Evaluation of Equipment Models of Clustered Photolithography Tools for Fab-level Simulation

Jung Yeon Park, James R. Morrison, and Kyungsu Park

Department of Industrial and Systems Engineering
KAIST, South Korea
Presentation Overview

• Motivation

• System Description: Clustered Photolithography Tool (CPT)

• Equipment Models
  ➢ Linear model
  ➢ Affine models
  ➢ Flow line models (Improved)

• Numerical Experiments
  ➢ Description (Three types of simulation)
  ➢ Results

• Concluding Remarks
Motivation
Motivation (1) - CPT

• Clustered photolithography tools (CPT)
  • Cost up to $120 million\textsuperscript{[1]}, typically $20 – 50 million
  • Often the fabricator bottleneck
  • Key contributor to fab throughput capacity and cycle time
Motivation (2) – Fab-level Simulation

• High construction costs
  • Fabs must be well-designed and operated efficiently

• Fab-level Simulation
  • Essential decision support technology
  • Examples:
    • Detailed AMHS models (Jimenez et al. 2008, Hsieh et al. 2012)
    • Studies of fab behavior in relation to changes in lot size (Schmidt et al. 2006)
    • Cycle time reduction (Zarifoglu et al. 2008)

• Equipment models are key components of fab-level simulation
Motivation (3) – Features of Equipment Models

• **Tradeoff: Fidelity vs. complexity**
  • More detailed models lead to greater fidelity, but require longer computation times
  • Require more modeling effort
• **Fab conditions can often change (different lot sizes, new toolsets, changing product mix, etc.)**
  • Models often trained on specific set of input data, may not be robust when input conditions change

• **Goal:** Comparison of CPT Models for use in fab-level simulation
  • **Accurate:** Predict throughput with less than 1% error
  • **Expressive:** Incorporate fundamental behaviors
  • **Computation:** Very quick to calculate results
  • **Robust:** Less dependent on input data
System Description
**System Description (1) – CPT**

**Clustered Photolithography Tool**

- Multi-cluster tool, robot in each cluster, IF buffers, STK buffer
- Scanner is often the CPT bottleneck
- Largely deterministic process times
- Process time can vary by product
- Setups between lots (reticle changes, pre-scan setup, ...)
- Wafer handling robot decision policy
System Description (2) – Performance Metrics

- **Notation**
  
  $a_i$: Arrival time of lot $i$ to the tool
  $S_i$: Start time of lot $i$ in the tool
  $C_i$: Completion time of lot $i$ from the tool

  Lot class: $k_1 \in \{1, \ldots, K\}$
  Number of wafers in lot $i$: $W_i$

- **Performance measures**
  
  Cycle time of lot $i$: $CT_i = C_i - a_i$
  Lot residency time of lot $i$: $LRT_i = C_i - S_i$
  Throughput time of lot $i$: $TT_i = \min(C_i - S_i, C_i - C_{i-1})$

  ![Diagram showing computation times and throughput times for three lots](image)
Equipment Models
Equipment Model (1) – Linear Model

- Referred to as the Ax equipment model or linear model
- Used to study recipe dedication in CPTs in an ASIC fab model[3]
- Lot indices per class: $L(k_1) = \{i | k(i) = k_1\}$
- $k_1$ is current lot class
- Time between wafer completions: $A^{k_1}$
- Parameter estimation:
  
  \[
  A^{k_1} = \frac{\sum_{i \in L(k_1)} (c_i - \max(a_i, c_{i-1}))}{\sum_{i \in L(k_1)} w_i}
  \]

**Complete Model:**

\[
\tilde{s}_i = \max\{a_i, \tilde{c}_{i-1}\}
\]

\[
\tilde{c}_i = \tilde{s}_i + A^{k_1} \times w_i
\]
Equipment Model (1) – Linear Model

- **Pros:**
  - Simple to understand
  - Fast computation

- **Cons:**
  - Exactly matched to single module tool, not for CPT
  - New lots enter only when the tool is empty (No parallelism)
Equipment Model (2) – Affine Model

- Referred to as the Ax+B model

\[ A^{k_1} = \frac{\sum_{i \in L(k_1)} (C_i - \max(a_i, C_{i-1}))}{\sum_{i \in L(k_1)} W_i} \]

• Basic model provided in AutoSched AP \[4\]

Lot indices per pairs of classes:

\[ L(k_1, k_2) = \{ i | k(i) = k_1, k(i-1) = k_2 \} \]

\[ S_i = \max(\tilde{a}_i, C_{i-1}) \]

\[ \tilde{C}_i = \tilde{S}_i + A^{k_1} \times (W_i - 1) + B^{k_1, k_2} \]

Complete model:

- First wafer delay:

\[ B^{k_1, k_2} \]

- Time between wafer completions:

\[ A^{k_1} \]

B is generalized to consider setups between classes
Equipment Model (2) – Affine Model

- **Pros:**
  - Simple to understand
  - Fast computation

- **Cons:**
  - Only one module per process, so not matched to CPT
  - New lots enter only when the tool is empty (No parallelism)
Equipment Model (3) – Flow Line Models

- Have been used for optimization and simulation modeling studies ([5] – [8])
- Series of sequential processes $P_1, ..., P_M$
- Buffers modeled as zero process time modules
- Fundamental assumption: CPT is process-bound
- Modifications for CPT modeling
  - Consider robotic workload in process times of modules
  - Consider setups – reticle setup, pre-scan setup
  - Different number of processes for different lot classes
- Two types of flow lines
  - Parametric flow line (PFL) : Known process times
  - Empirical flow line (EFL) : Unknown process times
Equipment Model (3) – Flow Line Models

**Notation**
- \( a_w \): Arrival time of wafer \( w \) to the tool, \( a_w \leq a_{w+1} \)
- \( X_{w,m} \): Entry time of wafer \( w \) into process \( m \) of the tool
- \( R(k, m) \): number of identical servers for process \( m \) for wafer class \( k \)
- \( \tau_m^k \): Deterministic process time for process \( m \) for wafer class \( k \)

**Modified Process Times**

<table>
<thead>
<tr>
<th>Parametric FL</th>
<th>Empirical FL</th>
</tr>
</thead>
</table>
| \[
S(k, m) = \begin{cases} 
\tau_m^k + 3\delta + 4\varepsilon, & m = PB \\
\tau_B^k + 2\delta + 4\varepsilon, & m = B \\
\tau_B^k + \delta + 2\varepsilon, & \text{otherwise}
\end{cases}
\] | \[
S(k, m) = \begin{cases} 
\frac{\sum_{i \in \mathbb{E}(k)} \sum_{w=2}^{W(k)} (X_{w,B+1} - X_{w-1,B+1})}{\sum_{i \in \mathbb{E}(k)} (W_l - 1)}, & m = B \\
\min_{w,k(w) = k} (C_w - X_{w,m}), & m = M \\
\min_{w,k(w) = k} (X_{w,m+1} - X_{w,m}), & \text{otherwise}
\end{cases}
\] |

**Elementary Evolution Equations**
- \( \bar{X}_{w,1} = \max\{a_w, \bar{X}_{w-R'(k,1),P(w)+1}, \bar{X}_{w-1,1}\} + \tau_s'(w, m) \)
- \( \bar{X}_{w,m} = \max\{\bar{X}_{w,m-1} + S(k, m - 1) + \tau_R'(w, m), \bar{X}_{w-R(k,m),m+1}, \bar{X}_{w-1,m}\} \)
- \( \bar{X}_{w,M} = \max\{\bar{X}_{w,M-1} + S(k, M - 1), \bar{X}_{w-R(k,m),M} + S(k, M), \bar{X}_{w-1,M}\} \)

**Start and Completion Times**
- \( \bar{S}(i) = \bar{X}_{\Omega(i,1),d(k)} \)
- \( \bar{C}(i) = \bar{X}_{\Omega(i,W(i)),M} + S(k, M) \)
Numerical Experiments
Numerical Experiments (1) – Detailed Model

- Use CPT data from industry\(^9\); create detailed CPT model using discrete event simulation
- Two interface buffers (IF), one pre-scan buffer (STK)
- Longest waiting pair (LWP) robot policy\(^{10}\): gives optimal steady state throughput
- Robot move time: 3s, pick/place time: 1s
- Deadlock avoidance rule
- Reticle alignment setup (for every lot) ~ Unif[240, 420]
- Pre-scan track setup (for lot class change) ~ Unif[210, 260]
- 15,000 lots × 30 replications
- Detailed model assumed to be exact
Numerical Experiments (2) – Simulation Description

- **Type I simulation**: Compare accuracy and computation
  - Models trained on one set of input data
  - Simulated on same set of data

- **Type II simulation**: Predict performance at current operating conditions
  - Models trained on one set of input data
  - Simulated on different set of data with same parameters

- **Type III simulation**: Robustness to changed operating conditions
  - Models trained on one set of input data
  - Simulated on different set of data with different parameters
  - Vary operating parameters: train size, lot size, loading level, pre-scan buffer size, penultimate bottleneck’s process times, pre-scan module process times, etc.

- LM - Linear Model, AF – Affine Model, PFL – Parametric flow line, EFL – Empirical flow line, DS – Detailed Model

Common practice to increase tool throughput
Numerical Experiments (3) – Type I Results

Type I Simulations

- LM and AF are trained on throughput but cannot handle parallelism.
- FL models are accurate on all metrics: CT, LRT, and TT.

All models can calculate throughput time accurately (<0.2%)

LM and AF are not accurate for both CT and LRT (~60%)
Numerical Experiments (3) – Type I Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Computation Time (ms)</th>
<th>Scaled Computation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>143</td>
<td>1</td>
</tr>
<tr>
<td>AF</td>
<td>177</td>
<td>1.24</td>
</tr>
<tr>
<td>PFL</td>
<td>83183</td>
<td>581</td>
</tr>
<tr>
<td>EFL</td>
<td>83596</td>
<td>584</td>
</tr>
<tr>
<td>DS</td>
<td>17372368</td>
<td>121273</td>
</tr>
</tbody>
</table>

- Tradeoff between accuracy and computational complexity
  - FL models approximately 500 times more complex than LM or AF
  - FL models still 200 times less computationally complex than DS
- Type II simulations show similar results to Type I
  - LM and AF correctly predict TT, not for CT or LRT
Numerical Experiments (3) – Type III Results

Type III Simulations
Baseline conditions: Train size = 3, lot size = {23, 24, 25}, loading = 0.95

LM and AF become inaccurate for TT (~5% for loading)

LM and AF can have TT percent errors up to 80% (for lot size)

- LM and AF cannot accurately predict TT anymore; inaccurate on all metrics
- FL models are accurate on all metrics again \(\rightarrow\) robust
## Numerical Experiments (3) – Accuracy Comparison

<table>
<thead>
<tr>
<th></th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linear Model</strong></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td><strong>Affine Model</strong></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td><strong>Flow Line Models</strong></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>LRT</td>
<td>TT</td>
</tr>
</tbody>
</table>

- Errors relative to detailed model
  - Error of 20%+
  - Error 5-20%
  - Error 0-5%
Concluding Remarks
Concluding Remarks

• CPT: Expensive & typically fab bottleneck

• Equipment models for CPT
  • Linear, affine, flow line models, and detailed model
  • Extension of affine model
  • Propose new method to compute module processing times for flow line
  • Assess models’ fidelity on cycle time, lot residency time, and throughput time
  • Robustness to changed operating conditions

• Future work
  • Improved models: Newer exit recursions, additional parameters
  • Implementation: Fab simulation, process optimization, etc.
References


