Determining the Slack Capacity of Testers thru Model Simulation

Herwina Richelle C. Andres
Operations Research – Test Equipment Engineering
Analog Devices, Inc.
General Trias, Cavite 4107, Philippines

Abstract

The complexity of semiconductor manufacturing and the rising demand for testing and wafer fabrication in the industry led to the objectives of this research as follows; to analyze the behavior and relationship between the variables in production and selected performance measures: Cycle Time (CT), Work-In-Process (WIP) and System Utilization and to design an optimal slack capacity model that will produce parts at an ideal CT. Noteworthy outcome of this study was a slack capacity percentage that served as basis for an equipment monitoring plan that the company may use in deciding on system acquisition. This led to the construction of a simulation model of the usual testing process in a semiconductor company for which the variability of the volume arrival rate, test rate and system downtimes were considered. Simulation software, FlexSim, was used in the building of the models. Results show that increasing arrival rate led to an increase in System Utilization; decreasing the number of systems and system downtimes (DT) will as well increase Utilization. Furthermore, as System Utilization increases, both CT and WIP Level increase. Moreover, the company must maintain 22.72% Slack Capacity. Otherwise, it will experience major changes on the performance measures; System Utilization, WIP and CT.

Keywords
Cycle Time, Work in Process, System Utilization

4. Introduction
The Semiconductor and Electronics Industries of the Philippines Inc. stated that 2010 would be ‘a growth year’ for the industry, after going through 6 major cycles since the 1970’s as exports are expected to start recovering from the 2009 worldwide recession, overcapacity and inventory burn. This is true as the big players here in the country experience high demand from both offshore and on-shore clients and are now operating at a full blown.

The basis for this study is Company A, from which all the data were gathered. Their performance shows the need to establish a headroom model due to the rising demand and decreasing headroom. The data shows that as the demand increases, its capacity is put into fuller use, thus, headroom decreases.

Company A is experiencing an increase in CT per quarter and the notable presence of work in process (WIP). Thus, the maintenance of appropriate headroom percentage is necessary to keep CT and WIP at a favorable value; this becomes the challenge. This situation paved the way in establishing the objectives of this study which are as follows: to determine the relationships between the three mentioned performance measures with the variables, arrival rate, and service rate and down time events, to determine the relationship of machine utilization with CT and WIP Level. In addition, the study aims to design an optimal headroom model that will produce parts at a favorable CT which will be based on the relationships that will be attained.

In Company A, a system is a composition of tester, handler and the interface between them which are board and contactor. It is used in testing of devices subjected under different conditions.

Figure 1 presents the major composition of the total time each system has. Its major sub composition is Downtime and System Availability. The former is the period of time that a system cannot do its intended function due to several failure issues. The latter on the other hand is said to be the period of time that the system is available to test parts. It is subdivided into two, namely, Volume Requirement and Slack Capacity. Volume requirement is the period of time the system is loaded for testing. Slack Capacity is the percentage of time that the system is “free-of-load”.

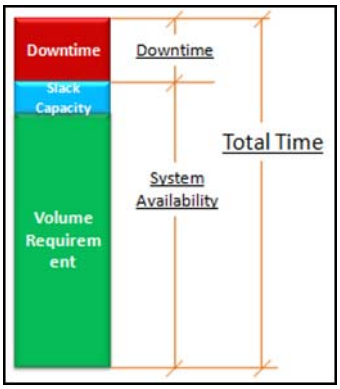

Figure 1: System Availability

Slack capacity affects significantly one of the performance measures that manufacturing facilities strictly monitors, Cycle Time (CT). Semiconductor companies are challenged to meet the aggressive demand of the market at a specified date. This is extremely reliant on the CT of lots which is a function of queuing time and processing time. A mathematical model founded on queuing theory will determine the relationships between the number of systems and the randomness of the following variables: volume arrival process, test process and system downtimes (DT).

Queuing Theory

A queuing model analyzes systems that provide service to random demands. In the perspective of Company A, it represents, firstly, the system's physical configuration, by specifying the number and arrangement of the systems, which provide service to the volume arrivals, and, secondly, the stochastic nature of the requirement, by specifying the variability in the volume arrival process and in the test process.

At the end, an overall analysis will be shown regarding the behavior and relationship of the selected performance measures: Cycle Time (CT), Work-In-Process (WIP) and System Utilization in responses to the randomness of the variables: volume arrival process, test process, system downtimes and number of systems; and to design an optimal slack capacity model that will produce parts at an ideal CT.

1.1 Examination of Factors Affecting Performance Measures

The following are the several factors that would be influential to the values of the performance measures:

- Volume Arrival Process
- Test Process
- System Downtimes (DT)
- Number of Systems

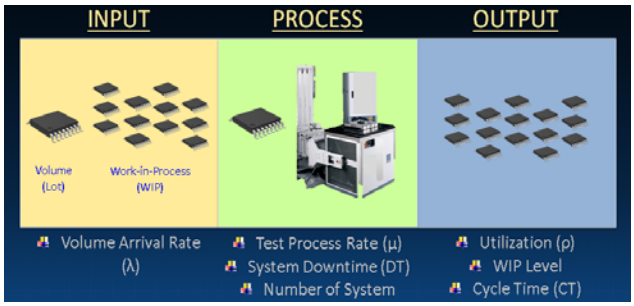

Figure 2: IPO of the System

1.1.1 Volume Arrival Process

The most universal assumption about the volume arrival process is that the arrival times follow a Poisson process which is a very good approximation in real systems. One of the properties of the Poisson process is that the times

between arrivals, the inter-arrival times, are exponentially distributed. A random variable X, is said to be exponentially distributed if its distribution function $F_X(x)$ is given by $F_X(x) = 1 - e^{\lambda x}$ for all $t \geq 0$, where $1/\lambda$ is the average value of X.

Arrival time distribution
The simple model assumes that the number of arrivals occurring within a given interval of time t follows a Poisson distribution with parameter $\lambda(t)$. This parameter $\lambda(t)$ is the average number of arrivals in time t which is also the

variance of the distribution. If n denotes the number of arrivals within a time interval t, then the probability function $p(n)$ is given by,

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad (1)$$

The arrival process is called Poisson input.

1.1.2 Test Process
This is the probability density distribution that determines the volume (lot) testing in the system. Test rate denotes the rate at which lots are being processed in a system. It is the reciprocal of the test time (TT). The statistical distribution observed by the test processing time is an exponential distribution which describes the times between events in a Poisson process; it is a process in which events occur continuously and independently at a constant average rate.

1.1.3 System Downtimes (DT)
System downtime is the act of disrupting an established order so it fails to continue processing. This downtime has significant effects on the performance measure. One study said that equipment downtime is the most significant contributor. It was stated that both random failures and preventive maintenance events increase the variability of the test processing times experienced by lots arriving at systems. That is if a lot arrives at a down system, the time spent waiting for the system to come back up becomes effectively part of the lot's test processing time. Variability of the latter, increases lot CT. The magnitude of this effect is influenced by the amount of variability in DT.

1.1.4 Number of Systems
Number of system defines the service capacity when multiplied with test rate. The test process mechanism of a queuing system is specified by the number of systems, each system having its own queue or a common queue and the probability distribution of lot's test time. System's service capacity is a function of the number of service facilities and system proficiency.

2. Experimental Section
Kendall Notation is a standard three-factor notation system for classifying the queue model that the system corresponds to. The first letter specifies the lot inter-arrival time distribution and the second one the test process time distribution. M pertains to the exponential distribution which stands for Memoryless. The third and last letter specifies the number of systems.

M/M/S model were simulated to analyze the impact of the different factors in the metric.

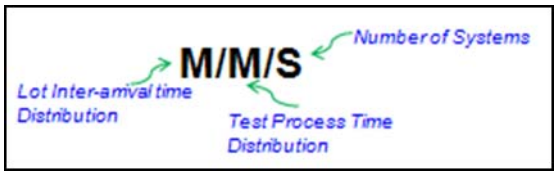

Figure 3: Kendall Notation

In order to determine the effects, we have considered the following scenarios in the simulation model:

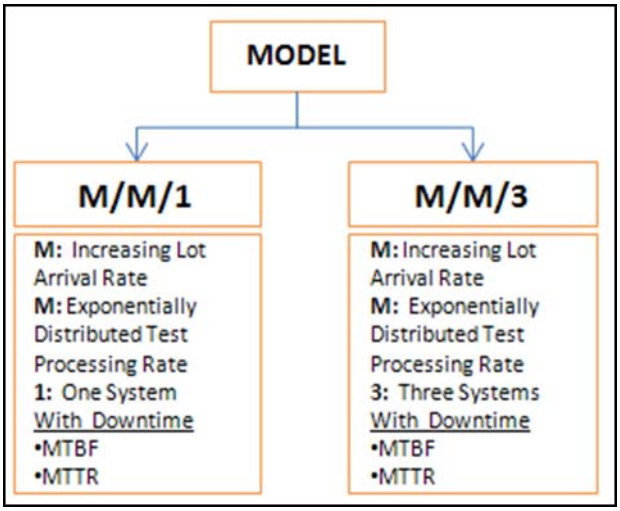

Figure 4. Simulated Scenarios

The scenarios in Figure 4 were modeled using FlexSim Simulation Software. It claims to be the most powerful tool for modeling, analyzing, visualizing, and optimizing any imaginable process - from manufacturing to supply chains, abstract examples to real world systems, and anything in between. A snapshot of the model is shown in Figure 5.

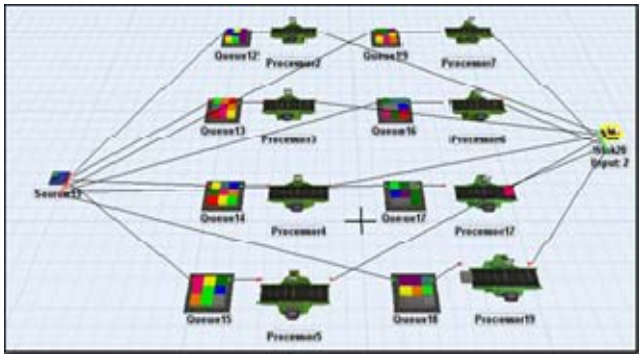

Figure 5. Simulation Model

2.1 Scenario 1: M/M/1

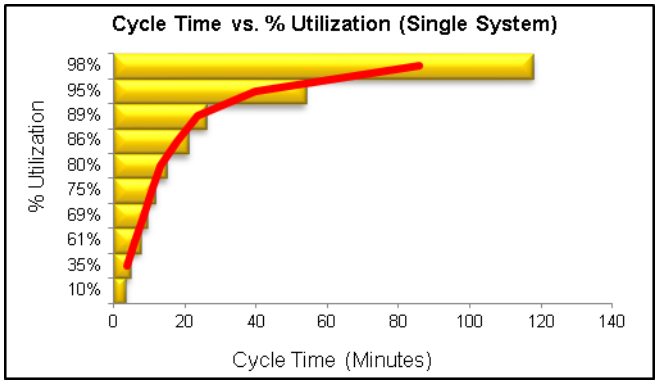

Figure 6. Analysis on Cycle Time with Increasing Lot Arrival Rate

Figure 6 shows the relationship of the CT to machine utilization. As the machine was utilized, the CT tended to get longer. This is because CT is a function of queue time and processing time. The more utilized the machine was, the more it could not accommodate more arrivals. Thus, queue started to build up which led to a much longer queue time.

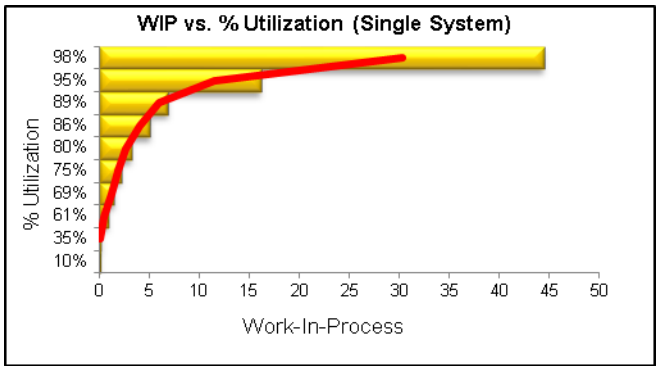

Figure 7. Analysis on WIP Level with Increasing Lot Arrival Rate

Figure 7 shows the relationship of the WIP level to machine utilization. This relationship justifies the statement made regarding Figure 6. The results showed that WIP level tended to increase the more the machine was utilized. Since the machine had high utilization, and arrivals were increased, the machine went beyond its capacity. Thus WIP built up in the system. Figures 6 and 7 also showed that a big leap on CT starts when System Utilization is pushed beyond 80%. Their relationship therefore, is directly proportional.

2.2 Scenario 2: M/M/3

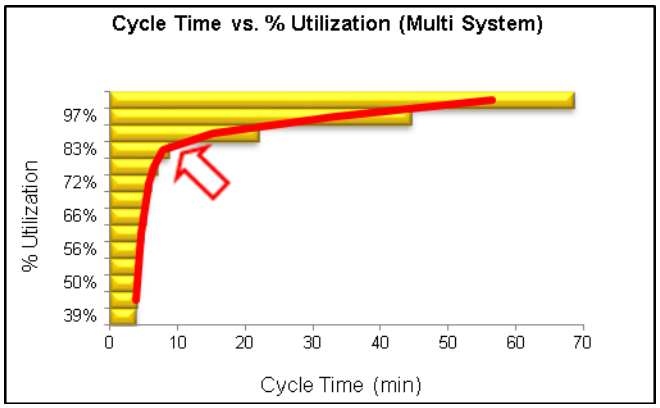

Figure 8. Analysis on Cycle Time with Increasing Lot Arrival Rate4.1 Text Sections and Headings

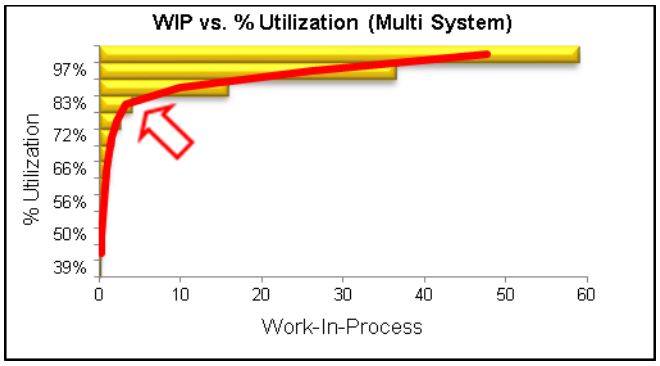

Figure 9. Analysis on Work-In-Process with Increasing Lot Arrival Rate

For multisystem, the same relationship in single system was observed; factors are directly proportional. Figures 8 and 9 showed that a big leap on CT starts when System Utilization is pushed beyond 83%.
The figures above illustrate the relationship of the performance measures to one another and its relationship to the variables. Another variability considered here was the number of machines. From the analysis of one server, adding up machines had no effect on the relationship. It can be noted from the findings that all observed a direct proportional relationship. As the arrival rate increased, the machine utilization increased too even if the machines had more capacity. The more utilized the machines were, the longer the CT. Since WIP level increased as utilization increased, it can be understood why CT got longer since as mentioned, it is a function of queue time or the waiting time of the WIP.

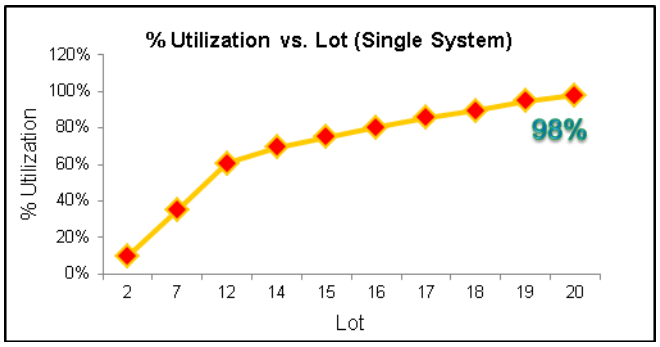

Figure 10. Analysis on System Utilization with Increasing Lot

Having systems loaded beyond their defined capacity, maximum percentage utilization is 98%. Thus leaving 2% slack capacity. It is expected that a system without downtimes will be fully utilized, and the more lots arrive, the more the graph will be asymptotic, the remaining portion of course will be attributed to the idle time. The observed relationship of the system utilization and lot arrival rate is directly proportional meaning as the lot arrival rate is increased it corresponds also to an increase in the CT.

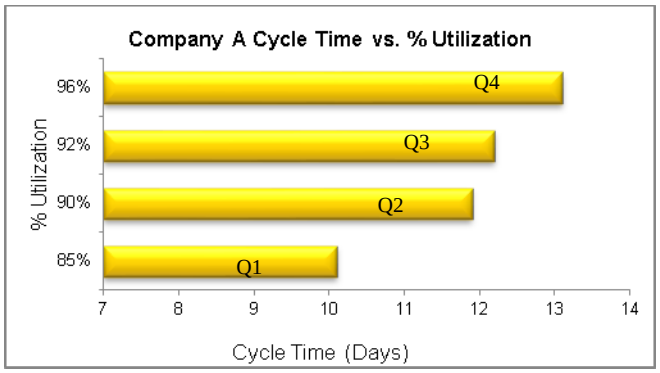

Figure 11. Actual CT of Company A against System Utilization

Results of the simulation reflected in the actual performance of Company A. As systems are being more utilized, days to finish testing gets longer.

3. Results and Discussion

The relationships of the variables and metrics are summarized in the Figures 12 and 13. They show that increasing volume (lot) arrival rate led to an increase in System Utilization; decreasing the number of systems and System DT will as well increase Utilization. Furthermore, as System Utilization increases, both CT and WIP Level increase.

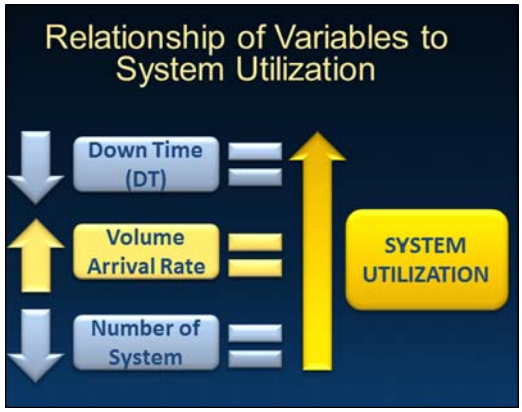

Figure 12: Relationship of Variables to System Utilization

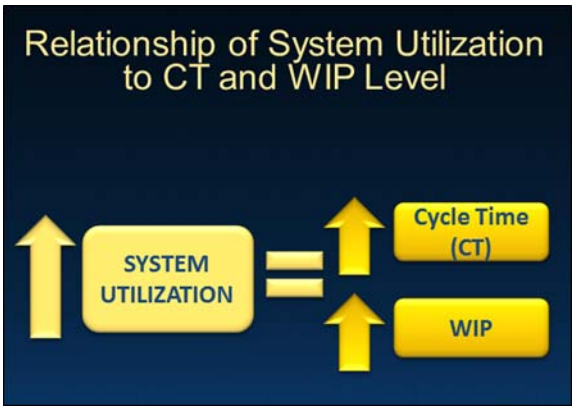

Figure 13: Relationship of System Utilization to Cycle Time and WIP Level

The table shown below is the final Equipment Monitoring Plan for the area taken into account. The limitations of the trial version of the software prevented the modeling of the actual scenario. The table shows that slack capacity should be maintained at 22.72% since the input lots were promised to be delivered after seven days of production. This was on the assumption that the payoff of late delivery would cost Company A significantly. The value is not standardized for all semiconductor companies, as each has different priorities. For example, Company B can allow at least 3 days delay rather than acquire the corresponding number of Systems to deliver on time, while Company C will acquire Systems to deliver to their customers on the declared date.

Table 1. Equipment Monitoring Plan

INPUT LOTS	SYSTEM UTILIZATION	SLACK CAPACITY	CYCLE TIME IN DAYS	WIP LOTS	ADDITIONAL SYSTEMS NEEDED
62	66.69%	33.31%	6.41	0.00	0.00
63	72.76%	27.24%	7.23	1.00	0.11
64	72.88%	27.12%	7.32	3.00	0.32
65	72.93%	27.07%	7.38	3.00	0.32
66	***77.28%***	***22.72%***	***7.45***	***3.00***	***0.32***
67	77.37%	22.63%	7.72	4.00	0.42
68	77.96%	22.04%	8.02	4.00	0.42
69	78.76%	21.24%	8.13	4.00	0.42
70	80.38%	19.62%	8.17	5.00	0.53
71	80.38%	19.62%	8.32	6.00	0.63
72	82.20%	17.80%	8.39	7.00	0.74
73	83.08%	16.92%	8.46	7.00	0.74
74	85.49%	14.51%	8.47	7.00	0.74
75	85.51%	14.49%	8.83	8.00	0.84
76	87.77%	12.23%	8.85	8.00	0.84
77	89.36%	10.64%	9.00	10.00	1.05
78	90.39%	9.61%	9.18	10.00	1.05
79	91.15%	8.85%	9.21	16.00	1.68
80	93.65%	6.35%	9.39	19.00	2.00
81	96.19%	3.81%	9.93	21.00	2.21
82	97.85%	2.15%	10.24	25.00	2.62

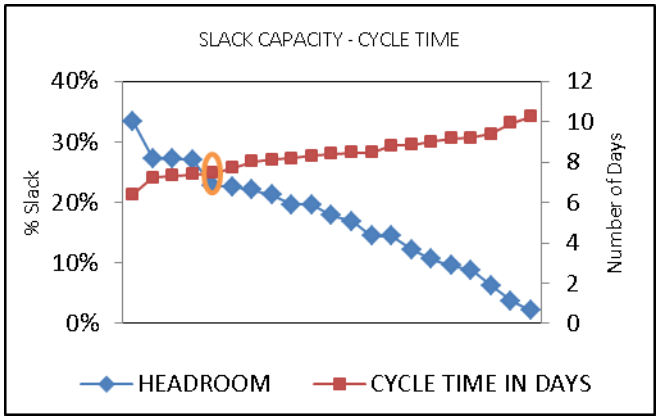

Figure 14. Slack Capacity versus Cycle Time

Based on Table 1, the optimal headroom level is at 22.72% where input loaded is at 66 lots. Figure 14 clearly shows that the chosen value is at the equilibrium point. The greater slack capacity is the lesser Cycle Time is; this shows an inverse relationship.

4. Conclusion

Results show that increasing volume (lot) arrival rate led to an increase in System Utilization; decreasing the number of systems and system downtimes (DT) will as well increase Utilization. Furthermore, as System Utilization increases, both CT and WIP Level increase.

It is concluded that the company must maintain at least 22.72% slack capacity. Otherwise, the company may experience major leap on the performance measures. It can also be concluded from the simulation that an additional investment on equipment would lead to a decrease in CT.

Moreover, the optimal headroom model yielded to the formulation of an equipment monitoring plan that will serve as the basis for capital acquisition. The outcome of this study will benefit different semiconductor companies for it will provide them with a design component and a simulation model which can be used in monitoring their performance in terms of machine utilization, CT and WIP level.

5. Recommendations

More experiments could be done to show how CT can be kept ideal. The scope of the project could also be expanded to perform more detailed analyses by including additional factors such as lot release/dispatch pattern and batch sizing. Furthermore, a study may be done to analyze the issues in complex semiconductor manufacturing processes. One issue that can be considered is the analysis of buffer lots in front of each tester so as to decrease the overall value of CT. It can apply a just-in-time system, as testing is considered a bottleneck operation for semiconductor companies.

If the company desires to see the cost significance of the table, they could specify the unit cost, the percentage tradeoff for every day the delivery is late, the holding cost of WIP, the value for each lot not delivered and the equipment cost. Substituting the corresponding values to the generated results from the headroom model could present a figural analysis of the project.

Acknowledgements

The authors of this paper would like to acknowledge the support of the Mr. Rex Aurelius Robielos, Dean of the School of IE-EMG and Abigail Ann Tagle, a graduate of MIT for their valuable comments and suggestions.

References

Beng Hui, Dennis. “An Analysis on the Dynamics of Work in Process(WIP) in a High Volume Manufacturing System of a Semiconductor Company using System Dynamics”. Department of Industrial Engineering, College of Engineering. De La Salle University-Manila. 2009.

Brown, Fowler, Gold, and Schoemig, Alexander. “Measurable Improvements in Cycle-Time-Constrained Capacity” IEEE. 1997.

Chance, Fowler, and Robinson, Jennifer. “Supporting Manufacturing with Simulation: Model Design, Development and Deployment”. IEEE. 1996.

Choi, Houshyar, and Kumar, Anil. “A Simulation Study of an Automotive Foundry Plant Manufacturing Engine Blocks”. pp 1035-1040.

Biography

Herwina Richelle C. Andres graduated with a Bachelor of Science in Industrial Engineering as Cum Laude and Gold Medalist at the Mapua Institute of Technology (MIT). She is currently taking her Master of Engineering major in Industrial Engineering specializing in Production Systems at the University of the Philippines. Herwina Richelle is currently with Analog Devices Inc. as an Operations Research Engineer.