Overlay Improvements Using a Real Time Machine Learning Algorithm

Emil Schmitt-Weaver*, Michael Kubis, Wolfgang Henke, Daan Slotboom, Tom Hoogenboom, Jan Mulkens, Martyn Coogans, Peter ten Berge, Dick Verkleij, Frank van de Mast

ASML Netherlands B.V., De Run 6501, 5504 DR Veldhoven, the Netherlands

ABSTRACT

While semiconductor manufacturing is moving towards the 14nm node using immersion lithography, the overlay requirements are tightened to below 5nm. Next to improvements in the immersion scanner platform, enhancements in the overlay optimization and process control are needed to enable these low overlay numbers. Whereas conventional overlay control methods address wafer and lot variation autonomously with wafer pre exposure alignment metrology and post exposure overlay metrology, we see a need to reduce these variations by correlating more of the TWINSCAN system's sensor data directly to the post exposure YieldStar metrology in time. In this paper we will present the results of a study on applying a real time control algorithm based on machine learning technology. Machine learning methods use context and TWINSCAN system sensor data paired with post exposure YieldStar metrology to recognize generic behavior and train the control system to anticipate on this generic behavior. Specific for this study, the data concerns immersion scanner context, sensor data and on-wafer measured overlay data. By making the link between the scanner data and the wafer data we are able to establish a real time relationship. The result is an inline controller that accounts for small changes in scanner hardware performance in time while picking up subtle lot to lot and wafer to wafer deviations introduced by wafer processing.

Keywords: machine learning, overlay, real time, process control, inline metrology, systematic error, random error

1. INTRODUCTION

Historically scanner overlay control has relied on feedback from monitor lots to maintain a state of optimum overlay performance. With the introduction of BaseLiner [1,2] for improved scanner stability and matching, a full field TWINSCAN exposure is performed on etched reference wafers followed by a full wafer dense overlay measurement with YieldStar metrology. The goal of the BaseLiner control model is to mitigate over correction, which would induce instability, while at the same time allowing the most amount of correction to be applied toward scanner drift control. To address process instability and drift, ultra small μDBO (Diffraction Based Overlay) targets were introduced to allow for in chip overlay metrology [3]. Combine this with integrated metrology to decrease the turn around time between TWINSCAN exposure and YieldStar metrology [4], the user is given greater control toward maintaining a state of optimum stability while extending the security of catching process excursions. To optimize control we compare the capability of Machine Learning to that of a Weighted Moving Average (WMA), the corrections from which are applied to the product by the TWINSCAN system with "scanner knobs" that allow for high order overlay corrections [5]. When it comes to lot to lot and wafer to wafer control, both BaseLiner and WMA control models rely on the scanner to correct for any external or internal induced instabilities, be that from the fabrication environment, hardware that controls the inner workings of the TWINSCAN system or the pre-exposure wafer alignment metrology performed with the TWINSCAN SMASH sensor [6].

In this paper we propose a novel technique to close the gap between external and internal induced scanner instabilities and the effects they can have on lot to lot and wafer to wafer overlay. We do this by correlating overlay performance in time with TWINSCAN context and sensor data using machine learning algorithms. This can be data from the fabrication environment, performance of parts from within the TWINSCAN system or by picking up on alignment signatures across multiple wavelengths of light, amongst others. By making this connection we are able to pickup on interrelated changes

[*] Correspondence: emil.schmitt.weaver@asml.com

Metrology, Inspection, and Process Control for Microlithography XXVIII, edited by Jason P. Cain, Martha I. Sanchez,
Proc. of SPIE Vol. 9050, 90501S · © 2014 SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2046914

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx

between inputs that contribute to lot to lot and wafer to wafer variation, effectively opening the door to a sub 5nm overlay budget.

2. SOURCES OF OVERLAY ERROR

The primary source of overlay errors can be traced back to exposure, process and metrology (Figure 1). From these the total contribution can be binned into two categories. The first category is systematic error and the second category is random error [7]. Systematic errors come from historically consistent small changes in process or system hardware performance over time. Random errors come from unique and sudden changes to the process or system performance over time. For example a source of random error can come from outside the TWINSCAN system, such as alignment residuals due to process, or from inside the TWINSCAN system, such as the part of the hardware systematic error that cannot be accounted for with a modeled fit.

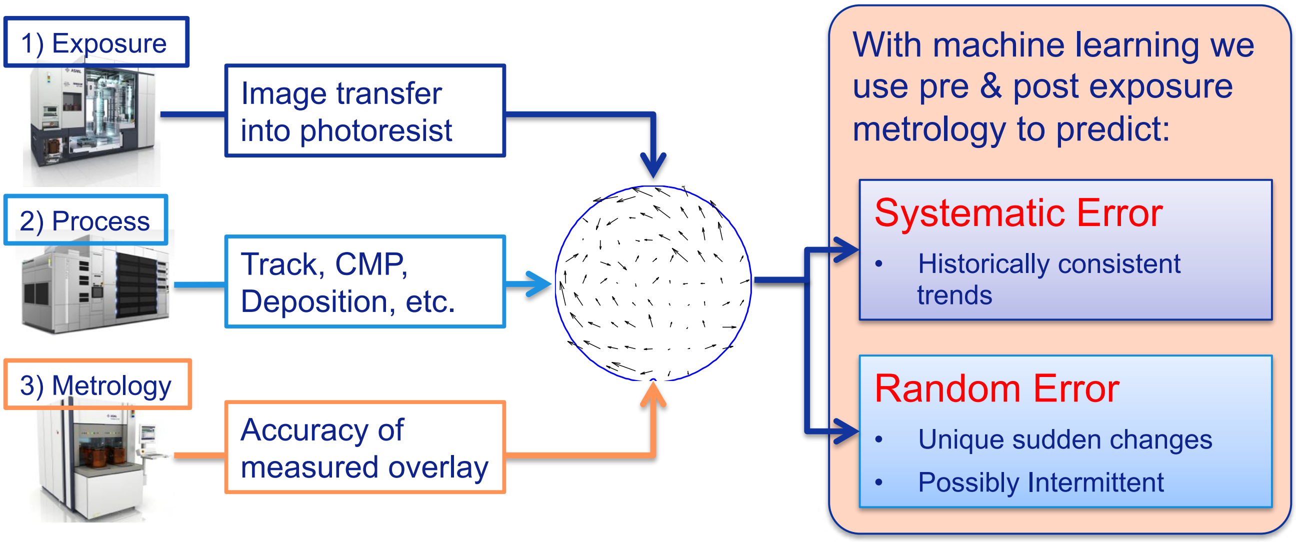

1) Exposure
Image transfer into photoresist
2) Process
Track, CMP, Deposition, etc.
3) Metrology
Accuracy of measured overlay
With machine learning we use pre & post exposure metrology to predict:
Systematic Error
• Historically consistent trends
Random Error
• Unique sudden changes
• Possibly Intermittent

Figure1: the primary sources of overlay errors can be categorized as systematic or random [7].

With machine learning we fit a function (Equation 1) to the measured YieldStar overlay metrology with respect to the context and sensor data of the TWINSCAN system in time. It is known that the context and sensor data that we use as inputs to the trained function contribute to the TWINSCAN system's exposure side overlay error. In other words, the inputs account for known systematic and random errors measured by a YieldStar system as overlay metrology. The output from the function is a predicted estimate of the to be measured overlay. It is to this estimate we fit an appropriate TWINSCAN model [5]. This allows us to account for our predicted estimate of overlay in terms of adjustable "scanner knobs".

3. APPLICATION OF MACHINE LEARNING

3.1 Assembling a Database

To Train (Section 3.2) our machine learning algorithm and Test (Section 3.3) our trained function we need to assemble a database of pre and post exposure data. Pre exposure data is what we know before the wafer is exposed. This can be context data like chuck number and wafer sequence, as well as TWINSCAN metrology such as wafer alignment. Post exposure data is what we only could know after the image is exposed. This can be TWINSCAN system wafer and reticle stage dynamics as well as YieldStar overlay metrology.

Once collected the database is sorted with respect to time and binned into two groups. The first group is labeled Training and the second is labeled Testing (Figure 2). The Training group is used to train the machine learning algorithm and is broken down into two subgroups, points used for training and those used for validation, the reason for which is explained in Section 3.2. The Testing group is used to test the trained function (Section 3.3). The amount of data required, be that lots of wafers over time and sample density per wafer, is strongly dependent upon the type and complexity of the machine learning algorithm to be trained. It is important to note that the diagram in Figure 2 is only a visual example and

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx

is in no way intended to represent the amount of data required to train our machine learning algorithm or test our trained function.

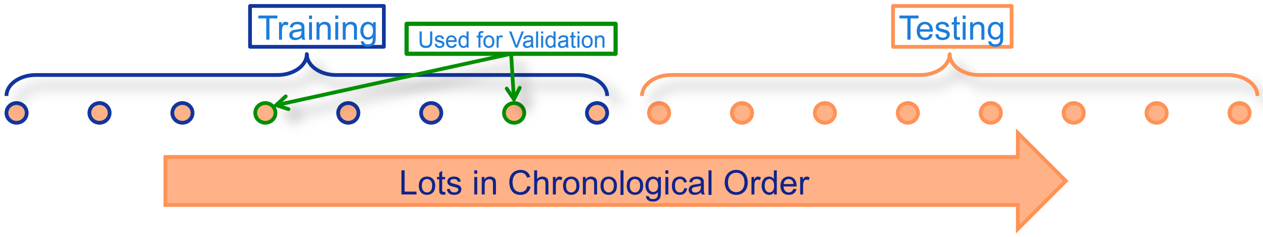


Figure 2: example of data sorted by time and broken into groups for training our machine learning algorithm and testing our trained function.

3.2 Training machine learning overlay controller

Machine Learning is a branch of artificial intelligence that involves the construction and study of systems that can learn from data. There are several approaches to machine learning such as support vector machines (SVM), Neural Networks (NN), *k*-means & principal component analysis (PCA), to name a few [8]. Each approach has its benefits and disadvantages.

In this paper we use a single machine learning approach known as Neural Networks. Specifically we use a time series Nonlinear AutoRegressive neural network with eXogenous inputs (NARX) [9], see Equation 1, to fit a function to the Training dataset (Figure 2).

$$y(t) = f\left(u(t-n_u),\ldots,u(t-1),u(t),y(t-n_y),\ldots,y(t-1)\right) \tag{1}$$

In Equation 1 the input and output signals for the network are $u(t)$ and $y(t)$ respectively at time t (the role of which is to index statistically independent samples), where the next value of the dependent output signal $y(t)$ is regressed on previous values of the output signal and previous values of the independent exogenous input $u(t)$, see Figure 3. While n_u and n_y are the input and output sequence, and f is a nonlinear function that is approximated by Multilayer Perceptron [9].

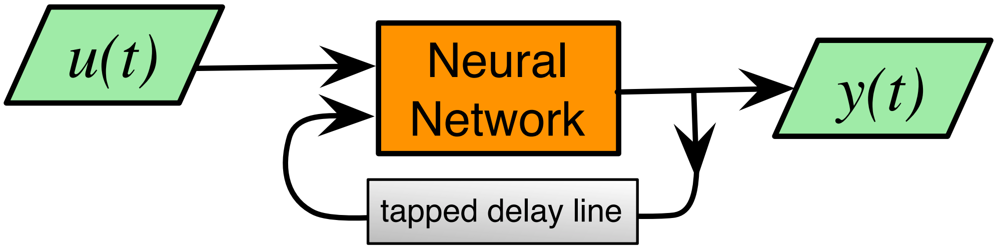


Figure 3: Predict series $y(t)$ given n_y past values of $y(t)$ and another series $u(t)$

While being mindful to the dangers of overfitting the neural network Training dataset, we implement several methods of generalization. Most notably we implement automated regularization with a Bayesian framework [10]. Furthermore we use early stopping techniques by randomly subdividing the Training dataset into two groups of points. Training points used for computing the gradient and updating the network weights and biases, and Validation points used to monitor the training process, as shown in Figure 2. When the Mean Square Error (MSE) and gradient of the training and validation groups reach a minimum (left plot in Figure 4), we trigger the training process to stop [11]. In this case after 7 epochs, an epoch is a measure of the number of times all the training vectors are used once to update the weights. To validate our trained network, we regress between what we measured and what our network predicted with data from the training and validation groups, see middle and right plot in Figure 4.

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx

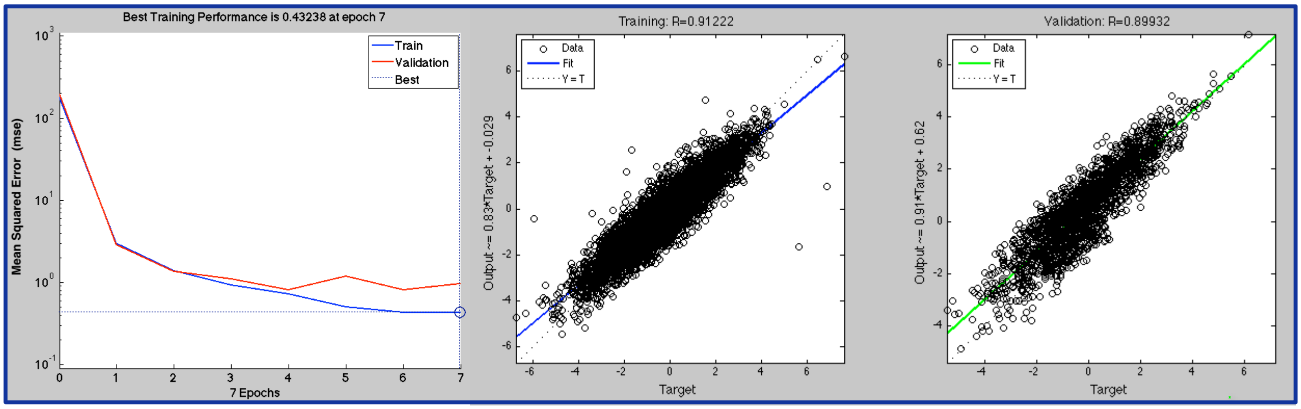


Figure 4: Example of MSE descent and accompanying linear regression plots when training was triggered to stop

In Figure 5 we show a flow diagram for the Training dataset used in our research study. At Step 1 wafers enter the TWINSCAN system. With Step 2 pre exposure context data (chuck number and wafer sequence) and sensor data (pre exposure wafer alignment and post exposure dynamics) come from TWINSCAN sensors. For Step 3 data comes from inline YieldStar overlay metrology. In Step 4 the customer defines what overlay process corrections were applied. Step 5 the wafer exits the track. At Step 6, Steps 2,3 & 4 feed data to a computing platform. It is here that we train the algorithm with current and historical data. At Step 7, the output from Step 6 is a sufficiently trained algorithm. Depending on the design of the network and intended implementation, the frequency of the training can be anything from once a day to once a week or less.

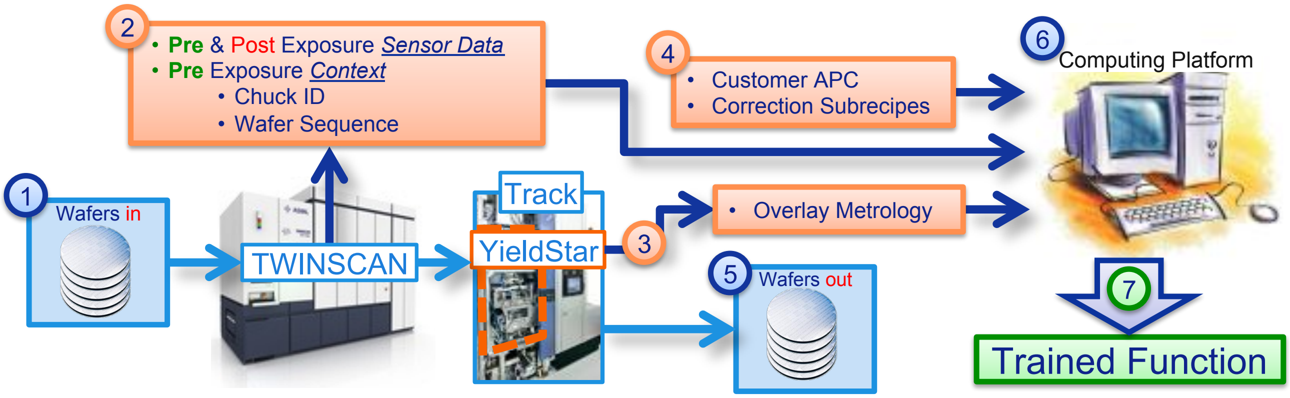


Figure 5: Example of data flow used to train a machine learning algorithm

3.3 Testing machine learning overlay controller

To test the trained function we use data “blind” to that which was used for training. That is the data used for testing was collected after the most recent data point used for training. This is our Testing dataset from Figure 1.

In Figure 6 we show the data flow for testing the trained function with the data from the Testing group. This is the same data flow the function will see in production, so we can conclude that the simulated performance is representative of a real world application. However unlike a real world application, with the Testing dataset we have the actual measured post exposure overlay to compare against. The point to point correlation between simulated and actual overlay will be the metric for benchmarking our trained function’s performance.

Step 1 in Figure 6 shows the wafers entering the TWINSCAN system. In Step 2 the pre exposure TWINSCAN context and metrology is sent to the computing platform at Step 4 for use in the trained function and when deemed necessary, retraining of the function. For Step 3 (a one time action) to satisfy the trained tapped delay lines (TDL), we send historical data from Steps 6 to the trained function on the computing platform. At Step 4 the computing platform uses inputs from Steps 2 and 3 in the trained function to generate an estimate of the overlay signature. In Step 5 an

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx

appropriate TWINSCAN model [6] is fit to the estimated overlay signature and applied to the wafer in real time. At Step 6 the post exposure TWINSCAN metrology is sent to the computing platform for future use in the Trained function and when deemed necessary, retraining of the function. For Step 7 overlay metrology is sent to the computing platform for when it is necessary to retrain the function. At Step 8 the wafers exit the track.

The same way BaseLiner [1,2] needs to know field size and chuck number before it can apply its correction, our trained function needs pre exposure TWINSCAN metrology and context from Step 2 (alignment & chuck number), post exposure metrology from Step 6 (wafer and reticle stage dynamics). Its important to note that information from Steps 6 is historical to satisfy the trained TDL.

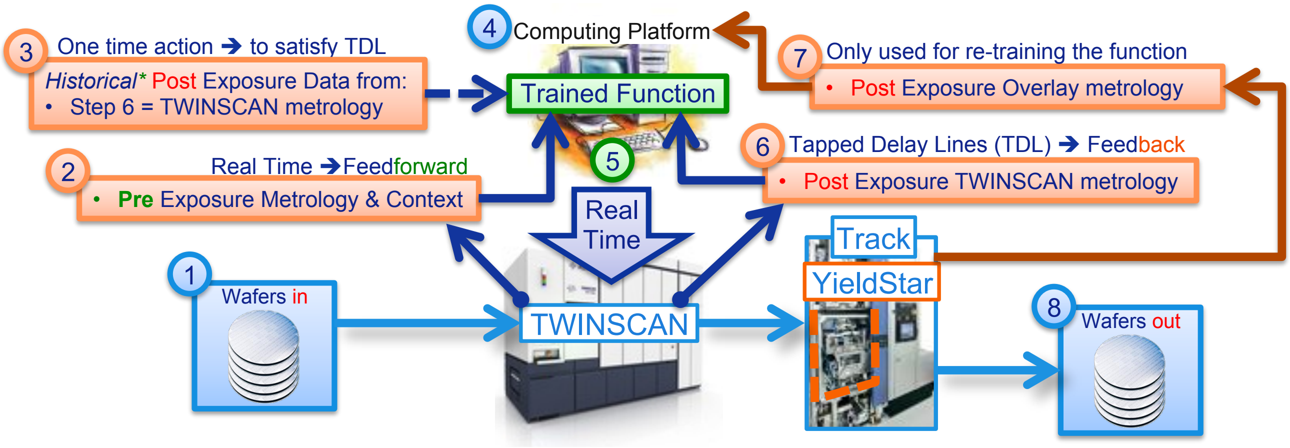

4 Computing Platform
3 One time action → to satisfy TDL
Historical* Post Exposure Data from:
• Step 6 = TWINSCAN metrology
Trained Function
7 Only used for re-training the function
• Post Exposure Overlay metrology
2 Real Time → Feedforward
• Pre Exposure Metrology & Context
5
Real Time
6 Tapped Delay Lines (TDL) → Feedback
• Post Exposure TWINSCAN metrology
Track
YieldStar
1 Wafers in
TWINSCAN
8 Wafers out

Figure 6: Example of data flow into a trained function to test real world performance

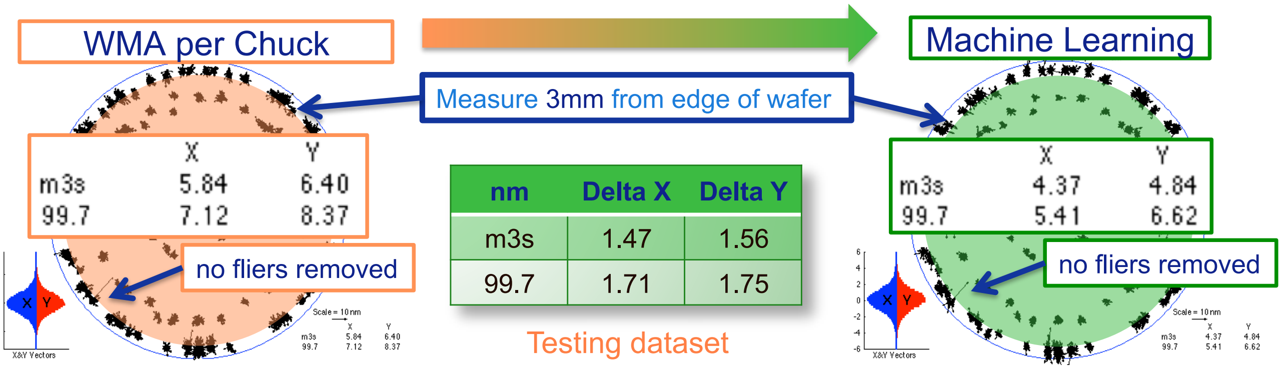
4. RESULTS FROM PRODUCTION LOT CASE STUDY

SK Hynix provided the on product overlay data for our proof book analysis. The front end 3X DRAM layer was exposed on an NXT:1950i TWINSCAN system and measured with an inline T200 YieldStar system. All valid metrology points were used, that is no fliers were removed. In total the study used 40 lots of sparse data, 4 wafers per lot, collected over 5 days. The dataset was broken into two halves, with the first 20 lots binned as the Training dataset and the second 20 lots binned as the Testing dataset.

In Figure 7 we see 40 stacked wafers from the 20 lots used for the Testing dataset defined in Figure 1, controlled with a Weighted Moving Average (WMA) per chuck, compared to the same dataset controlled with our trained function. The WMA corrections are made up of the last three lots, with 50% of the correction coming from the most recent point in time, followed by 30% and 20% sequentially. Both the WMA and Machine Learning corrections were fed forward only after a 30min delay from the time of TWINSCAN exposure. A sample scheme optimization [12] technique was used which allows us to represent the predicted response in terms of a full wafer layout, such that an appropriate TWINSCAN model [5] can be used to feedforward the corrections. To measure as much lot to lot and wafer to wafer variation, the sample scheme intentionally selected points up to 3mm from the edge of the wafer. No fliers were removed from the collected metrology before training the machine learning algorithm and testing the trained function. This was done to test the functions ability to predict overlay error from the inputs without introducing noise due to over fitting.

The result in Figure 7 shows the measured metrology after applying the feedforward TWINSCAN modeled correction. With no fliers removed the mean plus three sigma (m3s) for the Machine Learning trained function improved overlay X by 1.47nm and overlay Y by 1.56nm. This brings overlay X m3s from 5.84nm for WMA to 4.37nm and overlay Y m3s from 6.40nm for WMA to 4.84nm with Machine Learning.

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx



WMA per Chuck

	X	Y
m3s	5.84	6.40
99.7	7.12	8.37

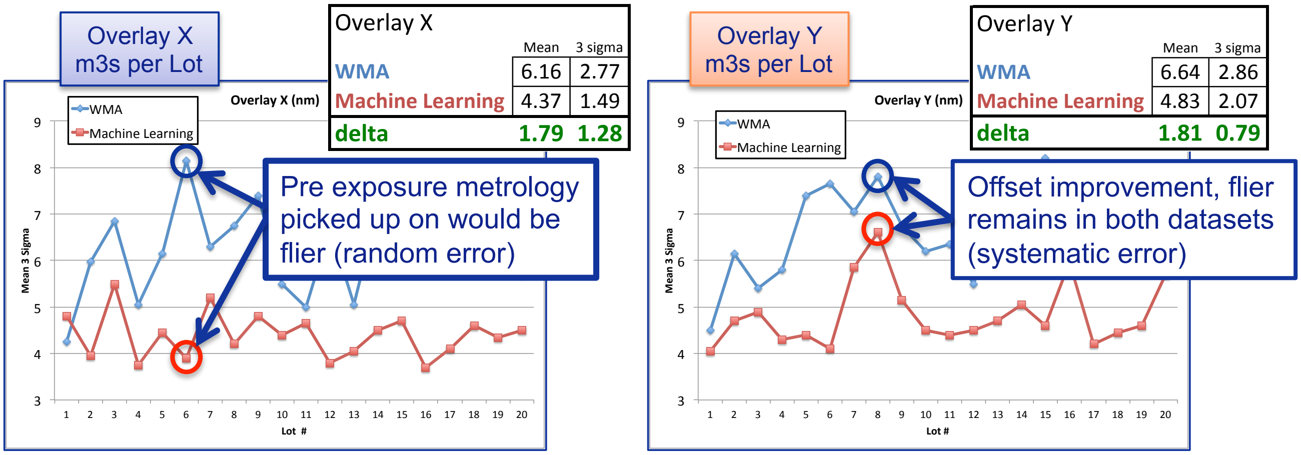
nm	Delta X	Delta Y
m3s	1.47	1.56
99.7	1.71	1.75

Machine Learning

	X	Y
m3s	4.37	4.84
99.7	5.41	6.62

Figure 7: Stacked 40 wafer residual from 20 lot Testing dataset after applying WMA and Machine Learning trained function

The benefit of Machine Learning over WMA comes with the use of pre and post exposure information to pickup on trained lot to lot wafer to wafer contributors to systematic and random overlay error. In Figure 8 we plot the residual overlay from Figure 7 sequentially per lot. The results show an improvement of 1.79nm in mean and 1.28nm in 3 sigma of overlay X and for overlay Y an improvement of 1.81nm in mean and 0.79nm in 3 sigma. Specifically we can look at the points circled for overly X and overlay Y. For overlay X we show a lot that was able to use pre exposure metrology to correct for a flier (random error). For overlay Y we show a lot that has a flier in both WMA and Machine Learning, only showing an improvement in offset (systematic error).



Overlay X

	Mean	3 sigma
WMA	6.16	2.77
Machine Learning	4.37	1.49
delta	1.79	1.28

Overlay Y

	Mean	3 sigma
WMA	6.64	2.86
Machine Learning	4.83	2.07
delta	1.81	0.79

Figure 8: Residual overlay from Figure 7 plotted sequentially per lot

5. SUMMARY AND CONCLUSION

Machine Learning techniques have the potential to close the gap between external and internal induced scanner instabilities and the effects they can have on lot to lot wafer to wafer overlay. For this paper we use a time series Nonlinear AutoRegressive neural network with eXogenous inputs (NARX) to build a correlation between context and TWINSCAN metrology to that of YieldStar metrology. The results of which improved systematic and random overlay error by roughly 1.5nm m3s over existing methods, thereby opening the door to sub 5nm on product overlay performance. Future work will look at how Machine Learning can be integrated into ASML Application activities in close collaboration with our customers.

6. ACKNOWLEDGEMENTS

The authors would like to thank SK Hynix customers CM Lim, HG Lee & SJ Han for there help in obtaining the on product overlay data, as well as ASML colleagues Christophe Fouquet, Alek Chen, Henry Megens, Kevin Ryan, Thomas Theeuwes, Paul Luehrmann & coworkers in the Holistic Applications Development group for there assistance and feedback.

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx

REFERENCES

[1] Busch, J., Parge, A., Seltmann, R., Scholtz, H., Schultz, B., Knappe, U., Ruhm, M., Noot, .M, Woischke, D., Luehrmann, P., “Improving lithographic performance for 32 nm.” Proc. SPIE 7638, Metrology, Inspection, and Process Control for Microlithography XXIV, 763805 (April 01, 2010); doi:10.1117/12.848613; http://dx.doi.org/10.1117/12.848613

[2] Vanoppen, P., Theeuwes, T., Megens H., Cramer, H., Fliervoet, T., et al, “Lithographic scanner stability improvements through advanced metrology and control,” Proc. SPIE 7640, Optical Microlithography XXIII, 764010 (March 10, 2010); doi:10.1117/12.848200; http://dx.doi.org/10.1117/12.848200

[3] Bhattacharyya, K., Ke, C.-M., Huang, G.-T., Chen, K.-H., et al, “On-product overlay enhancement using advanced litho-cluster control based on integrated metrology, ultra-small DBO targets and novel corrections,” Proc. SPIE 8681, Metrology, Inspection, and Process Control for Microlithography XXVII, 868104 (April 10, 2013); doi:10.1117/12.2011878; http://dx.doi.org/10.1117/12.2011878

[4] Chen, K.-H., Huang, J., Yang, W.-T., Ke, C.-M., Ku, Y.-C., et al, “Litho process control via optimum metrology sampling while providing cycle time reduction and faster metrology-to-litho turn around time,” Proc. SPIE 7971, Metrology, Inspection, and Process Control for Microlithography XXV, 797105 (March 28, 2011); doi:10.1117/12.879218; http://dx.doi.org/10.1117/12.879218

[5] Mulkens, J., Kubis, M., Hinnen, P., Graaf, R., Laan, H., et al, “High order field-to-field corrections for imaging and overlay to achieve sub 20-nm lithography requirements,” Proc. SPIE 8683, Optical Microlithography XXVI, 86831J (April 12, 2013); doi:10.1117/12.2011550; http://dx.doi.org/10.1117/12.2011550

[6] Hinnen, P., Depre, J., Tanaka, S., Lim, S.-Y., Brioso, O., et al, “Integration of a new alignment sensor for advanced technology nodes,” Proc. SPIE 6520, Optical Microlithography XX, 652023 (March 26, 2007); doi:10.1117/12.712084; http://dx.doi.org/10.1117/12.712084

[7] Ham, B.-H., Yun, S., Kwak, M.-C., Ha, S. M., Kim, C.-H., Nam, S.-W., “New analytical algorithm for overlay accuracy,” Proc. SPIE 8324, Metrology, Inspection, and Process Control for Microlithography XXVI, 83240A (March 29, 2012); doi:10.1117/12.918002. http://dx.doi.org/10.1117/12.918002

[8] Alpaydin, E., [Introduction to Machine Learning], The MIT Press, Cambridge, Massachusetts, chapter 1, (2010), ISBN 978-0-262-01243-0.

[9] Siegelmann, H.T., Horne, B.G., Giles, C.L., "Computational capabilities of recurrent NARX neural networks," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, Vol.27, No.2, 208-215, (Apr 1997). doi:10.1109/3477.558801; http://dx.doi.org/10.1109/3477.558801

[10] MacKay, D.J.C., “Bayesian interpolation,” Neural Computation, Vol. 4, No. 3, 415–447 (May 1992). doi:10.1162/neco.1992.4.3.415; http://dx.doi.org/10.1162/neco.1992.4.3.415

[11] Prechelt, L., “Early Stopping - But When?,” Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, Vol. 7700, 53-67 (2012). ISBN 978-3-642-35289-8, doi.10.1007/978-3-642-35289-8_5; http://dx.doi.org/10.1007/978-3-642-35289-8_5

[12] Chiu, C.-F., Huang, C.-Y., Shieh, J., Chiou, T.-B., Li, A., Shih, C.-L., Chen, A., "Impacts of overlay correction model and metrology sampling scheme on device yield", Proc. SPIE 8324, Metrology, Inspection, and Process Control for Microlithography XXVI, 83241S (March 29, 2012); doi:10.1117/12.916601; http://dx.doi.org/10.1117/12.916601

Downloaded From: http://proceedings.spiedigitallibrary.org/ on 12/22/2015 Terms of Use: http://spiedigitallibrary.org/ss/TermsOfUse.aspx