Lithography Process Model Building Using Locally Linear Embedding

Pardeep Kumar[a], Alan E. Rosenbluth[b], Babji Srinivasan[a], Ramya Viswanathan[c], and Nihar R. Mohapatra[a]

[a]*Indian Institute of Technology-Gandhinagar, VGEC Campus, Chandkheda, Ahmedabad, Gujarat, India, nihar@iitgn.ac.in*
[b]*IBM T.J. Watson Research Ctr., Yorktown Heights, NY*
[c]*IBM East Fishkill, Hopewell Junction, NY*

***Abstract*—Practical models of lithographic processes are usually empirically calibrated, making their accuracy dependent on the total number of samples used to build the models, and more specifically on the selection of a representative set of samples for calibration. An inadequate number of samples can adversely impact model accuracy, but a broadly comprehensive set will excessively increase measurement cost. Lithography process models based on samples which are picked uniformly from populated regions of the original pattern space and are truly a representative set will improve model prediction accuracy, as is highly desirable for model based optical proximity correction (OPC) simulations. We propose a robust approach for sample plan selection for lithography process model building using locally linear embedding (LLE). The effectiveness of the proposed method is verified by simulating some critical layers in 14-nm and 22-nm complementary metal oxide semiconductor (CMOS) technology nodes. Experimental results show that without compromising model accuracy, LLE can provide a competitive representative sample plan selection in a single shot, in comparison with hundreds of random cross-validation experiments as an alternative.**

Keywords—Lithography process model, OPC, LLE, CM1 model.

INTRODUCTION

Lithography simulation is an indispensable but compute intensive task in the process flow of sub-wavelength semiconductor manufacturing. The success of lithography simulation depends upon the performance of the process models, and efficient process models are of paramount importance for large mask designs. The problem of intractable computational time that would arise from use of truly physical models can be avoided by using empirically calibrated phenomenological models, such as the CM1 class of models from Mentor Graphics [1], which is used as an example in the present work. Although empirically calibrated models like CM1 are widely used in the industry, they must be accurately calibrated against experimental data, and the accuracy of these models relies on the representative sample plan selection used for model calibration. The importance of the number of data points needed for building an empirical model is illustrated in Fig. 1. Simple printing behavior can be captured with fewer data points. Fig. 1(b) shows schematically that a new model can be built with reduced data set and acceptable degree of accuracy in comparison with true model (shown in Fig. 1(a)) if redundant data is used. However, excessively frugal data collection for model building is also not a good practice because it will not be

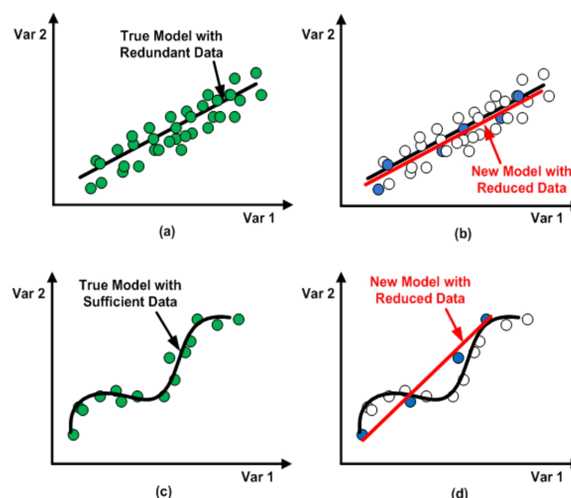

Fig. 1: Effect of number of data points when building an empirical model. Green solid points are data points for true model, white circles (blue solid points) are the discarded (required) data points for new model. Fig. 1(a) and Fig. 1(b) represent the case when process behavior is linear whereas Fig. 1(c) and Fig. 1(d) represent the scenario of a complex process behavior.

able to capture the overall process behavior and will lead to an incorrect model as shown in Fig. 1(d). Fig. 1(c) shows the true model for a complex process behavior with sufficient data points. Hence, selection of a sufficient number of data points for model building plays a key role in model performance.

Another important consideration is how to select a representative set of samples for model calibration. Lithography process models used for full-chip OPC simulation should be robust enough to accurately fit features that span a wide variety of shapes and sizes. If the sample plan is limited to only a few feature shapes and sizes, the model will generally capture well the printing behavior for those geometries, but accuracy is likely to be poor for geometries that differ from this narrow calibration set. Thousands of samples are required for building a process model to support a large variety of designs. Selecting such a large number of samples by manually inspecting the individual samples is a tedious job and not good practice. Repeated trials across multiple splits can be used to assess model performance on out-of-training patterns [2]. Different CM1 models can be built based on randomly selected sample sets and a robust model can be selected out of these models. However, the randomization

978-1-4673-7860-4/15/$31.00 ©2015 IEEE

approach with a large number of trials does not generate an objectively-defined set of representative samples for model calibration, and simulation time will also increase.

We show here that Locally Linear Embedding (LLE) algorithm can play an important role in sample selection for empirical process model building. LLE extracts information about potentially encountered patterns by deriving mutual similarity information across a highly oversampled pool of candidate patterns (which need not be printed or measured until after LLE-based down-selection) in a high dimensional feature space, and then mapping these potential samples to low dimensional space in a way that preserves the original local structure of the samples as closely as possible. From the low dimensional space the representative samples can be selected in one shot, which helps save computational time.

CM1 Models

CM1 models are a class of compact resist models used for full chip lithography simulations. CM1 models use two dimensional aerial images of the mask features as input and perform (during calibration) an optimization of resist model parameters to obtain model functions which provide two dimensional "resist response profiles" and an optimum threshold value (T). Note that "resist response profile" does not refer here to a prediction of the developed resist relief surface, but rather to an abstract response surface resembling a level set function whose contour at value T predicts the perimeter of the printed feature. During use the constant threshold T is then applied on the resist response profile calculated for mask layout features, in order to predict the critical dimension (CD) values on silicon wafers. The two dimensional resist response surface on the silicon wafer, R(x,y), is a linear combination of different modeling terms (M_i). The general form of these compact models can be defined as follows [1]:

$$R(x,y) = T \quad \text{at the print contour} \tag{1}$$

$$R(x,y) \equiv \sum_i c_i M_i(x,y) \tag{2}$$

$$M(x,y) \equiv \left[\left(\nabla^k I_{\mp b}(x,y)\right)^n \otimes G_{s,p}(x,y)\right]^{1/n} \tag{3}$$

Here, c_i is the coefficient value for i^{th} modeling term (M_i), I(x,y) is the aerial image of the mask pattern, k is the order of differentiation, b is a neutralization cutoff constant, G(x,y) is a Gauss-Laguerre function kernel, p is the kernel order, s is a diffusion length, and ⊗ is the convolution operator. The positive and negative signed model terms for non-zero b are intended to phenomenologically represent acid and base neutralization respectively, and the sign choice indicates whether the M function output is truncated from above or below at cutoff level b. As is clear from (3), a large number of model terms can be obtained from different combinations of the parameter values of k and n, so that a CM1 model can take on $2^N - 1$ different model forms, where N is the number of terms [3]. CM1 is calibrated using commercial code (from Mentor Graphics) that employs various search algorithms to explore s and b parameter choices and then find the coefficient values such that the constant threshold (T) will give a minimum difference between measured CD and simulated CD. The iterative process of the search algorithm continues until the required level of accuracy is achieved or the maximum defined limit for iterations is reached.

LLE Algorithm

LLE is a dimension reduction technique that attempts to discover nonlinear structure in high dimensional data by exploiting the local invariance symmetries of derived linear reconstructions [4]. The basic idea of the LLE algorithm is shown in Fig. 2. First, it finds the neighborhood of each data point in the original high dimensional space and represents the local neighborhood in the form of a weight matrix by minimizing the cost function given in (4). This minimization of cost function is subject to two constraints given in (5) and (6). It then constructs the low dimensional embedding of the data based on the computed weights by minimizing the cost function given in (7).

$$\varepsilon(W) = \sum_i |X_i - \sum_j W_{ij} X_j|^2 \tag{4}$$

$$W_{ij} = 0 \quad \text{if } X_j \text{ is not a neighbor of } X_i \tag{5}$$

$$\sum_j W_{ij} = 1 \tag{6}$$

$$\Phi(Y) = \sum_i |Y_i - \sum_j W_{ij} X Y_j|^2 \tag{7}$$

where the X_i are real-valued input vectors with dimension D, W_{ij} is the derived weight matrix, and Y_i is the output vector of dimension d (where d < D). The weights W_{ij} represents the contribution of the *j*th data point to the *i*th reconstruction. The constrained weights for any particular data point are invariant to

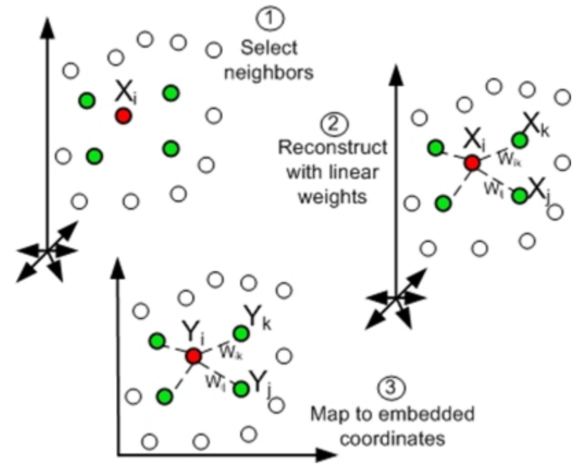

Fig. 2: Basic idea of the LLE algorithm [4].

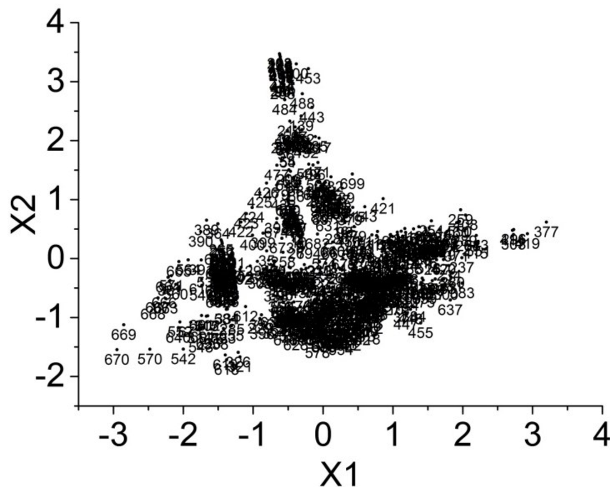

Fig. 3: LLE output for a 14nm candidate sample set. Samples of similar shapes and sizes are grouped together.

rotations, rescalings, and translations of that data point and its neighbors. LLE maps its inputs into a single coordinate system of lower dimensionality by exploiting adjacency information about closely located data points in the form of weight matrix and the optimization process does not involve local minima. It avoids solving nonlinear optimization equations and is based on sparse matrices algebra.

PROPOSED APPROACH

In this paper, we propose a systematic approach of sample plan selection for lithography model building using LLE algorithm. Individual samples are first represented by their associated set of CM1 model term values. LLE algorithm is then used to map this high dimensional data to lower dimension. An example two-dimensional output representation of a set of 701 samples from a 14nm CMOS technology node is shown in Fig. 3. Finding the representative samples in the original high dimensional space is a difficult task as it is very hard to visualize them with dimension greater than three. This two dimensional map approximately captures the distribution of samples in the original high dimensional space. Candidate samples of similar shapes and sizes are grouped together, and it becomes much easier to uniformly select the representative set of samples due the low dimensionality of the output mapping of candidate samples. For CM1 model building, calibration and verification samples can be selected from this two dimensional graph by using appropriate gridding schemes.

There can be other data analytics techniques such as clustering algorithms that can also be used to categorize the samples into different bins [5, 6] but these algorithms are often limited to grouping the samples into K clusters and are helpful if separate process models are required for each cluster. Classical techniques for dimensional reduction, such as principal components analysis (PCA) or multidimensional scaling (MDS), often fail when nonlinear structure cannot simply be regarded as a perturbation from a linear approximation. The principal advantage of using LLE is that it attempts to capture the true potentially nonlinear structure in high dimensional space and preserve that original neighborhood information in a low dimensioned embedded structure. Moreover, LLE can be employed without needing measured results for the candidate patterns. The proposed approach based on LLE algorithm gives a robust solution to select the truly representative set of samples for different mask levels in advanced CMOS technology nodes.

SIMULATION RESULTS

The proposed method is verified by simulating example critical layers in 14nm and 22nm CMOS technology nodes. As a control, LLE is compared against random selection of calibration data via cross-validation trials that are carried out within a larger set chosen by traditional engineering criteria. As usual, the mean model performance in such cross-validation trials lets us estimate the impact of restricting the number of calibration patterns (as is necessary due to the cost of accurate metrology) if the choice of calibration patterns is made in a random (unbiased) and unsystematic way. Figs. 4 and 5 show CM1 model accuracy results for 100 random models over 14nm and 22nm data sets, respectively. Overfitting is a major concern with empirical models. An empirical model should not fail over the verification set of samples. In Fig. 5, the verification error RMS for few of model numbers are approximately twice the calibration error RMS value. It represents the fact that for these models the candidate samples for model building were not the good representative set. To provide a particularly stringent benchmark, we also include the performance of the best model from the set of 100, as determined with post factum knowledge of the verification outcomes, yielding a reference that represents an exceptional, statistically extreme level of performance within the set of cross-validation splits. Table 1 shows the simulation time taken by CM1 models for 14nm and 22nm data sets. LLE based model need one simulation run and it saves the simulation time up to 100x compared to model selection from random experiment. Fig. 6 shows accuracies achieved for the 22nm data set, comparing the LLE based CM1 model against best and average outcomes from 100 random models, and against a CM1 model in which representative samples were chosen by manual engineering judgment. In term of accuracy, LLE based model shows best results for verification set as compared to other approaches. Table 2 summarizes the error RMS results of CM1 model, for both the 14nm and 22nm data sets. For both data sets the LLE based model gives better performance in terms of accuracy as considered in the cross-validation approach.

TABLE 1: CM1 SIMULATION TIME FOR 14NM AND 22NM DATA.

	14nm Calibration Model Time (s)	22nm Calibration Model Time (s)
Typical random model	3784	1808
Model selection from random experiment	≈100×3784	≈100×1808
Manual sampling based model	3424	1858
LLE based model	3892	1760

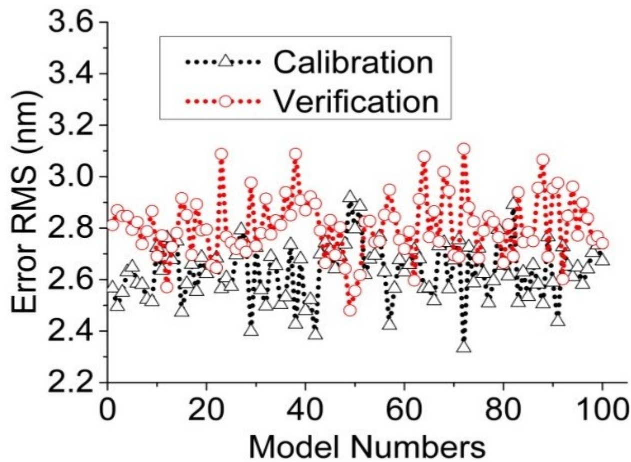

Fig. 4: Accuracy results of 100 random CM1 models for 14nm sample set. Vertical axis is the RMS error against measurements in each model's predictions of the printed widths of a large number of patterns. Black triangles are the error in predicting the printed widths with which the model is calibrated, and red circles are the error in predicted dimensions of set-aside verification patterns.

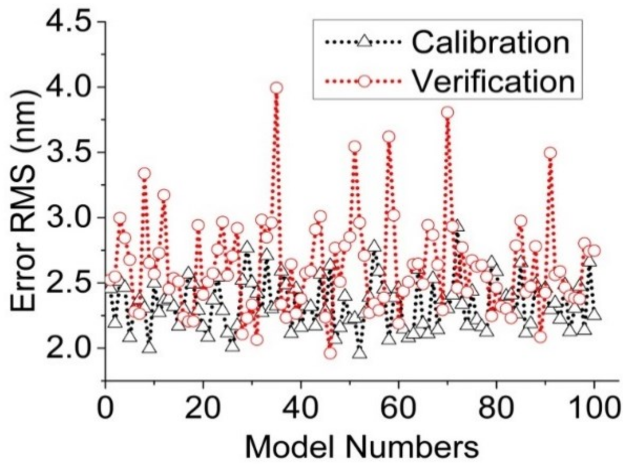

Fig. 5: Accuracy results of 100 random CM1 models for 22nm sample set. (See Fig. 4 caption.)

TABLE 2: CM1 SIMULATION ERROR RMS (NM) RESULTS.

	14nm data		22nm data	
	Calibration	Verification	Calibration	Verification
A	2.68	2.65	2.28	2.24
B	2.63	2.81	2.34	2.62
C	2.82	2.61	2.43	2.31
D	2.75	2.71	2.60	2.02

A = Cross-validation best, B = Cross-validation average,
C = Manual sampling, and D = LLE based sampling

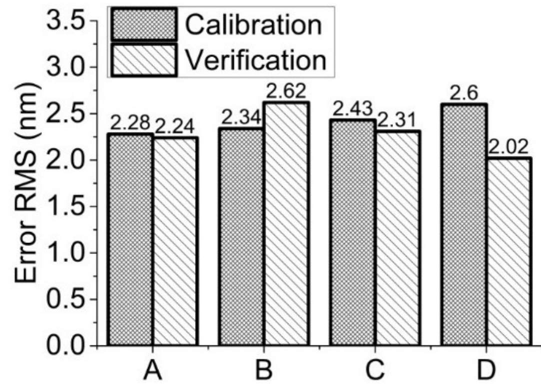

Fig. 6: Comparison of simulation results for 22nm sample set. A = Cross-validation best, B = Cross-validation average, C = Manual sampling, and D = LLE based sampling.

CONCLUSION

LLE-based results achieve the best performance in terms of verification accuracy of all the approaches considered. Our simulation results (summarized in Table 1 and Table 2) indicate that without compromising model accuracy, LLE can be used to select a highly competitive representative sample plan in a single shot, bypassing hundreds of random cross-validation experiments and therefore saves the computational time. The proposed method also has the benefit of being a deterministic approach that avoids stochastic uncertainty in representative sample plan selection, providing a more systematically assured pattern coverage.

REFERENCES

[1] Yuri Granik, Dmitry Medvedev, and Nick Cobb, “Towards standard process models for OPC”, Proceedings of SPIE, Vol. 6520, 652043, 2007.

[2] Chris Mack, “Improved methods for lithography model calibration”, Proceedings of SPIE, Vol. 6607, 66071D, 2007.

[3] Dmitry Vengertsev et al., “The new test pattern selection method for OPC model calibration, based on the process of clustering in a hybrid space”, Proceedings of SPIE Vol. 8522, 85221A-1, 2012.

[4] Sam T. Roweis and Lawrence K. Saul, “Nonlinear dimensionality reduction by locally linear embedding”, Science 290, 2323, 2000.

[5] Pardeep Kumar, Samit Barai, Babji Srinivasan, and Nihar R. Mohapatra, “Process model accuracy enhancement using cluster based approach”, *Physics of Semiconductor Devices*, pp 33-36, Springer International, 2014.

[6] Pardeep Kumar, Babji Srinivasan, and Nihar R. Mohapatra., “Fast and accurate lithography simulation using cluster analysis in resist model building”, *J. Micro/Nanolith. MEMS MOEMS* 14(2), 023506, 2015.