*Proceedings of the 2003 Winter Simulation Conference*
*S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.*

# INDIRECT ESTIMATION OF CYCLE TIME QUANTILES FROM DISCRETE EVENT SIMULATION MODELS USING THE CORNISH-FISHER EXPANSION

Jennifer E. McNeill
Gerald T. Mackulak
John W. Fowler

Industrial Engineering Dept.
Arizona State University
PO Box 875906
Tempe AZ 85287-5906, U.S.A.

## ABSTRACT

This paper introduces a new technique for estimating cycle time quantiles from discrete event simulation models run at a single traffic intensity. The Cornish-Fisher expansion is used as a vehicle for this approximation, and it is shown that for an M/M/1 system and a full factory simulation model, the technique provides accurate results with low variability for the most commonly estimated quantiles without requiring unreasonable sample sizes. Additionally, the technique provides the advantages of being easy to implement and providing multiple cycle time quantiles from a single set of simulation runs.

## 1 INTRODUCTION

On time delivery is a key metric for assessing the customer service level of a production facility, and the ability to generate accurate delivery dates is crucial, particularly in customer service driven industries (Gordon 1993). One technique for improving this metric would be to develop more accurate estimates of average cycle time and, moreover, cycle time quantiles for a given traffic intensity. Estimates of the average cycle time give a feel for the expected value of the cycle time, but do not take into account the variability in the system or the skewness of the cycle time distribution. An intelligently selected set of quantiles from the distribution, on the other hand, provides the decision maker with a complete picture of the cycle time distribution (Chen and Kelton 1999). Therefore, accurate quantile estimates provide much more information from which to quote customer lead times.

Discrete event simulation models have traditionally been used to generate estimates of average cycle time, and much work has been done on reducing the simulation run time for large scale production systems in an attempt to obtain these estimates more quickly. However, even with decreasing run times, there are still not efficient and easily implemented methods for obtaining accurate estimates of cycle time quantiles. Much of the reason for this is that quantiles are more difficult to compute than simple averages and can require excessive data storage.

A direct quantile estimate is one in which the estimate is a function of the data itself. Order statistics are most traditionally used for this purpose. To obtain a cycle time quantile estimate from a discrete event simulation model using order statistics, the cycle time values are simply collected and ordered from low to high. The desired quantile is then directly selected from the sorted data. For example, to estimate the $95^{th}$ quantile of cycle time from 100,000 observations, select the $95{,}000^{th}$ largest data point. Clearly, a drawback of this solution technique is that all the observations must be stored and then sorted to obtain the estimate, and even with rapidly increasing computing power, sorting and storing the hundreds of millions of samples required to estimate some quantiles is still unreasonable (Chen and Kelton 1999).

Jain and Chlamtac (1985) developed the $P^2$ algorithm to estimate quantiles without storing the individual observations, but the algorithm is cumbersome to implement. Heidelberger and Lewis (1984) also developed techniques for reducing data storage in quantile estimation, but their algorithm requires that the simulation model be rerun for each quantile estimated. Jin , Fu, and Xiong (2003) suggest a new quantile estimator and show that the error probability for this estimator goes to zero with a large enough sample size, but the sample size required for the estimation grows exponentially as the problem dimension increases. Also, Chen and Kelton (1999) developed the zoom-in algorithm for quantile estimation, which uses the concepts of order statistics and $\phi$-mixing to provide upper and lower bounds on the desired quantile at each iteration of the algorithm. However, their approach also requires sample sizes on the order of tens of millions.

An indirect quantile estimate is one in which the estimate is a function of data parameters (i.e. sample mean, sample variance, etc.) rather than the data itself. Indirect estimation techniques have the advantage of not requiring as much data storage, but may be less accurate or have higher variance than the direct estimation techniques. Avramidis and Wilson (1998) suggest a technique for reducing bias and variance of quantile estimators using correlation induction techniques. Hesterberg and Nelson (1998) also exploited control variates with known quantiles to reduce the variance in estimating selected quantiles of a distribution. The technique showed significant reduction in MSE for extreme quantiles (.9, .95, and .99), but requires the use of control variates with known quantiles.

An indirect quantile estimation technique that provides good accuracy, low variability, and which is easy to implement would be extremely useful. This paper introduces a technique for indirectly estimating cycle time quantiles from a discrete event simulation model run at a single traffic intensity. Results from models of a simple M/M/1 queueing system and from a full scale semiconductor manufacturing system are presented.

## 2 CORNISH-FISHER EXPANSION

The Cornish-Fisher expansion, developed by Cornish and Fisher (1937), relates sample data to the standard normal distribution in order to generate a quantile estimate. Equation 1, given below, gives the first four terms of the Cornish-Fisher expansion, where y* is the normalized quantile estimate, $g_1$ is the sample skewness, $g_2$ is the sample kurtosis, and $z_\alpha$ is a quantile draw from the standard normal distribution.

$$y^*=z_\alpha+1/6(z_\alpha^2-1)g_1+1/24(z_\alpha^3-3z_\alpha)g_2-1/36(2z_\alpha^3-5z_\alpha)g_1^2 \quad (1)$$

The expansion begins with the desired quantile draw from the standard normal distribution and then makes adjustments to this quantile based on the sample skewness and kurtosis from the desired distribution. Therefore, it is important to have accurate, consistent estimators of the sample skewness and kurtosis. Equations (2) – (4), shown below, give estimators for the first four cumulants of a sample distribution (Kenney and Keeping 1954). These cumulants, in turn, are used with equations (5)-(6) to obtain estimates of sample skewness and kurtosis (Kenney 1954). In equations (2)-(6), N represents the number of sample data points, and $m_r=1/N\Sigma(x_i-m_1)^r$.

$$k_2=Nm_2/(N-1) \quad (2)$$

$$k_3=N^2m_3/(N-1)(N-2) \quad (3)$$

$$k_4=N^2[(N+1)m_4-3(N-1)m_2^2]/(N-1)(N-2)(N-3) \quad (4)$$

$$g_1=k_3/k_2^{3/2} \quad (5)$$

$$g_2=k_4/k_2^2 \quad (6)$$

To obtain a quantile estimate of the cycle time distribution from a discrete even simulation model using the Cornish-Fisher expansion, running totals of the appropriate sums of squares, sums of cubes, etc. must be kept in order to calculate the $m_r$ values. Then, at the conclusion of each simulation run, equations (2)-(6) are used to obtain estimates of the sample moments, which are, in turn, plugged into equation (1) to obtain a normalized quantile estimate. To translate the quantile estimate back to the cycle time distribution from the normalized distribution, the quantile is simply multiplied by the sample mean and divided by the sample standard deviation Only minimal data storage, necessary for sample moment calculation, is required, and upon completion of a single set simulation runs, any quantile can be calculated without further simulation effort using Equation (1).

## 3 RESULTS

To assess the performance of the Cornish-Fisher expansion as an indirect quantile estimator for discrete event simulation models, its performance was compared to the direct estimation technique using order statistics. Quantiles were estimated for a simple M/M/1 system, in which the theoretical quantiles are known, and for a full scale model of a semiconductor manufacturing facility in which the theoretical quantiles are not known. Experimentation for the M/M/1 system was performed using a discrete event simulator written in C++, while experimentation for the full factory model was performed using the commercial simulation package Factory Explorer.

### 3.1 M/M/1 System

Figure 1 shows the results of estimating cycle time quantiles for an M/M/1 system using direct quantile estimation (order statistics). Each point on the line represents a separate set of 30 simulation runs. For each of the five lower traffic intensities (.5, .6, .7, .8, and .9), 1,000,000 cycle time observations were recorded at each simulation run, while at the highest traffic intensity, .97, 2,000,000 cycle time observations were recorded at each run. Additional observations were collected at the highest traffic intensity point since the variability in the system increases dramatically there, compared to the lower traffic intensity points, and, therefore, extra observations are required to drown out the initial bias induced by the empty and idle starting conditions.

The solid line in Figure 1 shows the mean relative percent deviation from the theoretical value across each of the design points, while the dotted lines represent the 95%

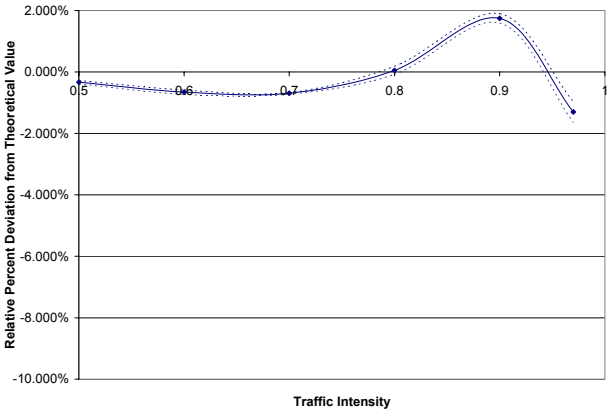


Figure 1: 95% Confidence Interval on the Relative Percent Difference from the Theoretical 90th Quantile of Cycle Time for an M/M/1 System Obtained Using Order Statistics

confidence interval on the same relative percent deviation. To build the confidence interval for a given traffic intensity, the quantile estimates for each of the simulation runs were calculated. Then, the estimates were grouped into sets of five, in order of simulation run, and the average of each group was calculated. By the central limit theorem, those averages are approximately normally distributed, and the standard equation for building a confidence interval around a mean was employed. As illustrated in Figure 1, the confidence interval around the quantile estimate is very tight, even in regions of high variability using the direct estimation technique.

Using the direct estimation as a benchmark, quantile estimates made using the Cornish-Fisher expansion, as shown in equation (1), were then calculated. Equation (1) gives only the first four terms of the expansion, as additional terms of the expansion were not found to add significantly to the results. In fact, using the first 6 terms of the expansion resulted in poorer quantile estimation. Figure 2 shows the results of this experimentation for an M/M/1 system. To ensure a fair comparison between the direct and indirect estimation techniques, common random numbers were employed, and the same number of simulation runs and observations were collected at each traffic intensity.

As expected, in Figure 2 the confidence interval is wider than the confidence interval obtained using order statistics, and it gets wider as the traffic intensity, and therefore variability in the system, gets higher. However, it is still relatively narrow, never getting wider than 2% of the theoretical cycle time value at any traffic intensity.

Figure 4 illustrates the effectiveness of the Cornish-Fisher expansion in estimating different cycle time quantiles for the M/M/1 system. As with the previous figures, experimentation consisted of 30 runs of 1,000,000 observations each at all traffic intensities except .97, at which

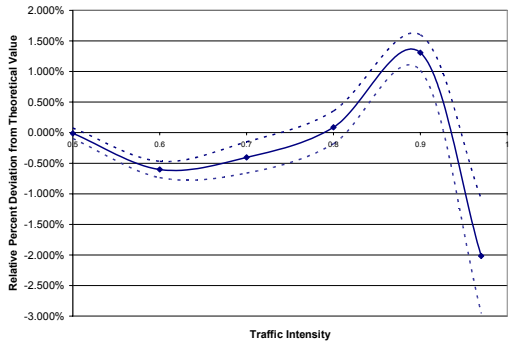


Figure 2: 95% Confidence Interval on the Relative Percent Difference from the Theoretical 90th Quantile of Cycle Time for an M/M/1 System Obtained Using the Cornish-Fisher Expansion

2,000,000 observations were collected at each run. The figure illustrates that, for this system, the Cornish-Fisher expansion does a better job estimating some quantiles than others. For instance, the 30th quantile tends to be underestimated by at least 5% and by as much as 12% at the higher traffic intensities, while the 70th and 90th quantile are estimated within 3% of the theoretical value at each of the simulated traffic intensities. In general, Figure 3 demonstrates that as the quantile being estimated gets lower and lower, the accuracy of the estimation technique for an M/M/1 system tends to get poorer and poorer. However, the likelihood of a production manager estimating the 30th quantile of cycle time is very small. Rather, it is more likely the most desirable quantiles to estimate will be at the other extreme, the 70th quantile or higher.

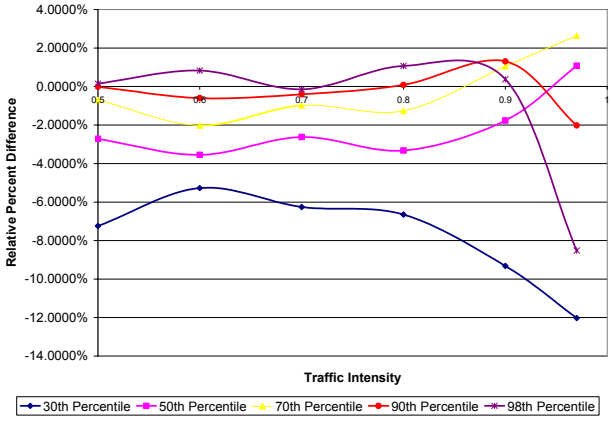


Figure 3: Relative Percent Difference Between Estimated and Theoretical Quantiles for an M/M/1 System

Despite the fact that Figures 1 and 2 illustrate that using order statistics provides slightly more accurate quantiles with slightly less variability for the M/M/1 system, Figure 3 highlights an advantage of using the indirect estimation technique. To generate cycle time quantile estimates using the

Cornish-Fisher expansion, only one set of simulation runs was required to obtain all the quantile estimates in Figure 3. Once the sample moments at each of the traffic intensities were calculated from the simulation runs, the $z_\alpha$ value in equation (1) was simply changed to produce estimates for all the quantiles shown in the figure. Using direct estimation, additional computing work would have been required to generate the same information. For each simulation run, for each quantile, the correct element of the sorted cycle time array would have to be identified. Table 1 illustrates these differences between direct estimation and indirect estimation using the Cornish-Fisher expansion.

Table 1: Differences in Requirements between Direct Estimation Using Order Statistics and Indirect Estimation Using the Cornish-Fisher Expansion

| | Direct Estimation | Indirect Estimation |
|---|---|---|
| **Data Storage** | Cycle time observations (possibly) | None |
| **Purpose of Simulation** | Collect sample of cycle time observations | Collect sample moments of cycle time |
| **Post- processing?** | Sort cycle time values and collect desired percentile | Evaluate Corning-Fisher approximation |
| **Want new percentile?** | Rerun post-processing for each simulation run at each traffic intensity | Change $z_\alpha$ parameter and recalculate the Corning-Fisher expansion |

### 3.2 Full Factory System

In addition to testing the Cornish-Fisher expansion on a simple M/M/1 queueing system, experimentation was also performed on a more complex, full factory model. The model was adapted from Testbed Data Set #1, obtained from the website of the Modeling and Analysis of Semiconductor Manufacturing lab at Arizona State University (http://www.eas.asu.edu/~masmlab), and experimentation was performed using the commercial simulation package, Factory Explorer. Details about the system are given in table 2.

Table 2: Description of Full Factory Model

| | |
|---|---|
| Product type | Non-volatile memory |
| Number of products | 1 |
| Number of processing steps | 232 |
| Number of tool groups | 83 |
| Number of operator groups | 32 |
| Rework modeled? | Yes |
| Machine breakdown modeled? | Yes |
| Machine loading/unloading? | Yes |

Figure 4 illustrates the results of using the Cornish-Fisher expansion to estimate cycle time quantiles for the this system, so the estimates could not be compared to the

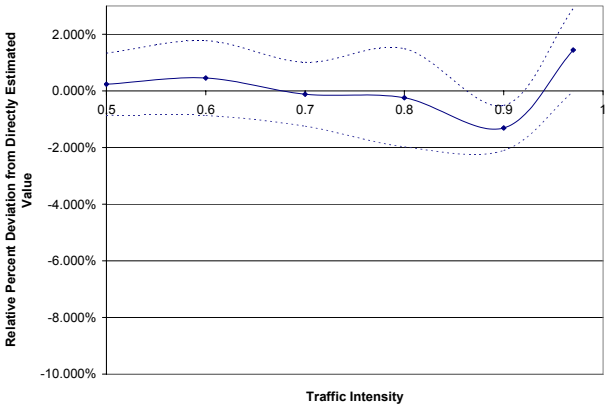


Figure 4: 95% Confident Interval on the Relative Percent Difference between the Indirectly and Directly Estimated 90th Cycle Time Quantile for the Full Factory Model

true quantile values as they were during the experimentation on the M/M/1 system. Instead, direct estimates of the quantiles at each traffic intensity were obtained using order statistics and considered to be best available quantile estimates. The indirect estimates obtained via the Cornish-Fisher expansion were then compared to these direct estimates to assess their accuracy. Figure 3 shows the 95% confidence interval on the relative percent difference between the directly and indirectly estimated 90th cycle time quantiles for the full factory model. To obtain both the direct and indirect estimates for all traffic intensities other than .97, the simulation was run for 5 years. As was done with the M/M/1 system, at the .97 traffic intensity the simulation was run for twice as long, or 10 years. Fifteen replications were made at each traffic intensity.

Figure 4 shows that quantiles obtained using indirect estimation are almost as good as those obtained using direct estimation. In fact, across all traffic intensities, the indirect estimate never varies by more than 2% from the direct estimate, and the width of the 95% confidence interval is never greater than 3% of the directly estimated quantile value.

As was done with the M/M/1 system, the accuracy of the Cornish-Fisher expansion in estimating different cycle time quantiles was evaluated for the full factory system. Figure 5 shows the relative percent difference between the directly and indirectly estimated values for the 30th, 50th, 70th, 90th, and 97th quantiles of the sample cycle time distribution As with the previous results, each data point is the result of 15 simulation replications of 5 or 10 years each, depending on the traffic intensity.

For the lower traffic intensities (.5, .6, .7, and .8), the Cornish-Fisher expansion estimates for all five cycle time quantiles are within 1% of the values obtained using direct estimation. At the higher traffic intensities (.9 and .97), the indirect estimates begin to vary more from their directly estimated counterparts, but they are never more than 5%

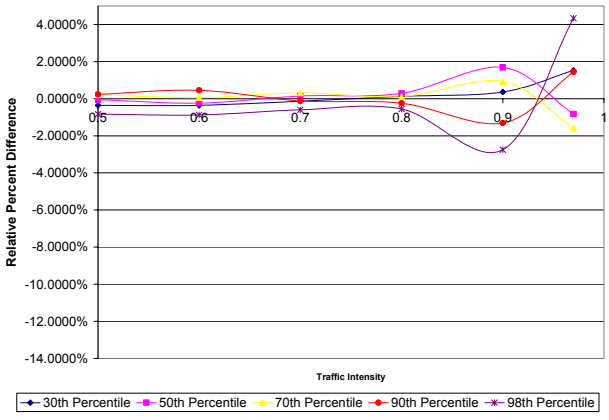


Figure 5: Relative Percent Difference Between Indirectly and Directly Estimated Cycle Time Quantiles for the Full Factory Model

different than the direct estimates. Additionally, as opposed to the results from the M/M/1 system, there does not appear to be any indication for the full factory that the Cornish-Fisher expansion predicts certain quantiles better than others. All five of the estimated quantiles are predicted with approximately the same level of accuracy when compared to the directly estimated values.

Finally, Figure 5 again illustrates the fact that indirect quantile estimation using the Cornish-Fisher expansion can save significant effort over direct estimation, especially when multiple quantile estimates are required. To generate the data points for the figure, both direct and indirect estimates for each quantile at each traffic intensity were required. For the indirect estimates, once the sample moments at each traffic intensity were known, it was trivial to generate the estimates of the different quantiles using the Cornish-Fisher expansion. Therefore, from a single set of computing runs, all five sets of quantile estimates were generated. However, to obtain the same set of information for the direct estimates, a short computer program that identified the correct element from the sorted array of cycle time values had to be run at each traffic intensity for each quantile estimate.

## 4 CONCLUSIONS

The results from the M/M/1 system and the factory model illustrate that using the Cornish-Fisher expansion in conjunction with discrete event simulation models can lead to accurate estimates of the most desirable quantiles for manufacturing systems. For the M/M/1 system, accuracy decreases as the quantile being estimated tends toward the lower extreme of the distribution, but these quantiles are less likely to be useful in a manufacturing setting. In the full factory model, the accuracy, when compared to directly estimated quantiles, did not depend at all on the quantile being estimated. Additionally, confidence intervals around the estimates show that their variance is quite small, provided that the initial bias in the simulation model is no longer affecting the sample moment calculations. Also, despite the fact that direct estimation using order statistics provides slightly more accurate quantile estimates with less variability, the gains in accuracy are offset by the ease of implementation and data storage requirements. The Cornish-Fisher expansion provides a quantile estimation technique which is easy to implement and has the advantage of being able to generate multiple quantile estimates from a single set of simulation runs. Additionally, using the Cornish-Fisher expansion to estimate quantiles requires minimal data storage, resulting in an extremely compact representation of the system

## 5 FUTURE WORK

Future work in this area includes generating quantile estimates for the entire cycle time throughput curve rather than for a single point on the curve at a time. Further testing could also be performed on systems in which the theoretical quantiles are known (i.e. Jackson network queueing models). Additionally, investigations into the required sample size to obtain a given confidence interval width could prove to be interesting.

## ACKNOWLEDGMENTS

This research has been supported in part by grant DMII 0140441/0140385 from the National Science Foundation and by grant 2001-NJ-878 from the Factory Operations Research Center that is jointly funded by the Semiconductor Research Corporation and by International SEMATECH. Additional thanks go to professors Barry Nelson and Bruce Ankenman from Northwestern University for their creative and technical insights during this research.

## REFERENCES

Avramidis, Athanassios N., and James R. Wilson. 1998. Correlation-Induction Techniques for Estimating Quantiles in Simulation Experiments. *Operations Research* 46 (4): 574–591.

Chen, E. Jack, and W. David Kelton. 1999. Simulation-Based Estimation of Quantiles. *Proceedings of the 1999 Winter Simulation Conference, P.A. Farrington, H.B. Nembhard, D.T .Sturrock, and G.W. Evans, eds.* 428-434.

Cornish, E.A. and R.A. Fisher. 1937. Moments and Cumulants in the Specification of Distributions. *Revue de l'Institut International de Statistique,* 5: 307-320

Gordon, V.S. A Note on Optimal Assignment of Slack Due Dates in Single Machine Scheduling. *European Journal of Operational Research* 70: 311-315.

Hesterberg, Timothy C., and Barry L. Nelson. 1998. Control Variates for Probability and Quantile Estimation. *Management Science* 44(9): 1295-1312.

Heidelberger, P., and P.A.W. Lewis. Quantile Estimation in Dependent Sequences. *Operations Research* 32(1): 185-209.

Jain, Raj, and Imrich Chlamtac. 1985. The $P^2$ Algorithm for Dynamic Calculation of Quantiles and Histograms without Storing Observations. *Communications of the ACM* 28(10): 1076-1085.

Jin, Xing, Michael C. Fu, and Xiaoping Xiong. 2003. Probabilistic Error Bounds for Simulation Quantile Estimators. *Management Science* 14(2): 230-246.

Kenney, J.F., and E.S. Keeping. 1954. *Mathematics of Statistics, Part 1*. D. Van Nostrand Company, Inc. Princeton, New Jersey.

## AUTHOR BIOGRAPHIES

**JENNIFER E. MCNEILL** is a Ph.D. student in the Industrial Engineering department at Arizona State University. Her research interests are in discrete event simulation methodologies and manufacturing applications. Prior to beginning her PhD studies, she served as an intern in the Operational Decision Support Technologies group at Intel, and was recently awarded the SRC/Intel Fellowship.

**GERALD T. MACKULAK** is an Associate Professor of Engineering in the Department of Industrial Engineering at Arizona State University. He is a graduate of Purdue University receiving his B.Sc., M.Sc., and Ph.D. degrees in the area of Industrial Engineering. His primary area of research is simulation applications within manufacturing with a special focus on semiconductor manufacturing.

**JOHN W. FOWLER** received the Ph.D. degree in Industrial Engineering from Texas A&M University. He is a Professor of Industrial Engineering at Arizona State University (ASU) and is the Center Director for the Factory Operations Research Center that is jointly funded by International SEMATECH and the Semiconductor Research Corporation. Prior to his current position, he was a Senior Member of Technical Staff in the Modeling, CAD, and Statistical Methods Division of SEMATECH and an Adjunct Assistant Professor in the Graduate Program in Operations Research of the Mechanical Engineering Department at the University of Texas at Austin. He spent the last year and a half of his doctoral studies as an Intern at Advanced Micro Devices. His research interests include modeling, analysis, and control of manufacturing (especially semiconductor) systems. He is the Co-Director of the Modeling and Analysis of Semiconductor Manufacturing Laboratory at ASU. The lab currently has had research contracts with NSF, SRC, International SEMATECH, Intel, Motorola, Infineon Technologies, ST Microelectronics, and Tefen, Ltd. Dr. Fowler is a member of ASEE, IIE, INFORMS, POMS, and SCS. He is an Area Editor for SIMULATION: Transactions of the Society for Modeling and Simulation International and an Associate Editor of IEEE Transactions on Electronics Packaging Manufacturing.