ECE 225
Lecture 12
Interconnect Design under Thermal, Power, Reliability & Variability Constraints

Prof. Kaustav Banerjee
Electrical and Computer Engineering
University of California, Santa Barbara
Voltage Drop Effects in Power Distribution Networks

Voltage Drop Effects

- Voltage-drop is an inseparable aspect of DSM power distribution network (PDN)
  - resistive voltage drop $\Delta V$ (on-chip resistance)
  - switching noise, $Ldi/dt$ (off-chip inductance)
- Degrades device switching speed and DC noise margins
- Can cause functional failure in dynamic logic and timing violation in static logic
- 10% voltage-drop $\Rightarrow$ 8% increase in device propagation delay (0.18$\mu$m tech.)
Challenges in P/G Network Design

- Satisfying the electromigration (EM) rules for each power routing segment
- Minimizing the area consumed by PDN
- Limiting the worst-case voltage-drop (~10% of $V_{dd}$ for current technology)
  - Power network wire-sizing ($\uparrow$ routing area)
  - Decoupling-cap insertion ($\uparrow$ substrate area)
Local Power Distribution

Standard Cell Row
(Substrate)

M1

M2

$C_{decap}$

$C_L$

$C_{decap}$

$C_{decap}$

$C_L$
Technology Scaling Impacts

- Scaling of line dimension ($R$, $L$, $C$)
- Scaling of chip frequency, power, and # of power pads
- Interconnect thermal effects
  - Electromigration rules
  - $R$
- Thin-film scattering and barrier thickness effects in DSM interconnects
  - $R$
Global/Semi-global PDN IR-Drop

Allocation of 5% and 10% of the routing area for wire sizing to minimize the voltage-drop:

Voltage-drop increases rapidly with technology scaling....

Interconnect temperature has a major impact on the voltage-drop of global segments.....
Full Chip IR-Drop

Local worst-case IR-Drop for 50 and 100 switching cells

Total chip worst-case IR-Drop 10% routing, 5% chip area

**T**—considering temperature effect  **S**—considering thin-film effects
Within-Die Temperature Variations

- Substrate power generation distribution is generally non-uniform
  - Functional block clock gating
  - System-level power management
  - Non-uniform distribution of gate sizing and switching activities in different blocks
- Substrate thermal profile is non-uniform
  - Thermal time constant is of the order of $ms$
  - Switching activities at the block level are more important
  - Introduces non-uniformity in the global interconnect thermal profile
Interconnect Thermal Profile

A. Ajami et al., DAC 2001

\[
\frac{d^2 T_{\text{line}}}{dx^2} = -\frac{Q}{k_m}
\]

\[
\frac{d^2 T_{\text{line}}(x)}{dx^2} = \lambda^2 T_{\text{line}}(x) - \lambda^2 T_{\text{ref}}(x) - \theta
\]

\[
Q = q_1 - q_2
\]

\[
\lambda \text{ and } \theta \text{ are constants}
\]

f (L, t_m, k_m, t_{\text{ins}}, k_{\text{ins}}, I_{\text{rms}}, R_E)
Non-Uniform Temperature-Dependent Delay

\[ D = R_d (C_P + C_L + \int_0^L c_0(x)dx) + \int_0^L r_0(x)(\int_x^L c_0(\eta) d\eta + C_L) dx \]

\[ D = D_0 + (c_0L + C_L)\rho_0\beta \int_0^L T(x) dx - c_0\rho_0\beta \int_0^L xT(x) dx \]

\( D_0 \) is the Elmore delay model at reference temp.
Implications for Delay and IR-Drop

Impact on delay estimation.....

T1: positive exponential gradient
T2: negative exponential gradient
For a fixed T_Low

Impact on IR-drop analysis.....

Worst-case voltage-drop ($V_{IR}/V_{dd}$) increases in the presence of thermal gradients

Ajami et al., TCAD 2005

Ajami et al., JAICSP, 2005
Impact on Buffer Insertion

Buffer movement in a 6660 um line (180 nm node)

(A) Standard
   \[ x = 3330 \text{ um} \]

(B) Variable wire resistance
   \[ x = 3430 \text{ um} \]

(C) Variable driver resistance
   \[ x = 81.5 \text{ um} \]

(D) Variable driver & wire resistance
   \[ x = 154.6 \text{ um} \]

Delay improvement after thermally-aware buffer insertion

Ajami et al., ICCAD 2001
Interconnect Thermal Effects and Reliability

Self-Heating or Joule Heating

\[ \Delta T_{\text{self-heating}} = (T_m - T_{\text{ref}}) = I_{\text{rms}}^2 R \theta_j \]

\( \theta_j \) is the effective thermal impedance:

\[ \theta_j = \frac{t_{\text{ins}}}{K_{\text{ins}} L W_{\text{eff}}} \]

\( W_{\text{eff}} \) is the effective metal width to account for quasi-2D heat conduction.
Interconnect Scaling Trends

Implications of Technology Scaling on Thermal Effects

- Low-k materials have lower thermal conductivity than silicon dioxide.

- Scaling trends that cause increasing thermal effects are:
  - increasing current density
  - increasing interconnect levels
  - low-k dielectrics
  - increasing thermal coupling
Impact of Line Width Scaling Using Low-k

DC Conditions

Standard Dielectric

\[ \Delta T [^\circ C] \]

\[ \text{Power [W]} \]

- As \( W \) decreases, \( \text{SH} \) increases.

Low-K Dielectric

\[ \Delta T [^\circ C] \]

\[ \text{Power [W]} \]

- Low-k increases \( \text{SH} \) by 10-15%
Electromigration Reliability

- Transport of mass in metal interconnects under an applied current density
- EM lifetime reliability modeled using Black’s equation given by,

\[ TTF = A \ j_{avg}^{-2} \ \exp\left(\frac{Q}{k_B \ T_m}\right) \]

- EM stress data gives a technology limit of the current density \( j_{avg} \) - for a required failure rate and a desired lifetime at a reference temperature \( T_{ref} (\sim 100^\circ C)\)
- The \( j_{avg} \) limit does not comprehend self-heating
VLSI Reliability: Lifetime Estimation and Design Rule Generation

- Product Lifetime (in years)
  - Projected lifetime under field stress conditions
  - Measured region
  - Projected region
  - Failure times under accelerated stress conditions

- Field stress conditions
- Accelerated stress conditions
  - Stress Parameter (Voltage, current density etc, @ a given temperature)

Design Rules
Accelerated EM stress data yields $A$, $Q$, and $n$ in Black’s equation, and a value of log-normal $\sigma_{LN}$.

EM stress data + Black’s equation gives a technology limit to the maximum allowed current density ($j_{\text{avg}}$) for the required failure rate and a desired lifetime at a reference temperature $T_{\text{ref}}$ ($\sim 100$ °C).

Typical goal: achieve a 10 year lifetime.

The $j_{\text{avg}}$ limit does not comprehend self heating, and is therefore not self-consistent.
Current Density Definitions

- **Peak, Average, and RMS current densities:**

  \[ j_{peak} = \frac{I_{peak}}{A} \quad j_{avg} = \frac{1}{T} \int_{0}^{T} j(t) \, dt \quad j_{rms} = \sqrt{\frac{1}{T} \int_{0}^{T} j^2(t) \, dt} \]

  \( A \) is the cross sectional area of interconnect, \( T \) is the time period of the current waveform.

- **For an unipolar waveform:**

  \[ j_{avg} = r \cdot j_{peak} \quad j_{rms} = \sqrt{r} \cdot j_{peak} \]

  \( r \) is the duty factor

- **EM is determined by** \( j_{avg} \), **and self-heating by** \( j_{rms} \).
Self-Consistent Design

Impact of Self-Heating on EM

- EM lifetime given by: \( T_{TTF} = A j^{-n} \exp \left( \frac{Q}{k_B T_m} \right) \)
- Due to self-heating: \( T_m = T_{ref} + \Delta T_{self-heating} \)
- \( \Delta T_{self-heating} = (T_m - T_{ref}) = I_{rms}^2 R \theta \)
- \( j_{rms}^2 = \frac{(T_m - T_{ref}) K_{ins} W_{eff}}{t_{ins} t_m W_m \rho_m (T_m)} \) (1)

\( \theta \) is the effective thermal impedance given by,

\[ R_{\theta} = \frac{t_{ins}}{K_{ins} L W_{eff}} \]

\( W_{eff} \) is the effective metal width to account for quasi-2D heat conduction.
Typically, design rules specify $j_{\text{avg}}$ from EM and $j_{\text{rms}}$ from self-heating separately.

Self-consistent approach: comprehends EM and self-heating simultaneously.

In order to achieve an EM reliability lifetime goal mentioned earlier, the lifetime at any $j_{\text{avg}}$ and metal temperature $T_m$, should be equal to or greater than the lifetime value (e.g., 10 year) under the design rule current density ($j_0$).

$$\exp \left( \frac{Q}{k_B T_m} \right) \geq \exp \left( \frac{Q}{k_B T_{\text{ref}}} \right)$$

(2)
Self-Consistent Design

What is Self-Consistent Design?

n Using the relationship between \( j_{\text{avg}} \), \( j_{\text{rms}} \), \( j_{\text{peak}} \) and \( r \) for an unipolar waveform described earlier, it can be shown that,

\[
\frac{j_{\text{avg}}^2}{j_{\text{rms}}^2} = r \quad (3)
\]

n Incorporating the \( j_{\text{rms}}^2 \) and \( j_{\text{avg}}^2 \) values from (1) and (2) in (3) yields the self-consistent equation,

\[
r = j_0^2 \frac{\exp \left( \frac{Q}{k_B T_m} \right)}{\exp \left( \frac{Q}{k_B T_{\text{ref}}} \right)} \frac{t_{\text{ins}} t_m W_m \rho_m (T_m)}{(T_m - T_{\text{ref}}) K_{\text{ins}} W_{\text{eff}}}
\]

This is a single equation in the single unknown temperature \( T_m \). Once the self-consistent \( T_m \) is obtained, the corresponding \( j_{\text{rms}} \) and \( j_{\text{peak}} \) and \( j_{\text{avg}} \) values can be calculated.
Coupled Analysis

Unipolar Case:

Coupled equation:

\[
r = j_0^2 \frac{\exp\left(\frac{Q}{k_B T_m}\right)}{\exp\left(\frac{Q}{k_B T_{ref}}\right)} \frac{t_{ins} t_m W_m \rho_m(T_m)}{(T_m - T_{ref}) K_{ins} W_{eff}}
\]

Once \( T_m \) is calculated, \( j_{rms} \) and \( j_{peak} \) can be obtained

\[
\frac{j_{avg}^2}{j_{rms}^2} = r
\]

Once \( T_m \) is calculated, \( j_{rms} \) and \( j_{peak} \) can be obtained.
Coupled Analysis

Unipolar Case: Metal 6, 180 nm node, $j_0 = 0.6 \text{ MA/cm}^2$
Interconnect Delay and Power Dissipation

Performance Optimization: Global Wires

- Increasing chip complexity and scaling
  - number and (electrical) length of global lines & latency increase

- Repeater insertion used for lowering the signal delays

- Delay of optimally repeatered line is linear in its length

- High-performance designs: large number of repeaters (> $10^6$ for sub-100 nm designs)
Increasing Number of Repeaters

Source: Intel Corporation
Optimal Repeater Insertion

- Long interconnects can be optimally buffered 

\[
\text{\textbf{s}}_{\text{opt}} \quad \text{---} \quad \text{\textbf{l}}_{\text{opt}} \quad \text{---} \quad \textbf{s}_{\text{opt}}
\]

- Large size: could be \(\sim 450\) times the minimum sized inverters available in a relevant technology (K. Banerjee et al., DAC 1999)

- Can contribute to significant on-chip power dissipation!
Delay Vs Power Tradeoff for Global Wires

- All global wires are not on the critical path
- Optimal buffering is not always necessary
- Some delay penalty can be tolerated on non-critical wires
- **Potential for large power savings** by using smaller repeaters and larger inter-repeater wire length
Interconnect Delay Optimization

For $s = s_{opt}/2$ and $l = 2 l_{opt}$; Delay Penalty is only 25%
Power-Optimal Repeater Insertion

Banerjee and Mehrotra, TED 2002.

- A methodology to estimate repeater size and inter-repeater wire length which minimizes the total interconnect power dissipation for a given delay penalty
- Estimate power-optimal buffering schemes for various ITRS technology nodes
- Analyze relative importance of various components of power dissipation as a function of technology scaling
Methodology

delay per unit length:

\[
\tau = \frac{1}{\ell} \left( r_s (c_0 + c_p) + \frac{r_s}{s} c + r_s c_0 + \frac{1}{2} rc \ell \right)
\]

For \( \ell = \ell_{opt} \), \( s = s_{opt} \):

\[
\left( \frac{\tau}{\ell} \right)_{opt} = 2 \sqrt{r_s c_0 rc} \left\{ 1 + \sqrt{0.5 \left( 1 + \frac{c_p}{c_0} \right)} \right\}
\]
Methodology

\[ P_{\text{repeater}} = P_{\text{switching}} + P_{\text{leakage}} + P_{\text{short circuit}} \]

\[ P_{\text{switching}} = \alpha \left( s \left( c_p + c_0 \right) + I_c \right) V_{DD}^2 f_{clk} \]

\[ P_{\text{leakage}} = V_{DD} I_{\text{leakage}} = V_{DD} I_{\text{off}} W_{n_{\text{min}}} s \]

\[ P_{\text{short circuit}} = \alpha t_r V_{DD} I_{\text{peak}} f_{clk} = \alpha t_r V_{DD} W_{n_{\text{min}}} s I_{\text{short circuit}} f_{clk} \]

- \( \alpha \) is the activity factor (\(=0.15\))
- \( t_r \) = rise time of input voltage to change from \( V_{tn} \) to \( V_{DD} - V_{tp} \)
Methodology

If the delay penalty is $f$, then:

$$\frac{\tau}{I} = (1 + f)\left(\frac{\tau}{I}\right)_{opt}$$

$$\frac{P_{\text{repeater}}}{I} = k_1 \left(\frac{s}{l} (c_p + c_0) + c\right) + k_2 \frac{s}{I} + k_3 s$$

- **switching**
- **leakage**
- **short-circuit**

$$k_1 = \alpha V_{DD}^2 f_{clk}$$

$$k_2 = \frac{3}{2} V_{DD} I_{off_n} W_{n_{min}}$$

$$k_3 = \alpha V_{DD} W_{n_{min}} I_{\text{short-circuit}} f_{clk} \log_e 3 \left(1 + f\right)\left(\frac{\tau}{I}\right)_{opt}$$
Minimize \( \frac{P_{\text{repeater}}}{l} \) w.r. t. \( s \):

\[
\frac{P_{\text{repeater}}}{l} = k_1 \left( \frac{s}{l} \left( c_p + c_0 \right) + c \right) + k_2 \frac{s}{l} + k_3 s
\]

- Involves three non-linear equations with three unknowns: \( l, s, \) and \( dl/ds \)
- Used Newton-Raphson to solve numerically
Results: Power Vs Delay Penalty

- Normalized power per unit length reduces as delay penalty increases.
- Incremental reduction in P/l is high for small values of the delay penalty.
- 180 nm and 130 nm results are almost identical.
- However, normalized P/l decreases with technology scaling: leakage power.
Optimization Results: 5% delay penalty

For optimal power dissipation: repeater size needs to be reduced and inter-buffer wire length needs to be increased
Optimization Results: 5% delay penalty

Normalized power per unit length decreases rapidly with scaling: due to substantial increase in leakage current
Optimization Results: 5% delay penalty

Relative contributions of the three components of PD

- Leakage power is the dominant component for advanced nodes
- Short circuit power is also non-trivial across all nodes
Implications of Parameter Variations

Parameter (P-V-T) Variations

- Sub-130 nm CMOS Transistors
- Transistor channel length
- Power Supply Voltage
- Chip Temperature
- Metal Thickness
- Supply Voltage
- Cache
  - Core
  - 70°C
  - 120°C
- Random Dopant Fluctuations
Process Variations

- **Device Variations**
  - Channel Length (L)
  - Oxide Thickness ($t_{ox}$)
  - Dopant Density ($N_a$)

- **Interconnect Variations**
  - Line Width (w)
  - Spacing (s)
  - Metal Thickness (t)
  - Dielectric Thickness (h)
Device vs. Interconnect Variations

S. Nassif (IBM)
Sensitivity Analysis of Delay and Power

3 cases considered; % variations increase from case 1 to case 3

Variations have significant impact on leakage power and delay

Wason and Banerjee, ISLPED 2005
Interconnect Design Considering Variations

- Significant Implications for Interconnect Design
- Power-Optimal Repeater Insertion - tradeoff delay with power

Variations can significantly impact power-optimal repeater insertion

Wason and Banerjee, ISLPED 2005
The normalized delay increases under variations.

Power savings are higher for the same penalty in delay, as variations increase.

Importance of power-optimal repeater insertion increases.

Wason and Banerjee, ISLPED 2005
Implications for Power Estimations

- Error in power estimation can be significant if variations are neglected.

- \( \alpha = 0.01 \)
- \( f = 400 \text{ Mhz} \)
- 100 nm tech. node
- Max power = \( \mu + 3\sigma \)

Error %

Case 1
Case 2
Case 3