VLSI Multilevel Interconnect Conference, Santa Clara, CA, June 18-20, 1996.

Compensating for CMP Pad Wear Using Run by Run Feedback Control

Taber Smith, Duane Boning,
James Moyne[1], Arnon Hurwitz[2],
and John Curry[3]

MIT, Cambridge MA, 02139
[1]Univ. of Michigan, Ann Arbor, MI, 48109
[2]SEMATECH, Austin, TX, 78741
[3] Strasbaugh Inc., San Luis Obispo, CA, 93401

Abstract

A prototype hardware/software system has been developed and applied to the control of chemical-mechanical polishing (CMP). The control methodology uses a linearized model of the process, which is updated using an exponentially weighted moving average (EWMA) model adaptation strategy. This is coupled with multivariate recipe generation incorporating user weights for preferential input variability and output error tolerance, bounds on the input ranges, discretization of the machine settings, and optimal control parameter selection. In our control experiment a removal rate of 1600 Angstroms/minute was maintained for 500 wafers while maintaining tight control on the wafer-to-wafer uniformity (2%) and within-wafer uniformity (3.3%), on a single pad.

1.0 CMP Process Control

The control of CMP is chronically difficult, arising from poor understanding of the process, degradation (wear-out) of polishing pads, inconsistency of the slurry, and the lack of in-situ sensors [4]. Therefore, constant supervision and maintenance of the process is required in order to maintain a tight wafer-to-wafer uniformity and within-wafer uniformity. The lifetime of a pad (for a tightly monitored wafer-to-wafer uniformity) is typically on the order of 100-200 wafers (see [1] for example baseline process runs). This causes expensive machine downtime as well as wasted product and equipment consumables. Here we seek to employ run by run process control to substantially increase CMP pad life while maintaining tight control of the wafer-to-wafer and within-wafer uniformity, thus allowing a large decrease in monitor wafer usage, machine downtime, and amount of necessary process supervision. This control is achieved by monitoring the removal rate (corresponding to the amount of oxide polished during the step) and the within-wafer uniformity of that removal across the wafer on a run by run basis. The removal rate is determined as the difference between the measurement of oxide film thickness before and after polish at each of nine sites on the wafer, divided by the (fixed) polish time. The “Removal Rate” output is the average amount removed at each of the nine sites on a wafer. The “Nonuniformity” output parameter is computed for each wafer as the standard deviation of the amount removed over the nine sites on the wafer, divided by the average amount removed over the nine sites, times 100. The control goal was to maintain a target removal rate and within-wafer nonuniformity in the face of pad wear and equipment disturbances on a run by run basis.

2.0 The Run By Run Controller

Off-line experiments were performed to build empirical (static input-output) models of the process response. An optimal process recipe is selected as the initial recipe for process control. Lots of 10 wafers each were planarized in the tool, and measurements of oxide film removal rate and nonuniformity are made on wafers #9 and #10. This information is fed into the run by run controller, which adapts the process response models. These updated models are then used to generate a new process recipe which (a) achieves the best (weighted) trade-off among the multiple output targets, or (b) achieves all targets with the smallest (weighted) change in the recipe. The revised recipe is then used for the next lot of wafers [2].

2.1 Control Model Development

A central composite design was conducted with Polish Pressure (7-9 psi), Backpressure (0-2 psi), Table Speed (20-30 psi), and Pad Profile

(-1 to 1) as inputs. Second order polynomial regression models were constructed for removal rate and nonuniformity with adjusted R^2 of 96% and 82%, respectively. The response surfaces are plotted in Fig. 1 as a function of two of the major variables used for control. Each polynomial regression model was linearized around the operating point to generate a multivariate model for the gradual mode run by run controller:

$$y = Ax + c$$

where y is the output vector (removal rate and nonuniformity), x is the input recipe vector (Polish Pressure, Backpressure, Table Speed, and Pad Profile), A is a 2x4 matrix of model coefficients, and c is a vector of offset terms. In this controller, we only update the offset terms c, while the gain coefficients A remain fixed. The linear response surface for removal rate was very accurate (96% R^2) while the nonuniformity model was relatively poor (56% R^2).

2.2 The Run by Run Control Algorithm

The control scheme uses a dynamic model (on a lot by lot basis) which is formed by adapting the offset terms c, while the gain matrix A remains fixed. The model (offset term) is updated recursively by an exponentially weighted moving average (EWMA) update of the offset term based on the error between model prediction and measurement:

$$c_t = \square(y_t - Ax_t) + (1 - \square)c_{t-1},$$

where c_{t-1} is the offset term from the previous run and □ is the EWMA weight or filter factor.

Recipe generation is aided by several practical features which are detailed in [2]. First, in order to maintain the controller in a regime where the linear model is sufficiently accurate, and to incorporate machine setting limits, input constraints on recipe generation are incorporated via a heuristic which recursively reduces the linear equations, locking one input at each iteration, until a bounded recipe is found. Second, user preferences as to which inputs should vary more are accommodated by transforming the linear equation into a weighted coordinate system. Third, discrete recipe generation provides the best recipe given quantized machine settings by recursively locking inputs to their nearest possible setting. Finally, an optimal EWMA weight vector, □, is determined by using experimental data to classify a process and then using this information to run simulations over the range of EWMA weight vectors. The optimal EWMA weight is that vector which minimizes the mean squared error of the controlled process in simulation.

Figure 1. Response Surfaces

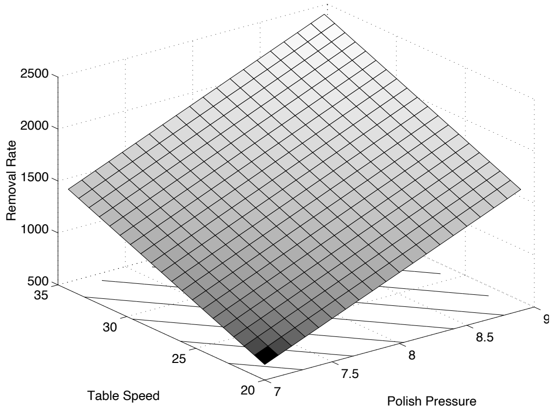


Figure 1a. Removal Rate

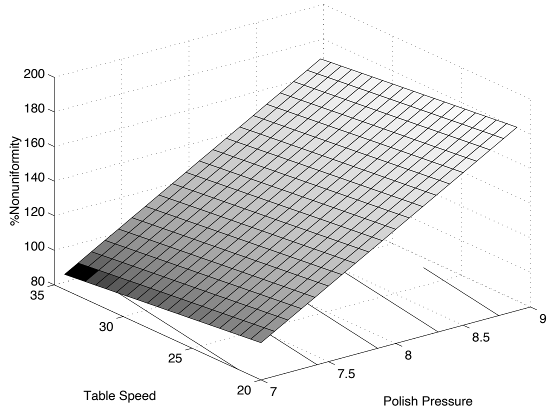
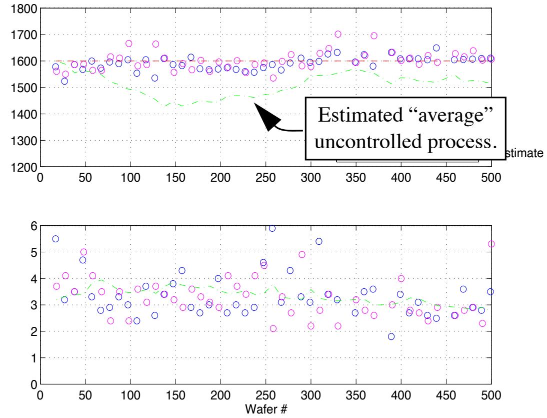


Figure 1b. Nonuniformity

3.0 500 Wafer Control Experiment

We now turn to experimental results which demonstrate the power of this control framework and the corresponding run by run control methods. The GCC control framework [3] was used on a Strasbaugh 6DS-SP dual head planarizer to test the possibility of extending the CMP pad life while maintaining a tight wafer-to-wafer uniformity. As can be seen in Fig. 2, 500 wafers were polished using the EWMA controller with no indication of pad wear or equipment disturbances in the removal rate and no increase in nonuniformity over the entire run. The within-wafer uniformity was 3.3% on average and the removal rate was maintained at 1597 Angstroms/minute with a wafer-to-wafer uniformity of 2.0%. An estimated baseline was determined from the input output data and the model adaptation algorithm given above. As can be seen in Fig. 2, the controlled removal rate provides much improvement over the estimated uncontrolled removal rate. Notice also that there is no indication of performance degradation even at the 500 wafer mark. These results demonstrate that the control framework can maintain tight industrial specifications for wafer-to-wafer uniformity while greatly extending the pad life for this process. This suggests the possibility for this methodology to extend CMP pad life to 1000 wafers and beyond, with very little process monitoring, scrap wafers, or machine downtime, while maintaining very tight control of material removal.

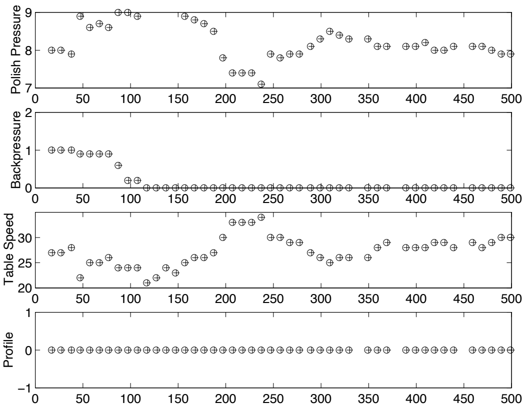
Figure 2. 500 Wafer Run - Outputs



Estimated “average” uncontrolled process.

In addition to successfully controlling this process for more than 500 wafers, several other goals have been met. As can be seen by the input trajectories shown in Fig. 3, all the inputs are bounded (they are plotted in their allowable ranges) and discretized (shown by the discrete step changes). The user preference to vary Polish Pressure and Table Speed over Backpressure and Pad Profile is clearly seen in Fig. 3.

Figure 3. 500 Wafer Run - Inputs



4.0 Conclusions and Future Work

We have demonstrated the successful application of run by run control to compensate for pad wear in chemical-mechanical polishing, extending the pad life beyond 500 wafers with tight control of wafer-to-wafer uniformity. Future work will explore control methods based on changing polish time to control material removal, or a combination of polish time and other equipment settings.

Acknowledgments

This research has been supported in part by SEMATECH, Strasbaugh, and by ARPA under contract NC-001-0014.

References

[1] D. Boning et al., “Run By Run Control of Chemical Mechanical Polishing,” *IEEE Proc. of the 17th Int. Elect. Manuf. Tech. Symp.*, Oct. 1995.
[2] D. Boning et al., “Practical Issues in Run by Run Control,” Sixth Annual SEMI/IEEE ASMC, Boston, Nov. 1995.
[3] J. Moyne and L. McAfee, Jr., “A Generic Cell Controller for the Automated VLSI Manufacturing Facility,” IEEE Trans. Semi. Manuf., vol. 5, no. 2, pp. 77-87, May 1992.
[4] E. Del Castillo and A. Hurwitz, “Run to Run Process Control: a Review and Some Extensions,” submitted to J. Qual. Tech., Feb. 1994.