# Physical Understanding and Modeling of Chemical Mechanical Planarization in Dielectric Materials

by

Xiaolin Xie

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007



Author .....
Department of Physics
March 30, 2007

Certified by .....
Duane Boning
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by .....
David Litster
Professor of Physics
Thesis Supervisor

Accepted by .....
Thomas J. Greytak
Associate Department Head for Education
Department of Physics

# Physical Understanding and Modeling of Chemical Mechanical Planarization in Dielectric Materials

by

Xiaolin Xie

Submitted to the Department of Physics
on March 30, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Physics

## Abstract

Chemical mechanical planarization (CMP) has become the enabling planarization technique of choice for current and emerging silicon integrated circuit (IC) fabrication processes. This work studies CMP in dielectric materials in particular, which is widely used in device formation for isolation and in interconnect formation for dielectric planarization. The physical understanding of the process is essential for CMP tool engineers to design optimal consumables, for circuit engineers to make the layout design manufacturing friendly and for process engineers to better control the process. The major contributions of this work are a framework to study the physics of CMP and physically-based particle-level and die-level models of polishing and planarization.

A framework for studying the physics of CMP is established by analyzing the complex system and decoupling the interactions occurring at different scales. A particle-level CMP model is developed that bridges the microscopic polishing mechanisms to the macroscopic properties of the system. A physically-based die-level model is proposed by explicit modeling of the pad and pad surface asperities, with model parameters that are based on the physical properties of the pad rather than purely fitting parameters. A semi-empirical die-level CMP model, motivated by the new physically-based die-level model, is developed that improves upon previous pattern-density step-height models by making realistic assumptions and approximations, and improving the ease of computation. The model is applied to simulate polishing of either single-material or dual-material structures with either conventional or non-conventional slurries. The die-level models are then applied to engineering problems, including design for manufacturing, nanotopography impact, wafer edge roll-off effects, and motor current based endpoint detection.

Thesis Supervisor: Duane Boning
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: David Litster
Title: Professor of Physics

# Acknowledgments

It has been seven years since I first visited MIT to attend the open house, and I cherish every moment of my stay at this prestigious institute. This experience has not only taught me knowledge and research skills, but also shaped the person that I am today. I have had the great opportunity and privilege to learn from and to interact with many talented people, and I am truly grateful for their help and support.

First, I would like to thank my thesis advisor, Professor Duane Boning, for all the support and encouragement which were more than I could ever ask for. I am truly grateful for the opportunities and the guidance he gave me. In addition to the constant inspiration on research, he is such a noble person and serves as a perfect role model for my life. I am also blessed to have Professor David Litster as my co-advisor, and I have benefited tremendously from his encouragement, his broad vision, and his wisdom in research and life. I would also like to thank the thesis readers, Professor John Joannopoulos and Professor Eric Hudson, for serving on my thesis committee, for arranging meetings despite their busy schedule, for their encouragement, and for their suggestions and feedback on my work.

This thesis is the result of significant industrial collaboration, and I would like to thank all of my collaborators.

- Aaron Smith, Paul Allard, Neil Patel, Xiaoman Duan, and Stephen Swan, at National Semiconductor, for two great summers spent with them in South Portland, Maine, and for the guidance, discussion, and much help in the design of the new STI mask and experimental data collection.

- Brian Lee, Aaron Gower-Hall, Taber Smith, David White, Vikas Mehrotra, and Stephen Fisher, at Praesagus Inc. (now part of Cadence), for a great summer internship experience, many insightful discussions on CMP, and constant help including support on layout extraction.

- Scott Lawing at Rohm & Hass and Katia Devriendt at IMEC for their guidance and help on the endpoint detection work.

- Frank Meyer and Roland Rzehak of Infineon Technologies (now Qimonda), Peter Wagner of Siltronic AG, and Win Baylies of BayTech, for the comprehensive experimental design and data collection of the nanotopography impact work, as well as their help and discussion on the project.

- Manish Deopura, Pradip K. Roy, and Sudhanshu Misra, of Neopad, for giving me the opportunity to try out their innovative Neopad, and for the collaboration and discussion on evaluation of the pads.

- David Stein and Dale Hetherington of Sandia National Laboratories for their collaboration on the endpoint detection work.

I would like to thank the academic CMP community from which I have constantly drawn support and research ideas. I would like to thank Professor Ara Philipossian for his help and support, and it has been a great experience to collaborate with him and Jam Sorooshian on endpoint detection work. I would also like to thank his other students, Yun Zhuang, Yasa Shampurno, Daniel Rosales-Yeomans, Darren DeNardis, Ting Sun, and Hyosang Lee, of University of Arizona, for explaining their research work to me and for many discussions on CMP. I would like to thank Professor David Dornfeld for his help, and to thank his student Jianfeng Luo for discussions. I would like to thank Len Borucki of Araca for his generous sharing of his expertise and thoughts on CMP. It is fortunate for me to have a local CMP community in Boston, whose monthly meeting has been a constant source of inspiration. I would like to thank the participants: Vincent Manno, Chris Rogers, Robert White, Caprice Gray, James Vlahakis, Nicole Braun, Andrew John Mueller, and Jeffrey C. Thompson, at Tufts, Howard Stone at Harvard, and Gareth McKinley at MIT.

This thesis would not have been completed without the support of my colleagues at MIT. Specifically, I would like to thank the former and current members of the Boning group, for sharing one of the most important periods of my life and I feel privileged to shared in yours.

- Brian Lee, Tae Park, and Tamba E. Gbondo-Tugbawa, who taught and helped me to start in the field of CMP.

- Hong Cai, Ed Paul, Daniel Truque, Brian Tang, and Zhipeng Li, for various thoughtful discussion on CMP.

- Karen Mercedes Gonzales-Valentin Gettings, Nigel Drego, Karthik Balakrishnan, Daihyun Lim, Hayden Taylor, Kwaku Abrokwah, Shyam Vudathu, Ajay Somani, Shawn Staker, Tyrone Hill, Mehdi Gazor, and Ali Farahanchi, for many cherishable memories.

I would like to thank Xiaoman Duan, who treats me like family. I would like to thank Professor Thomas Greytak, Brian Caravan, and Maria Riefstahl in the physics department for their help. I would like to thank Sharlene Blake and Debroah Hodges-Pabon for their dedicated assistance. I would like to thank the laboratory staff at the Microsystems Technology Laboratories for their help with experimental work performed at MIT. And I would like to thank the MTL computer gurus for their advice and assistance on the various computer-related issues that came up.

I would like to thank Professor Marc Kastner for giving me the opportunity to spend one year in his group and for many inspiring conversations. I would like to thank Nicole Morgan and Marija Drndic from whom I learned a lot on quantum dots and research skills. I would like to thank Ghislain Granger for his help and for his great effort to teach me swimming. I enjoyed working in the same lab with Jessica Thomas, Yuri Khavin, Andrei Kogan, and Sami Amasha.

I would like to thank all of my friends who have been and will continue to be an important part of my life.

Finally, my endless thanks and love to my parents, Xunlin Xie and Xianling Yu, and my fiancee, Yuan Chen. For their love and sacrifice, I dedicate this thesis to them. To my parents, I owe everything and everything. They are the kind of parents who cherish every second talking with me over the phone, but always make the conversation succinct to avoid distracting me from my research. As I have been studying abroad, I have not been able to fulfill many duties of a son, and I feel deeply grateful and indebted for all of your support and encouragement. I am also truly grateful to have Yuan in my life, who has been a constant source of encouragement

and joy over the years. We have shared many happy moments and have grown through tough periods, and I wish to walk the rest of my life with you at my side.

This work has been supported in part by National Semiconductor, and the SRC/NSF (now SRC/SEMATECH) Engineering Research Center (ERC) for Environmentally Benign Semiconductor Manufacturing.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Chemical mechanical planarization (CMP) has become the planarization technique of choice for silicon integrated circuit (IC) fabrication. It is widely used in the front end for device isolation (including shallow trench isolation formation) and for building advanced device structures, and in the back end for dielectric planarization and metal damascene formation [21].

CMP was originally developed to meet the stringent demands for inter-level dielectric planarization for aluminum interconnect and has been driven by multiple industry demands afterwards. CMP technology has been able to advance thus far mainly by borrowing from the mature techniques of glass polishing and trial-and-error. Despite the many uses and large potential of CMP, the physics of the process is not well understood. The purpose of the thesis is to present a framework to study the CMP process at both the particle-level and die-level, to study the physics of the CMP process in dielectric materials, and to provide CMP models and model applications for improved die- and wafer-level uniformity and process control.

**Contribution of this thesis** is presented in Section 1.5.

Prior to this, this chapter provides an overview of the CMP process and its related equipment and consumables. Section 1.1 describes the background and motivation for CMP. Section 1.2 briefly explains the polishing process as well as each of the participating components. Section 1.3 reviews polishing equipment and related metrology tools. The three levels of CMP models, at the wafer-, die-, and particle-level, are introduced in Section 1.4. Finally, Section 1.6 outlines the structure of the rest of the

thesis.

## 1.1 Background and Motivation for CMP

Chemical mechanical planarization, also known as chemical mechanical polishing, was initially developed at IBM for the fabrication of circuit interconnect wiring, also referred to as back-end processes. Its primary initial motivation was to planarize the inter-level dielectric layers to reduce the topography for successive metalization layers and to allow the use of higher-resolution optics for photolithography that has stringent requirements for depth of focus, radiation wavelength, suppression of standing waves, layer-to-layer registration, and other parameters. In addition to dielectric planarization, CMP is also used in polishing multiple materials as part of a damascene process, as in the shallow trench isolation process or copper interconnect formation. To illustrate how CMP is used, two examples are described in the next two subsections, dielectric planarization and shallow trench isolation.

### 1.1.1 Dielectric Planarization: Single Material Polishing

The polishing of inter-level dielectric (ILD) layers is one of the earliest applications of CMP. The motivation is to achieve planarity and to meet the requirement of the lithography step. Although a state-of-art lithography tool is capable of refocusing after each print, a flatness with very small variations over at least each die or die stripe (about 20 $mm \times 20\ mm$) is desired. Depth of focus describes the ability of a lithography system to successfully resolve features over certain surface height variations. Depth of focus (DOF) decreases with the minimum feature size, and can be estimated by [1]:

$$DOF = 10.75\frac{b^2}{\lambda}. \tag{1.1}$$

Typical lithography technology uses 193 nm wavelength light [22], and we can project the depth of focus requirement based on minimum feature size projections in Table 1.1. Note that this is the total DOF budget for lithography; thus the planarity requirement for CMP is even higher.

Table 1.1: Depth of Focus for Decreasing Feature Size

| Year | Minimum Feature Size (nm) [22] | Depth of Focus ($\mu m$) |
|---|---|---|
| 2005 | 80 | 356 |
| 2006 | 70 | 273 |
| 2007 | 65 | 235 |
| 2008 | 57 | 181 |
| 2009 | 50 | 139 |
| 2010 | 45 | 113 |
| 2011 | 40 | 89 |
| 2012 | 36 | 68 |
| 2013 | 32 | 57 |

The dielectric planarization process flow is illustrated in Figure 1-1. The metal layer is first deposited; then the layer is patterned and etched to form interconnect wires or lines; silicon oxide is uniformly deposited using Chemical Vapor Deposition (CVD); lastly, the oxide layer is flattened or planarized using CMP. The global flatness allows an accurate lithography step afterwards and prevents any topography from accumulating as multi-level metal structures are fabricated.

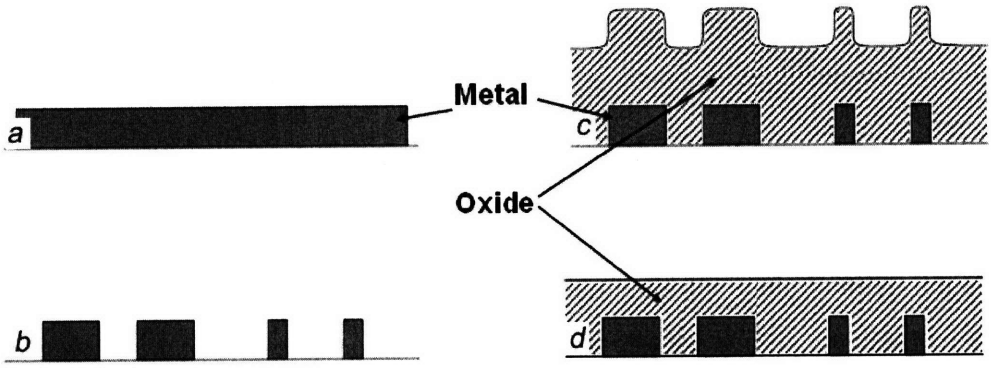



Figure 1-1: Oxide polishing process flow summary: (a) metal deposition, (b) metal etch, (c) oxide deposition, (d) oxide CMP

Similar to dielectric planarization, the CMP process can be used to polish any other single material when high local planarity and good long-range or global flatness is required.

### 1.1.2 Shallow Trench Isolation (STI): Damascene Process

CMP is also useful for the removal of overburden material in damascene processes. In the damascene approach, features are created by etching trenches where features are to be located, depositing material into the trenches, and then using CMP to remove overburden material that is also deposted on the raised area. Damascene processes are useful when the feature material cannot be etched effectively (such as copper interconnects [21]), but are also used for other processes, such as shallow trench isolation [1].

Shallow trench isolation (STI) is the only isolation scheme for semiconductor manufacturing with active area pitches in the sub-0.25 $\mu$m regime. STI is preferred over local oxidation of silicon (LOCOS) because it has near zero field encroachment, good latch-up immunity, better planarity, and low junction capacitance [23]. STI is also highly scalable, with the trench-fill capabilities being the only major challenge to scaling.

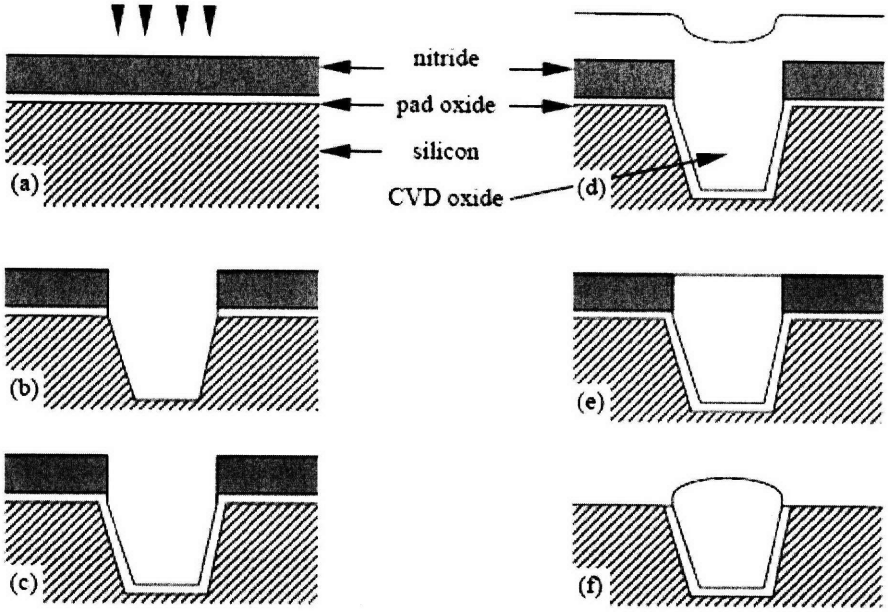



Figure 1-2: STI process flow summary: (a) pad oxide and nitride deposition, (b) anisotropic trench etch, (c) trench sidewall passivation, (d) trench fill, (e) CMP planarization, (f) nitride and pad oxide strip [1].

Figure 1-2 shows a typical STI process flow. First a thin (10 $nm$) pad oxide and

a blanket nitride film (150 $nm$) are deposited on a virgin silicon wafer. The isolation trenches are etched such that the desired trench depth (depth from silicon surface) is achieved (typical depth is 500 $nm$). A fill dielectric is deposit in the trench, and the CMP process is used to polish off the overburden dielectric, down to the underlying nitride, where the nitride serves as a polishing stop layer. After CMP, the nitride layer is then removed using an etch process, resulting in active area regions surrounded by field trenches. An SEM of a typical shallow trench isolation structure is shown in Figure 1-3.

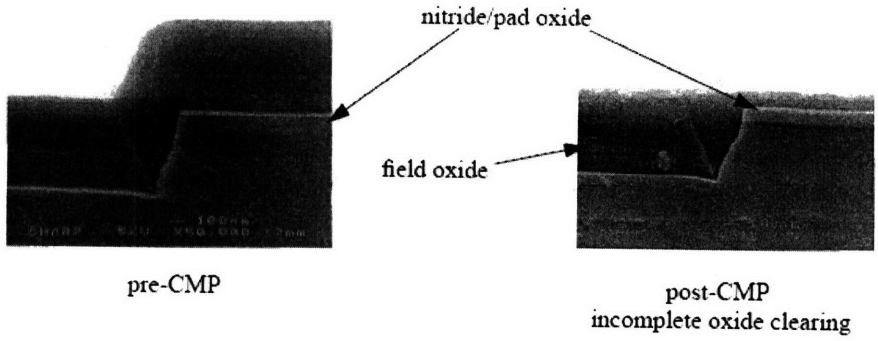



Figure 1-3: SEM of typical STI process pre- and post-CMP ( from [2] ).

### 1.1.3 Other Applications of CMP

From the two previous examples, the ability of CMP to planarize the wafer surface enables building multi-layer structure as well as forming inlaid or damascene structures such as those in STI. CMP can be used in microelectronic manufacturing whenever a high degree of planarization is demanded. Thus, the "P" in CMP often stands for planarization rather than polishing.

Besides dielectric polishing and STI, CMP is also used in metal polishing as part of back end process. Tungsten and aluminum are among the earliest metal elements to be polished in IC fabrication to achieve better planarity. In current technology, copper is the metal used in interconnect due to its improved conductivity and reduced electromigration. However, copper cannot be easily dry etched, which is the traditional way to form interconnect. Thus, copper damascene enabled by CMP is the dominant technique used to form interconnect in ICs [15].

As new materials and more sophisticated devices are introduced in IC manufacturing, CMP is finding many new applications, such as building advanced transistor structures, nonvolatile memories, silicon-on-insulator processes [21], and wafer bonding [24]. While being used to flat the ever-shrinking IC devices, CMP has also been used in the fabrication of MEMS devices, to smooth surface roughness[25], to prevent topography build-up [26], to planarize MEMS structures [27], or to form high aspect ratio MEMS structures [28]. The ability to achieve planarity also makes CMP a critical step in photonic crystals fabrication [29].

As the CMP process gains in popularity, familiarity and stability, more and more applications of this planarization technology will emerge. The challenges lie in the ability of CMP to handle new materials with a range of chemical and mechanical properties. An in-depth study of the physics of CMP can help guide application and future development of the technology.

## 1.2 Removal Mechanism

CMP is a process that combines chemical reactions on the wafer surface and mechanical force to remove surface materials and to achieve planarity. CMP can be thought of as chemically aided mechanical polishing, in which material removal is primarily due to a three-body contact, as illustrated in Figure 1-4. First the wafer surface is softened or modified by the chemical solution, and second, the soft surface layer is removed by abrasive particles held by a polishing pad. Without chemical interactions or surface modifications, the wafer surface is too hard to be polished; without mechanical polishing, the chemical attack of the surface material is usually self-limiting, resulting in near zero static etch rates. Fundamentally, CMP depends on the synergy between chemical and mechanical interactions near the surface of the wafer.

In this simple picture, four components are involved: the wafer surface, the chemical solution, abrasive particles in the slurry, and the polishing pad.

- The **Wafer surface** is the critical region of the object (a wafer) being polished. The wafer surface typically includes one or more thin films which undergo polish.

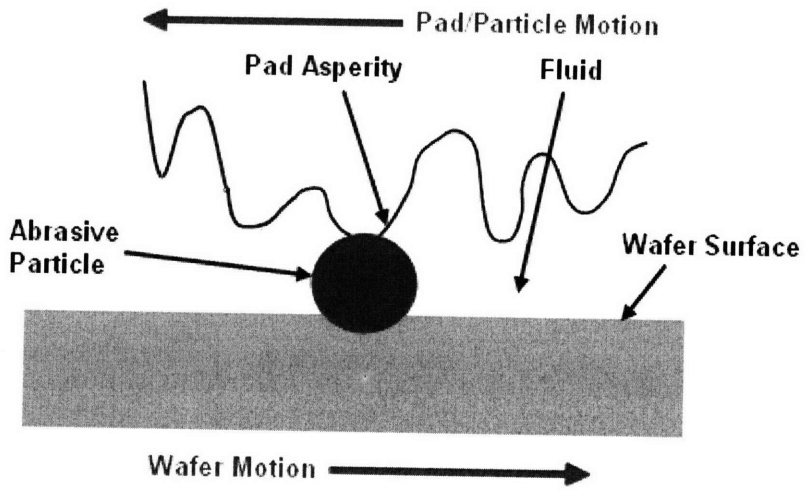



Figure 1-4: Material removal is due primarily to three-body contact (From [3] ).

The wafer surface can consist of a single material, such as silicon dioxide in a thick inter-level dielectric, or a mix of several materials, such as the stack of silicon nitride, silicon dioxide, and silicon present during during some steps of an STI CMP process.

- The **Chemical solution** is the liquid component of the polishing slurry. The chemical solution transports abrasive particles to the surface, and helps transport biproducts away from the surface. The other function of the chemical solution is to chemically soften or modify the wafer surface. It usually has a high pH value for polishing dielectrics, and a low pH value for polishing metals.

- **Abrasive particles**, the other component of the slurry, remove the weakened surface materials. For dielectric polishing, the abrasives are typically made of silica or ceria, while for metal polishing, they are typically made of silica or aluminum. The size of abrasives ranges from 50 $nm$ to a few hundred nanometers in diameter.

- The **Polishing pad** also transports fresh chemical solution to the wafer surface and carries removed debris away. The pad is largely responsible for the mechanical part of CMP. The pad holds and forces abrasive particles against

the wafer surface. The pad also exerts higher pressure on raised areas on the wafer surface, resulting in a higher removal rate in raised regions which enables planarization (or reduction in surface topography).

The material removal rate is often described by Preston's equation [30]:

$$RR = K \cdot P \cdot V \tag{1.2}$$

where $RR$ is the removal rate, $K$ is a constant also referred to as Preston's coefficient, $P$ is the local pressure on the wafer surface, and $V$ is the relative velocity of a point on the surface of wafer with respect to the pad.

Preston's equation is an empirical law discovered in glass polishing. For most of the data obtained in practice, especially in dielectric CMP, Preston's law provides a reasonably good fit to the data. Preston's equation suggests a linear dependence of removal rate on local pressure and relative velocity, explicitly highlights these mechanical components in CMP. The rest of the contributions to CMP removal rate, especially including the chemical effects, are lumped into the constant $K$.

Preston's equation partly explains the planarization ability of CMP. The raised areas on the wafer surface compress the polishing pad more than do the recessed areas, and the resulting higher pressure in raised areas contributes to a higher removal rate serving to flatten the wafer surface topography.

## 1.3 CMP Polishing Equipment and Metrology Tools

This section provides an overview of how CMP is implemented and what type of measurements we have to monitor the CMP process.

### 1.3.1 Polishing Tools

A pictorial view of a rotary CMP tool is shown in Figure 1-5. A wafer is held on a wafer carrier such that the surface to be polished faces and is pressed down against a polishing pad. The pad is typically made from a porous polyurethane, and is attached to a rotating table. The wafer carrier is rotated with some constant angular velocity

in the same direction as the pad or table. The carrier may also exhibit a slow lapping or oscillating motion across the pad, in addition to rotation; the primary purpose is to extend the area utilization and life time of the pad. A slurry composed of abrasive particles suspended in a chemical solution is deposited on the pad during polish, and is transported to the pad-wafer interface by the pad.

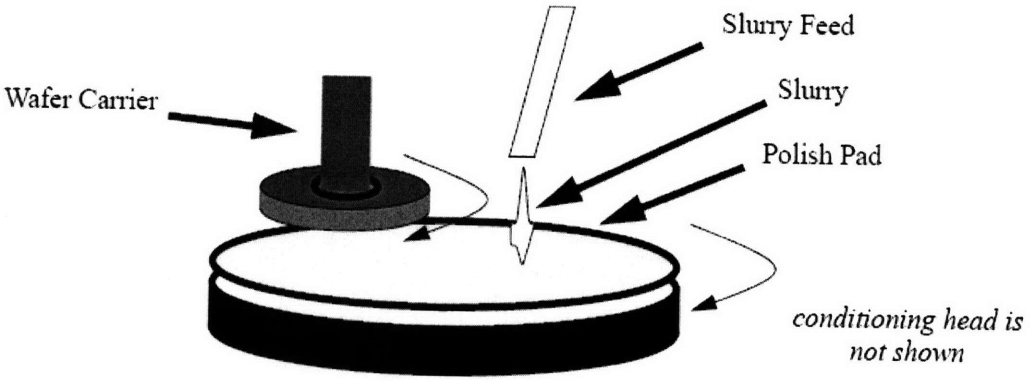



Figure 1-5: Diagram of a typical CMP tool [1].

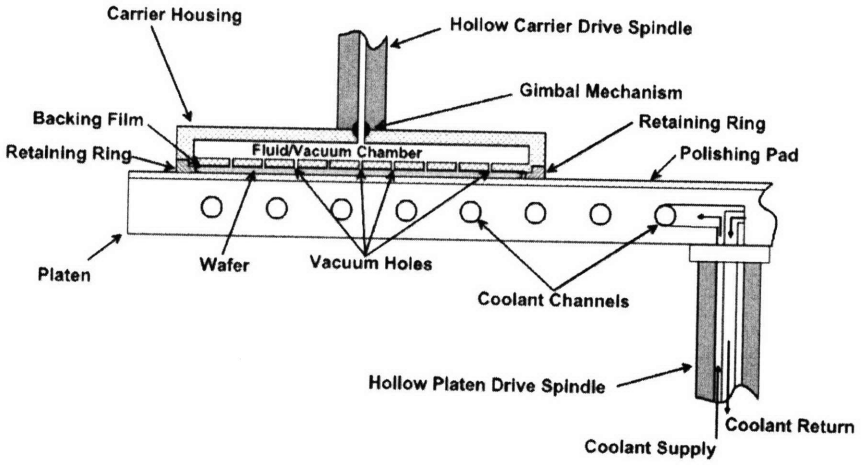



Figure 1-6: Pictorial view of a typical CMP tool [3].

The CMP tool refers to the machine used for the CMP process. The term "consumable set" typically refers to the pad and slurry (i.e., items that are consumed during the CMP process). In addition to rotary CMP tools, there are also linear CMP tools that use a rotating wafer carrier contacting a CMP pad moving on a linear belt [31]. However, the fundamental mechanism of removal (synergistic chemo-

mechanical processes created by wafer surface contact with a CMP pad, CMP slurry particles, and slurry chemistry) is similar.

One key criterion in designing a CMP polishing system is a uniform and consistent polishing process both spatially and in time. This is achieved via a collection of CMP equipment subsystems [32]:

- A mechanical drive system is able to control the relative surface velocity within one percent or better of the target speed. Sometimes the relative velocity, however, is intentionally set to vary across the wafer to compensate for other wafer level nonuniformities.

- A down force system controls the pressure distribution to be within in one percent or better across the wafer. One approach is to divide the wafer area into a few concentric zones, and to apply different pressure on each of the different zones.

- A thermal management system is used to maintain a stable and uniform temperature distribution during CMP. Temperature affects chemical reactions and is reported to have significant impact on oxide polishing [33] as well as metal polishing [34]. A spatial temperature variation causes non-uniform removal rate, and a nonstable temperature during CMP can result in over-polishing or under-polishing.

- A pad conditioning system regenerates or dresses the polishing pad surface to maintain a stable working condition, via either in-situ or ex-situ operation. The rapid decay of polishing rate without conditioning is well known [35] and well studied [36] [37]. Conditioning keeps the pad surface in a known functioning state, and helps to achieve more consistent CMP performance.

- A slurry distribution system delivers slurry to the wafer-pad surface evenly and efficiently. A low slurry flow rate may cause starvation of slurry in some regions which can result in a slow polishing rate, or can cause surface scratching due to

the lack of lubrication. On the other hand, a high slurry flow rate can increase the cost of ownership significantly.

### 1.3.2 Metrology Tools

The CMP process is used because of its ability to polish and to planarize surfaces. To study or to control CMP process, one needs to measure the film thickness or wafer surface profile before and after CMP. A number of methods are readily available in the semiconductor manufacturing industry.

Thickness of stacked dielectric films can be measured by optical methods, such as spectroscopic ellipsometry or reflectometry. An ellipsometer reflects a beam of light of known polarization off the stack film, collects the reflected light, and measures the polarization change, which is determined by film thickness and the refractive index of the reflecting film material. The KLA-Tencor ASET-F5x system is capable of measuring materials across a continuous wavelength spectrum from 190 $nm$ to 800 $nm$, with lateral spatial resolutions down to about down to 0.1 $\mu m$. In contrast to measuring very fine patterned feature structures, reflectometry tools, which analyze the intensity of reflected light to map out film thickness, provide faster measurement over larger spot sizes. The ADE AcuMap$^{TM}$ II, for example, can take up to 40,000 film thickness measurements across a wafer in under 90 seconds.

The surface profile or topography can be measured by a high resolution profiler (HRP), which moves a stylus across a single line trace or a designed area to map out the surface height profile. HRP measurement is similar to atomic force microscopy (AFM), having somewhat lower resolution but capable of measuring across larger scans or larger areas.

A scanning electron microscope (SEM) is able to provide both film thickness and surface profile or height information. SEM systems scan a cross-section of sample with a very fine electron beam, and obtains an image of the cross-section by analyzing the reflected and transmitted electrons. SEM can provide a very high resolution image (down to nanometer resolution). However, cross sectional SEM images are destructive to the sample, and suffer from slow measurement.

### 1.3.3 Real Time Monitoring Tools

Endpoint refers to the point in time at which a running process is designed or desired to stop. In dielectric polishing, endpoint refers to the time when a certain planarity or thickness is achieved; in the STI process, endpoint typically refers to the time at which all of the overburden oxide has been removed. Achieving consistent endpoint often relies on good control of removal rate and simply ends the process at a pre-specified time; this approach is susceptible to run-to-run variation in IC manufacturing. Real time monitoring tools try to detect the endpoint in real time to enable better per-wafer control and consistency.

Several CMP endpoint detection techniques use the frictional force between the pad and the wafer to generate an endpoint signal; some monitor the friction directly[38]; others monitor the motor current used to drive the platen or polishing head [39]; and others monitor a thermal signal. Other endpoint detection methods including monitoring high frequency acoustic emissions during CMP [40], and monitoring the amount of ammonia generated when polishing the nitride film in STI [41].

Directly measuring film thickness in real time is another method with several advantages. However, integrating a optical reflectrometry unit in a CMP polisher requires extensive hardware reconfiguration. The Applied Materials ISRM$^{TM}$ optical endpoint system has a transparent optical window in the polishing pad for optical measurement [42]. The KLA-Tencor Precice$^{TM}$ endpointing system monitors copper thickness using eddy-current measurement for thick copper films, and switches to optical measurement for residual copper and barrier layers [43]. Because patterned wafer film stacks can be quite complicated, interpretation of time- or spatial-averaged optical or other film thickness measurements can be challenging.

## 1.4 CMP Models

Various CMP models have been proposed in the literature in an effort to understand the process [44]. These models can be categorized into one of the following three classes: particle-level, die-level, and wafer-level models. Each of these categories is

introduced briefly below.

### 1.4.1 Particle-Level Models/Physical Understanding

Particle-level models seeks to understand the material polishing mechanism of CMP, and to find the dependence of output variables, such as removal rate and surface quality, on various input variables, such as applied pressure, chemical pH, abrasive size, and other parameters. Physical understanding of the CMP process enables better design and control of the technology. If we have a a correct and proven particle-level model, then engineering approaches can be used to define inputs (such as wafer materials and process flow), boundary conditions on output variables (such as maximum surface roughness), and a utility function to be optimized (incorporating various factors such as costs, performance, and environmental impact).

In a particle-level model, CMP is usually studied in an ideal scenario: a blanket wafer with a single surface material is assumed, uniform chemical concentration, uniform or single abrasive sizes, and other simplifications are typically made. Physical understanding of CMP can also be approached empirically by isolating a few input and output variables and analyzing their dependence experimentally. In reality, due to the vast number of potential inputs and the complexity of a typical CMP system, the problem is approached by a mixture of both empirical and theoretical methods. A particle-level model can also serve as a foundation on which to build die-level and wafer-level models.

### 1.4.2 Feature/Die-Level Models

Feature and die-level models focus on the planarization capability of CMP, i.e., modeling the reduction or evolution in the height of topographical structures on the wafer surface. Feature-level models study the polishing of only one or a few structures with known and detailed geometric shapes, while die-level models study the polishing of an entire die or IC chip. Because they deal with a simple case, feature-level models are able to focus in detail on how the existing structure or feature is planarized, by modeling the transportation of chemicals and abrasives, pressure distributions,

and/or other factors. In a real product die, however, there are millions to billions of individual structures, and modeling each of them separately is not feasible. Die-level models typically resort to statistical or approximate approaches to describe and analyze the resulting topography across the entire die.

Die-level models were initially introduced to address the phenomenon that more densely packed structures are polished slower than less densely packed regions. Early die-level models focus on statistically analyzing measurements from well-defined experiments, to build empirical models of film thickness as a function of feature size or layout pattern density. In contrast to the empirical or semi-empirical approach, a bottom-up approach focuses on building feature-level models based on physical understanding of the process, and then builds die-level models by making approximations to the feature-level or detailed physical models. Die-level models often make assumptions such as uniform slurry flow across the wafer, and benefit from the boundary condition that dies are arranged periodically on the wafer.

Die-level models can help process engineers estimate viable process operating windows, identify potential weak spots in the polished chip, and choose CMP setup conditions to improve the process. Die-level models are particularly useful at the layout design stage: an IC designer can make their design more fab-friendly with the models, so that money and time can be saved later. This design for manufacturing (DFM) concept is gaining adoption in the semiconductor manufacturing industry.

### 1.4.3 Wafer-Level Models

Wafer scale models seek to address cases where the typical assumptions of die-level models fail due to tool limitations, such as non-uniform distribution of pressure, slurry concentration, and temperature. Pressure distribution is known to be highly non-uniform near the wafer edge, which results in a typical edge roll-off profile [45]. Another cause of non-uniformity is that the dies near the wafer edge are often missing some of their neighboring dies, so that the patterned environment near the wafer edge is different than for central dies. Slurry is a critical component of the CMP process; however, an even delivery of slurry across the wafer is difficult to achieve, which

results in non-uniform slurry concentration. Slurry transportation also has an effect on thermal cooling, and its variation can cause non-uniform temperatures across the wafer. Wafer scale models help the tool manufacturers to design better polishing tools as well as process engineers to better control the CMP process.

## 1.5 Thesis Contributions

The contributions of this thesis can be divided into three categories: physical understanding of the CMP process, die-level CMP models, and the application of die-level CMP models.

### 1.5.1 Understanding the Physics of CMP of Dielectric Materials

A physical understanding of CMP is critical for continued development of the process, both to improve the current technology and to expand CMP to the polish of new materials. A particle-level model is also a key building block of feature-level, die-level and wafer-level CMP models.

A modeling framework is contributed to study the physics of CMP. While the framework should be applicable to both dielectric and metal CMP, the work here focuses on dielectric polishing. The framework takes a top-down approach to break the process into interactions between the participating components occurring at different scales. A particle-level model is proposed by studying all of the interactions and integrating both the chemical and mechanical contributions. The model is used to study how the material polishing is affected by the pad Young's modulus, the abrasive size distribution, and the applied pressure.

### 1.5.2 Developing Die-Level CMP Models

Die-level CMP models can be used to improve process control and save time and cost in the design-to-product cycle (i.e., supporting design for manufacturing or DFM), and to evaluate the performance of CMP consumables and guide their design.

This work adopts an explicit framework for die-level modeling of CMP, and contributes two new die-level CMP models. The first is a semi-empirical model that improves upon existing pattern-density step-height models [46] by assuming a continuous dependence of pressure on step height due to the existence of pad surface asperities. The model is applied to the cases of single-material polishing and dual-material polishing, with conventional and non-conventional slurries. The fast computation enabled by this approximate chip-scale model makes it an appropriate tool for process control and DFM. The second model is a physically-based model, which is established by explicitly modeling the response of the pad including both bulk and surface asperities. It is capable of modeling CMP scenarios having complicated structure (such as non-flat initial surface heights or non-uniform step-height across the chip). The physical properties of the pad are used as model parameters, and thus can be directly linked with the performance of CMP. The physically-based model can be used to verify the assumptions of other die-level models, and offer the potential for improved pad and slurry engineering.

### 1.5.3 Applying Die-Level CMP Models

The third set of contributions of this thesis are three engineering applications of the die-level CMP models, together with demonstration of our die-level modeling methodology. This modeling methodology consists of experiments on test wafers, model calibration, and model simulation using tuned models. A new STI test mask is introduced, and its advantages over previous test masks are is discussed. The usage of the CMP model in design for manufacturing is described as an illustration of the methodology.

The first application using the die-level models is a study of nanotopography, which refers to nanometer scale height variations that exist on a lateral millimeter length scale on un-patterned virgin silicon wafers. The interaction of this height variation with the CMP process can result in thinning of the surface film, which is a critical concern in shallow trench isolation processes. An extensive experiment is performed using sets of 200 $mm$ epi wafers with distinct nanotopography signatures and

using different CMP processes, and experiments are conducted using both blanket and patterned wafers. The interaction of the CMP process with initial wafer nanotopography is assessed and studied using several different methods. Our conclusion is that nanotopography can have a real and measurable effect on CMP, although that effect is much smaller than layout pattern and wafer level non-uniformity effects.

The second application is the study of non-uniform polishing near the wafer edge. Using a reasonable approximation, a contact wear model is applied to simulate the CMP process near the wafer edge. This work helps to explain how a wafer edge roll-off profile can be generated during CMP, and how existing edge roll-off affects further CMP process steps.

In the last application, a friction model is proposed to explain the measured motor current, which is used as an endpoint signal. The friction model assumes that the friction is proportional to the surface roughness, and the die-level CMP model is used to simulate the surface topography evolution during the CMP process. The friction model prediction agrees reasonably well with the distinctive motor current traces observed for different wafer layouts.

## 1.6 Thesis Outline

This thesis work focuses on the physical understanding of the CMP process, building powerful die-level models, and applying them to solve real problems in CMP. In Chapter 2, the CMP process is broken down into a set of individual interactions, and an integrated model is built based on the physics of each interaction. Chapter 3 details the development of the new die-level models and their implementations, and the die-level models are compared with experimental data. Chapter 4 describes the application of die-level models to a set of real engineering problems, including nanotopography impact, wafer edge roll-off, and endpoint detection. Conclusions and suggestions for future work are presented in Chapter 5.

# Chapter 2

# Physical Understanding of CMP

Despite the extensive research on dielectric CMP, the mechanism of material removal remains poorly understood. CMP is a complex system; understanding the physical interaction of the large number of elements is a daunting task. Significant experimental efforts to date have not directly revealed the physics of CMP for several reasons. First, most of the experiments have been on a macroscopic level and do not directly probe the microscopic polishing mechanism. In the measurement of macroscopic full-wafer experiments, the results are aggregates of those on a microscopic level, which blurs the effects of individual factors and interactions. Second, in the design of macroscopic full-wafer experiments, it is difficult to control the input variables to be uniform either across the wafer or in time throughout the CMP process, and monitoring the non-uniformity can also be challenging. Third, it is difficult to isolate all of the input variables. For example, the concentration of any chemical additives affects the pH of the solution, which could affect the stability of suspended slurry abrasives. In addition, the physical properties of the surface film are linked with the material type. As a result, there are only a few candidate materials and the properties of them can not be varied individually. Finally, limitations in the availability of commercially available CMP tools and consumbles further restricts the range of available experimental data.

This chapter is intended to establish a framework for physical understanding of CMP mechanisms. The key mechanisms are explored based on physical arguments, and support for these mechanisms are drawn from the available literature and experimental evidence. Section 2.1 reviews the four participating components and their

properties, including the surface film, slurry chemical solution, slurry abrasive particles, and polishing pad. Section 2.2 summarizes important input and output variables of the CMP process, as well as related experimental data. Section 2.3 reviews the existing particle-level models for CMP. In section 2.4, the removal mechanism of CMP is discussed: by considering the various interactions of participating components, and a comprehensive model is proposed.

## 2.1 Participating Components

On the particle or most detailed microscopic level, the CMP process involves the rubbing of the polishing pad against the wafer surface, with slurry present between the pad and the wafer. The slurry consist of both a chemical solution and abrasive particles. Before discussing CMP models, we need to have a good understanding of each of these participating components. In this section, we review the basic properties of these CMP components.

### 2.1.1 Film Material

The family of materials being polished in CMP has steadily increased over the past ten years. Silicon dioxide is the first dielectric material of interest, and may be either thermally grown or deposited using physical vapor deposition or chemical vapor deposition. Among the other dielectric films typically polished in IC fabrication are silicon nitride, silicon oxynitride, and polycrystalline silicon. In metal CMP, materials of the polished film can be aluminum, tungsten, copper, and various barrier metals such as tantalum or tantalum nitride. Here we focus on the two materials frequently used in dielectric CMP, silica and silica nitride.

Silica ($SiO_2$) is widely used in IC fabrication to provide insulation between metal lines. Silicon dioxide has historically been used as the insulation material due to its unique properties: it is the only native oxide of a common semiconductor which is stable in water and at elevated temperatures, it is an excellent electrical insulator, it serves as a mask to common diffusing species, and is capable of forming a nearly perfect electrical interface with its substrate. $SiO_2$ is formed by strong, directional

covalent bonds with a tetrahedral structure as shown in Figure 2-1. The $Si - O$ bond distance is 0.16 $nm$, and the oxygen atoms are electronegative. The bond angle between $Si - O - Si$ is nominally about 145°, but it can vary from about 100° to 170° without much change in bond energy, and the rotation about the axis is nearly free. This property enables $SiO_2$ to form many different crystalline structures, and usually results in amorphous materials which lack long-range order. Thick silica layers are usually deposited using chemical vapor deposition (CVD) or physical vapor deposition (PVD), and the material is usually in an amorphous glass form rather than in crystal form.

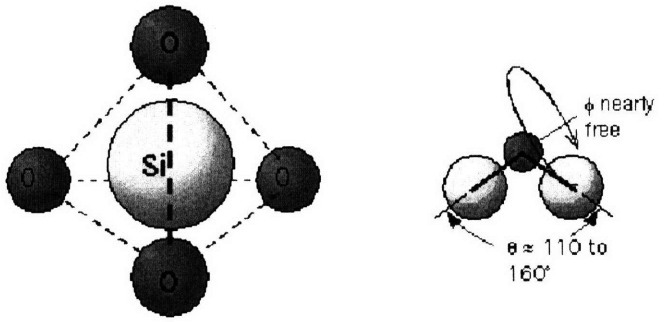



Figure 2-1: Diagrams of (a) the tetrahedral structure and (b) bond angle of silicon dioxide.

Silicon nitride $Si_3N_4$ is a hard and dense material. Each $N$ atom has three $N - Si$ bonds, and thus the bond angle is not as flexible as $Si - O - Si$. Nitride is also deposited using a CVD process, and has an amorphous crystalline structure.

### 2.1.2 Chemical Solution

The slurry is a key component of any CMP process, and consists of both a chemical solution and abrasive particles. Before reviewing these separately, it is worth noting the interaction between them. Abrasives are sub-micron particles suspended in the chemical solution. The solution pH value affects the surface charges on abrasive particles, and thus the electrostatic stability of the slurry. As a result, a specific abrasive material allows only a certain range of pH values for well-behaved, stable slurries. Various additives, such as high molecular organic compounds, are often

introduced to intervene in the interaction and increase the pH range.

The most important contribution of the chemical solution is to provide chemical agents to attack or modify the wafer surface, and water is one of the most important ingredients. Tomozawa et al. [47] find that the polishing rate using slurries which contain no water yields a near zero removal rate. In water, silica undergoes a reversible depolymerization reaction, which can be expressed as

$$(SiO_2)_x + 2H_2O \rightleftharpoons (SiO_2)_{x-1} + Si(OH)_4 \tag{2.1}$$

This surface modification is critical, as it results in a film that is more readily removed by mechanical abrasive action.

Silicon nitride is believed to be converted to oxide first [48], as in the following equation, before the resulting oxide layer is removed:

$$Si_3N_4 + 6O_2 \rightleftharpoons 3SiO_2 + 4NH_3(g) \tag{2.2}$$

Additional chemical agents, such as pH adjusters, inhibitors and accelerators, are added to maximize polishing rate, minimize defect rate, optimize material selectivity (i.e., maximize the polishing rate of one material and suppress the other), or optimize topography selectivity (i.e., maximize the polishing rate of raised features and suppress that of recessed regions). A high material selectivity is typically desired in damascene processes such as STI CMP, so that the polish process still "stop" on an exposure of an underlying film. A high topography selectivity is desired in general and can be enhanced using ceria abrasives and certain inhibitors [14]. The chemical solution also plays a mechanical role in CMP, providing lubrication between the wafer surface and the pad, transporting waste material, and maintaining stable temperature.

### 2.1.3 Abrasives

The abrasive particles remove the chemically modified layer on the wafer surface and expose fresh material for further chemical attack. The abrasives are generally thought to provide a mechanical contribution to material removal, especially in metal

polishing, in which abrasives are responsible for removing the soft oxidized metal surface. In dielectric CMP, the chemical bonding between abrasive particle and wafer surface is also important, and different abrasive materials result in vastly different polishing rates [49].

In designing or choosing slurry abrasives, the variable parameters are abrasive material type, abrasive morphology, and particle size distribution. Cook [49] reviews glass polishing rates using different abrasive materials, and observes thatthe rates can vary by orders of magnitude. Although the removal mechanism is unknown, an empirical relationship suggests that the removal rates vary with the material-oxide bond strength and isoelectric point (IEP) of the abrasive material. In dielectric CMP, abrasive particles are mostly made from silica or ceria, to obtain reasonable removal rates and low defectivity or wafer surface damage.

In addition to dependence on the abrasive material, the removal rate also depends on the method of preparation of abrasive particles. Abrasive particles can be manufactured via a fuming process or a colloid process, and the resulting particles have substantially different morphologies (shapes), as shown in Figure 2-2.

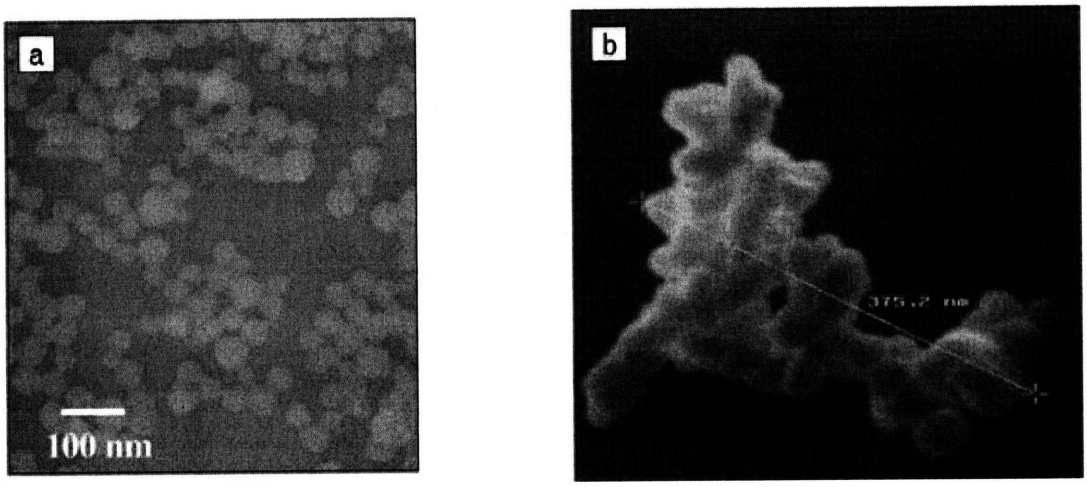



Figure 2-2: Transmission electron microscopy images of 45 *nm* nonporous silica particles synthesized by a sol-gel process [4]. (b) Field Emission Scanning Electron Microscopy (SEM) of fumed silica aggregates [5].

Most of the abrasive particles used in CMP have a size ranging from 20 *nm* to

200 $nm$, which is much smaller than the particles used in traditional glass polishing [49]. Various size selection methods have been employed to achieve a narrow size distribution. Figure 2-3 shows the measurement of size distribution of a colloidal silica slurry with mean size of 75 $nm$ using laser-light scattering and transmission electron microscopy (TEM), a narrow distribution of abrasive size is observed.

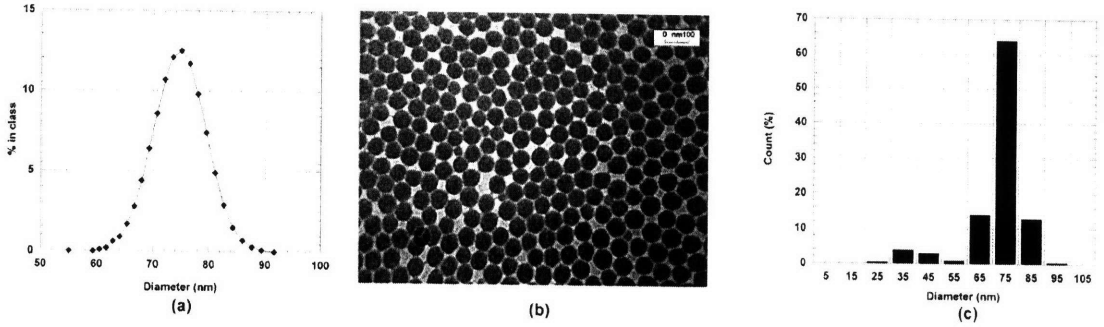



Figure 2-3: (a) Abrasive size distribution estimated using a laser scattering technique; (b) TEM of the colloidal silica abrasives; and (c) histogram of abrasive size obtained from the TEM measurement [6].

### 2.1.4 Polishing Pad

Most commercially available polishing pads are made of polyurethane, because the material satisfies several criteria [32], including sufficient mechanical integrity, chemical resistance to survive the polishing environment, and being hydrophilic. While the pad and slurry particles both contact and polish wafer surface during CMP, the wafer surface and slurry are also acting to polish or wear the pad at the same time. The pad material should be able to sustain the pressure and shear stress during CMP, and acceptable levels of hardness and modulus are required. The pad is also constantly being attacked by the chemicals in the slurry, which can be highly alkaline in dielectric CMP or highly acidic in metal CMP. Strong oxidizers such as hydrogen peroxide are often used in the slurry as well, and thus chemical resistance of the pad is a prerequisite for a long pad usage lifetime. One important function of the pad is the transportation of the slurry across the macroscopic wafer and pad areas. A hydrophilic surface facilitates the formation of a slurry film between the pad and the wafer, and prevents starvation of chemicals in some regions. One more desirable

criterion is the flexibility to vary pad properties for different polishing applications. One family of materials which satisfy these criteria is polymer, and the most commonly used polymer is polyurethane. Readers seeking a thorough understanding of polyurethane can refer to Szycher [50]. In CMP pad design, the polyurethane properties can be modified by the several approaches [32], including control of hard and soft segments, urethane stoichiometry, pad thermal history, amount of porosity, varying pad thickness, and stacked or composite pad approaches (Figure 2-4).

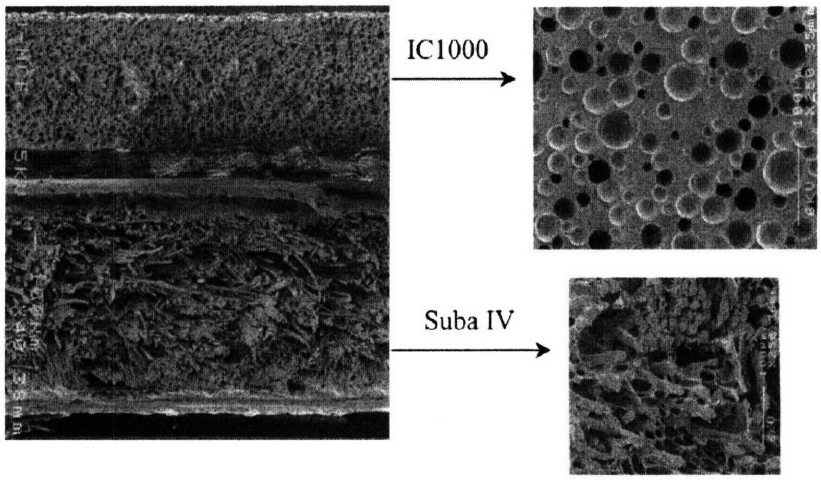



Figure 2-4: A cross-sectional SEM image of an IC1000 on Suba IV stacked pad [7].

Pad surface properties are of particular interest in CMP. The pad surface is designed to aid slurry transportation across the wafer, and the pad surface typically includes large machined grooves (0.50 $mm$ wide and 0.76 $mm$ deep spaced at 3 $mm$ pitch [51]). The pad also contains spherical pores with diameter between 30 $\mu m$ and 50 $\mu m$, to facilitate slurry transport at the microscopic level. The pad surface needs to be constantly conditioned to form and maintain surface asperities, as these surface asperities are an important part of CMP, and it has been observed that the polishing rate decays if unconditioned [35]. Hu et al. [7] analyze both fresh and used pads, and find a reduced volume of exposed pores on used pads and corresponding mechanical damage on asperity surfaces. Borucki [36] proposes a mathematical model to explain the polish rate decay by modeling the evolution of the asperity height distribution during CMP, and the model shows good agreement with experiment.

## 2.2 Process Input and Output Variables

CMP is a complicated process involving a large number of variables. Here we separate these into output variables and input variables [20]. The output variables measure the performance and quality of the planarization process. The input variables are process or consumable parameters chosen to affect the chemical and/or mechanical components of the CMP process. This section reviews these variables and examines those variables that are most strongly related to particle-level CMP models.

### 2.2.1 Output Variables

Table 2.1 lists a number of CMP output variables [20]. Each of these is discussed briefly below.

Table 2.1: Output Variables [20]

| | |
|---|---|
| Polish Rate | Surface Quality |
| Selectivity | Roughness |
| Planarization | Particles |
| Polish Rate Uniformity | Corrosion Resistance |
| Feature Size Dependence | Surface Damage |
| – polish rate | Structural |
| – planarization rate | Electrical |
| – damage | Stress |

- The polish rate is the film thickness removed on a blanket (unpatterned) wafer in a unit of polish time, usually expressed in the units of ($\text{Å}/min$). A fast polishing rate means a higher throughput rate of the CMP process and reduces the number of CMP polishers, which are very expensive tools, needed in the fab. A stable and well-calibrated polishing rate also enables good control of the process.

- The selectivity is defined as the ratio of the removal rate of one material to another. In the STI CMP process, the selectivity is the ratio of oxide removal rate to that of nitride. Often the second material is used as a stopping layer, such as a silicon layer in STI CMP, and a larger value of selectivity is preferred.

In other cases, a selectivity value close to one might be preferred, so as to achieve uniform polishing (avoiding dishing into the higher rate material).

- The main purpose of CMP is to achieve local and global planarization. Local planarization refers to the reduction of local step height, which is the height difference between a raised region and a trench region at the feature scale. Global planarization refers to flatness over longer distances, such as across an entire IC chip.

- Surface quality describes the surface roughness after the CMP process. A rough interlayer dielectric film is more susceptible to low breakdown strength and high leakage [20]. Most CMP processes are able to achieve a surface with an average roughness $R_a$ of a few Å across a micron lateral distance.

- Surface damage refers to large defects caused by mechanical scratches, usually associated with abnormally large polishing particles or other debris.

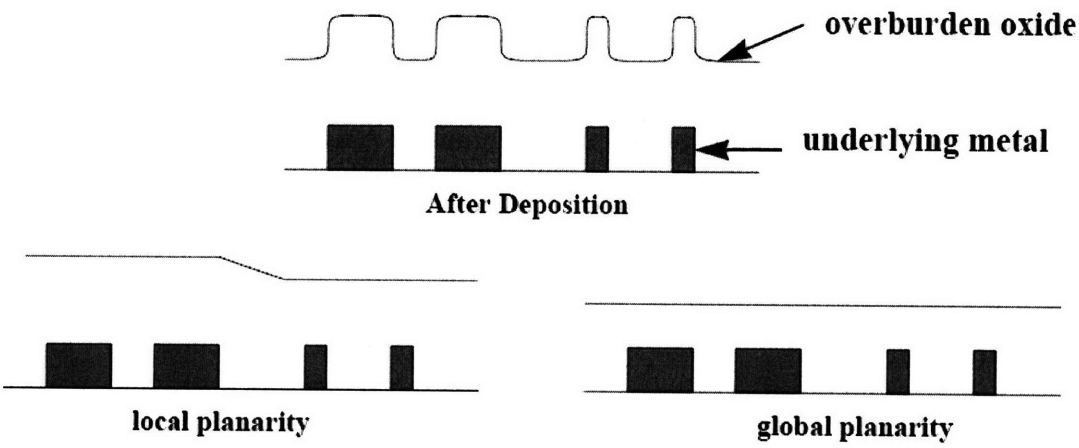



Figure 2-5: Illustration of local and global planarization. Local planarization refers to the reduction of local step height, which is the height difference between a raised region and a trench or recessed region at the feature scale. Global planarization refers to flatness over longer distance [1].

Among the output variables described above, the polish rate is of particular interest in particle-level CMP models; surface damage can also be understood based on a particle-level model.

### 2.2.2 Input Variables

The CMP process is affected or controlled by a large number of inputs [20]. A few key consumable and process input variables commonly used to tune the process are summarized below

- As previously discussed, the polishing pad is made of a polymer material, due to its resistance to chemical erosion and mechanism stress. The Young's modulus of the material affects its ability to planarize. A harder pad, whose Young's modulus is larger, has better planarization; however, a harder pad can cause more wafer-level non-uniformity due to the existence of wafer bow. The softness of the pad can be tuned with different pad pore density, pad thickness, and by stacking of the pad.

- The pad asperity distribution significantly affects the polish rate of CMP. Pad asperities aid slurry transportation, and their tips are directly in contact with the wafer and interact with the abrasive particles.

- Slurry pH affects the dissolution rate and surface properties of the wafer surface material. The properties of the pad material may also be affected by pH of the slurry.

- Slurry chemicals have one or more of several functions, including to enhance chemical modification of the wafer surface, to modify the mechanical properties of pad or abrasive particles, to modify the material removal rate selectivity, and to balance or maintain the slurry stability.

- Abrasive materials used in dielectric CMP are mainly silica and ceria. Cook [49] has shown the strong dependence of removal rate on abrasive particle material type.

- Abrasive size distribution is known to affect removal rate and defect rate [6] [8] [9] [52].

- Applied pressure and relative velocity are the most common input process variables to affect removal rate [30].

### 2.2.3 Empirical Relationships between Input and Output Variables

In this section, we review the experimental evidence of the dependence of output variables on inputs in CMP.

#### Removal Rate versus Pressure and Velocity

Preston [30] found a linear relationship between removal rate and the product of applied pressure and relative velocity, $P \cdot V$, in glass polishing. The linear relationship has been observed in dielectric CMP using silica-based slurries [8] [53] [54], as illustrated in Figure 2-6.

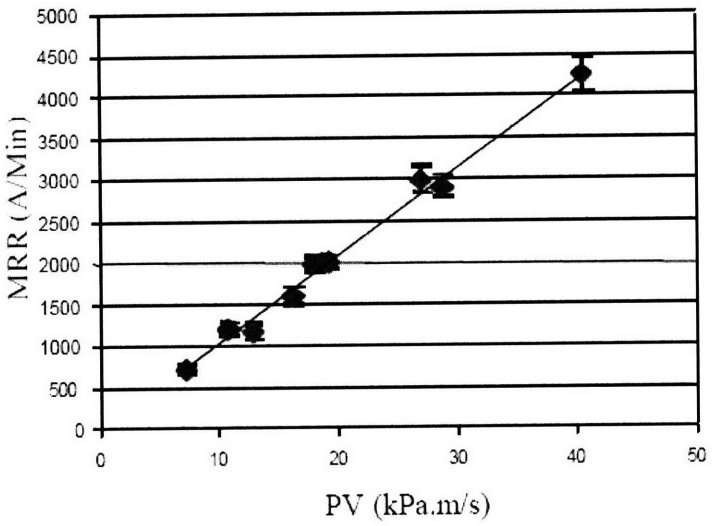



Figure 2-6: Material removal rate vs $P \cdot V$, for an IC1000 pad and 80 $nm$ silica slurry. [8]

#### Removal Rate versus Abrasive Size

In dielectric CMP, the removal rate can be affected by slurry abrasive size. For exmaple, the removal rate has been observed to peak at certain abrasive size, in experiments using silica based slurry [8] [6], as seen in Figure 2-7.

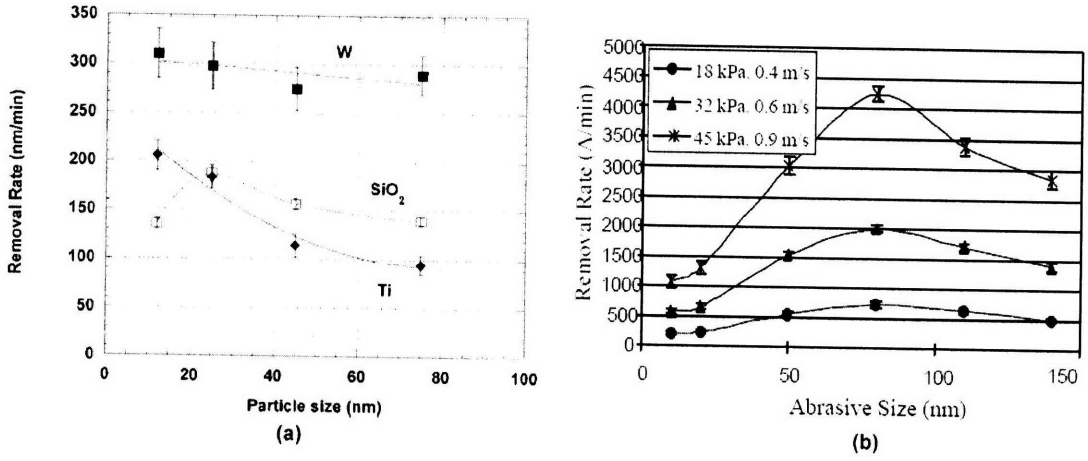



Figure 2-7: (a) Removal rate vs. particle size for tungsten ■ , titanium ♦ , and oxide □ [6]. (b) Removal rate vs. abrasive size [8]. Both use a silica-based slurry.

### Defect Rate versus Abrasive Size

The defect rate does not show any clear dependence on the average size of abrasives [9], but is seen to be proportional to the count of large particles in the slurry [52] [9]. An example of defect counts (on a blanket wafer) for different slurry and particle sizes is shown in Figure 2-8. Here $D99$ refers to the size of abrasives at the 99th percentile.

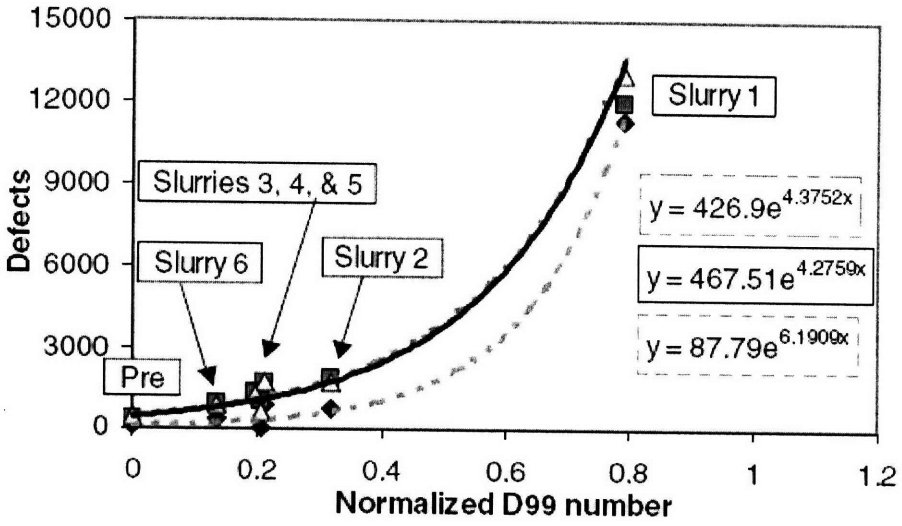



Figure 2-8: Variation of defects with D99 particle size (size of abrasives at 99% percentile) for various slurries at a constant down force [9].

## 2.3 Review of Particle Scale CMP Models

Particle scale models focuses on the two most important output variables of CMP: material removal rate and surface qualities (surface defects and scratching). The models try to explain how these two outputs are affected by various inputs, such as pressure, relative velocity, pad material, abrasive size distribution, abrasive material, and other parameters. Good reviews on particle scale CMP models include [49] [44] [55].

This section reviews the existing particle scale models in the literature. First, we review the empirical models, which attempt to identify the dependence of the output on one or a few input parameters by conducting controlled experiments. Second, we review the stress-enhanced erosion models, which treat CMP as a chemical erosion process enhanced by stress. Third, we review the models based on indentation mechanisms, which assume that the wafer surface is being mechanically plowed by abrasive particles, and relates the removal rate to the indentation depth of abrasives into the wafer surface. And finally, we review a chemical bond-dependent "pulling" mechanism, referred to as a chemical-tooth.

### 2.3.1 Empirical Models

The most widely used CMP model of material removal was introduced by Preston [30] in studying glass polishing. The Preston equation states

$$RR = K \cdot P \cdot V, \tag{2.3}$$

where $RR$ is the material removal rate, $P$ is the applied pressure, $V$ is the relative velocity between the wafer and the pad, and $K$ is the Preston coefficient together lumping all other effects. Although the equation is identified in glass polishing, the linear dependence of removal rate on the product of pressure and velocity is observed in dielectric CMP using conventional slurry [8] [53] [54].

A few revised versions of Preston's equation have been proposed for use in CMP. Maury et al. [56] observe two polishing regimes that are distinguished by the value

of $P \cdot V$. The removal rate has linear dependence on $P \cdot V$ in both regimes, and the slope of the regime $P \cdot V < 50\ psi \cdot m/min$ is larger than that of the regime $P \cdot V \geq 50\ psi \cdot m/min$. Wrschka et al. [57] observe a nonlinear dependence on $P$ and $V$ when polishing aluminum, and propose a generalized form:

$$RR = K \cdot P^{\alpha} \cdot V^{\beta}, \tag{2.4}$$

where $\alpha$ and $\beta$ are two fitting parameters. This generalized version of Preston's equation, however, is not widely used, because the complication of the model and additional parameters do not significantly improve the model accuracy. Another reason for the popularity of the basic Preston equation lies in its simple linearity, which makes it easier to use or to build die-level or wafer-level models.

The empirical models are useful in practice. When a process engineer is in charge of a new tool set or a process, a few test runs will tell him/her how to control removal rate by setting proper values of applied pressure and relative velocity. The limitation of the empirical models is that they do not reveal the removal mechanism, however, they do provide us some intuition into the CMP process, and provide experimental relationships which can be used to test predictions of different physically-based models.

### 2.3.2 Stress-enhanced Erosion Models

A CMP model proposed by Runnels [58] assumes that CMP is a stress-enhanced erosion process and that the removal rate is proportional to the product of normal and shear stresses:

$$RR = C \cdot \sigma_t \cdot \sigma_n, \tag{2.5}$$

where $\sigma_t$ is the shear stress, $\sigma_n$ is the normal stress, and $C$ is a coefficient. Runnels estimates the material removal rate by solving the hydrodynamics of the slurry film, and the model implies the same $RR \propto P \cdot V$ relationship as the Preston equation. Tseng et al. [59] estimate the shear stress by slurry flow mechanics, and the normal stress by considering the elastic contact between the abrasive and wafer. The

dependence of the removal rate to pressure and velocity is suggested to be

$$RR = C \cdot P^{5/6} \cdot V^{1/2}. \tag{2.6}$$

Zhang et al. [60] argue that adhesive force between abrasive and wafer surface dominates and normal stress is nearly constant, and the model implies

$$RR = C \cdot \sqrt{P \cdot V}. \tag{2.7}$$

Zhang et al. [61] further assume plastic deformation contact between abrasive and wafer surface. The indentation depth can be estimated by the equilibrium of the applied load, the adhesive force, and the abrasive-wafer contact pressure. The estimated value agrees well with experiment where the indentation depth is empirically estimated by the post-CMP surface roughness.

### 2.3.3 Indentation Mechanism Models

Cook [49] proposes a physical model in which a silica abrasive under pressure causes a Hertzian elastic indentation on the wafer surface, and the amount of glass being removed is proportional to the product of the distance the particle travels and the cross-section area of the indentation. A Hertzian penetration depth [62] can be calculated as

$$R_s = \left( \frac{3}{8} \cdot \frac{L}{E\sqrt{R}} \right)^{2/3}, \tag{2.8}$$

where $R_s$ is the penetration depth, $R$ is the abrasive radius, $E$ is the Young's modulus of the wafer material, and $L$ is the load on the abrasive. It is assumed that a monolayer of closely packed abrasives of the same size is formed between the wafer and the pad, and all of the pressure on the wafer is transmitted through the abrasives. If the relative velocity is $V$, the volume of material removed in unit time is

$$RR = \frac{1}{2E} \cdot P \cdot V, \tag{2.9}$$

which has the same linear dependence on pressure and velocity as Preston's equation. The model also predicts that the removal rate is inversely proportional to the Young's

modulus of the wafer material, and that it is independent of abrasive size and chemical solution.

Liu et al. [63] propose a similar model by considering the adherence of the abrasive particle to the surface, which is related to the hardness values of wafer and pad, as well as the bending of the abrasive particle, and found removal rate to be:

$$RR = C\left(\frac{H_w}{H_w + H_p}\right)\left(\frac{E_s + E_w}{E_s E_w}\right) \cdot P \cdot V, \tag{2.10}$$

where $H_w$ is the Brinell hardness value of the wafer surface, $H_p$ is the Brinell hardness of the pad, $E_s$ is the abrasive Young's modulus, $E_w$ is the wafer Young's modulus, and $C$ is a coefficient which accounts for all other effects.

Shi et al. [64] propose a model which assumes that the total removal rate equals the product of the number of abrasive in contact with the pad and the wafer, and the removal rate per abrasive. The removal rate per abrasive is assumed to be only proportional to velocity $V$, and the number of abrasives in contact is assumed to be proportional to the area of the pad in contact with the wafer. The dependence of contact area on pressure is estimated using the result of Herzian contact of a single pad asperity, and the removal rate is shown to have the following dependence:

$$RR = C \cdot P^{2/3} \cdot V. \tag{2.11}$$

Yu et al. [65] also consider the pad asperities and assume that the removal rate is proportional to the area swiped by the pad in unit time. Greenwood's model [66] is used to describe the pad asperities. Yu's model suggests a linear relationship between contact area and applied pressure, consistent with the Preston equation.

Luo et al. [67] propose a model to describe the interaction between wafer, pad, and abrasives. The model assumes that asperities have uniform height and that only the large abrasive particles are responsible for the polishing, and the following relationship is obtained:

$$RR = C_1 \cdot (1 - \Phi(3 - C_2 P^{1/6})) \cdot P^{1/3} \cdot V, \tag{2.12}$$

where $C_1$ and $C_2$ are fitting constants, and $\Phi(x)$ is the cumulative probability function

of a standard Gaussian distribution.

Qin et al. [68] propose a comprehensive metal CMP model by assuming that asperity height follows a Gaussian distribution, and that the chemistry of the slurry results in forming a soft thin film on metal surface. The soft thin film is easier to be polished than the metal film, and thus the model suggests two polishing regimes as observed by Maury et al. [56].

### 2.3.4 Chemical-Tooth Mechanism

Silicon oxide used in IC manufacturing is a form of silicate glass, and thus oxide polishing is a specialized glass polishing process demanding high planarization and low defectivity. Reviewing the mechanism of glass polishing helps us to gain some insight into the CMP mechanism. In glass polishing experiments, materials other than silicate have been studied, which provides more data to test different proposed models.

The interest in studying glass polishing arises from the need to build optical lens with highly smooth surfaces. Holland [69] and Izumitani [70] review several proposed glass polishing mechanisms. First is a wear mechanism, which proposes that the material is removed via mechanical wear. Second is a material flow mechanism, which suggests that the glass material at the peaks flows into the valleys under different pressures until the surface is flat. And third is a chemical mechanism, in which the glass material at the peaks dissolves faster due to the higher pressure.

To determine the dominating contribution, Izumitani [70] examines how polishing rate depends on material properties of various glasses, such as hardness, softening point, and chemical durability. Izumnitani uses silicate ($SiO_2$-based) and borate ($B_2O_3$-based) glasses with different concentration of modifier ions (such as $Pb^{2+}$ and $La^{3+}$). Although these are not used in IC fabrication, the range of their physical properties enables a good experimental design to study the polishing mechanism. The experiment shows that the removal rate has little correlation with hardness and softening point, and the removal rate varies predominately with the change of chemical durability. The result implies that the chemical contribution is important and CMP

is not purely a wear process.

Cook [49] presents a chemical tooth mechanism for material removal. In this mechanism, the abrasive material attaches to $Si$ through bonding with the $-O-$ atom, and then detaches or pulls the $Si$ atom away from the wafer surface. Later Osseo-Asare [71] proposes a model based on the chemical tooth mechanism, and suggested that mass transportation is responsible for the material removal.

### 2.3.5 Summary of Particle Scale Models

Our review of the literature shows that a wide variety of mechanisms have been proposed to explain the removal of oxide in CMP, and to relate removal rate to pressure, velocity, particle, pad, slurry, and other parameters. The experimental support for some of the these models is limited. For example, the range of pressures and velocities typically studied is not sufficient to conclusively establish alternatives to the Preston equation.

In the next section, we seek to develop a particle scale CMP model framework, which both decomposes the problem into key elements, and suggests specific mechanisms for those elements based a physical arguments and the available experimental data.

## 2.4 Physics of CMP Material Removal Mechanism

This section presents a framework for understanding the material removal mechanism of CMP, which incorporates both mechanical and chemical aspects. The framework analyzes the pad-abrasive-wafer interaction from a top-down approach, which bridges from the macroscopic variables to the microscopic phenomenon. Based on the framework, each interaction is analyzed and a physically-based particle-level CMP model is presented.

### 2.4.1 Modeling Framework

The material removal in CMP results from the interactions among pad, abrasive, chemical solution, and wafer surface. These interactions are inter-twined, and it is

essential to find a way to decompose and structure the interactions for understanding, and then to reintegrate these components to form the model. Fortunately, the dimensions of each component are orders of magnitudes away from each other: the size of most wafers ranges from 150 $mm$ to 300 $mm$, the size of most chips ranges from a few $mm$ to 20 $mm$, the size of pad asperities is around 50 $\mu m$, the diameter of abrasive particles are about 50 $nm$, and the chemical solution is of course liquid. Thus, the removal mechanism can be decomposed into four scales and interactions, as illustrate in Figure 2-9.

- Let us start with a system of only the polishing pad and the wafer, as in Figure 2-9 (a). The size or diameter of pad asperities (around 50 $\mu m$) is much smaller than a typical chip size (about 10 $mm$ on a side), so the pad asperities and their effects can be described statistically. Thus, the wafer-pad interaction can be approached by solving the interaction between the wafer and a single asperity, and averaging over all asperities given a distribution of their dimensions or properties. The analysis of the pad and wafer interaction can also provide information about the average fluid thickness, contact area, and pressure distribution in the contact regions where the pad asperity surfaces come in contact with the wafer. The size of the contact area between a pad asperity and the wafer can be estimated to be around 5 $\mu m$ in diameter.

- Add abrasives. The size of abrasives (50 $nm$) is much smaller than that of the pad asperity and that of the chip, so the addition of abrasives does not substantially perturb the wafer-pad contact area, pressure distribution, or slurry fluid thickness. The abrasives trapped in the contact area between the asperity and the wafer are responsible for material removal, and three separate issues need to be understood. First, the nature of the abrasive-pad interaction and second, the behavior of the abrasive-wafer interaction, must be considered. Third, the concentration of abrasives trapped in the contact areas between the wafer and the pad must be understood; this concentration is determined by the dynamics of abrasives entering and leaving the contact areas.

- Add chemical solution. The presence of the chemical solution may affect the coefficients of some physical properties (such as the Young's modulus of the pad asperities), but it does not change the nature of the above interactions. Its main contribution is to chemically modify the surfaces of the abrasives and the wafer, and to assist in the material removal of by-products.

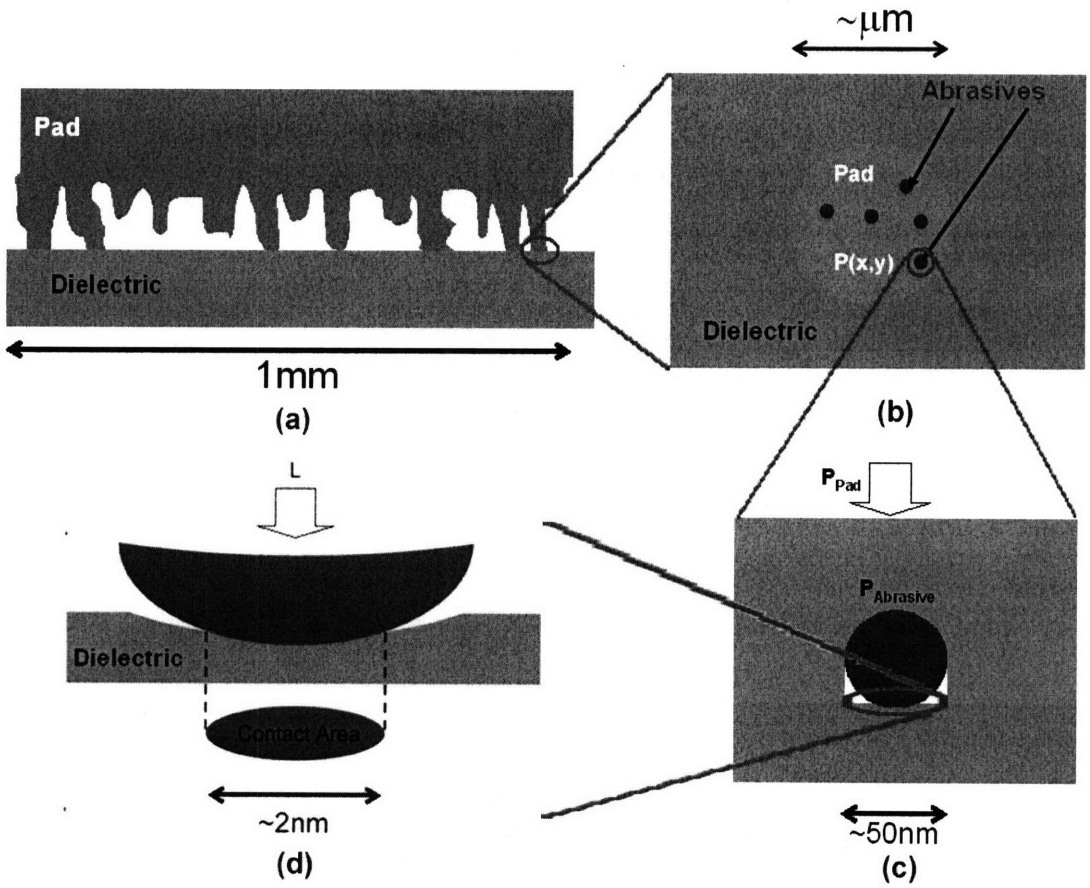



Figure 2-9: The complex system of CMP can be broken down into four pairs of interactions, which occur at different length scales.

Once each interaction is properly modeled, we can define the material removal rate in a bottom-up fashion. Our goal is to obtain the following relationships are obtained by modeling the interactions outlined above.

- The analysis of the chemical-mechanical interactions between the particle and wafer, mediated by the slurry chemistry, gives the removal rate $K^*(A_c, v_{ab,wf})$,

where $A_c$ is the contact area and $v_{ab,wf}$ is the relative velocity between the abrasive and wafer. $K^*$ may depend on the contact pressure $P_c$ as well, but it is ignored as a first order approximation. $K^*$ also depends on the materials of the wafer film and abrasives, as well as the slurry chemistry.

- Analysis of the abrasive-wafer interaction enables us to solve for the contact area $A_c(\phi, P_a)$ and the contact pressure $P_c(\phi, P_a)$, where $\phi$ is the abrasive diameter and $P_a$ is the average pressure on the abrasive.

- Modeling of the pad-abrasive interaction provides the pressure $P_a(\phi, P_{Pad}, q)$ on an individual abrasive, if the pressure on the pad asperity is $P_{Pad}$ and the concentration of abrasives in the contact area is $q$.

- Analysis of pad-abrasive-wafer interaction allows us to solve for the relative velocity $v_{ab,wf}$ and abrasive concentration on the pad $q(\phi, n(\phi), P_{Pad})$, where $n(\phi)$ is the abrasive concentration in the slurry.

- Finally, modeling of the pad-wafer interaction provides the incremental pressure density of the contact area $S(P_{Pad})$, i.e., the fraction of area with pressure between $P_{Pad}$ and $P_{Pad} + dP_{Pad}$ is $S(P_{Pad})dP_{Pad}$.

Once the individual interactions are solved, the total removal rate $K$ is

$$K = \int_0^\infty d\phi \int_0^\infty dP_{Pad} \cdot q(\phi, n(\phi), P_{Pad}) \cdot S(P_{Pad}) \cdot K^*(A_c(\phi, P_a(\phi, P_{Pad}), q), v_{ab,wf}) \tag{2.13}$$

The framework can also be used to estimate the defect rate due to scratches from large abrasive particles. If a defect occurs when the abrasive-wafer contact pressure exceeds a certain threshold, i.e., $P_c(\phi, P_{Pad}) > P_c^*$, the defect rate is

$$D = \int_0^\infty d\phi \int_0^\infty dP_{Pad} \cdot q(\phi, n(\phi), P_{Pad}) \cdot S(P_{Pad}) \cdot H(P_c(\phi, L) - P_c^*) \tag{2.14}$$

where $H(x)$ is the unit step function.

Each interaction will be modeled in turn in the next five sections.

### 2.4.2 Material Removal Mechanism

Several material removal mechanisms have been proposed as summarized in Section 2.3, and can be classified as either abrasive indentation models, in which abrasives plow the wafer surface [49] [67] [63], or chemical tooth models, in which abrasives bond and pull surface atoms away [49] [71]. Both models agree on the importance of the hydrolysis reaction between the chemical solution and the wafer surface; however, they differ on the interaction between the abrasive and the wafer surface. In the plowing mechanism, the abrasives mechanically indent into the softened wafer surface, and remove the wafer material as they move along. In the pulling mechanism, the bonding between abrasives and the wafer surface layer is essential.

We adopt the chemical tooth model in this work; our belief for this being the operative mechanism in dielectric CMP is based on the following evidence.

- Cook [49] shows that the material removal rate is independent of the hardness of the abrasive material in oxide CMP, but the removal rate positively correlates with the bond strength between the abrasive material and oxide. Note that this also suggests that a stress-induced dissolution of the wafer surface under particle pressure is not the mechanism of oxide CMP, as such dissolution would not be expected to depend on the abrasive-oxide bond strength.

- The post-CMP profile is measured to have an RMS variation of a few angstroms which corresponds to a few layers of silicon atoms, the surface does not exhibit discernable lateral directionality or micro scratches, and the variation is independent of abrasive size [6]. A gradual pulling mechanism might generate more spatially random point removals than a succession of many random mechanical plowing. And in the plowing mechanism, the indentation depth, which can be estimated by the RMS surface variation, is expected to depend on the abrasive size.

In the chemical tooth model [49], the material removal results from the following two steps.

- Bonding between the abrasive particle with $OH$ surface groups, $-M-OH$, and the silica surface, $-Si-O^-$:

$$-Si-O^- \ + \ -M-OH \ \rightarrow -Si-O-M- \ + OH^-. \tag{2.15}$$

The concentration of $-M-OH$, $[-M-OH]$, is proportional to the contact area between the abrasive particle and the wafer surface. $[-M-OH]$ is also proportional to the relative velocity of the abrasives with respect to the wafer, as faster movement enables more abrasives to come into contact with more wafer areas.

- Pulling of $Si$ from the surface.

$$-Si-O-Si-O-M- \rightarrow -Si-O \ + \ Si-O-M- \tag{2.16}$$

The pulling of the $Si$ atom away from the wafer depends on the chemical reaction between the wafer surface and the chemical solution. The stress caused by the contact pressure may weaken the bonds between the surface $Si$ atom and the bulk, and affect the pulling rate.

Let $p$ denote the probability of removing a silicon atom when an abrasive contacts the wafer surface, let $d$ denote the distance between neighboring silicon atoms, and let $v_{ab,wf}$ denote the relative velocity between the abrasive and the wafer. Thus, if $A_c$ is the contact area between abrasive and wafer, the removal rate is

$$K^*(A_c, v_{ab,wf}) = p \cdot A_c \cdot (v_{ab,wf}/d) \cdot d = p \cdot A_c \cdot v_{ab,wf}. \tag{2.17}$$

The probability $p$ likely depends on the contact pressure; however, the relationship is not known yet, and in the first order model proposed here, $p$ is assumed to be independent of pressure.

### 2.4.3 Abrasive-Wafer Interaction

The abrasive-wafer contact is illustrated in Figure 2-10. The contact area $A_c$ and the contact pressure $P_c$ can be solved in the abrasive-wafer interaction, which can be

considered as a typical Herzian contact, giving

$$\begin{cases} A_c &= \pi \left(\frac{3}{4}\frac{\phi L}{E'}\right)^{2/3} \\ P_c &= \frac{1}{\pi}\left(\frac{4E'}{3\phi}\right)^{2/3} \cdot L^{1/3} \end{cases} \tag{2.18}$$

where $\phi$ is the radius of the abrasive particle, $L$ is the applied load, and $E'$ is the effective Young's modulus of the wafer $E_w$ and the abrasive $E_a$, i.e., $\frac{1}{E'} = \frac{1-\nu_a^2}{E_a} + \frac{1-\nu_w^2}{E_w}$, where $\nu_a$ and $\mu_w$ are the Poisson's ratios for the abrasive and pad respectively. If $P_a$ denotes the average pressure on an abrasive, i.e., $P_a = L/(\pi\phi^2)$, we can rewrite the contact area and the contact pressure as

$$\begin{cases} A_c &= \pi\phi^2 \left(\frac{3\pi}{4}\frac{P_a}{E'}\right)^{2/3} \\ P_c &= \left(\frac{4E'}{3\pi}\right)^{2/3} \cdot P_a^{1/3} \end{cases} . \tag{2.19}$$

Thus, for an abrasive with size $\phi$ under pressure $P_a$, the contact area increases in proportion to $P_a^{2/3}$, and the contact pressure increases in proportion to $P_a^{1/3}$ and is independent of $\phi$. We note that the product $P_c \cdot A_c$ is simply the total load on the asperity, i.e., $P_c \cdot A_c = \pi\phi^2 P_a = L$.

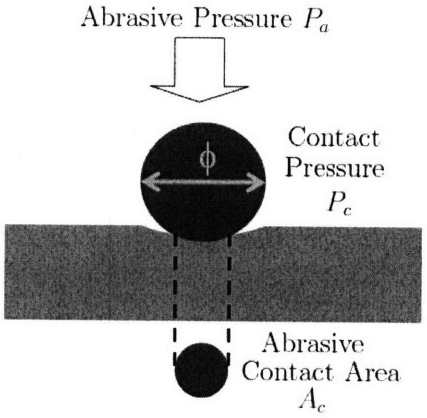



Figure 2-10: Diagram showing the interaction between abrasive and wafer surface.

### 2.4.4 Pad-Abrasive Interaction

The pad-abrasive interaction is studied to estimate the abrasive pressure $P_a$, given abrasives with size distribution $q(\phi)$, the asperity pressure $P_{Pad}$, and the Young's modulus of the pad asperity $E_p$. Here are the assumptions.

- The pad asperity is assumed to be an elastic body.
- The contact area is estimated to be several microns in diameter, which is much larger than the abrasive diameters (about 50 $nm$). Thus, it is reasonable to approximate the interaction as between an abrasive and an infinitely large elastic body with pressure $P_{Pad}$.
- The Young's modulus of the abrasive particles is much larger than that of the pad asperities, so the deformation of the abrasives can be ignored.
- The penetration of the abrasives into the wafer surface is expected to be very small, and is neglected for the calculation here. This assumption will be verified in the study of abrasive-wafer interaction.

The contact wear model, which is discussed in Appendix B, is used to simulate the interaction between the pad and abrasives. First, the case of a single abrasive is studied; second, a system of two abrasives with the same size is studied to observe their interaction; third, a system of $N$ abrasives all having the same size is studied; and last, a system of abrasives with different sizes is considered.

### Single Abrasive

In the single abrasive case, the abrasive particle is assumed to be spherical shape, and we are interested in the dependence of abrasive pressure $P_{a,s}$, where the subscript $s$ indicates the single abrasive case, on abrasive size $\phi$, pad pressure $P_{Pad}$, and Young's modulus of pad $E_p$. Before starting the contact wear model computation, we can deduce some aspects of the relationship by studying the contact wear formula,

$$w(x,y) = \frac{1-\nu_p^2}{E_p} \int d\xi \int d\eta \frac{P(\xi,\eta)}{\sqrt{(x-\xi)^2+(y-\eta)^2}}, \tag{2.20}$$

where $w(x,y)$ is the pad surface profile (pad height displacements) and $P(x,y)$ is the local pressure between pad and abrasive as a function lateral position $(x,y)$.

First, consider the case with an abrasive of size $\phi$, and a second case with abrasive having size $2\phi$. If $w(x,y)$ and $P(x,y)$ are the solution for the first case, $2w(2x,2y)$

and $P(2x, 2y)$ can be shown to be the solution for the latter case. The comparison indicates that the two cases have the same pressure distribution, and generalizing the case suggests that the abrasive pressure is independent of abrasive size, i.e., $P_{a,s} = P_{a,s}(P, E)$.

Next, we compare $P_{a,s}(P_{Pad}, E)$ and $P_{a,s}(P'_{Pad}, E')$ with the constraint $P_{Pad}/E = P'_{Pad}/E'$. In the contact wear formula, the pad displacement depends only on $P/E$. Thus, the two cases have the same displacement and the pressure scales with applied pressure, i.e., $P_{a,s}(P'_{Pad}, E') = (E'/E) \cdot P_{a,s}((E/E')P'_{Pad}, E)$. Thus, only the dependence of abrasive pressure on pad pressure for a given $E_0$ needs to computed, and the general case can be then obtained by

$$P_{a,s}(P_{Pad}, E) = \left(\frac{E}{E_0}\right) \cdot P_{a,s}\left(\frac{E_0}{E} P_{Pad}, E_0\right). \tag{2.21}$$

The contact wear model is used to simulate the pad-particle interactions with $E = 120\ MPa$, $\nu = 0.3$, $\phi = 40\ \mu m$, and pressure values from 50 $psi$ to 1600 $psi$. The pad profiles and pressure distributions can be obtained and are shown in Figure 2-11. The relationship between the abrasive pressure and the pad pressure is shown in

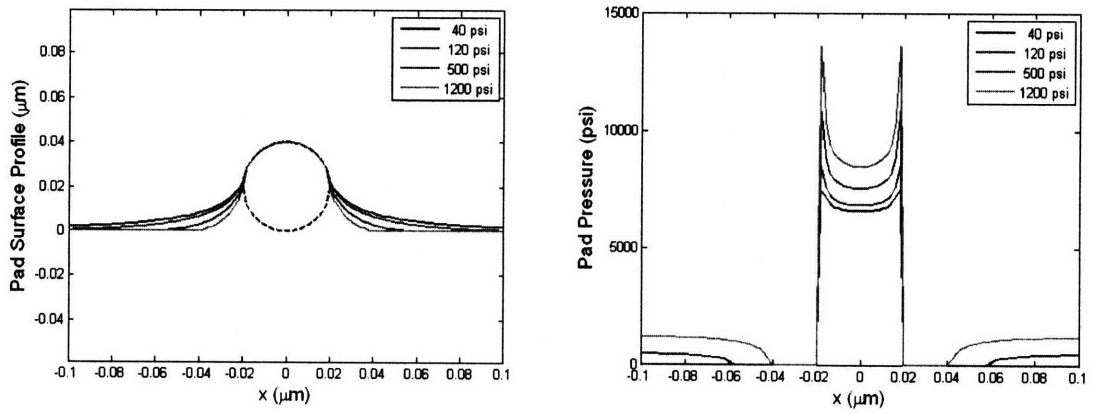



Figure 2-11: Pad surface $w(x, y)$ and pressure $P(x, y)$ profiles under different applied pad pressures $P_{Pad}$.

Figure 2-12, and an empirical fit of $P_{a,s}$ to $P_{Pad}^{1/2}$ agrees reasonably well for the range of pressures used.

Next, we generalize the problem to a truncated abrasive with height $h$, as illus-

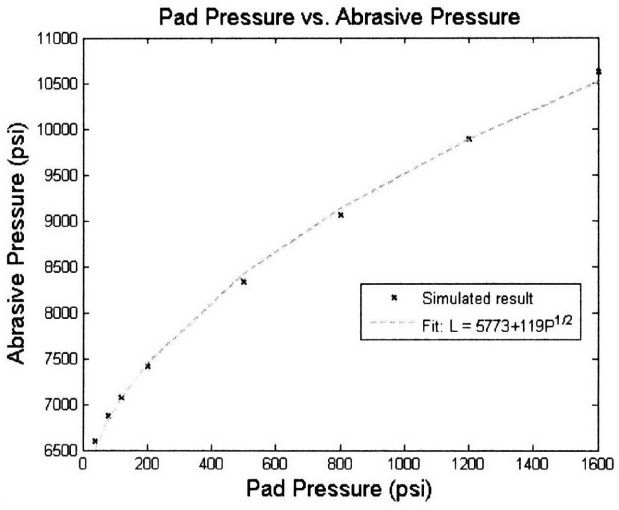



Figure 2-12: The dependence of abrasive pressure $P_a$ on pad pressure $P_{Pad}$, and empirically the abrasive pressure is found to be proportional to the square root of the pad pressure.

trated in Figure 2-13. We define the truncated ratio $r = h/\phi$, so that $r = 1$ is the non-truncated case. The truncated abrasive can be used to model the case when the abrasive penetrates into the wafer surface and the pad only sees a truncated abrasive. A second reason will become clear in later discussions. The above analysis of the dependence of abrasive pressure on $E$ still holds, and only the dependence of $P_{a,s}$ on $P_{Pad}$ and $r$ needs to be numerically determined.

$$P_{a,s}(P_{Pad}, E, r) = \left(\frac{E}{E_0}\right) \cdot P_{a,s}\left(\frac{E_0}{E} P_{Pad}, E_0, r\right). \tag{2.22}$$

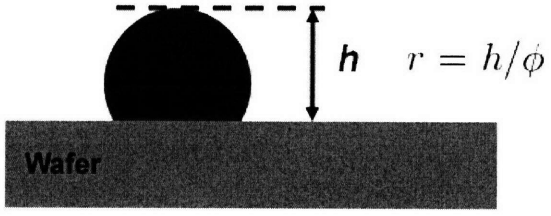



Figure 2-13: Illustration of a truncated abrasive, where the ratio $r$ is defined as $h/\phi$.

Using the same parameters as in the un-truncated case and the range of $r$ values

from 0 to 1, the result of contact wear simulation is shown in Figure 2-14. Although the results are not available in closed form, they can useful in two ways. First, the figures show that the dependence of load on $r$ and pressure is a smooth function; thus in computation of the aggregate model, the values of load can be easily interpolated from a few fully computed cases. Second, the analysis provides numerical estimation of the magnitude of abrasive load and pressure. Figure 2-14 shows that the abrasive pressure can be much higher than the pad pressure in case of high $r$ value. The magnification of the pressure is due to the fact that a single abrasive shields the nearby area, as indicated in Figure 2-15 (a).

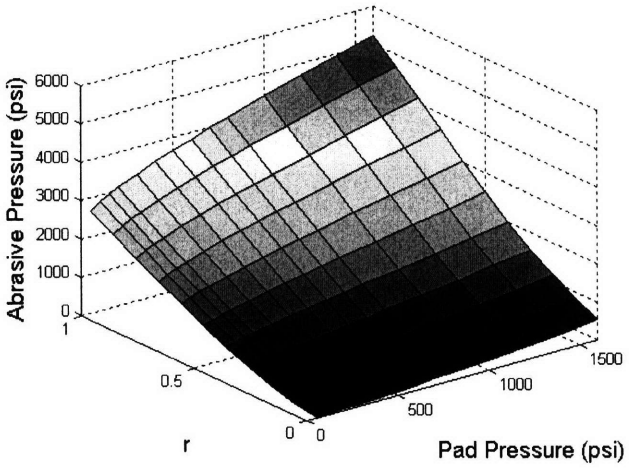



Figure 2-14: The dependence of abrasive pressure on pad pressure and truncated ratio $r$.

The shielding radius $R(\phi, P_{Pad}, E, r)$ is defined as the distance between the center of the particle contact area to the closest point where the pad contacts the wafer, as illustrated in Figure 2-15. $R(\phi, P_{Pad}, E, r)$ can be shown to satisfy

$$R(\phi, P_{Pad}, E, r) = \left(\frac{\phi}{\phi_0}\right) \cdot R\left(\phi_0, \frac{E_0}{E} P_{Pad}, E_0, r\right). \tag{2.23}$$

Under the same pad pressure, the larger the shielding radius, the larger the average pressure on an abrasive, and the abrasive pressure is approximately proportional to $R^2$. In the cases of two abrasives, the abrasives have little effect on each other if their distance is larger than $2R$, and the result can be applied to the cases of multiple

abrasives.

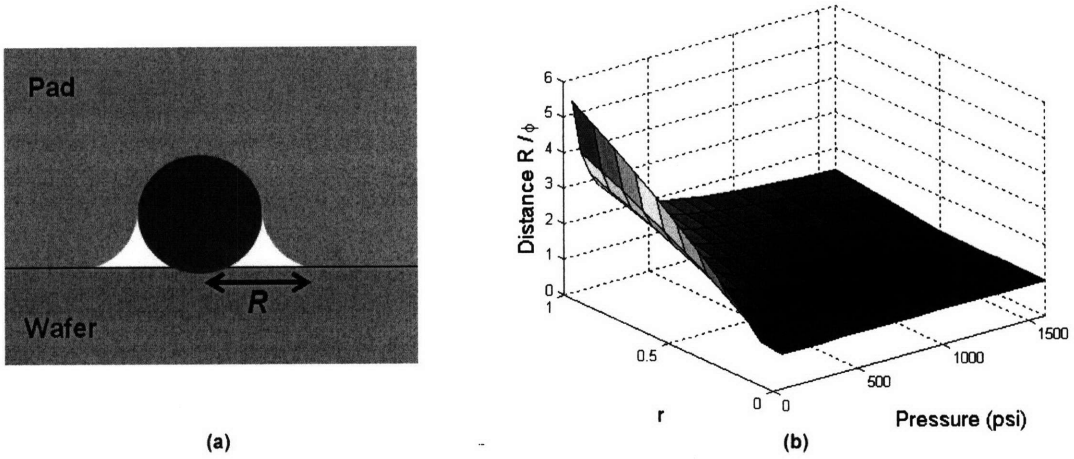



Figure 2-15: (a) Illustration of shielding radius $R$. (b) The dependence of shielding radius $R$ on $P_{Pad}$ and $r$.

### Two Abrasive Particles Separated a Distance

Now we study the case of two abrasive particles with the same size and separated by some distance. The results for pad displacement and abrasive pressure as a function of separation distance are illustrated in Figure 2-16 (a), where the model parameters used are $E = 120\ MPa$ and $P = 500\ psi$. The distances in the simulated cases range from 1 $\phi$ to 5 $\phi$, and the result is shown in Figure 2-16 (b). The pressure on each abrasive decreases with the smaller separation distance, as the closer the two abrasives are the more they shield each other. Figure 2-15 (b) shows that the shielding radius $R$ equals approximately $2\phi$ when $r = 1$ and $P = 500\ psi$, which implies little interaction between the two abrasive when the distance is larger than $2R = 4\phi$. In the two abrasive case, when the distance is $4\phi$, the abrasive pressure is shown to be almost the same as that in the single-abrasive case.

### Multiple Abrasives of Same Size and Density

This section studies the cases of multiple abrasives of the same size and varying area density $\rho$, where the area density is defined as the ratio of cross-section area of abrasives to the total area. When the area density of abrasives is very low, the

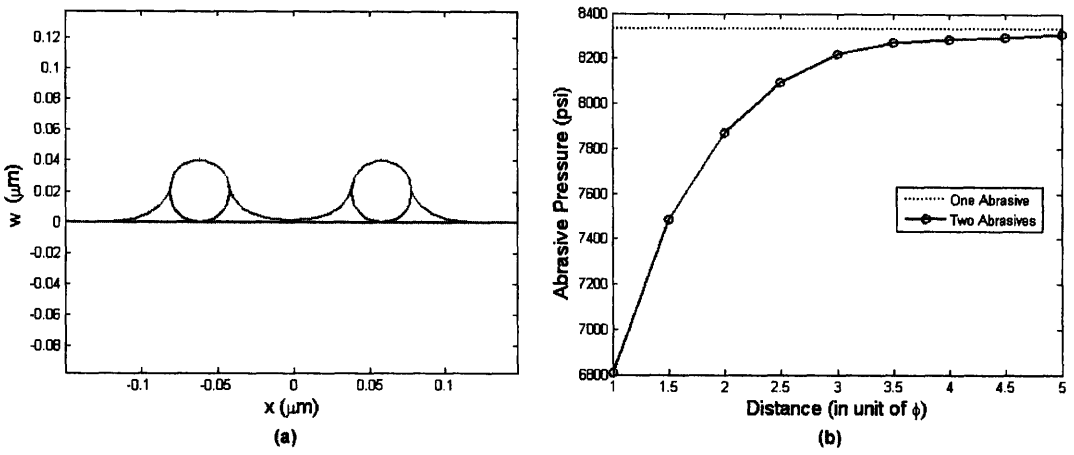



Figure 2-16: (a) Illustration of the interaction between the pad and two abrasives separated by $3\phi$. (b) The dependence of abrasive pressure on separation distance between two abrasives, and the dotted line shows the abrasive pressure in the case of a single abrasive.

abrasives are far away from each other and each of them can be treated as a single abrasive. When the area density is high, the abrasives share all of the applied load and the pad does not touch wafer surface, in which case $P_a = P_{Pad}/\rho$. The area density value dividing these two regimes can be estimated by arranging abrasives on square lattices. The abrasives have little interaction with each other when the lattice length is two times the shielding radius $R$. In that case, the area density is $(\pi\phi^2/4)/(2R)^2 = \frac{\pi}{16}(\frac{\phi}{R})^2$.

In simulation, abrasive positions are randomly generated, and an illustration of a top-down view is shown in Figure 2-17 (a). The same model parameters are used as in the previous section, in which case the shielding radius can be estimated as $2\phi$ from Figure 2-15, and the corresponding area density is $(\pi\phi^2/4)/(4\phi)^2 = \pi/64 = 4.9\%$. For this study, the area density is chosen to be larger than 5%. From the simulation, the average pressure on an abrasive can be estimated, and its ratio to the pad pressure is plotted as a function of area density in Figure 2-17 (b). Two different sizes of abrasives are used for the simulation, as plotted in different colors. The abrasive pressure in the case of a single abrasive is plotted, as well as is the curve $1/\rho$. The result agrees with our analysis:

- For large area densities, the ratio of abrasive pressure to pad pressure fits $1/\rho$ very well.
- The abrasive pressure is independent of abrasive size.
- From the figure, the cross point of single abrasive pressure and $1/\rho$ is 6%, which is close to the estimated 4.9%.

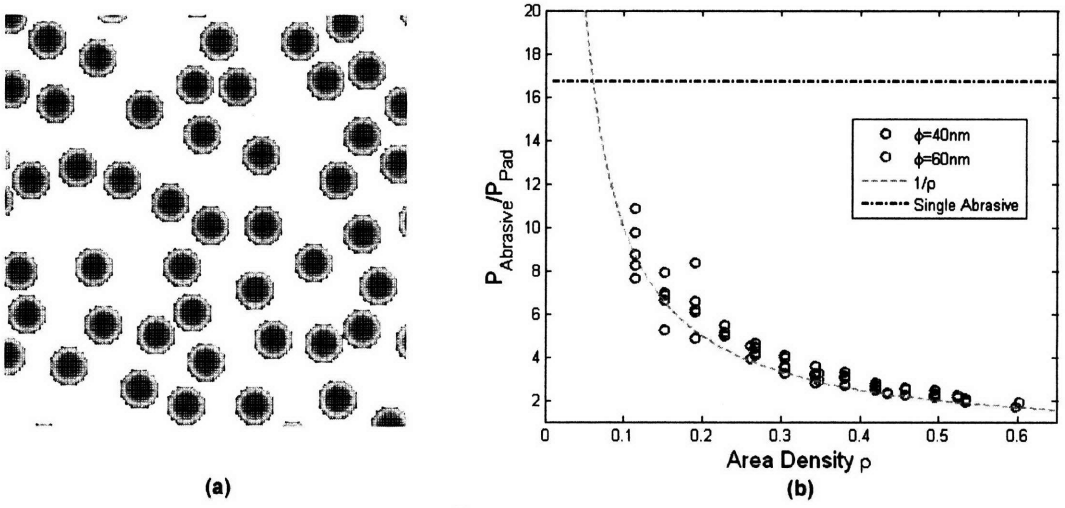



Figure 2-17: (a) Illustration of randomly packed abrasives with the same size. (b) The dependence of average abrasive pressure (in units of $P_{Pad}$) on the area density of abrasives.

Thus, the abrasive pressure for abrasives with the same size $\phi$ can be estimated as $P_{Pad}/\rho$ when the area density is large, and $P_{a,s}(P_{Pad}, E)$ when $\rho$ is small. The transition point can be chosen at $\rho = P_{Pad}/P_{a,s}(P_{Pad}, E)$ to ensure a smooth transition. The analysis can be generalized to truncated abrasives as well, resulting in estimated abrasive pressures given by Equation 2.24.

$$P_a(P_{Pad}, E, r, \rho) = \begin{cases} P_{a,s}(P_{Pad}, E, r) & when \quad \rho < P_{Pad}/P_{a,s}(P_{Pad}, E, r) \\ P_{pad}/\rho & \rho \geq P_{Pad}/P_{a,s}(P_{Pad}, E, r) \end{cases} \tag{2.24}$$

### Multiple Abrasives with a Statistical Size Distribution

In the case of abrasives with a statistical distribution of size, the pressure on each abrasive depends on the sizes of its neighbors and varies case by case, and the com-

putation to simulate all possible configurations is a daunting task. A first order approximation is to consider a single-size-approach that ignores the effect of size distribution and treats all abrasives as identical. This case was analyzed in the previous section, and the abrasive pressure can be obtained by Equation 2.24.

In this section, a two-size-approach is proposed, with the following assumptions and approximations.

- Assume the abrasive concentration in the contact area is $q(\phi)$, i.e., the number of abrasives between $\phi$ and $\phi + d\phi$ per area is $q(\phi)d\phi$. Then the total number of abrasives per area $N$ can be calculated as

$$N = \int_0^\infty d\phi \cdot q(\phi). \tag{2.25}$$

  The area density of abrasives $\rho$ is

$$\rho = \int_0^\infty d\phi \cdot q(\phi) \cdot \frac{\pi}{4}\phi^2. \tag{2.26}$$

- The abrasives are divided into two groups based on their size, $\phi \geq \phi^*$ and $\phi < \phi^*$. Abrasives in the same group are approximated as being identical in size. Let $n_+$, $\phi_+$, and $\rho_+$ denote the number fraction, size, and area density of the larger abrasives, and let $n_-$, $\phi_-$, and $\rho_-$ denote those of the smaller group.

$$\begin{aligned}
n_+ &= \frac{1}{N}\int_{\phi^*}^\infty d\phi \cdot q(\phi) \\
n_- &= \frac{1}{N}\int_0^{\phi^*} d\phi \cdot q(\phi) \\
\phi_+ &= \frac{1}{n_+}\int_{\phi^*}^\infty d\phi \cdot q(\phi) \cdot \phi \\
\phi_- &= \frac{1}{n_-}\int_0^{\phi^*} d\phi \cdot q(\phi) \cdot \phi.
\end{aligned} \tag{2.27}$$

$$\begin{aligned}
\rho_+ &= \int_{\phi^*}^\infty d\phi \cdot q(\phi) \cdot \frac{\pi}{4}\phi^2 \\
\rho_- &= \int_0^{\phi^*} d\phi \cdot q(\phi) \cdot \frac{\pi}{4}\phi^2
\end{aligned} \tag{2.28}$$

- The area density is assumed to be large, i.e., the pad is supported entirely by the abrasives. This is a reasonable assumption as most CMP processes are operated in a closely-packed abrasive particle case.

- In the two-size-abrasive system, the smaller abrasives are approximated as a film of thickness $\phi_-$, and the larger abrasives are approximated as truncated abrasives with size $\phi_+$ and truncation ratio $r = 1 - \phi_-/\phi_+$. The approximation flow is illustrated in Figure 2-18. Thus, the pressure on the larger abrasives can be estimated by the result for identical truncated abrasives in the previous section.

$$\begin{cases} P_+ &= P_a(P_{Pad}, E, r = 1 - \frac{\phi_-}{\phi_+}, \rho_+) \\ P_- &= \frac{1}{\rho_-}(P_{Pad} - P_+ \cdot \rho_+) \end{cases} \tag{2.29}$$

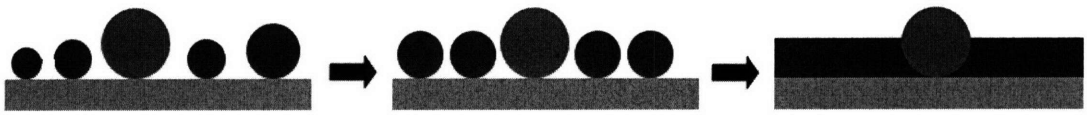


Figure 2-18: When calculating the pressure on the large abrasives(the red one), the effects of the neighboring abrasives can be approximated as a uniform film with the average abrasive diameter as its thickness.

This two-size-approach offers a convenient tool to approximate and study the impact of abrasive size distribution. The first application is to estimate the proportion of active abrasives, i.e., the abrasives which have non-zero pressure and are active in polishing. On the other hand, the inactive abrasives have zero pressure and do not contribute to polishing. The total load is supported only by the active abrasives, so the active area density is $\rho_+$ and abrasive pressure is $P_{Pad}/\rho_+$, which is different from the one-size-approach. The inactive abrasives are small abrasives, and its proportion can be estimated by finding the maximum $\phi^*$ subject to $P_- = 0$. In other words, at $max_{P_-=0}(\phi^*)$, all the abrasives in the smaller group with $\phi < \phi^*$ have zero pressure, but including any marginally larger abrasives makes $P_-$ positive. Let $\phi_{min}(P_{Pad}, q(\phi)) = max_{P_-=0}(\phi^*)$ be the minimum size of active abrasives; $\phi_{min}$ depends on pad pressure and abrasive size distribution.

The same approach can also be used to estimate the number of abrasives per area

$N(\hat{P})$ with pressure larger than a given value $\hat{P}$. If the value $\hat{\phi}^*$ can be found such that $P_+(\hat{\phi}^*) = \hat{P}$, $N(\hat{P})$ can be estimated as $n_+(\hat{\phi}^*)$. Denote $P_+^{-1}(P)$ as the inverse function of $P_+(\phi^*)$, and we have $N(\hat{P}) = n_+(P_+^{-1}(\hat{P}))$.

Assume that the abrasive size observes a Gaussian distribution $q(\phi) = \frac{C}{\sigma} \cdot exp\left(-\frac{(\phi-\phi_0)^2}{2\sigma^2}\right)$, where $C$ is a constant. If the abrasive area density is $\rho$, $C$ can be determined and $q(\phi)$ can be written as

$$q(\phi) = \left(\frac{2}{\pi}\right)^{\frac{3}{2}} \cdot \frac{\rho}{\sigma(\sigma^2 + \phi_0^2)} \cdot e^{-\frac{(\phi-\phi_0)^2}{2\sigma^2}} \tag{2.30}$$

Using the distribution, the following can be derived.

$$\begin{aligned}
n_+ &= \frac{4}{\pi} \cdot \frac{\rho}{\phi_0^2 + \sigma^2} \cdot (1 - \Phi(x^*)) \\
n_- &= \frac{4}{\pi} \cdot \frac{\rho}{\phi_0^2 + \sigma^2} \cdot \Phi(x^*) \\
\phi_+ &= \phi_0 + \frac{e^{-\frac{1}{2}x^{*2}}}{\sqrt{2\pi} \cdot (1 - \Phi(x^*))}\sigma \\
\phi_- &= \phi_0 - \frac{e^{-\frac{1}{2}x^{*2}}}{\sqrt{2\pi} \cdot \Phi(x^*)}\sigma \\
\rho_+ &= \frac{1}{\phi_0^2 + \sigma^2} \cdot \left[(1 - \Phi(x^*))\phi_0^2 + \sqrt{\frac{2}{\pi}} e^{-\frac{x^{*2}}{2}} \phi_0 \sigma + \left(1 - \Phi(x^*) + \frac{x^*}{\sqrt{2\pi}} e^{-\frac{x^{*2}}{2}}\right)\sigma^2\right] \\
\rho_- &= \frac{1}{\phi_0^2 + \sigma^2} \cdot \left[\Phi(x^*)\phi_0^2 - \sqrt{\frac{2}{\pi}} e^{-\frac{x^{*2}}{2}} \phi_0 \sigma + \left(\Phi(x^*) - \frac{x^*}{\sqrt{2\pi}} e^{-\frac{x^{*2}}{2}}\right)\sigma^2\right]
\end{aligned} \tag{2.31}$$

where $\Phi(x)$ is the cumulative density function of a standard Gaussian distribution, and $x^* = \frac{\phi^* - \phi_0}{\sigma}$.

To show the dependence of $P_+$ and $P_-$ on $\phi^*$, the following parameters are used: $E = 120\ MPa$, $\nu = 0.3$, $P_{Pad} = 400,800\ psi$, $\phi_0 = 50\ nm$, $\sigma = 10\ nm$, and $\rho = 0.9$. The results are shown in Figure 2-19 (a). The figure shows $P_-$ remains zero until some value of $\phi^*$, which has been defined as $\phi_{min}$. The figure also indicates that the value of $\phi_{min}$ decreases with increasing $P_{Pad}$. The decreasing part of $P_+/P_{Pad}$ is due to the decrease in $r$, as shown in Figure 2-19 (b).

The dependence of $\phi_{min}$ on pad pressure can be estimated by repeating the simulation over different pressure values. The result in shown in Figure 2-20 (a). Here

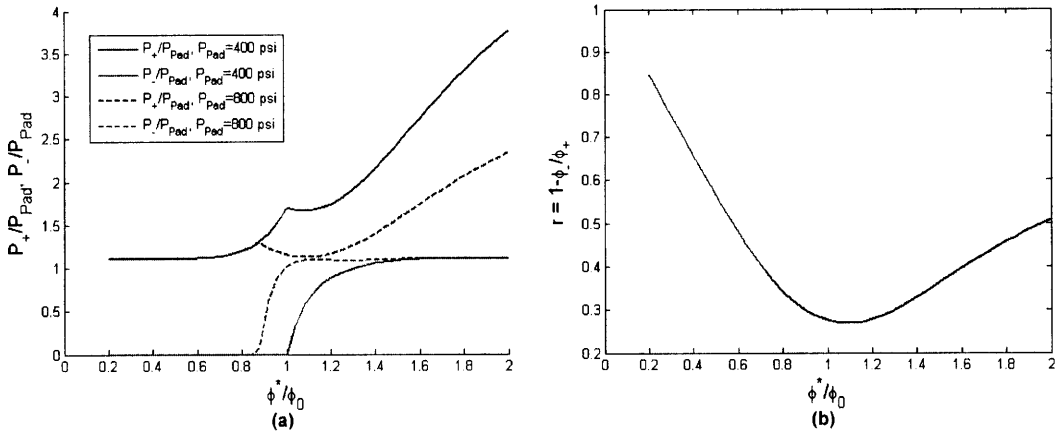



Figure 2-19: (a) The dependence of $P_+/P_{Pad}$ and $P_-/P_{Pad}$ on $\phi^*/\phi_0$ for two different values of $P_{Pad}$. (b) The dependence of the truncated ratio $r = 1 - \frac{\phi_-}{\phi_+}$ on $\phi^*/\phi_0$.

the decreasing trend is expected, as the pad surface bends less with smaller pressure and thus touches fewwer abrasives. Using the result of $\phi_{min}$, the dependence of $\rho_+$ on pad pressure can also be estimated; as shown in Figure 2-20 (b), the area density increases with $P_{Pad}$, which indicates that more abrasives are in contact under higher pressures.

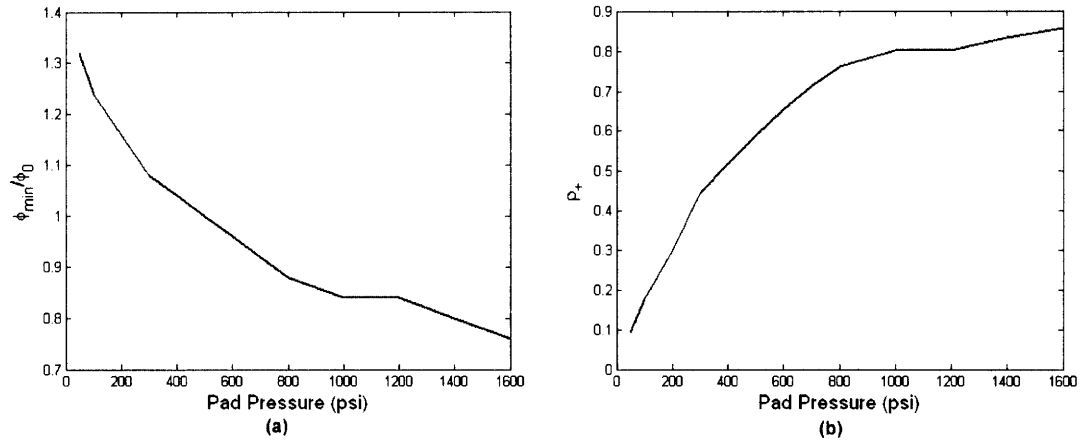



Figure 2-20: (a) The dependence of $\phi_{min}$ on $P_{Pad}$, (b) The dependence of $\rho_+$ on $P_{Pad}$.

### 2.4.5 Pad-Abrasive-Wafer Interaction

Our study of the pad-abrasive-wafer interaction focuses on two issues: the abrasive occupation rate in the contact area and the relative velocity of abrasive with respect

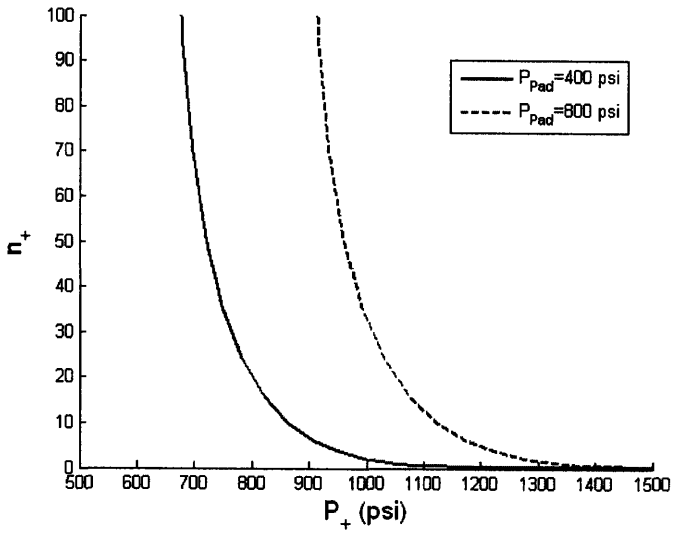



Figure 2-21: The dependence of $n_+$ on $P_+$. $n_+$ is in the units of number of abrasives per unit area.

to the wafer.

### Abrasive Occupation Rate

The concentration of abrasives in the contact area results from the balance between the rates of abrasives entering and leaving this area, as illustrated in Figure 2-22.

- The entering rate is proportional to the rate of abrasives hitting the asperity edge and the absorption probability $p_{ab}$. The number of abrasive hitting the asperity edge equals the abrasive concentration $n$ in the slurry, times its velocity relative to the pad $\vec{v}_{out}$, where the subscript "out" denotes that the abrasive is outside of the contact area.

- The emission rate is proportional to the product of the abrasive concentration $q$ in the contact area and the relative velocity of the abrasive of the abrasive $\vec{v}_{in}$ with respect to the pad, while it is within the contact area.

We seek to model several developments in order to estimate the abrasive entrance rate. First, we need the velocity distribution $n(\vec{v}_{out}, \phi)$ of abrasives in the slurry, where $\vec{v}_{out}$ is the velocity of the abrasive relative to the pad, and $\phi$ is the diameter

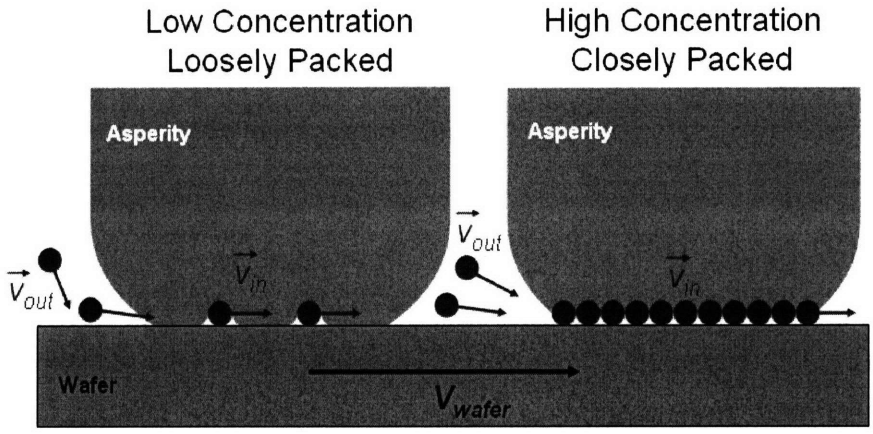



Figure 2-22: Illustration of absorption and emission dynamics of abrasives in the contact area between wafer and pad asperity.

of the abrasive. Second, we assume an effective hitting area $\partial S$. For an absorption probability $p_{ab}(\vec{v}_{out}, \phi, d\vec{S}, \vec{v}_{wafer})$, the rate of abrasives with diameter $\phi$ entering the contact area is then

$$\int d\vec{v}_{out} \cdot n(\vec{v}_{out}, \phi) \cdot \oint_{\partial S} \left(-d\vec{S} \cdot \vec{v}_{out}\right) \cdot p_{ab}(\vec{v}_{out}, \phi, d\vec{S}, \vec{v}_{wafer}). \tag{2.32}$$

Similarly, if the velocity distribution in the contact area is $q(\vec{v}_{in}, \phi)$, the rate of abrasives with diameter $\phi$ leaving the contact area is

$$\int d\vec{v}_{in} \cdot q(\vec{v}_{in}, \phi) \cdot \oint_{\partial S} d\vec{S} \cdot \vec{v}_{in}. \tag{2.33}$$

From the equilibrium of the two rates, the concentration of the abrasive particles in the contact area can be determined.

Solving the problem is difficult, but we can obtain some qualitative conclusions from the dynamics. In the slurry, the abrasive particles are driven by the fluid flow, whose velocity is close to $\vec{v}_{wafer}$ near the wafer surface, and there is little friction resistance. In the contact area, however, the abrasive particles are driven through its contact with the wafer surface. As the Young's modulus of the pad material is much less than those of abrasive and wafer materials, the abrasives are mainly embedded into the pad asperity and their indentations into the wafer are small. Therefore, the abrasives trapped between the contact area likely have a smaller force and much

more frictional resistance than abrasives in the slurry. As a result, $\vec{v}_{out}$ is expected to be much larger than $\vec{v}_{in}$, and the concentration of abrasives in the contact area is expected to be higher and possibly much higher than that in the slurry. Thus, unless the abrasive concentration in the slurry is very low, the abrasives are closely packed in the contact areas, as illustrated in Figure 2-22. Friction measurement shows that the friction force curve is nearly flat for abrasives with weight percentage concentration between about 2% to 15% [72]. Our qualitative reasoning based on the particle dynamics agrees with the experiment results.

#### Relative Velocity of Abrasives to Wafer

As the case of occupation rate, it is difficult to derive the relative velocity of abrasives with respect to the wafer. Qualitatively, the abrasives are embedded in pad asperities, and such abrasives are likely to have a small relative velocity with respect to the pad, as noted in the discussion of $\vec{v}_{in}$ above. Therefore, the relative velocity of abrasives with respect to the wafer is approximately the relative velocity of the pad with respect to the wafer. Liu et al. estimated the velocity of abrasives as

$$\vec{v}_{abrasive} = \frac{HB_{pad}}{HB_{wafer} + HB_{pad}}\vec{v}_{wafer} + \frac{HB_{wafer}}{HB_{wafer} + HB_{pad}}\vec{v}_{pad}, \tag{2.34}$$

where $HB$ is the Brinell hardness values of the pad or wafer surface. The wafer surface has a substantially higher value of Brinell hardness than that of the polishing pad, thus the same conclusion is drawn.

### 2.4.6 Pad-Wafer Interaction

Recently efforts have been reported to measure the asperity distribution on the pad surface [11], the wearing of asperities during polishing [36], the cutting of asperities during conditioning [10], and the impact of asperity distribution on material removal [37]. The asperity distribution depends on both pad material and pad conditioning. Although the distribution varies, asperities are of around 40 $\mu m$ in height and of around 50 $\mu m$ in width or diameter. Figure 2-23 shows an SEM image of a pad cross section.

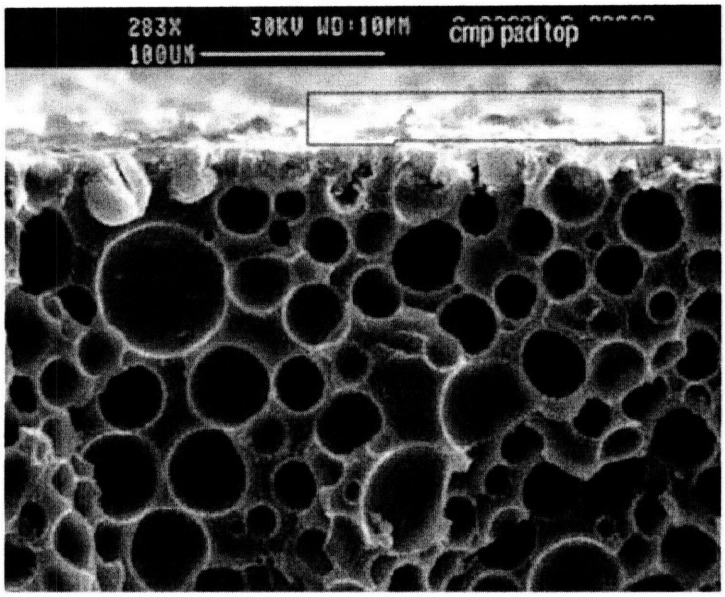



Figure 2-23: Scanning electron micrograph cross-section of a used, conditioned void-filled polyurethane polishing pad. Surface asperities can be seen at the top of the image. The scale bar at the top center is 100 $\mu m$ (0.1 $mm$) long. Voids average about 30 $\mu m$ in diameter and occupy about 60% of a planar cross-section [10].

The pad surface can be considered as a nominally flat surface covered with asperities of various shapes and different heights. It is usually easier to assume some asperity shape and use the height of the asperity as a characteristic parameter. The problem can then be broken down to two steps.

First, we assume a certain asperity shape and solve the elastic deformation problem of a single asperity with height $h$ when pressed upon the wafer surface. If the asperity deformation is $\delta$, we can express the following terms as functions of $\delta$, as illustrated in Figure 2-24: the contact area $a(\delta)$, the total load as $L(\delta)$, and the pressure distribution in the contact area $P(x, y; \delta)$. Another way to express the pressure distribution is the ratio of area having some pressure $s(P; \delta)$, i.e., the fraction of some area having pressure between $P$ and $P + dP$ is $s(P; \delta)dP$, where $s(P; \delta)$ denotes the pressure distribution of a single asperity and the lower case is used to be distinguish from the pressure distribution $S(P; \delta)$ considering all contact areas.

Second, we assume an asperity height distribution or probability density function $\xi(h)$, i.e., the number of asperities per unit area with height between $h$ and $h + dh$

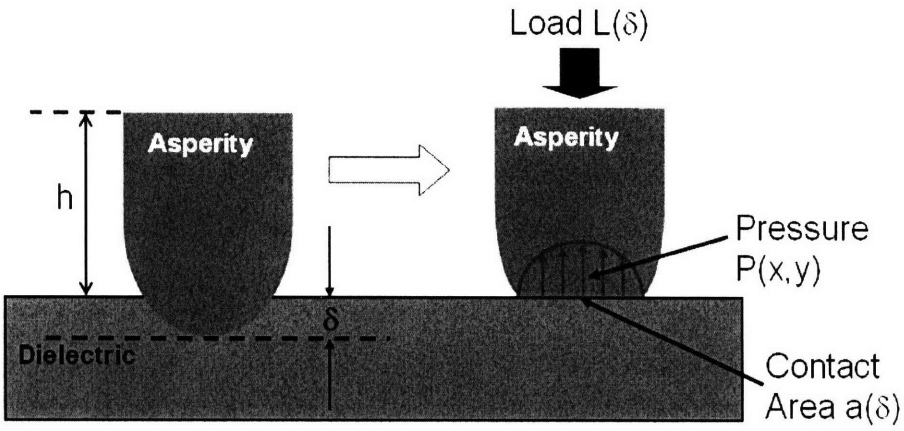



Figure 2-24: Diagram of a single asperity being compressed.

is $\xi(h)dh$. If the distance between the wafer and the nominal surface of the pad is $d$, the asperities with height greater than $d$ will be in contact with the wafer surface. The number of asperities in contact is

$$n = N \int_d^\infty \xi(h)dh, \tag{2.35}$$

where $N$ is the number of asperities per unit area.

For the asperity with height $h > d$, the deformation is $\delta = h - d$. The total contact area is

$$A = N \int_d^\infty a(h-d)\xi(h)dh. \tag{2.36}$$

For an applied pressure of P, the distance $d$ can be obtained as

$$P = N \int_d^\infty L(h-d)\xi(h)dh. \tag{2.37}$$

Similarly, we can obtain the area density distribution as

$$S(P) = N \int_d^\infty s(P; h-d)\xi(h)dh \tag{2.38}$$

Greenwood [66] assumes that the asperities have spherical surfaces, all with the same radius $R$, and the contact is Hertzian. Greenwood does not consider the pressure distribution, but it is not hard to calculate. Based on the same assumptions as

Greenwood and using the Hertzian results from Timonshenko [19], we find:

$$\begin{cases} a(\delta) &= \pi R\delta \\ L(\delta) &= \frac{4}{3}\frac{E}{1-\nu^2}R^{1/2}\delta^{3/2} \\ P(x,y) &= \begin{cases} P_c\left(1-\frac{\pi(x^2+y^2)}{a(\delta)}\right) & when \quad \pi(x^2+y^2) \le a(\delta) \\ 0 & \pi(x^2+y^2) > a(\delta) \end{cases} \\ s(P,\delta) &= \begin{cases} \frac{2P}{P_c^2}a(\delta) = \frac{\pi^3R^2}{2}(\frac{1-\nu^2}{E})^2 \cdot P & when \quad 0 < P \le P_c \\ 0 & P > P_c \end{cases} \end{cases} \tag{2.39}$$

where $E$ is the Young's modulus, $\nu$ is the Poisson's ratio of the asperity, and $P_c = \frac{3}{2}\frac{L(\delta)}{a(\delta)}$ is the pressure at the center of the contact area. Here, it is assumed that wafer material is much more rigid than that of the pad asperity. In the general case, $E/(1-\nu^2)$ can be replaced by $1/\left(\frac{1-\nu_1^2}{E_1}+\frac{1-\nu_2^2}{E_2}\right)$.

Measurements have shown that the asperities distribution follows an exponential decay [11] for large asperity heights, or

$$\xi(z) = \xi_0 e^{-\beta z}, \tag{2.40}$$

where $\beta$ characterizes the decay.

If the applied pressure is $P_0$, the distance between the wafer and the nominal pad surface can be determined by

$$d = \frac{1}{\beta}\log\left(\frac{E}{1-\nu^2}\frac{N\xi_0}{P_0}\sqrt{\frac{\pi R}{\beta^5}}\right). \tag{2.41}$$

The total contact area and pressure distribution are found to be

$$\begin{cases} n &= \frac{1-\nu^2}{E}\sqrt{\frac{\beta^3}{\pi R}}\cdot P_0 \\ A(P_0) &= \frac{1-\nu^2}{E}\sqrt{\pi R\beta}\cdot P_0 \\ S(P) &= \frac{1}{2}P_0\left(\frac{1-\nu^2}{E}\right)^3\sqrt{\pi^5\beta^3R^3}\cdot P\cdot e^{-\beta\frac{\pi^2R}{4}\left(\frac{1-\nu^2}{E}\right)^2P^2} \end{cases} \tag{2.42}$$

The density distribution $S(P)$ has a peak at

$$P_{Peak} = \frac{2}{\pi}\frac{P_0}{A} = \frac{E}{1-\nu^2}\sqrt{\frac{2}{\pi^2\beta R}}, \tag{2.43}$$

which is independent of $P_0$.

From Figure 2-25, we can estimate an example pad surface asperity distribution as $\phi(z) = 0.1exp(-0.34z)$ and $N = 2 \times 10^8/m^2$. We also estimate an asperity radius of $r = 50\ \mu m/m$. If we choose Young's modulus $E/(1-\nu^2) = 119MPa$ and vary the applied pressure from 1 *psi* to 9 *psi*, the contact area ratio and area density distribution $S(P)$ are shown in Figure 2-26. If we choose pressure at 5 *psi* and vary Young's modulus, the result is shown in Figure 2-27. If we fix pressure at 5 *psi*, Young's modulus at 119 $MPa$, and vary $\beta$, the result is shown is Figure 2-28.

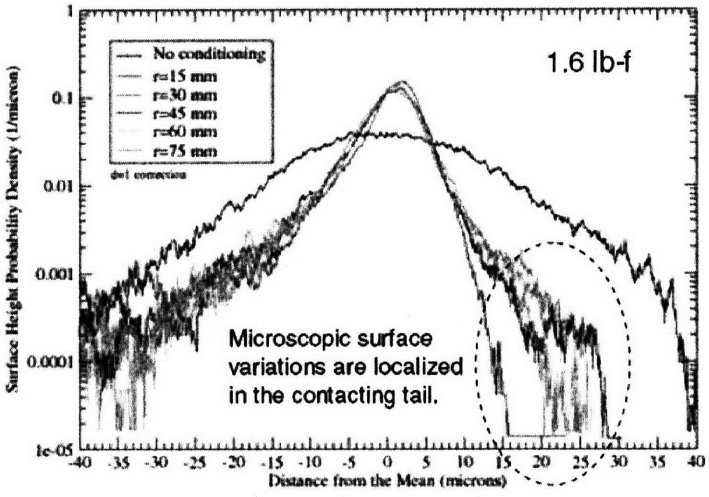



Figure 2-25: Surface asperities height distribution obtained by interferometry measurement [11]

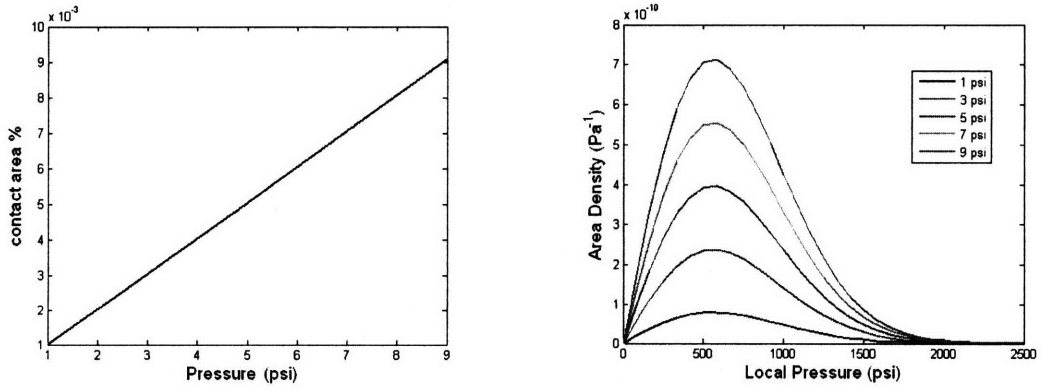



Figure 2-26: (a) The dependence of contact area fraction $A(P_0)$ on applied pressure. (b) The area density distribution of pressure $S(P)$ for different applied pressure.

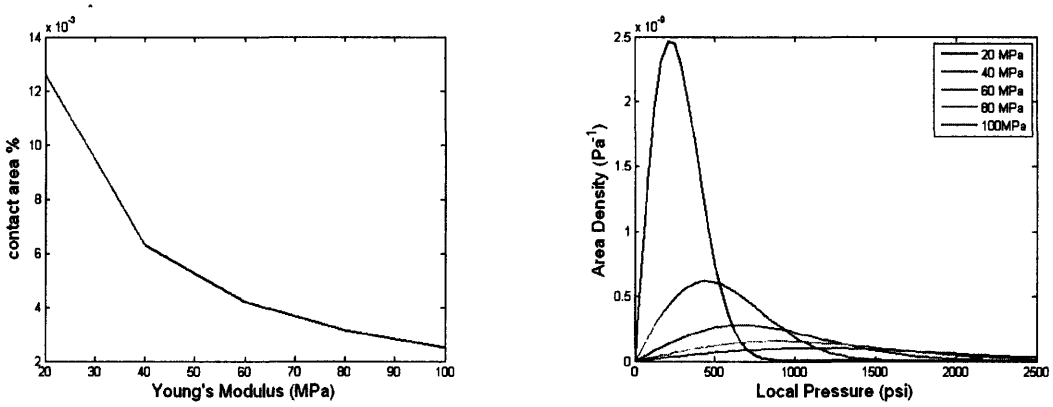



Figure 2-27: (a) The dependence of contact area fraction $A(P_0)$ on Young's modulus. (b) The area density distribution of pressure $S(P)$ for different Young's modulus.

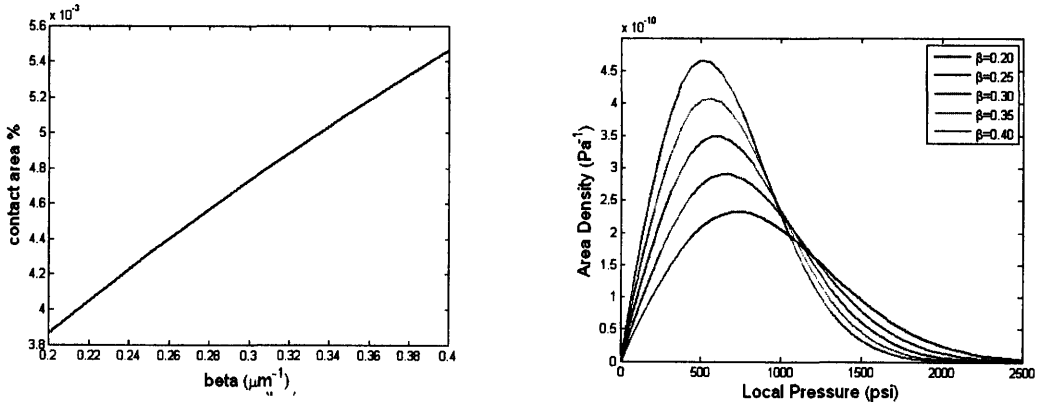



Figure 2-28: (a) The dependence of contact area fraction $A(P_0)$ on values of $\beta$. (b) The area density distribution of pressure $S(P)$ for different value of $\beta$.

Contact area ratio can be measured using confocal reflectance interference contrast microscopy (C-RICM) [12], and Figure 2-29 shows the C-RICM image sequence of VP3000$^{TM}$ under increasing pressure. Figure 2-30 shows the dependence of contact ratio on pressure for three different polishing pads [12]. The experimental result is consistent with our modeling of pad-wafer interaction. The measurement defines "in contact" to be when the asperity reaches a 4.7 $\mu m$ optical slice or range in height near the surface of the wafer, and as a result, the measured contact ratio is higher than the model prediction.

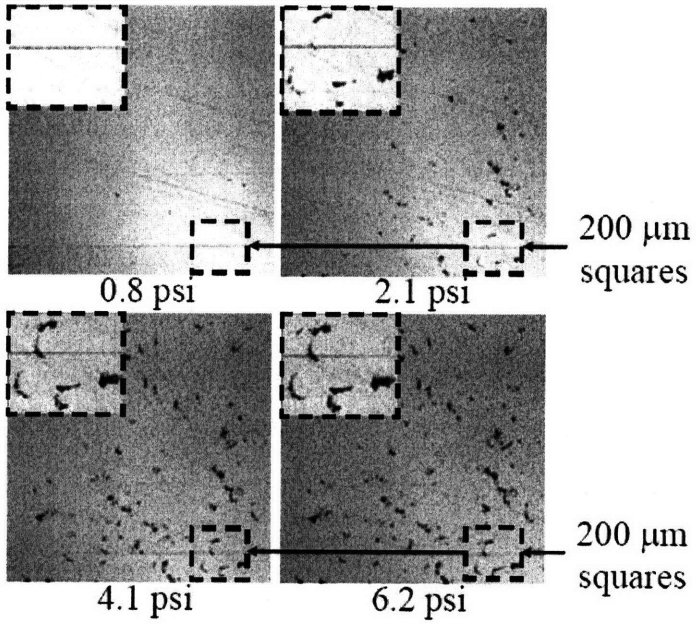



Figure 2-29: Confocal reflectance interference contrast microscopy (C-RICM) image sequence of VP3000$^{TM}$ pad conditioned for 30 $min$ at increasing applied pressure. Straight lines across images are manufacturing scratches on cover slip, excluded from contact area calculations. (Plan-Apochromat 10x/0.45 objective, optical slice thickness 4.7 $\mu m$) [12].

### 2.4.7 Particle-Level Model of CMP

A particle-level CMP model can finally be generated by integrating the contributions of the interactions and models discussed above. Below is a summary of of the individual model components.

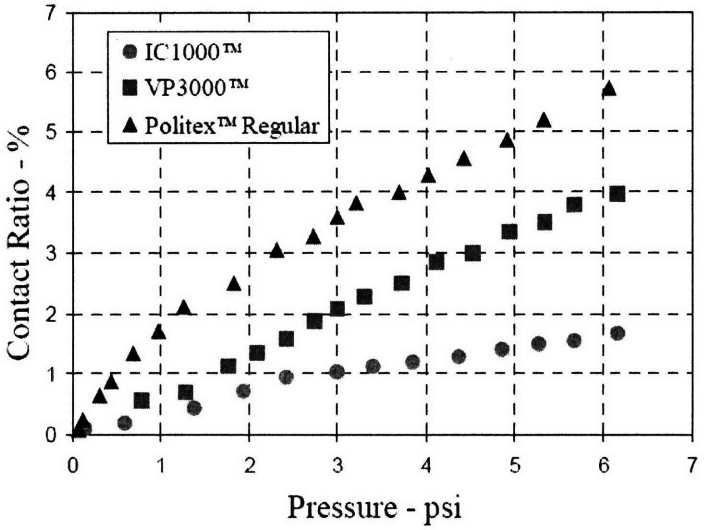



Figure 2-30: Contact ratio response to pressure for IC1000$^{TM}$ and VP3000$^{TM}$ pads conditioned for 30 *min* and Politex$^{TM}$ Regular pad, measured by C-RICM contact method. Contact areas at 3 psi are 1.0%, 2.1%, and 3.6% for IC1000, VP3000, and Politex pads respectively [12].

- Based on the chemical tooth mechanism, the removal rate of a contact area fraction $A_c$ with contact pressure $P_c$ is

$$K^*(A_c, P_c) = p \cdot A_c \cdot v_{ab,wf}, \tag{2.44}$$

where $p$ is the probability of removing a silicon atom when in contact, and $v_{ab,wf}$ is the relative velocity between the abrasive and the wafer.

- The contact area between an abrasive and the wafer surface can be determined using the Hertzian contact as

$$A_c = \pi\phi^2 \left(\frac{3\pi}{4}\frac{P_a}{E'}\right)^{2/3}, \tag{2.45}$$

where $\phi$ is the diameter of the abrasive and $P_a$ is the average pressure on the abrasive.

- In the study of the pad-abrasive-wafer interaction, $v_{ab,wf}$ is estimated to be nearly the same as the relative velocity $v$ between the pad and wafer.

- In most CMP processes, the abrasive concentration is reasonably high so that

the applied load is supported entirely by the abrasives. In the first order approximation, in which all abrasives are assumed to be identical in size, the abrasive pressure is $P_a = P_p/\rho$, where $\rho$ is the area density of abrasives.

- The defect rate $d(P_p, P_a^{th})$ can be determined once the abrasive concentration distribution in the contact area is known.

- The area distribution of pad pressure can be determined from the pad-wafer interaction.

$$S(P) = \frac{1}{2}P_0\left(\frac{1-\nu_p^2}{E_p}\right)^3\sqrt{\pi^5\beta^3R^3}\cdot P\cdot e^{-\beta\frac{\pi^2R}{4}\left(\frac{1-\nu_p^2}{E_p}\right)^2P^2}, \tag{2.46}$$

where $E_p$ is the Young's modulus of the pad, $\nu_p$ is the Poisson's ratio, $R$ is the radius of the pad asperity tip, and $\beta$ is the exponent in the asperity height distribution.

Combining these dependences, the removal rate $K$ can be solved as

$$K = \frac{1}{\sqrt{3}}\Gamma(\frac{2}{3})\left(\frac{2\pi^2}{3E'^2}\right)^{1/3}p\cdot P_0\cdot v\cdot\left(\frac{1-\nu_p^2}{E_p}\right)^{1/3}\cdot\beta^{1/6}\cdot R^{1/6}\cdot\rho^{1/3}, \tag{2.47}$$

where $E'$ is the effective Young's modulus of wafer and abrasive, $p$ is the removal probability in the chemical reaction, $P_0$ is the applied pressure, $v$ is the relative velocity between pad and wafer, $E_p$ is the Young's modulus of the pad asperity, $\beta$ is the exponent of asperity height distribution, $R$ is the asperity radius, and $\rho$ is the area density of abrasives in the contact area.

To model the effect of the abrasive size distribution, the pad-abrasive interaction can be considered using the two-size-approach. If the size distribution is Gaussian with mean $\phi_0$ and variance $\sigma^2$, we can estimate $\rho_+(P_{Pad}, E, \phi_0, \sigma)$, the area density of active abrasives, and $N(\hat{P}; P_{Pad}, E, \phi_0, \sigma)$, the number of abrasive with pressure larger than $\hat{P}$. Thus, the total removal rate is

$$K = \left(\frac{6\pi}{E'}\right)^{2/3}\cdot p\cdot v_{ab,wf}\cdot\int_0^\infty dP_{Pad}\cdot S(P_{Pad})\cdot P_{Pad}^{2/3}\cdot\rho_+^{1/3}(P_{Pad}, E, \phi_0, \sigma). \tag{2.48}$$

The total number of abrasives with contact pressure higher than $\hat{P}$ is

$$N_{total}(\hat{P}) = \int_0^\infty dP_{Pad} \cdot S(P_{Pad}) \cdot N(\hat{P}; P_{Pad}, E, \phi_0, \sigma). \tag{2.49}$$

Thus, if the yield stress of the dielectric material is $P_Y$, the defect number is expected to be proportional to $N_{total}(P_Y)$. Unfortunately, $\rho_+(P_{Pad}, E, \phi_0, \sigma)$ and $N(\hat{P}; P_{Pad}, E, \phi_0, \sigma)$ do not have analytical form, and we cannot obtain a closed form solution for the removal rate and defect rate.

### 2.4.8 The Dependence of CMP on Input Variables

The particle-level model predicts that removal rate is proportional to the product of applied pressure and velocity. This is consistent with the well-known Preston equation [30], and experiments [8] [53] [54] on dielectric CMP supports the relationship.

The model also suggests that the removal rate is inversely proportional to $\left(\frac{E_p}{1-\nu^2}\right)^{1/3}$. Empirically, softer pads do have higher removal rates [12], although the exact relationship is not available due to limited experimental data.

The model suggests that removal rate is proportional to $\beta^{1/6}$. Borucki [11] found that conditioning increases the value of $\beta$, and in the CMP process, polishing without conditioning suffers from decaying removal rate, which supports the model trend. Again, the exact relationship has not been tested in experiment, and neither have the relationships between removal rate and $R$ or $\rho$.

Experiments [8] [6] show that removal rate varies with abrasive size; however, the relationship is not explicitly included in our particle-level model. Revisiting the interactions in the modeling framework, there are two places where the abrasive size can appear. The area density of abrasives in contact areas may depend on the abrasive size, as the entering and exiting mechanism can be affected by the size of abrasive particles. Also, if the function $K^*$ depend on contact pressure $P_c$, the removal rate will also depends on the abrasive size. Future extensions to the model could potentially include asperity size dependences.

## 2.5 Summary

In this chapter, a framework is established to study the physics of CMP. The framework approaches the multi-scale problem in a top-down fashion, and decomposes the problem into pairs of interactions, which occur at different scales. Each interaction is studied, and they are integrated to build an overall particle-level model. The dependence of removal rate on various parameters predicted by the model agrees with experiments in general. Although the framework is developed to model CMP of dielectric materials, it should be applicable to metal CMP and only the material removal mechanism component needs to be re-modeled. The particle-level model of CMP helps explain key dependencies in the CMP process. In the next chapter, the particle-level model also helps inform the development of die-level CMP models.

# Chapter 3

# Die Level Modeling of CMP

An ideal CMP process has uniform polishing rate and leaves a perfectly flat surface afterwards. In practice, the CMP process suffers from various chip-level non-uniformity effects including with-in-die variation of film thickness and residual (non-zero) final step height (Figure 3-1). With-in-die variation of the film thickness can create undesirable long-range topography of an interconnect dielectric film, for example, which can cause problems for the lithography step, especially in chip manufacturing with multi-layer interconnect. In the STI CMP process, with-in-die variation in polishing surface films may cause failures, either by failure to clear overburden oxide, or through excessive erosion of nitride. The residual step height is a variation at the feature-scale, and in STI CMP can result in the failure to clear local features. In addition to topography variation, oxide dishing in STI CMP can cause undesirable leakage current in transistors formed inactive regions between the shallow trenches. As feature dimensions continue to shrinking, CMP induced pattern dependent yield and reliability issues become more complex [73].

Die-level CMP models seeks to explain and model the planarization of layout structures, and to predict the post-CMP surface profile for the entire IC chip. An accurate die-level model is an essential tool to improve product yield, minimize material wastes, and reduce environmental impact with shorter development cycle and less cost. A die-level CMP model provides the layout designer with instant feedback on how the chip will planarize in the fab and guides the designer in making the layout fab-friendly; the model helps product engineers to optimize the process by choosing

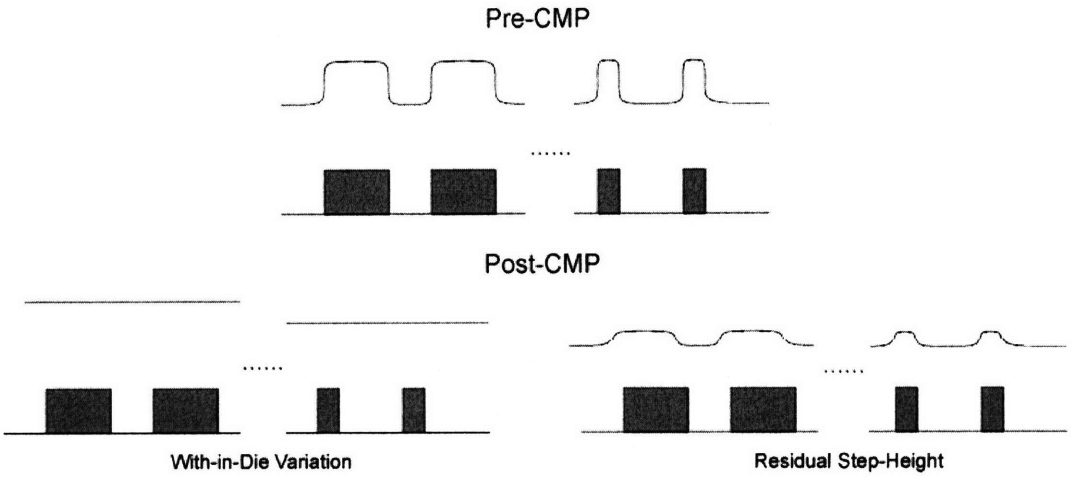



Figure 3-1: Illustration of die-level non-uniformity: (top) a diagram of the pre-CMP surface profile; (lower left) with-in-die variation of film thickness after CMP, and (lower right) residual step-height after CMP.

the proper thickness of deposited materials (oxide, copper, etc.) and process condition (pressure, velocity, etc.); and the model gives process engineers a better estimate of the process window and improves verification efficiency by pointing to potential weak spots. A physically-based die-level model can be used to improve design of consumables more efficiently than trial-and-error.

In this chapter, we focus on the physical basis and methods for die-level modeling of CMP. Section 3.1 reviews the previous feature-level and die-level models. Section 3.2 discusses the objectives of die-level CMP models and establishes the modeling framework. Section 3.3 proposes a physically-based model and a semi-empirical model. The physically-based model explicitly analyzes the elastic contribution of the pad including pad surface asperities, and the semi-empirical model improves upon the previous pattern-density step-height (PDSH) models by making realistic assumptions and approximations, and improving the ease of computation. Section 3.4 applies the semi-empirical exponential PDSH model to simulate the polishing of single material or dual material system, using either conventional or non-conventional slurries. Section 3.5 applies the physically based die-level models to study the effect of pad properties and applied pressure on planarization performance, as well as the impact of initial surface topography. Section 3.6 verifies the two die-level models by comparing

model predictions with experimental data.

## 3.1 Review of Feature-Level and Die-Level Models

Feature-level CMP models study closely the polishing of one or a few features, and are generally physically-based. Such feature-level models provide details of the surface profile evolution during CMP; however, they are computationally intensive and cannot be applied to the entire die, which contains tens of millions of features. Die-level models, in contrast, ignore or aggregate the details of individual features and focus instead on statistical or averaged descriptions of surface evolution across the chip. The die-level models draw insights from feature-level models, and approximate and apply them to model the entire die.

### 3.1.1 Feature-Level CMP Models

In the existing literature, the feature-level models typically assume a uniform chemical contribution and use Preston's equation [30] as the starting point, which proposes a linear dependence of removal rate on pressure. The mechanical contribution to feature evolution is attributed to the elasticity of the polishing pad. A contact wear approach has been used to model the polishing pad by Chekina [74] and Yoshida [75], in which the pad is assumed to be an infinitely-thick elastic body and the asperities are ignored. Given a surface profile, a contact model can determine the elastic deformation of the pad and the contact pressures on the wafer. Vlassak [76] models the asperity-deformation using Herzian contact and the bulk part of the pad using a contact wear model. These models, however, are typically intended to study polishing in one or a few feature structures. Modeling all of the feature structures on a die is a formidable challenge. For example, direct modeling of minimum feature sizes less than 0.13 $\mu m$ across a 20 $mm$ by 20 $mm$ die would require a matrix having at least $10^{11}$ elements. This motivates the need for approximate approaches for the prediction of surface topography evolution across the entire chip.

### 3.1.2 Die-Level CMP Models

Before reviewing the previous and new die-level models, let us address the concept of pattern-density, which is essentially a statistical measure of the local features on a chip. Figure 3-2(a) shows a cross-section of the wafer surface before an oxide CMP process. In a typical semiconductor process, the geometric features can have various shapes when the 2D surface is viewed from the top down. However, vertically the raised areas for any given film layer usually have the same height. Thus, they can be classified as either raised (up) or trench (down) areas, as illustrated in Figure 3-2(a). Local pattern-density $\rho^*$ is defined as the area fraction of raised area, or the ratio of raised area to total area. The limit of the pattern-density, as the averaging (total) area approaches zero, is either one or zero depending on whether one is in a raised or trench area. What is important in practice is the pattern-density value computed using a proper averaging function and total averaging area, where "proper" is based based on the physics of the CMP process. The pattern-density calculated in this way is referred to as the *effective pattern-density* $\rho$, in contrast to the local pattern-density, which is usually calculated by equally weighted averaging over a small square unit area:

$$\rho^* = \frac{Area_{raised}}{Area_{total}}. \tag{3.1}$$

In this chapter and the next, "pattern-density" always refers to effective pattern-density, $\rho$, in contrast to the local pattern-density, $\rho^*$.

A pattern-density based dielectric CMP model has been developed by Stine et al. [13]. Stine observes that the removal rate in a low pattern-density region is comparatively higher than that on a more dense region, and proposes that by choosing a good averaging function, the removal rate $K_u$ of raised/up area is inversely proportional to the effective pattern-density $\rho$, and the removal rate $K_d$ of the trench/down area is zero. When the local step height is reduced to zero, the removal rate everywhere

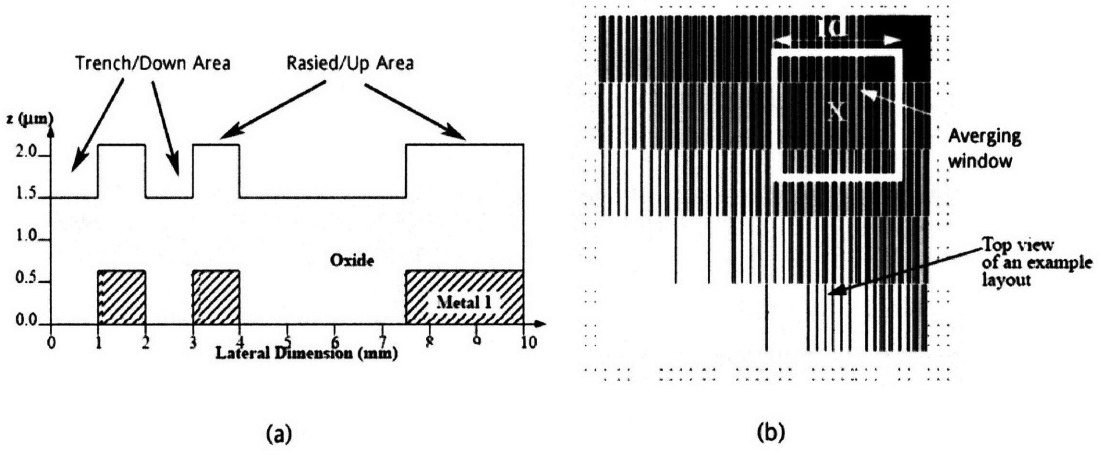



Figure 3-2: (a) Wafer cross-section for an oxide and metal interconnect structure, where the surface is classified as either raised or trench areas. (b) Top-down view of a chip layout illustrating the use of averaging window for the computation of effective pattern-density. [13]

is the blanket removal rate $K_0$. While step heights remain, the rates are given by:

$$\begin{cases} K_u &= \frac{dz_u}{dt} = \frac{K_0}{\rho} \\ K_d &= \frac{dz_d}{dt} = 0 \end{cases} . \tag{3.2}$$

The intuition behind the pattern-density model is the linear dependence on pressure described by the Preston equation. Considering an infinitely rigid pad and wafer surface structures with large step height, the pad only touches the raised regions of the wafer surface, and the average pressure is thus $P/\rho$, where $\rho$ is the pattern-density averaged over the whole die. In reality, the pad is not infinitely rigid and the local pressure is only affected by a finite neighboring area. In Stine's paper, a square averaging area with uniform weight is used, as illustrated in Figure 3-2, and the size of the area is a model parameter, which is called the planarization length $L_P$. Stine's model assumes that polishing occurs only in the raised areas until the local step height is completely removed, and afterwards both raised and trench areas are polished at the same blanket removal rate, so that $K_u$ becomes a function of lateral position $(x, y)$

on the die, and of vertical surface height $z$:

$$K_u(x,y,z) = \begin{cases} K_0/\rho(x,y), & z > z_0 - z_1 \\ K_0, & z \leq z_0 - z_1 \end{cases} \tag{3.3}$$

where $z$ is the oxide thickness, $z_0$ is the initial oxide thickness, and $z_1$ is the initial step height, as illustrated in Figure 3-3. A closed-form solution of oxide thickness can be obtained as in Equation 3.4.

$$z_u(t) = \begin{cases} z_0 - \frac{Kt}{\rho(x,y)}, & t < \frac{\rho(x,y)z_1}{K} \\ z_0 - z_1 - Kt + \rho(x,y)z_1, & t \geq \frac{\rho(x,y)z_1}{K} \end{cases} \tag{3.4}$$

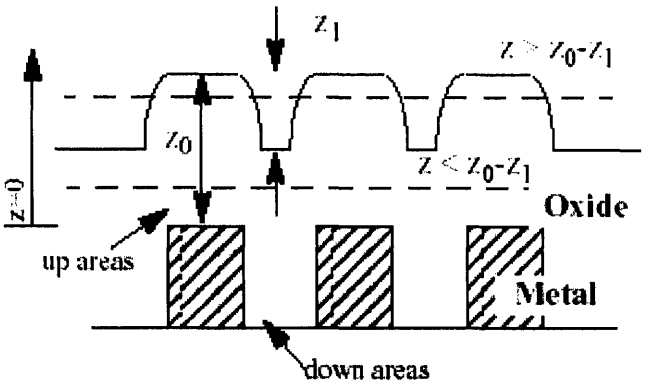



Figure 3-3: Definitions of terms used in model equations: oxide thickness $z$, initial oxide thickness $z_0$, and initial step height $z_1$. [13]

Smith [77] later introduced an extension, the pattern-density step-height (PDSH) model to Stine's pattern-density (PD) CMP model to address the fact that the step-height reduction is not linear for small step heights [78]. The original pattern-density based model assumes no down area polish until complete removal of the local step-height, and then removal of both raised and trench areas at the blanket removal rate. Based on observations by Grillaert et al. [78], Smith's model includes down area polishing once the local step-height is below a critical step-height $h^*$. The down area polishing rate increases in linear proportion to $max(h^* - h, 0)$, while the active area polishing rate decreases in linear proportion to $max(h^* - h, 0)$. The intuition of the model is based on the assumption that finite pad bending causes contacts in the trench

areas, which leads to material polishing in down areas even when the step-height is non-zero. The critical step-height corresponds to the maximum bending depth of the pad into a trench. Thus, the CMP process is separated into two phases depending on how large the step height is: in the first phase (for large step heights), there is only up area (active area) removal, which is exactly the same as in Stine's model; in the second phase (for small step heights), the polish rate of the up area decreases while the down area removal rate increases, and they eventually converge to the blanket removal rate as the step-height reduces to zero. The removal rate dependence on step-height of both models is illustrated in Figure 3-4, based on "removal rate diagrams" proposed by Boning et al. [79].

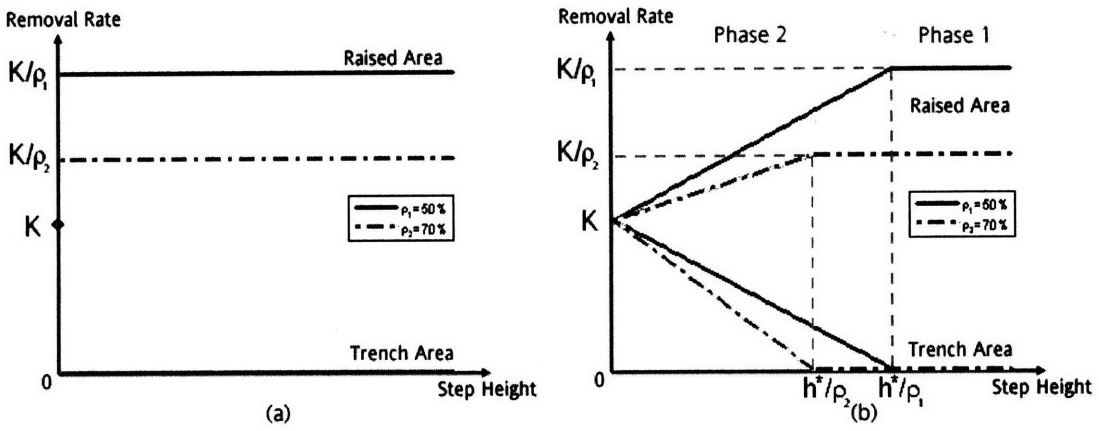



Figure 3-4: Diagrams illustrate the relationship between removal rate and step-height in: (a) pattern-density (PD) CMP model; and (b) pattern-density step-height (PDSH) CMP model.

The pattern-density step-height model results in an exponential reduction of step-height following an initial linear decrease, as summarized in Equation 3.5.

$$\begin{aligned} \Delta z_u &= \begin{cases} \frac{K}{\rho}t, & t \le t_c \\ \frac{K}{\rho}t_c + K(t-t_c) + (1-\rho)\frac{h^*}{\tau}(1-e^{-(t-t_c)/\tau}), & t > t_c \end{cases} \\ \Delta z_d &= \begin{cases} 0, & t \le t_c \\ K(t-t_c) - \rho\frac{h^*}{\tau}(1-e^{-(t-t_c)/\tau}), & t > t_c \end{cases} \end{aligned} \tag{3.5}$$

where $t_c$ is the time needed to reduce step-height to $h^*$, and $\tau = \rho h^*/K$ is a time

constant, which indicates the step-height reduction rate. This exponential decay in step height, following an initial linear step height reduction, is consistent with the data reported by Grillaert et al. [78].

Both models suggest that the majority of the non-uniform post-polish topography comes from the pattern-density variation across the chip. The pattern-density is calculated by averaging the local pattern-density over a square area with sides of length $L_P$, where $L_P$ is referred to as the planarization length. By doing so, all regions in the square carry the same weight in determining the pattern-density at the center of the squre. Ouma [80] [46] shows that it is essential to differentially weight the effects of neighboring features around a given point on a layout when computing the pattern-density at that point. This takes into account the long-range pressure distribution of the polish pad as it deforms around regions on the film surface. When calculating the pattern-densities of discrete unit cells, the weighted average is essentially a convolution of the local layout pattern-density map and an averaging filter. Due to the fact that dies are place periodically on the wafer, the calculation can be computed efficiently by fast Fourier Transformation, as illustrated in Figure 3-5.

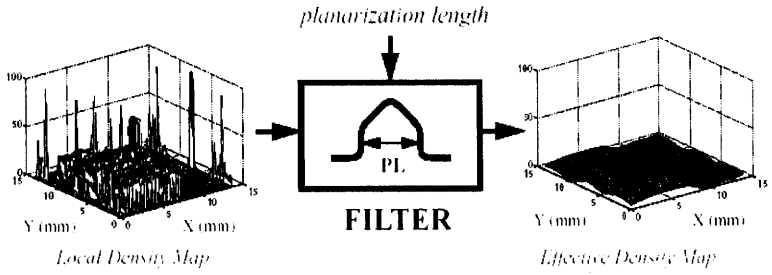



Figure 3-5: Using filtering to compute effective pattern-density map from local pattern-density [1]

Ouma studied filters of different shapes, including square, cylindrical, Gaussian, and elliptic shapes. A cylindrical filter is considered to account for axial symmetry; the Gaussian filter is considered to account for a neighboring interaction which decays with distance; and an elliptic filter is based on the shape of elastic deformation resulting from a small indenter. Ouma finds that the elliptic shape filter results in

smallest model error and has a clear correspondence to the polish pad deformation [80]. Smith's and Ouma's models are extensions of Stine's pattern-density based model, and integrating them together gives an effective pattern-density step-height (PDSH) CMP model that has been the basis for practical chip-scale modeling of oxide CMP.

## 3.2 Objectives and Framework of Die-Level CMP Models

In this section, we revisit the objectives and goals of die-level CMP models. In order to extend previous die-level models to account for important process and consumable dependencies, we then describe a generalization of these die-level models in which pressure and slurry contributions can be explicitly tracked. Section 3.3 then extends the die-level model further to explicitly account for details of the pad structure.

### 3.2.1 Objective of Die-Level CMP Models

The first objective of die-level models is to simulate the planarization of layout structures on a die. As discussed in the previous section, the wafer surface can be classified as either raised or trench area, and the surface profile can be described by the height $z_u(x, y)$ of the raised area, and by the local step-height $h(x, y)$. The estimation of $z_u(t; x, y)$ and $h(t; x, y)$ requires the initial conditions and dynamic equation for the removal rates $K_u = dz_u/dt$ and $K_d = dz_d/dt$. The initial conditions, $z_0(x, y)$ and $h_0(x, y)$, depend on the layout structure and process design, and are usually known. For a given CMP process, the removal rates depend on pattern-density and local surface topography, and can be represented as $K_u(\rho, z_u, h)$ and $K_d(\rho, z_u, h)$. The die-level CMP model depends on the dynamic equations for the removal rates, which must be found by either an empirical or a physically-based approach. Once $K_u(\rho, z_u, h)$ and $K_d(\rho, z_u, h)$ are known, the surface profile can be calculated by solving the dynamic

equations either analytically or numerically:

$$\begin{aligned} \frac{dz_u(t;x,y)}{dt} &= -K_u(\rho, z_u, h) \\ \frac{dh(t;x,y)}{dt} &= -K_u(\rho, z_u, h) + K_d(\rho, z_u, h). \end{aligned} \tag{3.6}$$

The second objective is to understand how the planarization performance can be affected or controlled by the properties of consumables (pad and slurry) and process conditions (applied pressure and velocity). The resulting physically-based model can then be used to improve consumable design, process control, and layout design.

### 3.2.2 Modeling Framework

The previous pattern-density step-height (PDSH) models are derived from the Preston equation, and the underlying assumption is the linear dependence of the removal rate $K = \frac{dz}{dt}$ on pressure $P$, where $z$ is the film thickness. Lee [1] presents a model framework in which the removal rate is a function of pressure, and in which the pressure is a function of pattern density $\rho$ and step-height $h$, i.e., $K(\rho, h) = K(P(\rho, h))$. This framework is adopted here and generalized to $K(\rho, z_u, h) = K(P(\rho, z_u, h))$.

The planarization achieved CMP results from the higher removal rate of raised areas compared to that of the trench areas. The physics of CMP, discussed in the previous chapter, implies that the removal rate depends on applied pressure, relative velocity, slurry concentration, and potentially other parameters.

First, the pad compresses more when pressed against the raised area than in the trench area, and the resulting higher pressure implies a higher removal rate on raised areas than on trench areas. This makes pressure an important, and indeed the dominant, factor in planarization.

Second, the relative velocity does not have any clear topography impact. A second-order effect may exist, in which the relative velocity affects the dynamic elastic response of the polishing pad, and Lee [1] observes that the planarization performance is affected by relative speed. In our work, we assume the velocity is fixed for a given process, so that second order effects are lumped into the overall estimate of effec-

tive pad and process parameters such as effective Young's modulus or planarization length.

Third, in polishing a single material, the chemical contribution may be tuned to have topography dependence, if the removal rate is sensitive to slurry concentration and the concentration has a topography dependence. The transportation of slurry may cause the slurry concentration to have topography dependence; however, the low aspect ratio of most features in CMP makes the difference negligible in most cases. Secondly, the slurry concentration is usually not uniformly distributed across the wafer, and thus a highly chemically sensitive slurry may cause undesirable wafer-level non-uniformity. Therefore, most slurries are designed to avoid a chemical concentration-driven topography dependence for the polishing of a single material type, and it is a reasonable assumption that the removal rate is primarily pressure dependent, or $K_{u,d} = K(P_{u,d})$. In contrast, the slurry chemistry generally has a strong impact on the pressure dependence of the removal rate, and in polishing of dual materials, the slurry chemistry strongly influences the relative removal rates for different materials (selectivity).

In the model framework used here, the problem is decomposed int two parts: the slurry-dominated dependence of removal rate on pressure $K(P)$ and the pad-dominated dependence of pressure on wafer surface topography $P(\rho, z_u, h)$. The dependence of removal rate is a focus of the particle-level CMP models; in die-level models the form of $K(P)$ is either determined empirically or assumed. Thus, most die-level CMP models are developed by estimating $P_{u,d}(\rho, z_u, h)$ by either empirical or physically-based approaches, integrated with the dependences of the removal rate on pressure $K(P)$.

### 3.2.3 Dependence on Slurry: $K_{u,d}(P)$

While the removal rate of a conventional CMP slurry is linearly proportional to pressure as suggested in the Preston equation, the removal rate of a non-conventional slurry, such as a ceria based slurry, exhibits a nonlinear dependence on pressure. The nonlinear dependence can be determined based on experimental measurements. For

example, Figure 3-6 shows the relationship between down force and polishing rate of a ceria-based slurry [14]. Hence, the relationship consists of two linear segments, where the slopes and transition pressure $P_t$ depends on both slurry abrasive particles and additive chemistry properties. The relationship can be described by Equation 3.7,

$$K(P) = \begin{cases} \beta_1 P, & when \quad P \leq P_t \\ \beta_1 P_t + \beta_2 (P - P_t), & when \quad P > P_t \end{cases} \tag{3.7}$$

where $\beta_1$ and $\beta_2$ are model parameters.

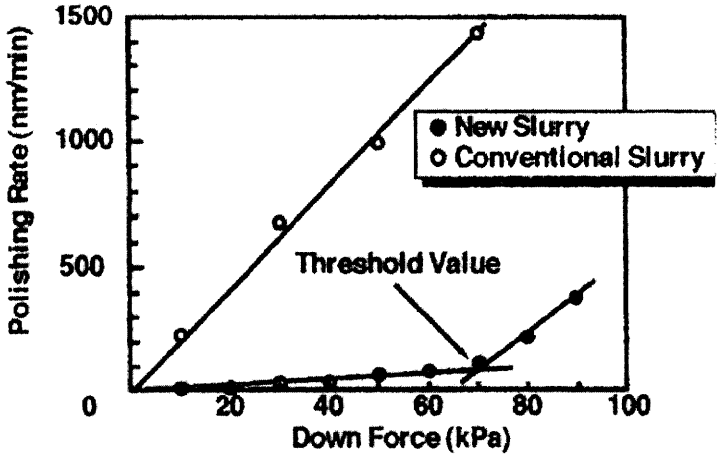



Figure 3-6: Relationship between down force and polishing rate [14].

In dual-material polishing, the material in the trench area may be different than in the raised area during some portion of the polishing step. The removal rate ratio of trench material to the non-trench area is referred to as the removal rate selectivity $s$, which is generally designed to be independent of pressure. Thus, when the blanket removal rate of the trench material has the pressure-dependence $K(P)$, the removal rates for the up and down areas are assumed:

$$\begin{cases} K_d(P_d) &= K(P_d) \\ K_u(P_u) &= \frac{1}{s} K(P_u). \end{cases} \tag{3.8}$$

## 3.3 Die-Level CMP Models

In this section, a physically-based die-level model is proposed which explicitly models pad and surface asperities, with model parameters that are based on the physical properties of the pad rather than being purely fitting parameters. The physically-based model can be used to simulate the surface evolution, as well as to study how planarization is affected by pad properties and process condition. However, the physically-based model is computationally intensive, and many practical applications require fast model simulation.

A semi-empirical die-level CMP model, motivated by the new physically-based die-level model, is developed that improves upon previous pattern-density step-height models by making realistic assumptions and approximations, and improving the ease of computation. The parameters of the PDSH model are calibrated for a specific process by polishing test wafers. One drawback of the PDSH model is that its application is limited to the case where the wafer surface is flat and the initial step-height is the same across the die. The physically-based CMP model is able to overcome this limitation.

### 3.3.1 Physically-based CMP Model

In this section, we propose a physically-based CMP model to describe the dependence of pressures on local pattern-density and step height, $P_{u,d}(\rho, z_u, h)$. It is worth noting that local pattern-density rather than effective pattern-density is used here and the value of $\rho$ only depends on discretization, which avoids the concept of effective pattern-density. The relationship to step height is obtained by analysis and modeling of the polishing pad. The polishing pad consists of the bulk material and the pad surface asperities, as illustrated in Figure 3-7. The bulk material can be treated as an elastic body. The surface asperities come in contact with the wafer surface, and the compression of the asperities depends on both the wafer surface profile and pad bulk bending.

The notations used in this discussion are illustrated in Figure 3-8. The wafer is

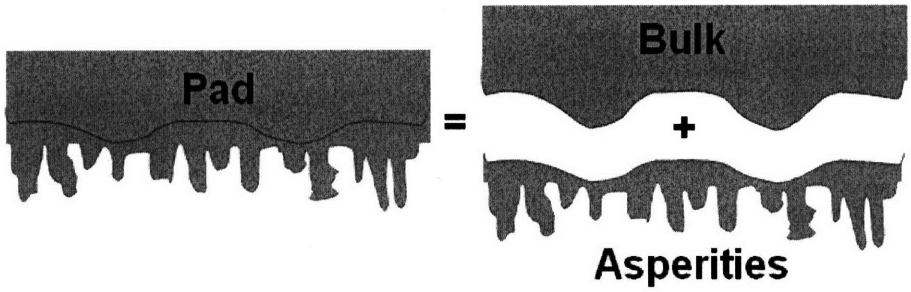



Figure 3-7: Polishing pad can be decomposed into bulk material and surface asperities.

assumed to sit face up in the direction of z-axis, and the polishing pad is pressed down onto the wafer surface. Here, $w(x, y)$ is used to describe the $z$-coordinate of the nominal separation point between the bulk and asperities of the pad, and $z(x, y)$ is used to describe the wafer surface. The distance between the wafer surface and nominal bulk pad position is $w(x, y) - z(x, y)$.

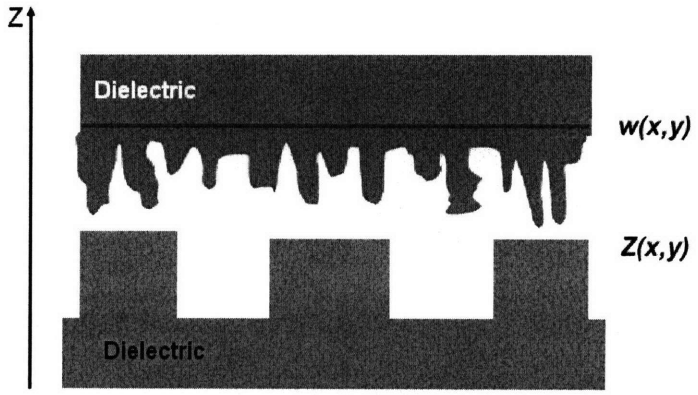



Figure 3-8: Polishing pad can be decomposed to bulk material and surface asperities.

### Modeling of Bulk Materials

The bulk material is assumed to be elastic, and can be modeled using a contact wear model, as described in Appendix B. In summary, the bulk surface displacement $w(x, y)$ and the pressure $P(x, y)$ satisfythe following relationship:

$$w(x, y) - w_0 = F(x, y) \otimes P(x, y) \tag{3.9}$$

where $F(x,y)$ is the deformation response to a point pressure. This deformation is proportional to $1/E$, where $E = \frac{E_p}{1-\nu_p^2}$ is the effective Young's modulus of the pad. The variable $w_0$ is a normalization constant, which depends on the applied pressure $P_0$.

### Modeling of Asperities

In modeling asperities, the asperities are assumed to have negligible width and a height distribution $l(\xi)$, where the height $\xi$ is measured from the nominal pad bulk surface $w(x,y)$. At the scale of the asperities, the pad bulk displacement near a set of asperities appears to be nearly constant, as pictured in Figure reffig:zcoord, and thus the displacement $w(x,y)$ can be referred as the "nominal plane" for the pad bulk interface. At location $(x,y)$, the distance between the wafer surface $z(x,y)$ and the nominal plane $w(x,y)$ is $w(x,y) - z(x,y)$, so the asperities of height $\xi$ larger than $w(x,y) - z(x,y)$ will contact the wafer surface and the amount of compression of the asperities is $\xi - [w(x,y) - z(x,y)]$. All asperities are assumed to obey Hook's law, i.e., the exerting force is proportional to the compressed amount. The expected value of $P(x,y)$ can be estimated by averaging across all of the asperities as follows:

$$\begin{aligned} P(x,y) &= \int_{w(x,y)-z(x,y)}^{\infty} k \cdot (\xi - [w(x,y) - z(x,y)]) \cdot l(\xi) d\xi \\ &= k \cdot \Psi(w - z) \end{aligned} \tag{3.10}$$

where $k$ is a spring constant, and $\Psi(x)$ is a derived asperity height distribution function, defined as $\Psi(x) = \int_x^{\infty} (\xi - x) l(\xi) d\xi$. $\Psi(z)$ can be calculated once the probability distribution of asperity height is known, and it is a strictly decreasing function and approach zero at infinity. For example, if the asperity height distribution is a delta function,

$$\begin{cases} l(\xi) &= \delta(\xi_c - \xi) \\ \Psi(\xi) &= (\xi_c - \xi) H(\xi_c - \xi); \end{cases} \tag{3.11}$$

where $H(.)$ is the unit step function. If the asperity height has an exponential distribution with parameter $\lambda$, then

$$\begin{cases} l(\xi) &= \frac{1}{\lambda} e^{-\xi/\lambda} \\ \Psi(\xi) &= \lambda \cdot e^{-\xi/\lambda} \end{cases} \tag{3.12}$$

When a feature of step-height $h(x, y)$ is pressed against the pad, Equation 3.10 implies that the up area pressure $P_u = k \cdot \Psi(w - z_u)$ and the down area pressure $P_d = k \cdot \Psi(w - z_d) = k \cdot \Psi(w - z_u + h)$. The total pressure is the sum of the two pressures weighted by their pattern-densities:

$$\begin{aligned} P &= \rho \cdot P_u + (1 - \rho) \cdot P_d \\ &= \rho \cdot k \cdot \Psi(w - z_u) + (1 - \rho) \cdot k \cdot \Psi(w - z_u + h) \end{aligned} \tag{3.13}$$

where $\rho$ is the pattern-density or area fraction of raised area.

If the asperity height distribution is a delta function, we can plug Equation 3.11 into Equation 3.13 and re-arrange the equations to obtain the following relationships:

$$\begin{cases} P_u &= \begin{cases} 0 & h > \xi_c \\ \frac{1-h/h^*}{1-(1-\rho)h/h^*} P & h \le \xi_c \end{cases} \\ P_d &= \begin{cases} P/\rho & h > \xi_c \\ \frac{1}{1-(1-\rho)h/h^*} P & h \le \xi_c \end{cases} \end{cases} \tag{3.14}$$

where $h^* = \xi_c - (w - z_u)$. This relationship is very similar to that of the previous PDSH CMP model illustrated in Figure 3-4 (b).

If the asperity height has an exponential probability density distribution, the resulting pressure relationships as a function of feature step heights are

$$\begin{cases} P &= k(\rho + (1 - \rho)e^{-h/\lambda}) \cdot \lambda e^{-(w-z_u)/\lambda} \\ P_d &= \frac{1}{1+\rho(e^{h/\lambda}-1)} P \\ P_u &= \frac{e^{h/\lambda}}{1+\rho(e^{h/\lambda}-1)} P \end{cases} \tag{3.15}$$

### The Physically-Based Model

The physically-based CMP model can be obtained by integrating the two parts together: the elastic bulk part of the pad, which is described by Equation 3.9, and the

asperities with exponential height distribution, which is described by Equation 3.15.

$$\begin{cases} P(x,y) &= k \cdot (\rho(x,y) + (1-\rho(x,y))e^{-h(x,y)/\lambda}) \cdot \lambda e^{-(w(x,y)-z_u(x,y))/\lambda} \\ w(x,y) &= F(x,y) \otimes P(x,y) + w_0 \end{cases} \tag{3.16}$$

In the above equations, $\rho(x,y)$ can be extracted from the layout design, and $z_u(x,y)$ and $h(x,y)$ are dynamically updated during the simulation of CMP. The two unknowns $P(x,y)$ and $w(x,y)$ can be calculated by solving the two equations iteratively. Once $P(x,y)$ is solved, $P_u(x,y)$ and $P_d(x,y)$ can be calculated using Equation 3.15.

In this model, there are three key model parameters: the effective Young's modulus of the pad $E$, the applied pressure $P_0$, and the asperity height distribution parameter $\lambda$ which we will refer to as the characteristic asperity height.

It is worth noting that the computed pressure distribution is proportional to $P_0$ as long as the ratio of $E$ to $P_0$ is fixed.

### 3.3.2 Exponential Pattern-Density Step-Height CMP Model

The previous PDSH model assumes two separate contact regimes, depending on whether the pad contacts the trench regions or not. As we discussed in the previous chapter, the pad surface is covered by asperities. The contact mechanism is not the bending of a continious elastic pad surface, but rather depends on the number of asperities in contact and how much each asperity is compressed. Thus, the transition between down area contact regimes is not a sharp transition as shown in Figure 3-4. In this section, an exponential PDSH model is proposed which assumes a smoother transition between the two regimes. Specifically, the model assumes an exponential dependence of trench or down area pressure on step-height as given by the following equation:

$$P_d = P_0 \cdot e^{-h/h_1}, \tag{3.17}$$

where $P_d$ is the pressure in the trench/down region, $P_0$ is the applied pressure, $h$ is the local step-height, and $h_1$ decribes he decay rate in the down area pressure. The variable $h_1$ describes how fast $P_d$ decreases with respect to step-height, and equals the value of step-height when $P_d = P_0/e$. The exponential dependence is assumed

because it smoothly decays to zero as step height approaches infinity, because the decay rate can be specified by a single parameter, and because a closed-form solution to the step height and thickness as a function of time can be obtained which speeds up computation.

Intuitively, $h_1$ depends on the pattern-density, as well as other layout parameters. Asperities more easily touch the trench of a low pattern-density region, thus $h_1$ will be larger. Empirically, we find that defining a characteristic step-height $h^* = h_1/\rho$ yields good agreement with experimental data. Equation 3.17 can then be rewritten

$$P_d = P_0 \cdot e^{-\rho h/h^*}, \tag{3.18}$$

where $h^*$ is independent of pattern-density $\rho$.

Next, we assume that the summation of both raised/up region and trench/down region pressures weighted by pattern-density equals the total applied pressure, i.e.,

$$P_0 = \rho P_u + (1-\rho)P_d, \tag{3.19}$$

where $\rho$ is the effective pattern-density. Although local pattern-density is a more intuitive choice than the effective pattern-density, the relationship is chosen to satisfy two extreme conditions: when $h = 0$, $P_u = P_d = P_0$ and the above equation holds; and when $h$ is very large, $P_d = 0$ and $P_u = P_0/\rho$. The pressure dependence on step-height is illustrated in Figure 3-9 (a).

$$P_u = \frac{1}{\rho}P_0 - \frac{1-\rho}{\rho}P_0 \cdot e^{-\rho h/h^*}. \tag{3.20}$$

In this model there are two empirical parameters: the planarization length $L_P$, which is used to compute the effective pattern-density, and the characteristic step-height $h^*$.

### 3.3.3 Physically Based PDSH model

The exponential PDSH model described in Section 3.3.2 is based on *assumed* exponential dependence on step height. Using the physically-based model of Section 3.3.1, a

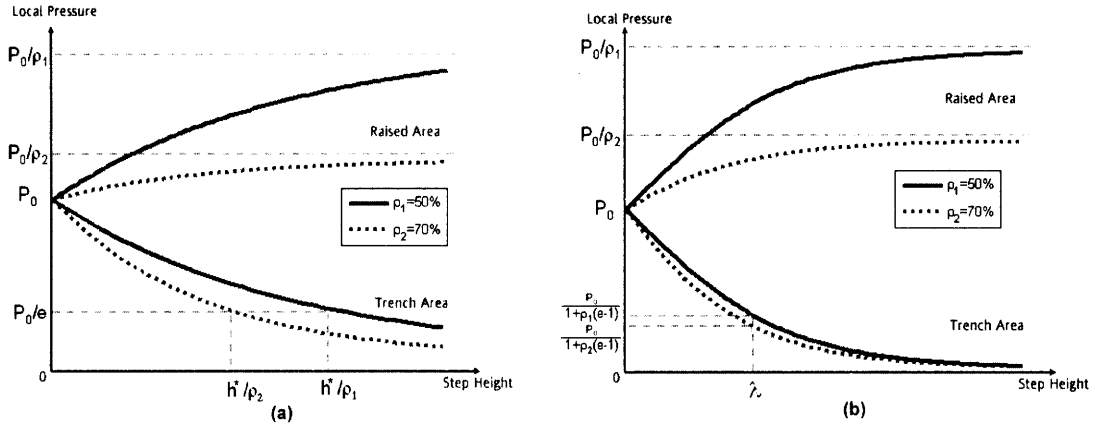



Figure 3-9: Relationship between local pressure and step-height (a) in the exponential PDSH model, and (b) in the physically based PDSH model.

different step-height dependence based on an exponential asperity height distribution can be generated. The resulting physically based PDSH model, using the dependence of pressure on step-height as in Equation 3.15, gives

$$\begin{cases} P_d &= \frac{1}{1+\rho(e^{h/\lambda}-1)} P_0 \\ P_u &= \frac{e^{h/\lambda}}{1+\rho(e^{h/\lambda}-1)} P_0 \end{cases} . \tag{3.21}$$

This model has two model parameters, $L_P$ and $\lambda$. Thus the empirical value $h^*$ is replaced with a physical pad parameter, $\lambda$. For comparison, we can plot the pressure versus step-height relationships for those two new PDSH models. Figure 3-9 (a) shows the exponential step height dependence, including the effect of the pattern-density on the the pressure decay with step height. Figure 3-9 (b) shows the pressure dependence on step height for Equation 3.21. Qualitatively, these two are similar, and both have a smooth dependence on step height, in contrast to Figure 3-4 which had a sharp break at some $h_c$.

## 3.4 Applying PDSH Die-Level CMP Models

In this section, the PDSH models are applied to simulate the polishing of either single material or dual material structures using either conventional or non-conventional slurry. The examples have illustrated a number of points and provide useful infor-

mation. In most cases the problem of surface evolution is solved analytically; the closed-form solutions of the PDSH models enable an easy implementation and fast computation. The surface evolution is illustrated by plotting the amount removed in the raised area $\Delta z_u$ and in the trench area $\Delta z_d$ as a function of time for different pattern-densities. Because the planarization of surface structures is of interest, the local step-height reduction is plotted, as well as the long range thickness variation $Range(z_u)$, which is defined as the difference between $z_u$ of the 90% pattern-density area and that of the 10% area. In dual material polishing, the steady-state removal rate and steady-state step-height are discussed. The planarization performance of a non-conventional slurry is compared with that of a conventional slurry. Finally, we discuss what $K(P)$ can result in optimal planarization performance, potentially providing some guidance to the design of future slurries with improved planarization capability.

### 3.4.1 Single Material Polishing with Conventional Slurry

The removal rate of conventional slurries have a linear dependence on pressure, and thus the removal rates in the exponential PDSH model can be written as

$$\begin{cases} K_d &= K_0 \cdot e^{-\rho h/h^*} \\ K_u &= \frac{1}{\rho} K_0 - \frac{1-\rho}{\rho} K_0 e^{-\rho h/h^*} \end{cases} \tag{3.22}$$

where $K_d$ is the removal rate in trench/down area, $K_u$ is the removal rate in raised/up area, and $K_0$ is the blanket removal rate. The relationship between removal rate and step-height is also illustrated in Figure 3-10(a).

Solving these equations gives a closed-form solution for the step-height $h$, amount removed in the raised/up region $\Delta z_u$, and the amount removed in the trench/down region $\Delta z_d$:

$$\begin{cases} \Delta z_d &= h^* \ln\left[1 + e^{-\rho h_0/h^*}\left(e^{K_0 t/h^*} - 1\right)\right] \\ \Delta z_u &= \frac{1}{\rho} K_0 \cdot t - \frac{1-\rho}{\rho} \Delta z_d \\ h &= h_0 + \Delta z_d - \Delta z_u \end{cases} \tag{3.23}$$

The surface evolution can be illustrated by plotting the amount removed in the

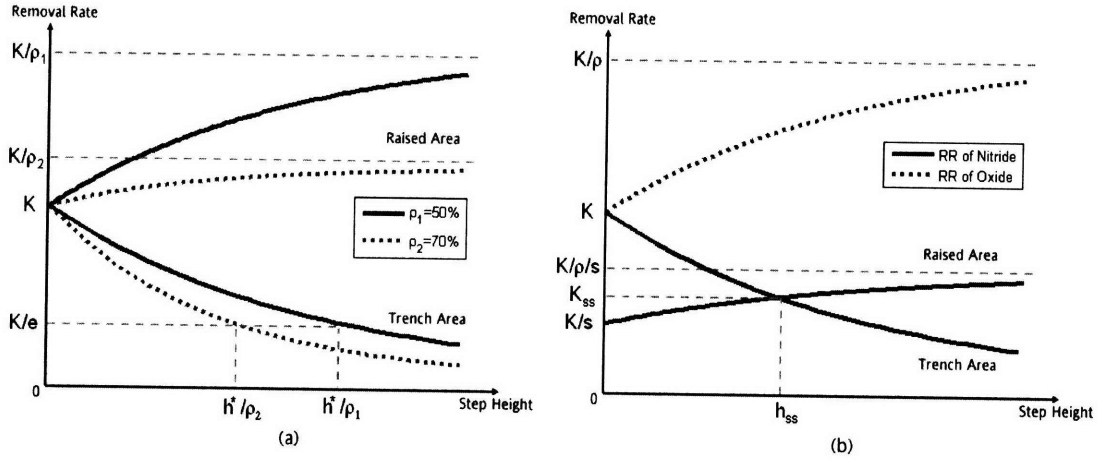



Figure 3-10: Diagrams illustrates the relationship between removal rate and step-height in the exponential PDSH model in polishing (a) oxide and (b) nitride.

raised area $\Delta z_u$ and that in the trench area $\Delta z_d$ as a function of time for different pattern densities, as in Figure 3-11. The parameters used are a blanket removal rate of 3000 Å/$min$, a characteristic step-height $h^* = 1000$ Å, and an initial step-height $h_0 = 6247$ Å. We also plot step-height and $Range(z_u)$ as a function of time in Figure 3-12, where the $Range(z_u)$ is the difference between $z_u(\rho = 10\%)$ and $z_u(\rho = 90\%)$. As $z_u$ captures the envelope of the wafer surface, $Range(z_u)$ describes the long-range topography variation across an entire chip having a large pattern density variation.

The result is intuitive: a low pattern-density region is polished faster than a high pattern-density region; a raised area is polished faster than a trench area; and step-height is reduced gradually. Two points are worth notice:

- Initially low pattern-density regions polish significantly faster than high pattern-density regions. When time is large ($t > 100s$), however, most regions polish at nearly the same rate as the blanket removal rate. This can also be observed by the plot of $Range(z_u)$, which increases initially, and remains nearly constant when for large times.

- Initially a raised region is polished significantly faster than a trench region, which effectively reduces the local step-height. When time is large, however, the step-height is small, and both regions are polished at nearly the same rate.

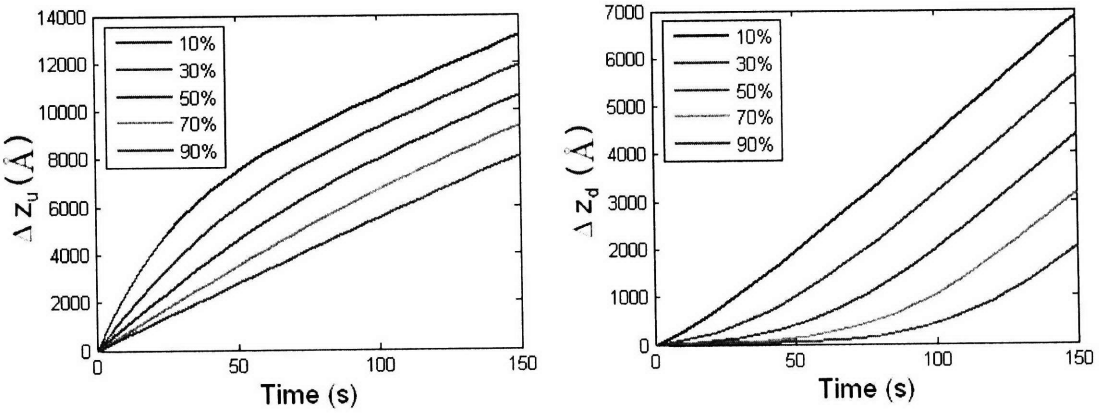



Figure 3-11: Amount of removal for different pattern densities as a function of time. (Left) $\Delta z_u$. (Right) $\Delta z_d$.

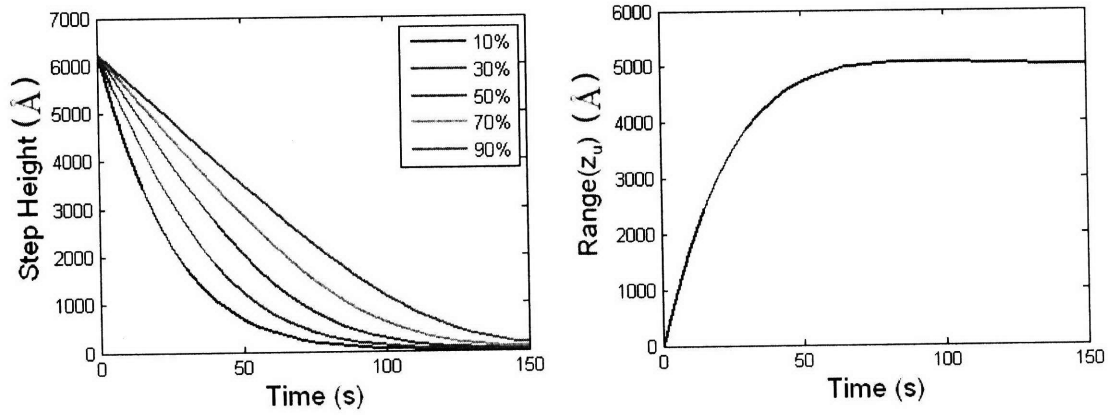



Figure 3-12: (Left) step-height of different densities as a function of time. (Right) $Range(z_u)$, which is defined as $z_u(\rho = 90\%) - z_u(\rho = 10\%)$, as a function of time.

The model has three parameters: blanket removal rate $K_0$, planarization length $L_P$, and characteristic step-height $h^*$. The effect of each of these parameters is considered below.

The blanket removal rate $K_0$ describes the overall polishing rate, and its effect on the model output is obvious, simply scaling the rate of surface evolution. A larger blanket removal rate is desirable, as it takes less time to finish a process and increases the throughput of the polishing tool.

The planarization length $P_L$ describes the spatial range over which neighboring structures can affect polishing of a given region. It defines the width of the spatial averaging filter used to calculate the effective pattern density. A larger value results in smoother distribution of effective pattern-density across the chip, and consequently less with-in-die variation, which is desirable.

The characteristic step height $h^*$ affects the partition of amount removed between raised area and trench area. We choose a different value of $h^* = 200$ Å, and compare the results with the case when $h^* = 1000$ Å, in Figure 3-13 and 3-14. Not surprisingly, a smaller $h^*$ results in faster polishing in raised areas and slower polishing in trench areas, which results in a faster reduction of local step-height. Thus, a smaller value of $h^*$ is desired, as it requires less time and deposited oxide to achieve a targeted local step-height. However, $h^*$ has little effect on topography variation over longer spatial distances, which is controlled by $L_P$.

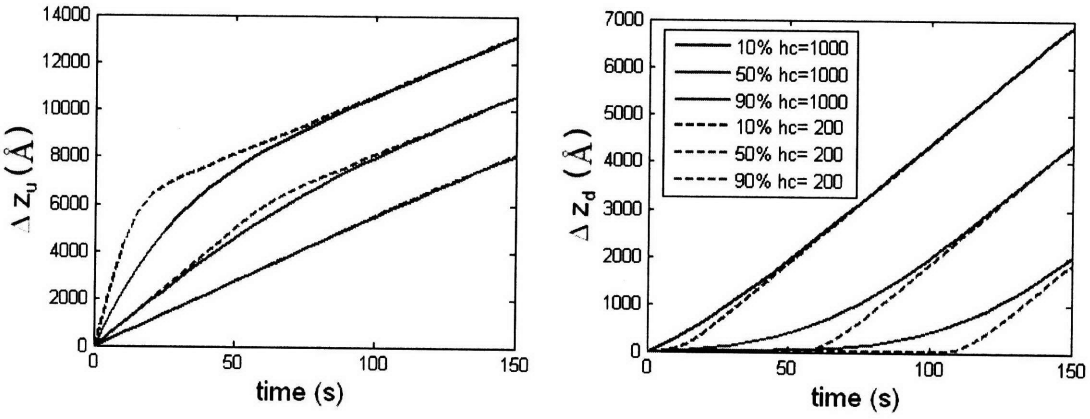



Figure 3-13: Comparison of model predictions using different values of $h^*$: $h^* = 1000$ Å (solid line) and $h^* = 200$ Å (dashed line).

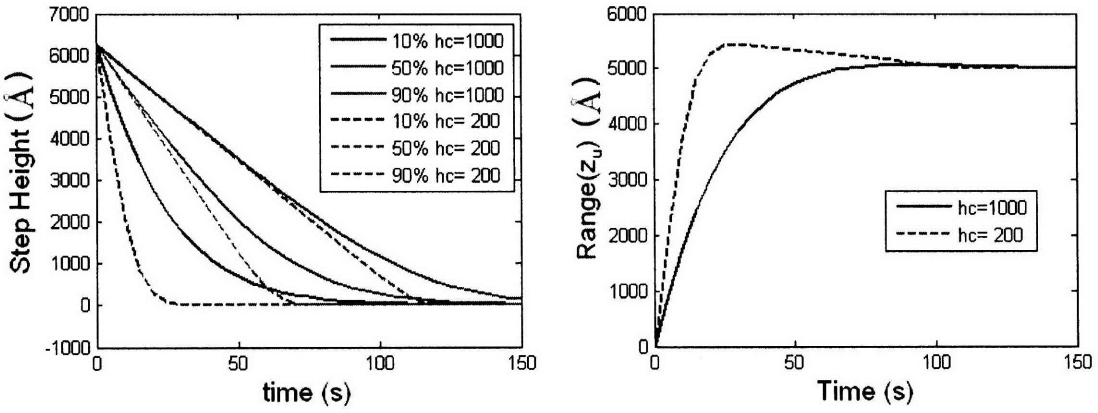



Figure 3-14: Comparison of model predictions using different value of $h^*$: $h^* = 1000$ Å (solid line) and $h^* = 200$ Å (dashed line)

If the physically-based PDSH model is used, the removal rates can be written as

$$\begin{cases} K_d &= \frac{1}{1+\rho(e^{h/\lambda}-1)}K \\ K_u &= \frac{e^{h/\lambda}}{1+\rho(e^{h/\lambda}-1)}K \end{cases} \tag{3.24}$$

These removal rate dependencies on step height are illustrated in Figure 3-15 (a).

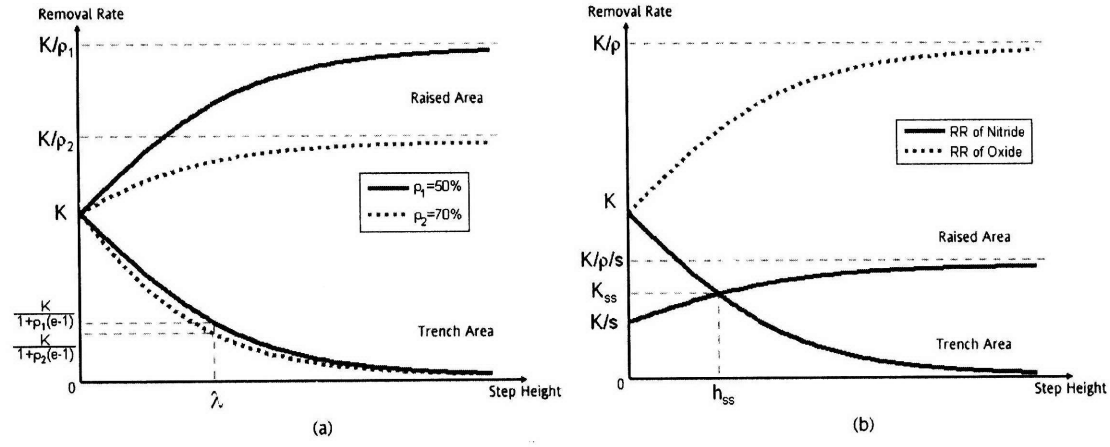



Figure 3-15: Relationship between removal rate and step-height in the physically-based PDSH CMP model in polishing: (a) single material polishing, and (b) dual material polishing.

Although no closed-form solution can be obtained, the step-height $h(t)$ is found to satisfy the following equation, which can be solved numerically using Newton's

method.

$$e^{\rho h(t)/\lambda}\left(1-e^{-h(t)/\lambda}\right)=e^{(\rho h_0-Kt)/\lambda}\left(1-e^{-h_0/\lambda}\right) \tag{3.25}$$

Then, the calculated $h(t)$ can be used to obtain a solution for $\Delta z_u$ and $\Delta z_d$.

$$\begin{cases} \Delta z_u &= Kt+(1-\rho)(h_0-h(t)) \\ \Delta z_d &= Kt-\rho(h_0-h(t)) \end{cases} \tag{3.26}$$

### 3.4.2 Dual Material Polishing with Conventional Slurry

The STI process flow, as reviewed in Section 1.1.2, can be separated into two stages. In the first stage, only oxide is polished. In the second stage when the overburden oxide is cleared, nitride in the raised area and oxide in the trench area are polished simultaneously. The first stage is single material polishing, which has been solved in Equation 3.23. Let the clearing time $t_c$ denote the time of transition from oxide polishing to dual material polishing, i.e., when the overburden oxide of thickness $z_0$ is just cleared. We can calculate $t_c$ by solving $\Delta z_u(t_c)=z_0$. Although $t_c$ does not have an analytic solution, it can be calculated using Newton's method. The solution also gives us the step-height $h(t_c)$ at time $t_c$, which is the initial condition for the second stage.

The second stage of STI CMP process involves dual material polishing, and we define the material selectivity $s$ as the ratio of blanket oxide removal rate to blanket nitride removal rate. The removal rate of the trench area, which is filled with oxide, is not changed. However, the removal rate of the raised area, which is now a nitride material, needs to be scaled by $1/s$. The set of equations need to be modified accordingly:

$$\begin{cases} K_d &= K_0\cdot e^{-\rho h/h^*} \\ K_u &= \frac{1}{s\rho}K_0-\frac{1-\rho}{s\rho}K_0e^{-\rho h/h^*} \end{cases} \tag{3.27}$$

The relationship of removal rates to step-height given by Equation 3.27 is illustrated in Figure 3-10(b).

One thing worth noting is that the steady state, when both the raised areas and the trench areas are polished at the same removal rate, in the polishing of dual material

structures is different from that of single material polishing. Rather than achieving local planarity and then polishing everywhere at the blanket removal rate, a steady state step height $h_{ss}$ results which is positive, and the steady state removal rate $K_{ss}$ is less than the blanket oxide removal rate. Intuitively, a positive step-height results in a higher pressure in a raised nitride area which balances the slower blanket removal rate of nitride (scaled by $1/s$). From the removal rate given by Equation 3.27, we can solve for the steady state step-height $h_{ss}$ and removal rate $K_{ss}$, resulting in

$$\begin{cases} h_{ss} &= \frac{h^*}{\rho} \ln\left[1 + (s-1)\rho\right] \\ K_{ss} &= \frac{K}{1+(s-1)\rho} \end{cases} \tag{3.28}$$

The result is plotted in Figure 3-16 for a range of selectivity and pattern-density values. A higher selectivity results in a slower steady state polishing rate as seen in the right figure, but it also causes a higher steady state step-height, as indicated in the left figure. The steady state removal rate is largest when $s = 0$ and at that point is uniform across different pattern-density areas. When the selectivity is large, the steady state removal rate is small and the difference across different pattern-density areas is also small. The non-uniformity across different pattern density regions of the steady state removal rate is largest around $s = 5$. Thus, the effect of having a slower polishing material present is complicated, and a higher selectivity is not always desirable.

For the dual material polishing stage, i.e., when $t > t_c$, integrating Equation 3.27 gives the following results.

$$\begin{cases} \Delta z_d(t) &= \Delta z_d(t_c) + h^* \frac{s}{1+(s-1)\rho} \ln\left[1 + (1+(s-1)\rho) e^{-\rho h(t_c)/h^*} \left(e^{\frac{K}{s \cdot h^*} \cdot (t-t_c)} - 1\right)\right] \\ \Delta z_u(t) &= \Delta z_u(t_c) + \frac{1}{s\rho} K \cdot (t - t_c) - \frac{1-\rho}{s\rho} (\Delta z_d(t) - \Delta z_d(t_c)) \\ h(t) &= h_0 + \Delta z_d(t) - \Delta z_u(t) \end{cases} \tag{3.29}$$

Figure 3-17 shows the model prediction for $\Delta z_u$ and $\Delta z_d$ during STI CMP. The plots are similar to that of oxide CMP except two key aspects. First, the removal rate of the raised area decreases significantly after the oxide is cleared, i.e., for $\Delta z_u > z_0$ as in the left plot, and this is due to the slower polishing rate of nitride. Second, the

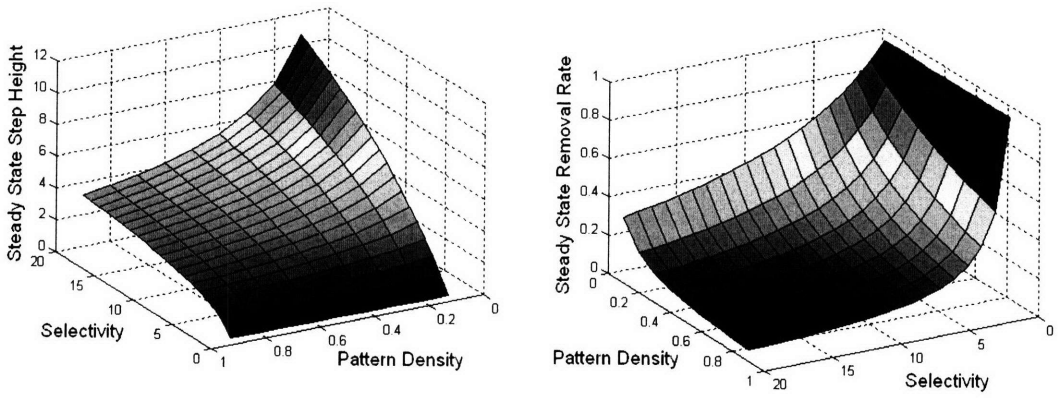



Figure 3-16: (Left): Steady state step-height (in units of $h^*$) vs. selectivity and pattern-density. (Right): Steady state removal rate (in units of oxide blanket removal rate $K$) vs. selectivity and pattern-density.

removal rate is no longer independent of pattern-density for large times, because the steady state removal rate depends on the pattern-density. The plots of step-height and $Range(z_u)$ are shown in Figure 3-18. The local step-height initially decreases, and once we start to polish nitride, the step-height rises and converges to the steady state step-height. The plot of $Range(z_u)$ is quite different from that for single material polishing. After $Range(z_u)$ increases in the single material polishing stage, the polishing of the areas which first advance to the nitride polish stage slows down significantly, which causes a decrease in $Range(z_u)$. When nitride is finally exposed everywhere, areas of different pattern densities are polished at different rates and $Range(z_u)$ begins to increase again. This behavior in $Range(z_u)$ has distinct and substantial changes at different points in time; in Section 4.4 we will take advantage of this effect to help explain observed changes in friction measurements taken during STI CMP and highlight the resulting opportunity for endpoint detection.

Modeling the dual material polishing case requires three parameters: the planarization length $L_P$, the selectivity $s$, and the characteristic step height $h^*$. Thus, modeling of the STI CMP process requires a total of six parameters: three for oxide polishing, and three for dual material polishing (or two if the same $h^*$ is used in both stages). Two different planarization lengths are used, as $L_P$ takes on very different values in the two polishing stages. Here is an intuitive explanation for different $L_P$

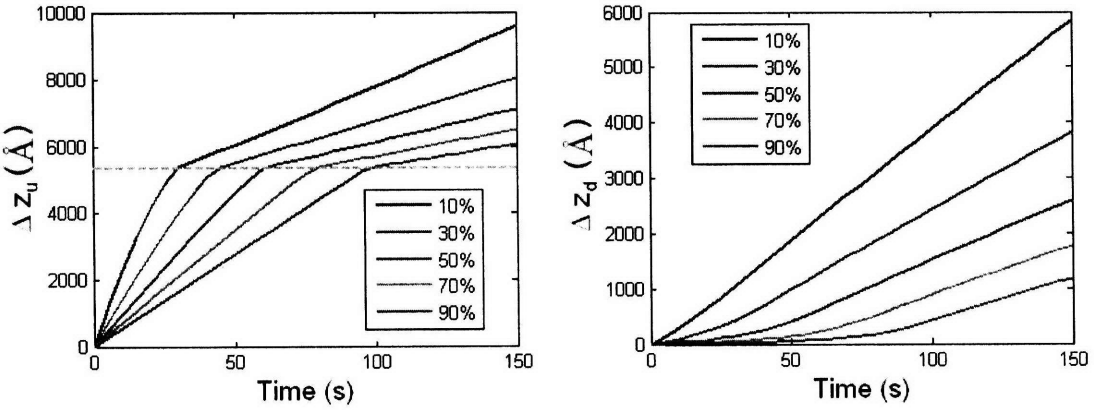



Figure 3-17: Amount removed during STI CMP for different pattern densities as a function of time: (left) $\Delta z_u$ and (right) $\Delta z_d$. The dotted line in the left figure shows $z_0$, the initial oxide thickness.

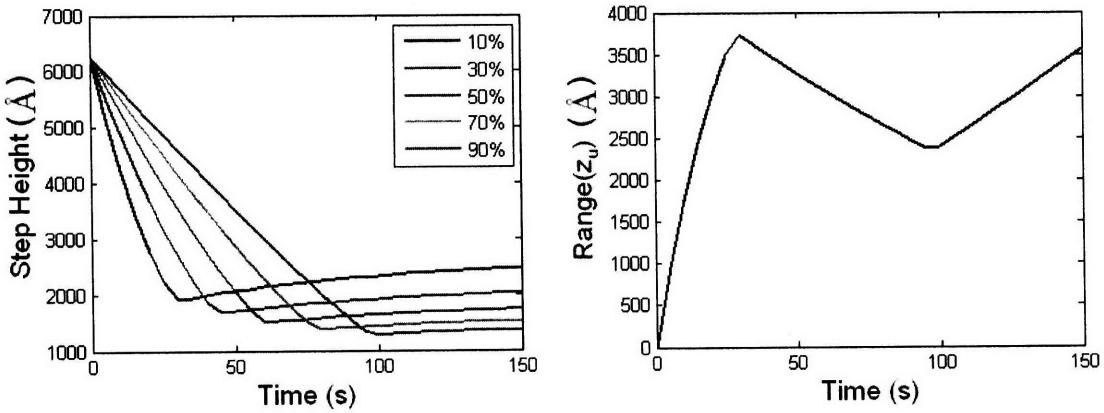



Figure 3-18: For STI CMP, (left) step-height of different densities as a function of time; (right) $Range(z_u)$, which is defined as $z_u(\rho = 90\%) - z_u(\rho = 10\%)$, as a function of time.

values. The planarization length captures the spatial range of neighboring topographical effects during CMP. During the stage of single material removal, the wafer surface has large local step-heights and feature-to-feature variations, which makes the neighboring interactions over short distances more important. During the dual material polishing, on the other hand, the surface is relatively smooth at both the feature scale and die scale, and the interaction over longer ranges becomes significant.

The effect of the characteristic step-height $h^*$ is shown in Figure 3-19 and 3-20, where two $h^* =$ values, 200 Å and 1000 Å, are chosen for comparison. Its effect is similar to that in oxide CMP; however, a smaller $h^*$ results in a smaller steady state step-height, which is desirable, and larger topography variation, which is not desirable.

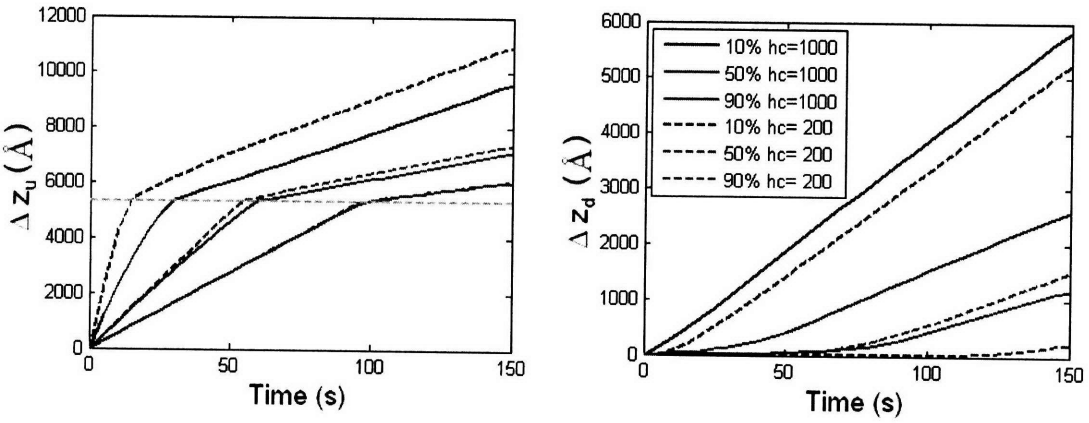



Figure 3-19: Comparison of model predictions for up and down areas using different values of $h^*$: $h^* = 1000$ Å (solid line) and $h^* = 200$ Å (dashed line).

The effect of selectivity $s$ is shown in Figure 3-21 and 3-22, where two $s =$ values, 4 and 10, are chosen for comparison. A higher selectivity causes slower removal in the dual material polishing stage, and less long-range topography variation. Figure 3-22 shows, however, that a higher selectivity results in a higher step-height in the dual material polishing stage.

If the physically-based PDSH model is used, the removal rates during the dual-

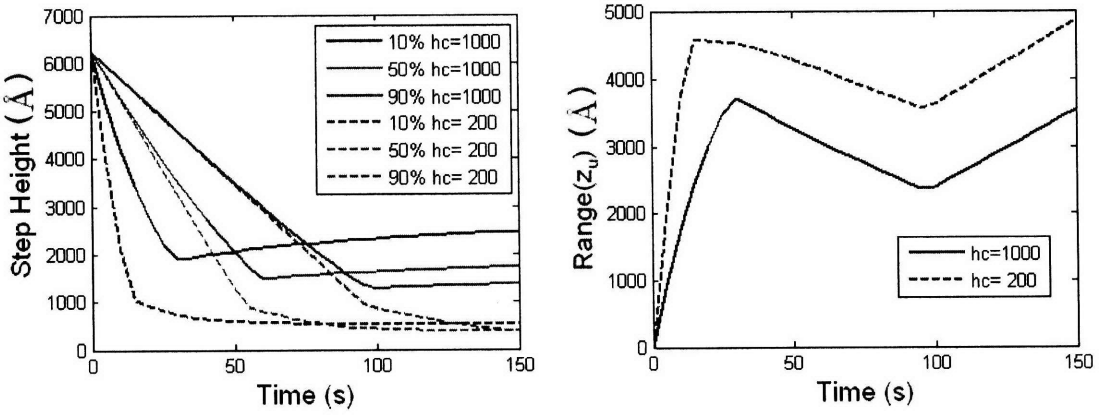



Figure 3-20: Comparison of model predictions for step height and long-range topography variation using different values of $h^*$: $h^* = 1000$ Å (solid line) and $h^* = 200$ Å (dashed line).

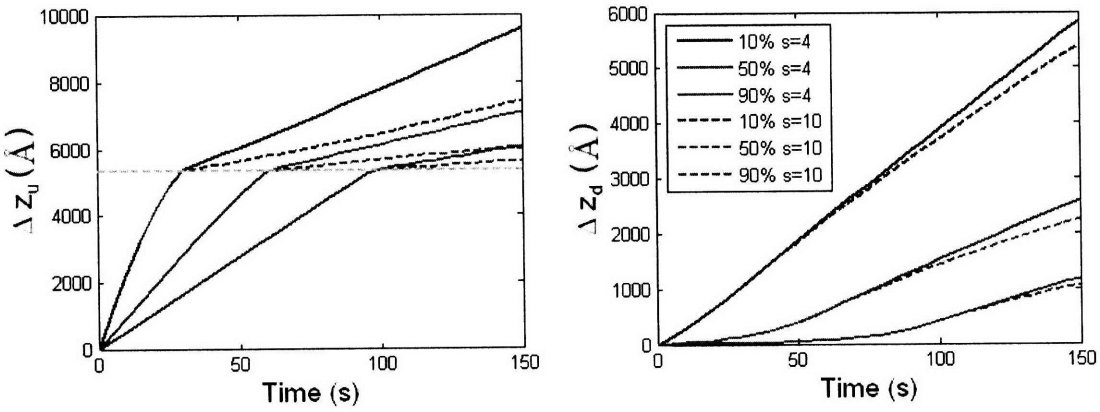



Figure 3-21: Comparison of model predictions for up and down areas using different values of selectivity $s$: $s = 4$ (solid line) and $s = 10$ (dashed line).

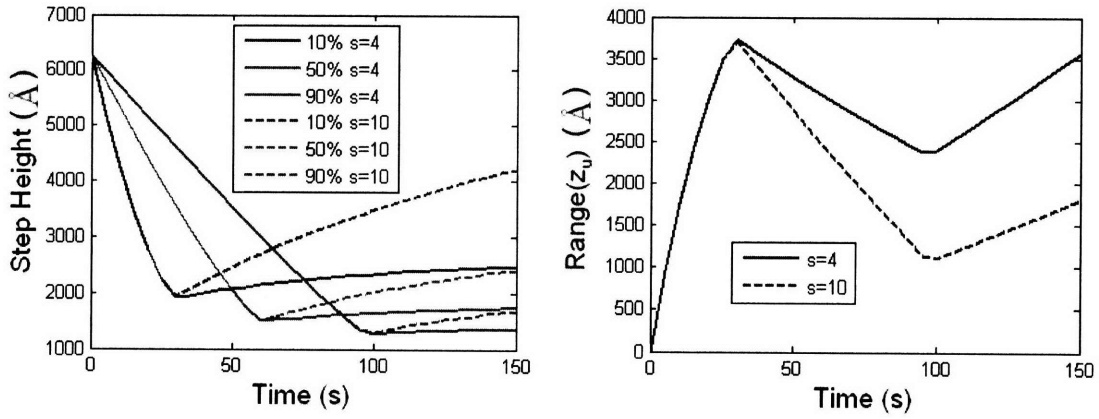



Figure 3-22: Comparison of model predictions for step height and long-range topography variation using different values of selectivity $s$: $s = 4$ (solid line) and $s = 10$ (dashed line).

material polishing stage are

$$\begin{cases} RR_d &= \frac{1}{1+\rho(e^{h/\lambda}-1)}K \\ RR_u &= \frac{1}{s}\frac{e^{h/\lambda}}{1+\rho(e^{h/\lambda}-1)}K \end{cases} \tag{3.30}$$

In the steady state, we have $h_{ss} = \lambda \ln(s)$ and $K_{ss} = \frac{K}{1+\rho(s-1)}$. The initial condition for the dual material polishing stage can be obtained by solving the single material polishing equation for $\Delta z_u(t_c) = z_0$, where $t_c$ is the clearing time. The step-height $h(t)$ during dual material polishing satisfies the following equation.

$$e^{\frac{s}{1+\rho(s-1)}\rho h(t)/\lambda}\left(1 - s \cdot e^{-h(t)/\lambda}\right) = e^{\frac{1}{1+\rho(s-1)}(s\rho h(t_c)-K(t-t_c))/\lambda}\left(1 - s \cdot e^{-h(t_c)/\lambda}\right) \tag{3.31}$$

Once $h(t)$ is obtained using Newton's method, $\Delta z_d(t)$ and $\Delta z_u(t)$ can be solved for using Equation 3.32.

$$\begin{cases} \Delta z_u(t) &= \Delta z_u(t_c) + \frac{K(t-t_c)+(1-\rho)(h(t_c)-h(t))}{1+\rho(s-1)} \\ \Delta z_d(t) &= \Delta z_d(t_c) + \frac{K(t-t_c)-\rho(h(t_c)-h(t))}{1+\rho(s-1)} \end{cases} \tag{3.32}$$

The predicted time evolutions of $\Delta z_u$ and $\Delta z_d$ are shown in Figure 3-23, which is similar to the evolution predicted by exponential PDSH model shown in Figure 3-17.

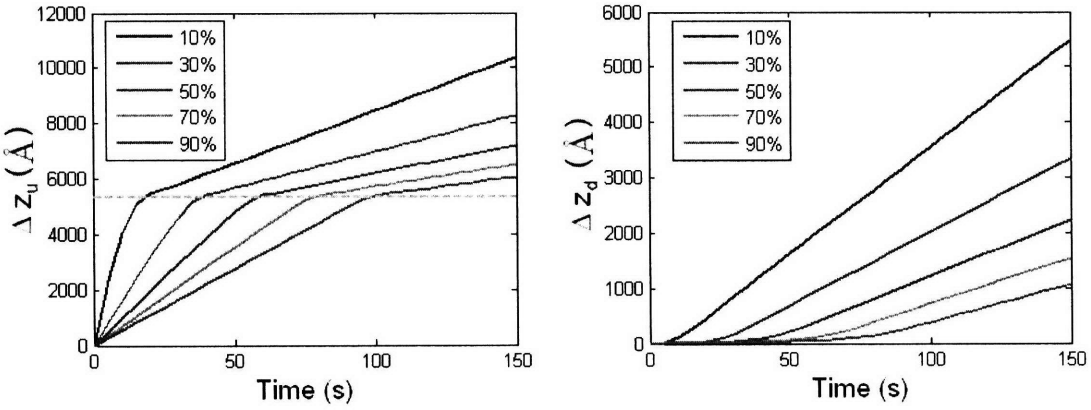



Figure 3-23: Amount remove during STI CMP for different pattern densities as a function of time, predicted by the physically-based PDSH model: (left) $\Delta z_u$ and (right) $\Delta z_d$.

### 3.4.3 Single Material Polishing with Non-Conventional Slurry

As an example application of the PDSH model to a non-conventional slurry, the removal rate dependence on pressure is assumed to be given by Equation 3.7, where the dependence of removal rate on pressure consists of two linear segments with transition pressure $P_t$ and the segment of lower pressures has the smaller slope. In this CMP process, $P_0$ is preferred to be larger than $P_t$ to achieve a reasonable blanket polishing rate, otherwise, a small blanket rate will clear the overburden oxide in the raised regions too slowly. To avoid this, the pressure on raised areas is kept larger than $P_t$ by setting $P_0 > P_t$. The removal rate equations are illustrated in Figure 3-24(a), and can be written as follows.

$$\begin{cases} K_d &= \begin{cases} r_p K \cdot e^{-\rho h/h^*} & \text{when} \quad h \geq h_t \\ K \cdot e^{-\rho h/h^*} - (1 - r_\beta) r_p K & \text{when} \quad h < h_t \end{cases} \\ K_u &= K\left(\frac{1}{\rho} - \frac{1-\rho}{\rho} e^{-\rho h/h^*}\right) - (1 - r_\beta) r_p K \end{cases} \tag{3.33}$$

where $r_p = P_t/P_0$, $r_\beta = \beta_1/\beta_2$, $K = \beta_2 P_0$, and $h_t$ is the transition step-height when $P(h_t) = P_t$ and $h_t = h^*/\rho \ln(P_0/P_t)$. The ceria blanket removal rate is $K_{ceria} = (1 - (1 - r_\beta) r_p) K$.

Solving the equations for the case when $h \geq h_t$ gives the following result. We define $t_t$ as the transition time when $h(t) = h_t$; then $t_t$ can be determined using

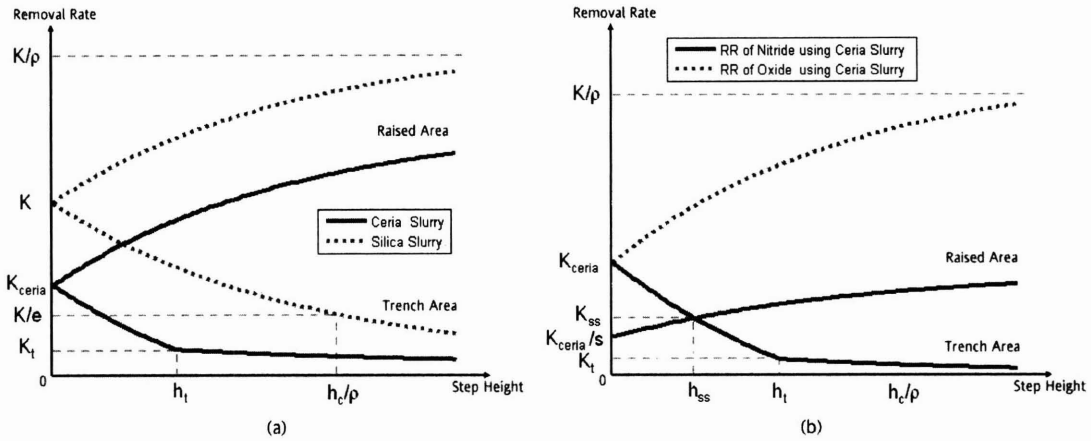



Figure 3-24: Relationship between removal rate and step-height in the exponential PDSH model for a CMP process using ceria slurry. (a) Comparison of polishing between a ceria slurry and silica slurry; (b) comparison between nitride and oxide polishing.

Newton's method.

$$\begin{cases} \Delta z_d(t) &= \frac{r_\beta h^*}{1-\rho(1-r_\beta)} \ln\left[1 + \frac{1-\rho(1-r_\beta)}{1-\rho r_p(1-r_\beta)}\left(e^{(1-\rho r_p(1-r_\beta))\frac{Kt}{h^*}} - 1\right) e^{-\rho h_0/h^*}\right] \\ \Delta z_u(t) &= \frac{1-\rho r_p(1-r_\beta)}{\rho} Kt - \frac{1-\rho}{\rho r_\beta}\Delta z_d(t) \\ h(t) &= h_0 + \Delta z_d(t) - \Delta z_u(t) \end{cases} \tag{3.34}$$

For the case when $t > t_t$, we have the closed form solutions for thickness and step height given by Equation 3.35. The amount removed and step-height evolutions are plotted in Figure 3-25 and 3-26, in which $r_p = 0.9$, $r_\beta = 0.1$, and the other parameters are chosen to be the same as those in the conventional slurry example.

$$\begin{cases} \Delta z_d(t) &= \Delta z_d(t_t) + h^* \ln\left[1 + e^{-\rho h_t/h^*}\left(e^{K(t-t_t)/h^*} - 1\right)\right] \\ & \quad -(1-r_\beta) r_p K(t-t_t) \\ \Delta z_u(t) &= \Delta z_u(t_t) + \frac{1-\rho r_p(1-r_\beta)}{\rho} K \cdot (t-t_t) \\ & \quad -\frac{1-\rho}{\rho} h^* \ln\left[1 + e^{-\rho h_t/h^*}\left(e^{K(t-t_t)/h^*} - 1\right)\right] \\ h(t) &= h_0 + \Delta z_d(t) - \Delta z_u(t) \end{cases} \tag{3.35}$$

The figures for the ceria-based non-conventional slurry look very similar to those for the conventional silica slurry. One difference is the dependence on pattern-density. The lines representing low pattern densities are much closer together than those

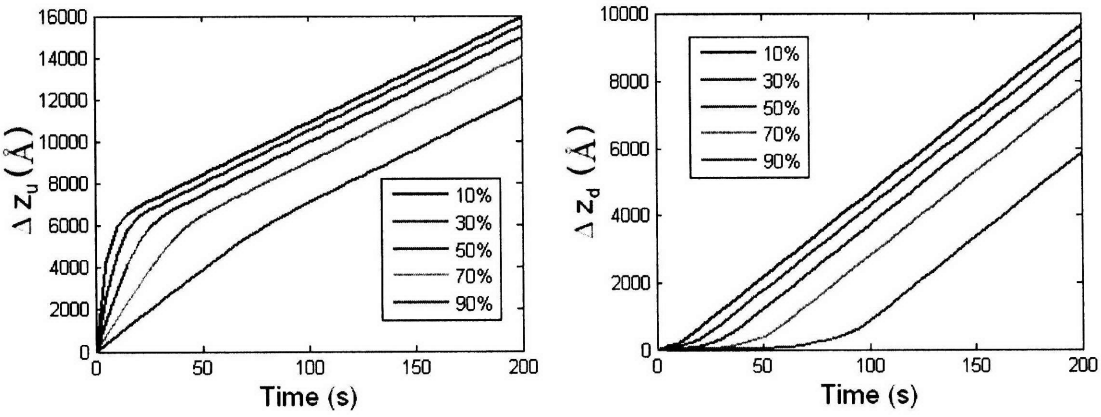



Figure 3-25: For STI CMP using a ceria-based slurry, amount removed for different pattern densities as a function of time: (left) $\Delta z_u$ and (right) $\Delta z_d$.

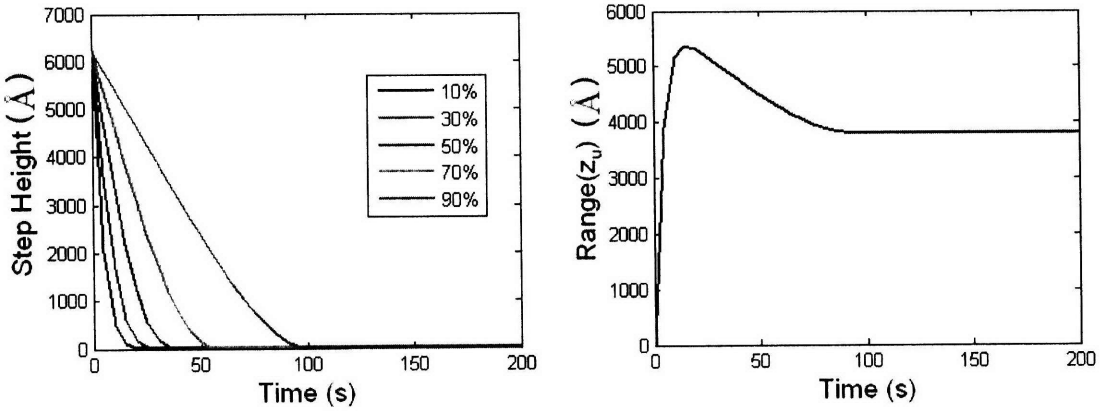



Figure 3-26: For STI CMP using a ceria-based slurry, (left) step-height of different pattern densities as a function of time; (right) $Range(z_u)$, which is defined as $z_u(\rho = 90\%) - z_u(\rho = 10\%)$, as a function of time.

representing high pattern densities. Comparisons between ceria-based slurry and silica-based slurry are shown in Figures 3-27 and 3-28. The blanket removal rates are chosen to be the same for a fair comparison. The ceria-based slurry shows a desired performance: a faster reduction of step-height is observed, resulting in smaller amount removed non-uniformity.

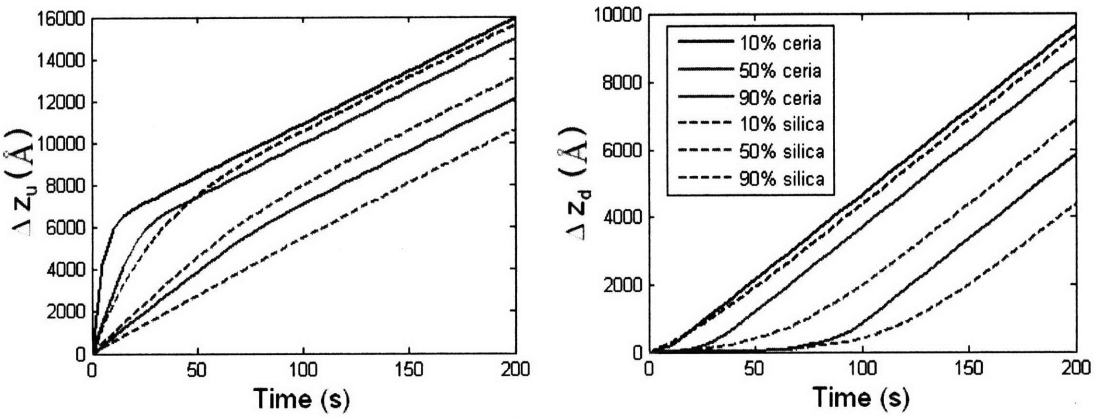



Figure 3-27: Comparison of a ceria-based (solid line) and silica-based (dashed line) slurries. For STI CMP using a ceria-based slurry, amount removed for different pattern densities as a function of time: (left) $\Delta z_u$ and (right) $\Delta z_d$.

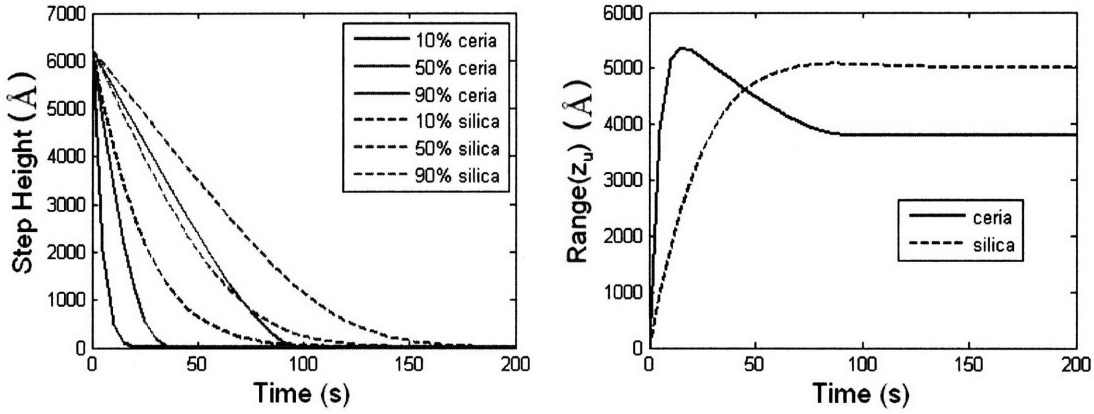



Figure 3-28: Comparison of a ceria-based (solid line) and silica-based (dashed line) slurries. (Left) step-height of different pattern densities as a function of time; (right) $\Delta z_u$ range, which is defined as $\Delta z_u(\rho = 10\%) - \Delta z_u(\rho = 90\%)$, as a function of time.

### 3.4.4 Dual Material Polishing with Non-Conventional Slurry

For the dual material polishing stage, if the local step-height is always less than $h_t$ in the dual material polishing stage, $t_c$ can be solved using $\Delta z_u(t_c) = z_0$ by Newton's method. For dual material polishing, we only need to scale the raised area removal rate by $1/s$, where $s$ is the selectivity, and the solution is obtained as below.

$$\begin{cases} \Delta z_d(t) &= \Delta z_d(t_c) - r_p(1-r_\beta)K(t-t_c) \\ & \quad + h^* \frac{s}{1+(s-1)\rho} \ln\left[1 + \frac{1+\rho(s-1)}{1+\rho(s-1)r_p(1-r_\beta)} e^{-\rho h(t_c)/h^*} \left(e^{\frac{K'}{s \cdot h^*} \cdot (t-t_c)} - 1\right)\right] \\ \Delta z_u(t) &= \Delta z_u(t_c) + \frac{1-r_p(1-r_\beta)}{s\rho} K \cdot (t-t_c) - \frac{1-\rho}{s\rho}(\Delta z_d(t) - \Delta z_d(t_c)) \\ h(t) &= h_0 + \Delta z_d(t) - \Delta z_u(t) \end{cases} \tag{3.36}$$

where $K' = \rho(s-1)r_p(1-r_\beta)K$. The problem can be more complicated in reality, as the local step-height could increase in the dual material polishing stage, especially with a high selectivity slurry. Given the modest computational demand, it is easy to perform time-stepping simulation using the instantaneous removal rate given by Equation 3.33. As an illustration, the model simulation is shown in Figure 3-29 and 3-30 with the parameters $r_p = 0.9$ and $r_\beta = 0.1$. In the figures, the results for a silica-based slurry are shown for comparison. The ceria-based slurry causes faster polishing in raised areas and slower polishing in trench areas compared with the silica slurry, which results in faster step-height reduction for the ceria slurry.

The steady state can be determined as as given by Equation 3.37. The resulting steady state step height and removal rates are illustrated in Figure 3-31 assuming $r_\beta = 0.1$ and $r_p = 0.9$. In comparison with a silica based slurry, as illustrated in Figure 3-16, the ceria slurry shows similar steady state removal rate, but much less steady state step height.

$$\begin{cases} h_{ss} &= \begin{cases} \frac{h^*}{\rho} \log\left(\frac{1-\rho+s\rho r_\beta}{1-(1-r_\beta)\rho r_p}\right) & \text{if} \quad h_{ss} \geq h_t \\ \frac{h^*}{\rho} \log\left(\frac{1+(s-1)\rho}{1+(s-1)(1-r_\beta)\rho r_p}\right) & \text{if} \quad h_{ss} < h_t \end{cases} \\ K_{ss} &= \frac{K}{s}\left(\frac{1}{\rho} - \frac{1-\rho}{\rho} e^{-\rho h_{ss}/h^*} - (1-r_\beta)r_p\right) \end{cases} \tag{3.37}$$

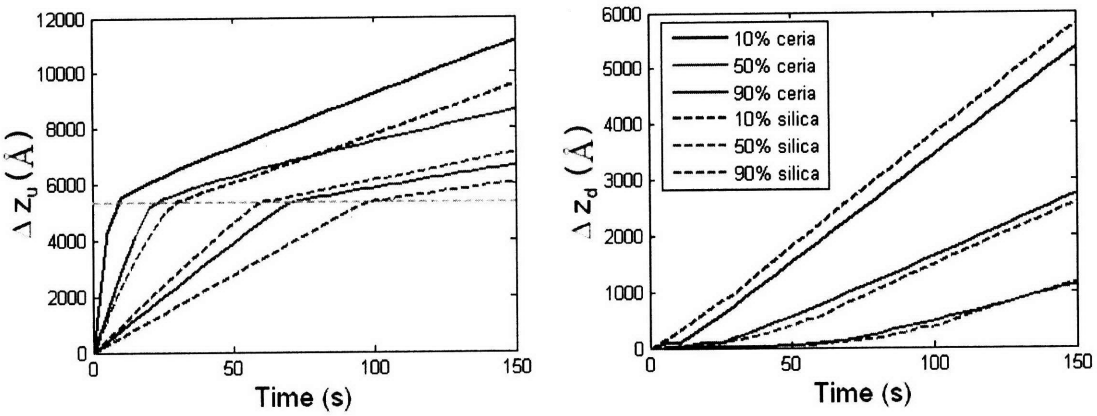



Figure 3-29: Comparison of ceria-based (solid line) and silica-based (dashed line) slurries. For STI CMP using a ceria-based slurry, amount removed for different pattern densities as a function of time: (left) $\Delta z_u$ and (right) $\Delta z_d$.

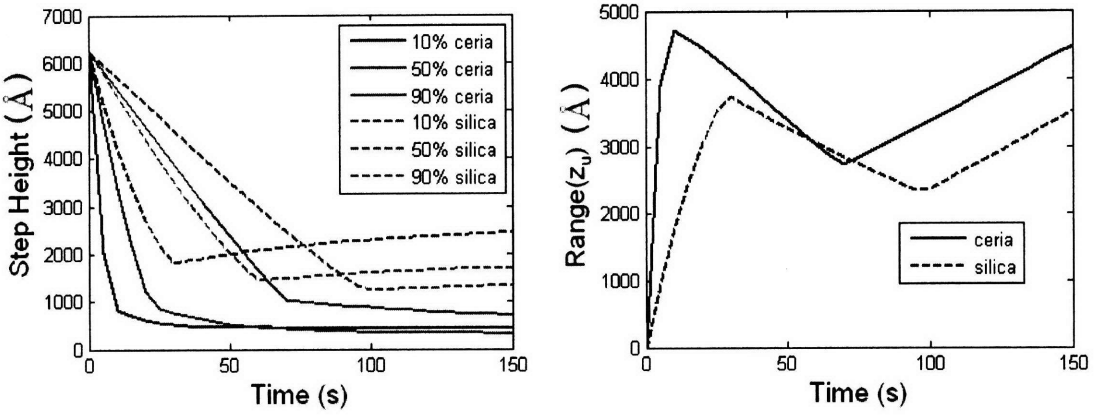



Figure 3-30: Comparison of ceria-based (solid line) and silica-based (dashed line) slurries. (Left) step-height of different densities as a function of time; (right) $\Delta z_u$ range, which is defined as $\Delta z_u(\rho = 10\%) - \Delta z_u(\rho = 90\%)$, as a function of time.

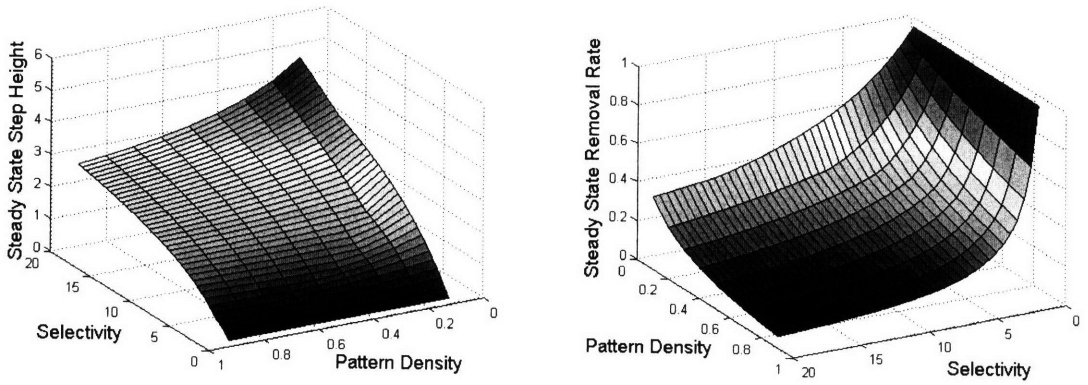



Figure 3-31: Steady state using ceria-base slurry. Left: steady state step-height (in units of $h^*$) vs. selectivity and pattern-density. Right: steady state removal rate (in units of oxide blanket removal rate $K$) vs. selectivity and pattern-density.

### 3.4.5 Optimal Pressure-Dependent Slurry

In the previous section, the two-segment behavior of ceria slurry (i.e., having two different rate versus pressure regions as defined by Equation 3.7) shows improvement over a conventional slurry with a single linear Prestonian rate dependence on pressure. A natural question to ask is what kind of rate versus pressure dependence would give the best polishing performance. The CMP process planarizes the surface, because higher pressure in raised areas results in higher removal rates. In polishing wafers with pattern structures, the pressure on the raised region with low pattern-density areas is higher and causes a higher removal rate in these low pattern-density areas, which results in with-in-die non-uniformity. An optimal slurry needs to not only efficiently reduce local step-heights, but also minimize within-die non-uniformity.

The advantage of a ceria-based slurry lies in its small removal rate under low pressure, which greatly enhances the ratio $K_u/K_d$. This property results in a low ratio of removal rate in the trench area to that in the raised area, and it greatly improves the efficiency of reducing the local step-height. The drawback is that, given the same blanket rate, the removal rate in raised areas has a stronger dependence on pattern-density, as seen in Figure 3-32. The larger difference in raised area removal rate could cause more within-die non-uniformity, which is undesirable.

The slurry could perform better, if the property of the low removal rate in trench

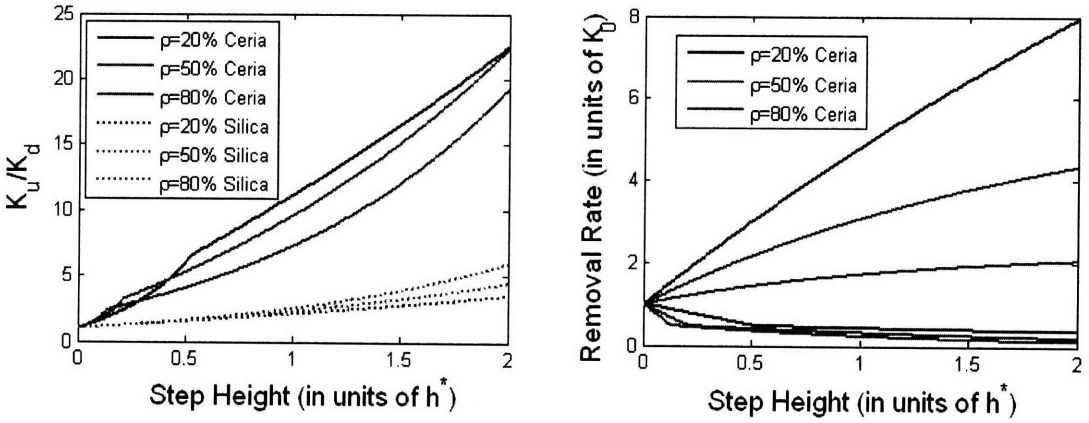



Figure 3-32: Comparison of removal rate dependence on step-height between ceria and silica slurries.

areas is maintained, and the removal rate is made more uniform in raised areas. One way to achieve this is to engineer a slurry that is similar to a ceria slurry, but which has a lower slope of removal rate versus pressure at high pressures. A three linear-segment form is used to describe the proposed dependence, as given by Equation 3.38. The dependence is illustrated in Figure 3-33.

$$K = \begin{cases} \beta_L P & \text{when} \quad P \leq P_L \\ \beta_L P_L + \beta(P - P_L) & P_L < P \leq P_H \\ \beta_L P_L + \beta(P_H - P_L) + \beta_H(P - P_H) & P_H < P \end{cases} \tag{3.38}$$

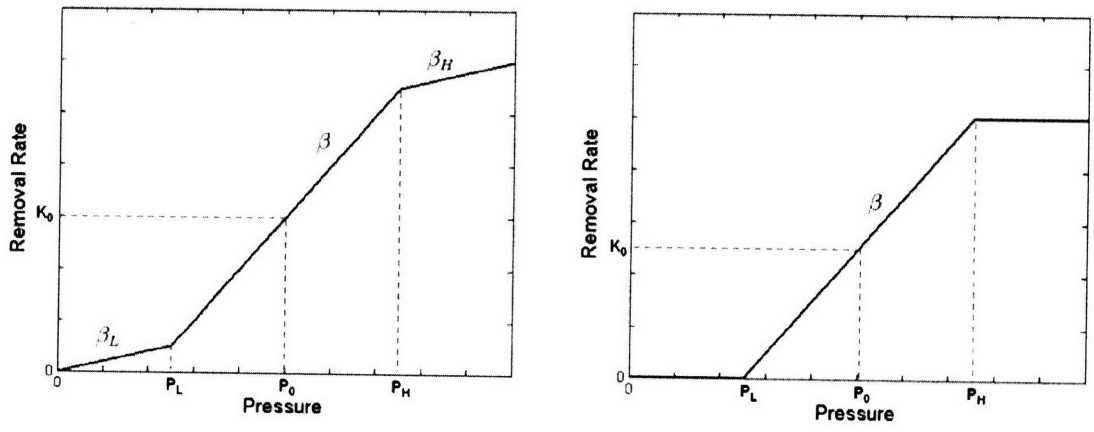



Figure 3-33: Illustration of the removal rate dependence on pressure of improved slurries with $L_P = 0.5P_0$, $P_H = 1.5P_0$; (a) $\beta_L = \beta_H = 0.2\beta$, and (b) $\beta_L = \beta_H = 0$.

For simplicity, let us consider the special case with $\beta_L = 0$ and $\beta_H = 0$, as in Figure 3-33(b). For a CMP process with a blanket removal rate $K_0$ at applied pressure $P_0$, we represent $P_L = r_L \cdot P_0$ and $P_H = r_H \cdot P_0$, and obtain the following equation.

$$K = \begin{cases} 0 & \text{when} \quad P/P_0 \leq r_L \\ \frac{P/P_0 - r_L}{1 - r_L} K_0 & r_L < P/P_0 \leq r_H \\ \frac{r_H - r_L}{1 - r_L} K_0 & P/P_0 > r_H \end{cases} \tag{3.39}$$

Figure 3-34 shows the removal rate dependence on step-height for two sets of parameters: the left plot uses $r_L = 0.8$ and $r_H = 1.2$, while the right plot uses $r_L = 0.8$ and $r_H = 2$. The two different $r_H$ values in these two examples result in different cut-off removal rates in the raised areas; in both examples, the presence of a cut-off rate forces $K_u$ of areas with different pattern densities to be tightly bounded.

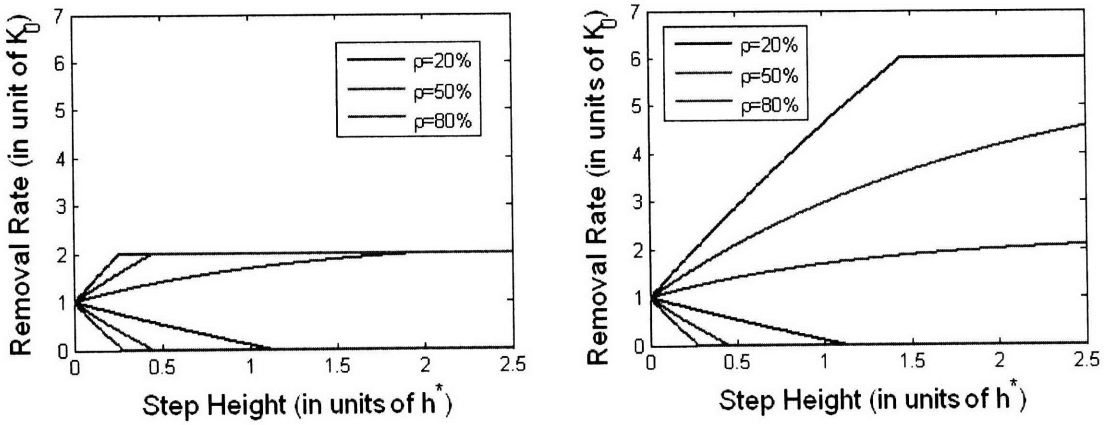



Figure 3-34: The dependence of removal rate on step-height for improved slurries. Left plot uses $r_L = 0.8$ and $r_H = 1.2$, while the right plot uses $r_L = 0.8$ and $r_H = 2$.

The results of model simulation are shown in Figure 3-35 and Figure 3-36, and the improved slurries show a large improvement over the conventional silica slurry, having faster step-height reduction and less post-CMP topogrpahy variation. Similar results can be seen in the case of dual material polishing, such as in STI CMP, as shown in Figure 3-37 and Figure 3-38.

The improved slurry, which is designed to be insensitive to pressure in both low pressure and high pressure regimes, more efficiently reduces step-height and signifi-

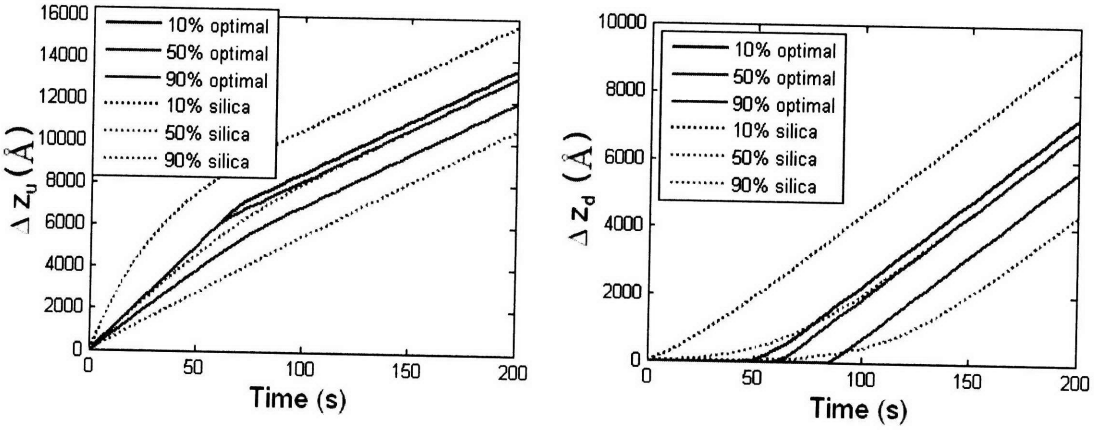



Figure 3-35: Model simulation for single material CMP using an improved slurry (solid) with $r_L = 0.8$ and $r_H = 1.2$, in comparison with that of a conventional slurry (dotted). (Left) $\Delta z_u$, and (right) $\Delta z_d$.

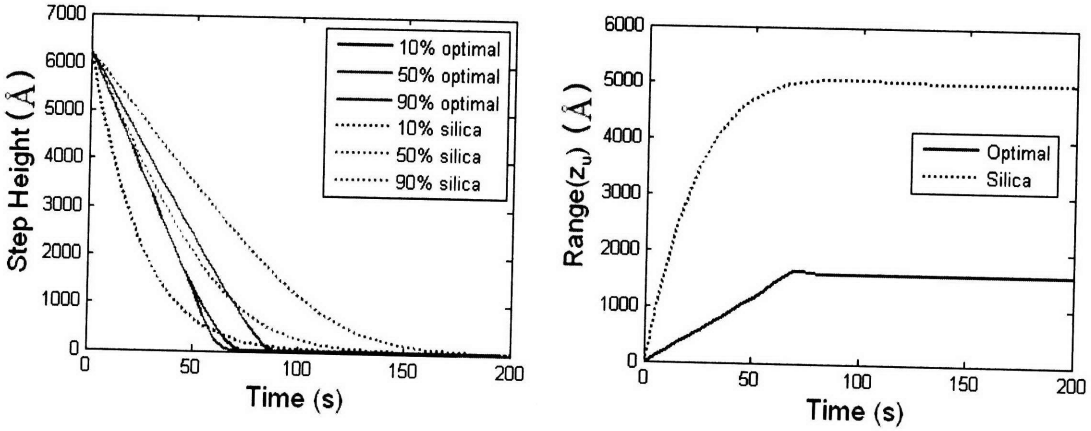



Figure 3-36: Model simulation for single material CMP using an improved slurry (solid) with $r_L = 0.8$ and $r_H = 1.2$, in comparison with that of a conventional slurry (dotted). (Left) step-height $h(t)$, and (right) $Range(z_u)$.

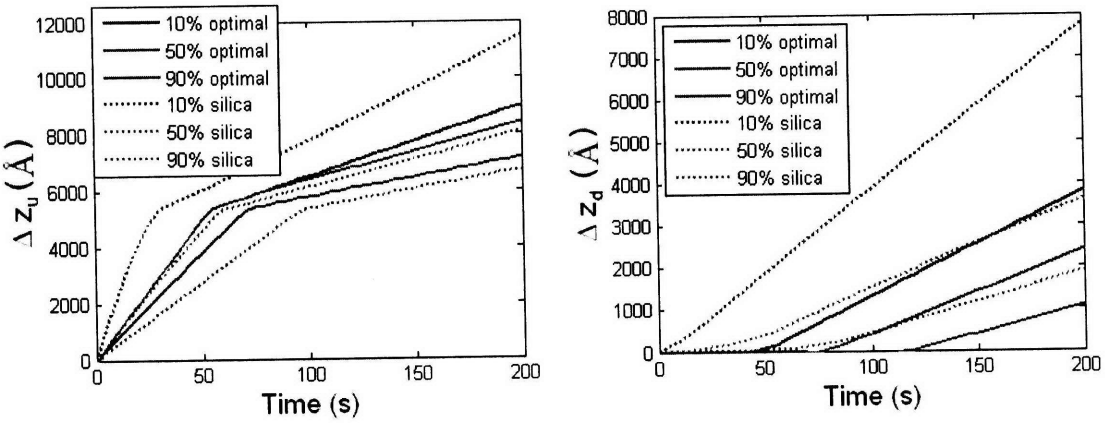



Figure 3-37: Model simulation for STI CMP using an improved slurry (solid) with $r_L = 0.8$ and $r_H = 1.2$, in comparison with that of a conventional slurry (dotted). (Left) $\Delta z_u$, and (right) $\Delta z_d$.

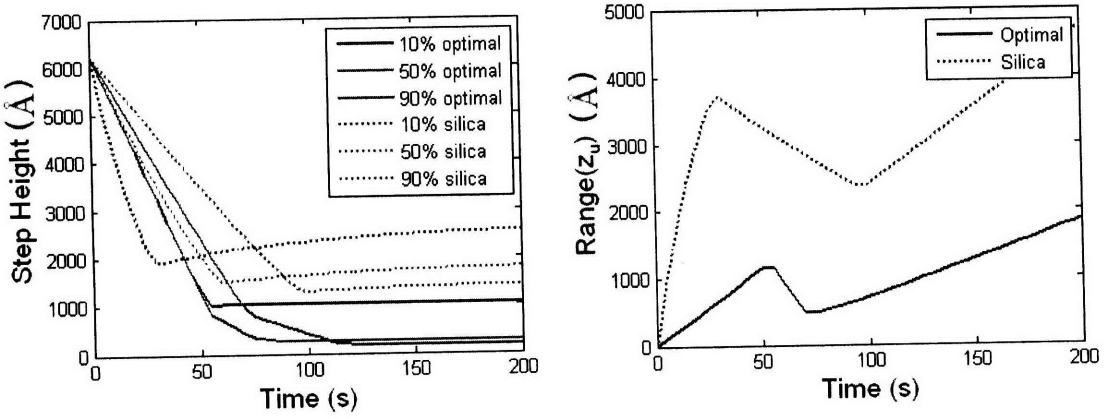



Figure 3-38: Model simulation of STI CMP using an improved slurry (solid) with $r_L = 0.8$ and $r_H = 1.2$, in comparison with that of a conventional slurry (dotted). (Left) step-height $h(t)$, and (right) $Range(z_u)$.

cantly improves post-CMP topography variation. If the dependence of removal rate on pressure is as described in Equation 3.38, here is a summary of how each parameter affects the die-level CMP can be summarized below.

- Small $\beta_1$ and large $P_{t1}$, as long as $P_{t1} < P_0$, effectively enhances the ratio of $K_u$ to $K_d$ and improves the step-height reduction efficiency, and thus are favorable. A possible drawback is a small blanket removal rate.

- Small $\beta_2$ and small $P_{t2}$, as long as $P_{t2} > P_0$, makes the removal rate of different raised areas more uniform and reduces the long-range post-CMP topography variation, and thus is favorable. As a negative, it reduces the ratio of $K_u$ to $K_d$ and results in less efficient step-height reduction. A good balance point depends on the process requirements, as well as on the pattern-density range of the product chip.

- Although it has not been discussed, small $r_L$ and $r_H$ are preferred. A smaller $r_L$ results in a larger ratio of $K_u$ to $K_d$ and improves the step-height reduction efficiency. A smaller $r_H$ results in more uniform removal rates in the raised areas.

Another concern is the robustness of the slurry performance against pressure variation across the wafer, especially near the wafer-edge. If $P_0 < P_{t1}$, $K_u$ becomes too small to effectively polish raised area below certain step-height. If $P_0 > P_{t2}$, $K_d$ becomes nearly the same as $K_u$ below a certain step-height, and CMP fails to planarize in both cases. Even if the condition $P_{t1} < P_0 < P_{t2}$ is always satisfied, the variation of $P_0$ across the wafer can cause different polishing performance across the wafer, and increase the wafer-level non-uniformity.

## 3.5 Applying the Physically-Based Model

The physically-based model has broader applications than the PDSH models, and this section focuses on applications in which the PDSH model cannot be used. First, the physically-based model is compared with the PDSH models, and the relationship

between the two sets of model parameters are discussed. Second, the physically-based model can simulate CMP processes for wafer having initial topography variation, which may arise, for example, due to nanotopography in STI or prior CMP or plating topography in multi-level metal CMP. Third, the physically based model is used to study how the planarization process is affected by pad properties. And last, the effect of applied pressure, which is an important process design and process control variable, is studied.

### 3.5.1 Comparison between physically-based and PDSH CMP Models

The physically-based model and PDSH models have different model parameters, so we first need to find the relationship between the two sets of model parameters. Then, the predicted surface evolutions are compared, and last, the dependence of PDSH model parameters on the initial layout structure, such as initial film thickness and initial step-height, is discussed.

#### The Model Parameters

For each model, the model parameters are determined by fitting to experimental data and minimizing the fitting error, which is the differences between model prediction and measurement data. In this way, one data set can be used to determine the parameters $L_P$ and $h^*$ of the PDSH model, as well as the parameters $E$ and $\lambda$ of the physically-based model. We are interested in the relationship between the extracted $(L_P, h^*)$ and $(E, \lambda)$. Without resort to experimental data, the relationship can be determined based on simulation studies in the following two steps:

1. Choose $(E, \lambda)$, and use the physically-based model to simulate the surface evolution at different time intervals during CMP.

2. The simulated results are treated as pseudo experimental data and used to calibrate the PDSH model, and the extracted $(L_P, h^*)$ is recorded.

The physically-based model is used to generate the pseudo data, because it is believed to better capture the dynamics of the CMP process.

Repeating these steps for different values of $(E, \lambda)$, we obtain the relationships $h^*(E, \lambda)$ and $L_P(E, \lambda)$, which are illustrated in Figure 3-39. The left figure reveals that the planarization length $L_P$ increases almost linearly with the effective Young's modulus $E$ and is not sensitive to the characteristic asperity height $\lambda$. The right figure shows that the characteristic step-height $h^*$ increases linearly with $\lambda$ and is not sensitive to $E$. The result confirms our intuition that the long-range pattern-density or neighboring effect comes from the bending of the pad bulk, and the pressure dependence on step-height results from the asperity height distribution.

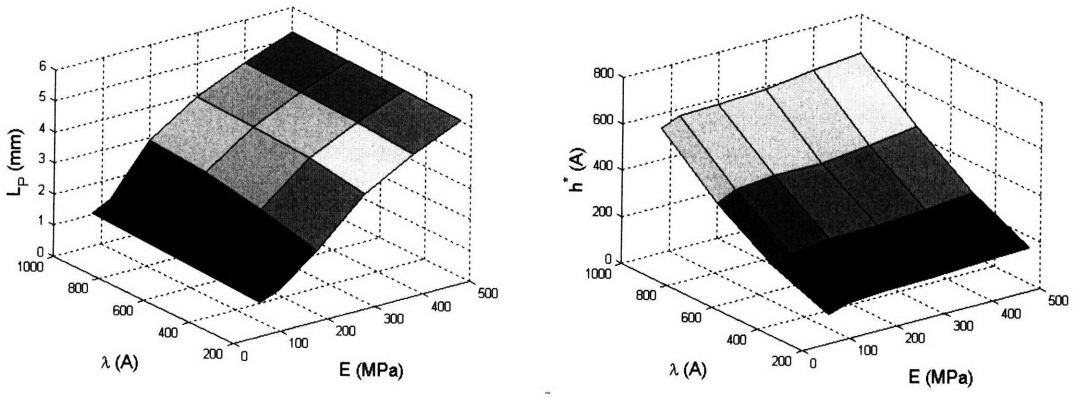



Figure 3-39: The dependence of PDSH model parameters on the parameters of the physically-based CMP model: (a) planarization length $L_P(E, \lambda)$ and (b) characteristic step-height $h^*(E, \lambda)$.

We can also obtain the inverse functions $E(L_P, h^*)$ and $\lambda(L_P, h^*)$, as shown in Figure 3-40. Although the functions are not in themselves meaningful, they can be useful in two ways. First, if certain values of $L_P$ and $h^*$ are desired, the inverse functions give an estimate on the values of $E$ and $\lambda$ to try. Second, the calibration of the physically-based model is time-consuming (as discussed in Section 3.6.3), and narrowing down the parameter ranges or initial success for parameter values can be helpful. Thus the data can be calibrated using the PDSH model first, and the extracted $L_P$ and $h^*$ can then be used to estimate the values of $E$ and $\lambda$.

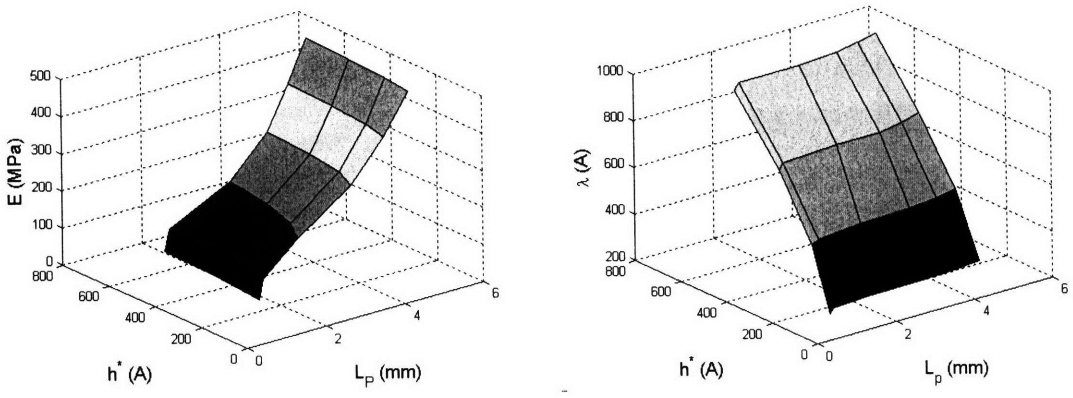



Figure 3-40: The dependence of physically-based parameters on those of the PDSH model: (a) Young's modulus $E(L_P, h^*)$ and (b) characteristic asperity height $\lambda(L_P, h^*)$.

### The Predicted Surface Evolution

Figures 3-41 and 3-42 show one comparison of the two models in terms of the predicted surface evolution. The parameters of the physically-based model used are a blanket removal rate $K = 6000$ Å$/min$, effective Young's modulus $E = 80\ MPa$, and characteristic asperity height $\lambda = 1000$ Å. In the model extraction using the PDSH model, the blanket removal rate is set to be 6000 Å$/min$ and the extracted parameters are planarization length $L_P = 1.37\ mm$ and characteristic step-height $h^* = 566$ Å, with a fitting error of 214 Å. The figures show a good agreement between the two models. It is worth noticing that the evolution of $Range(z_u)$ differs in the predictions. The PDSH model predicts that $Range(z_u)$ remains constant after about 40 seconds; however, the physically-based model predicts a slow decay even for long times, and suggests that CMP continues to globally planarize acriss different regions, albeit it at a slow rate.

Figures 3-43 and 3-44 show the comparison using a different set of parameters: blanket removal rate $K = 6000$ Å$/min$, effective Young's modulus $E = 500\ MPa$, and characteristic asperity height $\lambda = 250$ Å. The extracted PDSH model parameters are planarization length $L_P = 6.03\ mm$ and characteristic step-height $h^* = 80$ Å, with a fitting error of 275 Å. With a larger fitting error, the figures show less agreement than in the previous case. However, this serves as an excellent example to illustrate the

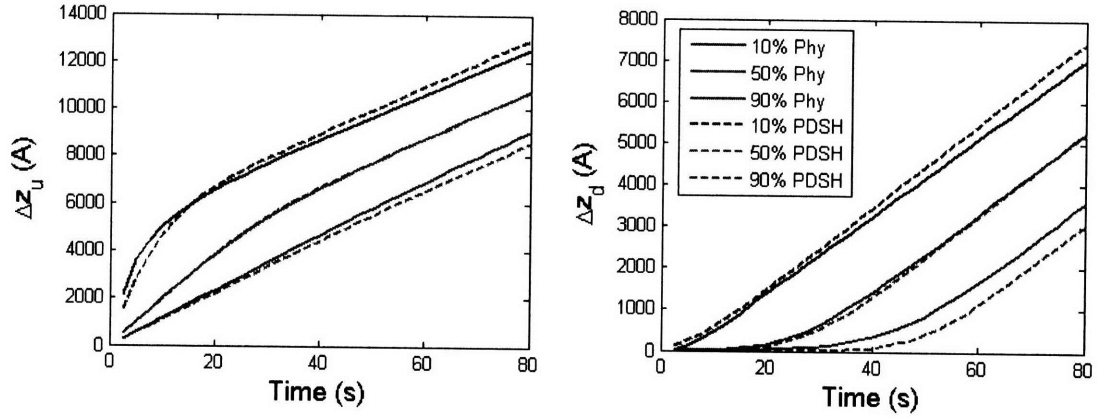



Figure 3-41: Comparison of the physically-based model and the PDSH CMP model predictions: (a) $\Delta z_u$, and (b) $\Delta z_d$. The prediction of the physically-based model is shown as solid lines, and that of PDSH model as dashed lines.

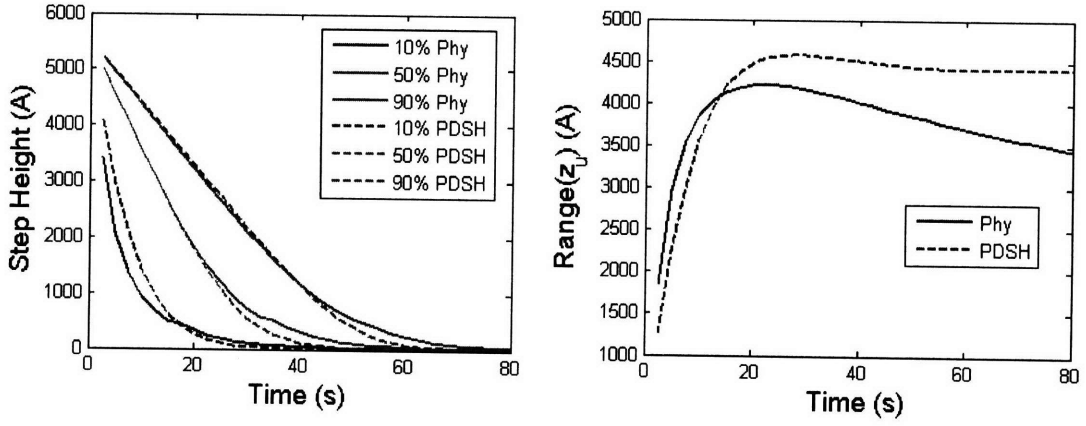



Figure 3-42: Comparison of the physically-based model and the PDSH CMP model predictions: (a) step-height, and (b) $Range(z_u)$, which is the different between $z_u(\rho = 90\%)$ and $z_u(\rho = 10\%)$. The prediction of physically-based model is shown as solid lines, and that of PDSH model as dashed lines.

different dynamics of the two models. From the perspective of the physically based model, at the beginning of the CMP process the profile of the wafer surface, which is defined by the raised area, is flat, and the low pattern-density areas are polished faster due to higher pressure. This faster thickness reduction in low pattern-density areas causes more bending of the pad bulk and results in less pressure and a slower polishing rate. The opposite is true for the high pattern-density areas. The PDSH model, however, captures only the average long-range behavior of the process because a single planarization length is used for the entire polishing process. For amount removed in both the raised and the trench areas, the PDSH model predictions for the 10% region initially under-estimate and later over-estimate the physically-based model simulation, while the opposite is true for the 90% region.

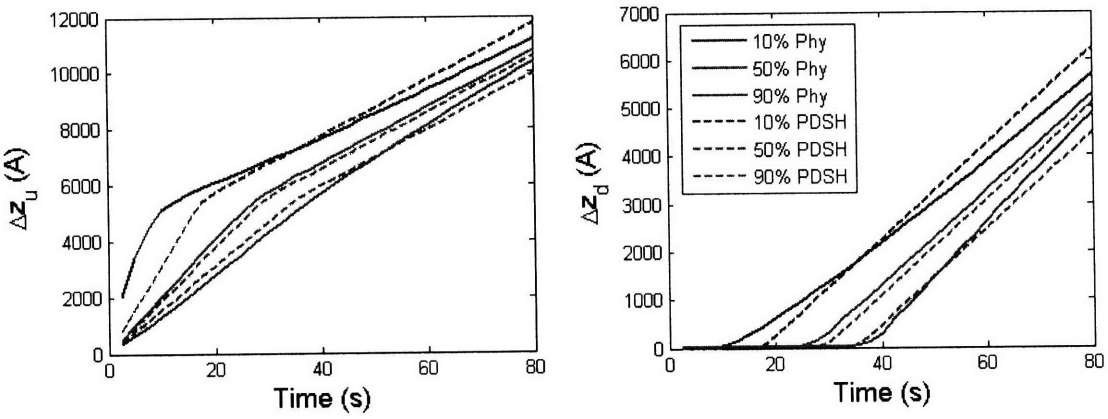



Figure 3-43: Comparison of the physically-based model and the PDSH CMP model predictions: (a) $\Delta z_u$, and (b) $\Delta z_d$. The predictions of the physically-based model are shown as solid lines, and those of the PDSH model as dashed lines.

### 3.5.2 The Impact of Initial Topography

In the previous study, the patterned feature structures are assumed to reside within or on a on flat surface. In practice, the virgin or starting wafer typically has topography variation of about 10 nm to 100 nm height differences over millimeter lateral distances, referred as nanotopography. This nanotopography can affect the STI CMP process used in device isolation. In the multi-layer interconnect metal polishing, the topography variation resulting from previous CMP stpes, or from etch and deposition

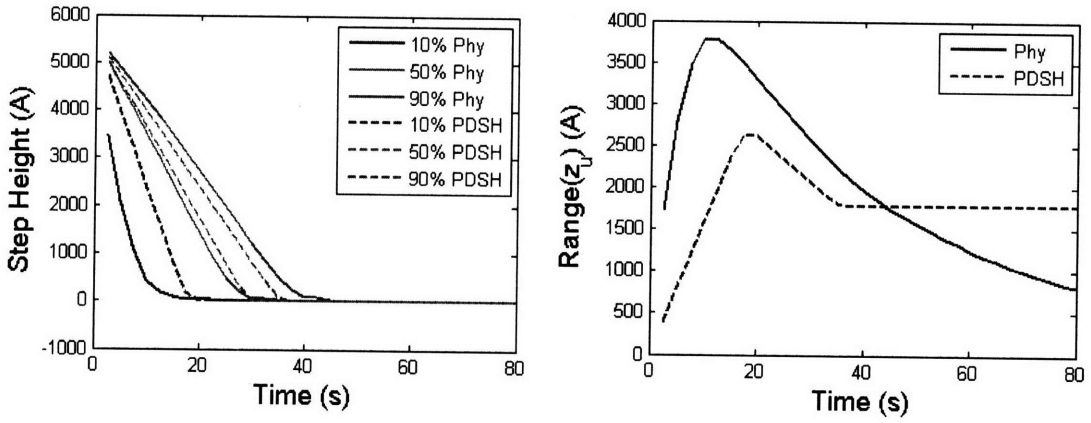



Figure 3-44: Comparison of the physically-based model and the PDSH CMP model predictions: (a) step-height, and (b) $Range(z_u)$, which is the different between $z_u(\rho = 90\%)$ and $z_u(\rho = 10\%)$. The predictions of the physically-based model is shown as solid lines, and those of the PDSH model as dashed lines.

non-uniformity and pattern dependencies, can affect the polishing of later layers. In this section, the physically-based model is used to study the impact of initial topography.

### Endpoint Criteria

CMP is used to planarize the wafer surface, and in practice, although perfect planarization is difficult and costly to achieve, the CMP process can be stopped once a certain endpoint criteria is satisfied. In oxide CMP, the local step-height remaining after CMP can be required to be less than a preset maximum value. Alternatively, one may polish until a target thickness remains in the up areas, e.g., to stop at a given $z_u(\rho = 50\%)$. In dual material polishing, such as STI and metal CMP, the material deposited on the raised areas is required to be completely removed. This section studies the impact of initial topography on the polishing time to achieve the endpoint and the impact on the surface profile at the endpoint.

### The Impact of Initial Topography on Single Material Polishing

To illustrate the impact of initial topography, a simulation of single material polishing with a flat initial surface, Figure 3-45 (a), is compared with polishing of a wafer with initial topography variation, as shown in Figure 3-45 (b).

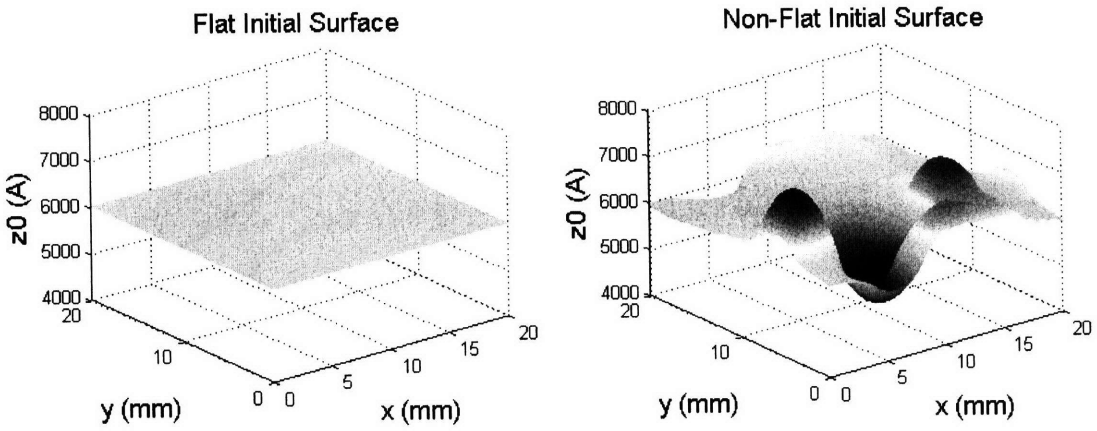



Figure 3-45: Initial surface topography: (a) without variation, and (b) with variation.

The model parameters used in the simulation are Young's modulus $E = 300\ MPa$, characteristic asperity height $\lambda = 500$ Å, selectivity $s = 4$, applied pressure $P_0 = 3\ psi$ and blanket removal rate $K_0 = 3000$ Å$/min$. Without initial topography variation and for structures with initial step height of 6000 Å, the polishing time $t_F$ to reach endpoint is 51.8 $s$ based on a remaining step height criterion of 100 Å. With starting wafer variation, the polishing time $t_F$ is 56.2 $s$. Thus, if the polishing time is chosen assuming no initial topography variation, polishing wafers having the variation will fail to meet the endpoint criteria. Figure 3-46 shows the surface profiles of the raised areas at the endpoint for both cases, and Figure 3-47 shows the step-height distribution. The comparisons clearly show the difference caused by initial topography. Figure 3-48 (a) shows the difference in raised area topography $z_u$ after CMP, and Figure 3-48 (b) shows the difference in the amount of material removed $\Delta z_u$. It is not hard to visually identify that these removals are positively correlated with the initial surface topography in Figure 3-45 (b).

The impact of initial topography can be further illustrated by comparing two extreme cases. Imagine the polishing of a two-level interconnect structure, where we focus on the polishing of the second level dielectric layer. The topography after polishing the first level dielectric defines the initial surface for polishing of the second level dielectric. We consider two cases, both having the same second level layout. In the first case, we assume the second layout has a pattern density $\rho$ that is identical

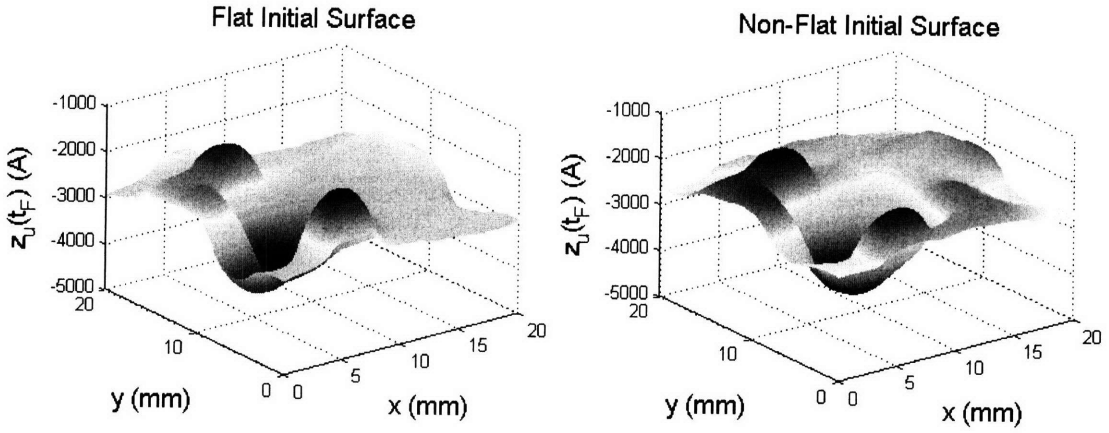



Figure 3-46: Plots of raised area topography across the die: (a) with flat initial topography, (b) with non-flat initial topography.

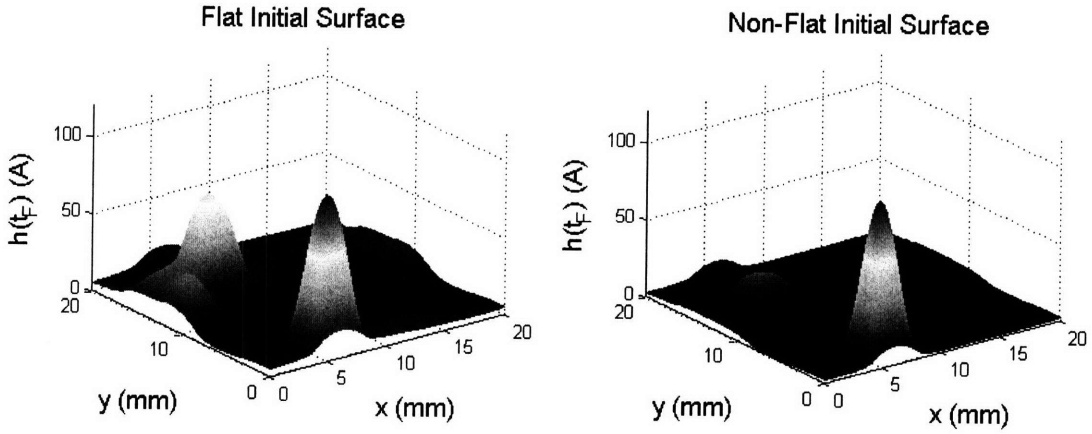



Figure 3-47: Plots of step-height across the die: (a) with flat initial topography, (b) with non-flat initial topography.

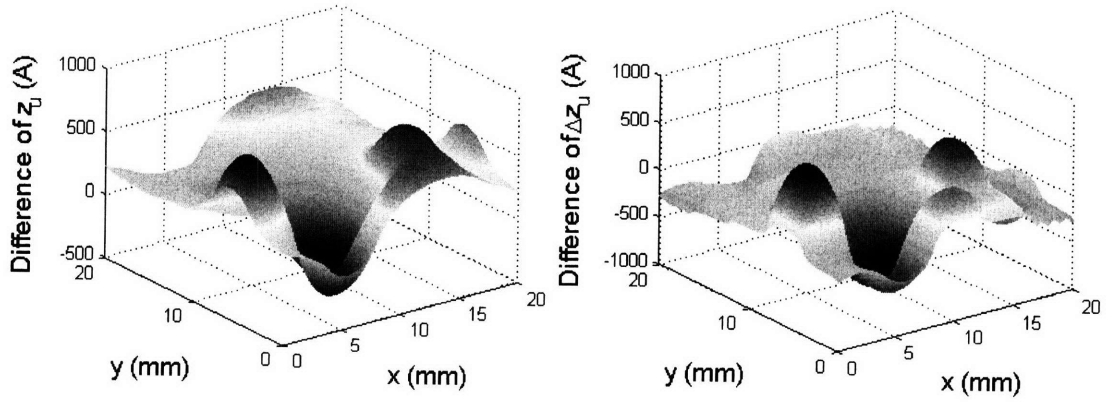



Figure 3-48: Comparison of polishing with and without initial topography variation. (a) the difference in post-CMP raised area topography $z_{u,nonflat} - z_{u,flat}$; (b) the difference in amount removal in raised area $\Delta z_{u,nonflat} - \Delta z_{u,flat}$.

(positively correlated) to the first level layout. In the second case, we assume the second level layout has a pattern density that "compensates" for the first level layout, such that its pattern density is $1 - \rho$ (or is negatively correlated) to that of the first level layout. We refer to these two cases as positively stacked and negatively stacked, respectively. The initial topography before polishing the second layer is shown in Figure 3-49.

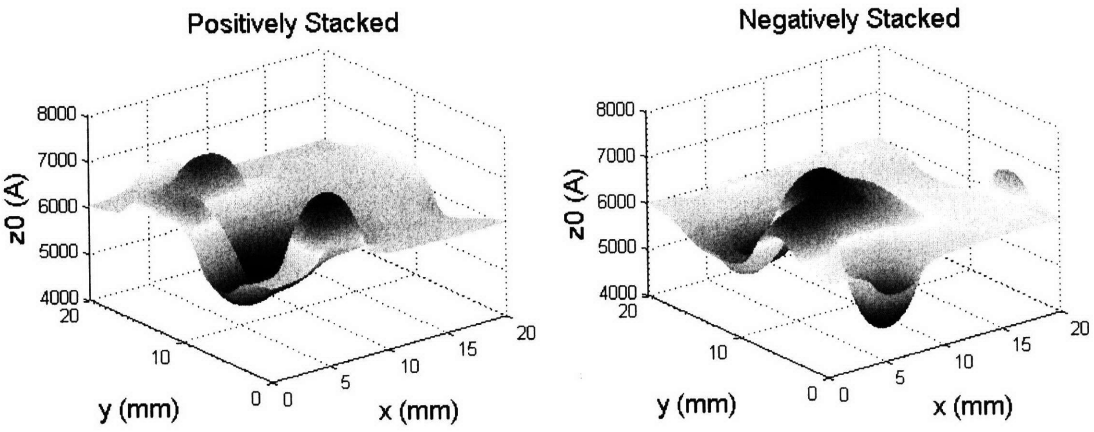



Figure 3-49: The initial topography map of (a) the positively stacked case, and (b) the negatively stacked case.

The same model parameters are used as in the previous example. The different initial topographies result in different polishing times to reach the endpoint: the positively stacked case requires 45.1 $s$ while the negatively stacked case requires 60.4 $s$. This substantial difference in endpoint time can be intuitively explained. First, the areas with higher pattern-density polish slower, and thus the initial topography before polishing the second layer is positively correlated with the first layer pattern-density. Second, we know that the area with higher initial topography will experience a higher pad pressure, which leads to a higher removal rate in these areas. Third, in the positively stacked case, the second layer pattern-density positively correlates with the initial topography. The higher removal rate due to the higher initial topography enhances the removal rate of all areas which also thus have a higher second-level pattern-density. The endpoint is mainly affected by the area with the slowest polishing rate, and thus, the polishing time to reach endpoint is shortened in the positively stacked case. For the negatively stacked case, the polishing of the areas with higher

second-level pattern-density is slowed down due to the negative correlation between the second-layer pattern-density and the initial topography. As a result, the polishing time required to reach endpoint is lengthened.

Figure 3-50 shows the raised area topography at the endpoint for both cases, where we see that the positively stacked case has much larger topography variation. For both cases, the initial topography has a peak-to-valley difference of 2699 Å. In the case of positively stacked first and second-level layouts, the pattern induced topography variations from the two layers are positively correlated and add to each other, and the peak-to-valley difference has increased to 4301 Å. On the other hand, in the case of negatively stacked layouts, the pattern induced variations are negatively correlated, and peak-to-valley difference in the final up area thicknesses has decreased to 1380 Å.

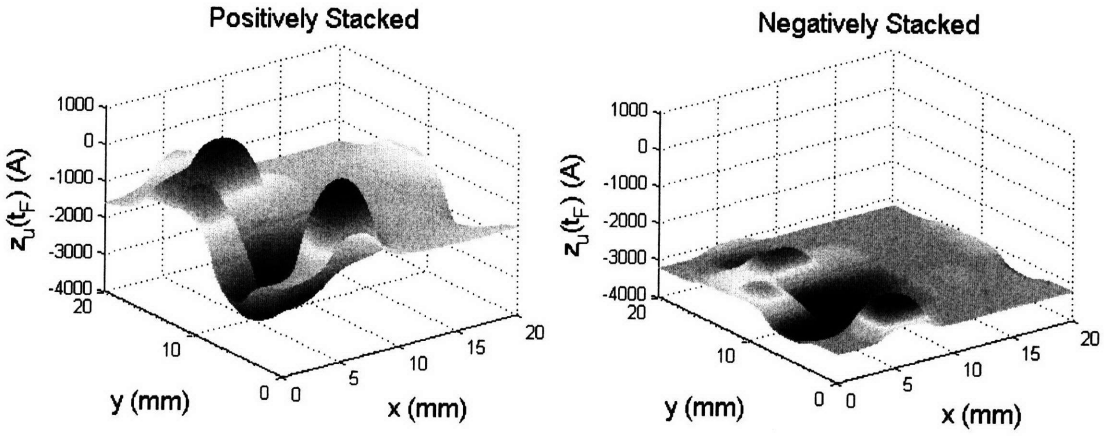



Figure 3-50: The raised area topography at the endpoint for (a) positively stacked case, and (b) negatively stacked case.

In previous work [81], an assumption of approximately additive topography across multiple layers was made based on an expectation that each layer polishes until the wafer appears "flat" to the CMP pad. In the example above, we see that the layout generates a chip-scale peak-to-valley range of 2699 Å, whereas the second level polish for the same layout adds only 1602 Å of additional final global variation.

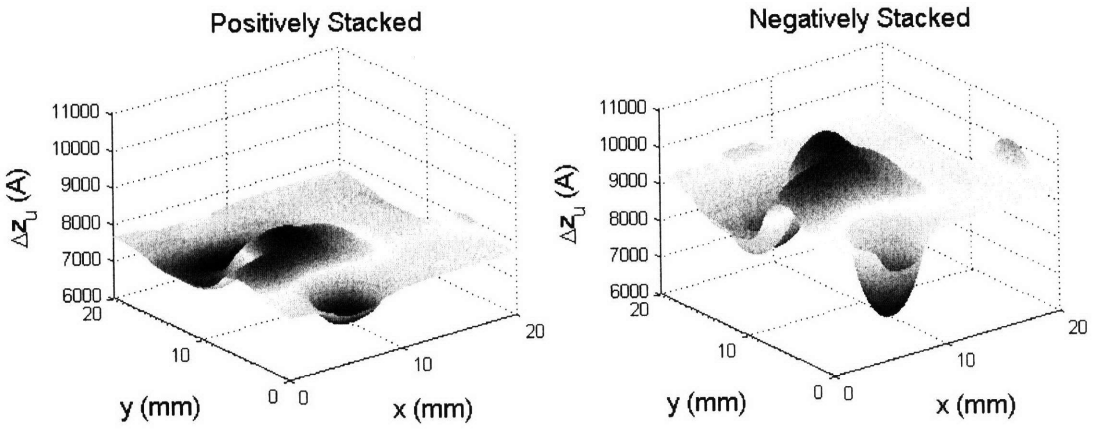



Figure 3-51: The amount removed on raised area at the endpoint for (a) positively stacked case, and (b) negatively stacked case.

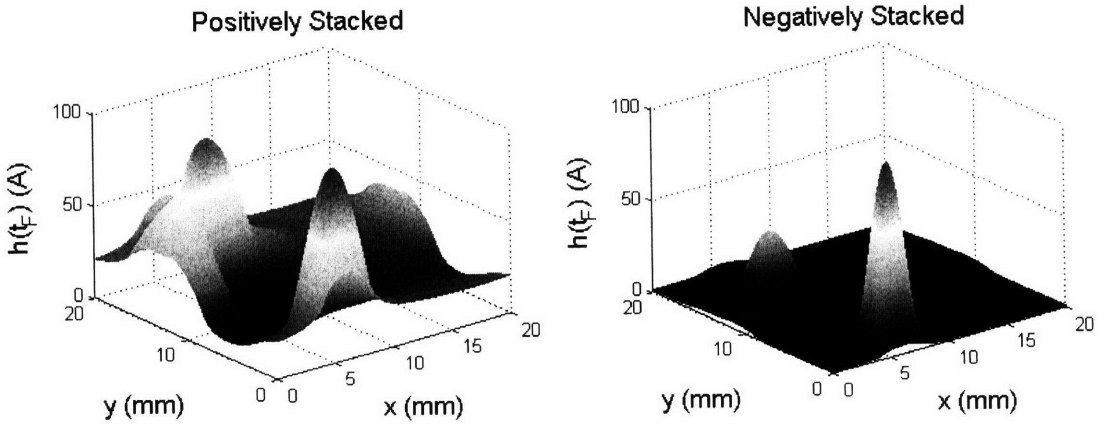



Figure 3-52: The step-height at the endpoint for (a) positively stacked case, and (b) negatively stacked case, for an oxide CMP process.

### The Impact of Nanotopography on STI

Nanotopography refers to the topography variation on a virgin wafer, which ranges from 10 nm to 100 $nm$ height difference over millimeter lateral distances. Its impact is illustrated by comparing two STI processes, where one has a flat initial surface and the other has initial nanotopography as plotted in Figure 3-53 (b). The STI test mask described in Section 3.6 is used for the simulation, and its pattern-density is shown in Figure 3-53 (a). The success of the STI process relies on the complete removal of overburden oxide in all raised areas, and thus the endpoint criteria is $max(z_u) = 0$.

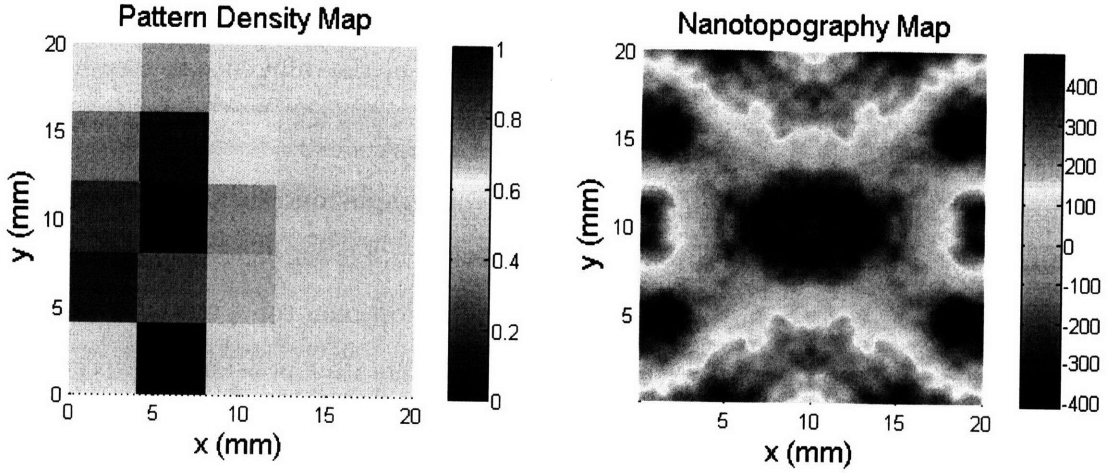



Figure 3-53: (a) Pattern-density map of the STI test mask. (b) The nanotopography map used in the comparison, where the vertical (height) are in Å.

The simulation shows that it takes 48.6 seconds to reach endpoint without nanotopography, and 50.1 seconds with nanotopography. Figure 3-54 (a) shows the difference of $z_u$ at the endpoint for the two cases, and Figure 3-54 (b) shows the difference in step-height. The additional variation caused by the nanotopography has a range about 400 Å, which can be significant in the STI process. Visually, the additional variations in $z_u$ and $h$ depend on both the density map and the nanotopography map.

### The Impact of Initial Topography on Multi-Level Metal Polishing

In multi-level copper interconnect, the topography variation resulted from the polishing of the metal level one leads to additional variation on metal level two [15],

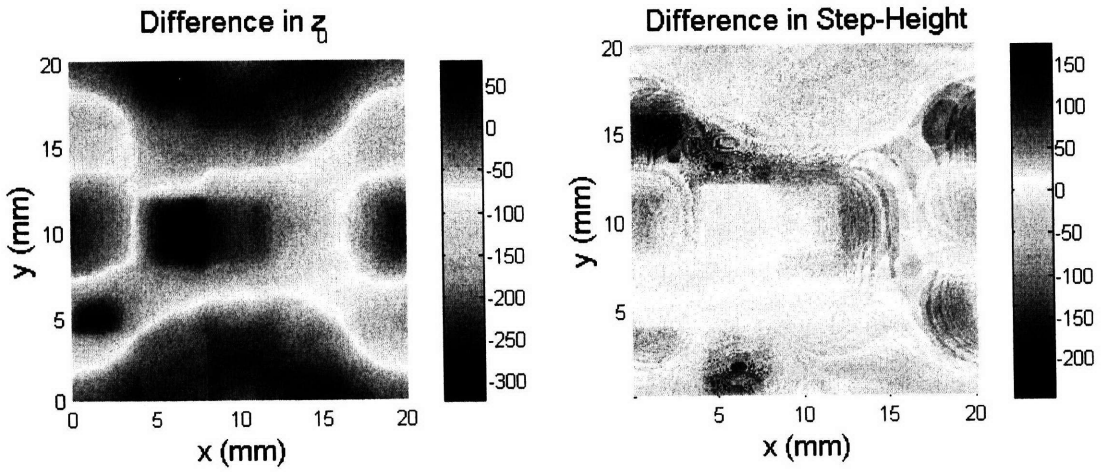



Figure 3-54: Comparison between the STI process with and without nanotopography. (a) The difference in $z_u$ due to nanotopography; and (b) the difference in step-height (both shown in Å units).

as illustrated in Figure 3-55. The accumulated topography and metal line thickness variation can lead to circuit performance degradation and lower yield. In this section, the physically-based model is used to illustrate the additional topography and metal line thickness variation caused by topography variation after polishing the previous layer.

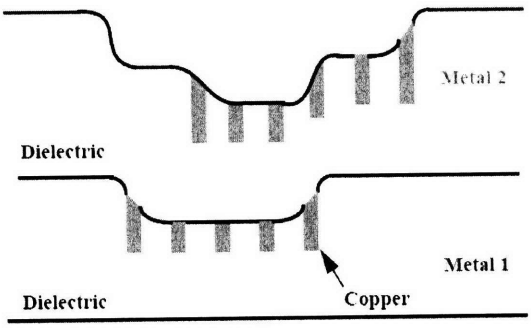



Figure 3-55: Cumulative non-uniformity effect [15]

A two-level copper interconnect structure is considered, where copper metal lines are formed in an oxide layer using a single damascene process with CMP. A simplified structure is assumed, without barrier metal layers or barrier metal CMP steps. We denote the pattern-density of the first layer as $\rho_1$ and that of the second layer as $\rho_2$. We consider two extreme cases with the same second layer pattern-density $\rho_2 = \rho$,

but in one case the pattern densities of the two layers are positively stacked $\rho_1 = \rho$ and in the other case they are negatively stacked such that $\rho_1 = 1 - \rho$. The STI test mask is used for the simulation, and its pattern density map is shown in Figure 3-53 (a).

For this example, the physically-based CMP model is applied to copper/oxide polishing, rather than oxide/nitride polishing as in the STI process. While additional effects in copper CMP may be present and need to be considered in a full copper CMP model [82], here we use the physically-based CMP models as a first order approximation, but one which is able to account for both long-range pad bending and pad-asperity driven dishing effects, to study multi-level topography concerns in the copper CMP case. The model parameters and initial structure parameters used are blanket removal rate of 3000 Å/*min*, characteristic asperity height of 500 Å, applied pressure of 3 *psi*, Young's modulus of 400 *MPa*, selectivity $s$ of 10, initial step-height equal to 5000 Å, initial copper thickness on raised area equal to 6000 Å, and initial copper thickness in the trench area of 5500 Å. The parameters are not extracted from any copper CMP process, and are chosen for illustration purposes only. Figure 3-56 shows the surface topography after polishing the first layer. The example is similar to the single material oxide case considered earlier in this section, where the key differences are due to the the selectivity of 10 used in the copper example here.

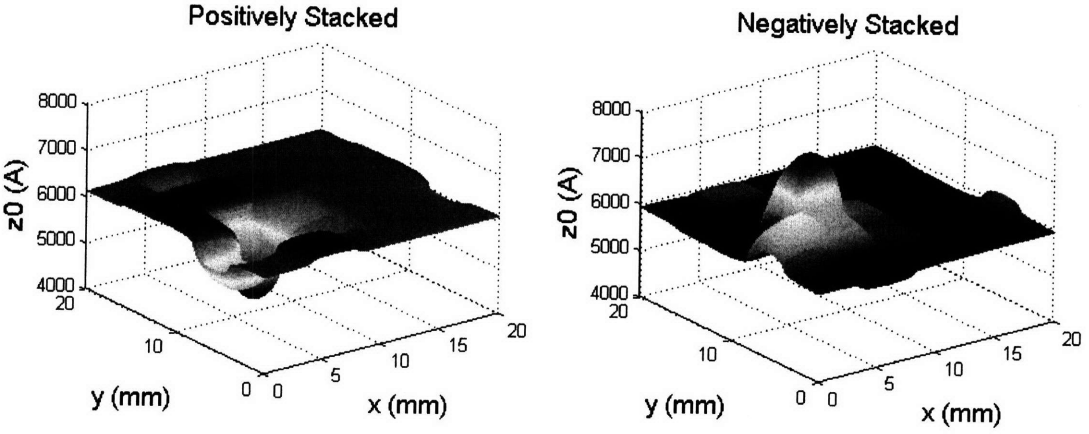



Figure 3-56: The topography after polishing the first copper interconnect level of (a) the positively stacked case and (b) the negatively stacked case.

In polishing the second layer, the simulations show that it takes 97.4 seconds to

reach the endpoint in the positively stacked case and 91.1 seconds in the negatively stacked case. At the endpoint for polishing the second layer, the surface topography of the raised areas is shown in Figure 3-57, and the remaining copper interconnect thickness is shown in Figure 3-58. In the positively stacked case, the pattern-induced topography accumulates, but the remaining copper thickness has less variation than that of the negatively stacked case. In the negatively stacked case, the pattern-induced variations from the two levels partially cancel each other and result in less total topography height variation; however, a large copper thickness variation is observed in the second layer.

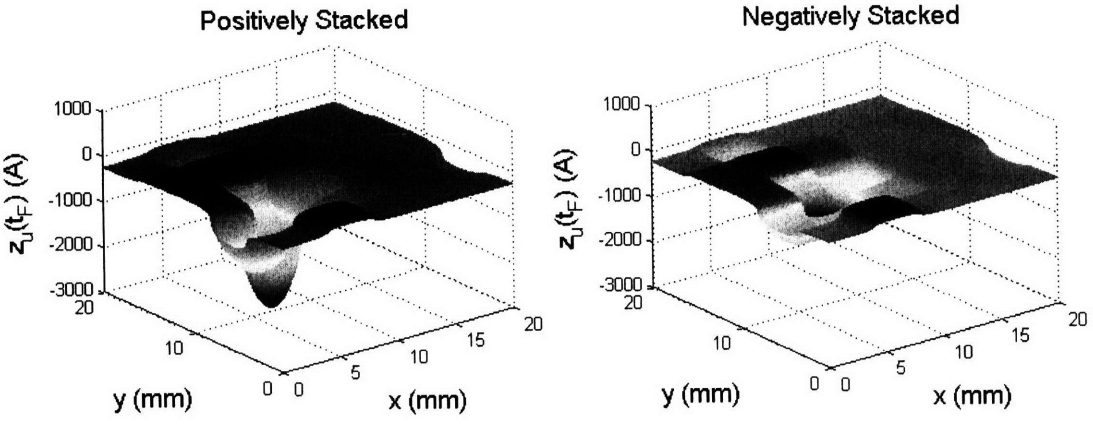



Figure 3-57: The surface topography of the raised areas after polishing the second copper interconnect level of (a) the positively stacked case and (b) the negatively stacked case.

The study illustrates the potential surface topography and metal thickness impact in multi-level copper CMP. The pattern-induced topography affects the polishing time to reach endpoint in the polishing of later layers, and causes additional topography variation and copper thickness variation. In the design of a multi-level interconnect layout, there exists a tradeoff between topography variation and copper thickness variation. The accumulated topography variation and the tradeoff make the layout design a challenging task, and physically-based CMP models are a valuable tool to optimize layout design.

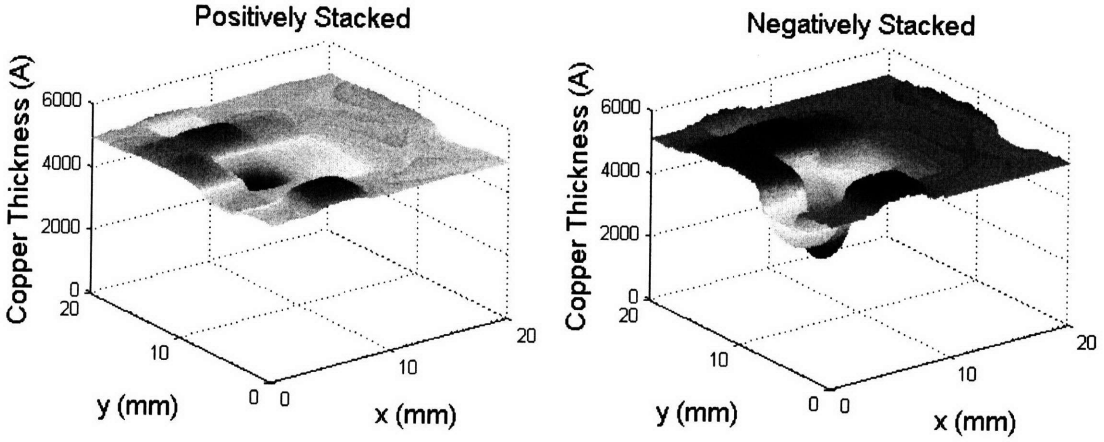



Figure 3-58: The remaining thickness of copper interconnect after polishing the second copper interconnect level of (a) the positively stacked case and (b) the negatively stacked case.

### 3.5.3 The Pad Properties

This section studies how the planarization performance is affected by pad properties, including the Young's modulus of the bulk and the characteristic asperity height.

#### The Effect of Pad Bulk Young's Modulus

Two simulations of oxide polishing are conducted with different values of Young's modulus: 200 $MPa$ and 400 $MPa$. The the other model and process parameters are $K = 3000$ Å, $P_0 = 3$ $psi$ and $\lambda = 500$ Å. To reach the endpoint condition of $h < 100$ Å, the polishing step with $E = 200$ $MPa$ needs 116 seconds, while that with $E = 400$ $MPa$ needs 104 seconds. The raised area topography and step-height at the endpoint are plotted in Figure 3-59. The step-height distribution is nearly the same for both cases, but a higher value of $E$ shows better global planarization performance, as seen in Figure 3-59.

Figure 3-61 shows the evolution of material removal in raised and trench areas, and Figure 3-62 (a) shows the evolution of step-height for regions with different pattern-densities of 10%, 50%, and 90%. The simulations using the two different values of $E$ show similar trends, but the case of $E = 400$ $MPa$ shows lesser difference between 10% and 90% regions, which indicates better global planarization. This can also be seen in Figure 3-62 (b), where the larger value of $E$ results in a smaller value of

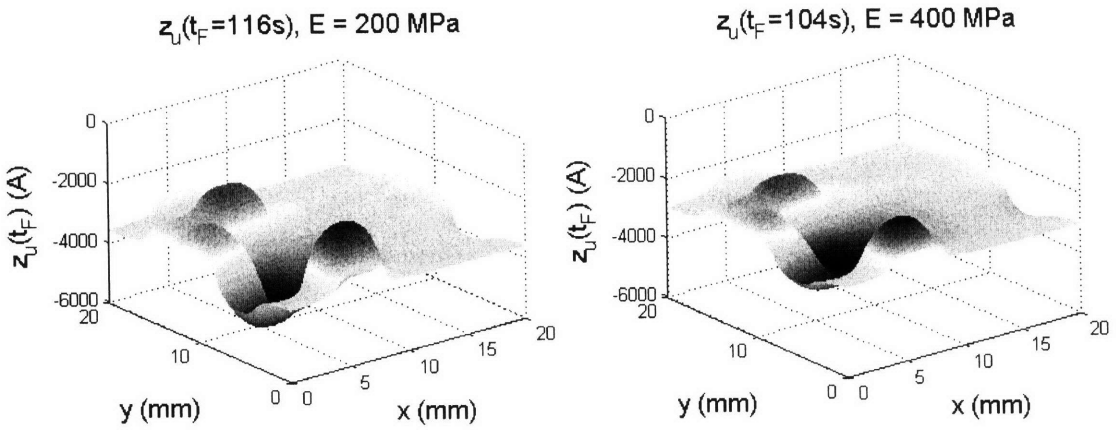



Figure 3-59: Plot of raised area topography at the endpoint for (a) 200 $MPa$, and (b) 400 $MPa$.

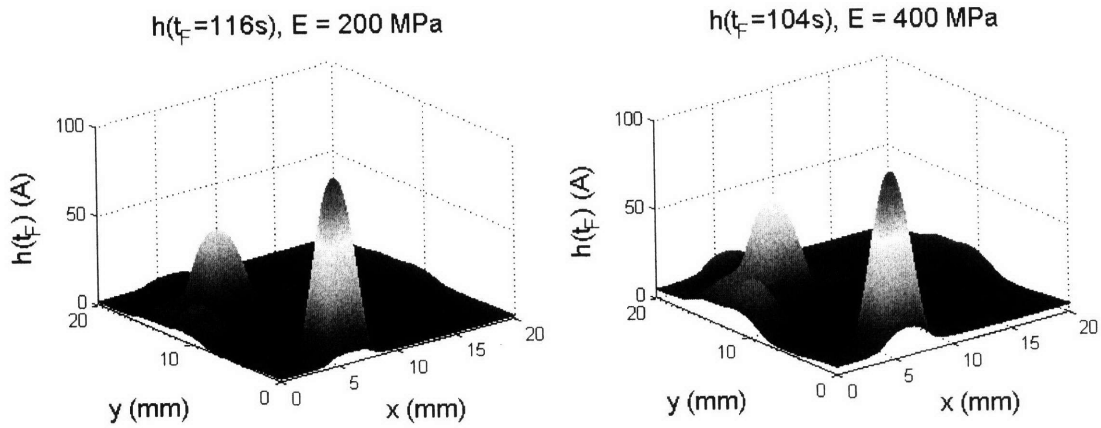



Figure 3-60: Plot of step-height at the endpoint for (a) 200 $MPa$, and (b) 400 $MPa$.

$Range(z_u)$.

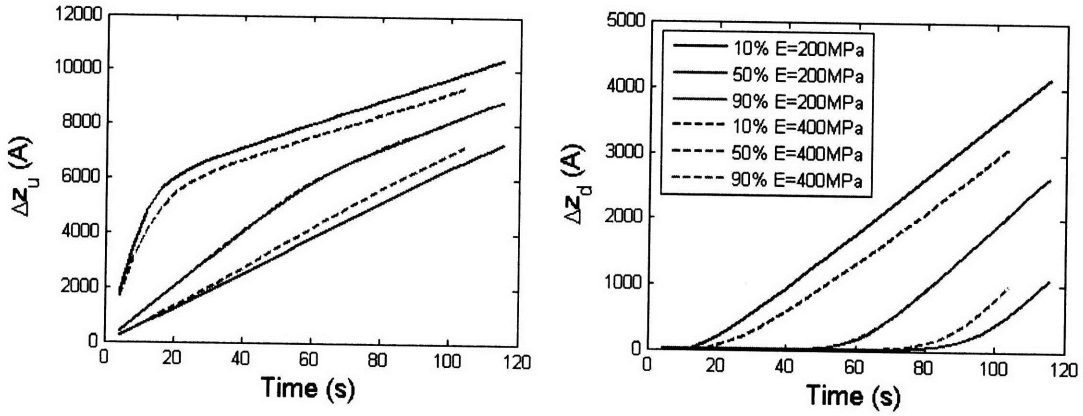



Figure 3-61: The material amount removed in (a) raised areas, and (b) trench areas, for regions with local pattern-densities of 10%, 50%, and 90%. The simulations with $E = 200\ MPa$ are shown as solid lines, and those with $E = 400\ MPa$ are shown as dashed lines.

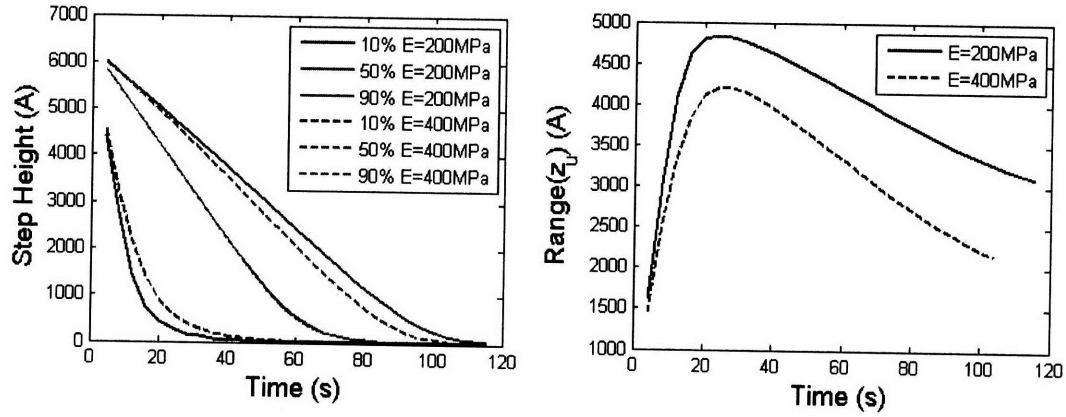



Figure 3-62: (a) Step-height evolution for regions with local pattern-densities of 10%, 50%, and 90%. (b) $Range(z_u)$, which is defined as $z_u(90\%) - z_u(10\%)$. The simulations with $E = 200\ MPa$ are shown as solid lines, and those with $E = 400\ MPa$ are shown as dashed lines.

### The Effect of Asperity Distribution

To study the effect and sensitivity to pad asperity distribution, two simulations of oxide polishing are conducted with different values of characteristic asperity height: $\lambda = 400$ Å and $\lambda = 800$ Å. The other parameters are $K = 3000$ Å, $P_0 = 3\ psi$ and $E = 300\ MPa$. To reach the endpoint criteria of $h < 100$ Å, the polishing with $\lambda = 400$ Å

needs 105 seconds, while that with $\lambda = 400$ Å needs 123 seconds. Thus, a smaller value of $\lambda$ leads to faster step-height reduction. The raised area topography and step-height at the endpoint are plotted in Figure 3-63. The step-height distribution is nearly the same for both cases; however, even with longer polishing time, the case of $\lambda = 800$ Å has larger residual step-height variation, as shown in Figure 3-64.

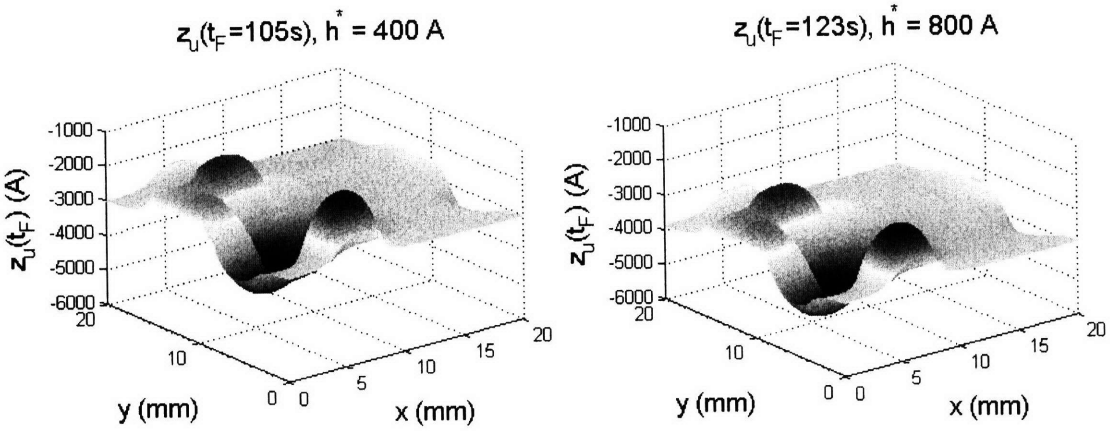



Figure 3-63: Raised area topography at the endpoint for (a) $\lambda = 400$ Å, and (b) $\lambda = 800$ Å.

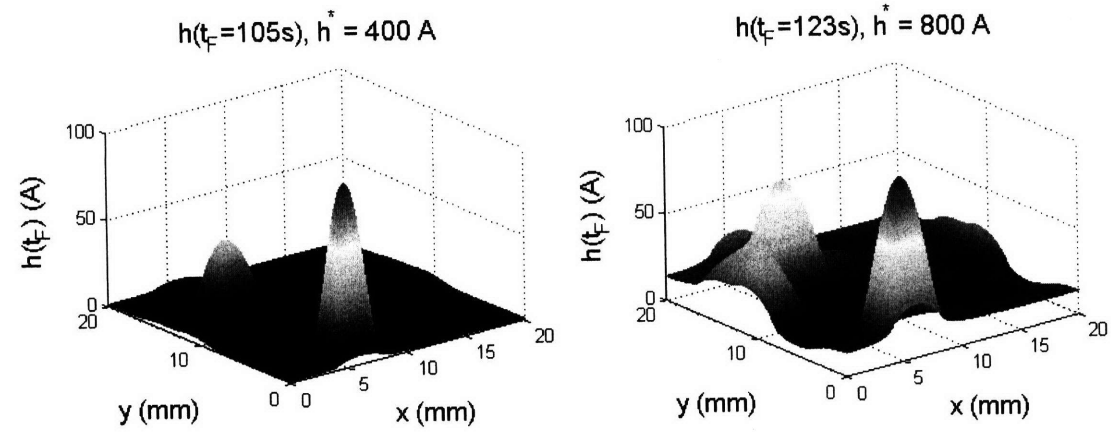



Figure 3-64: Step-height at the endpoint for (a) $\lambda = 400$ Å, and (b) $\lambda = 800$ Å.

Figure 3-65 shows the evolution of material removal in raised and trench areas. The plots of $\Delta z_u$ differ in the timings of the transition from larger slope to smaller slope, and the plots of $\Delta z_d$ differ also in the timing of change of slopes. The difference is more clearly seen in Figure 3-66 (a), which shows the evolution of step-height for regions with different pattern-densities 10%, 50%, and 90%,where we see that the

smaller value of $\lambda = 400$ Å results in faster step-height reduction. The value of $\lambda$ has little impact on the post-CMP global topography variation, as observed in Figure 3-66 (b).

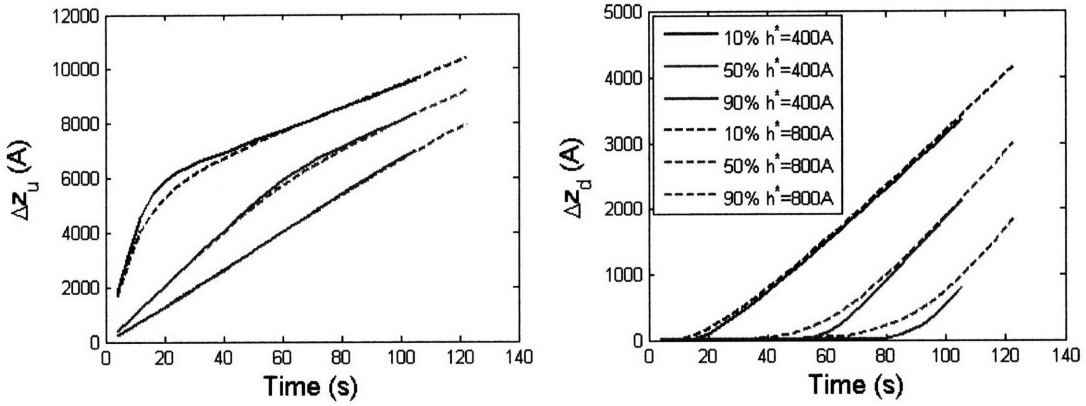



Figure 3-65: The material amount removed on (a) raised areas, and (b) trench areas for regions with local pattern-densities of 10%, 50%, and 90%. The simulations with $\lambda = 400$ Å are shown as solid lines, and those with $\lambda = 800$ Å are shown as dashed lines.

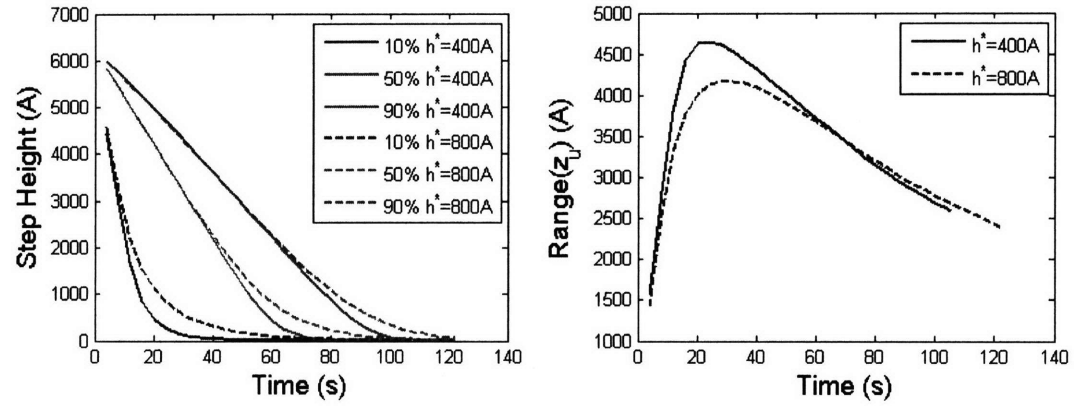



Figure 3-66: (a) Step-height evolution for regions with local pattern-densities of 10%, 50%, and 90%. (b) $Range(z_u)$, defined as $z_u(90\%) - z_u(10\%)$. The simulations with $\lambda = 400$ Å are shown as solid lines, and those with $\lambda = 800$ Å are shown as dashed lines.

### 3.5.4 The Effect of Applied Pressure on Planarization

In the physically-based model, the contact pressure is proportional to the applied pressure as long as the ratio of $E$ to $P_0$ is fixed. In other words, if the applied pressure

is set to $P_0'$, the resulting CMP process is equivalent to a CMP process using applied pressure $P_0$, effective Young's modulus $(P_0/P_0')E$ and removal rate $(P_0'/P_0)K_0$. For example, if the applied pressure is doubled, it is equivalent to a CMP process having a doubled blanket removal rate and a softer pad having half the Young's modulus, which leads to worse planarization performance as discussed earlier.

The contribution of applied pressure to removal rate is well understood, as summarized by the Preston equation; however, its contribution to planarization has only been studied empirically. Lee [1] studied oxide polishing using different pressures and relative velocities and used the experimental data to fit the PDSH model. The relationship between the extracted planarization length and applied pressure is shown in Figure 3-67 [1]. The extracted planarization length decreases with applied pressure, because a higher applied pressure is equivalent to a smaller Young's modulus $E$ and the planarization length is positively correlated with $E$.

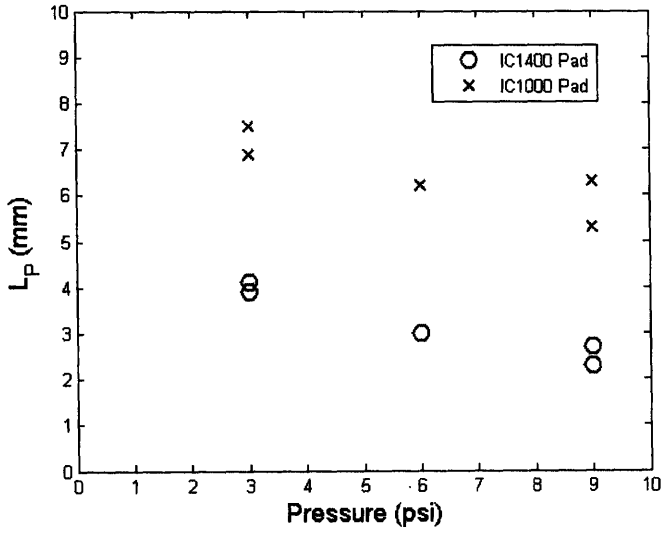



Figure 3-67: The relationship between the extracted planarization length $L_P$ and the applied pressure. (Data source [1].)

## 3.6 Verifying Die Level CMP Models

A typical method to verify a die-level CMP model is to compare model predictions with experimental results using test wafers which are specially designed to have var-

ious pattern structures on them to calibrate the CMP models. A set of test wafers are polished for different amounts of time, and the film thicknesses at various locations of the chip are measured before and after the CMP process. From the thickness data, the removal amount values can be calculated and used to calibrate both the physically-based and pattern-density CMP models. Once the physically-based die-level model is calibrated, it can be used to verify the assumptions and predictions of the approximate pattern-density step-height models.

### 3.6.1 Experimental Setup

The experimental data from an STI CMP process are summarized here. The experiment involves polishing five 200 *mm* wafers patterned with the MIT STI mask, whose pattern-density map is shown in Figure 3-68. All wafers begin with 90 Å of a thermally grown pad oxide on a p-type silicon substrate, followed by a 1190 Å silicon nitride deposition. Wafers are patterned and etched to obtain a trench depth of 5000 Å. Wafers are then subjected to a sidewall oxide layer growth of 250 Å using dry oxidation. This is followed by chemical vapor deposition of TEOS oxide for trench fill of 5750 Å. The five wafers are then subjected to the CMP process with polishing time splits of 5, 10, 15, 30, and 40 seconds. The CMP process uses a SpeedFam 5-head polisher with a down force of 7.4 psi. We use a Rodel IC-1400 pad, and Cabot SS-25 slurry. Before and after the CMP process, film thickness at various locations on the wafers, which are marked with red 'x' in the figure, are measured by a KLA-Tencor ASET F5 system.

### 3.6.2 Verifying Die-level Models with Experimental Data

The physically-based CMP model has four parameters: the blanket removal rate $K$, the effective Young's modulus of the polishing pad $E$, the characteristic asperity height $\lambda$, and the selectivity of nitride to oxide $s$. The model parameters are estimated by minimizing the sum of squared differences between experimental data and model predictions. The best fit is obtained with parameters $K = 2950$ Å$/min$, $E = 80\ MPa$, $\lambda = 1000$ Å, and $s = 4$. The mean square error of raised area data is 255 Å, and

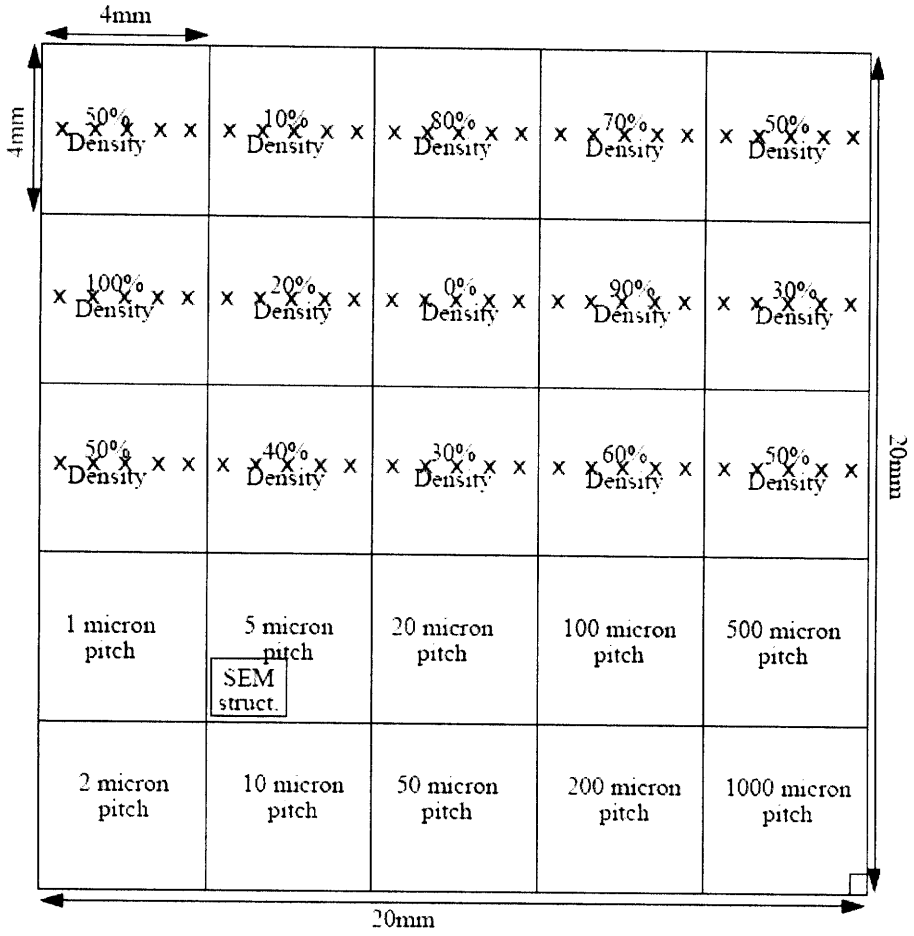



Figure 3-68: Pattern-density map of MIT STI test mask with measurement sites marked with red 'x'.

that of trench area data is 192 Å. The model predictions and experimental data are shown in Figure 3-69 and 3-70.

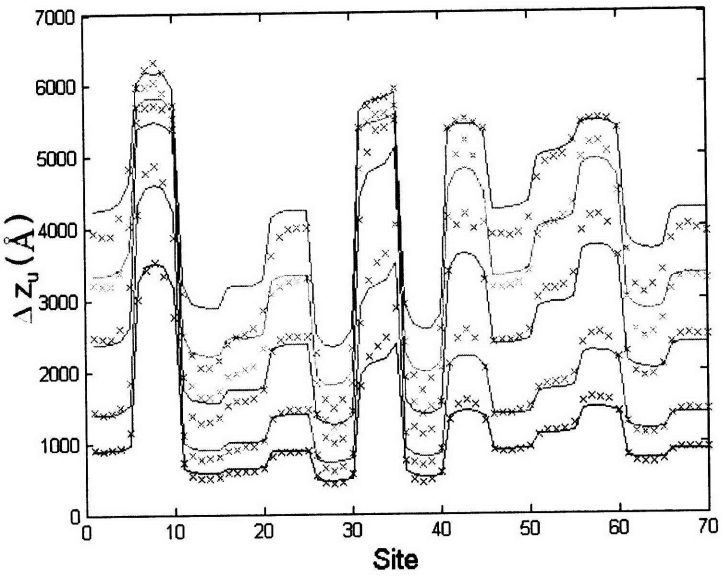



Figure 3-69: Fit of experimental data with the physically-based CMP model, where the amount removal in raised area is shown. Experimental data are plotted as 'x' and model predictions are plotted as lines.

The experimental data is also used to fit the exponential pattern-density step-height (PDSH) model. In the analysis, the data are separated into two groups: those where one is polishing only oxide and those where one is polishing dual materials. The model prediction and experimental data for the raised areas are shown in Figure 3-71, and those for the trench areas are shown in Figure 3-72. The model parameters to fit the data for oxide polishing are planarization length 1.01 $mm$, blanket removal rate 2744 Å/$min$, and characteristic step height 246.7 Å. The parameters to fit the dual material polishing data are planarization length 3.14 $mm$, selectivity 4.57, and characteristic step height 265.2 Å. The extracted blanket removal rate and selectivity are close to their measured values on blanket wafers, and the extracted values of characteristic height from both groups are close to each other. These observations are consistent with model assumptions; however, the extracted value of planarization length in the dual material polishing stage is nearly three times larger than that for oxide polishing.

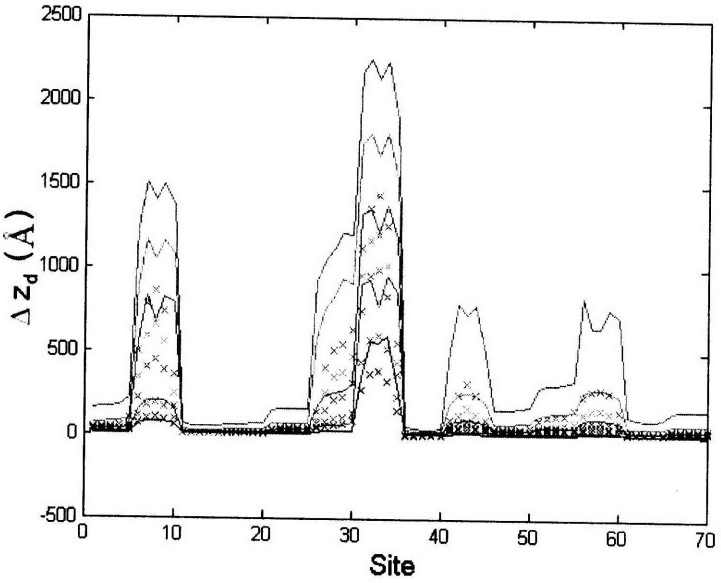



Figure 3-70: Fit of experimental data with the physically-based CMP model, where the amount removal in trench area is shown. Experiment data are plotted as 'x' and model predictions are plotted as lines.

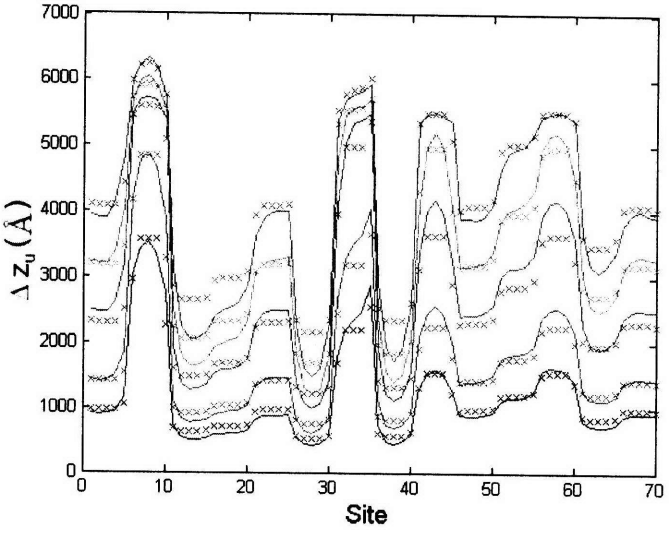



Figure 3-71: Fit of experimental data with the exponential PDSH CMP model. The amount removed in the raised areas is shown, and the fitting error is 215 Å.

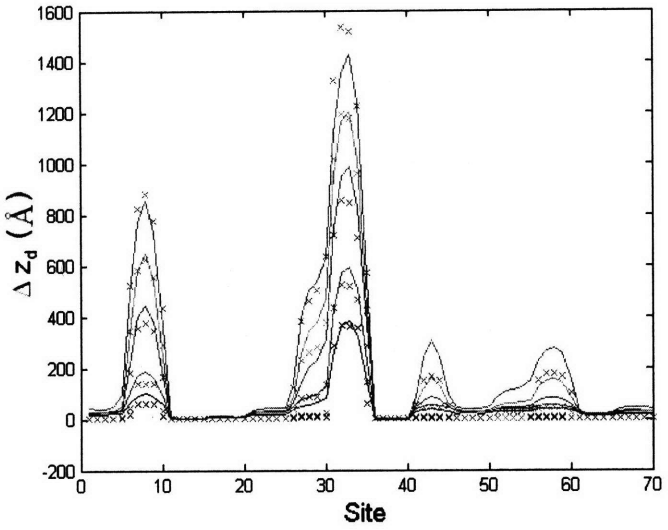



Figure 3-72: Fit of experimental data with the exponential PDSH CMP model. The amount removed in the trench areas is shown, and the fitting error is 47 Å.

### 3.6.3 Computational Requirements

The primary goal of the physically-based CMP model is to capture the effect of pad bulk and asperity properties on feature and die-level planarization. The PDSH models, in contrast, are approximations meant to enable rapid simulation of full-chip performance. Here, we briefly describe the typical computational requirements for fitting or extracting models based on experimental data, followed by those needed for simulating the evolution of surface topography.

A typical chip may be as large as 25 $mm \times 25$ $mm$ in size. Memory requirements for simulation using the PDSH model are storage for three arrays containing $\rho(x, y)$, $z_u(x, y)$, and $z_d(x, y)$ at a typical discretization of 40 $\mu m \times 40$ $\mu m$, or about 625 by 625 elements for each array. These memory demands are modest, and readily available on desktop PC's. The computation of surface evolution for a 60 second CMP process, using one second time steps, typically requires about a fraction of a second on a Pentium$^{TM}$ with 2 Gb main memory for the PDSH models. In the case of the physically based model, memory requirements are similar. However, the contact wear calculation requires additional iteratively solving of coupled matrices (as summarized in Equation 3.16), and the computational demand depends on the

error tolerance of the iterative method. The calculation of a 60 second time evolution, which uses one second time step and the relative error tolerance of $10^{-4}$, can require 30 to 200 minutes.

The fitting of either PDSH or the physically based model is more computationally demanding. Given data for up and down area thicknesses for multiple time splits, the simulation is performed for a guess of the model parameters, and an error minimization optimization loop is used to find the best fit set of model parameter values. The required computation time increases dramatically with the number of model parameters. For the PDSH model having three model parameters, a model extraction typically may require 0.5 to 2 minutes on our example desktop PC. The physically based CMP model having four parameters requires substantially more time for model parameter extraction, typically several days or longer.

## 3.7 Summary

In this chapter, two die-level CMP models are proposed with the same framework that decomposes the problem into the dependence of removal rate on pressure and the dependence of pressure on layout topography, i.e., $\frac{dz}{dt}(\rho, z_u, h) = \frac{dz}{dt}(P(\rho, z_u, h))$. A physically based model is proposed by explicitly modeling the pad structure including the bulk modulus and surface asperity structures. A semi-empirical exponential PDSH model improves upon the previous PDSH models by adopting more reasonable assumptions.

The exponential PDSH model can be solved analytically to estimate the wafer surface evolution in most cases, and even without a closed-form solution, the time-stepping simulation usually takes only a few seconds on a typical desktop computer. The exponential PDSH model is applied to simulate the polishing of single material or dual material systems using either conventional or non-conventional slurry. The non-conventional slurries show advantages in faster step-height reduction and more uniform post-CMP topography than the conventional slurries. An analysis suggests that a non-conventional slurry with low sensitivity to pressure at both low and high pressure values has improved planarization performance.

The physically based CMP model is first compared with the exponential PDSH model, and then applied to explore a number of applications beyond the capability of the PDSH models. The comparison shows good consistency between the predictions of the two models, and suggests that the planarization length $L_P$ depends on the Young's modulus of the pad $E$ and that the characteristic step-height $h^*$ depends on the characteristic asperity height $\lambda$. The study of the impact of initial topography shows that it affects the polishing time to reach endpoint, which makes the process control more difficult; and that it causes additional topography variation and in the case of copper CMP, additional variation in copper interconnect thickness. The physically based model suggests that a larger Young's modulus $E$ reduces with-in-die non-uniformity and a smaller characteristic asperity height improves step-height reduction. And finally, the physically based model implies that an increase (or decrease) of the applied pressure is equivalent to a decrease (or increase) of Young's modulus, which affects the with-in-die non-uniformity.

To verify the models, the experimental data of STI CMP process are used and both models yield good agreement with the data. With more model parameters, the exponential PDSH model shows smaller fitting error, but the physically based model is able to capture the details of surface profile.

# Chapter 4

# Applications of Die-Level CMP Models

Die-level CMP models play an important role in practice. A die-level CMP model can be calibrated using only a small set of experiments, and for the same or similar process conditions, the model is able to predict film thickness and step-height at any time during the CMP process. This information can be applied in many ways, such as in DFM (design for manufacturing), CMP process control, evaluation of CMP consumables, and optimization of the process. Compared with exploring these questions experimentally, the die-level models have an advantage in reducing development time and cost, improving manufacturing yield, and reducing the environmental impact of the process, all with more rapid feedback through simulation.

In this chapter, die-level CMP models are applied to a number of applications. Section 4.1 covers the methodology of calibrating and applying the model, test wafer design, and its application in DFM (design for manufacturing). Section 4.2 studies the nanotopography impact on CMP; the variation caused by initial topography on the starting wafer is analyzed. Section 4.3 investigates the cause of the roll-off profile at the wafer edge, and how to minimize its effect by properly configuring the tool setup. Finally, Section 4.4, a die-level model is used to understand the mechanism of motor current end point detection as used in STI CMP.

## 4.1 Die-Level CMP Model Methodology

For a given CMP process, the die-level CMP model captures the way the CMP process responds to different pattern structures. Thus, once the model parameters are properly estimated, the calibrated model is able to simulate the surface topography across the chip during CMP. A CMP modeling methodology proposed by Ouma [16] is illustrated in Figure 4-1. The methodology consists of the following four steps.

- Design of the experiment: test wafer patterns are designed to have a sufficient set of layout patterns to enable study the full range of pattern effects in the process.

- Conduct of the experiment: test wafers are polished by the specific process which needs to calibrated. Thickness of dielectric layers or local step-heights are measured before and after the CMP process. Experiments with different polishing time splits are performed to capture the dynamic evolution of the dielectric layer thickness and step heights.

- Data analysis: the experiment data are fed to the CMP model, and a set of model parameters are chosen to minimize the error between model prediction and experiment result.

- Model simulation: a calibrated model is then capable of predicting the result of polishing any product chip layout using the calibrated process.

In this section, first we illustrate the methodology flow, and then focus on two topics: a new test wafer design for the STI CMP process, and applying the models to design for manufacturing.

### 4.1.1 Illustration of the Die-Level Model Methodology

Here we discuss an example in which the die-level model methodology is applied to a product chip layout. The experimental data in this example are obtained via collaboration with National Semiconductor. The calibration experiment uses six test

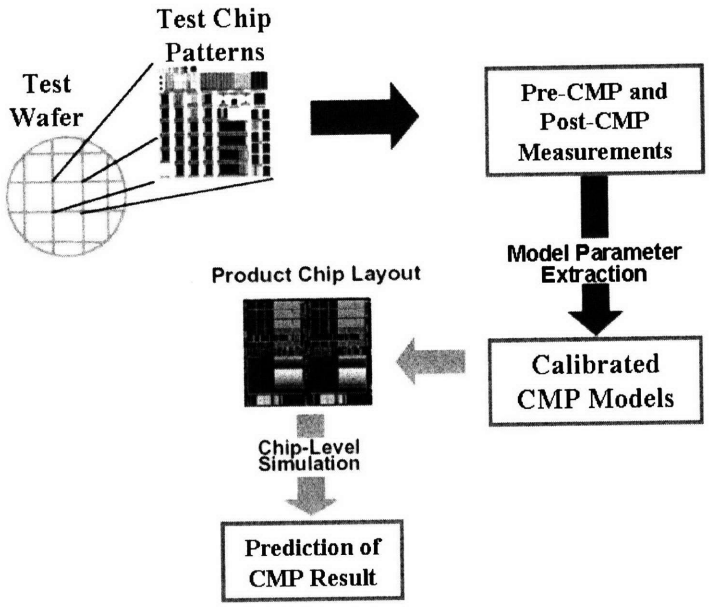



Figure 4-1: CMP modeling methodology [16]

wafers patterned using the MIT STI mask, whose pattern density map is shown in Figure 4-2, and then polished for a time splits of 10, 15, 20, 30, 40, 50 seconds. Each wafer is measured to obtain the pre-CMP and post-CMP film thickness based on the measurement plan shown in Figure 3-68.

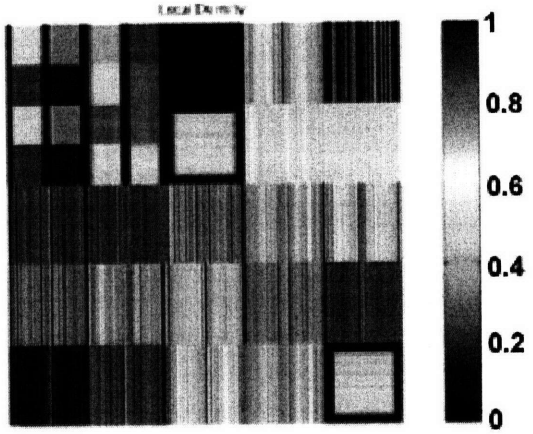



Figure 4-2: Pattern density map of MIT STI mask

The measured data are used to fit the die-level PDSH model, as shown in Figure 4-3, and the fitting error is 160 Å. The calibrated CMP model can then be used to

simulate the polishing of any product chip. For example, a product layout with the pattern-density map shown in Figure 4-4 is considered. The CMP model takes the pattern-density map as an input, and can be used to predict the result of polishing. Figure 4-5 (b) shows the predicted film thickness map after 60-second-polishing, and the result agrees well with the measurement film thickness, as shown in Figure 4-5 (a). The model can also predict other information about the CMP process which is difficult to measure directly, such as the oxide clearing time and nitride clearing time, as in Figure 4-6.

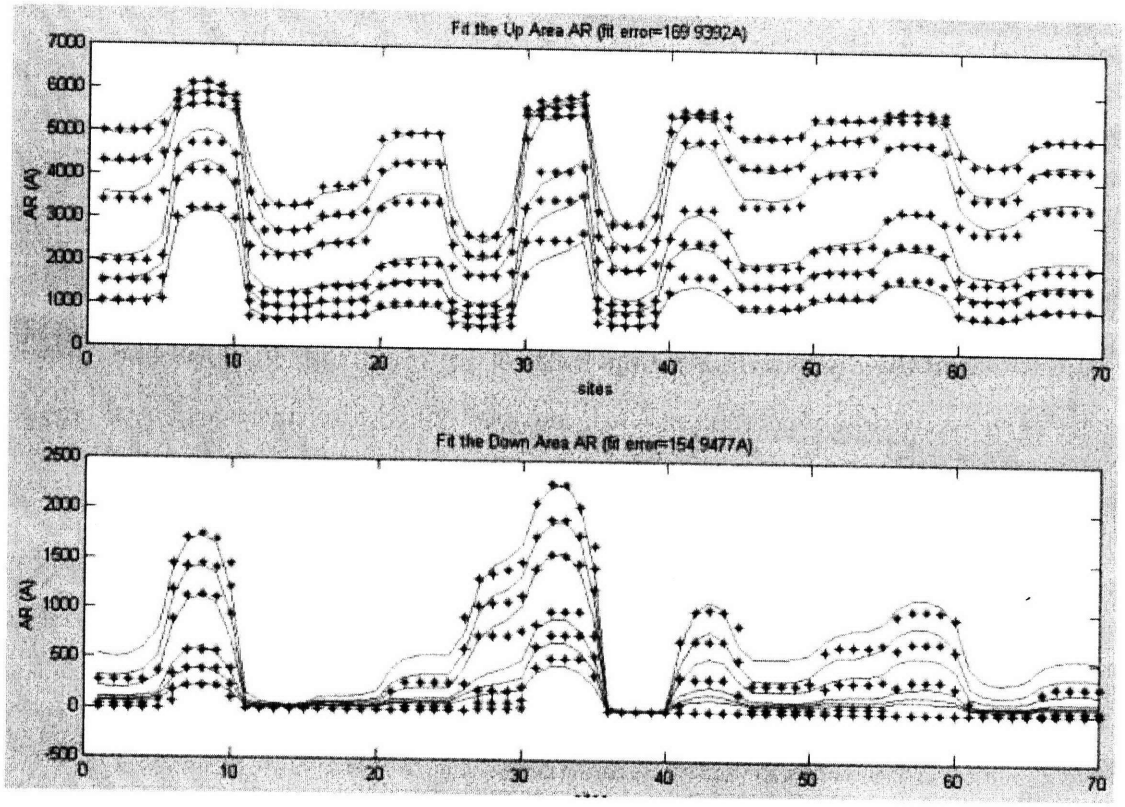



Figure 4-3: The comparison of the measured data (in lines) and the model predictions (in dots) of the amount removed in raised areas (top plot) and in trench areas (bottom plot).

### 4.1.2 New STI Test Mask

In the above methodology, the key underlying vehicle that enables characterization and modeling of the STI CMP process is a test mask. The test mask can be used for

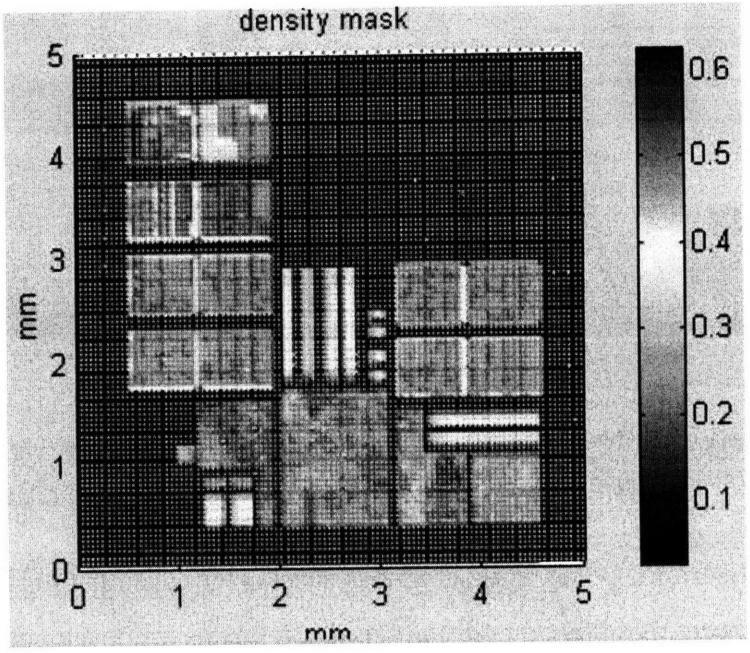



Figure 4-4: The pattern-density map of a product chip layout

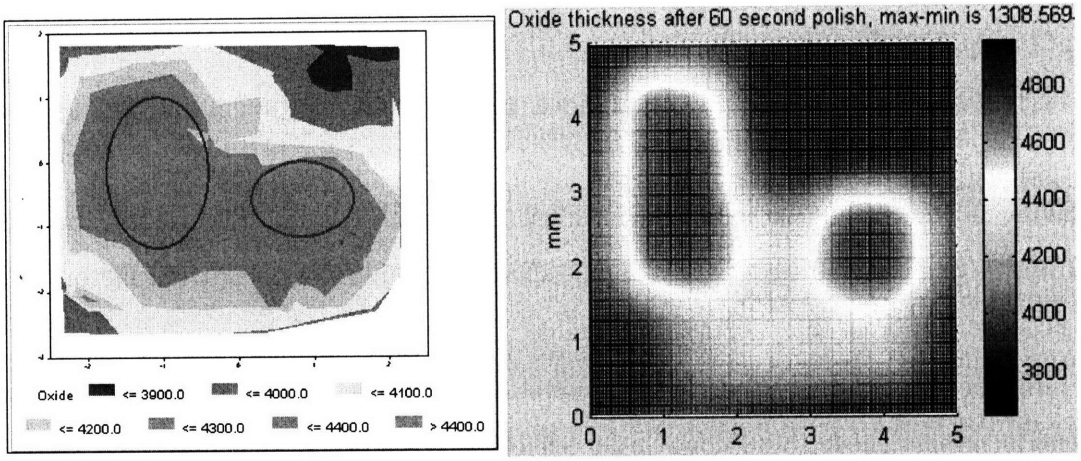



Figure 4-5: (left) The experimentally measured, and (right) the PDSH model predicted oxide film thickness after 60-second-polishing.

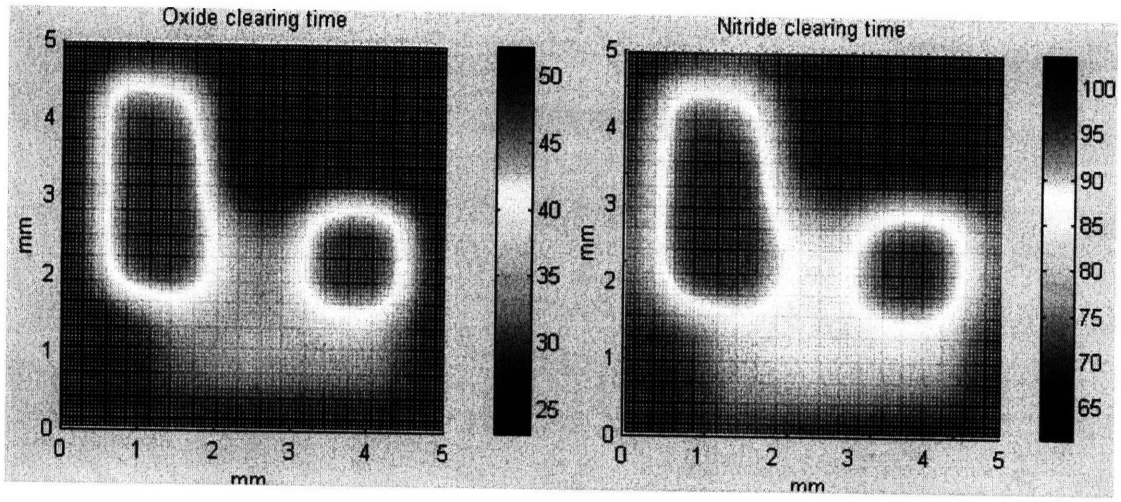



Figure 4-6: The predicted maps of (left) the oxide clearing time and (right) the nitride clearing time for a given product chip.

rapid characterization of CMP characteristics such as for consumable or tool comparisons. Furthermore, these test masks can serve as a means for developing and validating physical or semi-empirical models. The existing MIT "Comprehensive Dielectric Characterization Test Mask" has proven helpful in understanding the pattern-density dependence of STI CMP [1]. However, it has several limitations including large feature size, purely line array structures, and lack of a circuit-like region. The new STI characterization mask is designed to overcome these shortfalls and to support sub-90 nm technology advances.

The fundamental issue in designing a test mask is to understand what kinds of patterns actually matter in causing non-uniform topography during processing. Having a wide range of patterns does not necessarily mean that the key pattern problems can be identified. A test reticle should contain relevant pattern factors that a process of interest has dependency on.

The new STI mask is designed to understand the following five issues: the pattern-density dependence, the second order effects of post-CMP non-uniformity due to dishing and erosion, the edge-acceleration effects in small structures, the deposition bias in 2D structures, and how the modeling and calibration perform on realistic product layout. The new STI mask is designed according to these five goals, as shown in the floor plan of Figure 4-7(a). A pattern density map of the new STI mask

is shown in Figure 4-7(b), where the map is extracted using 40 $\mu m$ by 40 $\mu m$ cells.

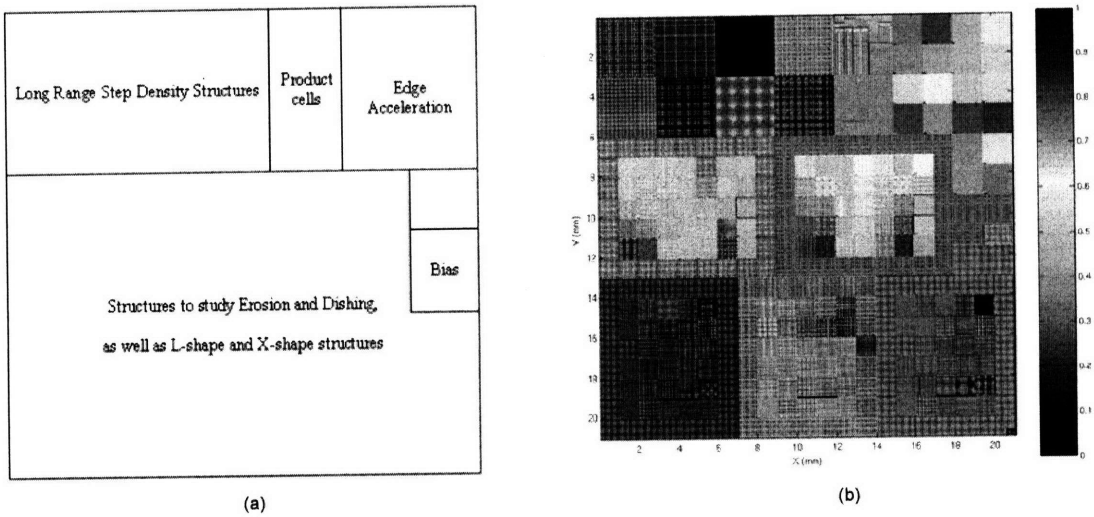



Figure 4-7: (a) Floor plan of the new STI mask, and (b) pattern-density map of new STI mask.

The first region is the pattern-density region. Previous research has shown that pattern-density is the dominant factor in causing die-level non-uniformity, and this region is dedicated to this long range dependency. The region is composed of eight 3 $mm$ $\times$3 $mm$ cells, where each cell has repeated rectangular structures. The pattern-density of cells ranges from 10% to 90%. The pattern-density cells are placed in a pseudo-random layout to achieve a good contrast of low and high densities as shown in Figure 4-8 (a). The rectangular structures are large, at about 100 $\mu m$, and are designed to be measured by optical metrology tools.

The second region is called the dishing and erosion region. This is the largest in area among the five, and is designed to study the second order effects on CMP of feature width and spacing. As pattern-density is the most significant factor, we have to keep the pattern-density constant while studying other factors contributing to dishing and erosion. Thus, the dishing and erosion region is divided into five blocks, and within each block it has the same pattern-density. We also add a 1 $mm$ buffer zone around this region to reduce the pattern-density interaction from the neighboring structures.

With structure parameters defined in Figure 4-8 (b), the local layout pattern-

density can be determined as

$$\rho = \frac{LW}{(L+S)(W+S)}. \tag{4.1}$$

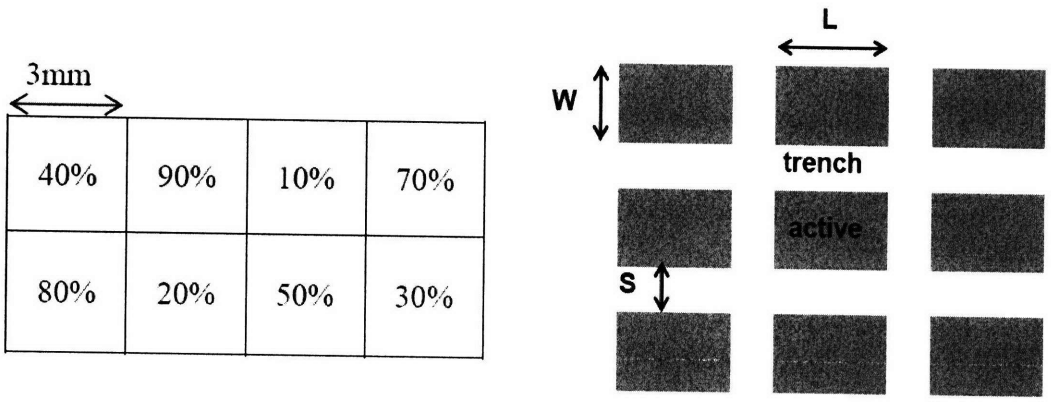



Figure 4-8: (a) Density configuration of pattern-density region. (b) Definition of structure parameters: line length ($L$), line width ($W$), and line space ($S$).

Thus, it is impossible to vary only one structure parameter while keeping the pattern-density constant in each block. To isolate the contributions of $L$, $W$, and $S$, we keep one variable constant, and vary the others. As a result, we further divide each block into seven region types: one for L-shape structures, one for X-shape structures, and five for keeping the following variables constant - $S$, $W$, $L$, area ($L \cdot W$), and ratio ($S : L : W$). Each region type has five variants.

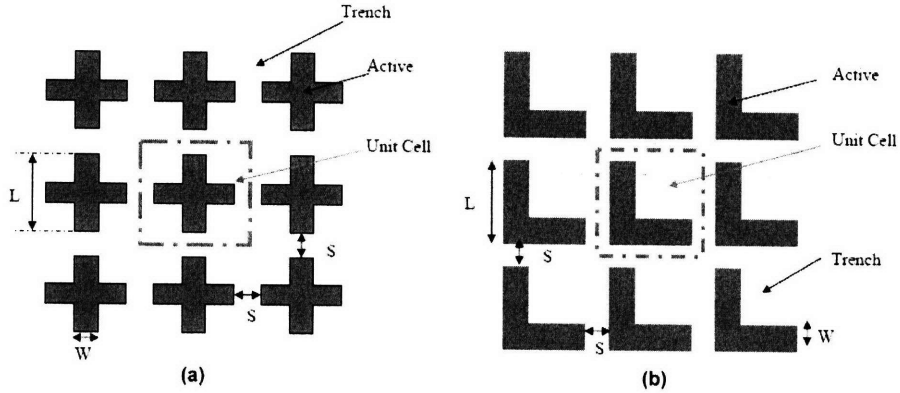



Figure 4-9: Diagrams of (a) X-shape and (b) L-shape structures.

The feature sizes in the region are from 5 $\mu m$ to 100 $\mu m$, with about half of them

designed to be measured optically and the other half having feature sizes less than 10 $\mu m$. Such a wide size range offers us the opportunity to compare the polishing of small structures to larger ones. As the sizes of some structures are less than 10 $\mu m$, the oxide pattern-density may differ from the local pattern-density due to deposition bias. In our mask, we try to keep the oxide pattern-density constant in the sub-region, so that all cells reach nitride at approximately the same time. Such a design provides a good opportunity to separate the oxide/nitride dual-material polishing from the initial oxide polishing stage.

The third region is the edge acceleration region. This section is designed to study the CMP process over structures of feature size from 0.5 $\mu m$ to 5 $\mu m$, and is intended to amplify the edge-acceleration effect in polishing small structures.

The fourth is the deposition bias region. As the feature size of the structures becomes smaller, the effect of deposition bias on pattern density becomes significant. For feature size less than 5 $\mu m$, the pattern densities of the oxide and nitride layers become quite different from that of as-drawn layout. A nitride deposition bias of $-0.07$ $\mu m$ and an oxide deposition bias of $-0.25$ $\mu m$ are used to carry out the calculation, based on characterization of a representative process at National Semiconductor. The bias structures are also rectangular in shape, which enables us to study the deposition bias in 2D structures.

The last region consists of product layout blocks. No matter how well the fitting of the test structures are, the direct verification of the CMP modeling comes from measurement on actual product chips. Hence, we leave a 3 $mm \times$ 6 $mm$ region for a partial product chip. In the current design, we have placed a partial logic chip layout and a partial memory chip layout. If product structures are too small, some measurement pads can be added within the product layout regions for optical measurement.

In this section, the design of the new STI mask have been reviewed, which serves two purposes. On one hand, it is an illustration of CMP mask design; while on the other hand, it introduces the new STI mask to interested readers. For other details on the new STI mask can be found in [83], and experiments, as well as data analysis,

using the new STI mask can be found in [84].

### 4.1.3 Design for Manufacturing

The flow of traditional manufacturing usually has only one direction: from chip design to fabrication. The process engineers are responsible for the yield of the device, as long as the designer follows a set of design rules. As the process becomes more and more challenging, simple design rules cannot ensure fabrication success, and the outcome will be a low yield even after many efforts of the process engineers. Furthermore, these problems might not be apparent until after a considerable amount of money and time has been devoted to the design and fabrication of the chip. For such cases, the approach of design for manufacturing is needed, which requires that the designs are made manufacturing-friendly. An accurate process model is a key component for DFM, in order to give instant feedback as to the manufacturability of the design. The difference between the design flows is illustrated in Figure 4-10. In design for manufacturing, the process model is calibrated and maintained by product engineers and process engineers, and the calibrated model provides instant feedback to any design and ensures the design to be fab-friendly.

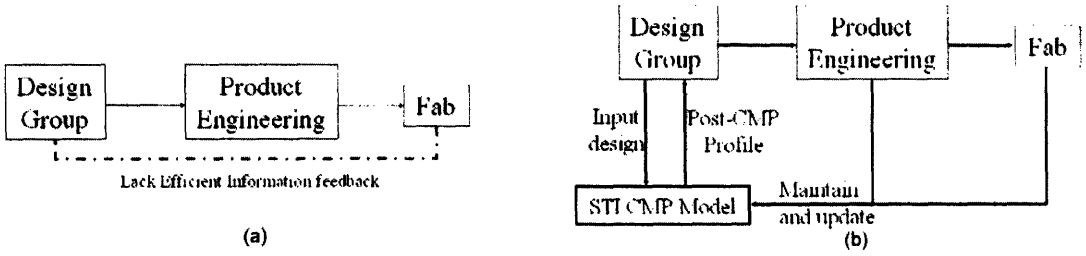



Figure 4-10: The left figure illustrates the traditional manufacturing, which lacks an efficient feedback channel. The right figure shows that the process model provides instant feedback and ensure the design is manufacturing friendly.

Figure 4-11 through Figure 4-13 shows snapshots of the implementation of an STI CMP model with a graphical user interface. This tool, implemented during an internship at National Semiconductor, encapsulates a die-level CMP model as described in Section 3.4. The model, once calibrated using the experimental data,

enables designers to input layout design information, and predicts various information about the CMP process and surface topography.

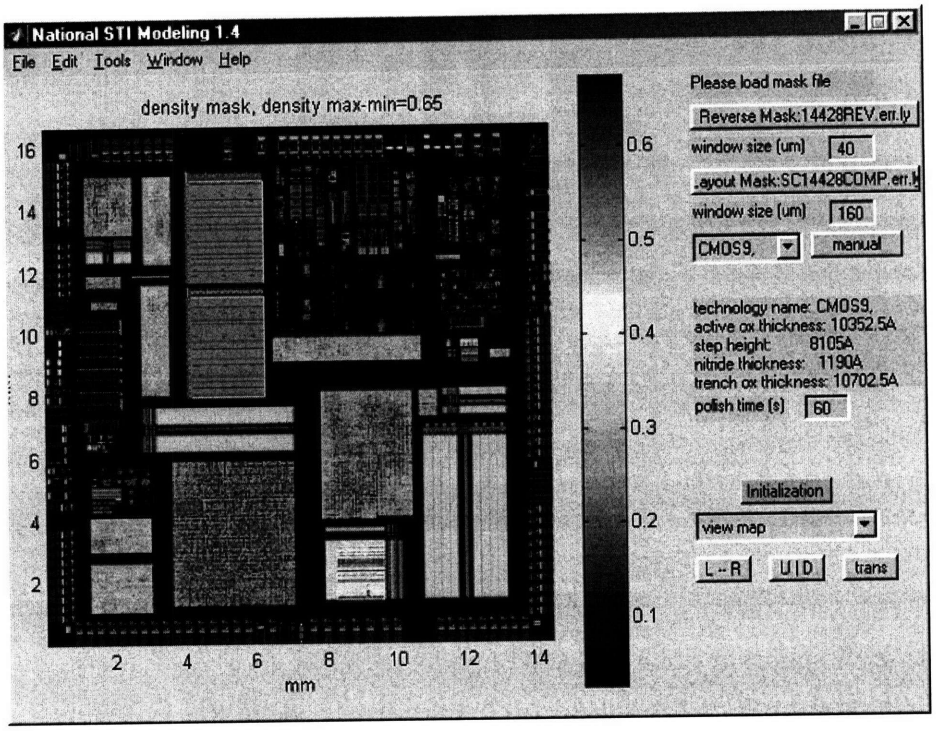



Figure 4-11: A snapshot of an STI CMP model with graphical user interface showing the pattern-density map of a product chip.

For example, given an input layout mask, the tool can plot its pattern density map, as shown in Figure 4-11. Among various predicted information, the most useful outputs for STI CMP are the map of oxide clearing time (Figure 4-12) and the map of nitride clearing time (Figure 4-11), where the oxide clearing time refers to the polishing time needed to clear the overburden oxide and the nitride clearing time refers to the polishing time that would result in complete (undesired) removal of the nitride layer. A successful STI CMP process requires a complete clearing of overburden oxide and no clearing of nitride, i.e., the maximum of oxide clearing time has to be smaller than the minimum of nitride clearing time. From Figure 4-12 and Figure 4-13, this example product chip is likely to fail in the fab, because any polishing time will result in either residual oxide or cleared nitride in some part of the chip. The tool has been in use for CMP problem screening within National Semiconductor. Additional outputs include chip maps of local step-height, time to clear nitride, and

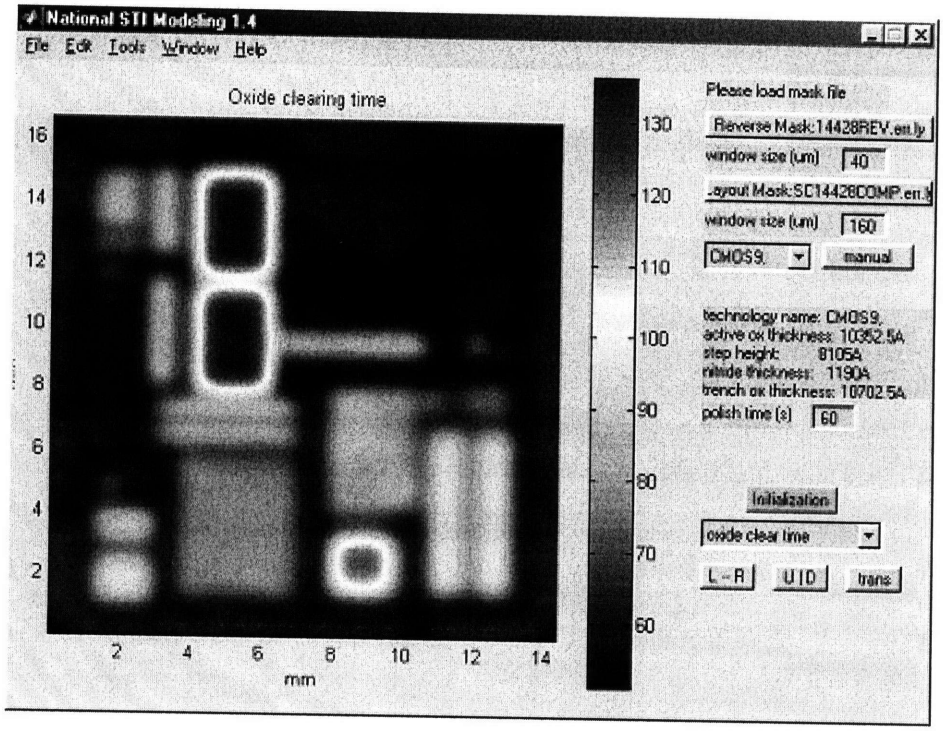



Figure 4-12: A snapshot of an STI CMP model with graphical user interface showing the map of oxide clearing time, corresponding to the chip shown in Figure 4-11.

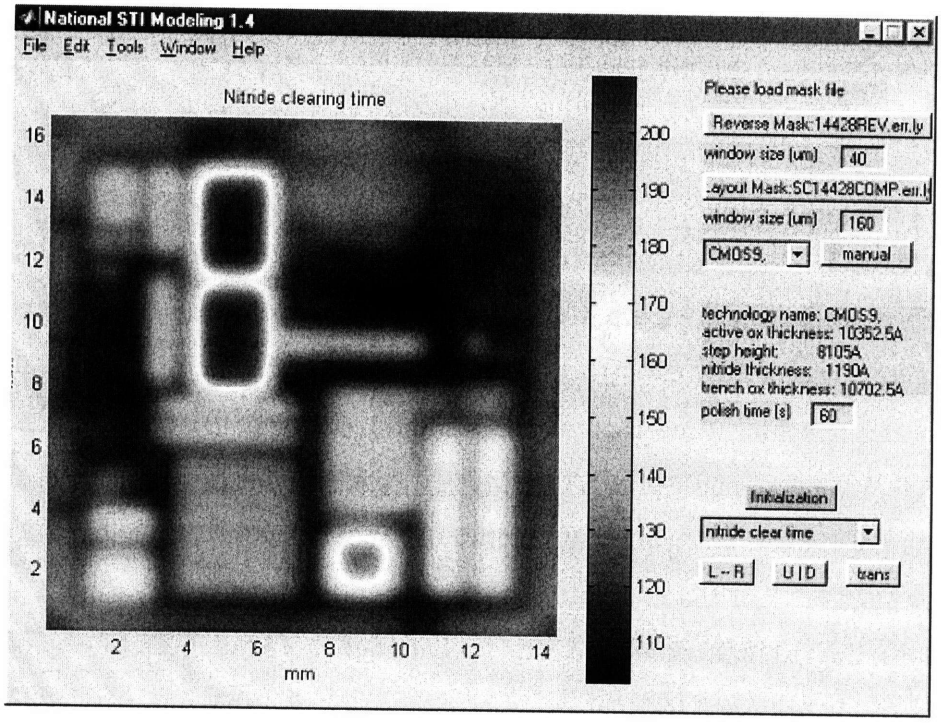



Figure 4-13: STI CMP model output showing the map of nitride clearing time, for the chip shown in Figure 4-11.

etc.

## 4.2 Nanotopography

Wafer nanotopography refers to subtle 10 to 100 $nm$ height variations which occur over lateral distances of 1 to 10 $mm$ on unpatterned silicon wafers [85] and Figure 4-14 shows the nanotopography map and the cross-section plot of a raw wafer finished with the standard single-sided polishing. Conformal thin films deposited on wafers with nanotopography have been shown to exhibit film thinning, or localized deviation in the polished film thickness, as illustrated in Figure 4-15, due to the planarization action of the CMP pad and process [86] [87] [88]. These deviations are of particular concern in modern shallow trench isolation (STI) structures fabricated using the CMP process [89]. Previous studies [89] [90] [91] focused on modeling or analysis of experimental data from polishing of blanket oxide films, where a clear nanotopography signature can be identified in post-CMP oxide thickness maps.

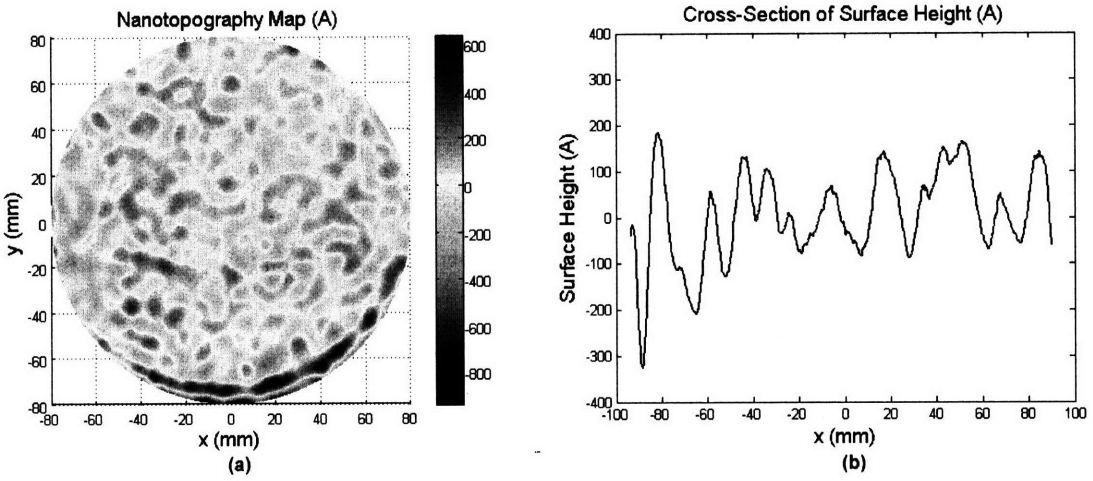



Figure 4-14: (a) Nanotopography map showing the surface across a virgin silicon wafer, and (b) cross-section of surface height.

In this section, our goal is to identify any nanotopography interactions with STI CMP, to estimate what proportion of the post-CMP wafer-level variation is contributed by nanotopography, and to test simulation models of nanotopography/CMP interaction. A new design of experiments (DOE) enables us to investigate the nanoto-

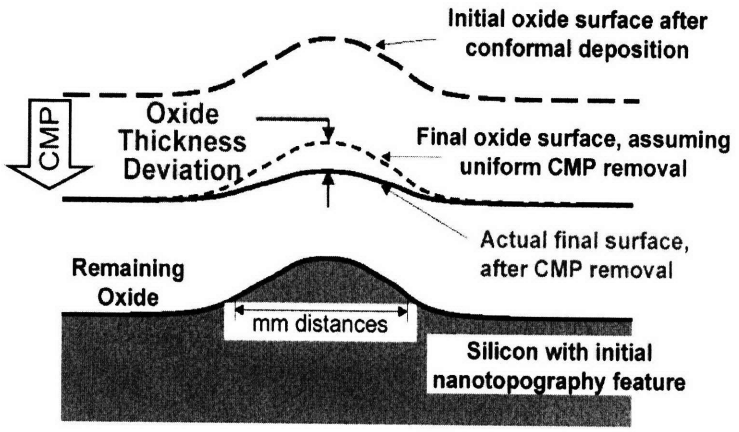



Figure 4-15: Film thinning resulting from the CMP of conformal films above nanotopography.

pography impact on polishing of both STI blanket wafers and patterned wafers. The experimental design will first be presented; this design enables us to study nanotopography impact of different wafer finishes or nanotopography signatures, different CMP processes, and different stages of CMP in a film stack. The experimental results and blanket wafer data analysis will then be presented. In analyzing the experiment data, we use both statistical analysis to identify the impact, and a contact wear model to simulate the CMP process [74] [75] [92]. Patterned wafer data analysis is presented, based on electrical measurements. Finally, conclusions are offered. The experiment, polishing, data collection, and analysis were performed in collaboration with Infineon (now Qimonda), as described in [93].

### 4.2.1 Experimental Design

A set of experiments was designed to study nanotopography and CMP interactions in STI layer stacks (oxide/nitride/oxide), in both blanket wafers and patterned wafers. Virgin 200 *mm* wafers of three types of wafer finishes were used in the experiments: standard single sided-polished (SSPs), improved single-sided polished (SSPi), and double-sided polished (DSP), in the order of reducing nanotopography. Among the raw wafers used in the experiment, SSPs wafers have an average standard deviation of nanotopography height of 249 Å, that of SSPi wafers is 104 Å, and that of DSP

wafers is 76 Å.

Before CMP, stacked films of oxide on top of nitride over a thin thermal oxide layer on silicon were applied on the blanket wafers. The blanket wafers were polished using four different processes, as listed in Table 4.1: two conventional polishes with different slurries, one with conventional pad and structure selective slurries, and one with fixed abrasive pad. The four processes will be referred to by their acronyms (LP, HP, SS, and FA) later. Half of the blanket wafer polishes were stopped in the oxide layer and the rest were polished into the nitride layer. Among a total of 60 virgin wafers started, we were not be able to obtain measurement data for nine wafers. The remaining 51 wafers are listed in Table 4.2 according to the wafer finishes, processes, and post-CMP stages. A similar DOE was used for polishing the patterned wafers, as shown in Table 4.3.

Table 4.1: Details of the four CMP processes used in the experiment.

| | Acronym | Process Name |
|---|---|---|
| 1 | LP | Conventional Low Planarization |
| 2 | HP | Conventional High Planarization |
| 3 | SS | Non-Conventional / Structure Selective |
| 4 | FA | Non-Conventional / Fixed Abrasive |

Table 4.2: Number of wafers polished in blanket wafer experiments.

| | SSPs | | SSPi | | DSP | |
|---|---|---|---|---|---|---|
| | Oxide | Nitride | Oxide | Nitride | Oxide | Nitride |
| LP | 2 | 2 | 2 | 2 | 2 | 2 |
| HP | 2 | 2 | 2 | 3 | 2 | 3 |
| SS | 2 | 2 | 2 | 1 | 2 | 3 |
| FA | 3 | 2 | 2 | 2 | 2 | 2 |

Table 4.3: Number of wafers polishing in patterned wafer experiments.

| | SSPs | SSPi | DSP |
|---|---|---|---|
| LP | 3 | 3 | 3 |
| HP | 3 | 3 | 2 |

The initial nanotopography of the bare Si wafers was measured using an ADE NanoMapper, an optical interferometry tool which produces a high resolution surface

height map consisting of 0.191 $mm$ by 0.218 $mm$ pixels. A nanotopography map is obtained by filtering the raw height map with a double-Gaussian high-pass filter with characteristic length of 30 $mm$. The post-CMP oxide or nitride thickness map was measured using an ADE AcuMap, which is a high speed imaging reflectometry tool. The post-CMP thickness map has a resolution of 2 $mm$ by 2 $mm$.

On the patterned DRAM wafers, bit cells form a latch region with unit of size 303 $\mu m$ by 163.5 $\mu m$, and arrays of these structures cover a majority of the wafer as seen in Figure 4-16. Each device is measured by applying voltages from $-0.2\ V$ to $-0.02\ V$ on the word line in $0.02\ V$ increments, and the voltage at which the data can no longer be correctly read is monitored. The electric test map records the number of successful readings (from 0 to 9) for each device. This electrical measurement is sensitive to the threshold voltage (VT), and seeks to relate any STI polishing imperfections to electrical device impact. A missing data point is marked with a solid red pixel when electrical measurement of the structure was not possible. Thus a high resolution electrical test map can be generated across the whole wafer, enabling us to explore spatial correlations between virgin wafer nanotopography and final device electrical parameters.

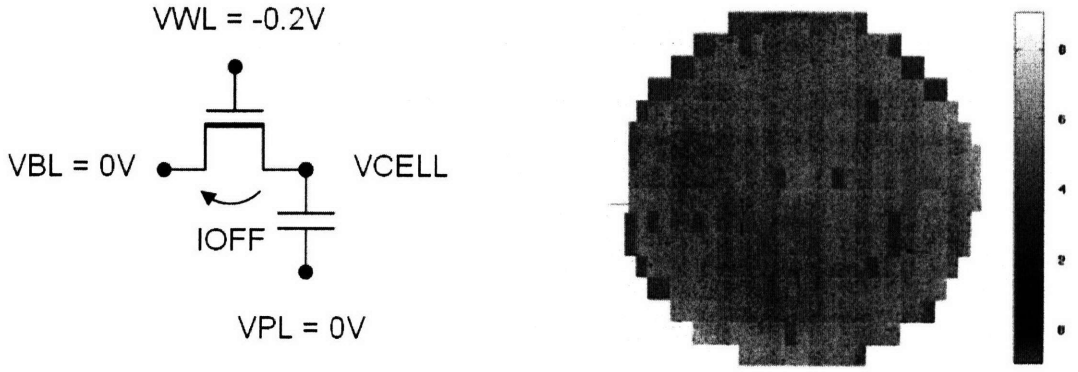



Figure 4-16: (Left) Bit cell electrical measurement. Increasing voltage steps are applied at VWL, and the VWL count (from 0 to 9 of $0.02V$ steps) of successful read steps is recorded. A low count corresponds to high cell leakage or low VT. Each test device consists of an array of bit cells with size 303 $\mu m$ by 153.5 $\mu m$. (Right) An electric test map showing VWL count across a wafer. A solid red color indicates that no experiment data is collected in the area.

### 4.2.2 Blanket Wafer Nanotopography Analysis

The blanket STI wafer study has three goals: to identify any nanotopography interactions with STI CMP, to estimate what proportion of the post-CMP wafer-level variation is contributed by nanotopography, and to test simulation models of nanotopography and CMP interaction.

#### Is There any Nanotopography Impact?

The first task is to investigate whether nanotopography has any impact on the amount removed in the CMP process, and how the impact varies for different wafer finishes and different CMP processes. We use the standard deviation of height or film thickness as a measure of surface roughness or amount removed (AR) variation. We plot how raw AR variation relates to initial surface roughness in Figure 4-17. The raw AR variations are found to be different for the different CMP processes, but they do not show a strong dependence on nanotopography. Considering that nanotopography consists of higher spatial frequency variations (compared to overall wafer trends), we filter the AR map with a 15 $mm$ double-Gaussian, and plot the filtered AR variation vs. nanotopography in Figure 4-18. We observe that the magnitude of filtered AR variation is much less than that of raw AR variation (we have removed the low-frequency CMP wafer-scale polish nonuniformity from the data) and increases slightly with the initial surface nanotopography. Next, we calculate the correlation between the nanotopography map and filtered AR map; this gives us a measure of the "print through" of nanotopography due to CMP-induced film thinning. To account for positioning variations and differences in the NanoMapper and AccuMap measurement tools, we varied the relative offset $(x, y)$ coordinates and angle between these two maps and calculate the maximal observed statistical correlation between the two maps. The results are displayed in Figure 4-19, showing the relationship between the correlation and nanotopography height as defined by the standard deviation of the initial nanotopography map.

There are two observations from Figure 4-19. The correlation coefficient generally increases with the standard deviation of nanotopography, which indicates that

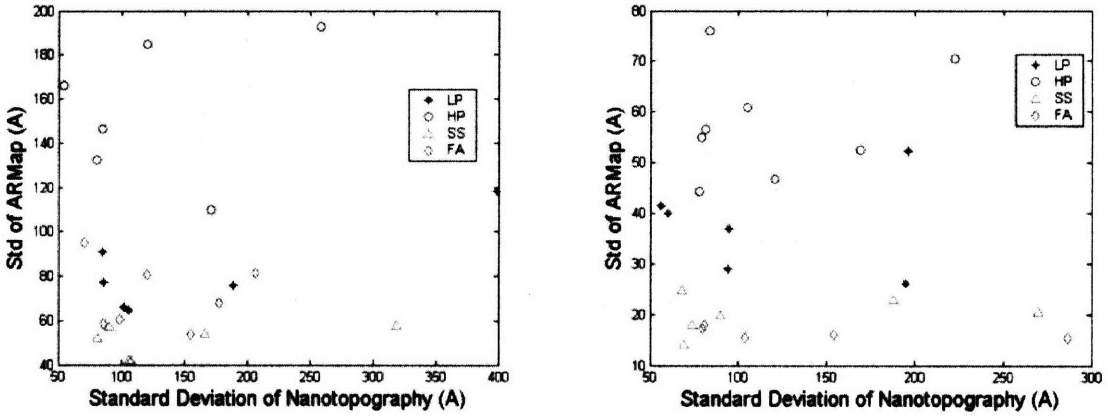



Figure 4-17: Uniformity of amount removed, measured by standard deviation of raw ARMap, vs. standard deviation of starting nanotopography. The left plot shows the oxide data, and the right shows nitride data.

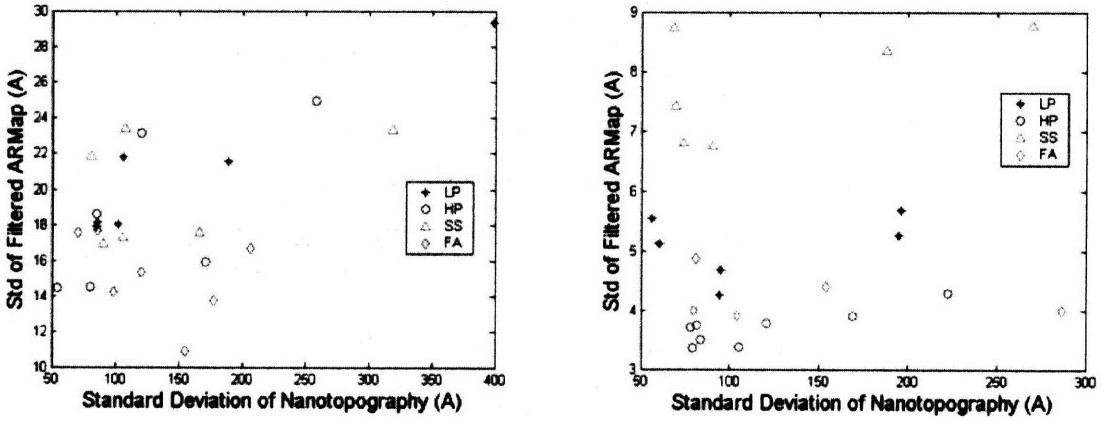



Figure 4-18: Uniformity of filtered amount removed, measured by standard deviation of the filtered ARMap, vs. standard deviation of the starting nanotopography map. The left plot shows the oxide data, and the right shows nitride data.

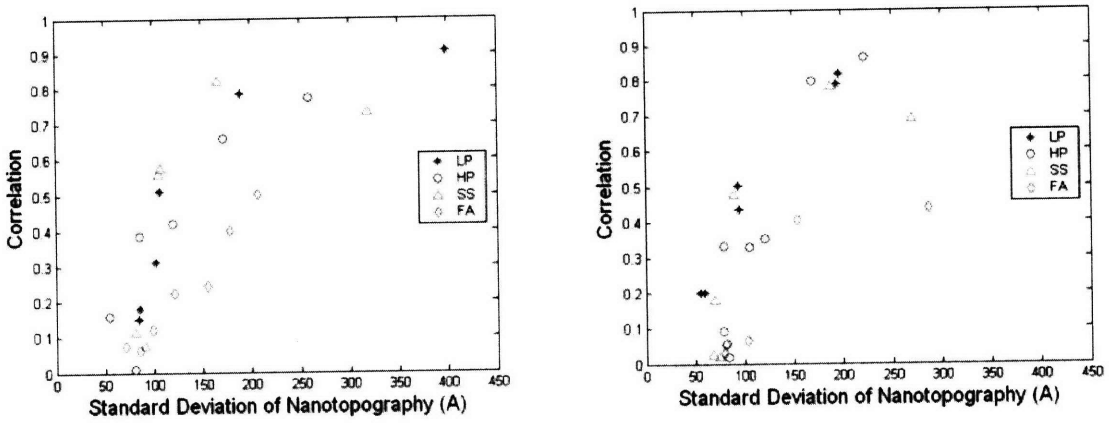



Figure 4-19: The correlation between nanotopography map and filtered oxide AR map depends on the initial nanotopography. The left plot shows the oxide data, and the right shows nitride data.

correspondingly stronger film thinning patterns are generated by larger initial nanotopography. Most of the SSPs wafers, which have the largest standard deviation of nanotopography values, have high correlation values above 0.8, indicating a strong impact from the initial wafer nanotopography. The spatial correlation can be observed in Figure 4-20, which shows the comparison between the nanotopography map and the filtered oxide AR Map for a polishing process stopped in the oxide layer, and Figure 4-21, which illustrates the case for a polishing process stopped in the nitride layer.

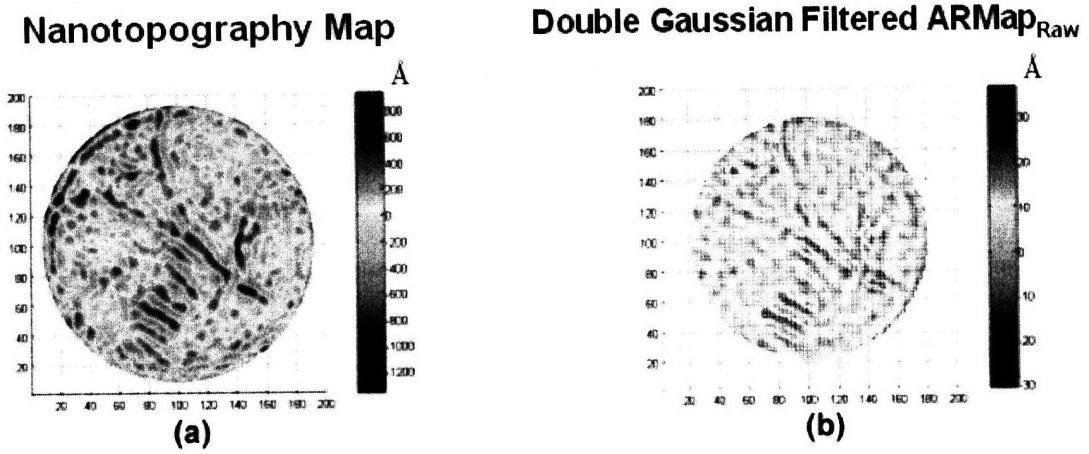



Figure 4-20: (a) The nanotopography map, and (b) the double-Gaussian filtered ARMap of a polishing process stopped in the oxide layer.

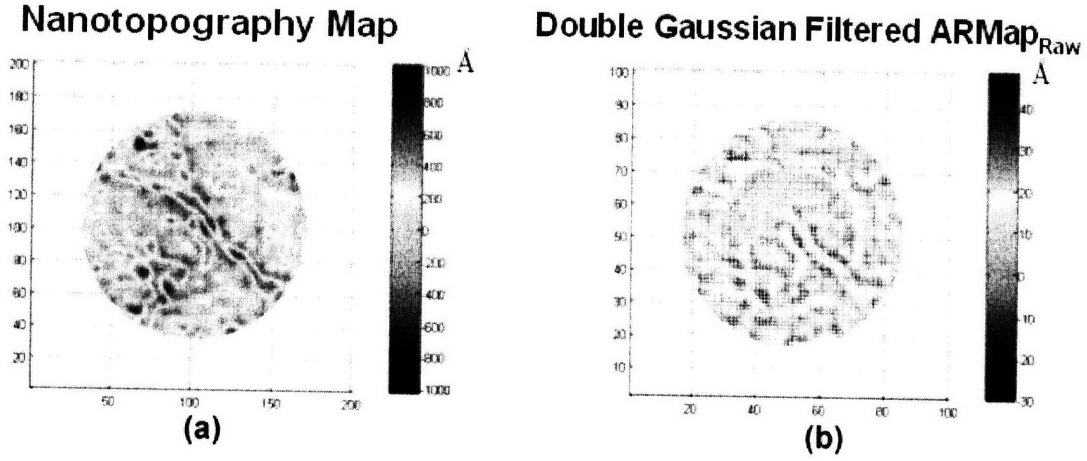



Figure 4-21: (a) The nanotopography map, and (b) the double-Gaussian filtered ARMap of a polishing process stopped in the nitride layer.

### How Large is Nanotopography Impact?

The next issue is what proportion of the total non-uniformity of material removed during CMP can be attributed to the starting nanotopography. This can be estimated from the correlation calculation above. In calculating the correlation, we are implicitly assuming a linear relationship between the starting nanotopography and the filtered amount removed:

$$ARMap_{DG\ Filtered} = \beta \cdot NanoMap + \epsilon \tag{4.2}$$

Thus, $std(\beta \cdot NanoMap)/std(ARMap_{DG\ Filtered})$ provides an estimate of nanotopography impact proportion (Figure 4-22). The ratios are larger for wafers with rougher initial surfaces and some of the ratio values are close to one, which implies that nanotopography contributes a large portion of the $ARMap_{DG\ Filtered}$ in those cases.

One also might want to know the nanotopography contribution to raw $ARMap$, and thus $std(\beta \cdot NanoMap)/std(ARMap_{Raw})$ is also calculated (Figure 4-23). This result implies that nanotopography impact contributes only a small proportion of total post-CMP thickness variation, although the ratio increases with $std(NanoMap)$ in general. The ratio values are less than 10%, and most of them are less than 5%. This fact is not surprising, as $ARMap_{DG\ Filtered}$ only accounts for approximately 10% of $ARMap_{Raw}$, according to Figures 4-17 and 4-18.

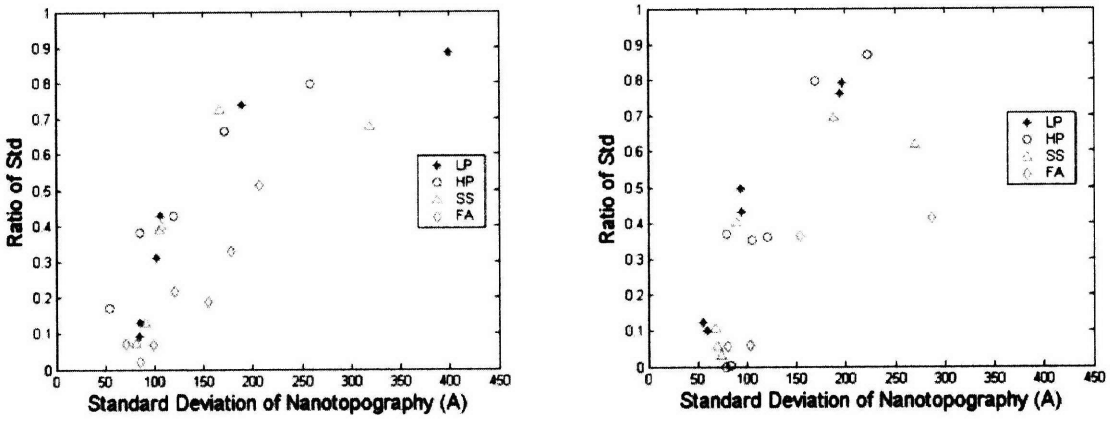



Figure 4-22: Left plot shows $std(\beta \cdot NanoMap)/std(ARMap_{DG\ Filtered})$ vs. $std(NanoMap)$ for oxide data, and the right plot shows that for nitride.

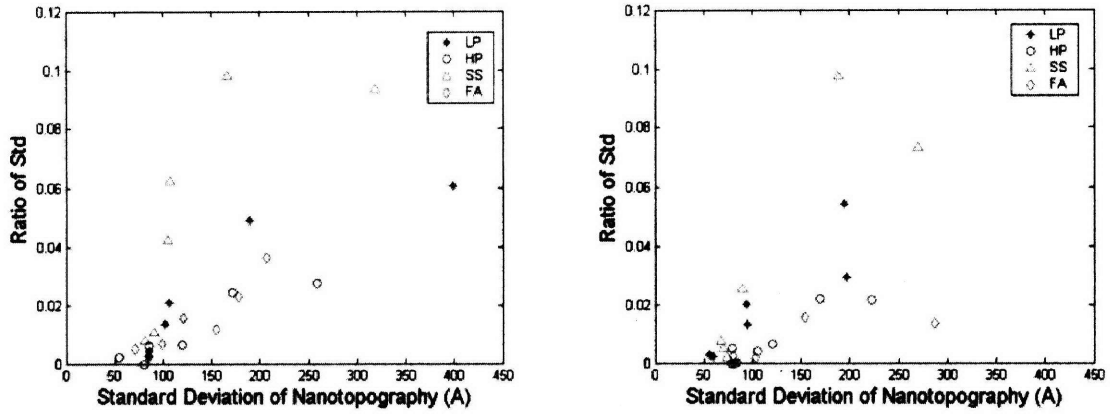



Figure 4-23: Left plot shows $std(\beta \cdot NanoMap)/std(ARMapRaw)$ vs. $std(NanoMap)$ for oxide data, and the right plot shows that for nitride.

### How to Model Nanotopography Impact?

The above discussion indicates that nanotopography contributes only to the high spatial frequency part of $ARMap$, and the low frequency part is the wafer level nonuniformity caused by CMP tool and process limitations.

$$ARMap_{Raw} = ARMap_{Low\ Freq} + ARMap_{DG\ Filtered} \tag{4.3}$$

In Figure 4-24 the $ARMap_{Low\ Freq}$ is displayed for oxide data for all blanket wafers polished for one of the processes (the low planarization or LP process). Some of these low frequency wafer level maps appear to be similar to each other. Low frequency maps for wafers polished in other processes (not shown here) also exhibit other spatial wafer-level similarities. These spatial similarities suggest that the low frequency pattern may be related to the specific CMP tool and process combination.

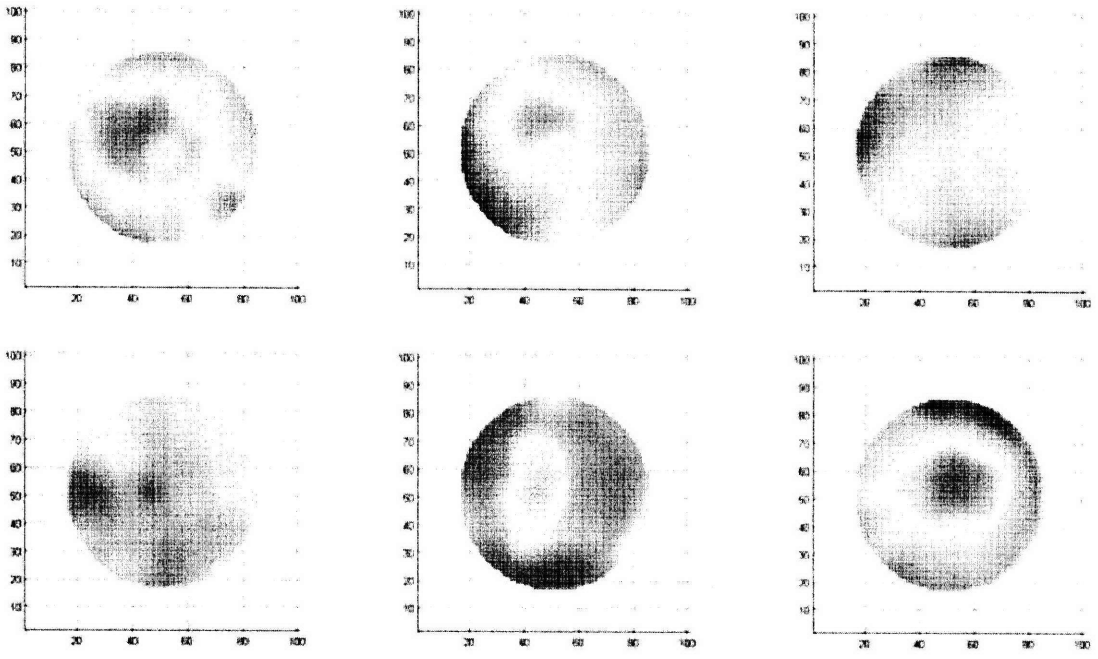


Figure 4-24: $ARMap_{Low\ Freq}$ of all LP wafers.

To model the nanotopography impact, the surface defined by the nanotopography map is taken, and a contact wear model is used for predicting the amount of material removed during CMP of oxide and nitride layers. Then we compare the

model predicted $ARMap_{CW}$ with $ARMap_{DG}$ Filtered. In contrast to Equation 4.2, where a simple linear dependence on the naanotopography height is assumed, here Equation 4.4 implies a potential non-linear functional dependence as computed by the CMP contact wear model.

$$ARMap_{DG\ Filtered} = ARMap_{CW}(NanoMap) + \epsilon \tag{4.4}$$

The contact wear model parameters consist of effective Young's modulus, DG filter length, and nitride selectivity, and are estimated by minimizing the root mean square (RMS) error of $\epsilon$. Two points should be noted in the parameter estimation procedure:

- The smaller the DG filter length, the smaller is the post-CMP variation as described by $ARMap_{DG\ Filtered}$, and thus smaller RMS errors are observed. Therefore, purely minimizing RMS error will lead to a zero value for the DG filter length. A better metric to drive the model fit is $RMS(\epsilon)/Std(ARMap_{DG\ Filtered})$, which is the proportion of unexplained variation.

- We assign one single set of parameters for each process, and parameters are extracted by minimizing the equal-weighted average of $RMS(\epsilon)/Std(ARMap_{DG\ Filtered})$ values of all polishes using the same process. DSP wafers have a relatively smooth initial surface and substantially smaller nanotopography impacts. Thus, only data from SSPs and SSPi are used to calibrate the CMP contact wear model parameters.

Seven cases are chosen to illustrate the nanotopography impact for different processes, wafers of different finishes, and in the different CMP stages. Figures 4-25 through 4-31 show the comparison between model predictions and experiment results; each figure contains three subplots: $ARMap_{DG\ Filtered}$, model predicted $ARMap$, and a cross-section comparison. In the plots, areas within 40 $mm$ from the edge are excluded. As $ARMap_{DG\ Filtered}$ displays only the high frequency part of the measured data, the values are centered around its mean, zero. In the cross-section comparison, $ARMap - mean(ARMap)$ is represented by the line and $ARMap_{DG\ Filtered}$ is represented by the

dots.

The case of wafers having SSPs wafer finish, being polished by LP process, and stopped in oxide layer is displayed in Figure 4-25, whereas in Figure 4-26 the polishing is stopped in the nitride layer. By either visual comparison of $ARMap_{DG\ Filtered}$ and predicted $ARMap$ plots or cross-section comparison, one can conclude that the model predictions agree well with experimental data for both Figures 4-25 and 4-26. Importantly, $ARMap_{DG\ Filtered}$ has much smaller variation in the nitride stage than that in the oxide stage, due to the slower polishing rate of nitride and nitride to oxide polish rate selectivity.

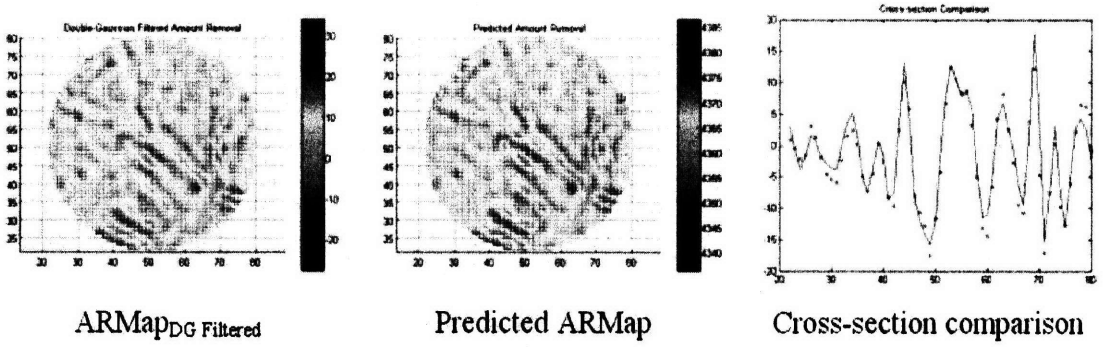



Figure 4-25: Comparison of experimental data and predictions, for the LP process, SSPs wafer finish, and oxide polishing.

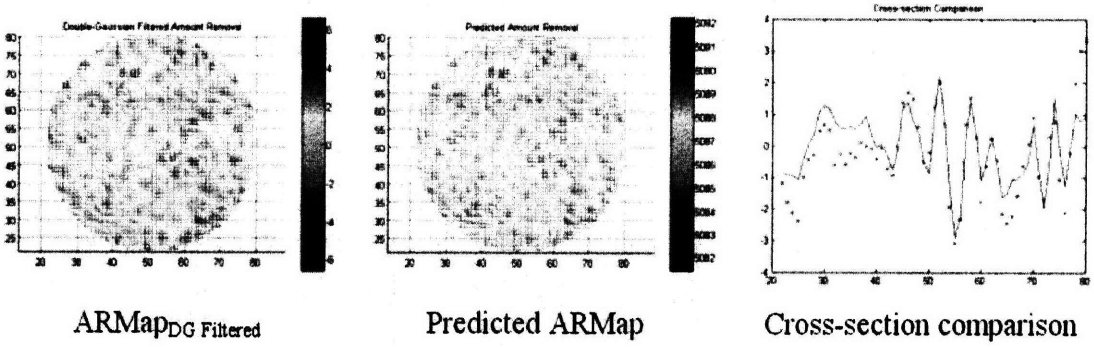



Figure 4-26: Comparison of experimental data and predictions, for the LP process, SSPs wafer finish, and nitride polishing.

The SSPs wafers being polished by LP, HP, SS, and FA processes and all polishing stopped in the nitride stage are displayed in Figures 4-26 (LP), 4-27 (HP), 4-29 (FA), and 4-29 (FA). For the first three processes, the model predictions agree well with

experimental results. The FA data are not well explained by the model, and the $ARMap_{DG\ Filtered}$ of the FA case has much smaller variation, roughly $-0.8$ Å to 0.8 Å compared with $-5$ Å to 5 Å for the other processes. The result implies a weak nanotopography impact in the fixed abrasive FA CMP process.

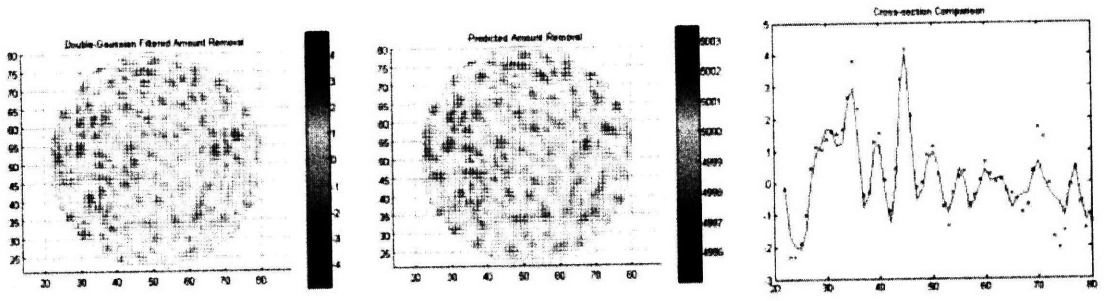



Figure 4-27: Comparison of experimental data and predictions, for the HP process, SSPs wafer finish, and nitride polishing.

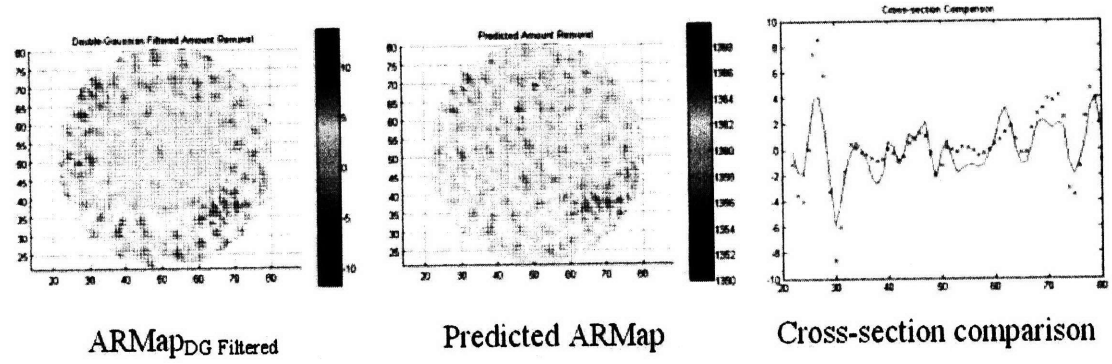
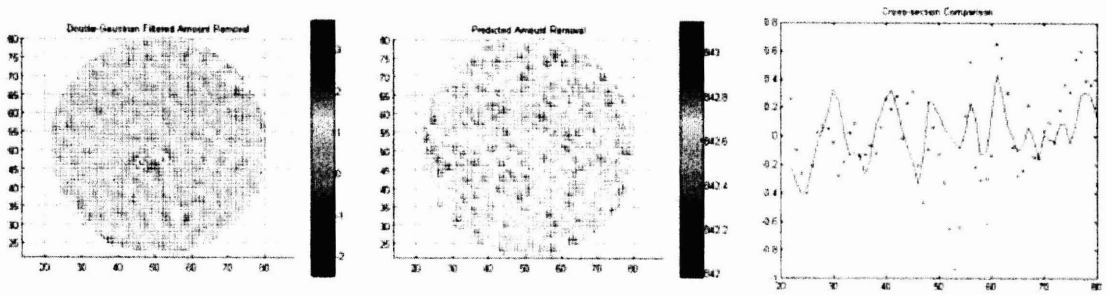


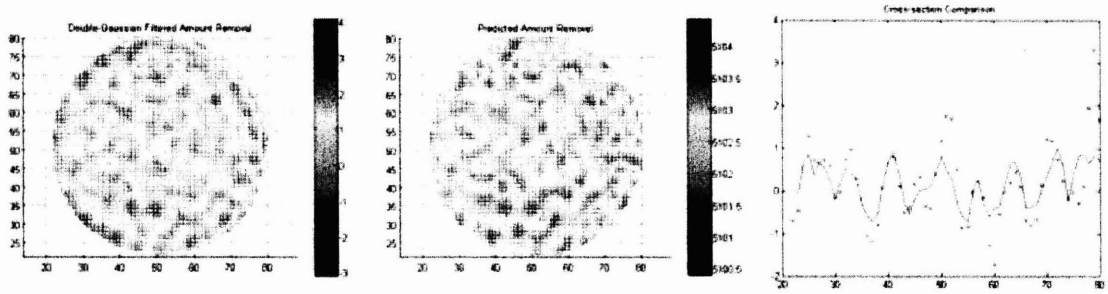

Figure 4-28: Comparison of experimental data and predictions, for the SS process, SSPs wafer finish, and nitride polishing.

Wafers having three different wafer finishes, polished by LP process and stopped in the nitride layer, are displayed in Figures 4-26 (SSPs), 4-30 (SSPi) and 4-31 (DSP). The model works well in the SSPs case, reasonably well in the SSPi case, and not so well in the DSP case. This fact suggests that as initial nanotopography decreases, its impact becomes comparable with the noise level.



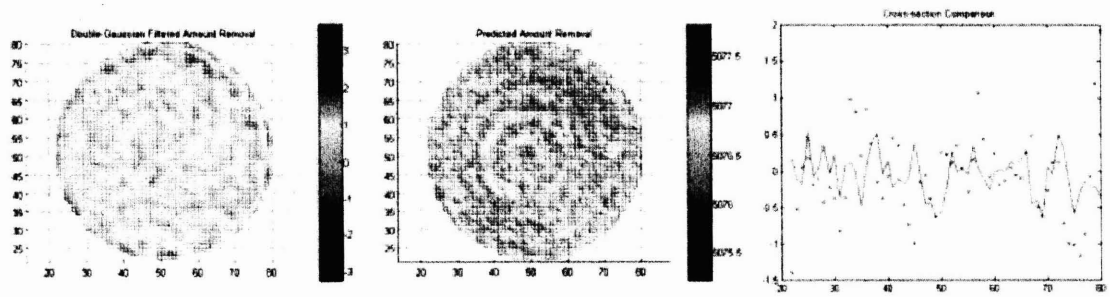
$ARMap_{DG\ Filtered}$ Predicted ARMap Cross-section comparison

Figure 4-29: Comparison of experimental data and predictions, for the FA process, SSPs wafer finish, and nitride polishing.



$ARMap_{DG\ Filtered}$ Predicted ARMap Cross-section comparison

Figure 4-30: Comparison of experimental data and predictions, for the LP process, SSPi wafer finish, and nitride polishing.



$ARMap_{DG\ Filtered}$ Predicted ARMap Cross-section comparison

Figure 4-31: Comparison of experimental data and predictions, for the LP process, DSP wafer finish, and nitride polishing.

### 4.2.3 Patterned Wafer Nanotopography Analysis

In the analysis of the patterned wafer data, a statistically based approach is used for estimating the nanotopography impact on patterned wafer electrical performance. Attempts to apply a model based rather than statistical analysis method would encounter several difficulties. First, we would need some function relating the electric test information to the post-CMP surface topography. Second, we would need very accurate and calibrated patterned wafer CMP simulation models integrating both layout pattern dependent and nanotopography effects [89]. This demand for extremely high accuracy (error less than 1 nm) is indicated by the blanket data analysis which suggests that the nanotopography impact contributes only a small fraction to the final nitride thickness variation; thus the small contribution might be difficult to identify with the presence of additional layout pattern-dependent variation. Our goal, then, is to search for any evidence of the starting nanotopography impact on the final electrical data, underneath the larger wafer-level and chip-layout dependent variations.

#### Correlation Analysis

In the blanket wafer data analysis, we observe spatial correlation between nanotopography map and the high frequency part of the oxide/nitride thickness map, and the correlation becomes weaker with less amount of starting wafer nanotopography. The same analysis is performed with pattern wafer data, and the result is shown in Figure 4-32. The left figure shows the analysis using raw e-test data, and the right figure shows the analysis using only the high spatial frequency component (filtered) of the e-test data. In either case, the correlation values show no noticeable trend, and the values group around zero.

One possibility explaining the low correlation is the presence of other variation sources, such as periodic variations (due to strong chip-scale layout pattern effects) and radial variations (due to wafer-level CMP polish nonuniformity). Removing these variations enhances the signal-to-noise ratio of our analysis. The periodic variation can be estimated by averaging the e-test data over all dies of each wafer. The dies

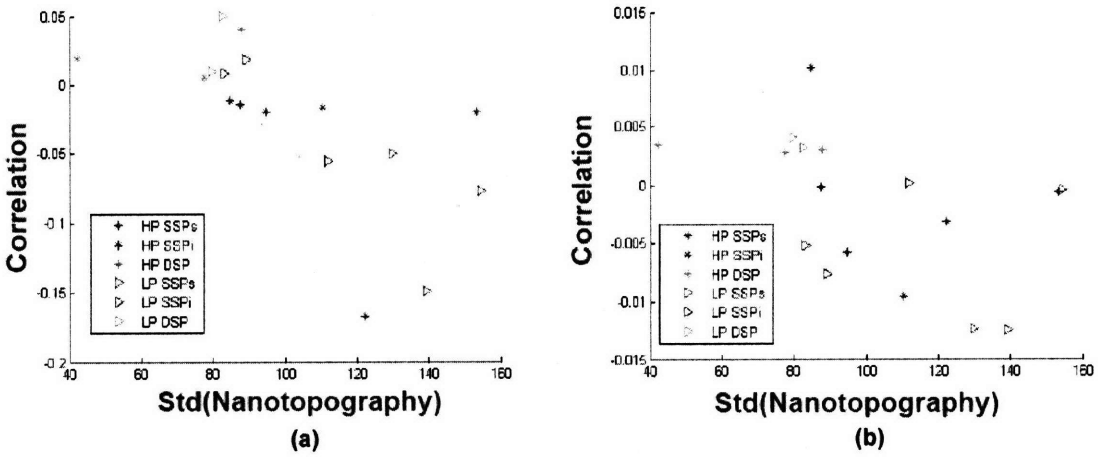



Figure 4-32: Correlation analysis of pattern wafer data using (a) the raw e-test map, and (b) only the high spatial frequency part.

near the wafer edge are excluded because they are partial dies and are susceptible to edge variations. The radial dependence can be estimated by averaging e-test data with the same radius. As an illustration, Figure 4-33 (a) shows a raw e-test map, (b) shows the e-test map excluding the periodic variation, and (c) shows the e-test map excluding both the periodic variation and radial dependence.

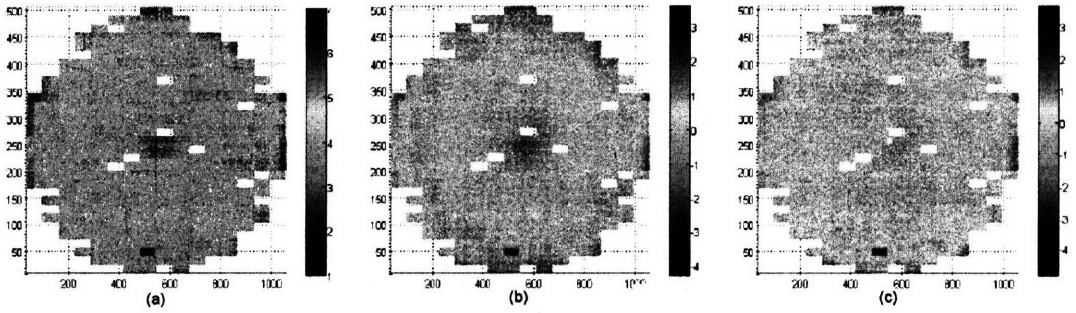



Figure 4-33: Illustration of removing periodic and radial variations. (a) Raw e-test map; (b) Periodic variation removed; and (c) Radial variation also removed.

The correlation analysis is applied to the e-test data excluding both the periodic and radial variation. The result is shown in Figure 4-34, where little correlation is observed. Thus, the spatial correlation analysis fails to detect any relationship between the starting nanotopography map and the e-test map.

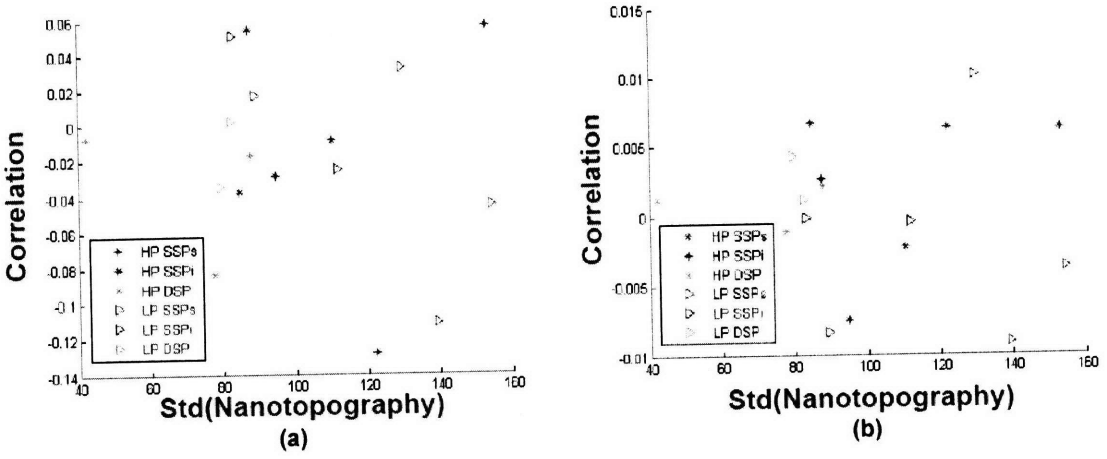



Figure 4-34: Correlation analysis of patterned wafer data with periodic and radial variations removed using (a) the raw e-test map, and (b) only the high spatial frequency part.

### Linear Factor Analysis

It has to be admitted that the existence of spatial correlation is a strong assumption as it implies a direct linear relationship between the e-test data and nanotopography height map. The question now is whether there is any nanotopography impact on e-test data. One direct measure of nantopography is its standard deviation $std(NanoMap)$, and we want to find out if it is a factor of the e-test result. The two metrics of e-test data used are the mean and standard deviation of the number of tests passed, $N_{pass}$. The mean value $Mean(N_{Pass})$ is calculated by averaging $N_{Pass}$ over all devices for each wafer, and $std(N_{pass})$ is calculated in a similar way. We plot $Mean(N_{Pass})$ vs. $std(NanoMap)$ in Figure 4-35 for LP and HP processes, with $std(N_{pass})$ plotted as the error bar. The figure shows that the within-wafer variations of $N_{Pass}$ are larger than the difference between wafers with different finishes, which suggests that wafer-level non-uniformity and pattern density induced non-uniformity are the dominant factors. The slope of the left figure (HP process) is $-0.006$, which suggests more nanotopography results in a lower number of tests passed on average; however, the 95% confidence level is $(-0.0165, 0.0046)$, which implies that the slope is not significantly from zero. The analysis of LP process data also shows a non-significant slope of $-0.0028$ with 95% confidence level $(-0.0156, 0.0100)$. In using the

raw e-test data, however, we are looking for small nanotopography effects inside of large variations created by die-level and wafer-level CMP nonuniformity. Thus, we need to refine our tests to improve their power to identify any subtle nanotopography effects.

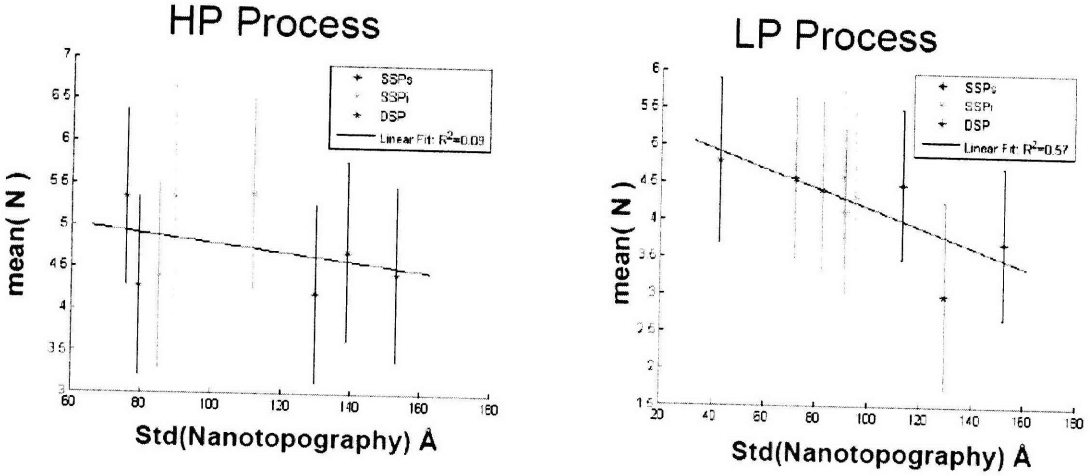



Figure 4-35: Linear factor analysis of raw e-test data. HP process (left): slope $= -0.006$ and 95% confidence level (-0.0165,0.0046), LP process (right): slope $= -0.0028$ and 95% confidence level (-0.0156,0.0100).

We apply the same analysis to the e-test data excluding periodic variation, which is the average of all dies within a wafer, and the result is shown in Figure 4-36. For both HP and LP processes, the slope values are still not significant at the 95% confidence level.

Now we apply the linear analysis to the periodic variation itself, and the results are shown in Figure 4-37. The slope for the HP process is $-0.0251$ with 95% confidence level $(-0.0473, -0.0030)$, and the slope for the LP process is $-0.0030$ with 95% confidence level $(-0.0044, -0.0016)$. In both cases, the slope values are significant with 95% confidence. The fact that nanotopography impact is observed in the periodic variation part of the e-test data implies that nanotopography impact is closely coupled with pattern-density induced variations.

The linear analysis shows the $mean(N_{Pass})$ decreases slightly with increasing $std(Nano)$, which indicates that a larger initial nanotopography is expected to cause lower device pass rate. The slope is significantly different from zero at the 95%

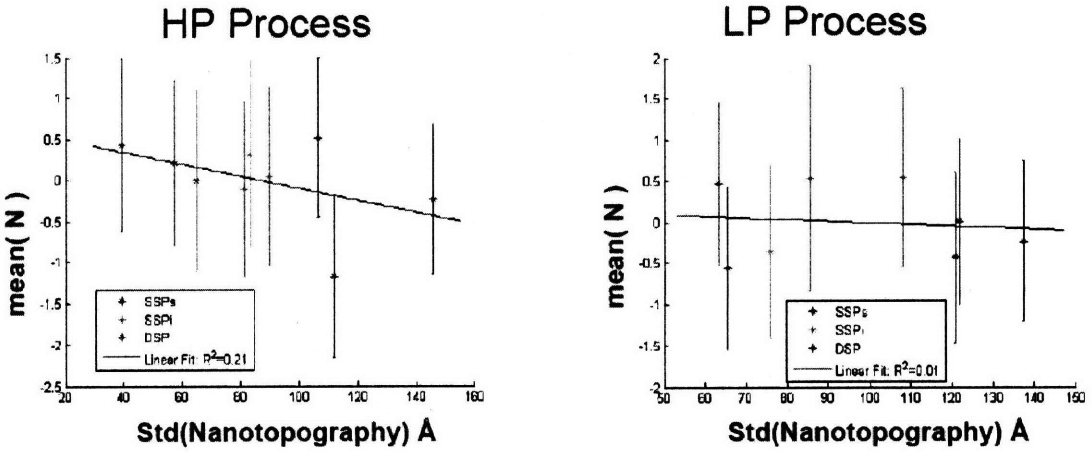



Figure 4-36: Linear factor analysis of residual e-test data. HP process (left): slope = −0.0074 and 95% confidence level (-0.0198,0.0051), LP process (right): slope = −0.0019 and 95% confidence level (-0.0179,0.0141).

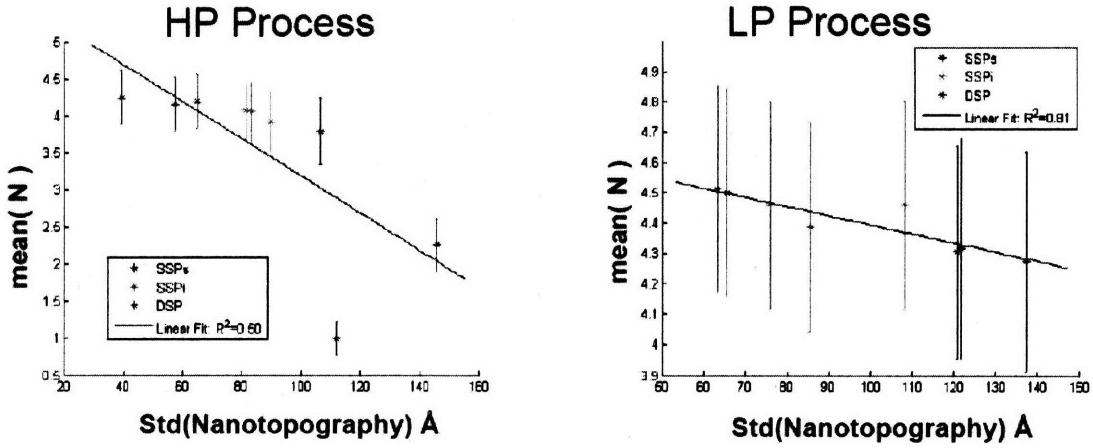



Figure 4-37: Linear factor analysis of mean-die data. HP process (left): slope = −0.0251 and 95% confidence level (-0.0473,-0.0030), LP process (right): slope = −0.0030 and 95% confidence level (-0.0044,-0.0016).

confidence level for both HP and LP processes. The large value of $std(N_{Pass})$ indicates that tool-induced and process-induced wafer-level variation, as well as pattern induced die-level variation, are the dominant factors.

### Nested Analysis of Variance (ANOVA)

The previous sections show that the nanotopography impact is not a dominant factor but that it does exist. In this section, we estimate the impact using nested analysis of variation (ANOVA). The nested ANOVA is used because the e-test data set suffers from several variation sources in a nested structure. Due to the design of the experiment, the data set has variation due to different wafer finish, which is directly related to the amount of nanotopography, and different CMP processes, either HP or LP process. In the measurement of $N_{Pass}$, there are lot-to-lot variation, wafer-level variation, and with-in-die variation, and these sources form a nested structure. ANOVA enables us to estimate the variance contribution of each variation source, and nested ANOVA is the proper way to deal with our data set. The result of nested ANOVA is shown in Table 4.4, indicating that with-in-die variation accounts for 69% of the total variation, and process variation accounts for about 12%. Although nanotopography accounts for only 4.4%, the analysis shows that it is a significant factor on a 99% confidence level, which is in agreement with previous section.

Table 4.4: Result of Nested ANOVA Analysis.

| Source | DOF | Mean Sq. | F | Pr>F | Variance | % of Variance |
|---|---|---|---|---|---|---|
| Total | 5955663 | 1.7 | | | 1.80 | 100 |
| Process | 1 | 355639 | 188.9 | 0.00 | 0.21 | 11.7 |
| Nanotopography | 2 | 176326.5 | 93.6 | 0.01 | 0.08 | 4.4 |
| Run-to-Run | 13 | 67023.9 | 35.6 | 0.03 | 0.14 | 7.9 |
| Wafer-Level | 447 | 1882.8 | 1.0 | 0.42 | 0.11 | 6.4 |
| With-in-Die | 5955200 | 1.3 | | | 1.25 | 69.6 |

## 4.2.4 Conclusion

In this study, we use an extensive set of blanket and patterned wafer experiments to explore the nanotopography impact on STI CMP. From data obtained by polishing

blanket wafers with an oxide/nitride/oxide film stack, it is observed that initial nanotopography contributes to the high frequency part of material removed variation; this nanotopography contribution accounts for less than 10% of the total post-CMP thickness variation. A contact wear CMP model works well to predict film thinning resulting from nanotopography, for data using LP, HP, and SS processes. For each process, oxide or nitride material removed maps can be predicted corresponding to the different initial wafer finishes with only one set of three parameters. The CMP model prediction for the fixed abrasive process agrees poorly with data, and requires further investigation. The analysis of patterned wafer data identifies nanotopography to be a significant factor, and larger starting nanotopography results in a smaller number of tests passed on average. A nested ANOVA of the data confirms nanotopography as significant at the 99% confidence level. While statistically significant, starting nanotopography is found to account for only 4.4% of the observed total variation in DRAM e-test measurements.

## 4.3 Wafer Edge Roll-off

CMP is an enabling technology to achieve local planarization and wafer-level uniformity. Recent studies have shown the existence of an important topographic issue: wafer edge roll-off. The edge roll-off is typically a convex profile near the wafer periphery and occurs within $1-5\ mm$ from the wafer edge, as illustrated in Figure 4-38. These profiles can exhibit a range of variations other than a simple convex profile as discussed later. In both prime wafer preparation and in IC fabrication, edge roll-off may result from CMP polish nonuniformity. In addition, this edge roll-off can potentially impact uniformity in subsequent CMP processes [17].

The edge roll-off can result from the inherent discontinuities of the process tool geometry at the wafer edge. Figure 4-39 shows the configuration of a typical rotary CMP tool. The wafer carrier holds the wafer facing down, which is polished against the polishing pad. Figure 4-39 also schematically pictures the geometry near the wafer edge. The wafer is held by a retaining ring, which is usually a few millimeters away from the edge of the wafer. In a typical setting, different pressures are applied

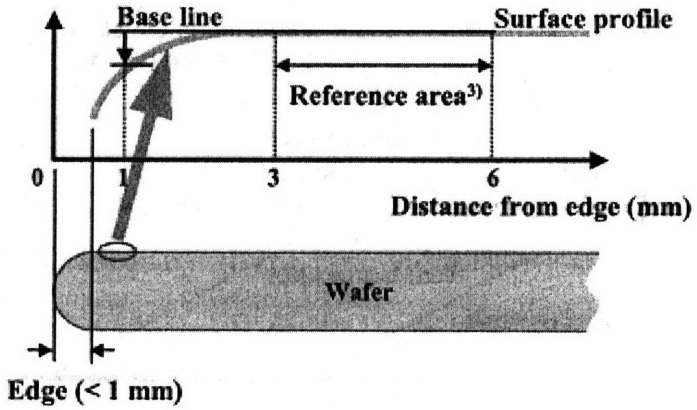



Figure 4-38: A schematic showing wafer edge roll-off, a deviation in the wafer geometry from a flat baseline level, near the edge of the wafer [17].

to the wafer and ring, with the ring usually under higher pressure to prevent the wafer from slipping out. The pad bends around the wafer edge due to the existence of the gap and retaining ring; thus the wafer edge is polished non-uniformly due to a localized pressure concentration.

The factors causing the non-uniform pressure distribution are gap size, pad stiffness (effective Young's modulus), and pressures on the wafer and retaining ring. In this paper, we study how these factors influence the edge roll-off profile. First, we present some background and examples of wafer edge roll-off. Second, we review the contact wear model used in the simulations. Then, we present our simulation results in a three part discussion. The first part focuses on static cases, which is the relationship between the tool and pad factors and pressure distribution at a single time instant. Second we consider dynamic cases, where the evolution of edge profile is simulated over time with different values of the factors. The first two parts help us to understand how edge roll-off is generated; the third part examines the influence of initial wafer profile on subsequent polishing.

Several studies have focused on the relationship between wafer scale CMP nonuniformity and tool geometry [17] [94] [95]. Fukuda et al. [17] found that wafer flatness in the peripheral area of a wafer was relatively worse than that of the inside region, which strongly suggests that edge roll-off deteriorates wafer flatness at the wafer edge.

They also found that an optimum polishing condition could cover several wafers with limited roll-off variation. Castillo-Mejia et al. [94] developed a stress-based model for the qualitative description of removal rates during silicon dioxide CMP on Applied Materials' Mirra polisher, and used the model to study the wafer level CMP nonuniformity affected by the wafer-retaining ring separation and the altering pressure exerted by the retaining ring. Soroooshian et al. [95] demonstrated via experiments that variations in wafer geometry as measured by overall shape and size of wafer-ring gap can significantly impact the extent of within wafer non-uniformity in pressure and hence inter-layer dielectric removal rate.

Measurements of several wafer edge profiles are shown here to give us some intuition about the real wafer edge profile. The measurements are done on blanket starting wafers from various wafer vendors. The starting wafer is usually single-sided or double-sided polished after it is sliced from the bulk and ground. Thus, the measurements on starting wafers can exhibit edge roll-off profiles, and these initial topographic profiles will influence further CMP processes.

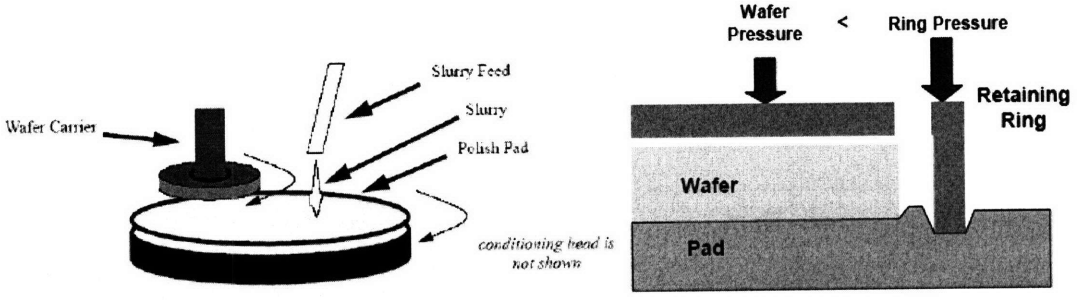



Figure 4-39: Illustrations of a typical rotary CMP tool set. The left diagram shows the whole tool, and the right diagram focuses on the geometry near the wafer edge.

Both the wafer front surface profile and the thickness are measured by an ADE WaferSight optical flatness gauge. A dataset of 59 front surface scans and 100 thickness scans was examined, and a few typical profiles are shown in Figure 4-40. The wafer front surface is the surface being polished and its profile is measured in unchucked condition. The profile shows mainly wafer warp and bow, and the edge of the wafer does not have any distinctive roll-off profile. In contrast, the thickness

scans have much smaller magnitude but exhibit distinctive roll-off profiles near the edge of wafer. This seeming puzzle comes from the fact that the wafer is under substantial pressure during the CMP process as schematically indicated in Figure 4-41, and thus the front surface profile that the pad sees results from the wafer thickness. The thickness profiles show many different patterns, which is probably due to the variety of CMP tools, polishing pads, and processing conditions used by different wafer vendors.

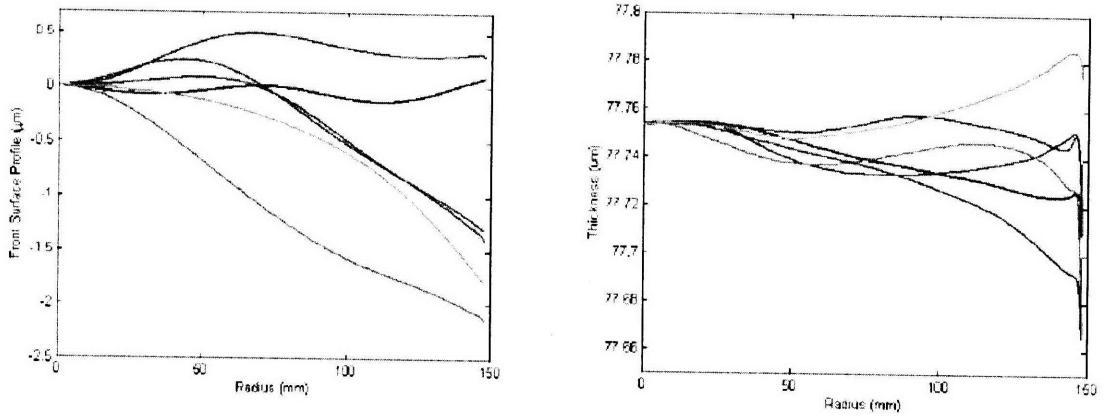



Figure 4-40: Left plot shows several measured front surface profiles; note that the range of variation is about a few microns. Right plot shows several measured thickness scans; here, the range of variation is approximately only 0.05 $\mu m$.

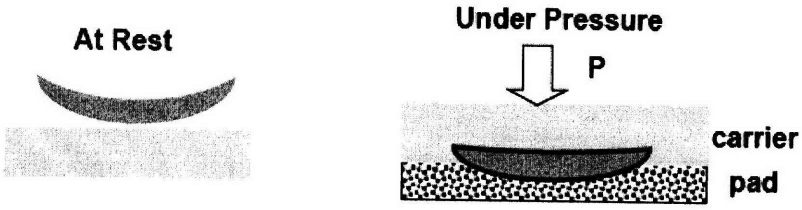



Figure 4-41: The diagrams show the difference between the wafer at rest and under pressure. We assume that the profile the pad sees is determined by the wafer thickness, based on the wafer carrier maintaining a flat back wafer profile.

### 4.3.1 Wafer Scale Contact Wear Model

In this section, we apply the contact wear model to simulate CMP on the wafer scale. Contact wear modeling of CMP was proposed by Chekina [74], Yoshida [75],

and Vlassak [96]. Its concept is to relate local pressures on the wafer surface to the polishing pad displacement, and use the local pressures to derive local removal rates, typically assuming the polish rate is linearly proportional to the pressure. As the film surface evolves through time, the pad displacement is changed, and thus local pressures and removal rates are modified. The simulation surface is discretized into elements, and a time-stepped algorithm is used to determine the final post-CMP film surface. Yoshida [75] describes a boundary element method to solve the matrix form of this equation.

Applying 2D contact wear simulation to the whole wafer is computationally expensive, as we need to discretize a 300 $mm$ wafer into sub-millimeter cells to study in detail the several millimeter area near the edge of the wafer. Instead we apply the model to an area of 60 $mm$ $\times$ 60 $mm$ at the wafer edge, as illustrated in Figure 4-42. The area is discretized into 0.1 $mm$ $\times$ 0.1 $mm$ cells, and we study the 10 $mm$ line across the gap at the center of the area. The distance from the area of interest to the closest boundary is 20 $mm$, and we assume that the interaction of elements more than 20 $mm$ apart is negligible. The assumption is reasonable as the interaction distance, of which planarization length can be one measure, is generally less than 10 $mm$, and in our simulation the pressure distribution and wear profile are nearly constant when the area of area of interest is 10 $mm$ inside the boundary.

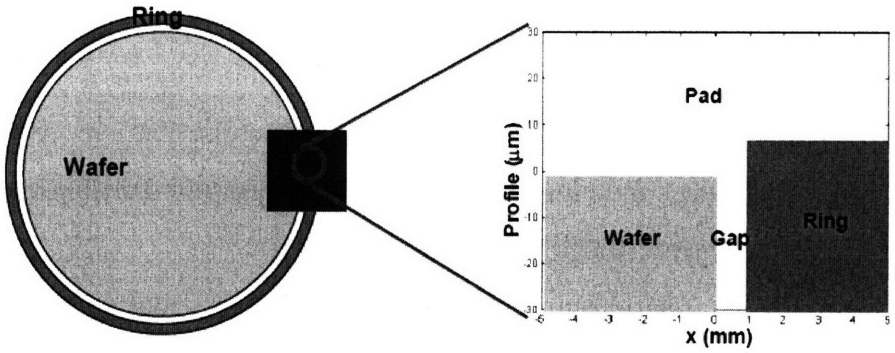



Figure 4-42: Illustration of simulated area, a 60 $mm$ by 60 $mm$ square centered at the wafer edge. We focus at the 10 $mm$ line across the gap, as shown on the right.

The contact wear model assumes a boundary condition of known contact surface; however, here the boundary condition is defined by the average pressures on the wafer

and retaining ring which are known. The relative height of the retaining ring to the wafer depends on applied pressure, the pad Young's modulus, and the gap. Therefore, the contact wear model is applied iteratively to solve the problem: a boundary profile is assumed to compute the pressure distribution, and the boundary profile is adjusted until the computed average pressures on the wafer and on the retaining ring match the specified wafer and ring average pressures.

### 4.3.2 Simulations - Static Case

To develop intuition as to how each factor influences the polishing at the edge of the wafer, we first study a set of static cases, in which we examine the pressure distribution response to different factors at one time instant. In Figures 4-43, 4-44 and 4-45, the same format is used: the gap starts at $x = 0$, the wafer is on the left of the gap and the ring on the right, and the wafer surface is set to a vertical level of zero. The wafer is assumed to be flat, and the wafer surface level drops to negative infinity at $x = 0$.

First, we vary the effective Young's modulus, and keep the gap size fixed at 1 $mm$, wafer pressure at 4 psi, and ring pressure at 5 $psi$. We simulate for three different effective Young's moduli, 2 $MPa$, 20 $MPa$, and 40 $MPa$. The pressure distribution and pad profile is shown in Figure 4-43. The pressure distributions we see the same for all values of Young's moduli, while the pad profiles are different; the softer the pad, the more the pad bends. The effective Young's modulus works as a scaling factor in the pad bending profile, without affecting the pressure distribution. Although in the static case the Young's modulus does not affect the pressure distribution, it does matter in the dynamic case, which will be discussed later.

Next, we study the relationship between the pressure distribution and gap size. We simulate for gap size of 0.5, 1.0, and 1.5 $mm$, and fix the Young's modulus at 20 $MPa$, wafer pressure at 4 psi and ring pressure at 5 $psi$. As the ring pressure is 1 $psi$ higher than the wafer pressure, the ring intrudes further into the pad than the wafer. The pad between the wafer and ring experiences two competing factors: it is pushed into the gap by the applied pressure, but it is lifted by the further intruded

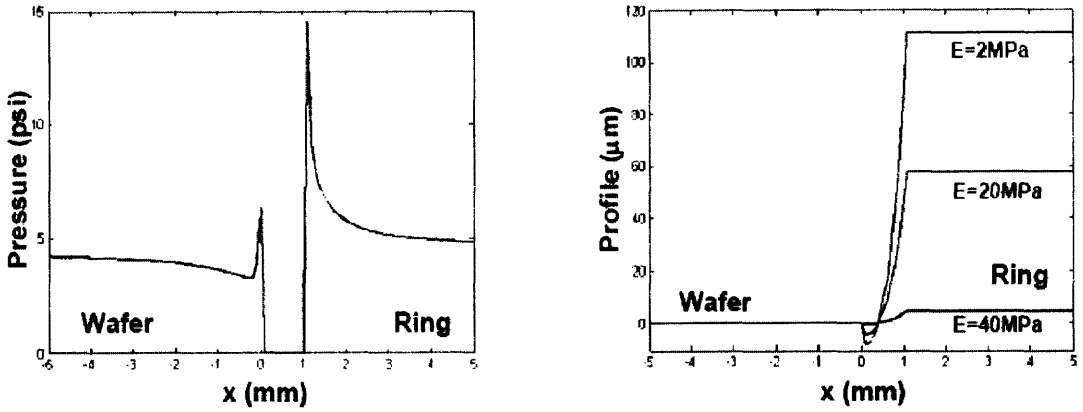



Figure 4-43: Static case to study the effect of pad Young's modulus. For all three simulations, gap is 1 *mm*, wafer pressure is 4 *psi*, and ring pressure is 5 *psi*.

ring. When the gap is large, as the case for a gap of 1.5 *mm*, the ring only lifts the part of the pad near it and most of the pad bends into the gap. As the gap becomes smaller, the pad is lifted more. When the gap is small enough, as the case for a gap of 0.5 *mm*, the pad is not touching the wafer edge and the pressure is decreasing as it comes closer to the edge.

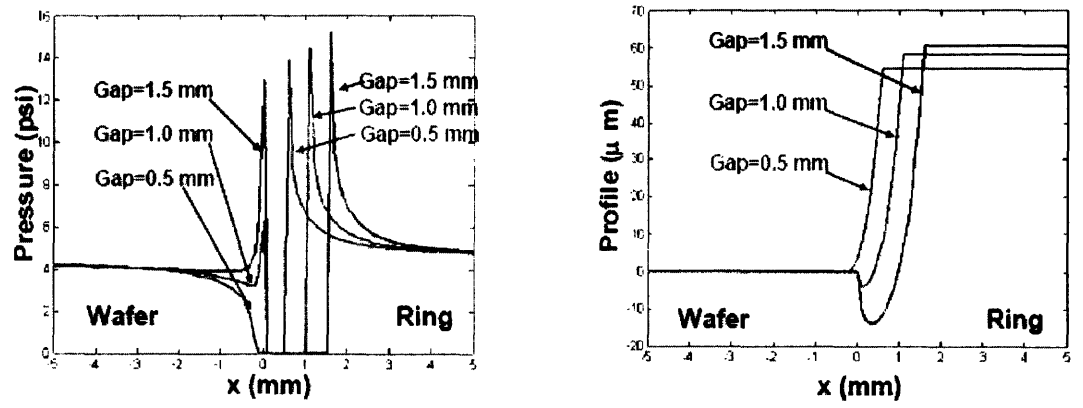



Figure 4-44: Static case to study the effect of gap size. For all three simulations, Young's modulus is 80 $MPa$, wafer pressure is 4 *psi*, and ring pressure is 5 *psi*.

Third, we focus on the influence of applied pressure. We simulate for the cases of wafer pressures of 2, 4, and 6 *psi*, and with the ring pressure 1 psi larger than the wafer pressure in each case. The effective Young's modulus is fixed at 20 $MPa$, and the gap size at 1.0 *mm*. Again we observe the competing factors of applied pressure and the gap size. As the applied wafer pressure decreases from 6 *psi* to 2 *psi*, the

contribution from the ring becomes more significant. The pad between the gap rises from being bent down to not touching the wafer edge, and as a result, the pressure distribution evolves from peaks at the edge to zero edge pressure.

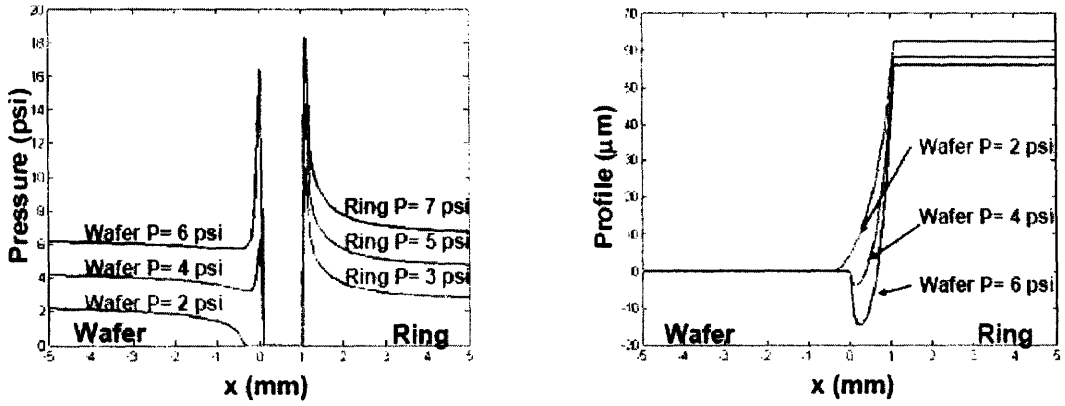



Figure 4-45: Static case to study the effect of pressures. For all three simulations, gap is 1 $mm$, pad Young's modulus is 80 $MPa$.

In summary, the simulations of the static cases show that the typical pressure distribution peaks at the wafer edge. There are two competing factors: the applied pressure pushes the pad into the gap while the ring lifts it. As a result, when gap size or applied pressures become smaller, the pressure at the edge decreases. The uniformity of the wafer pressure distribution, however, does not change monotonically with either pressure or gap size. In Figure 4-44, the wafer pressure distribution is most uniform when the gap size is 1.0 $mm$; and in Figure 4-45, the wafer pressure is most uniform when the wafer average pressure is 4 $psi$. The pressure distribution is not sensitive to the value of pad effective Young's modulus, for a flat surface.

### 4.3.3 Simulations - Dynamic Case

Unlike the static cases which focus on the pressure distribution at one time instant, dynamic cases study the time evolution of the surface profile during CMP. For all of the dynamic cases, we vary the value of one factor while keeping the other two constant. A sixty-second simulation is performed, and the graphs show the surface profiles every five seconds. The initial profile is always assumed to be flat, and

the topography variation after the sixty-second CMP is marked on the graph for comparison.

First, we simulate for different values of pad effective Young's modulus, 80 $MPa$ and 200 $MPa$, and the wafer pressure is kept at 6 psi, ring pressure at 7 psi and gap size at 1.5 $mm$. From Figure 4-46, we observe that both cases have a similar profile shape, but the stiffer pad (200 $MPa$) results in smaller roll-off value, 0.37 $\mu m$ compared to 0.62 $\mu m$ for the softer pad. Although Young's modulus is not a sensitive factor in the analysis of static cases, a stiffer pad requires less topography change to offset the non-uniform pressure distribution. Thus in the dynamic case, larger Young's modulus pads result in less edge roll-off as expected.

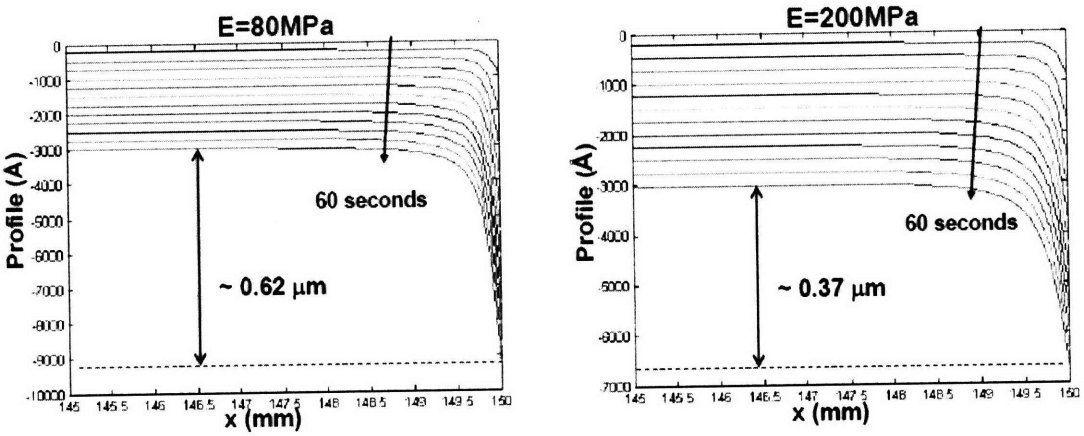



Figure 4-46: Dynamic case to study the effect of Young's modulus. For both simulations, gap is 1.5 $mm$, wafer pressure is 6 $psi$, and ring pressure is 7 $psi$.

Next we study how the surface evolution changes in response to different gap sizes, and the simulation results are shown in Figure 4-47. In the simulations, we choose gap sizes of 1.5 $mm$ and 0.5 $mm$, while the wafer pressure is kept at 6 $psi$, ring pressure at 7 $psi$, and effective Young's modulus at 80 $MPa$. In the static simulation, a large gap results in pressure peaked at the edge. Thus, we expect excessive polishing at the edge resulting in edge roll-off profiles. For smaller gap, we use different pressure values in the simulation from the ones in static cases. The dynamic case seems to be better matched with the median gap size in the static case, in which the pressure distribution has a small peak at the edge and a minimum value about 0.3 $mm$ inside. The profile

evolution in the dynamic case with smaller gap size shows a similar pattern, which is slightly over-polishing at the wafer edge, and the slowest polishing occurs about 0.3 $mm$ inside the edge.

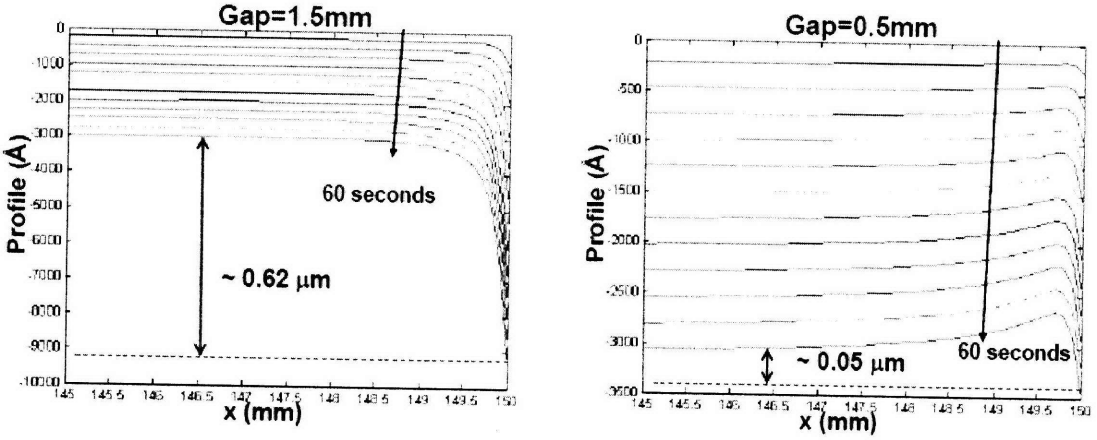



Figure 4-47: Dynamic case to study the effect of gap size. For both simulations, Young's modulus is 80 $MPa$, wafer pressure is 6 $psi$, and ring pressure is 7 $psi$.

Third, the simulations of different pressures are done by fixing the gap size at 1.5 $mm$ and the effective Young's modulus at 80 $MPa$. The ring pressure is always kept 1 $psi$ higher than the wafer pressure, and the two simulations use wafer pressure 6 $psi$ and 2 $psi$. The larger pressure results in excessive edge over-polishing, which is consistent with the static simulation result. The smaller pressure results in slight over-polishing at the edge and the slowest polishing taking place about 0.5 $mm$ inside, which seems to match the median pressure case in the static simulations.

These dynamic simulations suggest that: the larger the gap size, the faster the edge polishes; the higher the pressure, the faster the edge polishes; and the stiffer the pad, the more uniform the wafer polishes. A combination of large gap size, large applied pressure, and small Young's modulus will result in fast edge-polishing evolution, which is the case in Figure 4-46 (left). A combination of small gap size, small applied pressure and small Young's modulus will result in a slow edge polishing. We choose gap size 0.5 $mm$, wafer pressure 2 $psi$, ring pressure 3 $psi$, and Young's modulus 80 $MPa$, and the result is shown in Figure 4-49 (left). A uniform polishing can be achieved by a combination of median gap size, median pressures, and large

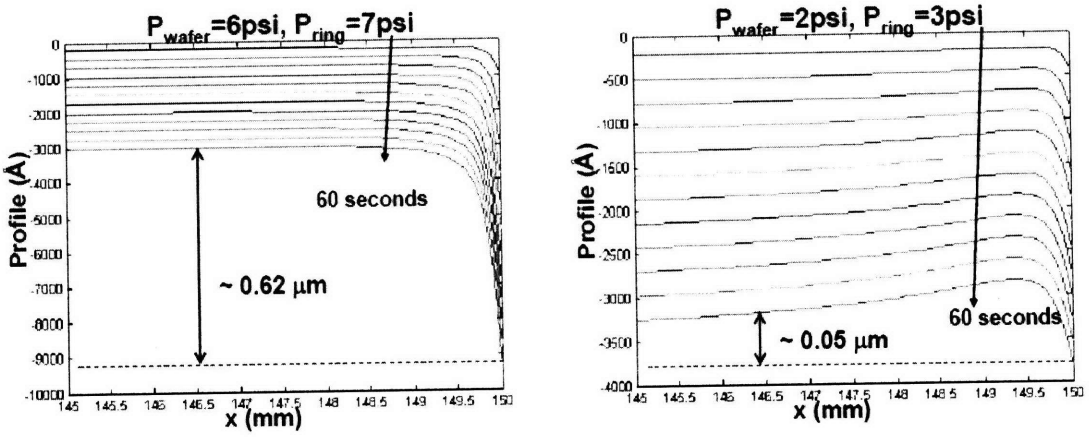



Figure 4-48: Dynamic case to study the effect of pressures. For both simulations, gap is 1.5 $mm$, and Young's modulus is 80 $MPa$.

Young's modulus. Figure 4-49 (right) shows the simulation result for gap size 1.0 $mm$, wafer pressure 4 $psi$, ring pressure 5 $psi$, and Young's modulus 200 $MPa$.

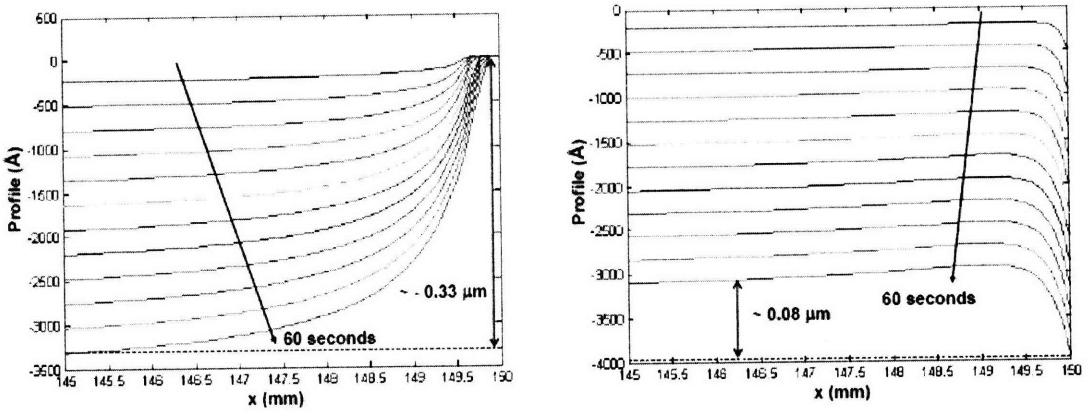



Figure 4-49: Left plot shows simulated surface evolution using a slow edge polishing setup: Young's modulus is 80 $MPa$, gap size is 0.5 $mm$, wafer pressure is 2 $psi$, and ring pressure is 3 $psi$. Right plot shows simulated surface evolution using a more uniform polishing setup: Young's modulus is 200 $MPa$, gap size is 1.0 $mm$, wafer pressure is 4 $psi$, and ring pressure is 5 $psi$.

In Figure 4-50, the measured thickness profiles near the wafer edge are compared with simulated profiles after 60-second polishing. The simulated results capture the range of observed variations in measured edge roll-off profiles.

We have studied how the edge roll-off profiles can be generated from the CMP process; the next question is how the initial topography profile will affect the perfor-

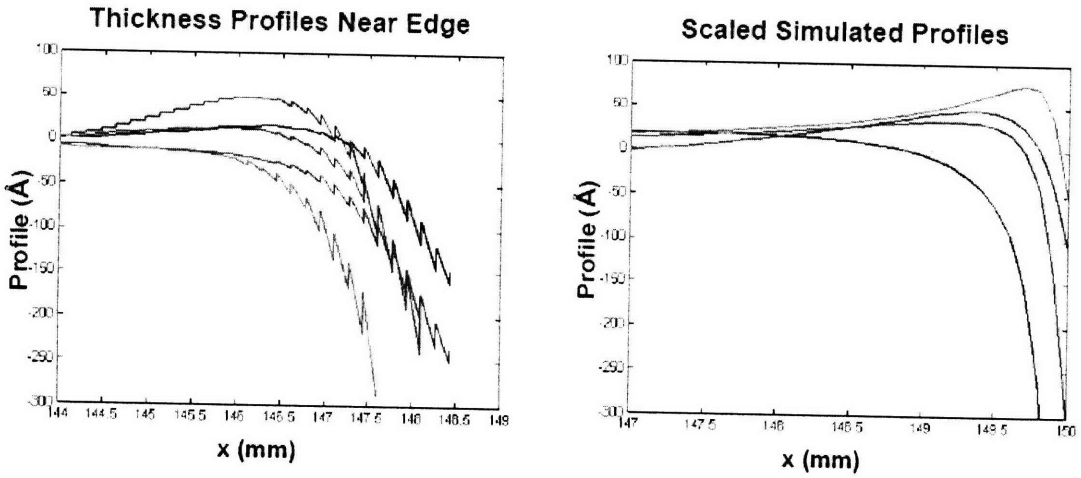



Figure 4-50: Left plot shows the measured thickness profiles near the wafer edge, and the right one shows the simulated profile after 60 – *second* CMP. The simulated profiles are able to capture different types of observed behaviors in measured edge profile. Note: the profile shows zig-zag artifacts due to resolution limits that are not present in the actual surface.

mance of later CMP. For a flat initial surface, as we studied earlier, a combination of median gap size, median pressures, and stiffer pad will give a nearly uniform polishing.

To study the impact on CMP of a wafer with a starting edge roll-off profile, we will consider two sets of CMP parameters. Both cases have the same effective Young's modulus, but the first set has higher pressures and larger gap. As discussed earlier, the first set should result in faster edge polishing, and the simulation results do indeed show a similar result as shown in Figure 4-51. The first case shows the surface evolving from a starting roll-off to a final profile with a similar edge roll-off pattern. The second case polishes slower on the edge, and the surface becomes flatter during CMP, as seen at right in Figure 4-51. Although the second case reduces the surface non-uniformity, its removal amount is more non-uniform than that of the first one (Figure 4-52.)

Thus, if the objective is to improve the surface flatness, (e.g., as preparation for lithography process), it is better to match a starting edge roll-off profile with a slow edge polishing setup. If the objective is uniform polishing of a surface film for example, then it is better to match a starting edge roll-off profile that rolls "down" near the edge with a fast edge polishing setup. Similarly, for an initial edge "roll-up"

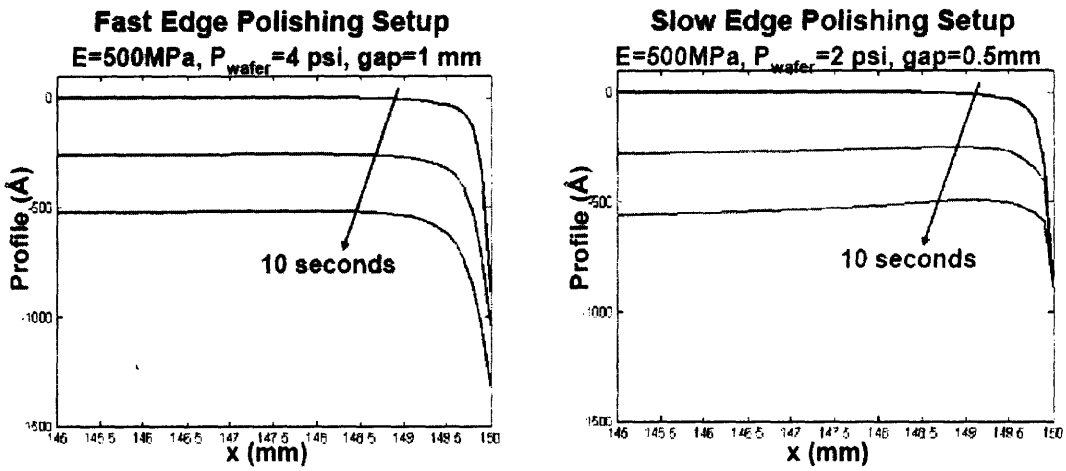



Figure 4-51: Dynamic simulations starting from initial edge roll-off profiles. The left plot shows simulation using a fast edge polishing setup, and the surface shape does not change significantly. The right plot shows simulation with a slow edge polishing setup, and the topography non-uniformity is reduced.

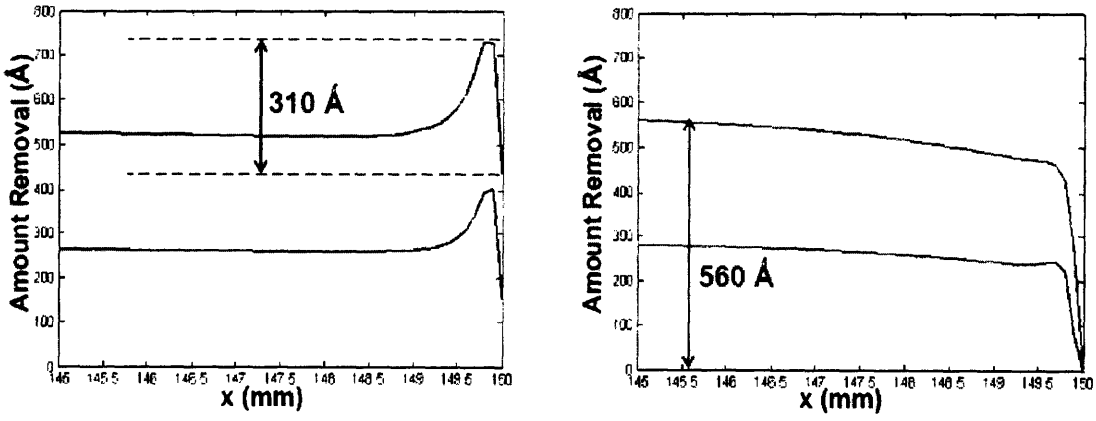



Figure 4-52: Same dynamic simulations as performed in Figure 4-51, but the amount removed is shown.

profile, it is better to use a slow edge polishing setup to achieve more uniform thin film polishing, as shown in Figure 4-53 and 4-54.

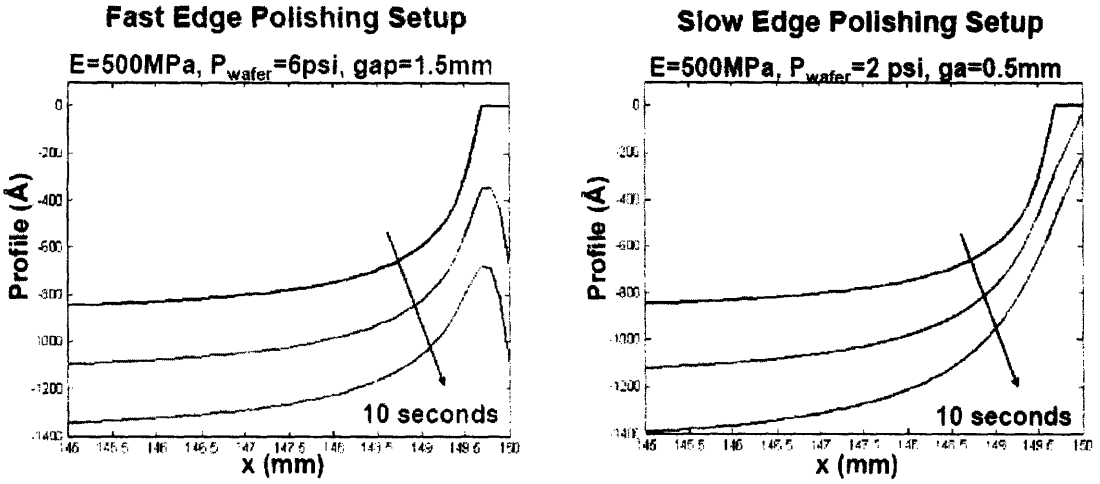



Figure 4-53: Dynamic simulations starting from an initial edge "roll-up" profile. The left plot shows simulation using a fast edge polishing setup, and the topography non-uniformity is thus reduced. The right plot shows simulation with a slow edge polishing setup, and the surface shape does not change significantly.

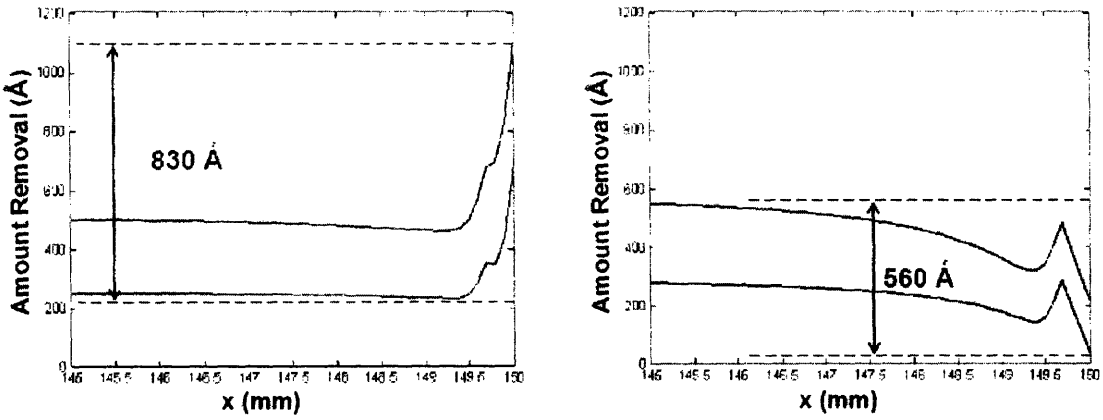



Figure 4-54: Same dynamic simulations as performed in Figure 4-53, but the amount removed is shown.

### 4.3.4 Conclusion

In this section, we studied how a wafer edge roll-off profile can be generated during CMP and how existing edge roll-off affects further CMP process steps. A contact

wear model is applied to simulate the CMP process near the wafer edge with reasonable approximation. The three factors studied are the effective Young's modulus of polishing pad, gap size between wafer and retaining ring, and the pressures on the wafer and the ring. The simulation results suggest better edge uniformity can be achieved by using stiffness tailored pads, median gap size, and median pressures. Edge polishing is also influenced by the initial surface profile. Better matching of the starting profile and tool set, such as edge roll-off profile with fast edge polishing tool setup, will provide more uniform polishing. In future work, obtaining experimental data on pre- and post-CMP can help to test and calibrate the model predictions.

## 4.4 Endpoint Detection

Frictional monitoring, or motor current endpoint detetion (EPD), has recently emerged as a feasible candidate for Shallow Trench Isolation (STI) Chemical Mechanical Planarization (CMP), and an effective EPD system has the potential of improving yield, increasing throughout, reducing wafer-to-wafer variability and improving planarity [97]. It is suspected that frictional effects generated by the pattern structures and various layered materials on the wafer will create distinct and characteristic responses for determining an appropriate endpoint. After noise filtration, the motor current signal shows distinctive characteristics. The purpose of this section is to provide a friction model to understand the motor current characteristics based on our STI CMP step-height pattern-density model [46] [1], and to provide insights for improving EPD.

In the following sections, first, we explain the endpoint in STI CMP; second we review the STI endpoint detection experiment setup and results; then we present the friction model of EPD motor current. After a brief review of our STI CMP step-height pattern-density model, we present the simulation results of EPD motor current, followed by a conclusion.

### 4.4.1 Endpoint in STI CMP

CMP is used in STI to remove the overburden oxide on the raised area, and the endpoint is defined to be the time when the overburden oxide is cleared completely

across the die and the wafer. The dependence of removal rate on pattern-density results in non-uniform polishing across the die, and the surface evolution during an STI CMP process is illustrated in Figure 4-55. Figure 4-55 (b) shows the cross-sectional diagram when the oxide starts to be cleared in low pattern-density areas but still remains in high pattern-density areas, also referred to as the "start" of the endpoint. Figure 4-55 (c) illustrates the completion of the endpoint interval when the oxide is cleared completely. In practice, the wafer is often over-polished beyond the ideal "endpoint" when the last position on the die has cleared to ensure a complete clearing of oxide across all dies on the wafer, but the over-polishing usually causes undesirable oxide dishing, nitride erosion, and topography variation, as shown in Figure 4-55 (d).

### 4.4.2 STI Endpoint Detection Experiment Setup and Results

The STI endpoint detection experiment involved endpoint characterization of reverse mask processed STI patterned wafers of four reticle sets (A through D). Wafers were 150 *mm* in diameter. All wafer sets began with 100 Å of a thermally grown pad oxide on a p-type silicon substrate. This was followed by a 2500 Å silicon nitride deposition for sets A and B, and a 1500 Å silicon nitride deposition for sets C and D. Sets A and B were then patterned and etched to obtain a trench depth of 5000 Å. Nominal trench depth for sets C and D was 3100 Å. All wafers were then subjected to a sidewall oxide layer growth of 250 Å via dry oxidation. This was followed by plasma-enhanced chemical vapor deposition of TEOS oxide for trench fill. Structural characteristics of the four sets are summarized in Figure 4-56. Reticles used for these sets provided an adequate range of mean oxide pattern-density (approximately 11.7 to 37.3 percent). This allowed for a thorough motor current endpoint investigation.

Luxtron's Optima 9300 CMP endpoint system was utilized for detecting motor current signals from the polishing tool. As motor current signals were detected during polish, the real-time controller (RTC) of the Optima 9300 allowed the system to record and recognize endpoint conditions as they occurred on the polishing tool. Once the system detected an endpoint, a relayed signal was sent to the polishing tool via a

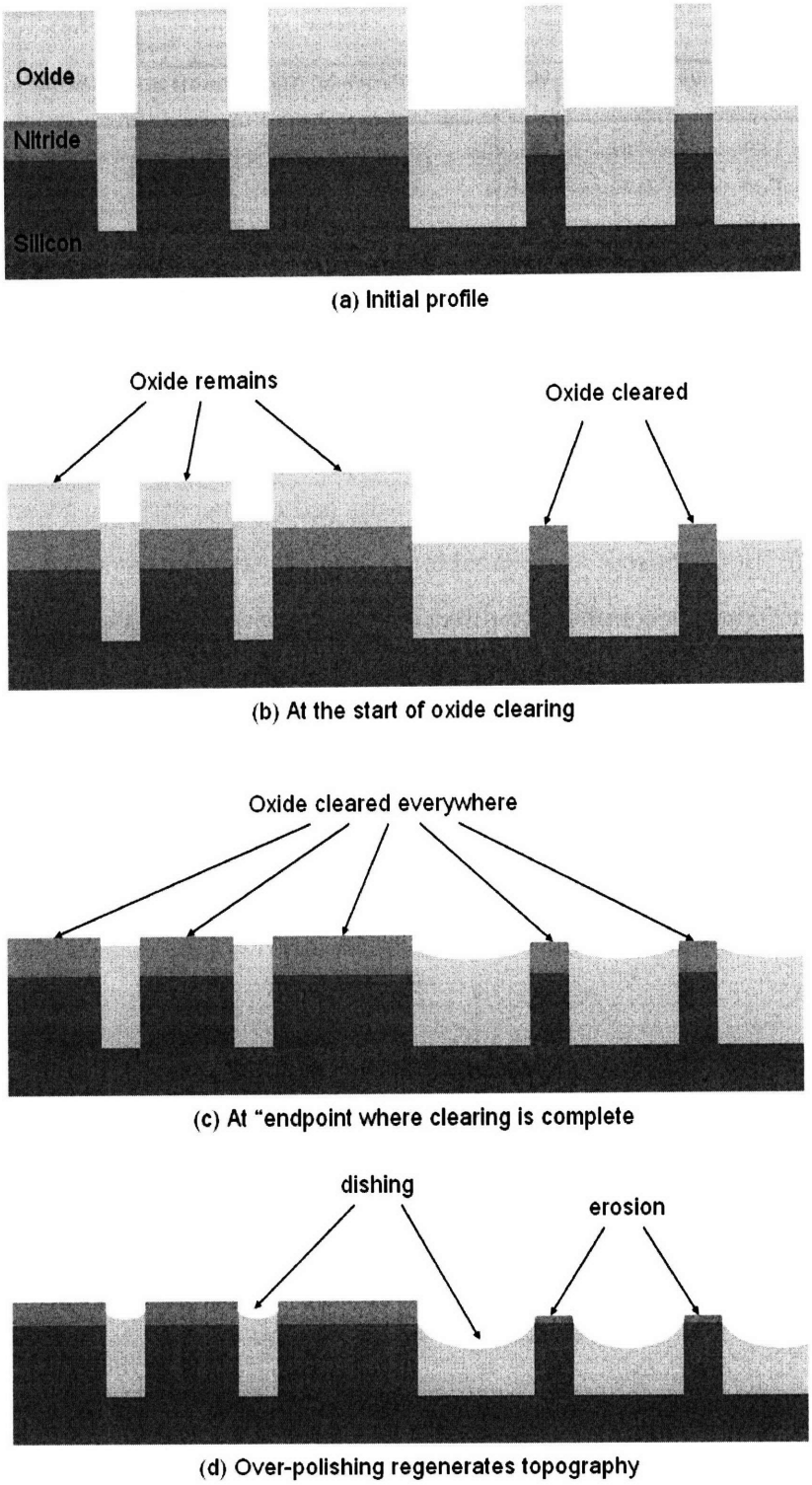



Figure 4-55: Illustration of the surface evolution and the endpoint of an STI CMP process: (a) initial profile; (b) at the start of oxide clearing; (c) at the "endpoint" where clearing is complete across the die and across the wafer; and (d) over-polishing which results in oxide dishing and nitride erosion, as well as the regeneration of topography.

| Reticle Set | Oxide (%) | | | | Nitride (%) | | | | Trench Depth (Å) | TEOS Trench Fill (Å) |
|---|---|---|---|---|---|---|---|---|---|---|
| | Density Variation | Max | Min | Mean | Density Variation | Max | Min | Mean | | |
| A | 13.8 | 19.4 | 5.6 | 11.7 | 95.8 | 98.8 | 3.0 | 36.2 | 5000 | 9000 |
| B | 17.4 | 26.4 | 9.0 | 19.0 | 35.5 | 53.3 | 17.8 | 28.4 | 5000 | 9000 |
| C | 15.9 | 30.7 | 14.8 | 24.3 | 59.1 | 86.1 | 26.9 | 41.3 | 3100 | 5900 |
| D | 25.2 | 48.4 | 23.2 | 37.3 | 62.2 | 87.3 | 25.1 | 41.4 | 3100 | 5900 |

Figure 4-56: STI patterned wafer information [18].

serial I/O interface [98]. The Optima 9300 was supplied with four current sensors, of which only two were used. One sensor was attached to the carrier-head and the other to the platen. Both sensors were capable of acquiring current signals at a frequency of 10 $Hz$, and could accommodate direct or alternating currents of up to 25 $A$. The sensors were also capable of detecting current changes of less than 2 $mA$. Figure 4-57 shows a block diagram of the set-up. Motor current signals were taken in three modes: platen signal alone (Channel A), platen and carrier-head signals (Channel A + Channel B), and the ratio of the platen to the carrier-head signal (Channel A / Channel B).

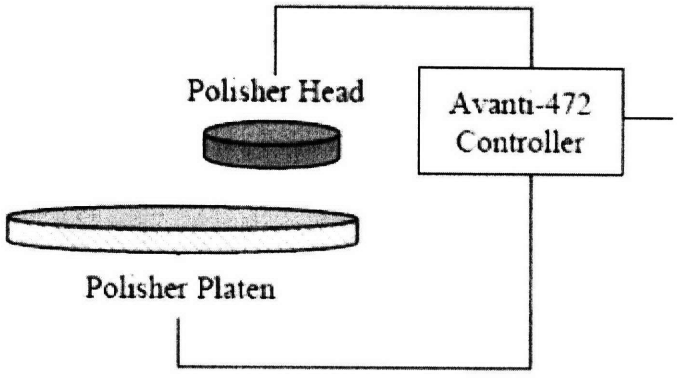



Figure 4-57: Block diagram of the Avanti-472 polisher with the Optima-9300EPD system.

Analysis of various signals indicated that Channel A (the platen motor current) provided the most distinctive signal change for endpoint detection. The other remaining signals (i.e. Channel A + Channel B and Channel A / Channel B) yielded

unusable results due to the lack of distinguishable curves. The current signal from Channel A shows a distinctive pattern as in the example of Figure 4-59a.

### 4.4.3 STI Endpoint Motor Current Model

We develop our endpoint detection signal model in two steps: first we explore the relationship between EPD motor current with frictional force, and then we present a model for the frictional force during STI CMP. We can view the carrier-head and platen as a system. The power injected into the system is from Channel A and Channel B currents, and the input power will dissipate through friction. We separate the total dissipation related to the frictional force between carrier-head and platen from other energy sinks, which can be viewed as approximately constant during CMP.

We first concentrate on the friction between carrier-head and platen. Let $\vec{f}(x,y)$ denote the friction force that the platen exerts on carrier-head at each location, $\vec{r}(x,y)$ denote the displacement from the center of the carrier-head to each position, and $\vec{R}(x,y)$ denote the displacement from the center of the platen to each position, which are illustrated in Figure 4-58. The torque due to friction with respect to the center of the carrier-head is

$$\tau_{carrier} = \int_{wafer} \vec{r}(x,y) \times \vec{f}(x,y) \cdot ds(x,y), \tag{4.5}$$

and the torque with respect to the center of the platen is

$$\tau_{platen} = -\int_{wafer} \vec{R}(x,y) \times \vec{f}(x,y) \cdot ds(x,y) \tag{4.6}$$

The friction $\vec{f}(x,y)$ between two materials depends on the materials, surface roughness, wetting, and relative velocity. For matched rotational velocity of the carrier-head and the platen, the relative speed of the head to the platen at any position on the wafer is of both the same magnitude and direction. The other factors at one point should only depend on its location in each die, i.e., we can write friction as a function of die coordinates $\vec{f}(\xi,\eta)$ . We rewrite the expressions for the carrier torque, with $\vec{r}(\xi,\eta)$ as the position of each point in its die coordinates and $\vec{r}_i$ as the

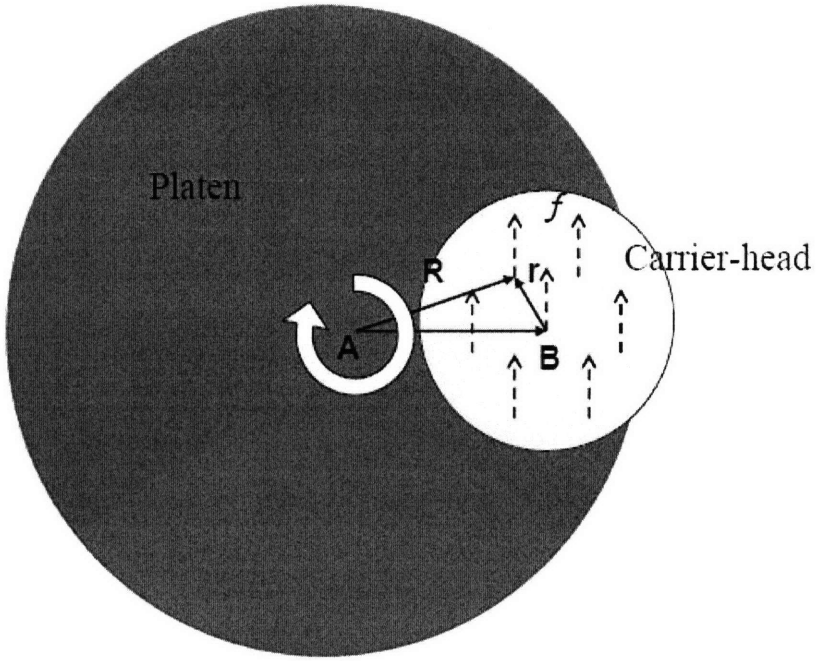



Figure 4-58: Illustration of friction force between platen and carrier-head. Here $f$ denotes the friction force exerted on the platen by the carrier-head; $R$ denotes the displacement from each point to the center of the platen A; and $r$ denotes the displacement from each point to the center of the carrier-head B. As small dies are symmetrically distributed on the wafer, the net torque with respect to B is almost zero, while all small torques add up to contribute to the net torque with respect to A.

displacement from the center of carrier-head to the coordinate origin of the $i^{th}$ die.

$$\tau_{carrier} = \sum_{cells} \int_{i^{th}\ cell} (\vec{r}_{cell}(\xi,\eta) + \vec{r}_i(x,y)) \times \vec{f}(\xi,\eta) \cdot ds(\xi,\eta) \tag{4.7}$$

We can set $\int_{i^{th}\ cell} \vec{r}_{cell}(\xi,\eta) \times \vec{f}(\xi,\eta) \cdot ds(\xi,\eta) = 0$, by properly choosing the origin of die coordinates, and we have $\tau_{carrier} = (\sum_{cells} \vec{r}_i(x,y)) \times \int \vec{f}(\xi,\eta) \cdot ds(\xi,\eta)$. On a 150 $mm$ wafer, there are usually hundreds of small dies nearly symmetrically distributed on the wafer, thus $\sum_{cells} \vec{r}_i(x,y) \approx 0$ and $\tau_{carrier}$ contains little information of friction force. In contrast, $\tau_{platen} = (-\sum_{cells} \vec{R}_i(x,y)) \times \int \vec{f}(\xi,\eta) \cdot ds(\xi,\eta)$ has all the contributions adding up, as illustrated in Figure 4-58. Hence, only motor current from Channel A contains information about the frictional forces which can serve as an EPD signal.

During the CMP process, two factors cause the change of friction force: the exposure of nitride layer and CMP induced topography variations. To characterize the friction force between platen and carrier head, we assume the frictional force is linearly proportional to the product of the averaged material frictional coefficient $\mu_{avg}$ and the long range standard deviation of raised area $\sigma_{Long\ Range}$, and we can write the frictional force as:

$$f \sim \mu_{avg} \cdot (1 + \beta \cdot \sigma_{Long\ Range}). \tag{4.8}$$

The averaged material friction $\mu_{avg}$ is based on the relative portion of different materials on the wafer surface, and is calculated by averaging the friction coefficient weighted by that material's exposed area. At the beginning of an STI polish, the exposed surface is generally 100% silicon dioxide; at exposure and clearing of the active areas, 50% or more of the exposed surface will be silicon nitride. When computing $\sigma_{Long\ Range}$, we first discretize the die into cells with size approximately equal to the planarization length, and calculate the standard deviation of average raised height of each cell. $\beta$ is a model parameter, and we take $\mu_{nit} - \mu_{ox}$ also as a model parameter, as the friction coefficients of oxide and nitride are very sensitive to slurry used [99].

### 4.4.4 Simulation and Discussion

The surface topography information can be obtained using the die-level CMP models, which are extensively discussed in the previous chapter. The STI CMP process can be separated into four steps in response to the evolvement of frictional force (Figure 4-59b). Before CMP, the raised areas of the deposited oxide have a uniform height across the wafer. As CMP starts to remove the raised oxide, a low pattern-density area is polished faster compared with the more dense area. In this first step, non-uniformity created by CMP increases the die-level wafer surface roughness. During the second step, as the step-height becomes smaller, the polish rate on the raised area begins to decrease, and the faster the area polishes, the earlier it starts to decrease its raised area removal rate. As a result, the non-uniformity declines as all the polish rates converge to the blanket polish rate. The third step starts when the fast-polishing area hits the nitride layer, and its removal rate is further decreased four- to five-fold due to slow nitride polishing. Thus, the surface roughness is expected to reduce several-fold in magnitude; however, as the nitride gets exposed, the average friction coefficient is changing too. As the experiment suggested, the net result is a slow decrease in friction force. In the last step when all raised oxide areas have finally cleared, the average friction coefficient becomes a constant again, while the nitride polishing starts to increase topographic non-uniformity slowly again, due to the dependence of steady state removal rate on pattern-density of the nitride layer. At this point, the friction trace begins to rise again, and is typically taken as the "endpoint signal" indicating that all oxide over active regions have been cleared.

The simulation is performed for the set of patterned wafers A to D, and we use the same initial structure parameters as in the experiment. For patterned wafers A and B, we assume an initial oxide layer of 9000 Å and a nitride layer of 2500 Å, while for patterned wafers C and D, we assume an initial oxide layer of 5900 Å and a nitride layer of 1500 Å. As for initial step-height, we assume 7000 Å for wafers A and B, and 4500 Å for C and D. The oxide and nitride layout pattern density distribution is an estimate based on the effective pattern density histogram distribution, as in Figure 4-

63, and the effective pattern density is calculated using an elliptical weighting function with a planarization length of 3.5 $mm$ [100]. For STI CMP step-height model, we use a blanket oxide removal rate of 3000 $\text{Å}/min$, and an oxide to nitride selectivity of 4. For the friction force model, we choose $\beta = 0.001\ \text{Å}^{-1}$ and $(\ \mu_{ox} - \ \mu_{nit})/\ \mu_{ox} = 12\%$, such that the simulated motor currents are close to measurement results.

The simulated results for each patterned wafer are drawn in Figures 4-59 through 4-62, along with the measured signal to compare. The simulated plots agree reasonably well with the experiment signals, and show the distinctive curves of the four steps. The estimated oxide clearing time, which marks the beginning of step 3, i.e., when oxide starts to be cleared, to the end of step 3, i.e., when all oxide gets cleared, is also listed with each plot as a reference. On each plot, the endpoint time is also indicated (EP), based on a simple detection of an increase in friction signal.

For wafers A to C, the simulated friction forces continue to decrease until all oxide is cleared, which indicates an effective endpoint detection. For wafer D, however, the rising of the motor current occurs during step 3, which results in an early detection.

The failure of endpoint detection of wafer D prompts us to study the limit of the EPD mechanism. We compare the pattern wafer C with D, as they have the same film thickness structure parameters. Their largest difference lies in their oxide pattern-density distributions, as shown in Figure 4-63. Wafer C has a center-peaked oxide pattern-density range from 0.15 to 0.29, while the oxide pattern-density of wafer D is very similar to that of wafer C, except a fat tail from pattern density of 0.29 to 0.38. As the larger pattern-density area polishes slower, the fat tail of wafer D results in a slow decreasing rate of roughness, after the nitride layer starts to be exposed. It is the slow roughness decrease compared with the rate of nitride exposure that leads to an early erroneous detection of endpoint.

Erroneous late detection of endpoint is also possible when the oxide pattern-density is inversely correlated with the nitride pattern-density, i.e., when for each point, the higher the oxide pattern-density, the lower is the nitride pattern-density.

Right after all oxide gets cleared, the more dense the nitride pattern-density, the

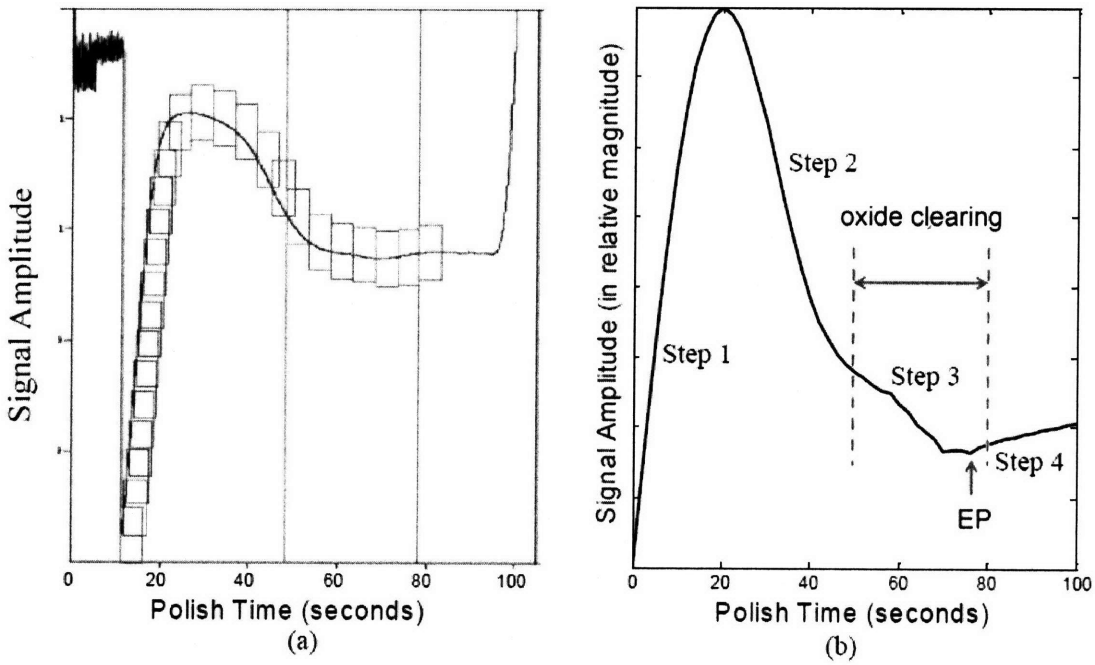



Figure 4-59: EPD motor current of patterned wafer A. (a) measured signal with an applied endpoint recipe; (b) predicted by friction model, estimated oxide clearing time from 49-80 seconds.

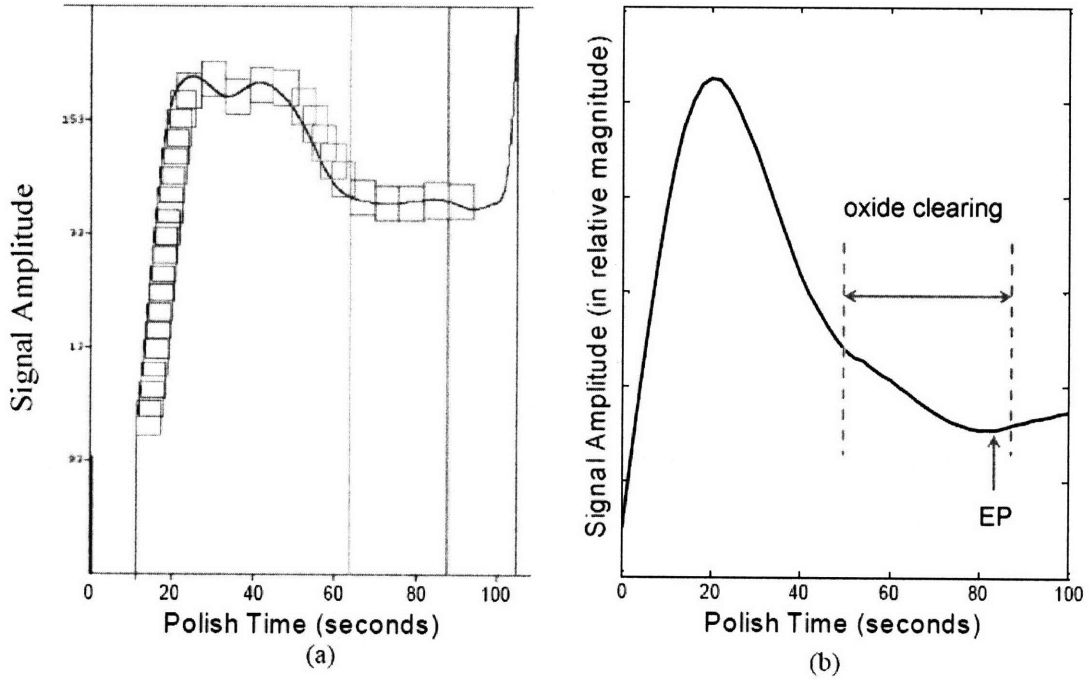



Figure 4-60: EPD motor current of patterned wafer B. (a) measured signal with an applied endpoint recipe; (b) predicted by friction model, estimated oxide clearing time from 50-86 seconds.

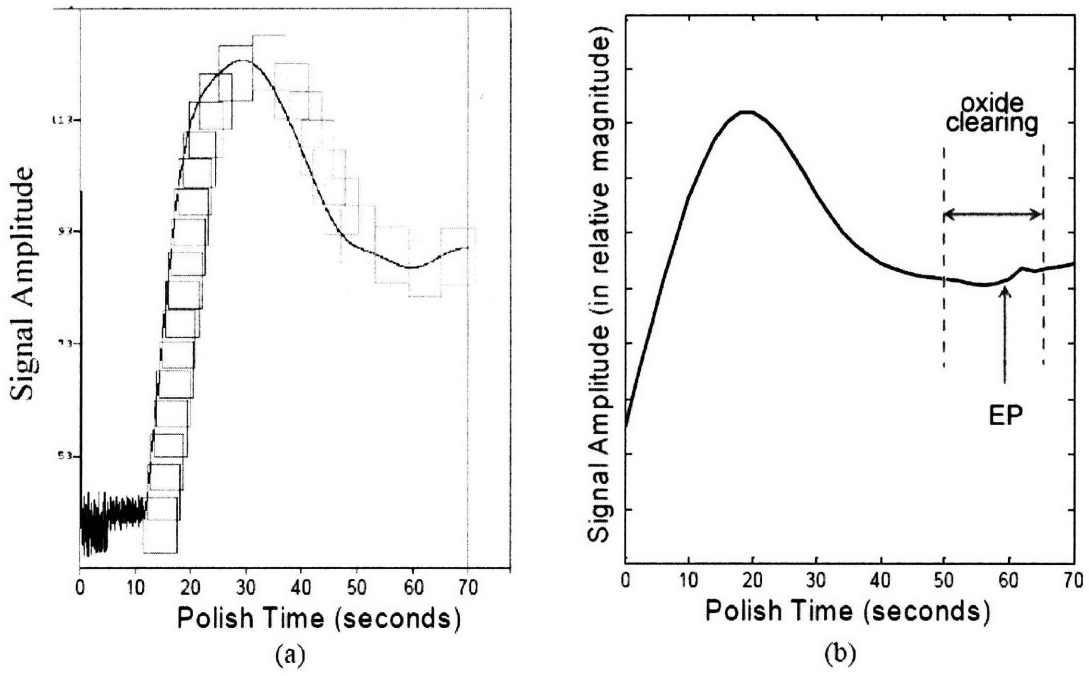



Figure 4-61: EPD motor current of patterned wafer C. (a) measured signal with an applied endpoint recipe; (b) predicted by friction model, estimated oxide clearing time from 50-65 seconds.

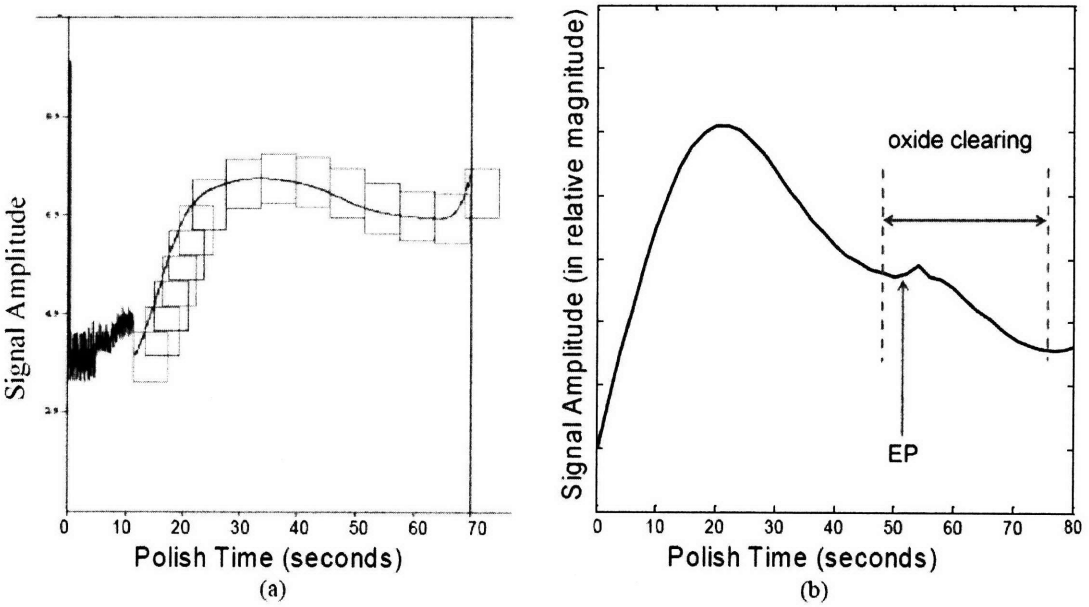



Figure 4-62: EPD motor current of patterned wafer D. (a) measured signal with an applied endpoint recipe; (b) predicted by friction model, estimated oxide clearing time from 48-75 seconds.

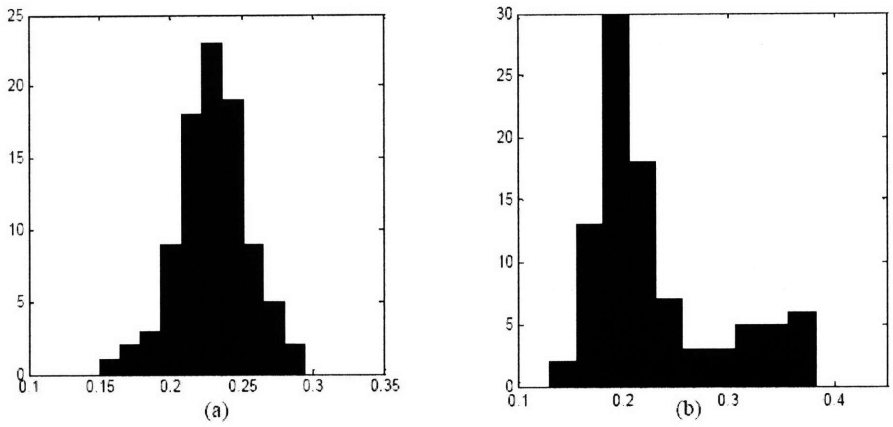



Figure 4-63: Oxide pattern-density distribution histogram of (a) wafer C and (b) wafer D.

slower it polishes, but the same point has a low oxide pattern-density and it has been polished more before. Thus, the polish will carry a momentum to reduce roughness before finally increasing it, as shown in Figure 4-64.

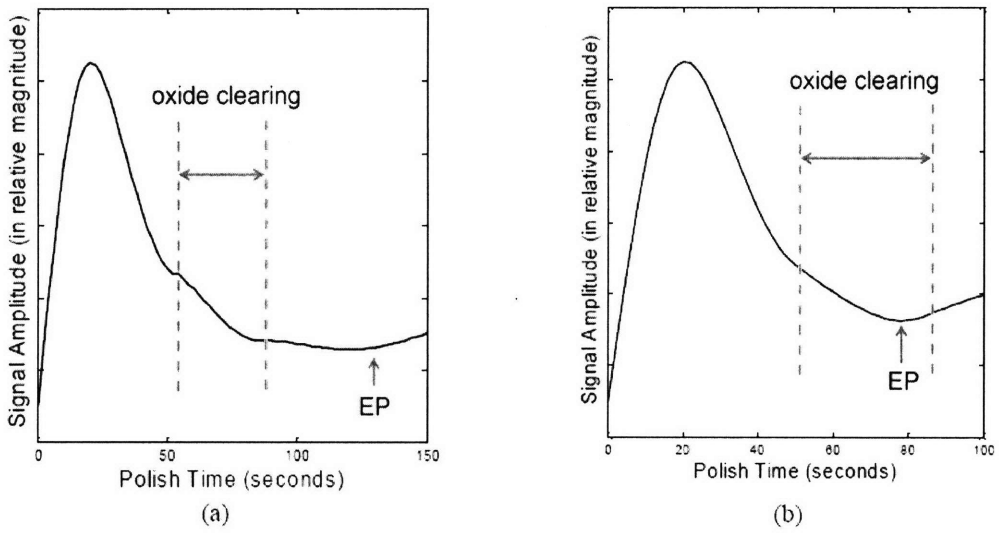



Figure 4-64: Late detection of endpoint due to nitride pattern-density being inversely correlated with oxide density (a), as compared with the case when positively correlated (b). The correct estimated oxide clearing time is 86 seconds.

### 4.4.5 Applying the Friction Model in Endpoint Detection

In the above discussion, the friction model has been used to explain the motor current trace measured during CMP. The friction model relates the friction force to the surface topography and the material exposure on the surface, and the topography evolution during CMP can be estimated using the die-level CMP model. Based on the friction model, the first sharp rise and fall of the motor current are caused by the topography evolution during the polishing of the overburden oxide. When the oxide on raised areas starts to be cleared, the topography variation is reduced and the nitride layer starts to be exposed, which affects the average friction coefficient. And finally when the overburden oxide is completely cleared, the non-uniform removal rate of polishing nitride slowly increases the topography variation. The friction model provides support to the typical practice which uses the second rise of the motor current as the endpoint signal for STI CMP.

The friction model can also used to assist the detection of endpoint. The second rise in the motor current trace, unfortunately, does not always correctly mark the endpoint, and in practice a ten-second over-polishing is typically used to ensure a complete clearing of the overburden oxide. This approach has disadvantages: a long over-polishing time can cause excessive oxide dishing and nitride erosion, and increase the risk of clearing the nitride layer and degrading the silicon surface; and a short over-polishing may fail to clear oxide in some areas, especially the in case of an early detection. Hence, an accurate detection of the endpoint is desirable, and the friction model can help to achieve the goal. For each layout design, the die-level CMP model can be used to simulate the topography evolution, and then the friction model can be used to predict the motor current trace and the correct feature in the trace (such as ten seconds after the second rise) to use as the endpoint. In this way, the motor current can be used more effectively and more accurately to determine the endpoint in STI CMP.

### 4.4.6 Conclusion

In this section, we present a simple friction model to explain the endpoint motor current based on the STI CMP step-height pattern-density model. The model confirms several observations of the EPD experiment, and the simulated signals of four patterned wafers agree reasonably well with the distinctive characteristic EPD motor currents measured. Based on our model, we further study the situation in which the EPD mechanism may yield false endpoint resulting in early detection or late detection. It needs to be emphasized that in the simulation we use approximate model parameters of the STI CMP process, as well as a rough estimation of oxide and nitride pattern-density distribution. The friction model can also be further developed and calibrated through specific designed experiments for a given STI CMP process. The use of the pattern dependent STI CMP process model can help relate observed motor current signals to the correct state of the wafer, and can prevent early or late endpoint detection.

## 4.5 Summary

This chapter describes several applications of die-level, pattern-density, and contact wear CMP models. The methodology of applying the die-level CMP models to any random chip layout is reviewed as a starting point, and a new STI mask design is described as an illustration of a test wafer for model calibration. Then, the models function as a virtual fab to provide instant feedback on any layout design, enabling the DFM methodology to make the chip design more fab-friendly.

The impact of nanotopography of blanket and patterned wafer polishing is assessed using experiments to study the variation caused by nanotopography, compared with process induced variation, run-to-run variation, wafer-level variation, and with-in-die variation. The nanotopography impact of blanket wafer polishing can be modeled using the contact wear model, and the prediction agrees well with experiment data. In pattern wafer polishing experiment, the nanotopography induced variation is statistically significant, although it only accounts for 4% of the total variation observed

in a DRAM product-level e-test parameter.

The contact wear model is also applied to study the polishing at the wafer edge, near the retaining ring of the wafer holder. The pressure profile near the wafer edge is affected by the configuration of wafer and retaining ring, as well as pad stiffness. Both the pressure distribution and surface evolution are simulated for different parameters to illustrate the dependence, and the potential for optimization to reduce edge nonuniformity.

The endpoint signal used in STI CMP is the motor current, which is shown to be directly linked with the friction force between the wafer and the pad. A friction model is proposed, where the friction is assumed to be proportional to surface roughness and fractions of exposed material types. The time evolution of the wafer surface profile can be obtained using the die-level CMP model, and the friction predictions agree with measured motor current traces. The integration of die-level pattern models with endpoint detection offers the possibility of improving the accuracy and applicability of friction-based endpoint to a wide range of product layouts.

# Chapter 5

# Conclusions

This thesis focuses on the physical understanding of the polishing and planarization mechanisms of CMP in dielectric materials, and establishes a framework for understanding and modeling CMP by analyzing the detailed structure and various interactions in the complex system of CMP. The framework has been used to develop particle-level and die-level CMP models, which are compared with experimental evidence. The die-level models are applied to engineering problems including nantopography impact, wafer edge roll-off, and motor current signals used for endpoint detection.

## 5.1 Thesis Contributions

The contributions of the thesis can be separated into three areas: understanding the physics of CMP and developing a particle-level model of CMP, modeling the planarization on the die-level, and applying the die-level CMP models to practical engineering challenges.

### 5.1.1 Understanding the Physics of CMP of Dielectric Materials

One key contribution to the physical understanding of CMP is the establishment of the modeling framework in Section 2.4.1, which is based on a detailed analysis of the complex CMP system. In the framework, the CMP process is decomposed into the following interactions: the pad-wafer interaction, the abrasive-contact-area dynamics,

the abrasive-pad interaction, the abrasive-wafer interaction, and the chemical reaction. These interactions occur at different spatial scales, and therefore they can be approached separately from empirical study or theoretical analysis. A particle-level model of dielectric CMP is proposed in Section 2.4.7 by studying all of the individual interactions:

- The material removal mechanism is studied in Section 2.4.2, and the chemical-tooth mechanism is adopted for modeling dielectric CMP based on the observed strong dependence of removal rate on abrasive material, and relatively weak dependence of polished wafer surface roughness on abrasive size.

- The abrasive-wafer interaction is studied in Section 2.4.3, and a Herzian contact is assumed to solve the contact area and the contact pressure as functions of the abrasive size and the pressure on the abrasive.

- The pad-abrasive interaction is studied in Section 2.4.4, and the contact wear model is used to study the dependence of the abrasive pressure on abrasive size distribution and the abrasive concentration in the contact area between pad asperity and wafer.

- The pad-abrasive-wafer interaction is discussed in Section 2.4.5 to estimate the relative velocity of abrasives to the wafer and the abrasive concentration in the pad-wafer contact areas.

- The pad-wafer interaction is studied in Section 2.4.6, and the Greenwood-Williamson model is used to estimate the contact area and the pressure distribution in the contact areas between the wafer and the pad with surface asperities having exponential height distribution.

The predicted dependence of removal rate on various input variables is discussed in Section 2.4.8, and most of the predictions agree with observations from experiments.

### 5.1.2 Developing Die-Level CMP Models

Section 3.2.1 identifies the objectives of die-level CMP models as developing the dynamic equations of removal rate $K_{u,d}(\rho, z_u, h)$, and the relationship between the planarization performance and physical properties of consumables. A modeling framework has been adopted in Section 3.2.2, in which the modeling of $K_{u,d}(\rho, z_u, h)$ has been broken down into the dependence of removal rate on pressure $K(P)$ and the dependence of pressure on pattern-density, surface topography, and step-height $P(\rho, z_u, h)$. The relationship $K(P)$ is discussed in Section 3.2.3 for polishing of single materials or dual materials using either conventional slurry or non-conventional slurry (in which a non-linear rate versus pressure dependence exists).

A physically-based die-level model is developed in Section 3.3.1 by explicitly modeling the polishing pad with surface asperities, and the model parameters are based on physical properties of the pad and the asperities. The physically-based model has the advantages of handling complicated initial topography and enabling the study of the effects of pad properties and applied pressure. In Section 3.5.2, the physically-based model is used to examine the additional topography variation caused by initial topography in oxide polishing, in STI CMP, and in multi-level interconnect polishing. Section 3.5.3 shows that a more rigid pad can reduce the post-CMP with-in-die variation, and a smaller value of the pad characteristic asperity height $\lambda$ can improve the step-height reduction. Section 3.5.4 explains how the applied pressure affects both the removal rate and the planarization performance of CMP.

A semi-empirical exponential pattern-density step-height (Exp-PDSH) model is proposed in Section 3.3.2 by making realistic assumptions and approximations, and improving the ease of computation compared to the previous PDSH model. In Section 3.4, the Exp-PDSH model has been illustrated to simulate the polishing of single-material or dual-material structures using a conventional or non-conventional slurry. In the illustration, analytical solutions of surface evolution are presented, and the effects of model parameters are discussed. Section 3.4.5 presents a class of slurries with tailored rate versus pressure dependencies which has the potential to improve

planarization in CMP.

The physically based model and the Exp-PDSH model are verified by experimental data in Section 3.6, and good agreement between the model predictions and experimental data are observed. The two models are also compared with each other in Section 3.5.1, and the comparison helps to explain the meaning of the Exp-PDSH model parameters. In particular, the planarization length $L_p$ is mainly affected by the Young's modulus $E$ of the pad, and the characteristic step-height $h^*$ is mainly affected by the characteristic pad asperity height $\lambda$.

### 5.1.3 Applying Die-Level CMP Models

A methodology for applying the die-level CMP models is reviewed and illustrated in Section 4.1.1. CMP test masks are an essential part of the methodology, and a new STI test mask, introduced in Section 4.1.2, is designed to contain relevant pattern factors that the current STI CMP process has dependency on. The developed CMP models have been applied to solve the following four problems in practice.

- Section 4.1.3 discusses the benefit of applying the die-level CMP models in design-for-manufacturing. Snap shots of a GUI simulator for STI CMP are shown to illustrate how the simulator can be used by a layout designer to gain instant feedback on the CMP performance for any layout design.

- In Section 4.4, the nanotopography impact in both blanket and patterned wafer polishing has been studied via a comprehensive experiment design and various data analysis methods. The study finds nanotopography to be a significant factor; however, its contribution to the total variance accounts for less than 10% observed variation in blanket wafer polishing, and less than 5% in patterned wafer polishing.

- In Section 4.3, a contact wear CMP model has been used to study the cause of a wafer edge roll-off effect and its impact on the uniformity of later polishing by simulation of both the instantaneous removal rate distribution and the surface evolution during CMP. The modeling work suggests that the edge roll-off profile

is caused by the detailed structures near the wafer edge, and the work points out a number of ways to reduce its impact on later polishing.

- Section 4.4 studies the usage of real-time motor current measurement as an endpoint signal in STI CMP. The motor current is shown to be proportional to the friction force between the polishing pad and wafer. A friction model is proposed, in which the friction is assumed to be proportional to the surface roughness and is affected by surface material type. The die-level CMP model is used to provide surface evolution information needed for friction calculation, and the model predictions agree reasonably well with measured motor current traces. The friction model can be used in combination with the motor current to more accurately determine the endpoint.

## 5.2 Value of Contributions

The contributions of this thesis on physical understanding of CMP are of value in the following ways:

- The framework for a particle-level model establishes a multi-scale structure to study the CMP process, and decouples various interactions of CMP. Thus it can serve as a road-map for future research on CMP which might empirically study or theoretically model each interaction.

- The framework can also serve as a platform to develop future particle-level CMP models. If empirical evidence suggests a different mechanism of an interaction or if the polishing of a new material needs to be modeled, only the corresponding model component focusing on that interaction needs to be replaced. For example, the framework can be used to model copper CMP, by replacing the chemical-tooth removal mechanism with a chemical-complex formation and indentation mechanism.

- The particle-level CMP model and physically-based die-level model link CMP performance directly with physical properties of the pad and the slurry, and

thus can be used to help design consumables with faster removal rates, smaller defect rates, and better planarization abilities.

The contributions in die-level CMP modeling are valuable in the following ways:

- The framework can serve as a basis for future die-level modeling work.
- The fast computation of the Exp-PDSH model makes it an efficient simulation tool in practice, especially for layout designers to get rapid feedback on the post-CMP information for any design, and for process engineers to better control the CMP process.
- The physically-based model can be used to simulate the CMP process whenever the impact of initial topography is a concern, such as the nanotopography impact in STI and polishing multi-level interconnect structures.
- In designing a CMP process, the physically-based model can be used to estimate the effect of different applied pressures.
- The physically-based model can also be used to design or optimize the polishing pad to achieve better planarization and faster step-height reduction.
- Both die-level models can be used to find the dependence of removal rate on pressure to improve planarization, which will be useful for slurry design or optimization.

The work on applying the die-level CMP model is of value in the following ways:

- The new STI mask improves upon the previous MIT STI mask by using more realistic shapes and a wider range of pattern factors. The new mask can be used to calibrate and study advanced oxide or STI CMP processes.
- The results of the edge roll-off study can be used to minimize the edge roll-off effect by improved tool design and process control.
- The proposed friction model can assist in more accurate endpoint detection based on the motor current signal.

## 5.3 Future Work

The proposed particle-level model predicts certain dependencies of removal rate on input variables, including pad modulus and asperity distribution. Experiments designed to study the dependence can be used to verify the proposed model, as well as future particle-level models. With the guide of the framework, further development of particle-level models might focus on the following interactions and experiments.

- The abrasive-wafer interaction and chemical reaction can be studied by wear of the wafer surface with an AFM stylus, which simulates a single abrasive particle. This experiment can empirically study how removal rate varies with abrasive material, abrasive size, applied load, pH value of slurry, and other parameters.

- The pad-wafer interaction can be further understood by both measurement of asperities distribution with higher resolution and direct measurement of asperities deformation, such as the dependence of contact area and deformation amount on applied pressure.

- The particle-level model approximates the abrasives in the contact area as being of identical size with a fixed area density and considers the contribution of the large abrasives in the tail of size distribution. Theoretically modeling of the dynamics of abrasives entering or leaving the contact area is challenging, and it is desirable to experimentally approach the problem, such as by studying the dependence of removal rate and friction on abrasive size and concentration of abrasives in the slurry.

In the development of the physically-based die-level CMP model, the width of the asperities is assumed to be negligible, and the assumption should be revisited for very small features. Further work can focus on modeling the effect of asperity width on polishing small features. Direct verification of the model is desirable, by design experiments to study the planarization performance of pads with different Young's modulus and different pad surface asperity distributions, where the Young's

modulus of the pad can tuned by changing the porosity or material and the asperity distribution can be altered by conditioning, for example,.

This thesis has contributed to the understanding and modeling of CMP; however, given the range of use of CMP and the complexity of the process, many interesting questions remain for future research.

# Appendix A

# Elasticity

Elasticity is a well-understood subject in classical physics [19]. The section reviews a few results related to the work in contact wear modeling.

## A.1 3D Elasticity Problem

Consider a small element $dx \cdot dy \cdot dz$ of an elastic body, as in Figure A-1. The body undergoes a deformation, and $u$, $v$, $w$ are the $x$, $y$, $z$ components of the displacement. The displacement $u$ not only varies as its $x$ coordinate changes, but also as its $y$ and $z$ coordinates change. Assuming it is a small deformation, we have

$$u_A = u_O + \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy + \frac{\partial u}{\partial z}dz. \tag{A.1}$$

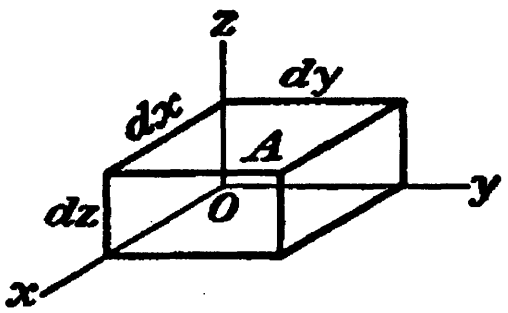



Figure A-1: A diagram shows a small element $dx \cdot dy \cdot dz$ of an elastic body [19].

Thus, strains $\epsilon_i$ and $\gamma_{jk}$ are introduced to describe the deformation of the elastic

body as follows.

$$\begin{aligned} \epsilon_x &= \frac{\partial u}{\partial x}, & \epsilon_y &= \frac{\partial v}{\partial y}, & \epsilon_z &= \frac{\partial w}{\partial z}, \\ \gamma_{xy} &= \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}, & \gamma_{yz} &= \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}, & \gamma_{xz} &= \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}. \end{aligned} \tag{A.2}$$

Stresses $\sigma_i$ and $\tau_{jk}$ are defined to describe the forces exerted on the elastic body as shown in Figure A-2. The $\sigma_i$'s are perpendicular components, and $\tau_{jk}$'s are tangential components.

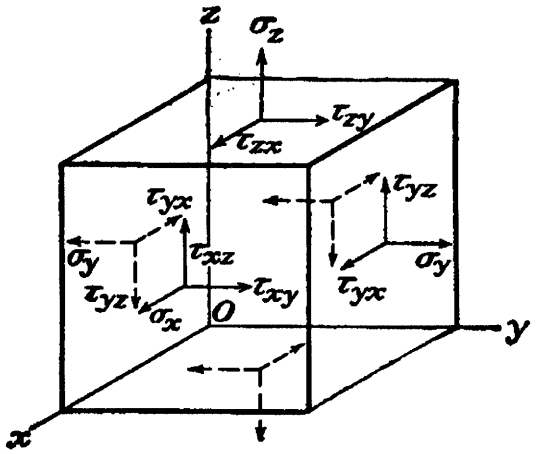



Figure A-2: A diagram shows stress components exerted on a elastic body $dx \cdot dy \cdot dz$ [19].

Given the strains, we can calculate the stress components for an isotropic elastic body.

$$\begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{xy} \\ \tau_{yz} \\ \tau_{xz} \end{bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu & 0 & 0 & 0 \\ \nu & 1-\nu & \nu & 0 & 0 & 0 \\ \nu & \nu & 1-\nu & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}-\nu & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}-\nu & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2}-\nu \end{bmatrix} \begin{bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \\ \gamma_{xy} \\ \gamma_{yz} \\ \gamma_{xz} \end{bmatrix}, \tag{A.3}$$

where $E$ is the Young's modulus and $\nu$ is the Poisson ration of the elastic material.

If the elastic body is in equilibrium, the balance of forces requires

$$\begin{cases} \frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} + f_x &= 0 \\ \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} + f_y &= 0 \\ \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} + f_z &= 0 \end{cases} \tag{A.4}$$

where $f_x$, $f_y$, $f_z$ are the $x$, $y$, $z$ components of external forces.

## A.2 Axial Symmetry

When the material exhibits axial symmetry, the equations can be simplified in polar coordinates. The strains are defined as

$$\begin{gathered} \epsilon_r = \frac{\partial u}{\partial r} \qquad \epsilon_\theta = \frac{u}{r} \qquad \epsilon_z = \frac{\partial w}{\partial z} \\ \gamma_{rz} = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial r} \end{gathered} \tag{A.5}$$

The constitutive equations become

$$\begin{bmatrix} \sigma_r \\ \sigma_\theta \\ \sigma_z \\ \tau_{xz} \end{bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu & 0 \\ \nu & 1-\nu & \nu & 0 \\ \nu & \nu & 1-\nu & 0 \\ 0 & 0 & 0 & \frac{1}{2}-\nu \end{bmatrix} \begin{bmatrix} \epsilon_r \\ \epsilon_\theta \\ \epsilon_z \\ \gamma_{xz} \end{bmatrix}, \tag{A.6}$$

The equilibrium equations become

$$\begin{cases} \frac{\partial \sigma_r}{\partial r} + \frac{\partial \tau_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} + f_r &= 0 \\ \frac{\partial \tau_{rz}}{\partial r} + \frac{\partial \sigma_z}{\partial z} + \frac{\tau_{rz}}{r} + f_z &= 0 \end{cases} \tag{A.7}$$

## A.3 Implications for Contact Wear Modeling

The contact wear model studies the deformation of an elastic body in contact with a rigid surface, and the modeling is based on the deformation of the elastic body due to a point pressure. Most previous work on the use of a contact wear model in CMP [74] [75] assume a semi-infinite elastic body, which is a rough approximation, to study the contact pressure distribution between the polishing pad and wafer. The elasticity theory can be used to study the deformation of a finite elastic body under

a point deformation, and the result enables the contact wear model to account for pad thickness. In addition, the contact under a point pressure has axial symmetry. As discussed in the previous section, the elastic contact with axial symmetry can be treated as a 2D problem and is computationally less demanding to solve numerically, providing further benefits for CMP modeling.

# Appendix B

# Contact Wear Model

The contact wear model [74] [75] describes how a semi-infinite elastic body deforms when it is pressed by a rigid surface with height $z(x, y)$. The contact is assumed to be non-sticky, i.e., all contact pressures $P(x, y)$ are positive. The rigid surface is not compressible, thus the elastic surface $W(x, y)$ cannot extend below the surface $z(x, y)$. And finally, force balance of the elastic body requires that the average of contact pressure equals the applied pressure. A summary of the boundary conditions are written as

$$\begin{cases} P(x,y) > 0 \\ \frac{1}{S}\int_S dx \cdot dy \cdot P(x,y) = P_0 \\ W(x,y) \geq z(x,y) \end{cases} \tag{B.1}$$

where $P_0$ is the average applied pressure on the elastic body.

The surface deformation under a point pressure is inversely proportional to the applied pressure [62]. Using a superposition rule, as illustrated in Figure B-1, we obtain the relationship between applied pressure and displacement as

$$w(x,y) - w_0 = \frac{1-\nu^2}{E}\int d\xi \int d\eta \frac{P(\xi,\eta)}{\sqrt{(x-\xi)^2 + (y-\eta)^2}} \tag{B.2}$$

where $E$ is the Young's modulus of the elastic material, $\nu$ is the Possion's ratio, $w(x, y)$ is the surface displacement, $P(x, y)$ is the applied pressure, and $w_0$ is the reference plane when $P(x, y) = 0$.

The contact wear model is useful to describe the contact between an elastic surface and a rigid body, and to compute the contact pressure distribution. This appendix

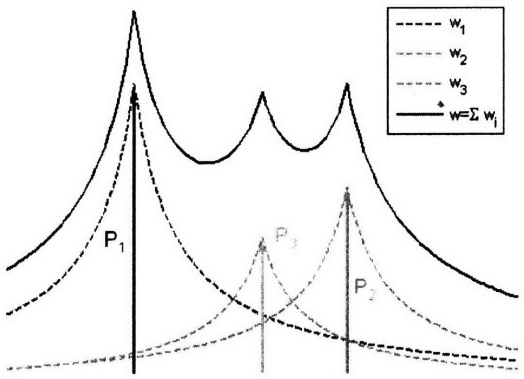



Figure B-1: Illustration of superposition in contact wear calculation

is intended to present an efficient implementation of the contact wear model and to expand the model to an elastic body with finite thickness. Section B.1 re-presents the problem in the discretization form; Section B.2 describes the algorithm for solving the contact wear problem; and Section B.3 discusses how to apply the model to an elastic body of finite thickness.

## B.1 Statement of the Problem in Discretized Form

In computation, we discretize the area of interest into squares regions, and we assume that the pressure and displacement do not vary much within each square, so that an average value of pressure and displacement can be used to represent the whole cell. Thus, the problem is reduced to the interaction between different cells. We denote the lower left coordinate of cell $i$ as $(x_{i,1}, y_{i,1})$, the upper right corner as $(x_{i,2}, y_{i,2})$, and the center as $(x_{i,0}, y_{i,0})$. The displacement of the $i_{th}$ cell $W_i$, which is assumed to be the same as that of its center $(x_{i,0}, y_{i,0})$, can be determined by Equation B.2 as

follows:

$$\begin{aligned} W_i - W_0 &= \frac{1-\nu^2}{E} \sum_j P_j \int_{x_{j,1}}^{x_{j,2}} d\xi \int_{y_{j,1}}^{y_{j,2}} d\eta \frac{1}{\sqrt{(\xi - x_{i,0})^2 + (\eta - y_{i,0})^2}} \\ &= \frac{1-\nu^2}{E} \sum_j P_j \int_{x_{j,1}-x_{i,0}}^{x_{j,2}-x_{i,0}} d\xi \int_{y_{j,1}-y_{i,0}}^{y_{j,2}-y_{i,0}} d\eta \frac{1}{\sqrt{\xi^2 + \eta^2}} \\ &= \sum_j F_{ij} \cdot P_j \end{aligned} \tag{B.3}$$

where $F_{ij}$ describes the deformation of the $i^{th}$ cell caused by unit pressure applied at the $j^{th}$ cell, and can be written as

$$F_{ij} = \frac{1-\nu^2}{E} \int_{x_{j,1}-x_{i,0}}^{x_{j,2}-x_{i,0}} d\xi \int_{y_{j,1}-y_{i,0}}^{y_{j,2}-y_{i,0}} d\eta \frac{1}{\sqrt{\xi^2 + \eta^2}}. \tag{B.4}$$

The integration can be solved analytically as

$$\begin{aligned} F_{ij} &= \frac{1-\nu^2}{E} \int_{x_1}^{x_2} d\xi \int_{y_1}^{y_2} d\eta \frac{1}{\sqrt{\xi^2 + \eta^2}} \\ &= \frac{1-\nu^2}{E} \left[ f(x_2, y_2) - f(x_1, y_2) - f(x_2, y_1) + f(x_1, y_1) \right], \end{aligned} \tag{B.5}$$

where $f(x,y) = y \ln(x + \sqrt{x^2 + y^2}) + x \ln(y + \sqrt{x^2 + y^2})$. It is more convenient to write the equation in the matrix form

$$[W]_{n\times 1} - W_0 = [F]_{n\times n} \cdot [P]_{n\times 1} \tag{B.6}$$

Many elements of matrix $F$ are identical: $F_{i,j} = F_{k,l}$ if $x_i - x_j = x_k - x_l$ and $y_i - y_j = y_k - y_l$, i.e., the value of $F_{i,j}$ only depends on the distance between the two cells. This motivates us to view the problem from a different perspective: $jth$ column of $F$ represents the response to the pressure on cell $j$. Each column describes the same response function, but varies due to the change of location. The above matrix calculation of displacement is equivalent to a convolution between the pressure distribution and the response function, assuming a periodic boundary condition.

Let us use $(i,j)$ to denote the cell that is $i$ units east and $j$ units north of the center cell, and its displacement and pressure can be represented as $\tilde{W}_{i,j}$ and $\tilde{P}_{i,j}$. Let $\tilde{F}_{i,j}$ describe the interaction between the center $O$ and the cell $(i,j)$, which has the

same value as $F_{O,(i,j)}$. Figure B-2(b) shows a 3D plot of $\tilde{F}_{i,j}$. Using $\tilde{W}_{i,j}$, $\tilde{P}_{i,j}$ and $\tilde{F}_{i,j}$, Equation B.6 can be written as a convolution calculation, which can be computed efficiently using fast Fourier transformation.

$$\begin{aligned} \tilde{W}_{x,y} - W_0 &= \sum_{\xi,\eta} \tilde{F}_{x-\xi,y-\eta} \cdot \tilde{P}_{\xi,\eta} \\ &= \tilde{F}_{x,y} \otimes \tilde{P}_{x,y} \end{aligned} \tag{B.7}$$

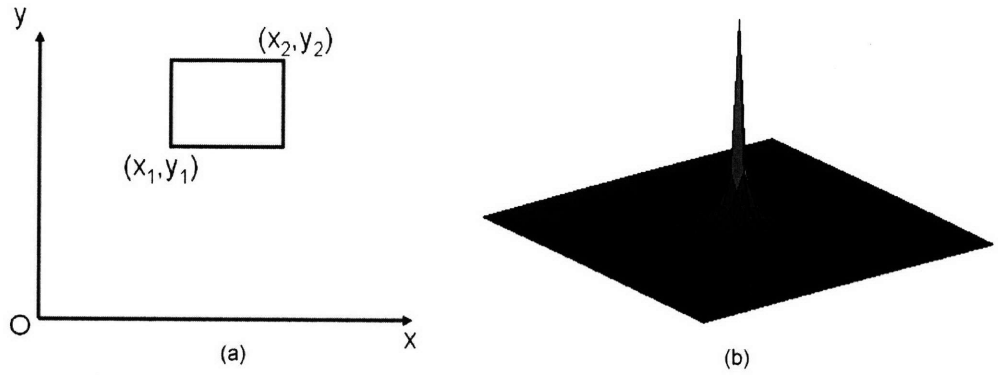



Figure B-2: (a) Diagram shows discretization of surface to compute the displacement of origin $O$ caused by the pressure on a square defined by lower left corner $(x_1, y_1)$ and upper right corner $(x_2, y_2)$. (b) Mesh plot of $\tilde{F}(i, j)$.

## B.2 Solving the Contact Wear Problem

Yoshida [75] proposes an approach for solving the contact wear problem in matrix format by noting that if the elastic body is in contact with the rigid surface, $W(x, y) = z(x, y)$; and if it is not in contact, $P(x, y) = 0$. If the cells in contact are known, the indexing of the cells can be arranged so that the contact cells appear first in vector $W$. We denote their displacements using vector $W_k$ and their pressure as $P_u$, and subscripts $k$ and $u$ are used to indicate that the displacements are known $(k)$ and the pressures are unknown $(u)$. Similarly, $W_u$ and $P_k$ are used to denote the cells that

are not in contact.

$$\begin{bmatrix} W_k \\ W_u \end{bmatrix} - W_0 = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \cdot \begin{bmatrix} P_u \\ P_k \end{bmatrix} \tag{B.8}$$

Carrying out the matrix multiplication yields

$$\begin{cases} W_k = F_{11}P_u + F_{12}P_k + W_0 \\ W_u = F_{21}P_u + F_{22}P_k + W_0 \end{cases} \tag{B.9}$$

Noting that $P_k = 0$, $W_k$ and $P_k$ can be solved as

$$\begin{cases} P_u &= F_{11}^{-1}W_k - W_0F_{11}^{-1}e \\ W_u &= F_{21}P_u + W_0 \end{cases} \tag{B.10}$$

where $e$ is a unit column vector whose elements are all 1, and $W_0$ can be determined by enforcing the condition that the average of $P$ equals $P_0$.

If the contact cells are not known, an initial guess is tested. If the calculated pressure of a contact cell is negative, the cell will be assigned as not-in-contact; if the calculated displacement of a not-in-contact cell intrudes into the rigid surface, it will be assigned as in contact. The guess is then tested, and the procedure is repeated, until no violation of the boundary occurs.

There are two problems with Yoshida's approach: the guess-and-test approach does not guarantee convergence of the solution and it often leads to oscillation states; and the full matrix approach is computationally inefficient. For example, if a 20 $mm \times$ 20 $mm$ die is discretized into 20 $\mu m \times 20$ $\mu m$ cells, there are $1000 \times 1000 = 10^6$ cells and the matrix $[F]_{10^6 \times 10^6}$ has $10^{12}$ elements, which is not feasible on readily available hardware.

### B.2.1 Sticky Approach

Empirically, a "sticky" approach solves the convergence problem in Yoshida's approach. The sticky approach is derived from the following physical intuition.

1. While applying the pressure on the elastic body, suppose that we also glue the entire elastic body on the rigid surface. In this case, the contact pressure can

be negative, which means the glue is holding them together.

2. We measure all the contact pressures, and at the same time, remove the glue where the pressure is negative.

3. The elastic body reaches a new equilibrium of contact pressures, and step 2 is repeated until no further negative pressures are found.

In this approach, the contact areas decrease monotonically after each iteration, and thus we always converge to the solution. The assumption here is that, as we "release" glued regions with negative pressures, the unglued areas do not re-touch the rigid surface. Intuitively, by setting the negative pressures to zero, the average pressure on the rest of the areas decreases and the smaller average pressure is not likely to increase or add to the number of contact points. No theoretical proof has been found for the assumption, although the assumption has not been violated when applying the approach to various complex topographies in our work.

The above steps can be translated into the contact wear computation as follows:

1. Initially the elastic body is assumed to be in contact with the entire rigid surface.

2. Calculate the pressure values, assign contact cells with negative pressure to the not-in-contact group, and release their pressure back into the pool of pressure to be reallocated to cells in contact.

3. Repeat step 2, until no violation of boundary conditions remains.

### B.2.2 Convolution Approach

The matrix form solution of the contact wear problem is limited to a small matrix size due to the huge amount of memory and computation power it demands. If the matrix computation can be replaced by a convolution calculation, the convolution calculation can be computed using the 2D discrete fast Fourier transformation, and the computational efficiency can be greatly improved. Consider the case of discretizing an area of interest into $n \times n$ squares. The matrix multiplication requires a memory of size

$O(n^4)$ and has a computational complexity of $O(n^4)$, while the convolution approach requires a memory of size $O(n^2)$ and has computation complexity of $O(n^2 \log(n))$.

It needs to be mentioned that the convolution calculation assumes a periodic boundary condition, while the matrix computation does not. This is not a big issue because we can choose a larger area for simulation and ignore the area near the edge, and in modeling a die structure, a periodic boundary is desired to reflect the periodic arrangement of dies on the wafer. The real problem is that the matrix computation in Equation B.10 cannot be easily replaced by convolution. Here the problem is approached in two steps: replacing the matrix multiplication with a convolution computation, and using an approximation method to solve $F_{11}^{-1}x$, where $x$ can be $W_k$ or $e$.

We have shown that the matrix multiplication in Equation B.6 can be computed using a convolution method as in Equation B.7. The computation of $F_{11}x$ can be achieved by using Equation B.8 with $P_u$ replaced by $x$ and $P_k$ set to zero.

$$\begin{bmatrix} W_k \\ W_u \end{bmatrix} = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \cdot \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} F_{11}x \\ 0 \end{bmatrix} \tag{B.11}$$

Thus, with $P_u = x$ and $P_k = 0$, the computation of $[F]_{n\times n} \cdot [P]_{n\times 1}$ can be carried out by convolution, and the computed $W_k$ is $F_{11}x$. Similarly, $F_{21}x$ can computed by setting $P_u = 0$ and $P_k = x$, and the calculated $W_u$ equals $F_{21}x$.

$$\begin{bmatrix} W_k \\ W_u \end{bmatrix} = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \cdot \begin{bmatrix} 0 \\ x \end{bmatrix} = \begin{bmatrix} 0 \\ F_{21}x \end{bmatrix} \tag{B.12}$$

The second step is to compute $F_{11}^{-1}x$, i.e., to solve $y$ from $F_{11}y = x$. Rather than solving the exact solution, obtaining an approximate solution is sufficient for most engineering problems. The Matlab© function **bicgstab** implements the biconjugate gradients stabilized method, which is ideal for the task.

The convolution approach greatly reduces the demand for memory and computation power, and the simulation of an area discretized into $1000 \times 1000$ elements can be solved in a few minutes using a typical Pentium© desktop computer.

## B.3 Elastic Body of Finite Thickness

The $1/r$ dependence in the contact wear model is valid if the elastic body has an infinite thickness, and is a good approximation if the thickness is much larger than the dimension of the area. In modeling the pad in contact with the wafer surface, the pad thickness is about 7 $mm$ and the size the wafer is usually much larger than this, and a correction may thus be necessary. In this section, the modeling of an elastic body with finite thickness is addressed.

With a finite thickness, the elastic body responds to a point pressure in a different way. However, as long as the elastic body is isotropic, the following equation still holds:

$$w(x,y) - w_0 = \int d\xi \int d\eta \cdot P(\xi,\eta) \cdot F(x-\xi, y-\eta), \tag{B.13}$$

where $F(x,y)$ captures the unit response of a point pressure. Once $F(x,y)$ is known for the finite thickness case, the contact wear problem can be solved in exactly the same way. Thus, the task is to obtain $F(x,y)$, given the Young's modulus $E$ and thickness $d$. Here, $F(x,y)$ is calculated by numerically solving the elasticity equations, which are reviewed in Appendix A. The elastic equations of axial symmetry A.6 and A.7 are used to simulate the finite-thickness elastic body. The simulation results for pads with several thickness values are shown in Figure B-3, and are compared with the $1/r$ dependence of the infinitely thick pad. We see that for a typical pad thickness $h_z = 7\ mm$, a finite thickness correction should be used for accurate results, and to avoid over-estimating the pad deflection.

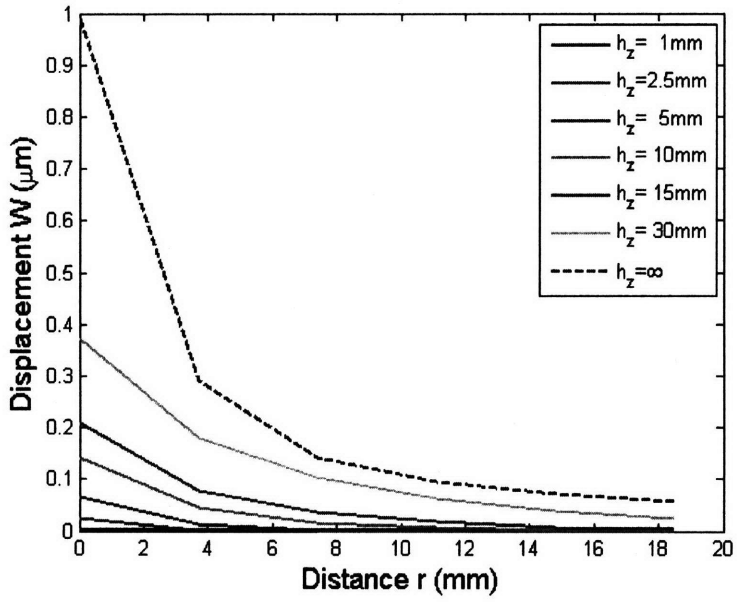



Figure B-3: Elasticity equations are used to simulate the pad response to point pressure for several values of pad thickness. The results are plotted and compared to the infinitely thick pad (dashed line), which has the $1/r$ dependence.

# Bibliography


[1] B. Lee. Modeling of Chemical Mechanical Polishing for Shallow Trench Isolation. *MIT Ph.D. Thesis, Department of Electrical Engineering and Computer Science*, May 2002.

[2] S. D. Hosali, A. Sethuraman, J. F. Wang, L. M. Cook, and D. Evans. Planarization Process and Consumable Development for Shallow Trench Isolation. *Proc. CMP-MIC Conf.*, pp. 52-57, February 1997.

[3] D. Evans. CMP Integration. *MRS Tutorial*, 2005.

[4] R. K. Singh, S. M. Lee, K. S. Choi, G. B. Basim, W. Choi, Z. Chen, and B. M. Moudgil. Fundamentals of Slurry Design for CMP of Metal and Dielectric Materials. *MRS Bulletin*, pp. 752-760, October 2002.

[5] H. Barthel, M. Heinemann, M. Stintz, and B. Wessely. Particle Sizes of Fumed Silica. *Particle and Particle Systems Characterization*, vol. 16, no. 4, pp. 169 - 176, October 1999.

[6] D. Bouvet, P. Beaud, P. Fazan, R. Sanjines, and E. Jacquinot. Impact of the Colloidal Silica Particle Size on Physical Vapor Deposition Tungsten Removal Rate and Surface Roughness. *Journal of Vacuum Science and Technology B*, vol. 20, no. 4, pp. 1556-1560, July 2002.

[7] H. Lu, B. Fookes, Y. Obeng, S. Machinski, and K.A. Richardson. Quantitative Analysis of Physical and Chemical Changes in CMP Polyurethane Pad Surfaces. *Mat. Char.*, vol. 49, pp. 35-44, 2002.

[8] C. Zhou, L. Shan, S. H. Ng, R. Hight, A. J. Paszkowski, and S. Danyluk. Effects of Nano-scale Colloidal Abrasive Particle Size on $SiO_2$ by Chemical Mechanical Polishing. *Mat. Res. Soc. Symp. Proc.*, vol. 671, paper M1.6.1, 2001.

[9] N. Chandrasekaran, T. Taylor, and G. Sabde. Effect of Ceria Particle-size Distribution and Pressure Interactions in Chemo-Mechanical Polishing (CMP) of Dielectric Materials. *Mat. Res. Soc. Symp. Proc.*, vol. 767, pp. F3.2.1, 2003.

[10] L. J. Borucki, T. Witelski, C. Please, P. R. Kramer, and D. Schwendeman. A Theory of Pad Conditioning for Chemical-Mechanical Polishing. *Journal of Engineering Mathematics*, vol. 50, pp. 1-24, 2004.

[11] L. Borucki, R. Zhuang, T. Sun, Y. Zhuang, A. Philipossian, and D. Slutz. Mechanical and Optical Analysis of Pad Surface Micro-texture Differences Caused by Conditioning. *International Conference on Planarization Technology*, October 2006.

[12] C. L. Elmufdi and G. P. Muldowney. A Novel Optical Technique to Measure Pad-Wafer Contact Area in Chemical Mechanical Planarization. *Mat. Res. Soc. Symp. Proc.*, vol. 914, paper 0914-F12-06, 2006.

[13] B. Stine, D. Ouma, R. Divecha, D. Boning, J. Chung, D. Hetherington, I. Ali, G. Shinn, J. Clark, O. Nakagawa, and S.-Y. Oh. A Closed-form Analytic Model for Ild Thickness Variation in CMP Processes. *Proc. CMP-MIC Conf.*, p. 266, February 1997.

[14] H. Nojo, M. Kodera, and R. Nakata. Slurry Engineering for Self-stopping, Dishing Free $SiO_2$-CMP. *IEDM*, pp. 349-352, 1996.

[15] T. E. Gbondo-Tugbawa. Chip-scale Modeling of Pattern Dependencies in Copper Chemical Mechanical Polishing Processes. *MIT Ph.D. Thesis, Department of Electrical Engineering and Computer Science*, 2002.

[16] D. Ouma, D. Boning, J. Chung, G. Shinn, L. Olsen, and J. Clark. An Integrated Characterization and Modeling Methodology for CMP Dielectric Planarization. *International Interconnect Technology Conference*, pp. 67-69, 1998.

[17] T. Fukuda, M. Tsujimura, Y. Nakai, T. Morimoto, M. Yoshise, S. Akiyama, and S. Kobayashi. The Impact of Edge Roll-off on CMP Performance. *Japan Electronics and Information Technology Industries Association*, no. EMR-3001, 2004.

[18] D. Boning, X. Xie, J. Sorooshian, A. Philipossian, D. Stein, and D. Hetherington. Relationship Between Patterned Wafer Topography Evolution and Sti CMP Motor Current Endpoint Signals. *Proc. CMP-MIC Conf.*, pages pp. 341–350, 2004.

[19] S. Timoshenko and J. N. Goodier. Theory of Elasticity. *McGraw-Hill, New York*, 1951.

[20] J. Steigerwald, S. Murarka, and R. Gutmann. Chemical Mechanical Planarization of Microelectronic Materials. *Wiley-Interscience*, 1997.

[21] D. Evans. The Future of CMP. *MRS Bulletin*, pp. 779-783, October 2002.

[22] International Technology Roadmap for Semiconductors. *International SEMATECH*, 2005.

[23] T. Gan. Modeling of Chemical Mechanical Polishing for Shallow Trench Isolation. *MIT Master of Engineering Thesis, Department of Electrical Engineering and Computer Science*, 2000.

[24] K. T. Turner. Wafer Bonding : Mechanics-based Models and Experiments. *MIT Ph.D. Thesis, Department of Mechanical Engineering*, 2004.

[25] D. L. Hetherington and J. J. Sniegowski. Improved polysilicon surface-micromachined micromirror devices using chemical-mechanical polishing. *Proceedings of SPIE*, vol. 3440, pp. 148-153, October 1998.

[26] R. Nasby, J. Sniegowski, J. Smith, S. Montague, C. Barron, W. Eaton, P. McWhorter, D. Hetherington, C. Apblett, and J. Fleming. Application of Chemical-Mechanical Polishing to Planarization of Surface-Micromachined Devices. *Proc. Solid State Sensor and Actuator Workshop*, pp. 48-53, 1996.

[27] C. Kourouklis, T. Kohlmeier, and H. H. Gatzen. The Application of Chemical-cmechanical polishing for Planarizing a SU-8/Permalloy Combination Used in MEMS Devices. *Sens. Actuators A Phys.*, vol. 106, pp. 263, 2003.

[28] J. Fleming and C. Barron. Novel Silicon Fabrication Process for High-Aspect-Ratio Micromachined Parts. *Proc. SPIE Micromachining and Microfabrication*, vol. 2639, p. 185, 1995.

[29] S. Y. Lin, J. G. Fleming, and E. Chow. Two- and Three-dimensional Photonic Crystals in III-V Semiconductors. *MRS Bulletin*, pp. 627-631, August 2001.

[30] F. Preston. The Theory and Design of Plate Glass Polishing Machines. *J. Soc. Glass Technology*, vol. 11, pp. 214, 1927.

[31] R. Jairath, S. Chadda, E. Engdahl, W. Krussel, T. Mallon, K. Mishram, A. Pant, and B. Withers. Performance of OnTrak System's Linear Planarization Technology (LPT) for Dielectric CMP Processes. *Proc. CMP-MIC Conf.*, pp. 194-201, February 1997.

[32] M. Oliver. Chemical-mechanical Planarization of Semiconductor Materials. *Springer, Berlin; New York*, 2004.

[33] N. H. Kima, P. J. Kob, Y. J. Seoc, and W. S. Lee. Improvement of TEOS-Chemical Mechanical Polishing Performance by Control of Slurry Temperature. *Microelectronic Engineering*, vol. 83, pp. 286C292, 2006.

[34] J. Sorooshian, D. DeNardis, L. Charns, Z. Li, F. Shadman, D. Boning, D. Hetherington, and A. Philipossian. Arrhenius Characterization of ILD and Copper CMP Processes. *Journal of the Electrochemical Society*, vol. 152, no. 2, pp. G85-G88, 2004.

[35] D. Stein, D. Hetherington, M. Dugger, and T. Stout. Optical Interferometry for Surface Measurements of CMP Pads. *Journal of Electronic Materials*, vol. 25, no. 10, pp. 1623 - 1627, 1996.

[36] L. Borucki. Mathematical modeling of polish-rate decay in chemical-mechanical polishing. *Journal of Engineering Mathematics*, vol. 43, pp. 105-114, 2002.

[37] A. S. Lawing. Pad Conditioning and Textural Effects in Chemical Mechanical Polishing. *Proc. CMP-MIC Conf.*, paper 2.C, 2005.

[38] N. Gitis, M. Vinogradov, and C. Gao. Quantitative Evaluation of CMP Processes and Materials Using A CMP Tester with Multiple Sensors. *Proceedings of the Second International Conference on Microelectronics and Interfaces*, February 2001.

[39] D. Boning, X. Xie, J. Sorooshian, A. Philipossian, D. Stein, and D. Hetherington. Relationship Between Patterned Wafer Topography Evolution and STI CMP Motor Current Endpoint Signals. *Proc. CMP-MIC Conf.*, pp. 341-350, February 2004.

[40] T. Kojima, M. Miyajima, F. Akaboshi, T. Yogo, S. Ishimoto, and A. Okuda. *IEEE Transactions on Semiconductor Manufacturing*, vol. 13, no. 3, pp. 291, August 2000.

[41] L. Li, C. Wei, J. Gilhooly, and C. Morgan. End Point Detection in Metal and Nitride-Containing CMP Processes. *Proceedings of the Second International Conference on Microelectronics and Interfaces*, February 2001.

[42] B. Adams, B. Swedek, R. Bajaj, K. Wijekoon, S. Nanjangud, A. Wiswesser, S. Tsai, D. Chan, F. Redeker, and M. Birang. Process Control and Endpoint Detection With In Situ Rate Monitor System in Chemical Mechanical Polishing of Cu Layer. *Proc. CMP-MIC Conf.*, pp. 558-562, 2000.

[43] V. Bhaskaran, C. Chen, R. Allen, K. Lehman, H. Chen, D. Watts, and B. Stephenson. Advanced In-Situ End-Point Control System for CMP Appli-

cations. *CMP Users Group Meeting, American Vacuum Society, Santa Clara, CA*, 2001.

[44] G. Nanz and L. E. Camilletti. Modeling of Chemical-Mechanical Polishing: A Review. *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 4, pp. 382-389, November 1995.

[45] X. Xie and D. Boning. CMP at the Wafer Edge – Modeling the Interaction Between Wafer Edge Geometry and Polish Performance. *Mat. Res. Soc. Symp. Proc.*, vol. 867, paper W.5.1.1, 2005.

[46] D. O. Ouma, D. S. Boning, J. E. Chung, W. G. Easter, V. Saxena, S. Misra, and A. Crevasse. Characterization and Modeling of Oxide Chemical-Mechanical Polishing Using Planarization Length and Pattern Density Concepts. *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, no. 2, pp. 232-244, 2002.

[47] M. Tomozawa, K. Yang, H. Li, and S. P. Murarka. Basic Science in Silica Glass Polishing. *Advanced Metallization for Devices and Circuits - Science, Technology and Manufacturability Symposium*, pp. 89-98, 1994.

[48] Y. Z. Hu, R. J. Gutmann, and T. P. Chow. Silicon Nitride Chemical Mechanical Polishing Mechanisms. *J. Electrochem. Soc.*, vol. 145, no. 11, pp. 3919-3925, November 1998.

[49] L. M. Cook. Chemical Processes in Glass Polishing. *J. Non-Cry. Solids*, vol. 120, pp. 152-171, 1990.

[50] M. Szycher. Handbook of Polyurethanes. *CRC Press, New York*, 1999.

[51] G. P. Muldowney. Modeling Cmp Transport and Kinetics at the Pad Groove Scale. *Mat. Res. Soc. Symp. Proc.*, vol. 816, paper K5.3, 2004.

[52] E. Remsen, S. Anjur, D. Boldridge, M. Kamiti, S. Li, T. Johns, C. Dowell, J. Kasthurirangan, and P. Feeney. Analysis of Large Particle Count in Fumed

Silica Slurries and Its Correlation with Scratch Defects Generated by CMP. *Journal of the Electrochemical Society*, vol. 153, no. 5, pp. G453-G461, 2006.

[53] Y. Homma, K. Fukushima, S. Kondo, and N. Sakuma. Effects of Mechanical Parameters on CMP Characteristics Analyzed by Two-Dimensional Frictional-force Measurement. *Journal of the Electrochemical Society*, vol. 150, no. 12, pp. G751-G757, 2003.

[54] Y. Homma. Dynamical Mechanism of Chemical Mechanical Polishing Analyzed to Correct Preston's Empirical Model. *Journal of the Electrochemical Society*, vol. 153, no. 6, pp. G587-G590, 2006.

[55] J. Luo and D. A. Dornfeld. Review of Chemical-mechanical Planarization Modeling for Integrated Circuit Fabrication: From Particle Scale to Die and Wafer Scales. *2002-2003 LMA Reports, University of California at Berkeley*, pages pp. 107–135.

[56] A. Maury, D. Ouma, D. Boning, and J. Chung. A Modification to Preston's Equation and Impact on Pattern Density Effect Modeling. *Advanced Metalization Conference, San Diego, CA*, October 1997.

[57] P. Wrschka, J. Hernandez, Y. Hsu, T. S. Kuan, G. S. Oehrlein, H. J. Sun, D. A. Hansen, J. King, and M. A. Fury. Polishing Parameter Dependencies and Surface Oxidation of Chemical Mechanical Polishing of Al Thin Films. *Journal of the Electrochemical Society*, vol. 146, no. 7, pp. 2689-2696, 1999.

[58] S. R. Runnels. Feature-scale Fluid-based Erosion Modeling for Chemical-mechanical Polishing. *Journal of the Electrochemical Society*, vol. 141, no. 7, pp. 1900-1904, July 1994.

[59] W. T. Tseng and Y. L. Wang. Re-examination of Pressure and Speed Dependence of Removal Rate during Chemical Mechanical Polishing Processes. *Journal of the Electrochemical Society*, vol. 144, pp. L15-L17, 1997.

[60] F. Zhang and A. Busnaina. The Role of Particle Adhesion and Surface Deformation in Chemical Mechanical Polishing Processes. *Electrochemical and Solid-State Letters*, vol. 1, no. 4, pp. 184-187, 1998.

[61] F. Zhang, A. A. Busnaina, and G. Ahmadi. Particle Adhesion and Removal in Chemical Mechanical Polishing and Post-cMP Cleaning. *Journal of the Electrochemical Society*, vol. 146, no. 7, pp. 2665-2669, 1999.

[62] K. Johnson. Contact Mechanics. *Cambridge University Press, Cambridge*, 1985.

[63] C. W. Liu, B. T. Dai, W. T. Tseng, and C. F. Yeh. Modeling of the Wear Mechanism during Chemical-Mechanical Polishing. *Journal of the Electrochemical Society*, vol. 143, no. 2, pp. 716-721, February 1996.

[64] F. G. Shi and B. Zhao. Modeling of Chemical-Mechanical Polishing with Soft Pads. *Appl. Phys. A*, vol. 67, pp. 249-252, 1998.

[65] T. K. Yu, C.C. Yu, and M. Orlowski. A Statistical Polishing Pad Model for Chemical-Mechanical Polishing. *International Electron Devices Meeting*, pp. 865-868, December 1993.

[66] J. A. Greenwood and J. B. P. Williamson. Contact of Nominally Flat Surfaces. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 295, no. 1442, pp. 300-319, December 1966.

[67] J. Luo and D.A. Dornfeld. Material Removal Mechanism in Chemical Mechanical Polishing: Theory and Modeling. *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 2, pp. 112-133, May 2001.

[68] K. Qin, B. Moudgil, and C. W. Park. A Chemical Mechanical Polishing Model Incorporating Both the Chemical and Mechanical Effects. *Thin Solid Films*, vol. 446, no. 2, pp. 277-286, January 2004.

[69] L. Holland. The Properties of Glass Surfaces. 1964.

[70] T. Izumitani. *Treatise on Materials Science and Technology*, vol. 17, pp. 115, 1979.

[71] K. Osseo-Asare. Surface Chemical Processes in Chemical Mechanical Polishing. *Journal of the Electrochemical Society*, vol. 149, no. 12, pp. G651-655, 2002.

[72] J. K. Vlahakis. private communication. 2006.

[73] S. Raghvendra and P. Hurat. DFM: Linking Design and Manufacturing. *18th International Conference on VLSI Design*, pp. 705- 708, January 2005.

[74] O. G. Chekina and L. M. Keer. Wear-Contact Problems and Modeling of Chemical Mechanical Polishing. *Journal of the Electrochemical Society*, vol. 145, no. 6, pp. 2100-2106, June 1998.

[75] T. Yoshida. Three-Dimensional Chemical Mechanical Polishing Process Model by BEM. *Electrochemical Society Proceedings of the Third International Symposium on Chemical Mechanical Planarization in IC Device Manufacturing*, vol. 99-37, pp. 593-604, 1999.

[76] J. J. Vlassak. A Model for Chemical-Mechanical Polishing of a Material Surface Based on Contact Mechanics. *Journal of the Mechanics and Physics of Solids*, vol. 52, pp. 847-873, 2004.

[77] T. Smith, S. J. Fang, D. Boning, G. B. Shinn, and J. A. Stefani. A CMP Model Combining Density and Time Dependencies. *Proc. CMP-MIC Conf. Conf.*, pp. 97-104, February 1999.

[78] M. Meuris J. Grillaert, N. Heylen, K. Devriendt, E. Vrancken, and M. Heyns. Modelling step height reduction and local removal rates based on pad-substrate interactions. *Proc. CMP-MIC Conf.*, pp. 79-86, February 1998.

[79] D. Boning, B. Lee, C. Oji, D. Ouma, T. Park, T. Smith, and T. Tugbawa. Pattern Dependent Modeling for CMP Optimization and Control. *Mat. Res. Soc. Symp. Proc.*, April 1999.

[80] D. O. Ouma. Modeling of Chemical Mechanical Polishing for Dielectric Planarization. *MIT Ph.D. Thsis, Department of Electrical Engineering and Computer Science*, November 1998.

[81] O. S. Nakagawa, S. Y. Oh, F. Eschbach, G. Ray, P. Nikkel, R. R. Divecha, B. E. Stine, D. O. Ouma, D. S. Boning, and J. E. Chung. Modeling of CMP-Induced Pattern-Dependent ILD Thickness Variation in Multilevel Metallization Systems. *Advanced Metalization Conference, San Diego, CA*, October 1997.

[82] H. Cai. Modeling of Pattern Dependencies in the Fabrication of Multilevel Copper Metallization. *MIT Ph.D. Thesis, Department of Materials Science and Engineering*, 2007.

[83] X. Xie and D. Boning. CMP Characterization Mask Set for Sti Processes Mask Documentation. 2003.

[84] X. Xie, T. Park, D. Boning, A. Smith, P. Allard, and N. Patel. Characterizing STI CMP Processes with an STI Test Mask Having Realistic Geometric Shapes. *Mat. Res. Soc. Symp. Proc.*, vol. 816, paper K9.4, 2004.

[85] K. Ravi. Wafer Flatness Requirements for Future Technologies. *Future Fab Int.*, vol. 7, pp. 207, 1999.

[86] C. S. Xu, E. Zhao, R. Jairath, and W. Krusell. Effects of Silicon Front Surface Topography on Silicon Oxide Chemical Mechanical Planarization. *Electrochem. Sol. State Lett.*, vol. 1, no. 4, 181-183, October 1998.

[87] B. Lee, D. S. Boning, W. Baylies, N. Poduje, and J. Valley. Modeling and Mapping of Nanotopography Interactions With CMP. *Mat. Res. Soc. Symp. Proc.*, vol. 732E, pp. I1.5.1, 2002.

[88] J. G. Park, T. Katoh, H. C. Yoo, and J. H. Park. Spectral Analyses of the Impact of Nanotopography of Silicon Wafers on Oxide Chemical Mechanical Polishing. *Jpn. J. Appl. Phys.*, vol. 40, pp. L857-860, 2001.

[89] X. Xie and D. Boning. Integrated Modeling of Nanotopography Impact in Patterned STI CMP. *Proc. CMP-MIC Conf.*, pp. 159-168, February 2003.

[90] D. Boning and B. Lee. Nanotopography Issues in Shallow Trench Isolation CMP. *MRS Bulletin*, pp. 761-765, October 2002.

[91] R. Schmolke, R. Deters, P. Thieme, R. Pech, H. Schwenk, and G. Diakourakis. On the Impact of Nanotopography of Silicon Wafers on Post-Chemical Mechanical Polished Oxide Layers. *Journal of the Electrochemical Society*, vol. 149, no. 4, pp. G257-G265, 2002.

[92] T. Tugbawa, T. Park, B. Lee, D. Boning, P. Lefevre, and L. Camilletti. Modeling of Pattern Dependencies for Multi-level Copper Chemical-mechanical Polishing Processes. *Mat. Res. Soc. Symp. Proc.*, vol. 671, pp. M4.3.1, 2001.

[93] X. Xie, D. Boning, F. Meyer, R. Rzehak, and P. Wagner. Analysis and Modeling of Nanotopography Impact in Blanket and Patterned Wafer Polishing. *Proc. CMP-MIC Conf.*, 2006.

[94] D. Castillo-Mejia, A. Perlov, and S. Beaudoin. Qualitative Prediction of $SiO_2$ Removal Rates during Chemical Mechanical Polishing. *Journal of the Electrochemical Society*, vol. 147, no. 12, pp. 4671-4675, 2000.

[95] J. Sorooshian, A. Philipossian, M. Goldstein, S. Beaudoin, and W. Huber. Impact of Wafer Geometry and Thermal History on Pressure and von Mises Stress Non-uniformity During Chemical Mechanical Planarization. *Jpn. J. Appl. Phys.*, vol. 42, pp. 6363-6370, October 2003.

[96] J. J. Vlassak. A Contact-Mechanics Based Model for Dishing and Erosion in Chemical-Mechanical Polishing. *Mat. Res. Soc. Symp. Proc.*, vol. 671, paper M4.6.1, 2001.

[97] D. L. Hetherington and D. J. Stein. Recent Advances in Endpoint and In-line Monitoring Techniques for Chemical-mechanical Polishing Processes. *Proc. CMP-MIC Conf.*, pp. 315-323, March 2001.

[98] Optima 9300 Series Chemical-mechanical Polishing Endpoint Controller User Manual. *Luxtron Corporation, Santa Clara, CA*, 1999.

[99] R. Brandes, T. Knothe, F. Klaessig, F. Menzel, W. Lortz, G. Varga, T. Shibasaki, and A. Philipossian. Metal-doped Silica Abrasive Slurries and their Effect on Friction and Removal Rate Characteristics of ILD and STI CMP. *Proc. CMP-MIC Conf.*, pp. 64-69, February 2003.

[100] B. Lee, D. S. Boning, D. L. Hetherington, and D. J. Stein. Using Smart Dummy Fill and Selective Reverse Etchback for Pattern Density Equalization. *Proc. CMP-MIC Conf.*, pp. 255-258, February 2000.