increases to 20 tools at ramp peak (10,000 WSPW).
    - Toolsets B and C begin with 1 tool at ramp start and increase to 10 tools each at peak.
- All toolsets have availability of 85%, with exponentially distributed downtime (MTBF is 40 hours and MTTR is 6 hours).
- Each toolset has two operations, with the following re-entrance sequence: A1→B1→C1 →A2→C2→B2.

- Process times at each operation are distributed uniformly with values ranging ±14% around their mean.
- No batching or cascading and zero rework or yield loss.

The baseline simulation starts with a 1,000 wafer starts per week (WSPW) and ramps by 1,000 each period, until the peak at 10,000 WSPW. The capacity of each toolset is set according to this baseline ramp. In addition to the baseline scenario, 5 other scenarios were simulated, to model a reduced range of starts during ramp, at 75%-95% of the baseline scenario.

Figure 4 shows Cumulative Output Difference Ratio (CODR), which is the ratio between each scenario and the baseline scenario (100% loading). For example, if the baseline scenario resulted in an output of 1,000 wafers during period 1, while the 95% loading scenario resulted in an output of 1,200 wafers during the same period, then CODR of 95% scenario is defined as follows:

$$CODR(t=1) = \frac{1200-1000}{1000} = 20\% \tag{7}$$

This means that the 95% scenario provides 20% more output during the first period relative to the baseline scenario. As can be observed in Figure 4, the 80-85% scenarios are not only better in terms of cumulative output over the entire ramp period (providing 5-6% extra output by the end of the ramp) – but also maintain their superiority during the post ramp period. Clearly, more output earlier is preferred over the other scenarios even if eventually the differences diminish (after a sufficiently long post ramp period.) The 75% scenario, on the other hand, cannot keep up with the cumulative output of more loaded scenarios and, eventually, the significant lower portion of starts results in output loss during the post ramp period. Therefore, the capacity margin needs to be carefully analyzed before it is applied, such that it minimizes long term output loss.

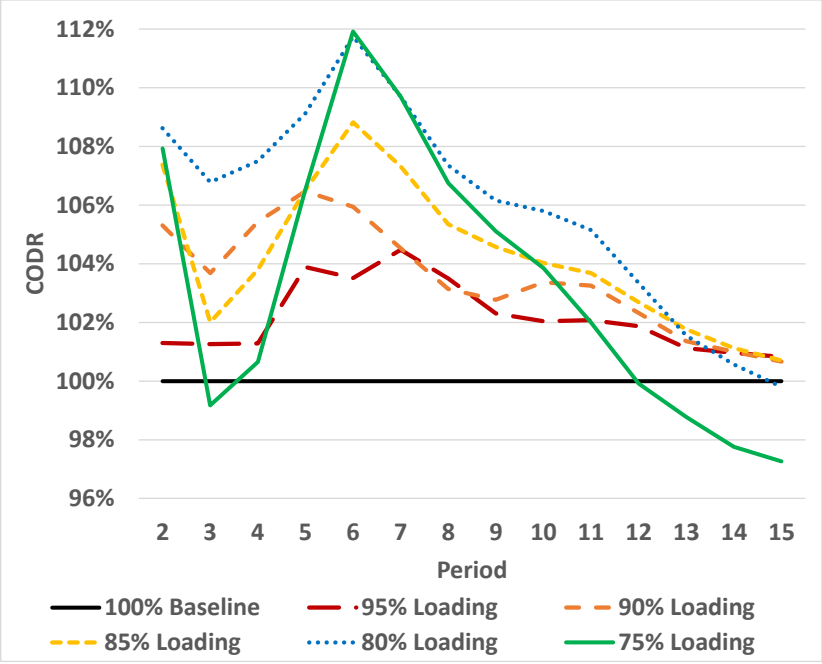

Figure 4: Average cumulative output, normalized versus baseline scenario (CODR).

6 CONCLUDING REMARKS

In this paper, a model has been proposed to address the question of optimal starts during ramp in order to maximize total output over a planning horizon, accounting for the installed capacity and expected cycle time. We devised an optimal solution to a special case (fixed constraint utilization) and supported our solution with simulation experiments. The simulation has incorporated real fab behavior in terms of factors such as re-entrant flow, multiple toolsets, tool installations occurring on different toolsets at different times, variance resulting from tool downtime events, etc.

Our key insights can be summarized as follows: (1) It is desirable to enforce a capacity buffer on the wafer starts introduced into the fab during ramp. The size of the buffer is dependent on the operating curve of the fab; (2) The sensitivity of the cumulative output function to deviations in the buffer is not steep – as long as deviations are within a small range (±5%) – but deteriorate fast when larger deviations occur; and (3) The gains achieved by enforcing a capacity buffer during ramp are expected on both output and cycle time, and during the ramp period as well as the post-ramp period.

In line with these conclusions, our recommendation for practitioners is straightforward even though hard to accept by managers: in ramp capacity planning and execution, do not give up that critical capacity margin!

Further work is in progress to extend the current work on two aspects: (1) extending the proposed model to account for varying utilizations (and subsequently cycle times) through the ramp; and (2) modifying the model to incorporate yields, their correlations with cycle time, and the overall impact of both on output.

REFERENCES

Benfer, R. H. 1993. *Learning During Ramping: Policy Choices for Semiconductor Manufacturing Firms.* M.S. Thesis, MIT, Cambridge.

Fayed, A. and B. Dunnigan. 2007. "Characterizing the Operating Curve – How Can Semiconductor Fabs Grade Themselves?" In *Proceedings of the 2007 International Symposium on Semiconductor Manufacturing (ISSM)*, 15-17 Oct, Santa Clara, CA: IEEE. DOI: 10.1109/ISSM.2007.4446827.

Grewal, N. S., A. C. Bruska, T. M. Wulf, and J. K. Robinson. 1998. "Integrating Targeted Cycle-time Reduction into the Capital Planning Process." In *Proceedings of the 1998 Winter Simulation Conference*, edited by D. J. Medeiros et al.,1005-1010. Piscataway, New Jersey: IEEE.

Haller, M., A. Peikert, and J. Thoma. 2003. "Cycle Time Management during Production Ramp-up." *Robotics and Computer-Integrated Manufacturing* 19(1-2):183–188.

Hopp, W. J. and M. L. Spearman. 2000. *Factory Physics: Foundations of Manufacturing Management*, 2nd ed. Boston, MA: McGraw-Hill.

Kalir, A.A. and K. Rozen, 2018. "Maximizing Output During Ramp by Integrating Capacity and Velocity." *IEEE Transactions Semiconductor Manufacturing*. 31(3):327-334.

Kim, S. and R. Uzsoy. 2008. "Integrated Planning of Production and Engineering Process Improvement." *IEEE Transactions on Semiconductor Manufacturing*, 21:390-398.

Leachman, R. C., and D. A. Hodges, 1996. "Benchmarking Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing* 9(2):158-169.

Manda, A.B., R. Uzsoy, K. G. Kempf and S. Kim, 2016. "Modeling the Impact of New Product Introduction on the Output of Semiconductor Wafer Fabrication Facilities." In *Proceedings of the 2016 Winter Simulation Conference*, edited by T.M.K. Roeder et al., 2547-2558. Piscataway, New Jersey: IEEE.

Terwiesch, C. and R. E. Bohn, 2001. "Learning and Process Improvement During Production Ramp-up," *International Journal of Production Economics* 70(1):1–19.

Tirkel, I., G. Rabinowitz, D. Price, and D. Sutherland. 2016. "Wafer Fabrication Yield Learning and Cost Analysis based on in-line Inspection." *International Journal of Production Research* 54(12):3578-3590.

Tirkel, I., N. Reshef, and G. Rabinowitz. 2009. "In-line Inspection Impact on Cycle Time and Yield." *IEEE Transactions on Semiconductor Manufacturing* 22(4):491-498.

AUTHOR BIOGRAPHIES

ADAR A. KALIR received his B.S. and M.S. degrees in industrial engineering and management from Tel-Aviv University, Israel, and his Ph.D. degree in industrial and systems engineering from Virginia Tech. He is a Sr. Principal Engineer at the Fab/Sort Manufacturing network of Intel Corp., responsible for the application of operational optimization in high volume manufacturing across Intel's factories, driving improvements in WIP management, production capacity and cycle time, equipment and capital productivity. He is also an Adjunct Associate Professor at Ben-Gurion University, Israel and serves as a co-chair of the IEEE Technical Committee on Semiconductor Manufacturing Automation (TC-SMA). His email address is adar.kalir@intel.com.

KOSTA ROZEN received his B.S. and M.S. in industrial engineering and management from Ben-Gurion University, Israel, in 2005. He is currently a Senior Simulation Engineer at Intel Corporation in Qiriat-Gat, Israel. In his position, he is responsible for full fab simulations, modeling and making recommendations to senior management. He is also an Adjunct Lecturer at Ben-Gurion University, Israel. His email address is kosta.rozen@intel.com.