Monitoring and Diagnostic Techniques for Control of Overlay in Steppers

Gary E. Flores, Warren W. Flack, Susan Avlakeotes
Ultratech Stepper, Inc.
San Jose, CA 95134

Mark Merrill
KLA Instruments Corporation
San Jose, CA 95161

1.0 Abstract

Semiconductor lithography manufacturing presents a major challenge for the application of classical Statistical Process Control (SPC) methodologies due to the complex nature of this process. For example, difficulties can occur due to inadequate data sampling, nonnormal error distributions, equipment or process instability and nonstationary random errors. Incorrect use of classical SPC techniques can result in the incorrect interpretation of process stability which can have a drastic impact on productivity. Photolithography provides additional SPC challenges due to the inherent multivariable nature of the output variables that are being controlled. This paper examines appropriate SPC and monitoring techniques for stepper control of overlay performance using in-process measurement and analysis equipment to address these issues.

The average run length of three charting techniques is compared to quantify the ability of each technique to detect process mean shifts. Shewart, Exponentially Weighted Moving-Average (EWMA) and Cumulative-Sum (CUSUM) charts are analyzed for a baseline process and mean shifts of 0.42, 0.85 and 1.25 standard deviations. These results illustrate the superior performance of a CUSUM chart over Shewart and EWMA charts. In addition, the Shewart chart with Western Electric rules produced false mean shift alarms for the baseline case. The EWMA is also observed to be sensitive to the selection of weighting factors. The effectiveness of plotting individual wafers is compared with plotting lot means. The plotting of individual wafers outperforms lot means in the determination of baseline shifts because of the larger population size of the individual charts.
2.0 INTRODUCTION

As the semiconductor industry has matured there has been a dramatic increase in the amount of data collected to address manufacturing issues [1]. In the area of photolithography, these data are supplied by in-process systems as well as independent post-process measurement systems. Overlay and critical dimensions (CD) are the predominant lithography processes measured. Proper characterization and analysis of these factors are essential for continuously improving product quality, productivity and device yield.

A major goal in lithography processes is the continuous improvement of overlay control. Matching overlay between stepper manufacturers’ and different generations of steppers can be a significant challenge [2]. Automated overlay metrology systems provide an opportunity to collect extensive data sets on lithography tools and processes. However, as a result of this large amount of data there is a need to improve and evolve monitoring and diagnostic techniques to maximize the data value. This requires the appropriate use of statistical analysis and statistical process control (SPC) techniques.

Figure 1 illustrates the typical relationship of process, stepper, metrology, analysis and SPC for a lithography process. The wafers at the stepper are processed and a subset is sampled using in-process measurement systems for monitoring overlay and critical dimensions. Next, the measurement data is analyzed using SPC techniques with appropriate overlay models and diagnostic routines. From this analysis, essential information on the stepper and lithography processes are derived. Typical information consists of possible process alarms, optimal stepper settings of focus, exposure and overlay parameters. The output of the analysis can be used to control the stepper and processes using standard control techniques.

Classical SPC principles such as Shewart Control Charts used in conjunction with Western Electric rules [3] are applied for optimal process control in many manufacturing environments. However, the process complexity of lithography manufacturing presents a major challenge for the application these techniques. For example, the lithographic alignment process has been observed to exhibit systematic errors rather than random errors [4]. An implicit assumption of SPC principles is that the processes and tools being monitored are Identically, Independent and Normally Distributed (IIDN) [5]. This implies that monitored processes and tools are from a normal data set and all measurements are uncorrelated over time.

Violation of the IIDN requirements can occur during overlay monitoring for a variety of reasons. For example, insufficient overlay data sampling may result for infrequently processed lithography levels. Insufficient data can be exhibited as a nonnormal overlay error distribution. A second scenario is a nonnormal distribution for processes and equipment that are unstable due to premature implementation in manufacturing. A third possibility is that individual measurements are not statistically independent and thus are correlated over time. An example would be a process or piece of equipment that has an inherent cyclic nature. In this case, sampled overlay
measurements would follow a nonrandom pattern. Incorrect use of classical SPC principles in these examples would prevent a proper determination of process stability.

Another challenge for applying SPC occurs due to the inherent multivariable nature of lithography processes. For the case of overlay control, there are typically six grid variables related to wafer staging errors and three or more variables pertaining to intrafield effects [2]. For optimal control, it is necessary to simultaneously monitor and control all of these variables. A potential problem is the high probability that cross-correlations will occur among the outputs being monitored and controlled [5]. Such a situation can lead to false alarms using single univariate control charts for each output [6]. Single output univariate charts also have the obvious difficulty of management and analysis of large quantities of charts. The Hotellings $T^2$ statistic provides a combined score that properly accounts for cross-correlation of multiple output variables. This statistic can be used to construct a single multivariate control chart that minimizes the occurrence of false process alarms as well as identifying process changes not detected using univariate charts [7, 8].

This paper examines appropriate SPC techniques for stepper control of overlay performance using in-process measurement and analysis equipment. The average run length of three charting techniques is compared to quantify the ability to detect process mean shifts. Shewart, Exponentially Weighted Moving-Average (EWMA) and Cumulative-Sum (CUSUM) charts are analyzed for a baseline condition in addition to three different mean shifts. The effectiveness of plotting individual wafers is also compared with plotting lot means.

### 3.0 Monitoring and Charting Techniques

There are a number of different types of charting techniques which are available in microlithography to monitor process factors. Each charting technique has certain advantages and disadvantages. The performance of a control chart is typically specified by the average number of readings that have been made before generating an alarm signal that a previously specified shift has occurred in the process mean. This is called the Average Run Length ($\text{ARL}_x$) where $x$ is the size of the shift in standard deviations. The most effective charting technique will have a short $\text{ARL}_x$. In addition, it is important to minimize false alarm signals even when the process is properly centered. A false alarm signal is specified by an $\text{ARL}_0$ with a shift of zero standard deviations. Since data collection is sparse in microlithography, it is typically more important to have a small $\text{ARL}_x$ than a large $\text{ARL}_0$.

The most commonly used control chart is the Shewart chart [3]. Here the value of a selected quality characteristic (data point) is simply plotted in time sequence and compared with standard three sigma control limits spaced above and below the process center line. A value above or below the control lines is an alarm signal for corrective action. Shewart charts are popular because they are the easiest charting technique to teach and interpret. This technique was originally
designed for high volume manufacturing environments where false alarms need to be minimized. For three sigma control limits the ARL$_0$ of a Shewart chart is 370. However, the technique has a poor sensitivity to detecting mean shifts. For example, a one standard deviation shift in the mean (ARL$_1$) can be shown to be 42 [6].

The ARL$_x$ of the Shewart chart can be reduced by adding rule testing for special alarm causes. The most common special tests are the Western Electric rules [3]. Adding special tests has a number of disadvantages. First, each organization creates rules or utilizes rule subsets different than the standard Western Electric rules. Second, rule testing increases the frequency of false alarms, or reduces the ARL$_0$. Finally, a rule based system dramatically increases the complexity of training and interpretation of the control charts which defeats the primary advantage of Shewart charts. Regardless, Shewart charts with rule testing remain the most common charting technique in use today.

Charting techniques based on moving averages can offer dramatic improvements over classical techniques such as the Shewart Chart. Using these techniques, the value of the current data point is combined with past measurements to produce a method better able to detect small shifts in the process average. Note that it is difficult to use rule based systems because the moving averages can be highly correlated [9]. An Exponentially Weighted Moving Average (EWMA) chart uses a weighting factor $r$ ($0 < r \leq 1$) which decreases measurement weighting exponentially going backwards in time. The EWMA response for the control chart is calculated using the following equation [10]:

$$\text{EWMA}_t = \sum_{i=1}^{t} r (1 - r)^{t-1} \bar{x}_i$$

where $\bar{x}_i$ is the subgroup or individual measurement mean and $r$ is the weight assigned to the most recent measurement. A small value of $r$ guards against a small shift in the mean. The EWMA chart is used extensively in time series modeling and forecasting [6]. Since it is based on a moving average of the data, it is insensitive to the requirement of normality and hence is a good technique for individual based observations.

A Cumulative-Sum (CUSUM) chart displays the cumulative sums of the difference between the current data and a target value. The cumulative sum response ($S_t$) for a control chart is calculated using the following equation [10]:

$$S_t = \sum_{i=1}^{t} \left( \frac{\bar{x}_i - \mu_0}{\sigma_x} \right)$$

where $\bar{x}_i$ is the subgroup or individual measurement mean, $\mu_0$ is the run average for a lot mean chart or wafer average for individual wafer chart, $\sigma_x$ is the pooled estimate of the standard deviation.
Both EWMA and CUSUM have a one sigma ARL$_1$ of around 10, which detects a mean shift about four times as fast as a Shewart chart [11]. Both charting techniques converge with the Shewart chart to produce a ARL$_0$ of 370. Unfortunately, these techniques are more difficult to teach and can not detect any shift, regardless of size, with fewer than two observations.

An alternative technique which is easy to teach and allows rapid detection of small mean shifts is the zone chart [12]. It is based on a Shewart chart where the one, two and three sigma zones from the process center line are plotted. When a data point falls in a zone, it is assigned a numerical score based on a predetermined weighting criteria. A typical weighting scheme would be 0 within the ±1 sigma zone, 1 in the 2 sigma bands, 2 in the 3 sigma bands and 4 outside of the ±3 sigma limits. This type of weighting is designated as (0,1,2,4). The numerical scores are summed sequentially in time and total is reset to zero whenever a new data point crosses the process center line. A process is declared out of control when the numerical score or Critical Run Sum (CRS) is greater than a predetermined value. For the (0,1,2,4) case, the CRS is typically 4. This particular system has a ARL$_0$ of 100 and a ARL$_1$ of 8.1 [9]. The frequency of false alarms is three times higher than either of the moving average charting techniques. However, the scoring chart system has the flexibility to easily alter the ARL$_0$ or ARL$_1$.

### 4.0 EXPERIMENTAL METHODS

An Ultratech Stepper model 1700 was used for the process monitoring alignment experiments in this study. The Ultratech Stepper is based on the 1x Wynne-Dyson lens design using broadband illumination which includes the $g$ and $h$ mercury lines as well as the continuum from 390 to 450 nm. All overlay experiments were based on a darkfield alignment system using a direct site-by-site mode [13]. The alignment targets for the stepper were located along the horizontal scribe across the top of the field. The target design is a cross with a width of 2.0 microns and a darkfield (valley) polarity. A single stepper was used in this study to eliminate any lens distortion and grid matching errors.

Bare silicon wafer substrates of 125mm diameter were used for the experimental work. Initially, a first level mask pattern was defined in a photoresist film. No intentional grid or interfield errors were applied at this level. A second level was aligned and exposed in the same photoresist. One group of wafers were run throughout the monitoring period without any process shift to establish a process and stepper baseline. Other groups of wafers had mean shifts artificially induced by programming grid offsets in the stepper. These shifts were applied to both the x and y axis of the stepper grid. This approach was intended to simulate the typical effect that might be encountered in a manufacturing environment. Each specific process shift was held constant for the remainder of the monitoring period to determine the ARL.

The photoresist-to-photoresist overlay pattern was then measured using a KLA 5700 Coherence Probe Microscope [14]. The overlay structures are a standard box-in-frame design. Previously,
characterization of the metrology system was performed to quantify the tool error. Multiple measurements were taken on the overlay targets during the tool setup to establish a three sigma tool variance of 7 nm. Additionally, the Tool Induced Shift (TIS) has been characterized and found to be less than 10 nm for layers up to 15 microns thick.

The wafer layout consisted of 19 fields as shown in Figure 2 with field dimensions of 34 mm in x and 16 mm in y. The five test fields selected for metrology are highlighted in grey. These sites were selected to reflect process variability over the entire wafer. In addition, five intrafield locations per field were selected at the center and four corners of the field as shown in the inset of Figure 2. Thus, the contributions of the field variability or lens related effects are included in the total variability for the measured overlay.

Each of these measurement levels acts as a source of process variation. In order to properly determine the standard deviation of the total process, it is necessary to use a nested normal model [7]. This involves determining estimates of process variation using pooled calculations of the standard deviation from all sources. This pooled standard deviation can then be used in the SPC charts. The sources of variation for this case include run-to-run or process variation ($\sigma_p^2$), variance between wafers in a run ($\sigma_w^2$), variance between fields on a wafer ($\sigma_d^2$) and variance between test site or intrafield locations in a field ($\sigma_t^2$). The variance of the average of all runs can then be calculated from the following relationship:

$$
\sigma_{avg}^2 = \frac{\sigma_p^2}{r} + \frac{\sigma_w^2}{rw} + \frac{\sigma_d^2}{rwd} + \frac{\sigma_t^2}{rwdt}
$$

where $t$ is the number of test sites per field, $d$ is the number of fields per wafer, $w$ is the number wafers per run and $r$ is the number of runs. For the metrology described in this paper, $t$ equals 5, $d$ equals 5 and $w$ equals 4.

With the exception of the intentionally induced process mean shifts, the stepper was allowed to run without any significant adjustments. This provided the ability to gauge the natural stepper variation during the controlled portions of the experiments. Sample wafers were then run on a periodic basis at the start of two shifts each day. A total of 4 wafers were aligned for each process group. Thus each lot or process group of wafers included a total 100 measurements per shift of operation.

5.0 RESULTS AND DISCUSSIONS

5.1 Detecting Process Mean Shifts

The average run length of three charting techniques were compared to quantify the capability to detect various size process mean shifts. The process shifts were induced by programmed grid
offsets in the mean x and y registration at a specified time. However, only x data will be presented
since the y data displayed similar type of behavior. Shifts of 0.42, 0.85 and 1.25 times the pooled
average standard deviation based on equation (3) were performed. Along with the process shifts, a
baseline monitor of the process was included for reference. Three separate analysis using
Shewart, EWMA and CUSUM Charts were made on the four controlled conditions. Note that the
zone chart was not adapted due to the lack of appropriate software tools.

The total population of overlay errors was found to have a standard deviation of 0.052 microns. In
terms of a nested normal model based on equation (3), this translated to 0.031 microns standard
deviation for run-to-run and 0.036 microns standard deviation for wafer-to-wafer. The process
mean shifts were selected to be 0.25, 0.50 and 0.75 times the standard deviation of the total errors.
This translates to 0.42, 0.85 and 1.25 standard deviations of the run-to-run standard deviation. For
purposes of this study, all the process mean shifts will be discussed in terms of run-to-run standard
deviation units.

The control charts can be calculated using either individual wafer means or lot subgroup means.
Using individual wafer charts the mean value of all 25 measurements on a wafer is plotted as a
single entry. For a lot mean chart the mean value for the subgroup of 4 wafers is plotted as a single
entry. In this case the mean lot value is based on 100 measurements.

Prior to applying the systematic mean shifts, the process and stepper baseline conditions were
monitored without any interventions to determine the natural variation and to validate the process
normality. Examination of the distribution of errors over the entire population illustrated a normal
distribution. Additionally, autocorrelation of the baseline condition was verified to eliminate the
possibility of time dependency. Therefore the overlay data meets the IIDN requirement for
analysis using SPC techniques.

Using the nested normal model described in the experimental methods section, the standard
deviations for the process (run-to-run), wafer-to-wafer, field-to-field, test site-to-test site were
 calculated. When using the individual wafer chart techniques, the control limits are based on the
pooled estimate of the standard deviation of the wafer-to-wafer variation. For the lot mean chart
techniques, the control limits are based on the pooled estimate of the standard deviation of the
run-to-run variation. In all cases, three sigma control limits were used for upper and lower control
limits.

5.2 Shewart and Western Electric Rules Analysis

The individual wafer control charts for the baseline and three different mean shifts are shown in
Figure 3. The Shewart charts include all relevant Western Electric rules for detecting mean shifts.
Western Electric alarm conditions are indicated on the charts by the numbers 1 through 6. The
baseline individual wafer control chart in Figure 3a shows several apparent false test alarms noted
by a 2 (nine points in a row within 1 standard deviation of the mean) for wafers 8 through 16. The
same test is flagged at wafers 64 through 72. Figures 3b, 3c and 3d illustrate the individual control
charts for the three induced mean shifts. Note that the process shift was induced at wafer 37 and maintained over the duration of the monitoring period. Process mean shifts were detected at wafer 83, wafer 54 and wafer 44 for 0.42, 0.85 and 1.25 sigma mean shifts respectively. As the mean shift increased the ARL decreased for the individual wafer chart.

The analogous lot mean control charts for the baseline and three mean shifts are shown in Figure 4. The upper and lower control limits were calculated using the nested normal model for the run standard deviation. The same Western Electric rules for process mean shifts were applied. In this case there are no false test alarms for the baseline process or the three mean shift cases. However the process mean shifts are flagged at the same point (wafer 71) for all three cases as shown in Figures 4b, 4c and 4d. In the lot mean case there is no reduction in detection time for the mean shift with increasing size. This occurs since special test 2 for a mean shift cannot be flagged until there are nine points in a row within 1 standard deviation of the mean. This requires a larger sample population for the lot mean case than the individual wafer case. Overall, the individual wafer chart outperforms the lot mean charts because of the population size. However, the individual charts are more susceptible to false alarms than the lot mean charts.

5.3 Cumulative Sum (CUSUM) Analysis

Cumulative Sum Control charts based on both wafer individual and lot mean samples were generated for the baseline and three shifts. The individual wafer control charts are shown in Figure 5. The interpretation of the CUSUM chart is based on the slope of the curve over time. When the process mean remains fixed and all errors are random in time, the slope of the curve will remain unchanged. However, any shift in the process mean results in a systematic change in the cumulative sum deviation and thus there will be a slope change. The baseline CUSUM chart in Figure 6a does illustrate some variation in the slope over time. However, application of a T-test for significance indicates that there are no detectable process shifts. For the three process mean shifts, the CUSUM charts in Figures 6b, 6c and 6d visually illustrate distinct slope changes in the CUSUM after 9 runs. Statistical interpretation of the CUSUM charts is done by applying a V shape mask on the chart to determine when the process mean has shifted by a specified value [11].

5.4 Exponentially Weighted Moving Average (EWMA) Analysis

As for the previous two analyses, EWMA Control charts based on both individual wafers and lot mean samples were generated for the baseline and three mean shifts. The individual wafers CUSUM charts are shown in Figure 6. An r value of 0.25 was selected because it has been demonstrated to be effective in guarding against moderate mean shifts [12]. Control limits were based on 3 times the average standard deviation from equation (3) were used. The appropriate selection of r is critical for effective application of this charting technique. At each point on the EWMA chart, the weighted average of all previous individual measurements is plotted using the above relationship. Hence the weights of individual measurements decrease exponentially going backward in time. Interpretation of these charts for mean shifts include individual points outside
the 3 sigma control limits as well as trend shifts in the EWMA profile. The EWMA for the baseline and 0.42 sigma mean shift show no points outside the 3 sigma control limits. For the 0.85 and 1.25 sigma there are process flags indicated at wafer 51 and wafer 39 respectively. As for the CUSUM charts, the EWMA profile provides a clear visual indicator of the process mean shifts.

5.5 Comparison of Charting Techniques
The ARL of the Shewart, CUSUM and EWMA can be compared by determining the number of runs required to detect the various process mean shifts. For the CUSUM chart, an analysis of the run length for each shift was determined by applying statistical testing using a V shape mask [11]. This involves specifying a desired mean shift for detection and using hypothesis testing to determine if the data has exceeded the desired shift, which we will designate as delta. The two delta values selected were 0.25 and 1.0 times the standard deviation. The selection of delta values will impact the $ARL_x$ and $ARL_0$ and are thus useful to examine. For the EWMA chart, the value of $r$ used along with the control limits have a strong impact on $ARL_x$ and $ARL_0$ [11]. Two conditions were evaluated for the EWMA chart, these are $r$ equal to 0.25 with 3 sigma control limits and $r$ equal to 0.05 with 2.75 sigma control limits. The smaller value of $r$ equal to 0.05 and 2.75 sigma control limits were selected since they have been noted as effective for detecting small mean shifts [6].

Figure 7 depicts the ARL in wafers based on individual wafer charts for the three mean shifts. Clearly, the CUSUM chart outperforms the other techniques for the mean shifts of 0.42 and 0.85 standard deviations. For the largest mean shift of 1.25 standard deviations, both the EWMA and CUSUM techniques are comparable. The EWMA technique proved to sensitive to the selection of $r$ and control limits for small mean shifts which suggests that the $r$ value needs optimization depending on the size of the shift. It is also evident the ARL for Shewart increases dramatically as the process mean shift decreased from 1.25 sigma to 0.85 sigma. Overall, the Shewart chart had the largest ARL which agrees with the theoretical predictions of $ARL_1$. A final observation is that the performance of the CUSUM chart shows an improvement in the ARL as delta, the specified process shift for detection increases.

Figure 8 depicts the ARL in number of wafers required to determine a process mean shift based on mean lot charts for all three mean shifts using Shewart, CUSUM, and EWMA. Again the CUSUM and EWMA charts outperform the Shewart techniques. Finally, comparing the individual wafer versus lot mean chart techniques, it is evident that individual wafer charts tend to outperform the lot mean charts. This is due to the larger population size of the individual charts.

6.0 CONCLUSIONS
The average run length of three charting techniques was compared for detecting various process mean shifts in a lithography process. The Shewart, Exponentially Weighted Moving Average
(EWMA) and cumulative sum (CUSUM) charts were analyzed for the baseline condition in addition to mean shifts of 0.42, 0.85 and 1.25 standard deviations. These results illustrated the superior performance of CUSUM and EWMA charts over Shewart charts with Western Electric Rules. In addition, the Shewart chart with Western Electric rules produced false mean shift alarms for the baseline case. The EWMA was also observed to be sensitive to the selection of weighting factors. The effectiveness of plotting individual wafers was compared with plotting lot means. The plotting of individual wafers outperforms lot means in the determination of baseline shifts because of the larger population size of the individual charts. These observed benefits of the CUSUM and EWMA charts are quite attractive for improving process monitoring and control of overlay in lithography. It is therefore important to consider using these techniques to supplement existing SPC methodologies. This can be achieved more readily now with the increased availability of process control software.

7.0 REFERENCES


Figure 1: Relationship of process, stepper, metrology, analysis and statistical process control.

Figure 2: Wafer layout and metrology sampling scheme for overlay monitoring. The enlarged field shows the intrafield sampling schemes.
Figure 3a and 3b: Shewart control charts with Western Electric rules of individual wafers for the baseline and 0.42σ mean shift.
Figure 3c and 3d: Shewart control charts with Western Electric rules of individual wafers for $0.85\sigma$ and $1.25\sigma$ mean shifts.
Figure 4a and 4b: Shewart control charts with Western Electric rules of lot means for the baseline and 0.42σ mean shift.
Figure 4b and 4c: Shewart control charts with Western Electric rules of lot means for $0.85\sigma$ and $1.25\sigma$ mean shifts.
Figure 5a and 5b: Cumulative sum control charts of individual wafers for the baseline and 0.42\(\sigma\) mean shift.
Figure 5c and 5d: Cumulative sum control charts of individual wafers for 0.85\(\sigma\) and 1.25\(\sigma\) mean shifts.
Figure 6a and 6b: Exponentially weighted moving average charts of individual wafers for the baseline and 0.42σ mean shift.
Figure 6c and 6d: Exponentially weighted moving average charts of individual wafers for $0.85\sigma$ and $1.25\sigma$ mean shifts.
Figure 7: Comparison of charting techniques for detecting process mean shifts based on individual charts. Mean shifts are defined in units of the total process standard deviation.

Figure 8: Comparison of charting techniques for detecting process mean shifts based on mean lot charts. Mean shifts are defined in units of the total process standard deviation.