A Method for Die-Scale Simulation of CMP Planarization

Thye-Lai Tung
TCAD, RN2-40
Intel Corporation
2200 Mission College Blvd.
Santa Clara, CA 95052

Abstract -- Chemical-Mechanical Polishing (CMP) is well known for its planarization capability. However, it suffers from long-range non-uniformity due to its sensitivity to pattern density. This paper shows that, by using basic building blocks and formulation techniques, CMP simulation can be done on a large dimension, namely the whole die.

I. INTRODUCTION

Chemical-Mechanical Polishing (CMP) has emerged as an indispensible process for achieving a high degree of global planarization in multi-level interconnects. It is naturally sensitive to layout density variations because it has to remove different amounts of oxide. Consequently, thickness non-uniformity is an issue. The long-range planarization distance that makes CMP so attractive also makes it difficult to predict what is going to happen for a given chip layout.

There are many issues involves in CMP, such as the slurry and rotation speed. However, as far as planarization is concerned, the most important factor is the mechanical interaction between the polishing pad and wafer surface. Some models have been proposed to analyze the problem [1-4], but they are restricted in many ways. This paper reports a numerical technique that models CMP planarization with a realistic die-level approach.

The three key steps in this method are: capturing the shape of as-deposited Interlayer Dielectric (ILD) oxide, condensing cross-sectional density data, and simulating polishing, as described in the following sections.

II. LAYOUT DATA SIZING AND SAMPLING

The importance of capturing the protrusion profile of as-deposited oxide is illustrated in Fig. 1. Two types of features are presented. Both have 50% "features density," but the one on the left has finer pitch and acquires more oxide. Therefore the remaining oxide is thicker remaining oxide after the polish step. The mechanism contributing to the discrepancy is the sidewall formation. Depending on the oxide deposition technique, sidewall coverage may range from 60% to 100% of the top coverage.

In our approach, a 3-D profile is actually constructed. This is the most time consuming step in the simulation process. Cutlines are taken at the top, middle, and the base of the protrusion to obtain representative cross-sections, as shown in Fig. 2. From there, intermediate values are interpolated. Below the base, it is 100% bulk oxide; therefore no data are needed.

Note that the three cross-sections are nothing but "sizing" the layout at different values. These values are a function of the oxide type, thickness and the metal height. In sizing up the layout data, the problem of eliminating overlaps and fusing structures must be addressed, as shown in Fig. 3. This problem is particularly serious in densely packed areas such memory cells.

Given layout data are unconstrained in shapes and sizes, it is impractical to use them as is in polish simulation. They are sampled and condensed into density numbers. Typically the size of the sampling window is in the order of 100µ by 100µ. Note that it is normal for the

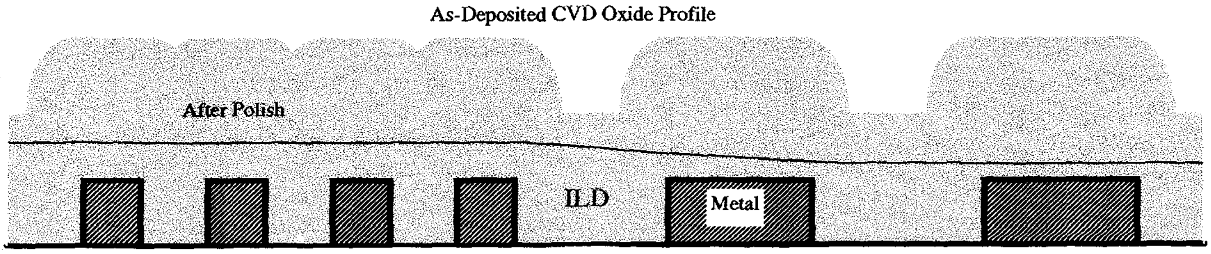

Fig. 1. Two types of lines with "50%" features density. The one on the left has a finer pitch and collects more CVD oxide at the side walls. For a given depth, there is more oxide to remove. Consequently oxide there remains thicker after polish.

0-7803-3775-1/97/$10.00 © 1997 IEEE.

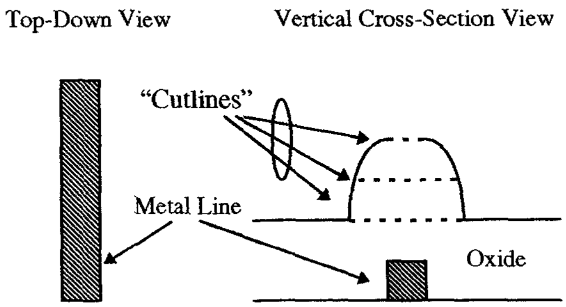

Fig. 2. The metal line is imaged from layout. When CVD oxide is deposited, it forms a protrusion over the line. The width of the protrusion at different cutlines can be viewed as the layout "sized" at different values.

sidewall of a feature to spill into an adjacent sampling window. At this point, we have a compact 3D representation of the oxide profile.

III. POLISH SIMULATION

In polish simulation, oxide is removed in small increments. As protrusions are being "planed" down, their contact areas with the polishing pad increase. The fraction of contact areas (per sampling window) can be interpolated from the cross-sectional data mentioned earlier. Our removal rate function is an extension to Preston's Law, which is:

$$R = KpL\frac{dS}{dt} \qquad (1)$$

It gives the removal rate R as a function of Preston's coefficient Kp, applied load L (contact pressure), and relative speed between the polishing pad and the oxide dS/dt. Since we do not explicitly deal with rotations of the wafer carrier and the polishing pad, the speed factor is lumped into Preston's coefficient. We also introduce the concept of contact area in our local removal rate model for features polishing:

$$R(x,y) = K'p\frac{P(x,y)}{A(x,y)} \qquad (2)$$

where $R(x,y)$ is the vertical removal rate at point (x,y), $K'p$ is the modified Preston's coefficient, P the total pressure, and A the fractional contact area (100% for oxide bulk, and variable for the protrusion section). Note that this equation implies the "volume" removal rate is independent of the contact area.

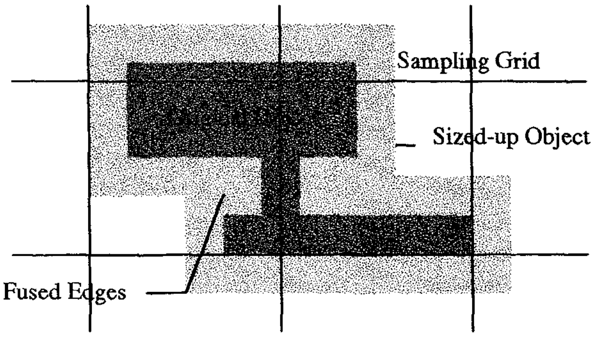

Fig. 3. During the sizing operation, overlapped edges must be fused together. Changes can be dramatic. Note that old and new objects may appear in different sampling windows.

Initially all protrusions are of the same height; thus $P(x,y)$ is uniform across the whole field. $R(x,y)$ varies as caused by different contact density. Thus as time progresses, protrusions are eroded non-uniformly across the field and become uneven in height. As the pad bends to conform with the new surface, pressure redistribution appears. Our pressure model is given as:

$$P(x,y) = P_a + \iint \Phi(x'-x, y'-y)h(x',y')dx'dy' \qquad (3)$$

where P is the total pressure at point (x,y), P_a is the applied pad pressure, and the double integral is an expression relating perturbation force to pad bending. $\Phi(x,y)$ is perturbation pressure at any point (x,y) due to a step change in height at $(x=0, y=0)$, and $h(x,y)$ the oxide height. A stiffer pad yields stronger force redistribution by Φ and hence increases the planarization distance. Conversely, a higher pad pressure reduces the relative contribution of Φ, and thus worsens oxide non-uniformity.

In [3], finite-element (FEM) solution is used to determine the force distribution over the whole die. It is seldom necessary to obtain a full-scale solution like that. Typically only a "point" solution, as presented in our approach, is sufficient. After that, linear superposition can be used to obtain whole-field solution for every time step.

There are many ways to derive an expression for Φ, such as using FEM to construct a lookup table, or utilizing the fundamental solution for an elastic foundation problem. In reality, dynamical effects must also play a role. Ultimately, parameters must be fine-tuned to match measurement data. For illustrative purposes, we consider a hypothetical function of the form:

$$\Phi(x \neq 0, y \neq 0) = 5.5 \times 10^7 \, dyne \cdot cm^{-2} \times (x^2 + y^2)^{-3/2} \qquad (4)$$

where x and y are in cm.

Since net perturbation must be zero to conserve forces, $\Phi(x_0, y_0)$, in discretized form, is constrained to be

$$\Phi(x_0, y_0) = -\sum_{i\neq 0}\sum_{j\neq 0}\Phi(x_i, y_j) \qquad (5)$$

where $x_i = i\ \Delta x$, $y_i = j\ \Delta y$, and Δx and Δy are the window dimensions.

Because the cut-off distance (where Φ becomes negligibly small) is large, the numerical integration in Eq. 3 is CPU intensive. However, note that the process resembles a 2D convolution, thus we can use fast Fourier transform (FFT) to speed up the solution. Once transformed into spatial frequency domain, convolution is replaced by multiplication, which is significantly faster for moderately-sized problems like those presented in the next section.

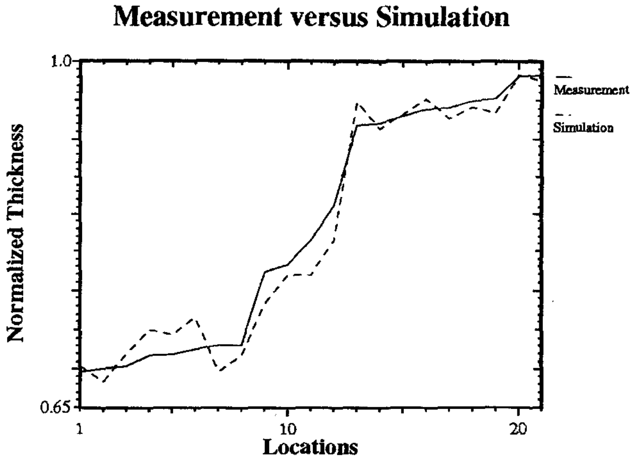

Fig. 4. Measured and simulated thickness values for 21 locations on one test chip. In general, they agree quite well. The points are ordered according to measurement values.

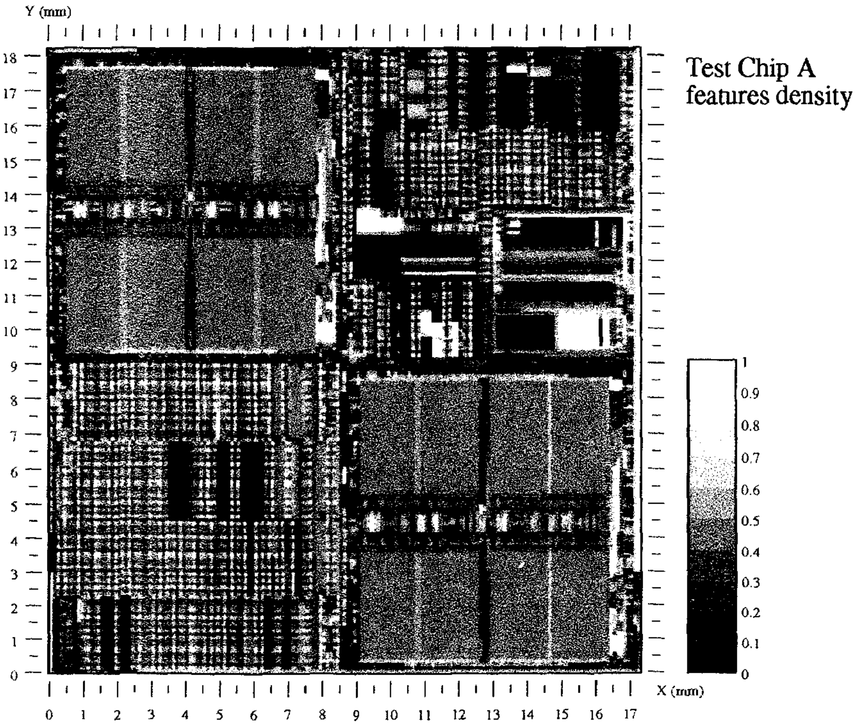

Fig. 5. Layout density of a metal layer. This test chip contains memory cells in the upper left and lower right quadrants. Their metal density is higher than the rest of the chip. However, this characteristic is not easily discernible in a gray-scaled map. The fine pitch spacing results in even more oxide to be removed from the memory cells.

Finally, for time-step iterations, we use the predictor-corrector approach. This explicit method can potentially generate unstable results. F_a is monitored for negative values. Physically no negative values are possible because the pad would rather separate from the oxide. A step is redone with smaller time increment if necessary. The simulation finishes when the desired target thickness is achieved at either the thinnest or thickest location, depending on the specification

IV. SIMULATION RESULTS

The accuracy of our simulation method is illustrated in Fig. 4, which shows how well simulation compares with measurement data.

To show how oxide profile would look like after polish, we demonstrate the simulation technique on a metal layer from another test chip. Fig. 5 plots the layout density. Memory blocks are located in the upper left and lower right quadrants. The simulation stops when the oxide thickness[1] reaches 1.3μ at the thickest location, which is inside one of the memory blocks. Fig. 6 displays the final thickness profile.

Evolution of thickness non-uniformity is illustrated in Fig. 7. Oxide thickness is plotted as a function of polishing time for different locations. The rapid divergence in the beginning ($t < 2$) is caused by uneven protrusion density distribution. The corrective force generated by the polishing pad cannot overcome this problem. After all protrusions have been removed, it enters the bulk polishing phase ($t > 2$) where contact areas are uniform throughout the die. At this point, the corrective force can

[1] Oxide thickness is measured from the top of a metal line.

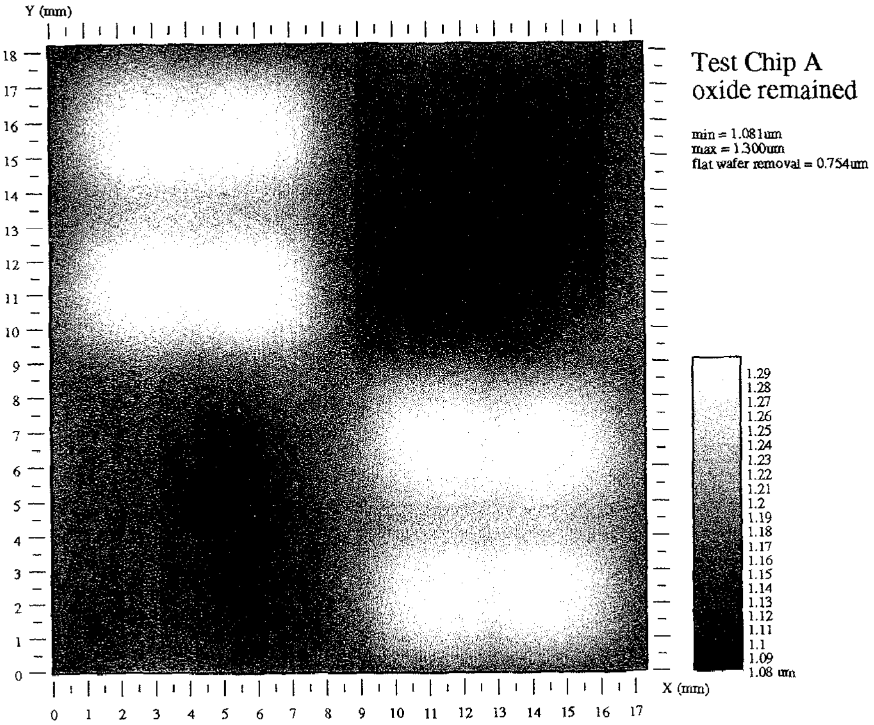

Fig. 6. Oxide thickness profile after polish. The memory rows have significantly thicker remaining oxide because they have a lot more oxide to be removed. The profile is obtained using force redistribution equation (4) and $P_a = 1.38\times10^5$ dyne/cm^2 (2psi).

act to reduce global non-uniformity, as evident from the converging lines.

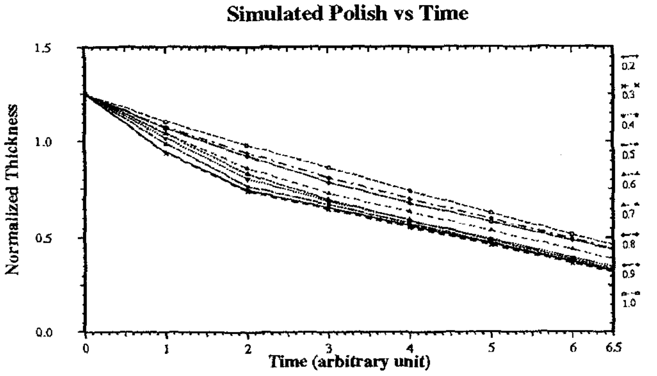

Fig. 7. This plot illustrates thickness non-uniformity as a function of polishing time for a third chip. Thickness values at 9 different locations diverge in the beginning, but start to converge slowly in the later stage (t > 2).

V. CONCLUSIONS

We have presented a realistic method for simulating CMP planarization on a die-scale level. A compact 3D representation of oxide protrusions is produced by sizing and sampling layout data. Efficient force calculation technique is used in the polish simulation step. We can determine the polishing characteristics before a chip is fabricated.

REFERENCES

[1] J. Warnock, “A Two-Dimensional Process Model for Chemi-Mechanical Polish Planarization,” J. Electrochem. Soc., Vol 138, p 2398-2402, 1991.

[2] P. Burke, “Semi-Empirical Modelling of SiO_2 Chemical-Mechanical Polishing Planarization,” 8th Intl. VMIC Proc., p379-384, 1991.

[3] Y. Hayashide *et al*, “A Novel Optimization Method of Chemical Mechanical Polishing (CMP), ”12th Intl. VMIC Proc., p 464-467, 1995.

[4] H. Ohtani, M. Murota, M. Norishima, H. Shibata, and M. Kakumu, “A Simple and Accurate Model Using Elastic Deformation Theory for Dielectric Chemical-Mechanical Polishing Process,” 12th Intl. VMIC Proc., p 447-452, 1995.