English

Research Express@NCKU Volume 1 Issue 3 - September 7, 2007
[http://research.ncku.edu.tw/re/articles/e/20070907/2.html]

Developing Virtual Metrology Technology

Fan-Tien Cheng*, Tung-Ho Lin

Institute of Manufacturing Engineering, NCKU
chengft@mail.ncku.edu.tw

Major basis for winning the 2006 NSC outstanding industry-university-cooperation award and to appear in IEEE Trans. Semicond. Manuf., Nov. 2007.

In semiconductor and TFT-LCD manufacturing factory, production equipment needs to be periodically monitored on-line to assure stable workpiece (wafer for semiconductor or glass for TFT-LCD industry) fabrication and high yield rate. In current practice of semiconductor manufacturing, equipment-monitoring is performed by periodically measuring one production wafer that is pre-selected within each lot (also called FOUP in semiconductor or cassette in TFT-LCD industry). But the quality of other production wafers beyond the measuring wafer is unknown. Thus, equipment abnormality may not be discovered in time and many defective production wafers may have been produced. This will result in a great wafer yield loss. A better approach is to apply virtual metrology (VM) technology, which can predict the process quality of every workpiece with process data of production equipment without physically conducting quality measurements. Through VM, the quality of each workpiece can be known immediately right after the process data are obtained to ensure prompt detection of equipment anomaly and avoid defective products.

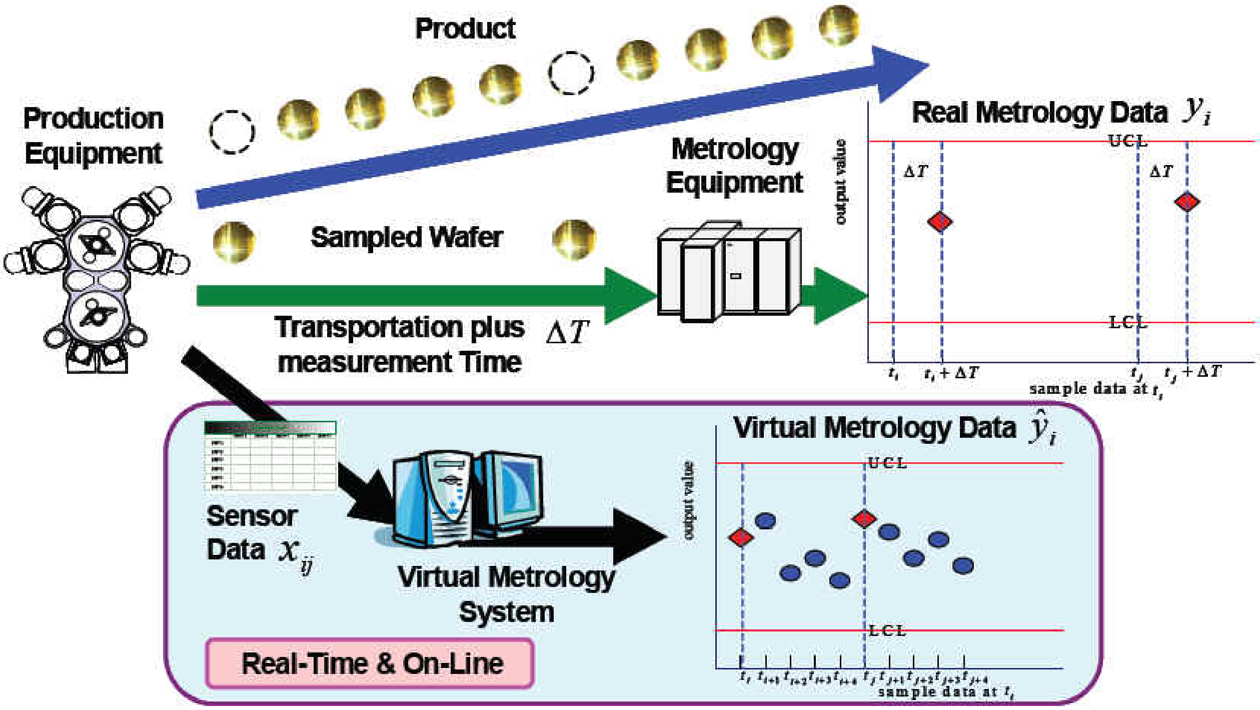

Fig. 1. Comparison between Real Measurement and VM.

Figure 1 depicts the comparison between real measurement and VM. The upper part of the figure is the common approach adopted by most semiconductor factories for wafer quality monitoring. In semiconductor production process, each FOUP contains 25 pieces of wafers. For monitoring the abnormality of production equipment and assuring the quality of production wafers, real measurement is performed only on the pre-selected production wafer within each FOUP. As shown in Fig. 1, t_i and t_j are the completion time of processing the pre-selected monitor wafers, and real measurement data of monitor wafers are obtained at time t_i +△T and t_j +△T. That is, after completing processing of sampled wafers, it still takes time △T (about 6 hours) to obtain their quality data. This approach assumes that the process quality of production equipment is unlikely to be suddenly unusual, and the measurement results of the sampled wafers can fully represent the product quality. However, in the real situation, only the quality of the sampled wafers is grasped. The quality of other production wafers beyond the measuring ones is unknown.

By contrast, through a VM system (VMS) as shown in the lower part of Fig. 1, conjecture values ($\hat{y}$) for not only the sampled wafers but also all the production wafers can be obtained using on-line process data from the production equipment. The real measurement values of the sampled wafers can only be known about six hours after all the wafers of the same lot have been finished processing. During this metrology-delay period, the quality of other production wafers beyond the measuring time is unknown. However, when VMS is adopted to monitor the process, the problems mentioned above can be solved.

This study proposes a dual-phase VM scheme which contains five modules as described below.

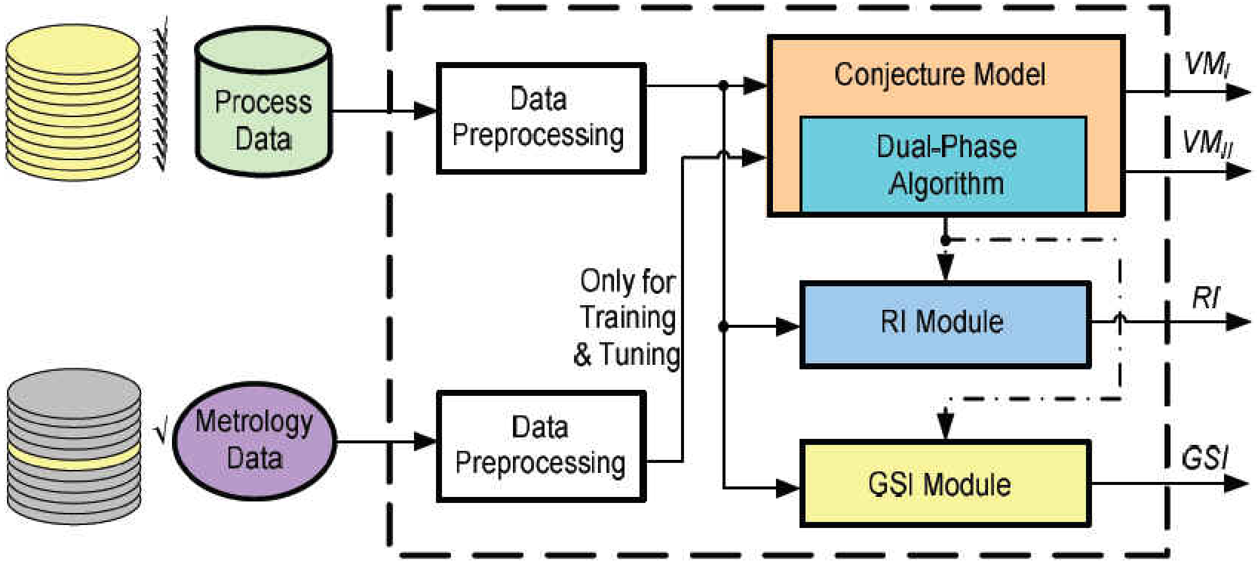

Fig. 2. Schematic Illustration of a Dual-Phase VM Scheme.

■**Data Preprocessing**Before performing VM conjecture value ($\hat{y}$) / reliance index (RI) / global similarity index (GSI), data preprocessing must be conducted to assure data quality and high conjecture accuracy.

■**Conjecture Model**The conjecture model is the kernel of the VM scheme. This work adopts simple recurrent neural network (SRNN) and multiple regression (MR) as the algorithms for establishing the VM conjecture and reference models.

■**Dual-phase Algorithm**The dual-phase algorithm is shown in Fig. 3. Observing Fig. 3, the Phase-I (VM_I) algorithm starts to collect the process data of each processing workpiece after the conjecture model is built. The workpiece's VM_{I} and its accompanying RI and GSI values are computed once the process data collection of the said workpiece is complete. This computation takes about a second only; therefore, promptness is assured.

As shown in Fig. 3, the Phase-II ($\mathrm{VM}_{\mathrm{II}}$) algorithm starts to collect the metrology data of monitor wafers and the pre-specified workpiece in a lot after the conjecture model is built. Correlation between the metrology data and the process data is checked via the workpiece ID once a complete set of metrology data is collected. If correlation check is successful, the set of process data and metrology data with the same workpiece ID will be applied for retraining or tuning the conjecture model. The $\hat{y}$ /RI/GSI models will be updated once they are re-trained or tuned. Finally, $\mathrm{VM}_{\mathrm{II}}$ and its accompanying RI/GSI values of each workpiece in the entire lot are re-computed. After updating the $\hat{y}$ /RI/GSI models in $\mathrm{VM}_{\mathrm{II}}$, these updated models should also be adopted to compute the subsequent VM_{I} and its accompanying RI/GSI values. Therefore, $\mathrm{VM}_{\mathrm{II}}$ is more accurate than VM_{I} .

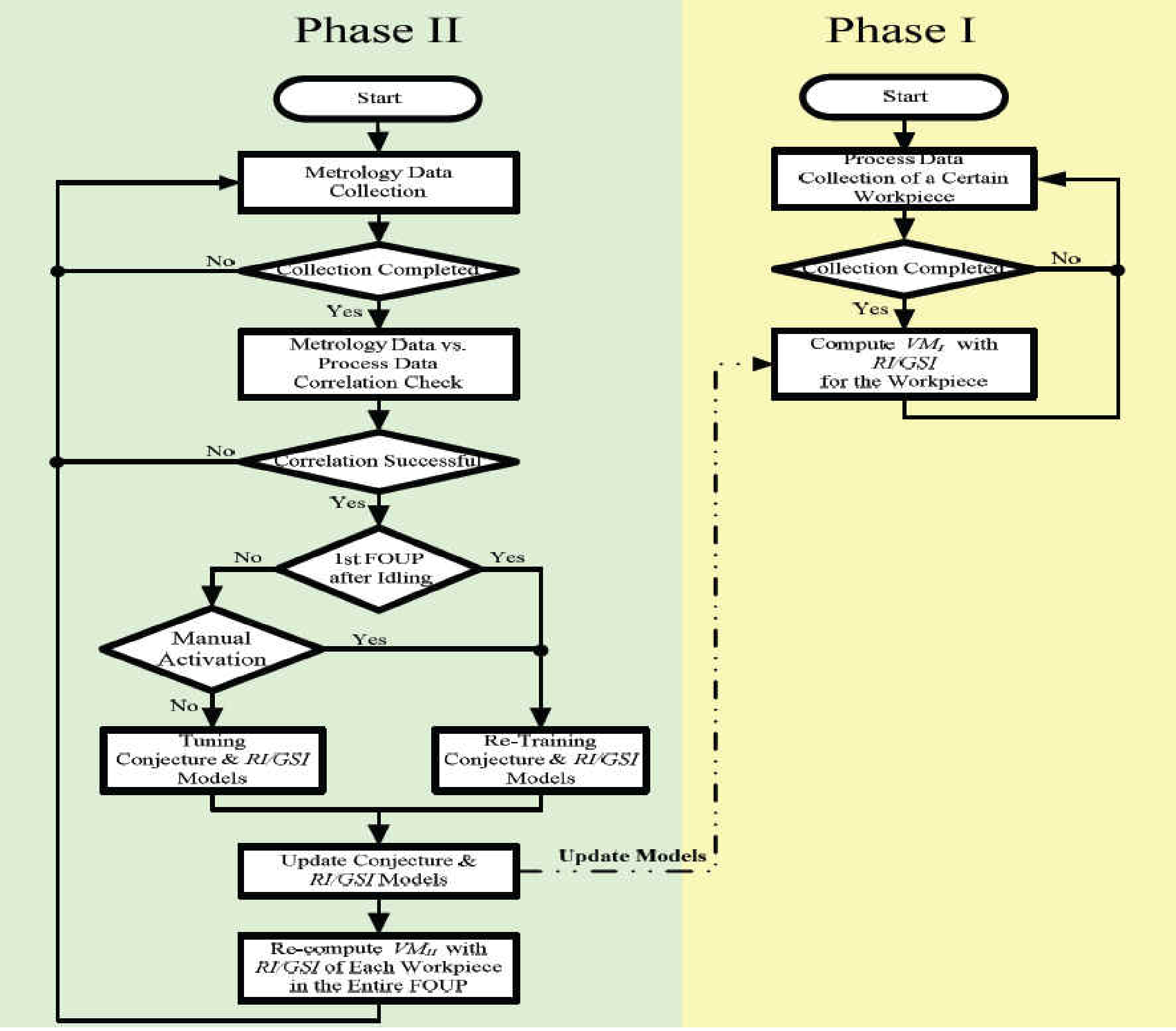
Phase II
Start
Metrology Data Collection
No
Collection Completed
Yes
Metrology Data vs. Process Data Correlation Check
No
Correlation Successful
Yes
No
1st FOUP after Idling
Yes
Manual Activation
Yes
No
Tuning Conjecture & RI/GSI Models
Re-Training Conjecture & RI/GSI Models
Update Conjecture & RI/GSI Models
Update Models
Re-compute VM_{II} with RI/GSI of Each Workpiece in the Entire FOUP
Phase I
Start
Process Data Collection of a Certain Workpiece
Collection Completed
No
Yes
Compute VM_{I} with RI/GSI for the Workpiece

Fig. 3. Dual-Phase VM Conjecture Algorithm.

■**Reliance Index**A RI module generates the VM accompanying RI value to estimate the reliance level of the VM value. The RI value is defined as lying between 0 and 1. The RI of each conjecture value can be derived by calculating the intersection area value of NN and MR normal distribution curves (overlap area A) as shown in Fig. 4. To distinguish how good the RI is, RI threshold value (RI_T) has to be defined. RI_T is defined as the RI value corresponding to the maximal tolerable error limit (E_L). If the RI exceeds the RI_T, then the VM value is reliable; otherwise, its reliability is relatively low, which means the VM conjecture result requires further verification.

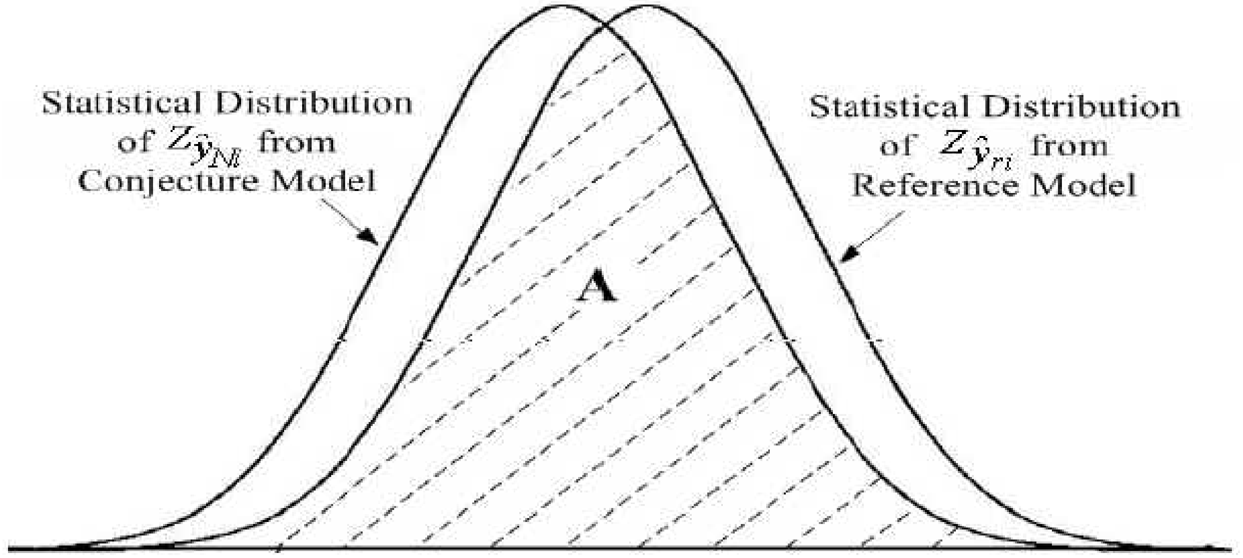
Statistical Distribution of $Z_{\hat{y}_{Ni}}$ from Conjecture Model
Statistical Distribution of $Z_{\hat{y}_{ri}}$ from Reference Model
A

Fig. 4. Statistical Distributions of $Z_{\hat{y}_{Ni}}$ and $Z_{\hat{y}_{ri}}$ for Defining the RI.

■**Similarity Index (SIs)**A SIs module calculates the GSI and ISI values to evaluate the similarity degree of the input-set process data. The GSI is defined as the degree of similarity between the set of process data currently inputted and all the historical sets of process data used in the conjecture model for training and tuning purposes. Moreover, the ISI of an individual process parameter is defined as the degree of similarity between this individual process parameter’s standardized process datum of the input set and the same process parameter’s standardized process data in all the historical sets that are used for training and tuning the predict model. The GSI and ISI values are utilized to assist the RI in gauging the reliance level and identifying the key process parameters that cause major deviation.

The procedure for determining reliance level of the VM conjecture results is shown in Fig. 5.

If both $RI_b > RI_T$ and $GSI_b < GSI_T$ (the subscript “b” stands for each sample set in the conjecturing phase) are true, then a green light is shown. The green light indicates that the NN and MR conjecture results are quite similar, and the degree of similarity between the set of newly entered process data and the sets of historical process data used for model-building is high, confirming strong confidence in the conjecture value.

If $RI_b > RI_T$ is true and $GSI_b < GSI_T$ is false, then a blue light is shown. This implies that, although the VMS provided a conjecture result, some deviations may occur in the process data owing to the high GSI_b. Accordingly, the process data with high ISI values must be examined to prevent excessive confidence.

If $RI_b > RI_T$ is false and $GSI_b < GSI_T$ is true, then a yellow light is displayed. This yellow light means that the conjecture value may be inaccurate. However, since the GSI_b is low, which implies a high degree of similarity, the situation may result from bad MR conjecture.

If both $RI_b > RI_T$ and $GSI_b < GSI_T$ are false, then a red light is displayed. This red light indicates a large deviation between the NN and MR conjecture result. Moreover, the high GSI_b is associated with a low degree of similarity. As such, it is confirmed that the conjecture value is unreliable. In this case, the parameter(s) with the greatest deviation can be identified from the corresponding ISI Pareto chart.

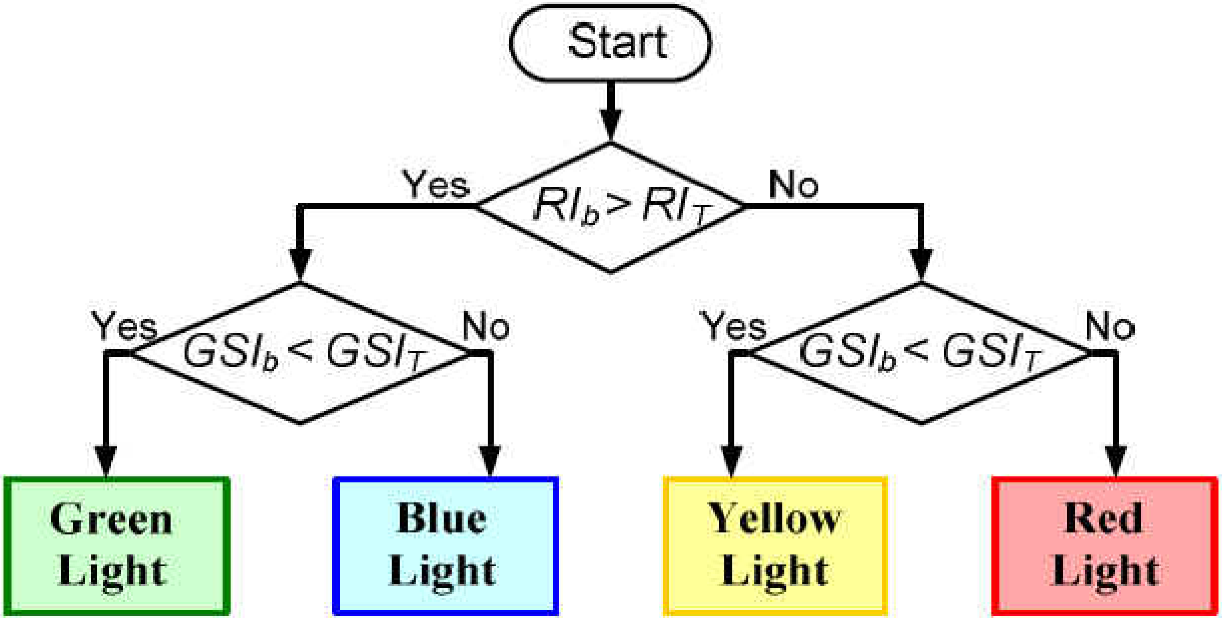

Fig. 5. Flow Chart for Determining Reliance Level of the VM Results.

This study proposes a dual-phase VM scheme. VM_I emphasizes promptness while VM_{II} improves accuracy. This dual-phase VM scheme is well suited for wafer-to-wafer (W2W) advanced-process-control (APC). This scheme may also be applied to eliminate the usage of monitor workpieces and further reduce production cost.

This work also proposes a novel method for evaluating the reliability of a VMS. The proposed method calculates a reliance index (RI) value between 0 and 1 by analyzing the process data of production equipment to determine the reliability of the conjecture results. Besides the RI, the method also proposes SIs. The SIs are defined to assess the degree of similarity between the input set of process data and those historical sets of process data used to establish the conjecture model. The proposed method includes two types of SIs: GSI and ISI. Both GSI and ISI are applied to assist the RI in gauging the reliance level and locating the key parameter(s) that cause major deviation, thus resolving the VMS manufacturability problem.

In semiconductor industry, the wafer size is getting bigger, but the dimension of production process is becoming smaller. In order to maintain the production yield, VM is required for achieving W2W APC. Monitor wafer consumables in a 300mm fab cost about NT$ 100,000,000 per month. If these consumables can be reduced, it will definitely have significant impacts to the semiconductor industry.

This VM technology was successfully transferred to Chi Mei Optoelectronics Corporation (CMO) and Taiwan Semiconductor Manufacturing Company, Ltd (TSMC). After deploying this VM scheme to CMO and TSMC, both of them are quite satisfied with the pilot-run results. Currently, CMO is planning for fab-wide deployment of VMS. This dual-phase VM scheme is ROC, USA, Japan, Korea, and China patents pending.