Congestion-Aware Power Grid Optimization for 3D circuits Using MIM and CMOS Decoupling Capacitors

Pingqiang Zhou
Karthikk Sridharan
Sachin S. Sapatnekar

University of Minnesota
Outline

- Motivation
- A new CAD solution to 3D power grid optimization
- Experimental results
- Summary
3D IC Design

Adapted from [Das et al., ISVLSI, 2003] by B. Goplen
3D Integration: Driving forces

- Improved global interconnect performance
- Reduce footprint / improve packing density
- Mixed-signal integration

[Al-Sarawi et al., 1998]
Power Supply Integrity in 3D

- Higher current density, faster current transients worsen supply noise
- Greater challenge in 3D due to via resistance, limited number of supply pins

**Power bottleneck:** a major problem for 3D

---

[IBM]
Traditional power delivery

• Requirements
  – $V_{dd}$, GND signals should be at correct levels (low V drop)
  – Electromigration constraints
    • Current density must never exceed a specification
    • For each wire, $I_i/w_i < J_{spec}$
  – $dl/dt$ constraints
    • Need to manage $dl/dt$ to reduce inductive effects

• Techniques for meeting constraints
  – Widening wires
  – Using appropriate topologies
  – Adding decoupling capacitances

• Already challenged for 2D technologies
  – Reliable power delivery hard
  – Decaps get leaky

• New CAD approaches necessary
Decoupling capacitances (decaps)

- The most powerful method to reduce transient noise
- Conventional decap technology: CMOS decap
- New considerations for CMOS decaps in 3D
  - Compete for area on device layer with landing pads of 3D vias
  - May increase footprint size
  - Get more leaky, due to T-leakage feedback

Any other option?
MIM decaps

![Power Grid Upper Supply Buss](image)

- MIM capacitor is implemented over metal!

---

<table>
<thead>
<tr>
<th>Decap</th>
<th>*Capacitance density (fF/µm²)</th>
<th>*Leakage density (A/cm²)</th>
<th>Congestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMOS</td>
<td>17.3</td>
<td>1.45e-4</td>
<td>—</td>
</tr>
<tr>
<td>MIM</td>
<td>8.0</td>
<td>3.2e-8</td>
<td>routing blockage</td>
</tr>
</tbody>
</table>

* Numbers deduced from Roberts et al., IEDM05 and PTM simulations
Our Contributions

- Apply newer decap technology - MIM decap
- Develop CAD solutions for inserting both MIM and CMOS decaps:
  - Sequence of linear programming based problem formulation
  - Linearized noise model based on adjoint sensitivity analysis
  - 3D congestion analysis and linear congestion model
Overall Algorithm Flow

Initial setup

3D layout info.  Technology parameters

Build 3D power grid

Transient power grid analysis

Noise metric \( S \neq 0 \)?

Optimization loop

NO  Stop

YES

LP based allocation of CMOS and MIM decaps
Power network modeling and analysis

● Power Network Modeling

● Modified Nodal Analysis

\[ G \mathbf{x}(t) + C \mathbf{x}'(t) = \mathbf{b}(t) \]

- \( \mathbf{x}(t) \): time varying vector of voltages and currents
- \( \mathbf{b}(t) \): time varying vector of independent current sources

● Adjoint Sensitivity Analysis

- Based on Tellegen’s theorem: the instantaneous power in any circuit is zero
- An approach to calculate the sensitivity of one objective function w.r.t all the parameters in the circuit
Power Noise Metric $S$

- Noise: optimize the integral of noise violation over time

\[
\begin{align*}
S(j) &= \int_{t_s}^{t_e} (V_{th} - V_j(1,p))dt \\
S &= \sum_j S(j)
\end{align*}
\]
Decap optimization: problem formulation

minimize $\alpha S(x_k, y_k) + (1-\alpha) P(x_k, y_k)$
subject to

$0 \leq x_k \leq C_{CMOS}^k$

$0 \leq y_k \leq C_{MIM}^k$

Congestion in grid $k \leq 1$

- $x_k$: CMOS decap added to grid $k$
- $y_k$: MIM decap added to grid $k$

Nonlinear optimization problem!
Sequence of linear programs: formulation

- **Objective**
  
  \[
  \min \alpha \Delta S + (1-\alpha) \Delta P
  \]

  - \(\Delta S = \sum_k (a_k \Delta x_k + b_k \Delta y_k)\) = change of violation area \(S\)
  - \(\Delta P = \sum_k (c_k \Delta x_k + d_k \Delta y_k)\) = change in leakage
  - \(\Delta x_k\): Newly added CMOS decap to grid \(k\)
  - \(\Delta y_k\): Newly added MIM decap to grid \(k\)

- **Constraints**
  - **Congestion constraint**
    
    \(\Delta \text{Cong}_k \leq \gamma \cdot \text{Cong}_k\)
  - **Decap resource constraint**
    
    \[
    0 \leq \Delta x_k \leq \min\{\Delta_{CMOS}, C_{CMOS}^k\} \\
    0 \leq \Delta y_k \leq \min\{\Delta_{MIM}, C_{MIM}^k\}
    \]
Congestion Analysis and Linear Model

● 3D congestion analysis


\[
\begin{cases}
F(p,q,r) = F(p-1,q,r) + F(p,q-1,r) + F(p,q,r-1) \\
F(p,1,1) = F(1,q,1) = F(1,1,r) = 1
\end{cases}
\]

● Linear congestion model

\[
\Delta \text{Cong}_k = \left( \sum_{i \in R_k, i \neq k} \frac{\Delta W_{k,i}}{\text{Cap}_k} \right) + \lambda_k \cdot \Delta y_k \\
= \sum_{i \in R_k} (\lambda_i \cdot \Delta y_i)
\]

- \(\Delta y_i\): the small MIM decap added to grid \(i\)
- \(\Delta w_{k,i}\): the # of routes moved out of grid \(i\) to grid \(k\) caused by \(\Delta y_i\)
- \(\text{Cap}_i\): the capacity of grid \(k\)

[Details in the paper]
Experimental Setup

- 90nm technology node
- 6 metal layers for each 2D tier
- Supply voltage: 1.2 V
- Voltage drop threshold: 0.12 V (10%)

### 3D Benchmarks

<table>
<thead>
<tr>
<th>Ckt</th>
<th># Nodes</th>
<th>Worst V droop (V)</th>
<th># nodes with noise violations</th>
<th>Violation Area S (V · ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ibm123</td>
<td>18,634</td>
<td>0.135</td>
<td>3330</td>
<td>13.739</td>
</tr>
<tr>
<td>ibm05</td>
<td>12,026</td>
<td>0.122</td>
<td>1359</td>
<td>72.260</td>
</tr>
<tr>
<td>ibm08</td>
<td>17,030</td>
<td>0.125</td>
<td>3191</td>
<td>41.305</td>
</tr>
<tr>
<td>ibm10</td>
<td>29,262</td>
<td>0.159</td>
<td>5935</td>
<td>91.286</td>
</tr>
<tr>
<td>ibm18</td>
<td>75,042</td>
<td>0.163</td>
<td>6392</td>
<td>108.649</td>
</tr>
</tbody>
</table>
Experimental Results

• Comparison of three optimization strategies

<table>
<thead>
<tr>
<th>Ckt</th>
<th>CMOS only</th>
<th>MIM only</th>
<th>CMOS + MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VNs</td>
<td>S (V · ns)</td>
<td>Lkg (mA)</td>
</tr>
<tr>
<td>IBM123</td>
<td>368</td>
<td>0.023</td>
<td>2.1</td>
</tr>
<tr>
<td>IBM05</td>
<td>24</td>
<td>0.049</td>
<td>2.7</td>
</tr>
<tr>
<td>IBM08</td>
<td>31</td>
<td>0.010</td>
<td>1.2</td>
</tr>
<tr>
<td>IBM10</td>
<td>351</td>
<td>0.182</td>
<td>1.6</td>
</tr>
<tr>
<td>IBM18</td>
<td>130</td>
<td>0.071</td>
<td>2.7</td>
</tr>
</tbody>
</table>

- VNs: number of violating nodes
- Lkg: leakage current
- maxC: maximum increment of congestion
- avgC: average increment of congestion

**CMOS+MIM** can achieve a good tradeoff between leakage power and routing congestion
Experimental results: ibm18 (cont.)

- **Violation Area**

- **Leakage**
## Experimental Results (cont.)

### Optimization results of power grid densities

<table>
<thead>
<tr>
<th>Cases</th>
<th>Power Grid Density</th>
<th>#Nodes</th>
<th># nodes with noise violations</th>
<th>Worst V droop (V)</th>
<th>Violation Area S (V · ns)</th>
<th>Decap (pF)</th>
<th>Lkg (mA)</th>
<th>maxC (%)</th>
<th>avgC (%)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case1</td>
<td>Normal</td>
<td>18634</td>
<td>3330</td>
<td>0.135</td>
<td>13.739</td>
<td>628</td>
<td>1.1</td>
<td>8.35</td>
<td>1.66</td>
<td>42.6</td>
</tr>
<tr>
<td>Case2</td>
<td>Denser</td>
<td>36433</td>
<td>4210</td>
<td>0.126</td>
<td>2.615</td>
<td>488</td>
<td>0.6</td>
<td>31.27</td>
<td>4.75</td>
<td>45.0</td>
</tr>
<tr>
<td>Case3</td>
<td>Densest</td>
<td>72114</td>
<td>4671</td>
<td>0.124</td>
<td>1.482</td>
<td>229</td>
<td>0.3</td>
<td>58.41</td>
<td>7.62</td>
<td>53.1</td>
</tr>
</tbody>
</table>

- Lkg: leakage current
- maxC: maximum increment of congestion
- avgC: average increment of congestion

1. Denser power grid $\implies$ smaller voltage droop
2. Denser power grid $\implies$ increased congestion
Summary

• Power delivery into a 3D chip is a critical problem for next-generation designs

• MIM decap is an efficient option for 3D power grid optimization

• A LP based decap allocation approach using both MIM and CMOS decaps

• Our algorithm can also be used to solve the 2D power grid optimization problem
Thank You!