SHEWMA: An End-of-Line SPC Scheme Using Wafer Acceptance Test Data

Chih-Min Fan, *Student Member, IEEE*, Ruey-Shan Guo, *Member, IEEE*, Shi-Chung Chang, *Member, IEEE*, and Chih-Shih Wei

***Abstract*—In this paper, an end-of-line quality control scheme based on wafer acceptance test (WAT) data is presented. Due to the multiple-stream and sequence-disorder effects typically present in the WAT data, an abnormal process shift caused by one machine at an in-line step may become vague for detection using end-of-line WAT data. A methodology for generating robust design parameters for the simultaneous application of Shewhart and EWMA control charts to WAT data is proposed. This SHEWMA scheme is implemented in a foundry environment and its detection and diagnosis-enhancing capabilities are validated using both numerical derivations and fab data. Results show that the SHEWMA scheme is superior to the current practices in detection speed. Its use is complementary to the existing in-line SPC for process integration.**

***Index Terms*—Exponentially weighted moving average (EWMA), multiple-stream, process integration, semiconductor manufacturing, sequence-disorder, statistical process control (SPC).**

I. Introduction

A. Motivation

DURING integrated circuit fabrication, various test structures are fabricated on a wafer to extract information on the process and device performance for yield management. Wafer acceptance test (WAT) data come from the electrical measurements of these test structures after completing the whole fabrication process. In current WAT practice, several sites located on the fixed locations of each wafer are selected, from which over 100 WAT parameters are measured. Statistical analysis and process diagnosis based on end-of-line WAT data provide an assessment of overall process performance and its impact on product yield.

Although quality control should be improved as early as possible, first in the design stage, followed by the manufacturing stage, quality control at the end-of-line stage still adds value for the following reasons [1], [2]:

- statistical stability at the in-line level does not guarantee stability of the entire IC fabrication process;
- in-line data are often not available owing to time-consuming or destructive data collection methods;
- product performance is usually strongly correlated with the WAT data;
- abundant WAT data are usually available.

Since WAT measurements reflect the overall results of the entire fabrication process, their statistical characteristics are usually complicated. Standard quality control techniques such as statistical process control (SPC) charts cannot be applied to WAT data without caution. For example, there are multiple variation components shown in WAT measurements: lot-to-lot, wafer-to-wafer, site-to-site, and residual variations, which are not adequately taken into account in the traditional Shewhart control chart. Furthermore, there are two complicating features of the WAT data generation process. First, individual lots of wafers may go through different streams of machines during their fabrication processes, which induce the machine-to-machine variation among lots, and violates the assumption of standard SPC practices that each lot is identically distributed. This is called the "multiple-stream" effect. Second, the cycle time from an in-line step to the WAT step varies among lots, which makes the WAT lot sequence not the same as the lot sequence in each in-line step. An abnormal trend such as a process shift occurring at one machine of an in-line step would be more difficult to detect at the end-of-line WAT step under the "sequence-disorder" effect.

B. Related Work

To the authors' best knowledge, end-of-line quality control activities in the current industry practices are summarized below:

1) "Brute Force" SPC Applications: This approach eliminates the sequence-disorder and multiple-stream effects by implementing data sequence trace-back and stratification before constructing control charts. It first traces back the WAT lot sequence at individual machines for each process step, then sorts and stratifies WAT data accordingly. Statistical inference techniques, such as an analysis of variance (ANOVA) or a SPC control chart, are then performed to detect potential faults. For example, AMD [3] has used the powerful data processing capability provided by advanced computers to screen and analyze all the production and process data. Although useful in the above example, the "brute force" approach may not be feasible in a multiple-product fab or a foundry fab, in which there may be more than 300 product types, 300 processing steps, and multiple machines at each step. A well designed database and powerful data processing capability must be provided to support this method.

Manuscript received February 17, 2000. This work was supported in part by the National Science Council of the Republic of China under Grant NSC 87-2416-H-002-018 and by Taiwan Semiconductor Manufacturing Corporation under Contract 85-S-047.

C.-M. Fan and S.-C. Chang are with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

R.-S. Guo is with the Department of Industrial Management and Business Administration, National Taiwan University, Taipei, Taiwan, R.O.C.

C. S. Wei is with the Taiwan Semiconductor Manufacturing Corporation, Hsinchu, Taiwan, R.O.C.

Publisher Item Identifier S 0894-6507(00)06369-7.

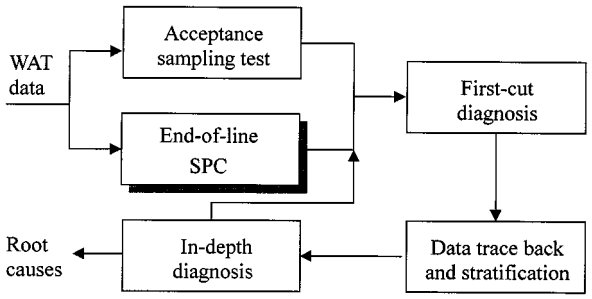


Fig. 1. Sequential detection and diagnosis approach for end-of-line quality control.

2) Sequential Detection and Diagnosis Approach: This approach detects and diagnoses the process abnormalities in a sequential manner. As shown in Fig. 1, quality of the incoming WAT data is first monitored in the acceptance sampling test module and the end-of-line SPC module. If there is an out-of-spec signal detected in the acceptance sampling test module or out-of-control signal detected in the end-of-line SPC module, a first-cut diagnosis function is then performed and potential process steps are identified. Based on the trace-back and stratification data at these critical steps, an in-depth diagnosis is performed. One current practice of end-of-line SPC is to review the C_{pk} values of key WAT parameters every two weeks. If any one of the C_{pk} values is less than the corresponding threshold values specified by engineers, a problem may have occurred and the corresponding control charts must be reviewed to see if there is any significant trend. The philosophy of this approach is to reduce the sequence-disorder and multiple-stream effects by plotting the C_{pk} data in a larger batch size such as lot data of two weeks so that a change in process mean and/or variance could be identified more easily. Another end-of-line SPC practice is the use of Shewhart control chart combined with variance decomposition as proposed by Philips Co. [4]. In their scheme, the multiple variance components in the end-of-line parametric measurements are first decomposed and then monitored in a batch-by-batch manner. In both practices, either control limits or batch size are usually determined empirically, which could degrade the detection performance. There should be a more rigorous rule for the design of these control parameters under different process conditions.

As for the research work in the general quality control field, cumulative sum (CUSUM) control chart [5] and exponentially weighted moving average (EWMA) control chart [6] are designed for a small shift or drift detection. Lucas and Saccucci [7] compared the effectiveness of the CUSUM and EWMA control charts and concluded that their performance are close to each other. They also validated that Shewhart control chart is superior to CUSUM and EWMA control charts in a large shift detection and suggested to adopt the combined Shewhart-CUSUM scheme [8] or combined Shewhart-EWMA scheme [7] to monitor various magnitudes of process shifts. However, none of them is designed specifically for the features of WAT data. As for the complicating features of WAT data—multiple-stream and sequence-disorder effects—only the former has been discussed in the literature. Montgomery [9], Nelson [10], and Mortell and Runger [11] developed several kinds of group control charts to detect the variation within each process stream as well as the variation among different process streams. But their methods are useful only in the case that each measurement can be easily stratified in terms of its corresponding stream. As for control charts for measurements with multiple variation components, total variation is first decomposed into various components by using ANOVA techniques and these components are then monitored individually [12]–[15]. With this variance decomposition technique, the sensitivity of control charts is greatly enhanced.

C. SHEWMA Scheme

To fully utilize the WAT data for end-of-line quality control, this paper adopts the "sequential detection and diagnosis" approach. With this approach, fewer control charts are needed for each WAT parameter as compared to those needed for the "brute force" SPC application. Furthermore, the data trace-back and stratification tasks are performed only at the critical process steps during the in-depth diagnosis stage. As a result, the root causes can be discovered much more efficiently without a significant amount of computing power, which is important in a multiple-product fab or a foundry fab. Since the Combined Shewhart-EWMA (CSE) scheme is very effective to monitor various magnitudes of process shifts, it is extended into a SHEWMA scheme in this paper for application to the end-of-line SPC module in Fig. 1. The proposed SHEWMA scheme is a methodology for generating robust design parameters for the simultaneous application of Shewhart and EWMA control charts to WAT data. By carefully designing the parameters of these charts against different process conditions and false alarm rate requirements, the SHEWMA scheme is able to identify the underlying trend from a multiple-streamed and sequence-disordered data sequence and optimize the trend detection performance. To be more specific, the goals and contributions of this paper include

- characterizing the features of WAT data;
- identifying the challenges of applying SPC to WAT data;
- designing robust SHEWMA parameters for industrial applications;
- implementing SHEWMA scheme within a foundry fab;
- integrating the detection function with the diagnosis function;
- validating the effectiveness of SHEWMA scheme using fab data.

D. Organization of this Paper

The remainder of this paper is organized as follows. In Section II, the WAT generation process is modeled and special challenges for SPC applications are discussed. Based on these results, in Section III, we describe the SHEWMA scheme for end-of-line SPC and its detailed design algorithm. Section IV then presents the evaluation results of SHEWMA performance. In Section V, the SHEWMA scheme is validated by fab data. The integration of SHEWMA detection function with diagnosis function is also discussed. Conclusions are finally made in Section VI.

II. Issues for Statistical Process Control at WAT

Statistical characteristics of WAT measurements are very complicated because they are taken at the end-of-line step and accumulate the effects of various sources of variation in the in-line processes. To detect special process disturbances in WAT data, it is necessary to characterize their generation and common variations.

A. Variations of WAT Data

Consider lots of wafers of the same part type, which are labeled by an index $i = 1, \cdots, I$ according to their sequence of finishing the WAT step. There are J wafers in each lot and K sites are sampled per-wafer at the WAT step. Let X_{ijk} be a WAT measurement taken from site k of the jth wafer in lot i, where $k = 1, \cdots, K$, and $j = 1, \cdots, J$. Each WAT measurement can be partitioned into four independent sources of variations such that

$$\begin{aligned} X_{ijk} - \overline{X}_{\bullet\bullet\bullet} = &(\overline{X}_{i\bullet\bullet} - \overline{X}_{\bullet\bullet\bullet}) + (\overline{X}_{ij\bullet} - \overline{X}_{i\bullet\bullet} \\ &+ (\overline{X}_{i\bullet k} - \overline{X}_{i\bullet\bullet}) \\ &+ (X_{ijk} - \overline{X}_{ij\bullet} - \overline{X}_{i\bullet k} + \overline{X}_{i\bullet\bullet}) \end{aligned} \quad (1)$$

where

$$\begin{aligned} \overline{X}_{\bullet\bullet\bullet} &\equiv \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk}, \\ \overline{X}_{i\bullet\bullet} &\equiv \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk}, \\ \overline{X}_{i\bullet k} &\equiv \frac{1}{J} \sum_{j=1}^{J} X_{ijk}, \quad \text{and} \\ \overline{X}_{ij\bullet} &\equiv \frac{1}{K} \sum_{k=1}^{K} X_{ijk}. \end{aligned} \quad (2)$$

The four terms on the right hand side of (1) correspond to lot-to-lot, wafer-to-wafer, site-to-site, and residual variations, where

1) *lot-to-lot variation* typically arises because different lots are processed on different machine in fabrication;
2) *wafer-to-wafer variation* results from the nonuniformity of batch processing machines or the nonideal repeat performance of single-wafer processing machines;
3) *site-to-site variation* results from the nonuniformity of each processing machine;
4) *Residual variation* is generated by the random disturbance on measurements and other unexplained variations such as the variation due to wafer and site interaction effect.

A set of WAT field data from a foundry fab is used to support our classification, where $I = 50$, $J = 24$, and $K = 5$. This set of data was empirically analyzed and judged by process engineers to be free from abnormal variations. Fig. 2 shows a multi-vari plot of a representative subset of the data, where a vertical line connects the largest and smallest observation within a wafer, and a horizontal tick represents the wafer mean over five sites. Both the lot-to-lot variation and the within-lot variation can be clearly observed, and the former is obviously more significant than the latter. Fig. 3 displays the within-lot data of five arbitrarily chosen wafers in a lot, where both wafer and site effects are observed. By applying the VARCOMP procedure of SAS program [16] to the complete data set, it is shown that the lot-to-lot, wafer-to-wafer, site-to-site, and residual variations contribute 66.0%, 10.9%, 11.4%, and 11.7% to the total variation, respectively. The SPC scheme design in this paper focuses on monitoring the WAT lot average sequence.

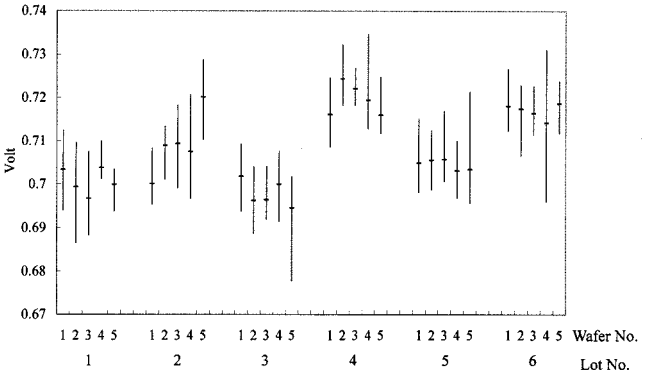


Fig. 2. Multi-vari plot for WAT data.

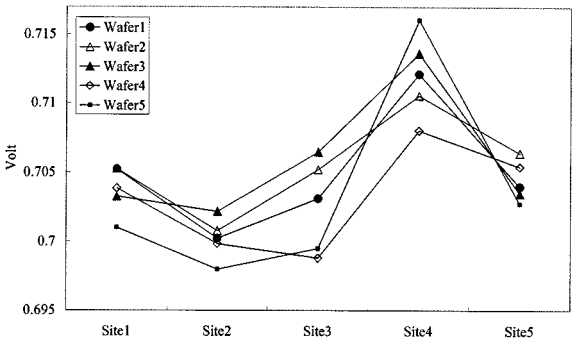


Fig. 3. Wafer effect and site effect within a lot.

B. Modeling the WAT Data

To simplify the notation, let $\{\overline{X}_i\} \equiv \{\overline{X}_{i\bullet\bullet}\}$ be a random sequence representing wafer lot averages of a WAT measurement item, where i is the lot output sequence index at the WAT step. Let us now analyze and develop models for the generation process of a WAT data sequence $\{\overline{X}_i\}$. Note that the processing of a lot may require more than 300 steps and each step may be processed by any one of a machine group. We define a *stream* as a sequence of machines that a lot goes through during its fabrication process. There are many possible streams in a fab and the resultant WAT measurements among different streams vary due to machine-to-machine variation. This is defined as the *multiple-stream effect*. In general, the cycle time from a process step p to the end-of-line WAT step also varies among lots. As a result, the lot with a sequence label n at step p very likely has a different lot sequence label i at the WAT step. This is defined as the *sequence-disorder effect*.

1) Modeling the Multiple-Stream Feature: When the process is in the in-control situation, $\overline{X}_i$'s are assumed to be

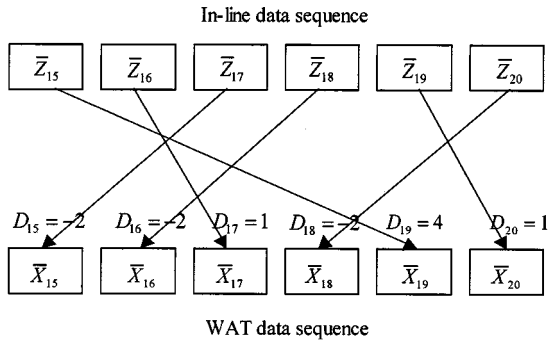


Fig. 4. Definition of sequence-disorder grade.

i.i.d. with a probability density function (p.d.f.) $f_0 \equiv N(\mu, \sigma^2_{\overline{X}})$. Now consider a model of multiple-stream effect for $\{\overline{X}_i\}$ with a shift in its process mean. Let machine m be one of the M identical but independent processing machines for step p. Assume that a special variation occurs at machine m at step p and it results in a shift of lot averages on $\{\overline{X}_i\}$. If lot i goes through machine m at step p during its fabrication process, the corresponding $\overline{X}_i$ is assumed to be i.i.d. with a conditional p.d.f. $f_s \equiv N(\mu + S\sigma_{\overline{X}}, \sigma^2_{\overline{X}})$, in which S is the magnitude of the shift in units of $\sigma_{\overline{X}}$. Otherwise, $\overline{X}_i$ is in-control and is i.i.d. with the conditional p.d.f. f_0. Assume a uniform probability that lot i goes through one of the machines for process step p, i.e., $1/M$. Thus, the probability density function of $\overline{X}_i$ under the mutiple-stream effect is

$$f_M = \frac{1}{M} f_s + \left(1 - \frac{1}{M}\right) f_0. \tag{3}$$

2) Modeling the Sequence-Disorder Feature: Let n be the in-line lot sequence label at step p and $\{\overline{Z}_n, n = 1, \cdots, I\}$ be the data sequence that reorders the WAT data sequence $\{\overline{X}_i\}$ according to the in-line sequence at step p. Define the sequence-disorder magnitude of a lot as $D_i \equiv i - n$, where n is the sequence label of a lot at step p and i is the sequence label of the same lot at the WAT step (Fig. 4). Then the range of D_i over all i, denoted as $R_p \equiv \max\{|D_i|, i = 1, \cdots, I\}$, is a characterization of the sequence-disorder effect from step p to WAT.

When a process is in-control, both $\{\overline{X}_i\}$ and $\{\overline{Z}_n\}$ are statistically the same, i.e., $\{\overline{X}_i\}$ and $\{\overline{Z}_n\}$ are both i.i.d. with probability density function f_0. Assume that a process shift of step p occurs at machine m starting from the n^*th lot and that the probability density function of $\overline{Z}_n$ is f_M for $n \geq n^*$. Due to the sequence-disorder effect, the probability that $\overline{X}_i$ has a shifted mean, i.e., lot i at the WAT step has an in-line sequence label $n \geq n^*$ at step p is defined as

$$a_i(n^*) \equiv \Pr(i - D_i \geq n^*) \tag{4}$$

and the p.d.f. of each $\overline{X}_i$ can then be inferred as

$$f_{\overline{X}_i} = a_i(n^*) f_M + (1 - a_i(n^*)) f_0. \tag{5}$$

C. Challenge for Trend Detection

To demonstrate the challenge in WAT trend detection, an example is created using the simulation model of Appendix A.

TABLE I
INPUTS TO THE SIMULATION EXAMPLE

Sequence-Disorder Range	Number of Machines	Magnitude Of Shift
$R = 15$	$M = 2$	$S = 1.5$
Total Data Points	Start Point of Shift	In-Control Density Function
$I = 50$	$n^* = 20$	$f_0 = N(0,1)$

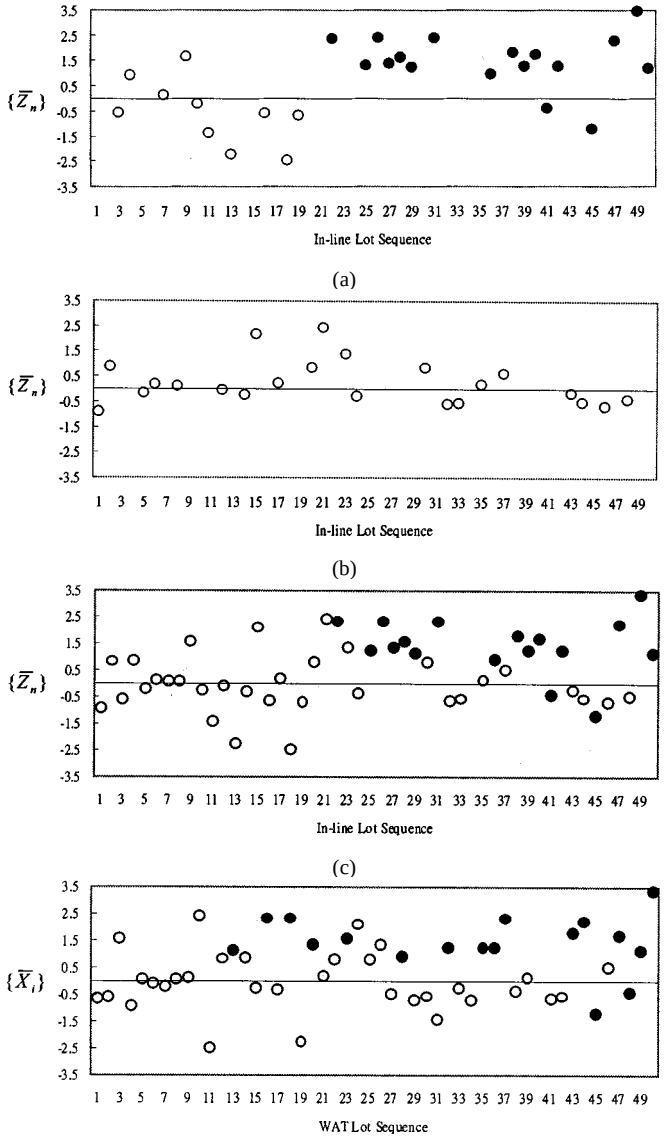


Fig. 5. Simulation to demonstrate the multiple-stream and sequence-disorder effects. (a) Machine A data sequence. (b) Machine B data sequence. (c) In-line data sequence. (d) WAT data sequence.

Model parameters are listed in Table I and the in-line sequences of individual machines are depicted in Fig. 5(a) and (b). It can be observed that after reordering WAT data of the abnormal machine A according to its in-line sequence, there is a significant shift pattern but not for the in-control machine B. Fig. 5(c) illustrates the multiple-stream effect by combining the data sequences of the two machines into one. Fig. 5(d) shows the WAT

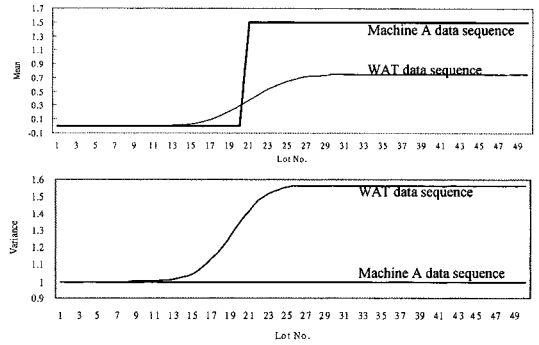


Fig. 6. Changes of mean and variance in the simulation case.

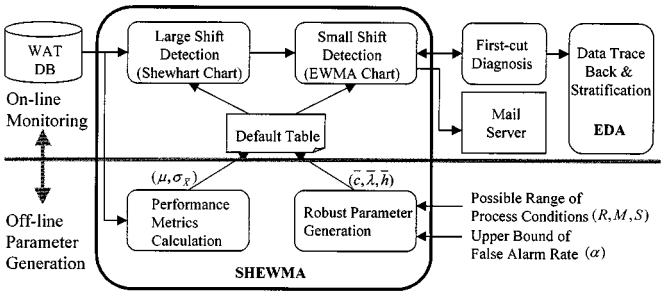


Fig. 7. SHEWMA system and environment.

sequence, where the sequence-disorder effect is added. It is visually clear that the trend pattern in Fig. 5(d) is ambiguous and difficult to identify.

Fig. 6 demonstrates, by using the simulation example, the changes in both mean and variance of WAT data sequence in contrast with those of the in-line sequence at machine A. The results in Fig. 6 are derived from (5)

$$E\{\overline{X}_i\} = \left(1 - \frac{a_{i(n^*)}}{M}\right)\mu + \frac{a_{i(n^*)}}{M}(\mu + S\sigma_{\overline{X}}) \qquad (6)$$

and

$$\mathrm{Var}\{\overline{X}_i\} = \left[\frac{a_{i(n^*)}}{M}\left(1 - \frac{a_{i(n^*)}}{M}\right)S^2 + 1\right]\sigma^2_{\overline{X}}. \qquad (7)$$

It can be seen that the in-line process shift ramps and then levels off in the WAT sequence because of the sequence-disorder effect. The slope of the ramp of $\mathrm{E}\{\overline{X}_i\}$ is $S/2MR$ approximately, while $\mathrm{E}\{\overline{X}_i\}$ and $\mathrm{Var}\{\overline{X}_i\}$ finally reach at the steady values $\mu+(S/M)\sigma_{\overline{X}}$ and $[(1/M)(1-(1/M))S^2+1]\sigma^2_{\overline{X}}$, respectively. It means the larger the sequence-disorder range, the smaller the slope of the ramp. Also, due to the multiple-stream effect, the magnitude of the leveling off part is M times smaller while the variance is $[(1/M)(1-(1/M))S^2+1]$ times larger than that of the original shift pattern. It is clear that an in-line trend pattern becomes ambiguous and difficult to detect from the WAT data sequence.

III. Design of SHEWMA Scheme for WAT Data

Our design aims at following goals:

1) to accumulate the evidence of any emerging trend in WAT data sequence and extract the trend pattern;
2) to be sensitive to real process disturbance without increasing the false alarm rate;
3) to be easy to implement and robust for various process conditions in a real fab.

A. Overview of SHEWMA System

The SHEWMA scheme has been implemented as a software system interfacing with the WAT database and the engineering data analysis (EDA) system. By monitoring the WAT data of each wafer lot, this system is designed to detect an abnormal trend and trigger the "sequential detection and diagnosis" functions (Fig. 1). The architecture and environment of the SHEWMA system is shown in Fig. 7. Data monitoring by the SHEWMA scheme is typically completed within a few seconds after each lot leaves a WAT tester. If a WAT warning message is generated, the system will send an e-mail message to the engineers in charge. Engineers will then query the corresponding control charts and perform a first-cut diagnosis. By tracing back the data in the EDA system, engineers may further correlate the control charts in the SHEWMA system to the in-line data of potential abnormal process steps and conduct an in-depth diagnosis.

The SHEWMA system itself consists of two subsystems: on-line monitoring and off-line parameter generation. In the on-line monitoring subsystem, lot averages $\{\overline{X}_i\}$ are queried from the WAT database with the outliers of raw data excluded. Then the large shift detection module (the Shewhart chart) tests if the average of a lot is in-control, and the small shift detection module (the EWMA chart) tests if there is any trend shown in $\{\overline{X}_i\}$. Warning messages from these two modules provide information about process shift size. If only the EWMA chart detects an abnormal trend, there could be a small process shift. When there is a large trend in the EWMA chart and a data point out of Shehwart control limits at the same time, a large process shift may have occurred.

In the Shewhart control chart, the monitoring statistic is the lot average sequence $\{\overline{X}_i\}$. In the EWMA control chart, the EWMA sequence is generated by

$$A_i = \lambda\overline{X}_i + (1-\lambda)A_{i-1} \qquad (8)$$

where $0 < \lambda \leq 1$, and the initial value A_0 is usually set as the process mean μ. In summary, SHEWMA scheme parameters form a triplet (c, λ, h), where c is the Shewhart control limit width, λ is the EWMA weighting factor, and h is the EWMA control limit width. Once the SHEWMA parameters (c, λ, h) and the long-term performance $(\mu, \sigma^2_{\overline{X}})$ are available, control limits are then set as

Shewhart chart:

$$\begin{aligned} SCL_U &= \mu + c\sigma_{\overline{x}} \\ SCL_L &= \mu - c\sigma_{\overline{x}} \end{aligned} \qquad (9)$$

EWMA chart:

$$\begin{aligned} ECL_U &= \mu + h\sqrt{\lambda/(2-\lambda)}\sigma_{\overline{x}} \\ ECL_L &= \mu - h\sqrt{\lambda/(2-\lambda)}\sigma_{\overline{x}} \end{aligned} \qquad (10)$$

To optimize the scheme performance, an Off-line Parameter Generation process is required; robust SHEWMA parameters

are generated based on possible process conditions and the bound of false alarm rate. Results of these parameters are then stored in the Default Table for on-line monitoring function. There are two modules in this subsystem. The first module, Performance Metrics Calculation, calculates the mean μ and variance $\sigma^2_{\overline{X}}$ of a historical WAT lot average sequence $\{\overline{X}_i, i = 1, 2, \cdots, I_0\}$. In specific, the moving range estimator of [9] is adopted to estimate the variance $\sigma^2_{\overline{X}}$,

$$MR_i \equiv |\overline{X}_{i+1} - \overline{X}_i|, \qquad i = 1, 2, \cdots, I_0 - 1$$
$$\overline{MR} = \frac{1}{I_0 - 1} \sum_{i=1}^{I_0 - 1} MR_i, \quad \text{and}$$
$$\sigma_{\overline{X}} \approx 0.887 \overline{MR}. \tag{11}$$

This estimator is unbiased, is robust with respect to shifts in the process mean [13], and can model the machine-to-machine variation among lots well. The second module, Robust Parameter Generation, generates the SHEWMA parameters that maximize the detection speed while keeping the false alarm rate lower than a required level α.

B. Robust Parameter Generation

In this paper, the average run length, ARL, serves as a performance metric of the SHEWMA scheme. ARL is a random variable characterizing the average number of observations that an SPC scheme takes to generate an out-of-control signal after the occurrence of a process change. In general, ARL is further classified into $ARL0$ and $ARL1$. $ARL0$ represents the average run length when the process is under normal condition while $ARL1$ represents the average run length when an abnormal situation occurs. The reciprocal of $ARL0$ has a meaning of false alarm rate. One possible way to design the SHEWMA parameters (c, λ, h) is to minimize $ARL1$ for a given set of process conditions and a specified upper bound of false alarm rate. The process conditions are denoted as a triplet (R, M, S), where R is the sequence-disorder range from the monitored step p to WAT step, M is the total number of machines in the monitored step p, and S is the potential magnitude of a shift (in the unit of $\sigma_{\overline{X}}$). In practice, there may be a wide range of process conditions in a real fab and exact process conditions (R, M, S) cannot be known *a priori*. For the feasibility of implementation, design of a robust set of SHEWMA parameters $(\overline{c}, \overline{\lambda}, \overline{h})$ is desirable so that the SHEWMA scheme results in satisfactory performance over various process conditions.

The design procedure is summarized in Fig. 8. Design inputs include a set of process conditions, $\Omega \equiv \{(R, M, S), R \in \mathbf{R}^+, M \in \mathbf{Z}^+, S \in \mathbf{R}^+\}$, and an upper bound of the false alarm rate, α. Design output is a robust selection of parameters, $(\overline{c}, \overline{\lambda}, \overline{h})$. There are two parts in the design procedure. The first part calculates the feasible parameters with false alarm rate α, from which the optimal parameters are generated by minimizing $ARL1$ under any given process condition triplet in Ω. The second part generates a robust design of parameters by minimizing the worst case detection delay of the SHEWMA scheme over all possible conditions in Ω. Interested readers may refer to Appendices B and C for more discussions.

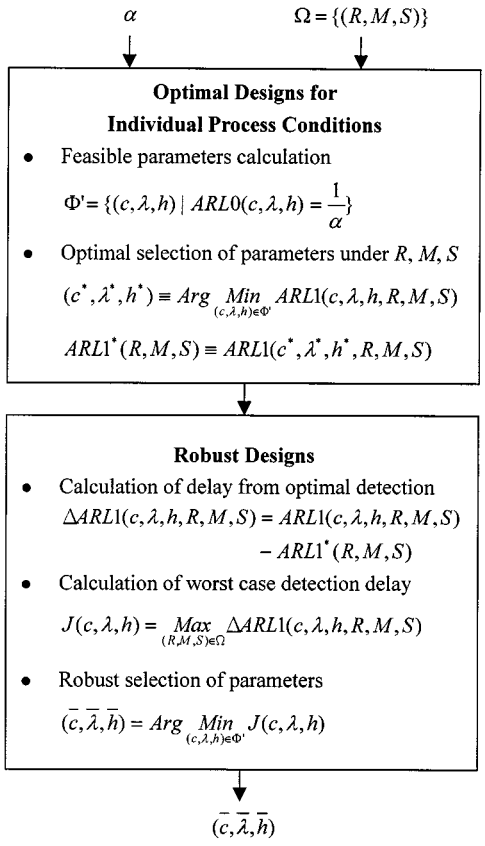


Fig. 8. Design procedure of the SHEWMA scheme.

IV. Performance Evaluation

The objective of robust SHEWMA design is to optimize the scheme performance for a wide range of possible process conditions in a real fab. This Section examines if the design is necessary and if the design objective is indeed achieved. First, a simple simulation example is used to highlight the optimal performance of the SHEWMA scheme in trend identification under one set of process conditions. Next, a more rigorous sensitivity study is performed, in which ARL performance against different scheme parameters and process conditions are presented. Finally, the effectiveness of robust SHEWMA design is demonstrated by comparing it with the combined Shewhart-EWMA (CSE) scheme, where the multiple-stream and sequence-disorder effects are not considered.

A. Simulation Example

In this simulation example, WAT data sequences are first generated as described in Appendix A. Lot averages $\overline{X}_i$ under an in-control condition is assumed to follow a normal distribution $N(0, 1)$. Tables I and II list the process conditions, requirement of false alarm rate and scheme parameters in this simulation study. The process conditions are the same as those of the example in Section II, where a shift occurs at the 21st data point of the in-line sequence. Under such process conditions and the requirement of false alarm rate of 0.27%, the optimal SHEWMA parameters are $(c, \lambda, h) = (3.25, 0.05, 2.693)$. To investigate the effect of weighting factor λ, Table II includes three more

TABLE II
FALSE ALARM RATE AND SHEWMA PARAMETERS IN THE SIMULATION EXAMPLE

False Alarm Rate		
α = 0.27% ; $ARL0$ = 370		
SHEWMA Parameters		
c	λ	h
3.25	0.01	2.069
3.25	0.05	2.693
3.25	0.35	3.055
3.25	1	3.250

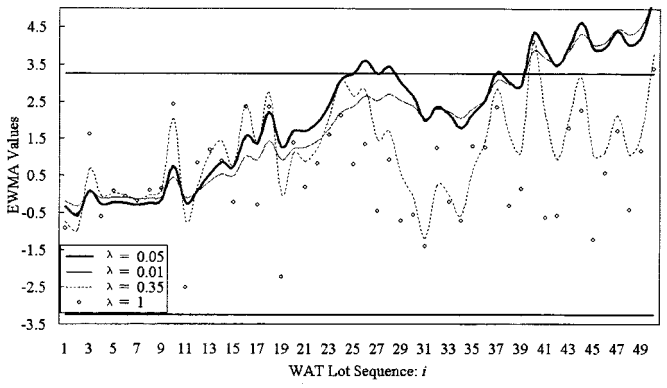


Fig. 9. SHEWMA chart of the simulated case in Fig. 5.

sets of SHEWMA parameters besides the optimal one for comparison.

Fig. 9 depicts simulation results of the four designs, where all the monitoring data are normalized to have the same control limits. As can be seen, the EWMA trend generated by the optimal parameters (the bold solid line) approaches the upper control limit the fastest among the four. The optimal SHEWMA scheme generates an out-of-control alarm at the 25th lot while the other three do not detect until the 39th, 40th and 49th lots respectively. It can also be observed that when the weighting factor is smaller than the optimal value, the EWMA trend pattern is clearer at a price of slower out-of-control detection. When the weighting factor is higher than the optimal value, the EWMA trend pattern becomes blurred without getting any benefit in the speed of out-of-control detection.

B. Sensitivity Analysis

ARL performance against different scheme parameters and process conditions is now characterized. The ARL's are numerically derived by using the methods described in Appendix B. To validate the accuracy of numerically derived ARL's, Monte Carlo simulations are conducted. The differences of ARL's between the numerical derivation and simulations are mostly within the 95 percent confidence intervals, i.e., two times of the standard deviations. Relative differences are all within 4%.

Consider a range of process conditions in a real fab as

$$\Omega = \{(R, M, S) | 0 \le R \le 50, 1 \le M \le 5, 0 \le S \le 2\} \tag{12}$$

and design seven cases for sensitivity analysis as listed in Table III. The $ARL1$ performance of SHEWMA scheme with

TABLE III
SEVEN TEST SCENARIOS FOR DESIGN AND ANALYSIS

Scenario	R	M	S
S1	0	3	1
S2	25	3	1
S3	50	3	1
S4	25	1	1
S5	25	5	1
S6	25	3	1.5
S7	25	3	2

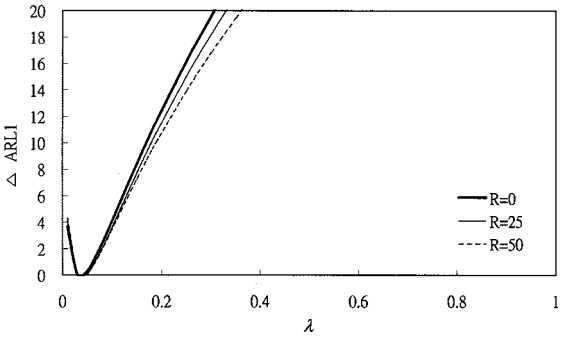


Fig. 10. Sensitivity of $ARL1$ to weighting factor λ under various R values with $M = 3$ and $S = 1$.

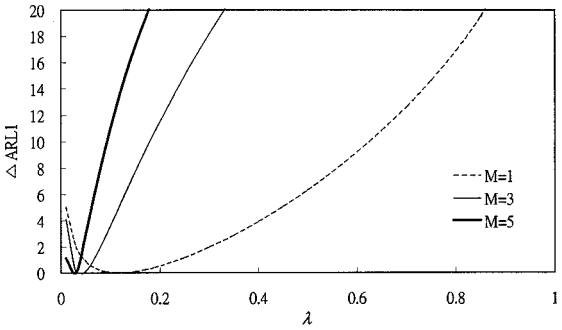


Fig. 11. Sensitivity of $ARL1$ to weighting factor λ under various M values with $R = 25$ and $S = 1$.

respect to the changes of weighting factor λ and the process conditions in Table III are examined. For all the seven cases, the Shewhart control limit width c is set to 3.25 while the EWMA control limit width h is determined based on λ so that the false alarm rate α is equal to a frequently used level of 0.27%.

Fig. 10 shows the sensitivity of $ARL1$ with respect to λ under various values of sequence-disorder range R. The vertical axis represents the delay from the optimal detection time, $\Delta ARL1$, which is defined in (B7) of Appendix B. It can be seen that the optimal λ value (λ^* = 0.03) stays the same, i.e., the SHEWMA parameter λ is insensitive to the change in sequence-disorder range R.

Fig. 11 shows that the optimal weighting factor λ decreases as the number of machines M increases, which means that a larger window size of moving average is needed to reveal the underlying trend under a stronger multiple-stream effect. This figure also shows that the sensitivity of $ARL1$ with respect to λ increases as M increases. Note that if λ is set to 0.13 by assuming $M = 1$, i.e., without considering the multiple-stream effect, there may be a detection delay up to 15 lots ($\Delta ARL1 \approx$

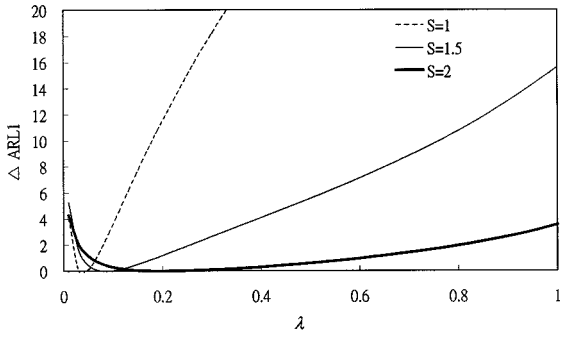


Fig. 12. Sensitivity of $ARL1$ to weighting factor λ under various S values with $R = 25$ and $M = 3$.

TABLE IV
ROBUST SHEWMA SCHEME VERSUS COMBINED SHEWHART-EWMA SCHEME

Scheme	Robust SHEWMA scheme	Combined Shewhart-EWMA scheme
False Alarm Rate	0.27%	0.27%
Range of Process Conditions	$0 \le R \le 50$ $1 \le M \le 5$ $0 \le S \le 2$	$R = 0$ $M = 1$ $0 \le S \le 2$
Robust Scheme Parameters	$(\bar{c}, \bar{\lambda}, \bar{h}) = (3.25, 0.03, 2.523)$	$(\bar{c}, \bar{\lambda}, \bar{h}) = (3.75, 0.19, 2.866)$

15) when $M = 5$. On the contrary, if the multiple-stream effect is taken into account, the λ value based on the robust design will be around 0.03, which yields a maximum detection delay of no more than 2 lots ($\Delta ARL1 \approx 2$) for M ranging from 1 to 5.

In Fig. 12, there is a quick decline in optimal λ as the magitude of shift S decreases. The sensitivity of $ARL1$ with respect to λ increases rapidly as S decreases. A detection delay up to 10 lots ($\Delta ARL1 = 10$) might occur if λ is set to the optimal value of 0.19 for $S = 2$ instead of the robust design value of 0.05.

C. Robust SHEWMA Scheme versus Combined Shewhart-EWMA Scheme

The robust SHEWMA scheme is compared with the CSE scheme for demonstrating its effectiveness. Table IV lists the false alarm rate, range of process conditions, and the resultant design parameters by these two schemes.

Results of the $ARL1$ performances with respect to the number of machines (M), sequence-disorder range (R), and shift size (S) are illustrated in Figs. 13–15, respectively. It can be seen that as the number of machines increases, the robust SHEWMA scheme is getting superior to the CSE scheme. However, the $ARL1$ performance seems independent of the sequence-disorder range. As for the magnitude of shift, the smaller the shift size, the better the performance of robust SHEWMA scheme. In summary, if there exists a multiple-stream effect and if the shift size is small, the superiority of our SHEWMA scheme over the CSE scheme will be most significant, with at least 10% reduction in detection time.

Consider the process conditions in (12). The worst case detection delays, $J_{c,\lambda,h}$'s as defined in (B6) of Appendix B, of SHEWMA and the CSE schemes are 2.4 and 20.3 respectively. Suppose that the average throughput of a foundry fab is 5 lots per part type per day. A SHEWMA scheme with the robust parameters $(\bar{c}, \bar{\lambda}, \bar{h}) = (3.25, 0.03, 2.523)$ will delay the generation of a warning message only by an average of $2.4/5 = 0.48$ day even in the worst case. The worst case detection delay for CSE scheme may be as long as $20.3/5 = 4.06$ days.

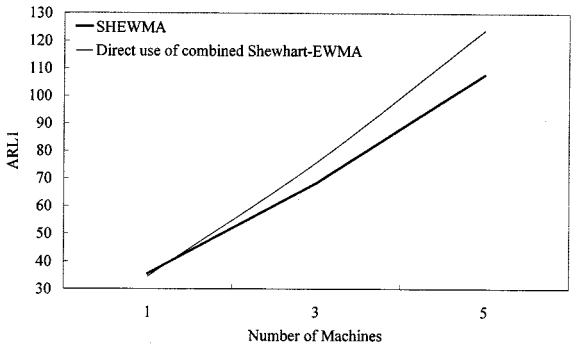


Fig. 13. Relation of $ARL1$ to M with $R = 25$ and $S = 1$.

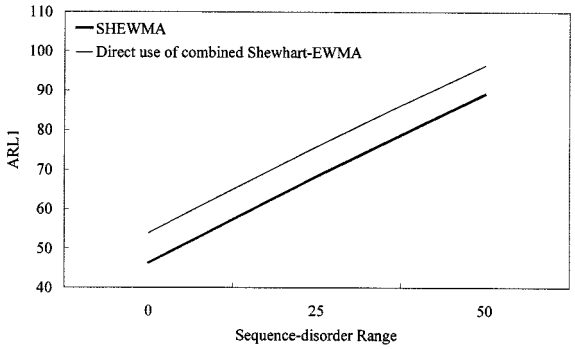


Fig. 14. Relation of $ARL1$ to R with $M = 3$ and $S = 1$.

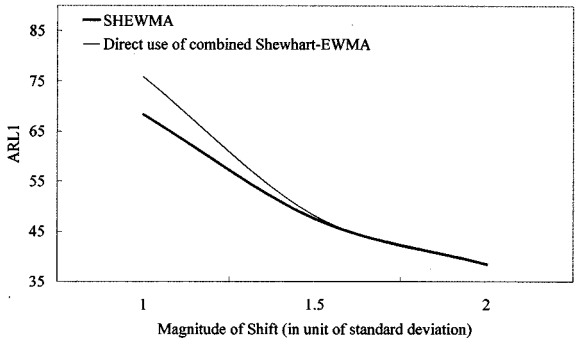


Fig. 15. Relation of $ARL1$ to S with $R = 25$ and $M = 3$.

V. APPLICATION TO A FOUNDRY FAB

The SHEWMA scheme has been implemented in a foundry fab following the schematic diagram of Fig. 7. To facilitate process integration, the SHEWMA system is further integrated with the EDA system, both of which are running on the same NT server. Process integration engineers can easily access the two systems through the Intranet.

A. Implementation Issues

The establishment of a baseline process model from WAT data follows the standard industrial practice as follows:

1) Identify and remove extreme data points by empirical rules.
2) Draw a probability plot to check if there is any outlier in the remaining data and if the data is of a normal distribution.
3) Remove all the outliers and calculate the mean and variance for setting SHEWMA control limits.
4) Exclude the out-of-control-limit data and recalculate the SHEWMA control limits until all the data points are within the control limits.

There are two salient issues in implementing the SHEWMA system.

1) Extraction of Process Condition Range Parameters: Based on process physics and by design of the test structures, each WAT item is only correlated to certain process steps. A straight forward adoption of the process condition range parameters of the whole fab as inputs to the robust parameter generation module may lead to an increase in worst case delay for fault detection. This is because the process condition range parameters of the whole fab is an upper bound to the process condition range parameter for a WAT item. To correctly extract process condition range parameters from fab data for individual WAT items, it is crucial to precisely identify the correlated process steps of each WAT item.

2) Integration with the Diagnosis Process: To find out the root cause after receiving a fault detection warning message from SHEWMA, a direct diagnosis method is to trace the WAT data sequence back to the processing machines in all the possibly faulty process steps. An efficient diagnosis obviously requires not only the information about which WAT items are correlated to which process steps but also information about how they are correlated. The extraction of such a knowledge base from empirical data and physical laws is key to the integration between fault detection by SHEWMA and the diagnosis process.

A solution to the two implementation issues exploits a mapping table which correlates WAT data and process steps. For each WAT item, its sequence-disorder range is extracted from the production data of its correlated process steps. Let P be a set of process steps correlated to a WAT item under investigation and $p \in P$ is the earliest step in the process flow among all steps in P. Suppose that there are I lots processed at step p in one day. The arrival day of each lot at WAT step is assumed to be within the range (mean arrival day $\pm\ 2\omega_P$), where ω_p is the standard deviation of the cycle time from step p to the WAT step. Thus the difference among the arrival days of the I lots at WAT step is at most $4\omega_p$ days. The sequence-disorder range R_p of step p is therefore estimated by

$$R_p \approx 4\omega_p T_p, \tag{13}$$

where T_p is the throughput at WAT step. The sequence-disorder range R_p is an upper bound to the sequence-disorder ranges of all other steps in P. Similarly, the sequence-disorder range of the last step in the process flow among all steps in P serves as a lower bound. As for the other two range parameters, the number of machines (M) and the magnitude of shift (S), they are empirically and easily determined.

TABLE V
RELATED PROCESS STEPS OF Rc_N+ AND Rs_N+

WAT Item	Related Process Step
Rc_N+	N+ S/D Implant, N+ S/D RTA, Contact Photo, Contact Etch, Ti Barrier, TiN Barrier, Ti/TiN RTA, W Plug, Thermal Process in Metal 1
Rs_N+	N+ S/D Implant, N+ S/D RTA, Thermal Process in Contact & Metal 1

With the use of the mapping table between WAT data and process steps, the SHEWMA scheme is integrated with an intelligent diagnosis system (IDS) for the purpose of first-cut diagnosis [17]. When an abnormal WAT symptom is detected by SHEWMA, the deviation of WAT data is then fed into the IDS system. The IDS system then calculates the fault causing possibility of individual process steps. As a result, a list of possible faulty process steps are generated for further in-depth diagnosis.

B. Case Study

A 0.26 μm logic device is selected with a focus on monitoring WAT items of Rs_N+ and Rc_N+, which represents the sheet and contact resistance of N+ structure respectively. The two WAT items are monitored for evaluating concentration and contact of NMOS drains/sources fabricated on each wafer. Process steps related to these two WAT items are listed in Table V. The process range condition parameters derived from using historical production data are $\Omega = \{(R, M, S) | 10 \leq R \leq 30, 1 \leq M \leq 3, 0 \leq S \leq 2\}$.

Data from 120 lots were collected over a period of 1.5 months. The first 50 lots are used for baseline process model construction while the last 70 lots are used for on-line monitoring. The false alarm rate is again set to 0.27%. The robust SHEWMA parameters are then generated as $(\overline{c}, \overline{\lambda}, \overline{h}) = (3.25, 0.11, 2.9)$. The long-term mean (standard deviation) for these 50 lots are 66.01 (0.768) and 69.93 (0.460) for Rc_N+ and Rs_N+, respectively.

C. Trend Detection via SHEWMA

Figs. 16 and 17 illustrate the SHEWMA control charts of the Rc_N+ and Rs_N+ respectively. Each figure has a Shewhart chart in part (a) and an EWMA chart in part (b).

In the application of SHEWMA scheme to Rc_N+ data, there are two warning messages generated by the EWMA chart at the 61st and 64th lots, but all the data points are within the Shewhart control limits. As EWMA is more sensitive to small shift detection while Shewhart is better in detecting a large deviation, it is deduced that Rc_N+ data may have a small shift.

In monitoring the Rs_N+ data, SHEWMA generates four warning messages, one from the Shewhart chart at the 65th lot and the other three from the EWMA chart at the 27th, 37th, and 64th lots, respectively. Under the same reasoning as the one above, the deviation of Rs_N+ may be conjectured as a small shift when the monitoring procedure is around the 27th

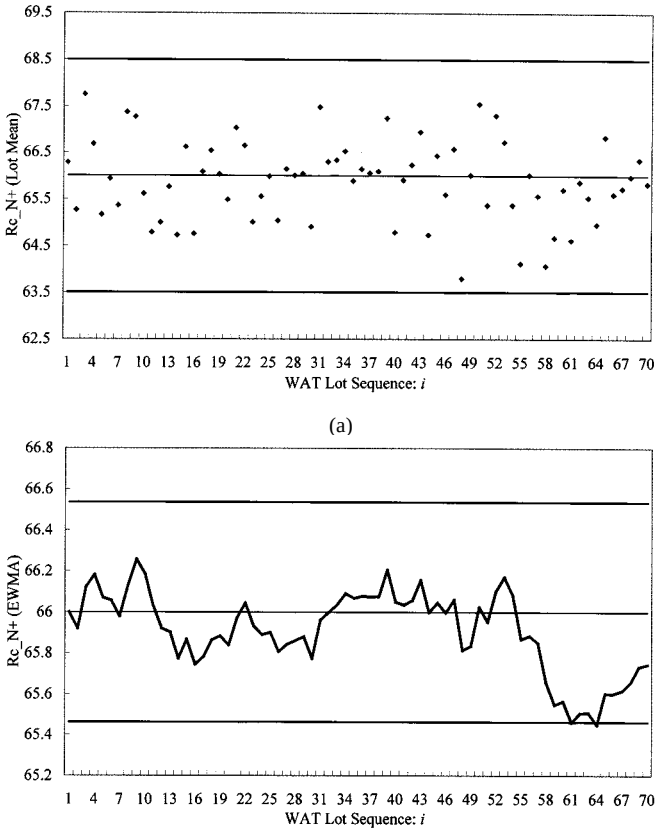


Fig. 16. Robust SHEWMA control charts of Rc_N+. (a) Shewhart chart at WAT ($c = 3.25$). (b) EWMA chart at WAT ($\lambda = 0.11$, $h = 2.899$).

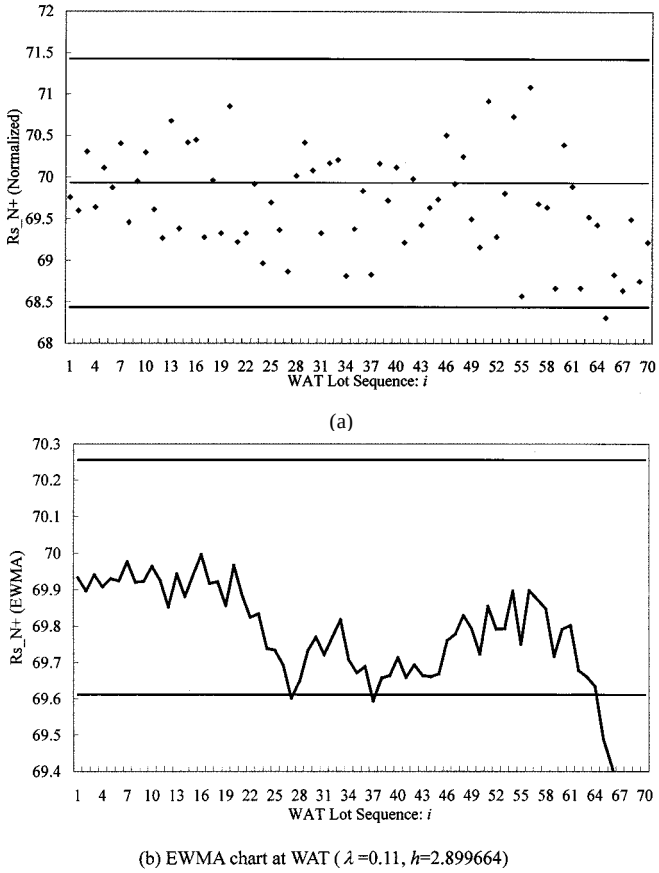


Fig. 17. Robust SHEWMA control charts of Rs_N+. (a) Shewhart chart at WAT ($c = 3.25$). (b) EWMA chart at WAT ($\lambda = 0.11$, $h = 2.899$).

to 37th lots and as a large shift when the monitoring procedes up to the 65th lot.

D. SHEWMA and Root Cause Diagnosis

The warning messages and control charts of SHEWMA trigger and assist engineers in root cause diagnosis. When reviewed by engineers, the EWMA control chart provides a visual trend pattern of WAT data sequence and facilitates intuitive estimation of the type and size of WAT data deviation. Such a visualization for each WAT item in turn gives indications for root cause diagnosis. For example, the EWMA values of Rs_N+ in Fig. 17(b) have two slight downward trends around the 27th lot and the 37th lot respectively and a large downward trend after the 64th lot. So, it is conjectured that a small process shift results in the two slight downward trends while another large process shift generates the large trend-down. In the EWMA chart for Rc_N+ [Fig. 16(b)], there is also a small trend-down after the 64th lot but no abnormal symptoms around the 27th lot and the 37th lot. It is therefore reasoned that the faulty process step affects Rs_N+ more than Rc_N+.

The Rs_N+ data sequence is then traced back to the N+ drain/source implant and RTA steps for in-depth diagnosis. It is found that N+ drain/source implant step is the root cause. There are four machines, M1 $\sim$ M4, for this step. In tracing back, the lot average sequence of Rs_N+ is stratified by the four machines [Fig. 18(a)] and reordered by the lot sequence at the step [Fig. 18(b)]. It can be clearly observed from Fig. 18(b) that M1 has a significant machine offset from the 29th to 36th lots in its in-line lot sequence as compared to the other machines. Also, there is a process shift occurred at M4 starting from the 62th lot in its in-line lot sequence. These results validate the first-cut diagnosis and the earlier conjectures that there is a small shift (around the 27th and 37th lots in end-of-line WAT lot sequence) and a large shift (after the 64th lots in end-of-line WAT lot sequence) in Rs_N+.

E. EOL SHEWMA Complementary to In-Line SPC

Can the fault be identified by using in-line SPC for the N+ drain/source implant step? The in-line SPC at the N+ drain/source implant step monitors the sheet resistance, which is taken from the test wafer every 12 h. Both Western Electric Rules (WER) and CSE schemes are adopted as the in-line SPC schemes.

The CSE charts for machine M4 at N+ drain/source implant step during the tracking time of the 70 lots under investigation are given in Fig. 19(a) and (b). During the period of process shift, there are 23 lots, from the 48th to the 70th lot in Fig. 18(b), processed by machine M4 for the N+ drain/source implant step. However, in the same period of time, only four data points, from the 13th sampling point to the 16th sampling point in Fig. 19, of sheet resistance are taken for in-line SPC. It can be seen that, using the in-line sheet resistance data, neither the CSE scheme nor the WER detects the large process shift in machine M4. As

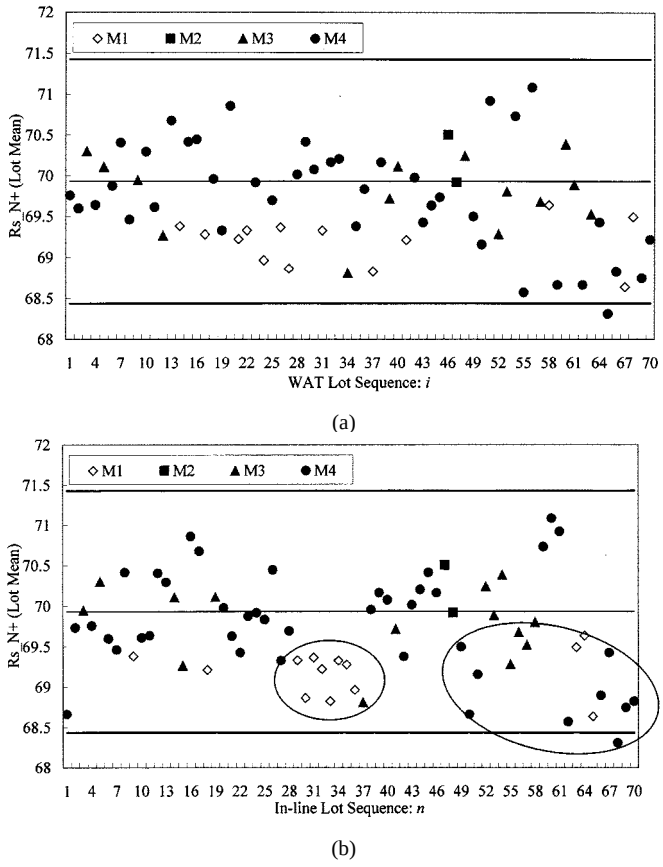


Fig. 18. Shewhart control charts stratified by processing machines. (a) Shewhart chart at WAT ($c = 3.25$). (b) Shewhart chart reordered by in-line lot sequence ($c = 3.25$) .

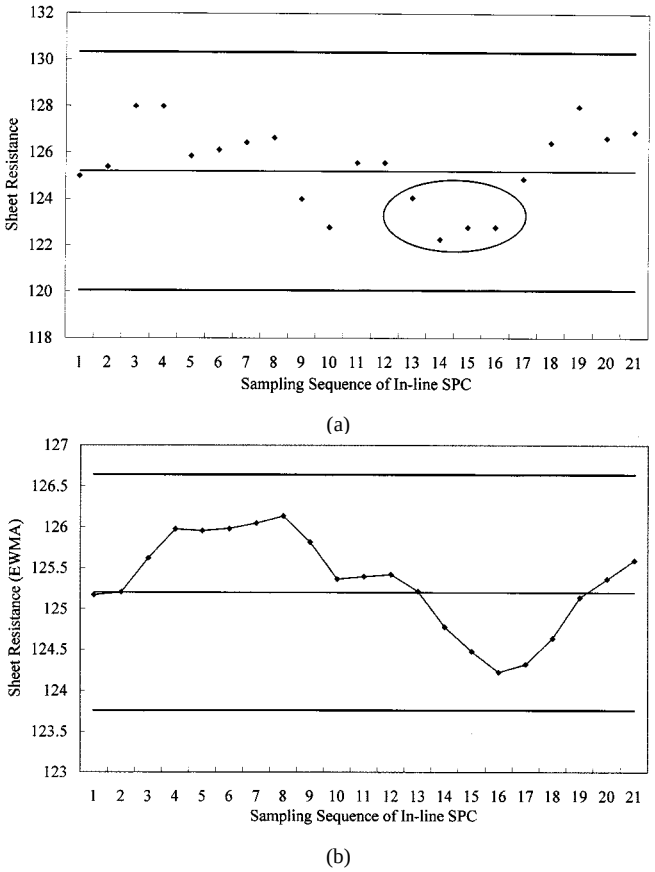


Fig. 19. SHEWMA control charts of M4. (a) Shewhart chart for in-line SPC ($c = 3$). (b) EWMA chart for in-line SPC ($\lambda = 0.15, h = 2.961$).

for the offset of machine M1, it is more difficult to detect by using in-line SPC because only two data points of sheet resistance are taken from M1 and the offset of M1 is much less than the magnitude of process shift in M4.

There are two reasons that the in-line SPC does not detect the process shift and machine offset in this case. First, the in-line measurements may be less sensitive to the process change as compared to the WAT measurements taken from product wafer. Second, the sampling rate in in-line level is much less than that of WAT. End-of-line SHEWMA is thus complementary to the in-line SPC for process integration.

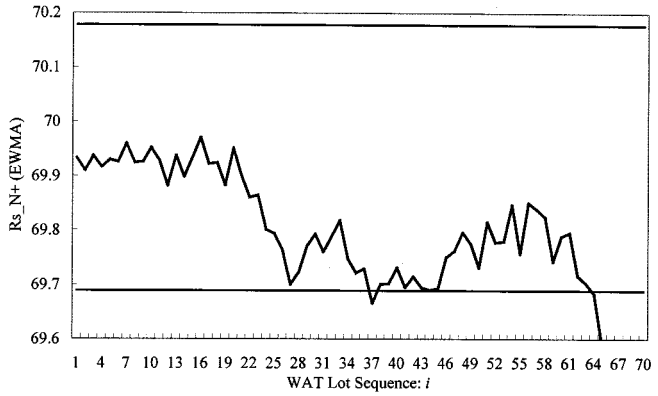
F. Necessity of Robust Design

Through this case study, SHEWMA has been validated as a useful scheme for end-of-line detection and assists in diagnosis in a real fab. To simplify the implementation of SHEWMA system, can the procedure of robust parameter generation be omitted? To answer the question, in addition to the robust SHEWMA parameters used so far, two additional sets of SHEWMA parameters, $(c, \lambda, h) = (3.25, 0.07, 2.789)$ and $(c, \lambda, h) = (3.25, 0.5, 3.052)$, are designed under the same false alarm rate requirement $\alpha = 1/370$. The two sets are selected to evaluate the selection of weighting factor λ, because it is related to the effective moving window size and is important for revealing out the underlying trend pattern and enhancing the detection speed of SHEWMA. In the one hand, to reveal the underlying trend, a large window size (a small weighting factor λ) is needed. As a result, a smaller value of λ (0.07) is adopted in the first of the two new sets. On the other hand, to enhance the sensitivity of EWMA values to process change, a larger value of λ (0.5) is adopted in the second of the two new sets. It is expected that the SHEWMA scheme with the empirically determined parameters will slow down the detection of small process shift.

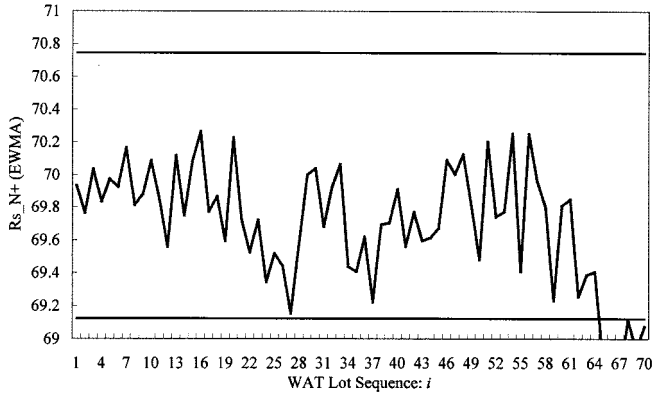
The EWMA control charts of Rs_N+ using these two sets of parameters are demonstrated in Fig. 20(a) and (b), respectively. It can be observed that both of them detect the large process shift at the 64th lot, which is the same as that by the robust SHEWMA parameters. However, for the small shift resulted by machine offset, the EWMA chart with a smaller weighting factor $\lambda = 0.07$ generates a warning message at the 37th lot, which delays the detection by ten lots as compared to the EWMA chart with robust parameters. In the EWMA chart with a larger weighting factor $\lambda = 0.5$, no warning message is signaled for the small shift around the 27th and 37th lots.

VI. Conclusions

In this paper, an end-of-line quality control scheme based on WAT data is presented. In particular, the design and implementation of an end-of-line SPC scheme and its integration with diagnosis function is detailed. The end-of-line SPC scheme, SHEWMA, considers the multiple-stream and sequence-disorder effects of WAT data and generates robust



(a)



(b)

Fig. 20. EWMA control charts with arbitrarily-chosen parameters. (a) EWMA chart with a small weighting factor ($\lambda = 0.07, h = 2.789\,395$). (b) EWMA chart with a large weighting factor ($\lambda = 0.5, h = 3.051\,675$).

design parameters for the simultaneous use of Shewhart and EWMA control charts.

The SHEWMA scheme has been implemented in a foundry environment for monitoring the lot-to-lot average performance and integrated with the fab EDA system for root cause diagnosis. Its detection and diagnosis-enhancing capabilities have been validated using both numerical derivations and fab data. Numerical results show the robust SHEWMA scheme reduces detection time by at least 10% for small shift detection as compared to the direct use of combined Shewhart-EWMA scheme without considering the multiple-stream and sequence-disorder effects. In addition, in the fab data validation study, the SHEWMA scheme reveals the potential faults that are not captured in the in-line step. Its use is complementary to the existing in-line SPC for process integration.

APPENDIX A

With the multiple-stream and sequence-disorder models combined, the p.d.f. of $\overline{X}_i$ can be related to f_s and f_0 of step p as

$$f_{\overline{X}_i} = a_i(n^*)\left\{\frac{1}{M}f_s + \left(1-\frac{1}{M}\right)f_0\right\} + (1-a_i(n^*))f_0. \tag{A1}$$

A simulation model for generating $\{\overline{X}_i\}$ based on (A1) is shown in Fig. 21. Inputs include the sequence-disorder range R, number of machines M, magnitude of shift S, total number

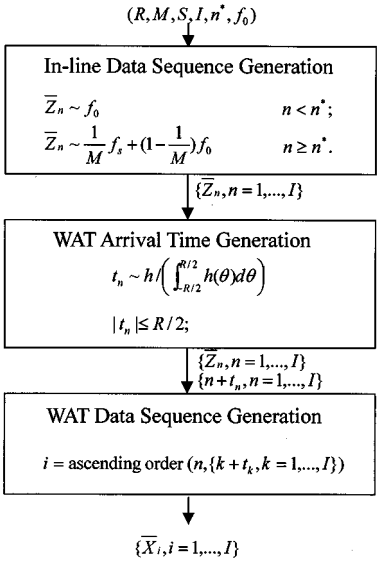


Fig. 21. Simulation of the WAT data generation process, where $f_0 = N(\mu, \sigma^2_{\overline{X}}$, $f_S = N(\mu + S\sigma^2_{\overline{X}}, \sigma^2_{\overline{X}})$, and $h = N(0, R^2/16)$.

of lots I, starting sequence label of the shift n^*, and the p.d.f. f_0. All these inputs are defined for step p.

The in-line data sequence $\{\overline{Z}_n, n = 1, \cdots, I\}$ is first generated with a p.d.f. f_0 for $n < n^*$ and with a p.d.f. f_M for $n \geq n^*$. Then, the sequence-disorder effect is added to $\{\overline{Z}_n\}$ for generating the WAT data sequence $\{\overline{X}_i, i = 1, \cdots, I\}$. Consider a lot at step p with a sequence label n. Let t_n be a random variable and $n + t_n$ be WAT step completion time of the lot. By sorting the WAT step completion time sequence $\{n + t_n\}$ of all lots $\{k + t_k, k = 1, \cdots, I\}$ in an ascending order, the WAT sequence label is obtained as

$$i = \Gamma(n, \{k + t_k, k = 1, \cdots, I\}) \tag{A2}$$

where $\pounds$ is an ascending order operator.

To ensure the sequence-disorder range is within R, the random variable t_n must fall in $((-R/2), (R/2))$. Assume that t_n is truncated normal, i.e., its p.d.f. is

$$g(t_n) = \begin{cases} h(t_n)/\left(\displaystyle\int_{-R/2}^{R/2} h(\theta)\, d\theta\right) & |t_n| \leq R/2 \\ 0 & |t_n| > R/2 \end{cases} \tag{A3}$$

where $h \sim N(0, R^2/16)$. Simulation results show that a WAT data sequence $\{\overline{X}_n\}$ generated according to (A2) and (A3) has the density function of its sequence-disorder grade D_i approximately $N(0, 0.576R^2)$. This approximation can be used to calculate $a_i(n^*)$ in (4).

APPENDIX B

In view of the fact that in (8), each EWMA value A_i is an interpolation of its former value A_{i-1} and the present lot average $\overline{X}_i$, the calculations of $ARL0$ and $ARL1$ are approximated by modeling the SHEWMA scheme as a Markov chain [18]. To be more specific, the transient probability between two Markov states is derived by the p.d.f. of WAT lot average sequence $\{\overline{X}_i\}$ and the SHEWMA parameters (c, λ, h). As de-

rived in (A1), the p.d.f. of $\{\overline{X}_i\}$ is generated by combining an in-control p.d.f. f_0 and an out-of-control p.d.f f_S with a fixed proportion, which is determined by the process conditions (R, M, S). As a result, when a process shift occurs, $ARL1$ is a function of both process conditions (R, M, S) and SHEWMA parameters (c, λ, h). However, when a process is in-control, the p.d.f. of $\{\overline{X}_i\}$ is exactly f_0. Therefore, $ARL0$ is only a function of SHEWMA parameters (c, λ, h).

A. Optimal Designs for Individual Process Conditions

Under a given process condition triplet (R, M, S) and the upper bound of false alarm rate α, the SHEWMA parameter design problem is to

$$\begin{aligned} &\underset{(c,\lambda,h)}{\text{Minimize}}\ ARL1(c,\lambda,h,R,M,S) \\ &\text{subject to } ARL0(c,\lambda,h) \geq \frac{1}{\alpha}. \end{aligned} \tag{B1}$$

Let Φ be the feasible set of parameters, where

$$\Phi \equiv \left\{(c,\lambda,h) | ARL0(c,\lambda,h) \geq \frac{1}{\alpha}\right\}. \tag{B2}$$

To minimize over $ARL1$ over Φ, the search space can be reduced from the set Φ to a set Φ', where

$$\Phi' \equiv \left\{(c,\lambda,h) | ARL0(c,\lambda,h) = \frac{1}{\alpha}\right\}. \tag{B3}$$

As a result, given a set of process condition (R, M, S), the optimal parameters are determined by

$$(c^*,\lambda^*,h^*)_{R,M,S} \equiv \arg \min_{(c,\lambda,h)\in\Phi'} ARL1(c,\lambda,h,R,M,S) \tag{B4}$$

and the corresponding optimal average run length is

$$ARL1^*(R,M,S) \equiv ARL1(c^*,\lambda^*,h^*,R,M,S). \tag{B5}$$

Reasoning of the search space reduction and the solution procedure for (B4) are briefly summarized in Appendix C.

B. Robust Design

The optimal SHEWMA parameters under one set of process conditions may not be optimal under another set of process conditions. Define a metric of maximum delay from the optimal SHEWMA detection by

$$J_{c,\lambda,h} = \max_{(R,M,S)\in\Omega} \Delta ARL1(c,\lambda,h,R,M,S) \tag{B6}$$

where

$$\begin{aligned} \Delta ARL1(c,\lambda,h,R,M,S) \equiv & ARL1(c,\lambda,h,R,M,S) \\ & - ARL1^*(R,M,S) \end{aligned} \tag{B7}$$

is the delay of $ARL1$ when using nonoptimal SHEWMA parameters as compared to using the optimal ones. The goal here is then to choose a robust set of parameters $(\overline{c}, \overline{\lambda}, \overline{h})$ that minimize the worst case detection delay over all process conditions in Ω, i.e.,

$$(\overline{c},\overline{\lambda},\overline{h}) = \arg \min_{(c,\lambda,h)\in\Phi'} J_{c,\lambda,h}. \tag{B8}$$

APPENDIX C

1) Search Space Reduction: Given a set of c and λ, choose h' so that $ARL0(c,\lambda,h') = (1/\alpha)$. The value h' is clearly a function of c and λ and let us denote it as $h'(c,\lambda)$. Intuitively, when c and λ are fixed, the tighter the h, the smaller the values of $ARL0$ and $ARL1$. Therefore, for all $h \geq h'$

$$ARL0(c,\lambda,h) \geq ARL0(c,\lambda,h'(c,\lambda)) = 1/\alpha \tag{C1}$$

and

$$ARL1(c,\lambda,h) \geq ARL1(c,\lambda,h'(c,\lambda)). \tag{C2}$$

To satisfy the requirement of false alarm rate $(ARL0 \geq 1/\alpha)$ and to increase the detection speed (minimize the $ARL1$) at the same time, the value of parameter h should be exactly $h'(c,\lambda)$. Consequently, the search space can be reduced from the set Φ to a set Φ', where

$$\begin{aligned} \Phi' &\equiv \{(c,\lambda,h) | h = h'(c,\lambda)\} \cdot \\ &= \{(c,\lambda,h) | ARL0(c,\lambda,h) = 1/\alpha\}. \end{aligned} \tag{C3}$$

2) Equation (B4) Solution Procedure:

P1) *Determine the range of search space:* Define $ARL0_{SE}(c,\lambda,h)$, $ARL0_E(\lambda,h)$ and $ARL0_S(c)$ as the $ARL0$'s of SHEWMA, EWMA and Shewhart schemes respectively. Let c' be a value at which $ARL0_S(c') = (1/\alpha)$. As the SHEWMA scheme adopts the Shewhart and EWMA control charts simultaneously, its run length is the minimum of the run lengths of Shewhart and EWMA control charts, i.e.,

$$ARL0_{SE}(c',\lambda,h) \leq ARL0_S(c') = 1/\alpha. \tag{C4}$$

Since $ARL0_{SE}(c,\lambda,h)$ is decreasing when c decreases,

$$ARL0_{SE}(c,\lambda,h) < ARL0_{SE}(c',\lambda,h) \tag{C5}$$

for all $c < c'$. It then follows from (C4) and (C5) that to satisfy the requirement of false alarm rate $(ARL0_{SE} = 1/\alpha)$, the feasible value of parameter c should be larger than or equal to c'. By definition of EWMA, the search space of weighting factor is $\lambda \in [0,1]$. Thus the search space in (C3) can be further expressed as

$$\Phi' \equiv \{(c,\lambda,h) | c \geq c', 0 < \lambda \leq 1, h = h'(c,\lambda)\}. \tag{C6}$$

P2) *Quantatize the searching space of* (c,λ): Closed-form design of SHEWMA parameters is not available because of the problem complexity. Therefore, (B4) is solved by numerical evaluation. The search spaces of c and λ are first quantatized into $\{c' + 0.05, c' + 0.1, \cdots, c_b\}$ and $\{0.01, 0.02, \cdots, 1\}$ respectively, where c_b is the upper bound of c and is empirically set as $c' + 1$.

P3) *Approximate the control limit width h' for each pair of (c, λ)*: Although the derivation of a closed-form expression of the function $h'(c, \lambda)$ is formidable, it can be numerically approximated by a bisection procedure [19] for various values of c and λ.

P4) *Search all the quantatized data points:* The solution to (B4) is solved by searching the $ARL1$'s over all quantatized data points $(c, \lambda, h'(c, \lambda))$. If the approximated optimal value of c occurs on or near c_b, add one to c_b and repeat the procedures P2–P4. Although such a brute force search approach takes a lot of time, it only needs to be implemented one time and all the searched data points can be used for robustness analysis in Appendix B.B.

ACKNOWLEDGMENT

The authors collaborated with TSMC colleagues to develop and implement the SHEWMA system. They would like to thank H.-H. Kung, J.-C. You, H.-P. Chen, and J.-H. Lee of Fab 2 for providing all the technical assistance, and J.-J. Wu and H.-C. Tseng of Fab 5 and C.-C. Chou of Fab 3 for their insights into the process integration problems. They would also like to thank W.-C. Kuo and Y.-H. Chan of the Automation Division for their assistance in the implementation of SHEWMA. Most of all, the authors would like to thank managers S.-H. Lin and K.-C. Hsu for both their continuous support of the collaboration project and their visionary guidance. Without the help from these TSMC colleagues, the authors could finish neither the SHEWMA system nor this paper.

REFERENCES

[1] J. H. Lee and K. C. Hsu, private communication.
[2] M. J. B. Bolt, J. Engel, C. L. M. Klauw, and M. Rocchi, "Statistical parameter control for optimum design and manufacturability of VLSI circuits," in *Proc. Int. Semiconductor Manufacturing Science Symp.*, 1990, pp. 99–106.
[3] J. Pak, R. Kittler, and P. Wen, "Advanced methods for analysis of lot-to-lot yield variation," in *Proc. Int. Symp. Semiconductor Manufacturing*, 1997, pp. E17–E20.
[4] N. H. Fiona, D. Montijn, and J. Herman, "Expert system for test structure data interpretation," in *Proc. Microelectronic Test Structures*, vol. 1, 1988, pp. 169–173.
[5] D. M. Hawkins and D. H. Olwell, *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer, 1998.
[6] J. S. Hunter, "The exponentially weighted moving average," *J. Qual. Technol.*, vol. 18, no. 4, pp. 203–210, Oct. 1986.
[7] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: Properties and enhancements," *Technometrics*, vol. 32, pp. 1–12, Feb. 1990.
[8] J. M. Lucas, "Combined Shewhart-CUSUM quality control schemes," *J. Qual. Technol.*, vol. 14, pp. 1–12, Apr. 1982.
[9] D. C. Montgomery, *Introduction to Statistical Quality Control*. New York: Wiley, 1991.
[10] L. S. Nelson, "Control chart for multiple stream processes," *J. Qual. Technol.*, vol. 18, pp. 255–256, Oct. 1986.
[11] R. R. Mortell and G. C. Runger, "Statistical process control of multiple stream processes," *J. Qual. Technol.*, vol. 27, pp. 1–12, Jan. 1995.
[12] W. H. Woodall and E. V. Thomas, "Statistical process control with several components of common cause variability," *IEE Trans.*, vol. 27, pp. 757–764, Dec. 1995.
[13] E. Yashchin, "Monitoring variance components," *Technometrics*, vol. 36, pp. 379–393, Nov. 1994.
[14] K. C. Rose and R. J. M. Does, "Shewhart-type charts in nonstandard situations," *Technometrics*, vol. 37, pp. 15–40, Feb. 1995.
[15] K.-S. Kim and B.-J. Yum, "Control charts for random and fixed components of variation in the case of fixed wafer locations and measurement positions," *IEEE Trans. Semiconduct. Manufact.*, vol. 12, pp. 214–228, May 1999.
[16] SAS Institute Inc., "SAS/STAT User Guide," SAS Institute Inc., 1994.
[17] R. S. Guo, C. K. Tsai, J. H. Lee, and S. C. Chang, "Intelligent process diagnosis based on end-of-line electrical test data," in *Proc. Int. Electronics Manufacturing Technology Symp.*, Oct. 1996, pp. 347–354.
[18] M. S. Saccucci and J. M. Lucas, "Average run lengths for exponentially weighted moving average control schemes using the Markov chain approach," *J. Qual. Technol.*, vol. 22, pp. 154–162, Apr. 1990.
[19] R. L. Burden and J. D. Faires, *Numerical Analysis*. Boston, MA: PWS-Kent, 1989.

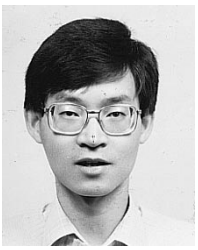
Chih-Min Fan (S'98) received the B.S. degree in control engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 1994, and the M.S. degree in electrical engineering from National Taiwan University (NTU), Taipei, in 1996. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, NTU.

In 1988, he worked as a Summer Intern with the R&D 0.18 μm Logic Department, Taiwan Semiconductor Manufacturing Co., Taipei. His research interests include statistical process monitoring and control, and the development of engineering data analysis system for semiconductor manufacturing.

Ruey-Shan Guo (S'90–M'91) received the B.S. degree from National Taiwan University (NTU), Taipei, Taiwan, R.O.C., in 1983 and M.S. degree from the Massachusetts Institute of Technology (MIT), Cambridge, in 1987. He received the Ph.D. degree in mechanical engineering with a major in manufacturing and a minor in solid state physics in 1991, also from MIT, and the M.B.A. degree from San Jose State University, San Jose, CA, in 1994.

From 1991 to 1995, he was a Senior Process Engineer with National Semiconductor Fairchild Research Center, Santa Clara, CA, where he worked in the area of factory layout design, contamination control, statistical quality control, and computer-aided manufacturing. Since 1995, he has been an Associate Professor in the Graduate Institute of Business Administration and Graduate Institute of Industrial Engineering, National Taiwan University. During his current position, he has many projects with TSMC, mostly in the areas of CIM, WAT quality control, CMP run-to-run process control, supply chain management and foundry business model. He teaches undergraduate and graduate courses in operations management, advanced quality control, enterprise resource planning, and supply chain management.

Dr. Guo served on the program committee and as session chair of ISSM from 1997 to 1999.



Shi-Chung Chang (S'83–M'87) received the B.S.E.E. degree from National Taiwan University (NTU), Taipei, Taiwan, R.O.C., in 1979, and the M.S. and Ph.D. degrees in electrical and systems engineering from the University of Connecticut, Storrs, in 1983 and 1986, respectively.

From 1979 to 1981, he was an Ensign in the Chinese Navy, Taiwan. He worked as a Technical Intern at the Pacific Gas and Electric Co., San Francisco, CA, in the Summer of 1985. During 1987, he was a Member of Technical Staff, Decision Systems Section, Alphatech, Inc., Burlington, MA. He has been with the Electrical Engineering Department, NTU, since 1988 and was promoted to Professor in 1994. His research interests include control and management of complex systems, high-speed networks, optimization theory and algorithms, and distributed decision making. He has been a principal investigator and consultant to many industry and government funded projects in the above areas, and has published more than 100 technical papers.

Dr. Chang is a member of Eta Kappa Nu and Phi Kappa Phi.

Chih-Shih Wei received the B.S.E.E. degree from National Taiwan University, Taipei, Taiwan, R.O.C., in 1979, and the Ph.D. degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 1986.

From 1979 to 1981, he was an Ensign in the Chinese Navy, Taiwan. From 1986 to 1990, he was with Intel, working on salicide, metal, and thin films process technology developments. In 1990, he joined the Taiwan Semiconductor Manufacturing Co. as a Member of Technical Staff and was in charge of the development of EPROM process technology. He then became the Manager of the Integration Department, the Manager of Etch/Lithography Department, and the Deputy Director of Fab 5, all at Taiwan Semiconductor Manufacturing Co. In 1999, he was transferred to the Vanguard International Semiconductor Co. as a Technology Development Division Director and he is now the Director of Fab1.