Integrationsaspekte der Simulation:
Technik, Organisation und Personal
Gert Zülch & Patricia Stock (Hrsg.)
Karlsruhe, KIT Scientific Publishing 2010

Hybride Modellierung und simulationsgestützte Optimierung in der Planung von Produktionsprozessen

Hybrid Modelling and Simulation-based Optimization in Planning of Production Processes

Ulrich Donath
Fraunhofer-Institut für Integrierte Schaltungen, Dresden (Germany)
Ulrich.Donath@eas.iis.fraunhofer.de

Sven Pullwitt
DUALIS GmbH IT Solution, Dresden (Germany)
SPullwitt@dualis-it.de

Abstract: The discrete event simulation is well established in the planning of production processes. However, an information deficit arises when sub-processes are executed according to complex recipes. For the mathematical modelling of these recipes often differential-algebraic equations (DAEs) can be used. Using the example of chemical mechanical polishing (CMP) of the semiconductor manufacturing, the integration of DAEs in the process simulation is demonstrated. Moreover, the aging of materials and the consumption of consumables are included in the modelling. A compromise of wafer yield and consumables consumption is determined by means of Pareto-optimization.

1 Einleitung

Im Rahmen der operativen Planung von Produktionsprozessen hat sich die Ablaufsimulation als wichtiges Werkzeug etabliert. Sie zeigt, ob die Aufträge in der angegebenen Reihenfolge von den vorgesehenen Maschinen und Arbeitskräften termingerecht ausgeführt werden können. Basis der Ablaufsimulation ist ein Modell der Produktion, das nach den Planungsvorgaben aufgebaut wird. State-of-the-Art Simulatoren (LINDEMAN, SCHMID 2007) stellen dazu eine Objektbibliothek zur Verfügung, deren einzelne Objekte durch grafische Symbole repräsentiert werden. Per Drag and Drop werden mit diesen Symbolen Objekte in einer Modellansicht instantiiert und nachfolgend grafisch verknüpft. Dabei werden die räumlichen Koordinaten der Produktionsanlage übernommen. Die Objekte sind entweder statische Ressourcen, wie Maschinen oder Lager, oder mobile Ressourcen, wie Operatoren

oder Transporter. Aufträge oder Produkte werden abstrakt als Entitäten aufgefasst und durch das Produktionsmodell bewegt. Die Wege werden in einem Verbindungsnetzwerk vorgegeben.

Ergebnis der bisher skizzierten Ablaufsimulation ist ein verifizierter Produktionsplan mit den Umlaufbeständen, dem Durchsatz von Produkten und der Auslastung der Ressourcen. Für die einzelnen Ressourcen werden neben dem Durchsatz die Statistik ihrer Zustände *idle*, *processing*, *blocked*, … sowie die Minima, Maxima und Mittelwerte von Inhalt und Verweilzeit aufgezeichnet.

Die Feinplanung bereitet Probleme, wenn Teilprozesse Engpässe bilden, die durch diverse Qualitäten der Produkte und unterschiedliche Prozeduren (Rezepte) zu ihrer Erzeugung entstehen. Ein Beispiel ist die chemisch-mechanische Politur (CMP) (WIKIPEDIA 2010a) in der Halbleiterfertigung. Hier wird die Polierzeit durch die Topographie des Wafers, Druck und Geschwindigkeit im Poliergerät, das Poliermittel und die Qualität des Poliertuchs bestimmt.

Die mathematischen Modelle der oben genannten Teilprozesse sind bekannt. Es sind Systeme differential-algebraischer Gleichungen (DAEs; WIKIPEDIA 2010b):

$$\underline{F}\left(\dot{\underline{x}}(t), \underline{x}(t), \underline{p}, t\right) = \underline{0} \mid [t_0, t_S],\ \underline{x}(t_0)\ . \qquad (1)$$

Die Verfasser integrieren diese Gleichungen und ihre Lösungen in die Ablaufsimulation. Die Ausführung des so konstruierten hybriden Modells zeigt detailliert den Produktionsablauf für den dedizierten Planungsabschnitt. Am Beispiel der Halbleiterfertigung, speziell des CMP, wird im Folgenden das Vorgehen bei der hybriden Modellierung demonstriert. Der Ablaufsimulator wird außerdem mit einem Optimierungswerkzeug gekoppelt. Dieses wird benutzt, um die Zielfunktionen der Produktion für einen breiteren Planungshorizont zu sichern.

2 Ereignis-diskretes Ablaufmodell der Chip-Fertigung

Die Halbleiterherstellung wird in drei Abschnitte gegliedert:

- Wafer-Herstellung
- Front-End-Prozess: Realisierung der elektrischen Funktionen der Chips
- Back-End-Prozess: Vereinzeln der Chips und Montage in Gehäusen.

Im Front-End-Prozess werden die elektrischen Funktionen der Chips in Planartechnik erzeugt, indem man in mehreren aufeinanderfolgenden Prozessschritten Schichten mit spezifischen elektrischen Eigenschaften übereinander auf den Wafer aufbringt (HILLERINGMANN 2008). Der Zyklus der Prozessschritte umfasst Lithographie, Ätzen, Dotieren, Reinigen, Abscheiden und Planarisieren.

Der Produktionsplan beschreibt die Reihenfolge der Prozessschritte. Je nach Prozessschritt werden die Fertigungsgeräte (Tools) mit einem oder mehreren Fertigungslosen, bestehend aus einem oder mehreren Wafern, beladen. Die Fertigungslose werden in hermetisch abgeschlossenen Transporthilfsmitteln zwischen den Fertigungsgeräten transportiert und gelagert. Die Fertigungsgeräte sind in ver-

schiedenen Zonen (Bay) gruppiert, so dass je nach Standort und Prozessschritt ein Transport innerhalb einer Zone oder zwischen den Zonen erforderlich wird.

Im ereignis-diskreten Modell des Front-End-Prozesses werden die Fertigungsgeräte durch Prozessoren mit definierten Bearbeitungszeiten (Process Times) abgebildet, die als Konstanten vorgegeben oder als Zufallsgrößen nach statistischen Verteilungsfunktionen generiert werden. Dem Beispiel, das hier eingeführt wird, ist zu Grunde gelegt, dass die Fertigungslose einzelne Wafer sind und die Geräte den jeweiligen Prozessschritt *wafer-by-wafer* ausführen. Jedem Prozessor ist eine Warteschlange vorgelagert, welche die Lagerkapazität der Ladestation des Fertigungsgerätes repräsentiert.

Der Transport von Fertigungsgerät zu Fertigungsgerät kann in unterschiedlicher Abstraktion modelliert werden:

a) Der Wafer wird ohne Hilfsmittel ohne Verzögerung von der Quelle zum Zielgerät übertragen.

b) Der Wafer wird von einem Transporter auf direktem Weg von der Quelle zum Zielgerät transportiert. Dabei werden die durchschnittlichen Transportzeiten (Transport Times) abgefragt, die vom realen System übernommen wurden.

c) Nach dem Vorbild des Produktionsprozesses wird ein Wege-Netz gebildet und entsprechend dem Verkehrsaufkommen im realen System (Anzahl der Transporter, Geschwindigkeiten, Abstände) parametriert. Der Wafer wird von einem Transporter auf dem Wege-Netz von der Quelle zum Zielgerät transportiert, wobei die Transportzeiten berechnet werden.

Das ereignis-diskrete Modell des Front-End-Prozesses wird gebildet, indem die einzelnen Prozessoren in der Modellansicht des Ablaufsimulators instantiiert und nach dem Produktionsplan verkoppelt werden. Ergänzt werden die Quellen und die Senken der Wafer, die nach der Statistik des vorgelagerten bzw. nachgelagerten Produktionsabschnitts erzeugt bzw. konsumiert werden. Der Transport wird nach der oben genannten Variante a) oder b) integriert. Die Wafer selbst werden abstrakt als Entitäten behandelt, die bei ihrer Erzeugung eine *identity number* und *wildcard attributes* erhalten, die im Prozessverlauf belegt werden.

3 Kontinuierliche Modelle für Teilprozesse

3.1 Modell für das Planarisieren

Über metallische Leiterbahnen oder andere aktive Gebiete des Chips wird Siliziumoxid SiO_2 aufgetragen, um diese von der nächsten aktiven Schicht zu isolieren. Dabei entsteht in Abhängigkeit von der Geometrie der überdeckten Strukturen ein Stufengebirge, dessen einzelne Erhebungen unterschiedlich dicht aneinander liegen. Ohne eine weitere Maßnahme könnte sich mit dem Auftrag der nächsten aktiven Schicht die Höhe des Stufengebirges akkumulieren. Als Folge sind Fehlfunktionen des Chips möglich. Die Oxidschicht wird daher vor dem nächsten Prozessschritt durch eine chemisch-mechanische Politur eingeebnet.

Beim CMP (Abbildung 1) wird der Wafer mit der zu bearbeiteten Oberfläche nach unten durch Unterdruck an einen Halter gezogen. Mit dem Halter wird der Wafer in Rotation versetzt und mit Druck über einen Poliertisch geführt, der sich ebenfalls dreht. Der Poliertisch trägt ein Poliertuch, auf dem ein flüssiges Poliermittel verteilt wird. Das Poliermittel enthält chemische Stoffe und Abrasivpartikel, die unter Druck mit dem Oxid an der Oberfläche reagieren und dieses durch die Rotationsbewegung abtragen.

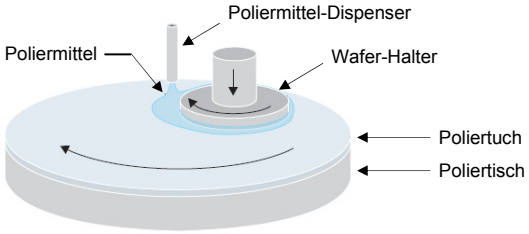

Abbildung 1: *Elemente des CMP (IC KNOWLEDGE 2010)*

Die Planarisierung des Wafers wird per Chip durch eine Menge von Differential-Gleichungssystemen (XIE 2007) beschrieben, die qualitativ identisch sind. Die einzelnen Gleichungssysteme (2) beschreiben die Schichtdicken- und Stufenhöhen-Reduzierung (Abb. 2) in Segmenten des Chips, die in einer (x,y)-Diskretisierung gebildet werden.

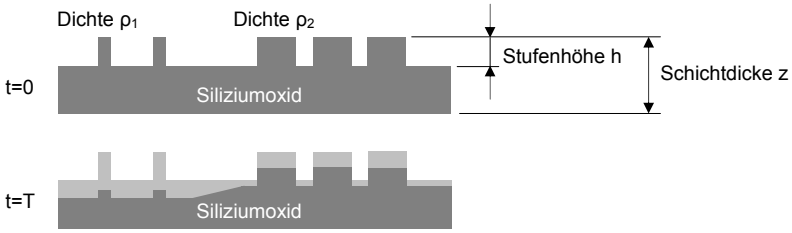

Abbildung 2: *Planarisierung des Wafers (MEYER 2003)*

$$
\begin{aligned}
d\left(z\left(t;x,y\right)\right)/dt &= -K_u\left(\rho,z,h\right)\\
d\left(h\left(t;x,y\right)\right)/dt &= -K_u\left(\rho,z,h\right)+K_d\left(\rho,z,h\right) \qquad (2)\\
K_d &= K_0\times e^{-\rho h/h_c}\\
K_u &= \frac{1}{\rho}K_0-\frac{1-\rho}{\rho}K_0\times e^{-\rho h/h_c}
\end{aligned}
$$

mit K_u und K_d Abtragsraten in up- und down-Gebieten und K_0 Abtragsrate für die plane Fläche.

Die Gleichungssysteme (2) werden gelöst, wenn der Wafer vom CMP-Prozessor aufgenommen wird. Vom vorgelagerten Prozessor sind die *wildcard attributes* des Wafers mit den Anfangswerten der Schichtdicken und Stufenhöhen $z_0(x,y)$ =

$z(t=0; x,y)$ und $h_0(x,y) = h(t=0; x,y)$ sowie der Dichten der Stufen $\rho(x,y)$ belegt worden.

3.2 Modellierung von Alterung und Hilfsstoff-Verbrauch

Die in (2) benutzte Abtragsrate für die plane Fläche K_0 kann nur für den einzelnen Prozessschritt als konstant betrachtet werden.

In Realität gilt für die Abtragsrate K_0:

$$K_0 = K_0(t) = c \times P \times V \tag{3}$$

mit P Wafer-Anpress-Druck, V Relativgeschwindigkeit von Wafer und Poliertisch und c Materialkennwert.

Während P und V entsprechend der Rezeptur des Prozessschrittes eingestellt werden, fasst der Materialkennwert c die Qualitäten des Poliermittels und des Poliertuchs zusammen.

Untersuchungen zeigen, dass der Materialkennwert mit

$$c = c(t) = c_0 \times e^{-d \times t_{elapsed}} \tag{4}$$

näherungsweise bestimmt werden kann, mit c_0 Initialwert, $t_{elapsed}$ akkumulierter Wert der Bearbeitungszeiten und d Skalierungsfaktor.

$c(t)$ und $K_0(t)$ werden in der Simulation nach jedem Lauf des CMP-Prozessors nach (4) und (3) neu berechnet. Entsprechend der Lösung von (2) reduziert sich der relative Abtrag auf dem Wafer. Bei vorgegebener Zielfunktion muss demzufolge die Bearbeitungszeit verlängert werden oder die Einflussgrößen P oder V müssen erhöht werden.

Die Verlängerung der Bearbeitungszeiten zieht eine Erhöhung des Poliermittelverbrauchs nach sich. Während der Simulation wird nach jedem Lauf eines CMP-Prozessors k der Poliermittelvorrat neu berechnet:

$$S = S - \sum_k s_{rate}^k \times t_{process}^k \tag{5}$$

mit S Poliermittel-Vorrat, s_{rate} Poliermittel-Verbrauchsrate und $t_{process}$ Bearbeitungszeit. Unterschreitet der Poliermittel-Vorrat ein Minimum, werden alle involvierten CMP-Prozessoren suspendiert.

3.3 Modellierung von Bearbeitungsstrategien

Die Planarisierung des Wafers soll in zwei Schritten ausgeführt werden. Im ersten Schritt wird eine grobe Glättung in einer konstanten Bearbeitungszeit vorgenommen. In einem zweiten Schritt wird die Planarisierung so lange ausgeführt, bis das Stufengebirge bis zu einem Grenzwert abgetragen ist. Die Bearbeitungszeit dafür ist variabel, wird jedoch auf ein Maximum beschränkt. Es soll ein CMP-Gerät verwendet werden, das für den ersten Schritt einen Poliertisch (Prozessor A) und für den zweiten Schritt zwei Poliertische (Prozessor B1, Prozessor B2) zu Verfügung stellt (Abb. 3).

Unter dem Aspekt der Alterung wird für Prozessor A gefordert, dass der relative Abtrag q ein Minimum q_{min} überschreitet:

$$q = 1 - \sum_i h_i \; / \; h_{i_0} > q_{\min} \tag{6}$$

mit h_{i0} Stufenhöhe vor Planarisierung und h_i Stufenhöhe nach Planarisierung.

Für die Prozessoren B1 und B2 gilt, dass

$$\forall h_i < h_{\max} \quad \text{in} \quad t_{process} < t_{\max} \tag{7}$$

erreicht wird.

Die Verletzung der Bedingungen (6) und (7) werden einem Dispatcher gemeldet. Eine Menge von parallelen Zustandsautomaten beschreibt dessen Reaktionen. In den beiden genannten Fällen wird eine Poliertuch-Aufbereitung bilanziert, indem für ihre Dauer die Ports der involvierten Prozessoren geschlossen werden. Im Fall der Grenzwertunterschreitung des Poliermittelvorrats (5) werden die Ports aller Prozessoren für die Dauer der Poliermittelbeschaffung und -aufbereitung geschlossen.

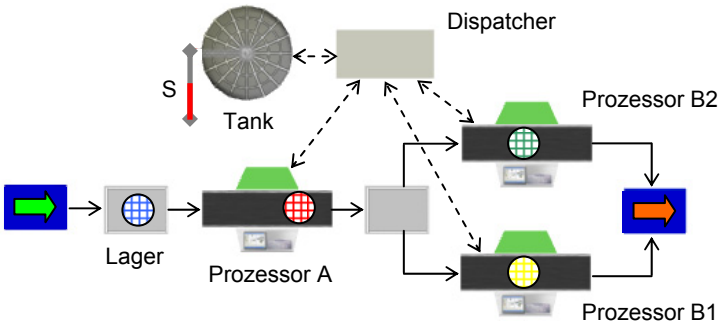

Abbildung 3: *Modell des CMP-Prozesses*

4 Pareto-Optimierung

Der Front-End-Prozess der Wafer-Fertigung soll hinsichtlich der Zielgrößen Wafer-Ausbringungsmenge und Kosten optimiert werden. Ein wesentlicher Kostenanteil wird durch den Poliermittelverbrauch beim Planarisieren der Wafer hervorgerufen, so dass die Optimierung zuerst für diesen Teilprozess vorgenommen wird.

Für das Planarisieren gelten damit die Ziele:

$$\#Wafer \rightarrow Maximum \quad \text{und} \quad Poliermittelverbrauch \rightarrow Minimum \; . \tag{8}$$

Die Maximierung der Wafer-Ausbringung und die Minimierung des Poliermittelverbrauchs in (8) werden mit den Gewichten 0.75 und 0.25 verfolgt. Freiheitsgrade für die Optimierung sind der Grenzwert q_{min} für den relativen Abtrag durch Prozessor A und die Zeitschranken t_{max} der Bearbeitungszeit für den Abtrag durch die Prozessoren B1 und B2.

Dabei gelten folgende Zusammenhänge:

- Je weiter die Toleranzgrenzen für den minimalen Abtrag und die maximale Bearbeitungszeit sind, desto mehr Poliermittel wird verbraucht.
- Je enger die Toleranzgrenzen liegen, desto eher muss das Poliertuch aufbereitet werden.
- Bei hohem Poliermittelverbrauch muss der Poliermitteltank nachgefüllt werden, was den Stillstand der Prozessoren verursacht und damit die Wafer-Ausbringungsmenge verringert.

Das Optimierungswerkzeug (KRUG 2002) kombiniert wahlfrei deterministische, stochastische und evolutionäre Suchstrategien. Die Suchstrategien werden parallel ausgeführt. Ergebnis der Suche sind die Pareto-Menge der Zielgrößen (Abbildung 4) und die optimalen Parameter für den CMP-Prozess (Tabelle 1).

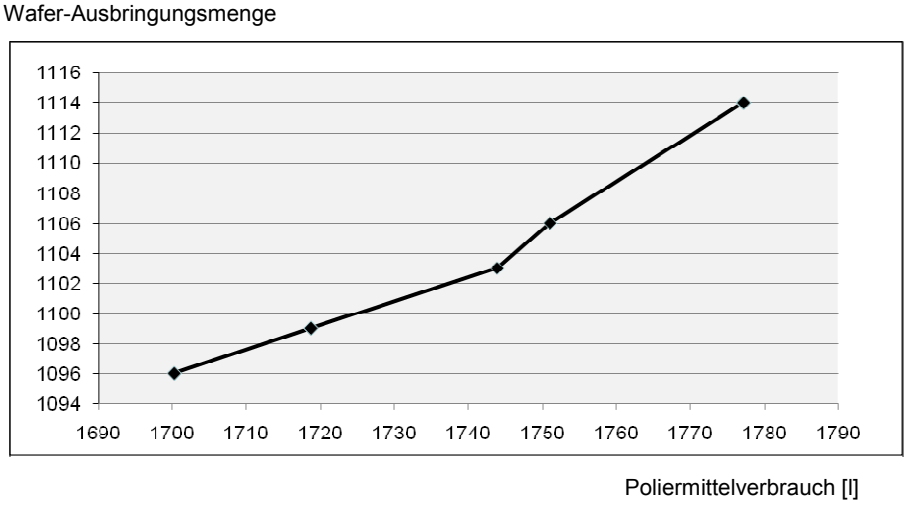

Abbildung 4: *Pareto-Menge der Zielgrößen*

Minimaler relativer Abtrag Prozessor A	Maximale Prozesszeit Prozessor B1 [s]	Maximale Prozesszeit Prozessor B2 [s]	Poliermittel-verbrauch [l]	Wafer-Ausbringung [#]
0.6	130	121	1700	1096
0.6	136	121	1718	1099
0.6	139	122	1751	1106
0.6	123	149	1743	1103
0.6	134	140	1777	1114

Tabelle 1: *Optimale Parameter für den CMP-Prozess*

5 Ausblick

Der Gleichungslöser für die in Abschnitt 3.1 exemplarisch dargestellte Problemstellung liegt als dynamische Bibliothek (Dynamic Link Library, DLL) vor. Diese Gleichungslöser-DLL kann über eine Interface-DLL von jedem Windows-basierten Ablaufsimulator geladen werden. Auf dieser Basis wurde die hybride Simulation des CMP-Prozesses mit den Simulatoren Flexsim (FLEXSIM 2010) und SpeedSIM (DUALIS 2010) ausgeführt. Die beiden Zugänge bestätigen die Portabilität der IT-Lösung. Die Arbeiten werden in Richtung der Integration und dem Test von Run-to-Run-Steuerungen und der Abbildung von Prozessleitsystemen bzw. SCADA-Systemen fortgesetzt. Die Kombination mit der multi-kriteriellen Optimierung ermöglicht dem Planer, den Produktionsprozess abschnittsweise zu optimieren.

Das Projekt wurde im Rahmen „Innovative technologieorientierte Verbundprojekte auf dem Gebiet der Zukunftstechnologien im Freistaat Sachsen“ vom Freistaat Sachsen und dem Europäischen Fonds für Regionale Entwicklung (EFRE) gefördert.

Literatur

DUALIS: SPEEDSIM. Der schnelle Simulator. http://www.dualis-it.de/downloads/Info-SpeedSIM-1106.pdf, Stand: 06.05.2010.

FLEXSIM: Flexsim 3D Simulation Software. http://www.flexsim.com/products/flexsim/FlexsimBrochure.pdf, Stand: 06.05.2010.

HILLERINGMANN, Ulrich: Silizium-Halbleitertechnologie. Wiesbaden: Vieweg + Teubner, 5. Auflage 2008.

IC KNOWLEDGE: Chemical Mechanical Planarization. http://www.icknowledge.com/misc_technology/CMP.pdf, Stand: 06.05.2010.

KRUG, Wilfried: Modellierung, Simulation und Optimierung für Prozesse der Fertigung, Organisation und Logistik. Delft u.a.: SCS, 2002. S. 160-180.

LINDEMAN, Marcus; SCHMID, Simone: Simulationswerkzeuge in Produktion und Logistik. In: PPS Management, Berlin, 12(2007)2, S. 28-35.

MEYER, Frank u.a.: Vom lokalen Stufenhöhenmodell zur erreichbaren Post CMP Topographie. In: Proceedings 11. CMP Nutzertreffen, Itzehoe, 2003. http://www.isit.fraunhofer.de/Veranstaltungen/Nutzertreffen

WIKIPEDIA: Halbleitertechnik. http://de.wikipedia.org/wiki/Halbleitertechnik, Stand: 06.05.2010. (=2010a)

WIKIPEDIA: Differential-algebraische Gleichung. http://de.wikipedia.org/wiki/Differential-algebraische_Gleichung, Stand: 06.05.2010. (=2010b)

XIE, Xiaolin: Physical Understanding and Modeling of Chemical Mechanical Planarization in Dielectric Materials. Cambridge, MA: Ph. D. Thesis, Massachusetts Institute of Technology. 2007. pp 109-111 and 120-125