

DETERMINING OPTIMAL LOT-SIZE FOR A SEMICONDUCTOR BACK-END FACTORY

Juergen Potoradi
Gerald Winz

Infineon Technologies Asia Pacific
168 Kallang Way
SINGAPORE 349253

Lee Weng Kam

Infineon Technologies (Integrated Circuit) Sdn Bhd
Free Trade Zone, Batu Berendam
Melaka, MALAYSIA

ABSTRACT

Modeling analysts are using a methodology that applies queuing theory logistics laws and simulation to factory performance analysis. These methods are being applied at semiconductor back-end factories, where a major focus is on achieving capacity increases with minimal equipment additions.

This paper describes this technical methodology and investigates an optimum lot-size for back-end factories based upon given throughput and cycle time targets. The analysis provides a recommended lot-size of 6800 for the overall production area, allowing the factory to maximize throughput while still meeting overall factory cycle time goals. The model indicates a potential 14% increase in throughput by selecting the optimal lot-size.

1 INTRODUCTION

Productivity improvement efforts in the semiconductor industry have historically focused on wafer fab operations. However, it has recently been recognized that the so-called "back-end" factories have great potential for improvement, particularly in the area of production logistics. Figure 1 shows the general material flow of semiconductor manufacturing from wafer fab through the elements of the back-end factory (pre-assembly through mark/scan/pack).

The challenge for back-end productivity improvement is to maximize equipment utilization and achieve high capital efficiency while maintaining cycle time targets. Cycle time is important to meet product delivery dates and to minimize feedback time about quality and yield to the wafer fabs. Of course, high utilization and short cycle time are two conflicting goals. Because high utilization leads in most cases to high WIP, Little's Law tells us this will increase cycle times. Queuing theory applications (Hopp and Spearman 1996) teach us that the cycle time depends not only on the utilization, but also on variability and raw

process time (RPT). The latter is largely determined by the lot-size in back-end operations. Other studies have also indicated that for production systems with reentrant flow, one of the three most influential factors on system performance is lot-size (Adachi *et. al.* 1989).

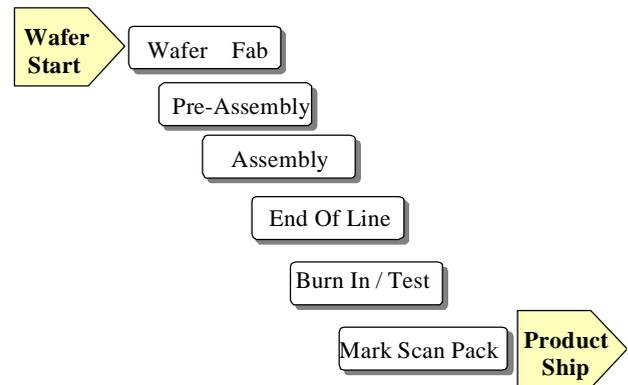


Figure 1: A Simplified Semiconductor Production Flow

Determining an optimum lot-size for the back-end factory is not a straightforward exercise since the optimal lot-size in assembly is generally different from the one in test. The material flow in pre-assembly and assembly is linear, much like the linked lines in the automotive industry. Therefore, these tend to have less variability. This is not true in the test area where re-entrant flow significantly increases non-uniformity of the arrival rates. Additionally, the back-end contains areas with stand-alone equipment and areas with dedicated auto-lines (similar to the automobile manufacturing industry). These inconsistencies in the material flow, the high degree of product variety of most logic-product back-ends, and the varying lot-sizes make production logistics difficult. The effective coordination of tools, process, WIP and manpower is critical to reducing variability.

2 ANALYTIC APPROACH: LOT-SIZE INFLUENCE

2.1 Cycle Time and Lot-size

When looking at the influence of the lot-size on a given type of back-end equipment, one might encounter a curve of the shape shown in Figure 2.

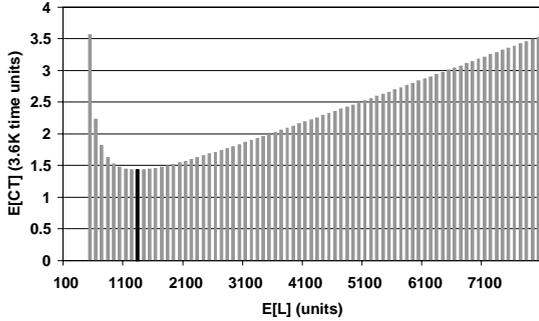


Figure 2: Cycle Time CT

This curve shows the relationship of cycle time and lot-size for a trim/form tool, but is indicative of many factory equipment types. What these types of equipment all have in common is a certain time that is needed for loading and unloading of a lot. That means there is a fixed amount of time that has to be spent for each lot, independent of the lot-size. For every type of equipment where the processing time of a lot consists of a fixed time for overhead and a time slice for each unit, the cycle time curve has this askew “U” shape.

In order to explain this curve, we consider the cycle time formula for M/GI/1 systems (Tran-Gia 1996). There the mean cycle time CT is given as a function of utilization ρ , service time B and the coefficient of variation of the service time c_B^2 :

$$E[CT] = \frac{1+c_B^2}{2} \frac{\rho}{1-\rho} E[B] + E[B]$$

(For an English derivation of this formula, see Hopp and Spearman (1996, pg. 295).) As a next step, we have to look at the lot-size dependency of the three factors B , ρ and c_B^2 .

The service time increases linearly for increasing lot-size L , according to the following formula. The constant overhead part is denoted by C .

$$E[B] = E[L] + C$$

For simplification, we assume that the time needed to process one unit equals one time unit. Typical values for the overhead C are 800 time units for testers and 1500 time units for trim/form tools. Figure 3 shows the service time curve:

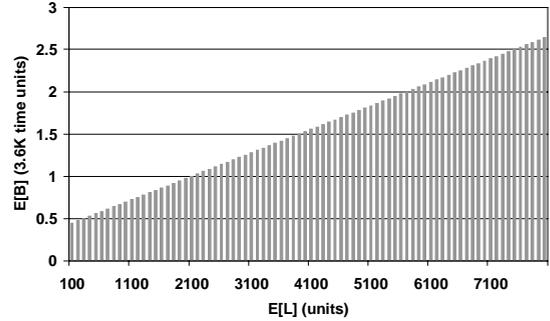


Figure 3: Service Time B versus Lot-size L

The utilization ρ is determined by the mean of the service time B and the inter-arrival time A of the lots. Consequently we just need the following formula for the mean inter-arrival time to get the lot-size dependency of ρ . The mean inter-arrival time is determined by the mean lot-size and the arrival rate based on units per unit of time, λ_{unit} .

$$E[A] = \frac{E[L]}{\lambda_{unit}}$$

So we get for the utilization ρ :

$$\rho = \frac{E[B]}{E[A]} = \frac{\lambda_{unit} (E[L] + C)}{E[L]}$$

Based on this formula we get a condition for λ_{unit} in order to have a stable system:

$$\lambda_{unit} < \frac{E[L]}{E[L] + C}$$

The utilization/lot-size curve for different values of C has a distinctive shape, as shown in Figure 4.

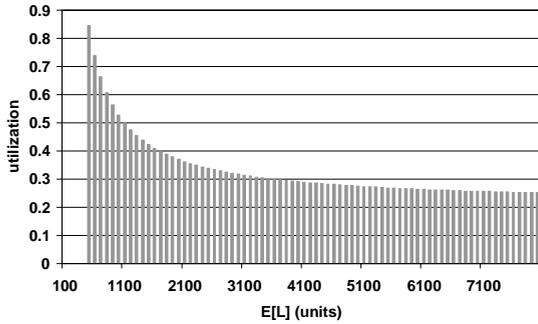


Figure 4: Utilization ρ versus Lot-size L

The curve in Figure 4 ($C=1500$, $\lambda_{unit}=0.21$) shows an overloaded system for lot-sizes below 500. In this case the condition for λ_{unit} does not hold for lots smaller than 500.

For this analysis we assume that a change of the mean lot-size has no influence on the distribution of the arrival and service processes. Therefore, the coefficient of variation c_B^2 , as well as the Markovian arrival process, is independent of the lot-size change. This might not be true in a real manufacturing system, since large-size lots might be treated differently from smaller-sized lots. For example, if an automated transportation trolley is used always at full capacity for lot transport to a tool, then the number of lots arriving at the same time depends on the lot-size. These kinds of dependencies are considered in the simulation models.

To summarize, it can be said that the lot-size has an influence on the utilization and on the service time. Accordingly the cycle time dependency on the lot-size is a function of both utilization and service time dependency. This can easily be seen from the curves in Figures 2, 3, and 4. The left part of the curve in Figure 2 is influenced by the utilization. The steep drop in the utilization is even amplified in the cycle time formula. The right part of this curve reflects the linear increase of the service time.

2.2 Changing Utilization and Variability

The curve in Figure 2 shows clearly that 1300 is the optimum lot-size for the described tool. But, what will happen if we start more material on the tool or the variability of the service time increases?

For increasing λ_{unit} (from 0.1 to 0.6) the cycle time curve shifts to the right upper corner (see Figure 5). This happens under the influence of the utilization, since the service time is not affected by a change in λ_{unit} . In Figure 5 we can see the minimum lot-size increasing faster than the linearly-increasing arrival rate λ_{unit} .

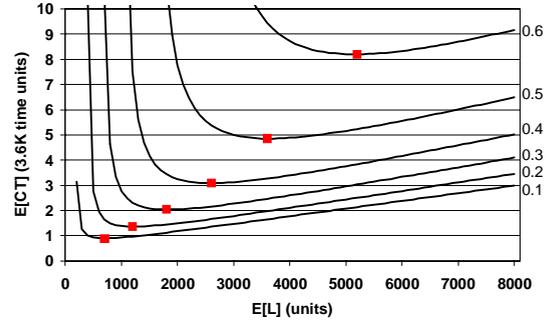


Figure 5: Cycle Time versus Lot-size for Different Arrival Rates

Compared to the significant influence of equipment utilization on the optimum lot-size, a higher variability of the service process has a rather negligible effect on optimum lot-size.

Figure 6 shows the cycle time curves for the coefficient of variation c_B^2 increasing from 0 to 4. Based on this we can say that the optimum lot-size is much more sensitive on a utilization increase than to a variability increase.

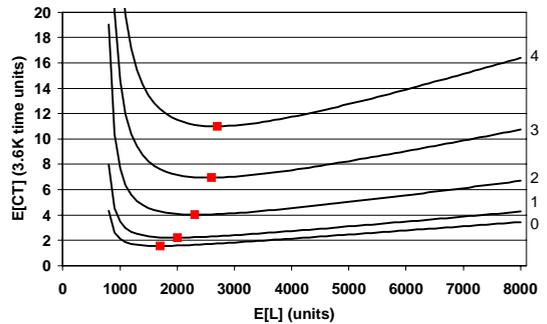


Figure 6: Cycle Time for Different Service Time Distributions

2.3 Throughput, Operating Curves

Since people on the production floor are very throughput oriented, the question might arise of what lot-size will give maximum throughput for a given cycle time. In this case we use the same cycle time equation to calculate, for different lot-sizes, the number of units that can be released in order to achieve a certain cycle time target.

Figure 7 shows the arrival rate λ_{unit} for lot-sizes between 100 and 8000 and for a target cycle time of 3.5 hours (one hour equals 3600 time units). This curve has an inverted “U” shape with an optimum lot-size at the maximum arrival rate λ_{unit} . This is also the maximum throughput since we consider only stable systems. The explanation of the “U” shape is again found in the steep

utilization curve for smaller lots and the increasing service time for larger lots.

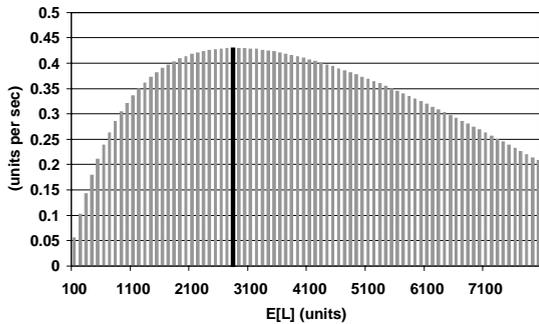


Figure 7: Arrival Rate that Gives a Cycle Time of 3.5 Hours for Each Lot-size

A summary of the various curves discussed thus far is given in Figure 8, showing the relationship between cycle time and arrival rate (throughput) (Hopp and Spearman 1996). At Infineon Technologies, this is referred to as the “operating curve”. In Figure 8 the operating curves for five selected lot-sizes are shown.

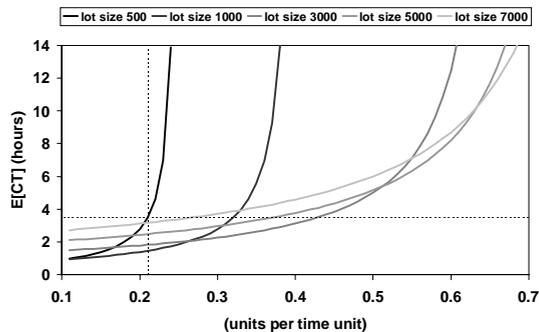


Figure 8: Operating Curves for Different Lot-sizes

The curve shown in Figure 7 can be viewed as a horizontal section at cycle time 3.5. Note that the curve for lot-size 3000 crosses the line furthest to the right, showing again that for a cycle time of 3.5, 3000 is the lot-size that gives maximum throughput. The curve in Figure 2 can be viewed as a vertical section at the arrival rate 0.21. To get the curves in Figure 5, vertical sections for the correspondent arrival rates have to be considered.

The application of queuing formulas is more complex for systems with more than one piece of equipment, several consecutive process steps, variable down times, setups, and other additional parameters. Nevertheless, these simple curves and formulas are very useful to explain simulated cycle-time curves of these complex systems. The simulated curves for a multi-step system consist more or less of superimposed curves of the single steps.

3 SIMULATION APPLICATION: LOT-SIZE INFLUENCE

3.1 Project Approach

This analysis is part of the long-term goal of introducing simulation as a tool for management decision support in the back-ends in Malacca (Malaysia) and Singapore. The specific question concerning the factory is: What is the optimum lot-size?

In order to answer this question for the Malacca back-end, two different models were built: one for the combined assembly/end-of-line areas and one for the burn-in/test area. For the pre-assembly area, lots arrive from the previous production department in a predetermined size. The “overhead” time is negligible in the mark/skan/pack area, so lot-size is not so important. Therefore these areas are not considered in this analysis.

The models were built, validated, and used for analysis by a team of process engineers from the specific production areas, in partnership with a simulation expert. This team structure, as well as regular meetings with production people and management, ensured that the simulation results met expectations and requirements. Details of this project management approach can be found in Chance, Robinson, and Fowler (1996).

3.2 Data

Data from the planning department were used as the core input data for the simulation models. More detail was added concerning scheduled and unscheduled down times, process flow, and variability. Since the static planning data uses just average values for calculations, any data about distributions (e.g., lot-size, lot arrivals, cycle time) had to be gathered from the manufacturing execution system (MES) and other standard factory reporting systems. Data about equipment states and process times was gathered manually and was restricted to the average values (more sophisticated CIM online data collection systems are not installed in this factory). For any missing data, triangular distributions were estimated (Law and Kelton 1991).

3.3 Validation

For a successful simulation project, two conditions are very important:

- models must be valid
- the model and the output must be trusted.

The model validation was done by comparing model output to actual historical data for a given period. We used cycle time and OEE (Overall Equipment Efficiency) output charts for the comparison. By comparing these parameters

we ensured that the variability in our model was comparable to the actual system. The cycle time is a function of utilization and variability (Hopp and Spearman 1996). Therefore, if we compare the utilization (OEE chart) and the cycle time, we can infer a match of the variability.

Proof that results from a simulation study are trusted is evidenced only by the factory managers using simulation output in their decision-making process. Simulation for simulation's sake is of no value in a production environment. The team structure and the regular meetings contributed significantly to meeting this goal.

3.4 Software

This project used the performance analysis software Factory Explorer®, from Wright Williams and Kelly (Chance 1996), which proved to be a very effective tool for modeling back-end operations. Building from previous modeling experiences within Infineon Technologies (Domaschke, *et. al.* 1998), the total time to train the factory analysts, build valid models, and conduct the analysis was less than three months.

3.5 Results

The optimum lot-size is dependent upon the objective function of the simulation exercise. Optimizing for the minimum cycle time, for example, will give a different answer than optimizing for maximum throughput. In order to answer both questions (i.e., determining the optimum lot-size for a given cycle-time-constrained capacity) the simulation output is shown as the operating curves described in section 2.3. Cycle-time-constrained capacity (Fowler and Robinson 1995) is defined as the maximum throughput rate sustainable for the factory for a given product mix, line yield, and equipment set, and a constraint on the average cycle time.

In Figure 9 the operating curves for different lot-sizes for the assembly/end-of-line area are shown.

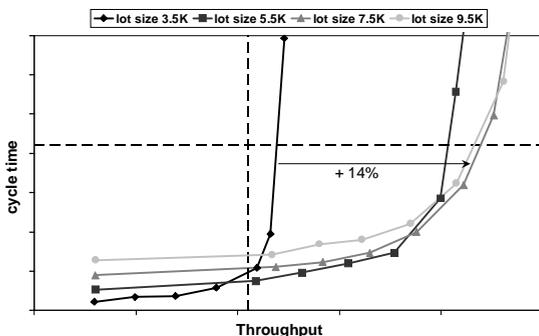


Figure 9: Operating Curves for Different Lot-sizes; Assembly/End-of-line

The horizontal line stands for the cycle time target, the vertical line for the planned throughput. From this analysis we can say the following:

- The lot-size of 3500 is the minimum lot-size for the planned throughput. Static calculations from the Factory Explorer® analysis indicate that a lot-size of 3000 will lead to an unstable model.
- The minimum cycle time for the planned throughput can be achieved with lot-sizes around 5500. Lots with 3500, 7500, and 9500 show a higher cycle time. However, this minimum cycle time is only slightly lower than the cycle time for a 7500 unit lot-size, and the latter has significantly higher throughput.
- There is a potential for a 14% increase in throughput by going from 3500 lot-size to 7500 lot-size and still meeting the cycle time target. The 14% throughput gain is due to the decreasing utilization for increasing lot-sizes, as discussed in section 2.1. The smaller increase from 5500 lot-size to 7500 lot-size is based on the flat utilization curve for the larger lots. Additionally, when the lot-size is increased above 5500, the bottleneck tool changes from a tool that is lot-size sensitive to a tool that is not lot-size sensitive.
- The throughput can not be further increased for the given cycle time target by increasing the lot-size to 9500 and higher. See also Figure 7 in section 2.3, where the maximum throughput is decreasing for lots bigger than the optimum.

Figure 10 shows operating curves for different lot-sizes for the burn-in/test area.

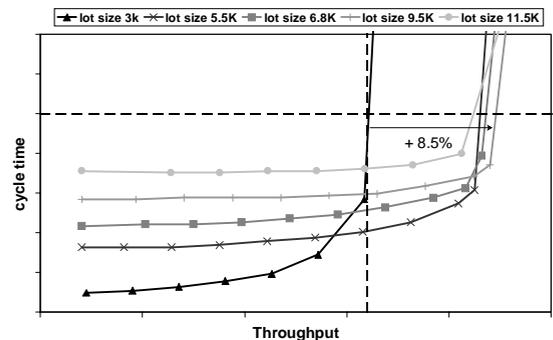


Figure 10: Operating Curves for Different Lot-sizes; Burn-in/Test

Again we have a horizontal line for the cycle time target and a vertical line for the planned throughput. For the burn-in/test area, we can say the following:

- 3000 is the minimum lot-size for the planned throughput. The figure shows clearly that any smaller lot-size will lead to an unstable system.
- The minimum cycle time along the vertical line can be achieved with lots around 5500 and less.
- The throughput could be increased 8.5% by going from 3000 lot-size to 9500 lot-size and will still meet the cycle time target. The reasons for the high throughput jump from 3000 lot-size to 9500 lot-size, and the relatively small changes for lots larger than 9500, are the same as in the assembly/end-of-line analysis. One reason is the increasing flatness of the utilization curve for increasing lot-sizes (see Figure 4 section 2.1). A second reason is the change of the bottleneck at lot-sizes around 5500 from a lot-size sensitive tool to a tool that is not lot-size sensitive.
- A further increase of the lot-size to 11500 and above will not increase the throughput for the given cycle time target. The maximum throughput lies around 9500 lot-size. See also Figure 7 in section 2.3, where the maximum throughput is decreasing for lots bigger than the optimum.

4 CONCLUSION

For this modeled factory, the optimum lot-size is 7500 for the assembly area and 9500 for the burn-in/test area. This allows the factory to maximize its throughput in the certain area while still meeting the established cycle time goal.

However, an additional operational consideration is that the factory operates the MTX ovens on a full-batch loading policy. The maximum batch-size for these ovens is 6800, so for the burn-in area this would be considered a locally optimum size for effective production operations.

Comparing the 6800 lot-size to the optimum lot-sizes for assembly/end-of-line (7500) and burn-in/test (9500), the simulation results show relatively minor differences in the throughput. Therefore, for the overall production area the recommended lot-size is 6800. This finding reinforces the fact that simulation should be used to analyze not only local improvements, but also the overall impact on the factory as one entity.

This study reinforces the importance of applying queuing theory and simulation together as one unified approach. Used separately, the Operating Curve Management (OCM) approach (using queuing theory)

simply indicated the benefit of increasing the lot-size at specific tools. Only after applying simulation in partnership with OCM was the team able to fix an upper bound on the optimum lot-size.

A general observation can be made concerning lot-sizes. For areas of the factory that are highly utilized, a larger lot-size is required to meet throughput. For areas less utilized, a smaller lot-size can be implemented to minimize cycle time.

This analysis demonstrates that simulation models can be rapidly built and effectively maintained to provide quick updates and recommendations as the factory changes. With part-time simulation analysts, the Malacca factory is able to easily provide continuous simulation results for managerial decision-making.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the technical contribution of Dr. Jennifer Robinson, Chance & Robinson, Inc., and the editorial assistance of Mr. Steven Brown, Infineon Technologies. We thank our many partners from the Malacca and Singapore factories for their active participation. Special thanks to Bong, Neo, Mohan, KG, Khoo and Melvin.

REFERENCES

- Adachi, T., J.J. Talavage, and C.L. Moodie. 1989. A rule-based control method for a multi-loop production system. *Artificial Intelligence in Engineering*, Vol. 4, No. 3, 115-125.
- Chance, F. 1996. Factory Explorer® users' guide.
- Chance, F., J.K. Robinson, and J.W. Fowler. 1996. Supporting manufacturing with simulation: model design, development, and deployment. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J.M. Charnes, D.J. Morrice, D.T. Brunner, and J.J. Swain, 114-121. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.
- Domaschke, J., S. Brown, J.K. Robinson, and F. Leibl. 1996. Effective implementation of cycle time reduction strategies for semiconductor back-end manufacturing. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D.J. Medeiros, E.F. Watson, J.S. Carson, and M.S. Manivannan, 985-992. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.
- Fowler, J.W. and J.K. Robinson. 1995. "Measurement and Improvement of Manufacturing Capacity (MIMAC) Designed Experiment Report", SEMATECH Technology Transfer #95062861A-TR.
- Hopp, W. J., and M.L. Spearman. 1996. *Factory Physics: Foundation of Manufacturing Management*. Chicago: Irwin.

Law, A. M., and W.D. Kelton. 1991. *Simulation Modeling & Analysis*. McGraw Hill.

Tran-Gia, P. 1996. *Analytische Leistungsbewertung verteilter Systeme*. Berlin: Springer.

AUTHOR BIOGRAPHIES

JUERGEN POTORADI is a Factory Modeling and Simulation analyst and project leader with Infineon Technologies (formerly Siemens Semiconductor Division). He is currently responsible for implementing simulation techniques and methodologies in the Asian back-end factories: Singapore and Malaysia. Mr. Potoradi received his undergraduate and graduate degrees in Computer Science from the University of Wuerzburg in Germany. He has extensive experience in simulation analysis of semiconductor wafer fab and back-end production operations. His email address is juergen.potoradi@infineon.com.

GERALD WINZ received his doctorate in engineering from Fraunhofer Institute, Dortmund. After logistics consulting for the German production and trade industry, he joined Infineon Technologies (then Siemens Semiconductor Division) in 1997. Dr. Winz is in charge of OCM introduction (Operating Curve Management, a queuing theory application) in the Asian back-end factories: Singapore and Malaysia. His email address is gerald.winz@infineon.com.

LEE WENG KAM has been in the Semiconductor industry for the past 20 years in various capacity as Production Maintenance Engineer, Development Mechanical Engineering Manager to Manufacturing Process Engineering Manager. Projector leader for the OCM implementation in Drams, Infineon, Melaka. His email address is weng-Kam.lee@infineon.com.