Statistical Circuit Optimization Considering Device and Interconnect Process Variations

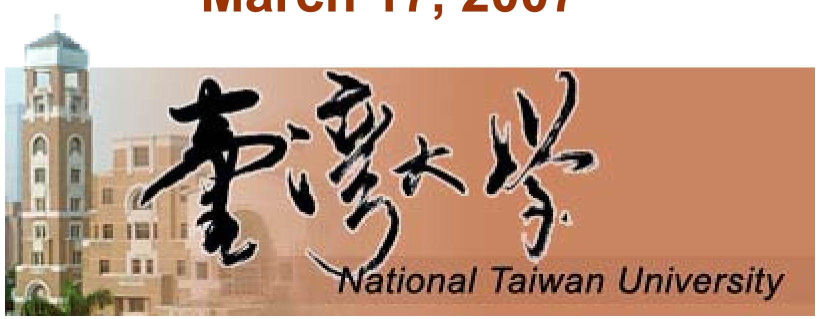
I-Jye Lin, Tsui-Yee Ling, and Yao-Wen Chang

The Electronic Design Automation Laboratory

Department of Electrical Engineering

National Taiwan University

March 17, 2007



Outline

- Introduction
- Deterministic Algorithm
- Statistical Algorithm
- Experimental Results
- Conclusions

Outline

- Introduction
- Deterministic Algorithm
- Statistical Algorithm
- Experimental Results
- Conclusions

Interconnect Process Variation

- Interconnect delay and reliability highly affect VLSI performance.
- The variability of interconnect parameters will raise up to 35%.
  - Sirvastava et al., Springer, 2005.
- The worst-case corner models cannot capture the worst-case variations in interconnect delay.
  - Liu et al., DAC 2000
- The interconnect optimization guided by statistical analysis techniques has become an inevitable trend.
  - Visweswariah, SLIP 2006

Previous Work in Statistical Optimization

- Statistical gate sizing with timing constraints using Lagrangian Relaxation.
  - Choi et al.," DAC 2005.
- Statistical power minimization by delay budgeting using second order conic programming.
  - Orshansky et al., DAC 2005.
- Statistical gate sizing using geometric programming
  - Patil et al., ISQED 2005.
- No statistical optimization work consider both interconnect and device sizing.

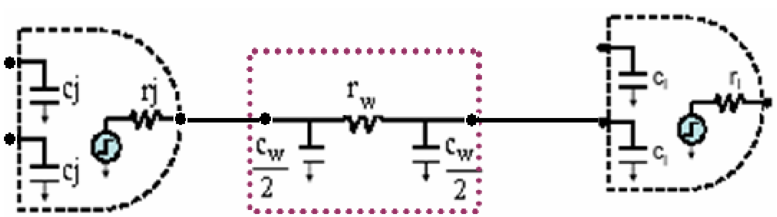
Comparison with Previous Work

	Orshansky's work (DAC 2005)	**Our work**
Sizing variable	Gate only	**Gate and wire**
Delay Model	Linear model (linear term)	Elmore delay model (**nonlinear** term)
Objective	Power	Area
Constraint	Timing	Power, timing, thermal

Due to the nonlinear term introduce by the Elmore delay model, the optimization using both gate and wire sizing will be much harder to solve.

Delay Model

- Our delay model and timing constraint:
  - Elmore delay model



$$D_i = R_g\,(\,C_w X_w L_w + C_g X_i)\,/X_j + R_w L_w(\,C_w X_w L_w/2 + C_g X_i)/(X_w)$$

Higher order (quadratic) terms!

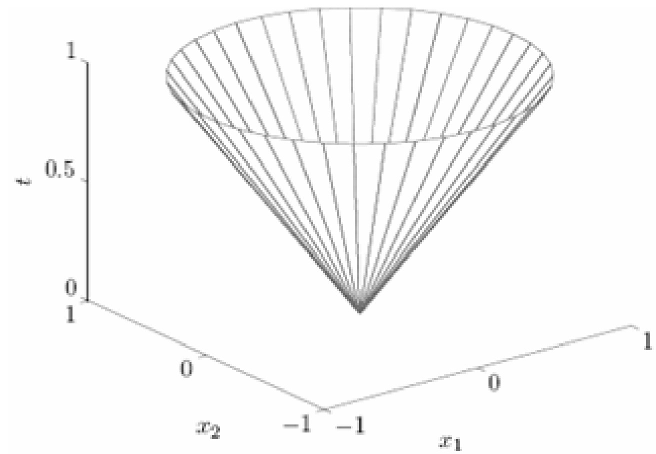
  - Timing constraint: $a_i \geq a_j + D_i$
  - a_i = arrival time of gate i

- Delay model and timing constraint used in previous work in DAC 2005: $a_i \geq a_j + d_i^0 + d_i$ linear terms!
  - d_i^0 = delay due to the sizing for maximum slack
  - d_i = slack added to node i due to the loading

Statistical Circuit Optimization with SOCP

- Second-order conic programming (SOCP)

$$\begin{aligned} \textbf{Minimize} \quad & f^T x \\ \textbf{subject to} \quad & \|A_i x + b_i\|_2 \le c_i^T x + d_i \\ & Fx = g \end{aligned}$$



  - Convex optimization
  - Theoretical runtime $O(N^{1.3})$
  - Orshansky (DAC 2005), Davoodi (DAC 2006)

- Second-order conic constraint:

$$\|\underline{A_i x + b_i}\|_2 \le c_i^T x + d_i$$

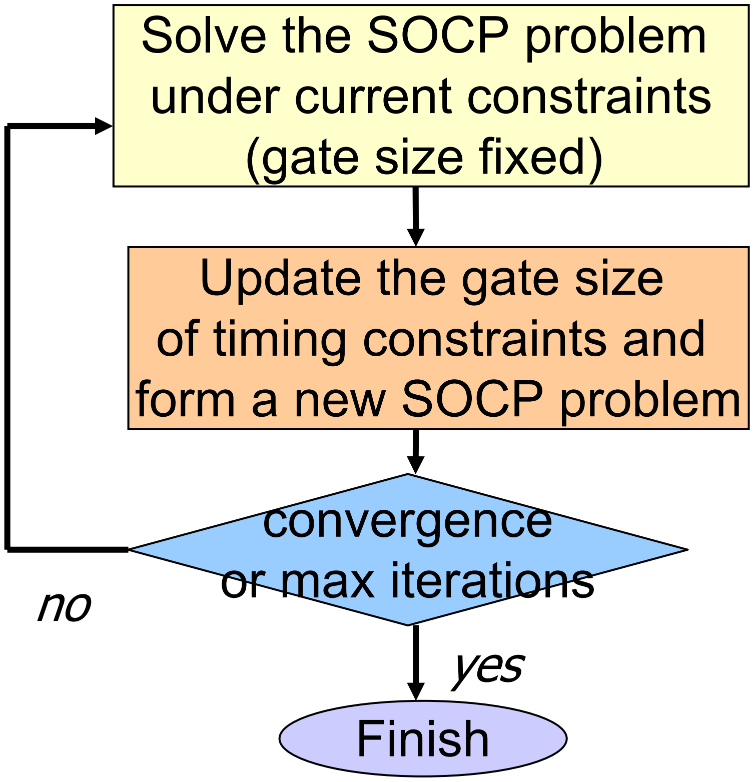
Linear terms!

Nonlinear (quadratic) terms are not applicable!

Approximation method

Approximation Method

- Fix the gate size in the timing constraint.
  - Reduce the timing constraint from quadratic order to linear order.
- Approximate the gate sizes by a two-stage flow.
  - Iteratively reduce the approximation errors.
  - The flow is similar to Sequential Linear Program (SLP).



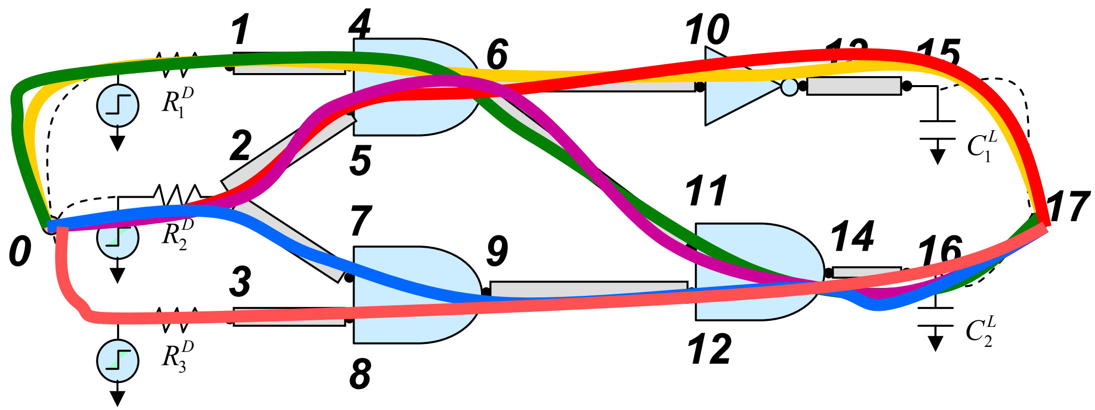
Our Contributions

- The first work of statistical optimization on circuit interconnect and devices
  - Previous work considers only circuit devices (gates).
  - Statistical optimization for considering both interconnect and devices is much harder.
- The first work that statistically optimizes the area with *thermal- and timing-constrained* parametric yields
  - Most existing statistical optimization considers only timing.
- The first work capable of analytically transforming the statistical RC model into an SOCP
  - Previous work uses linear delay model

Outline

- Introduction
- Deterministic Algorithm
- Statistical Algorithm
- Experimental Results
- Conclusions

Timing Constraint



Timing constraint:

$$\sum_{i \in \omega} D_i \le D^B, \ \forall \omega \in \Omega$$

- # of paths may grow exponentially to the circuit size.
- To reduce problem size, we distribute the timing information to each node.

$$\begin{array}{ll} D_i \le a_i & i = 1, \dots, s \ \ /* \ primary \ inputs */ \\ a_j + D_i \le a_i & i = s+1, \dots, n{+}s \ \ and \ \forall j \in input(i) \\ a_j \le D^B & j \in input(m) \ \ /* \ primary \ outputs */ \end{array}$$

Thermal Constraint

- Electron Migration (EM) lifetime reliability of metal interconnects is governed by the well-known Black's equation:

$$TTF = A^{\star} \cdot j^{-n} \cdot exp(\frac{Q}{k_B T_m}),$$

TTF: time-to-fail period
A* : a constant
j : average current density
Q : activation energy
KB : Boltzmann's constant.
Tm : metal temperature

- The design is reliable when

$$\frac{j_{avg}^2}{exp\left(\frac{Q}{k_B T_m}\right)} \leq \frac{j_0^2}{exp\left(\frac{Q}{k_B T_{ref}}\right)}.$$

j_0: specific current density
T_{ref}: specific metal temperature

$$T_m \leq \frac{Q \cdot T_{ref}}{Q - 2K_B T_{ref}(\ln j_0 - \ln j_{avg})} = T^{B'}.$$

$$T_m = \Delta T_{self-heating} + T_{environment}$$

Average Temperature of the Chip

- The average temperature of the chip, T_{avg}, can be estimated by:

$$T_{avg} = T_{air} + R_n\left(\frac{P_{tot}}{A}\right)$$

Power

P_{tot} : total power consumption of the chip
T_{air} : ambient temperature
R_n : thermal resistance of the substrate and the package
A : chip area

– Banerjee et al., ISPD 2001.

Power Constraint

- Need to constrain chip's temperature under a reasonable bound during the optimization:

$$\alpha_i c_i \le P_i^{B'}, \quad P_i^{B'} = P_i^B / V_{DD}^2 f$$

  - For simplicity, consider the dynamic power consumption only.
  - P_i^B: the power bound of the gate i
  - c_i : the downstream capacitance of the gate I
  - α_i : switching activity of component *I*

Deterministic Formulation

$$\text{Minimize} \quad \sum_{i=s+1}^{n+s} l_i \, x_i$$

l_i: **gate unit area or wire length**
x_i: **gate or wire size (sizing variable)**

subject to

$$T_{mi} \le T_i^B, i \in W,$$

Thermal constraint

$$\begin{cases} D_i \le a_i, \; i = 1, \ldots, s \\ a_j + D_i \le a_i, \; i = s+1, \ldots, n+s \; and \; \forall j \in input(i) \\ a_j \le D^B, \; j \in input(m) \end{cases}$$

Timing constraint

$$\alpha_i c_i \le P^{B'}, \; i = s+1, \ldots, n+s$$

Power constraint

$$L_i \le x_i \le U_i, \; \forall s+1 \le i \le n+s.$$

f: working frequency; α_i : switching activity of component *I*; c_i: load capacitance of component *I*; ω: path in the path set Ω.

Outline

- Introduction
- Deterministic Algorithm
- Statistical Algorithm
- Experimental Results
- Conclusions

Variation Models

- Introduce two process parameters as the variation sources: Inter-layer dielectric (ILD) thickness (H), and metal thickness (T).
- R and C can be approximated by the first-order Taylor expression:

R_{nom}/C_{nom}: nominal value of R/C
$\Delta T/\Delta H$: random deviation of metal thickness/ILD thickness

$$R = R_{nom} + a_1 \Delta T,$$
$$C = C_{nom} + b_1 \Delta T + b_2 \Delta H$$

  - a1, b1, b2 are sensitivities calculated by the differential differentiation of:

C_{gnd}/ε: normalized capacitance
S: space between parallel lines

$$R = \frac{\rho}{WT},$$
$$\frac{C_{gnd}}{\epsilon} = \frac{W}{H} + 3.28\left(\frac{T}{T+2H}\right)^{0.023} + \left(\frac{S}{S+2H}\right)^{1.16}$$

  - Srivastava et al., Springer 2005.

RC Variability

- Assume T and H are Gaussian, the variability magnitude of R and C can easily be calculated by:

$$\sigma_R^2 = a_1^2 \sigma_T^2$$

$$\sigma_C^2 = b_1^2 \sigma_T^2 + b_2^2 \sigma_H^2$$

- Apply the *interconnect delay variation metric* to calculate the variability of the product of R and C.
  - Well captured by a normal distribution with 1.2% average error of the mean delay and 3.8% average error of the standard deviation.
  - Blaauw et al., DAC 2004.

Statistical Formulation

Deterministic formulation

Minimize

$$\sum_{i=s+1}^{n+s} l_i x_i$$

subject to

$$T_{mi} \le T_i^B$$

$$\begin{cases} D_i \le a_i \\ a_j + D_i \le a_i \\ a_j \le D^B \end{cases}$$

$$\alpha_i c_i \le P^{B'}$$

$$L_i \le x_i \le U_i$$

Statistical formulation

Minimize

$$\sum_{i=s+1}^{n+s} l_i x_i$$

subject to

$$Prob\,(T_{mi} \le T_i^B) \ge \delta$$

$$\begin{cases} Prob\,(D_i \le a_i) \ge \zeta \\ Prob\,(a_j + D_i \le a_i) \ge \zeta \\ Prob\,(a_j \le D^B) \ge \zeta \end{cases}$$

$$Prob\,(\alpha_i c_i \le P^{B'} \ge \eta$$

$$L_i \le x_i \le U_i$$

– **δ/ζ/η**: Thermal/Timing/Power yield constraint

Transformation into SOCP

- **Theorem:** Given independent Gaussian random vectors a_i and bound vectors b_i, the parametric yield (η) problem is as follows:

$$\begin{aligned} &Minimize \\ &\qquad \sum x_i \\ &subject\ to \\ &\qquad \boxed{Prob\left(a_i^T x_i \le b_i\right) \ge \eta,} \end{aligned}$$

  the problem can be reformulated as an SOCP:

$$\begin{aligned} &Minimize \\ &\qquad \sum x_i \\ &subject\ to \\ &\qquad \boxed{\left(\bar{a}_i^T x_i\right) + \Phi^{-1}\left(\eta\right)\left(x^T \Sigma_i x\right)^{1/2} \le b_i.} \end{aligned}$$

  - Φ^{-1}: the cumulative density inverse function
  - Boyd and Vandenberghe, Cambridge, 2004.

Transformation Flow

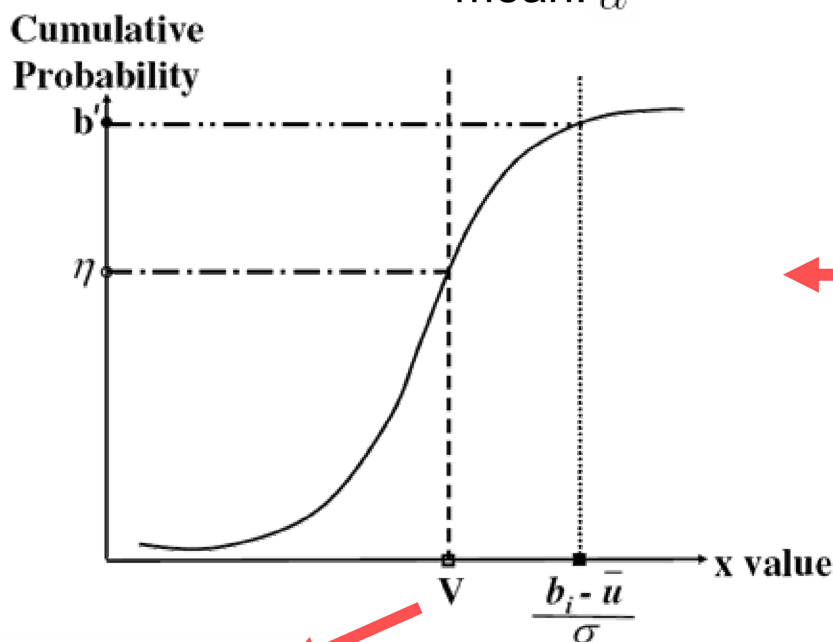
$$Prob\left(a_i^T x_i \le b_i\right) \ge \eta \xrightarrow{u = a_i^T x_i} Prob\left(\boxed{\frac{u-\bar{u}}{\sigma}} \le \frac{b_i - \bar{u}}{\sigma}\right) \ge \eta,$$

variance: σ^2
mean: $\bar{u}$

zero mean unit variance Gaussian variable

$$\Phi\left(\frac{b_i - \bar{u}}{\sigma}\right) \ge \eta$$

cumulative density function



$$\frac{b_i - \bar{u}}{\sigma} \ge \Phi^{-1}(\eta) \rightarrow \bar{u} + \Phi^{-1}(\eta)\,\sigma \le b_i$$

$$\left(\bar{a}_i^T x_i\right) + \Phi^{-1}(\eta)\left(x^T \Sigma_i x\right)^{1/2} \le b_i$$

Thermal & Power Constraints in SOCP Form

- Thermal constraint:

$$p\,(\,T_{mi} \le T_i^B\,) \ge \delta$$

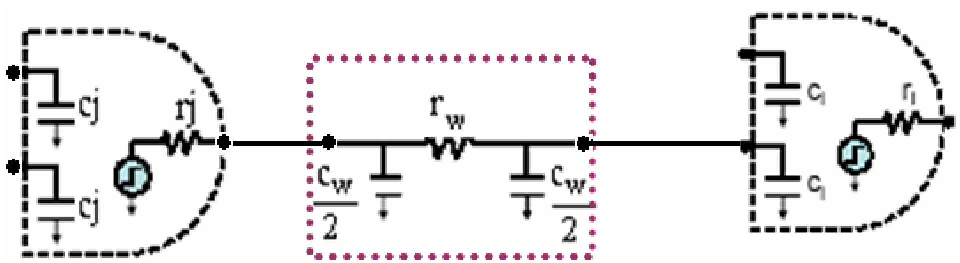
$$\bar{T}_{mi} + \phi^{-1}(\delta)\sqrt{\sigma_1^2 x_w^2 + \sigma_2^2 \frac{1}{x_w^2} + \sigma_3^2 + \sigma_4^2} \le T_i^B,$$

- Power (Thermal distribution) constraint:

$$p\,(\,\alpha_i c_i \le P^{B'}\,) \ge \eta$$

$$\alpha_i \bar{c}_i + \phi^{-1}(\eta)\sqrt{\sigma_5^2 \sum_{i=s+1}^{n+s} (x_i^2)} \le P^{B'},$$

Timing Constraint in SOCP Form



X_j: size of the driving gate
X_i: size of the loading gate
X_w: width of the interconnect
L_w: length of the interconnect (constant)

$$D_i = R_g (C_w X_w L_w + C_g X_i) / X_j + R_w L_w (C_w X_w L_w/2 + C_g X_i) / X_w$$

Only Xw is the sizing variable

$$D_i = R_j C_w X_w L_w + R_j C_i + R_w C_w (L_w)^2/2 + R_m L_w C_i / X_w$$

Timing constraint: $\mathrm{P}(a_j + D_i \leq a_i) \geq \zeta$

$$a_j + \bar{D}_i + \phi^{-1}(\zeta)\sqrt{\sigma_6^2 x_w^2 + \sigma_7^2 + \sigma_8^2 + \frac{\sigma_9^2}{x_w^2}} \leq a_i$$

Statistical Problem Formulation using SOCP

P : $Minimize$

$$\sum_{i=s+1}^{n+s} l_i x_i$$

σ_1	σ_2	σ_3	σ_4	σ_5
$\sigma_{\rho m} B_0^2 A_1^2$	$\sigma_{\rho m} B_1^2 A_0^2$	$\sigma_{\rho m} B_1^2 A_1^2$	$\sigma_{\rho m} B_0^2 A_0^2$	$\alpha_i^2 \sigma_{c_g}$

σ_6	σ_7	σ_8	σ_9
$\sigma_{(r_j)(c_w l_w)}$	$\sigma_{(r_j)(c_i)}$	$\sigma_{(r_w l_w)\left(\frac{c_w l_w}{2}\right)}$	$\sigma_{(r_w l_w)(c_i)}$

$subject\ to$

$$\bar{T}_{mi} + \phi^{-1}(\delta)\sqrt{\sigma_1^2 x_{mi}^2 + \sigma_2^2 \frac{1}{x_{mi}^2} + \sigma_3^2 + \sigma_4^2} \leq T_i^B,$$

Thermal constraint

$$a_j + \bar{D}_i + \phi^{-1}(\zeta)\sqrt{\sigma_6^2 x_w^2 + \sigma_7^2 + \sigma_8^2 + \frac{\sigma_9^2}{x_w^2}} \leq a_i,$$

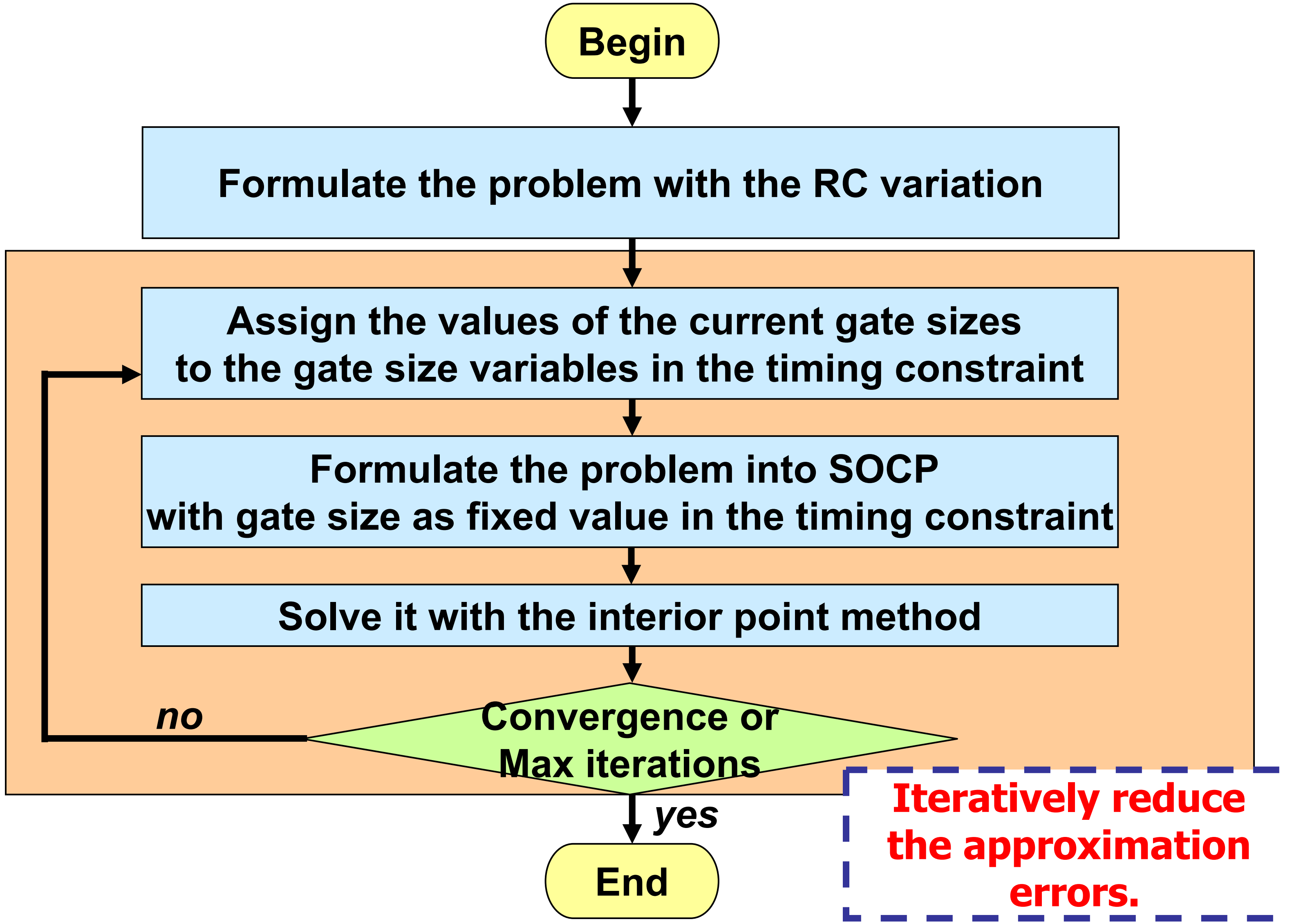
Timing constraint

$$\alpha_i \bar{c}_i + \phi^{-1}(\eta)\sqrt{\sigma_5^2 \sum_{i=s+1}^{n+s} (x_i^2)} \leq P^{B'},$$

Power constraint

$$L_i \leq x_i \leq U_i,\ \forall s+1 \leq i \leq n+s.$$

Program Flow



Outline

- Introduction
- Deterministic Algorithm
- Statistical Algorithm
- Experimental Results
- Conclusions

Experimental Setup

Circuit Name	Circuit Size		
	#Gate	**#Wire**	**#Total**
c17	11	12	23
c432	122	230	352
c499	246	396	642
c880	256	230	486
c1355	297	555	852
c1908	201	336	537
c2670	499	754	1253
c3540	429	1021	1450
c5315	927	1792	2719
c6288	1298	3596	4894

- Implemented in C++ & applied the MOSEK optimization tool to solve it.
- Tested on the commonly used ISCAS85 benchmark circuits in this area.
- Used Design Compiler & Astro with UMC 0.18μm technology library to synthesize and place the circuits.

Experimental Results

- Achieve 51%, 39%, and 26% area reductions for 70%, 84.1%, and 99.9% yield constraints, respectively.
- Avg. / Max. # of the running iterations: 5.6 / 10
- Timing constraint error bound: 2%

Circuit name	Deterministic			70% yield			
	area (μm^2)	Runtime / ite. (s)	Total runtime (s)	area (μm^2)	Area improv.	Runtime / ite. (s)	Total runtime (s)
c17	**7160**	**0.06**	**0.6**	**2892**	**59.61%**	**0.09**	**0.36**
c432	**47752**	**0.24**	**1.21**	**21543**	**54.89%**	**0.83**	**4.15**
c499	**127103**	**0.41**	**2.07**	**56957**	**55.19%**	**2.41**	**9.62**
c880	**152804**	**0.37**	**1.11**	**38346**	**74.91%**	**1.40**	**13.96**
c1355	**174896**	**0.58**	**5.79**	**84076**	**51.93%**	**3.87**	**19.33**
c1908	**96968**	**0.33**	**3.26**	**44350**	**54.26%**	**2.79**	**5.57**
c2670	**275967**	**0.74**	**7.39**	**121065**	**56.13%**	**7.32**	**14.64**
c3540	**362409**	**1.10**	**11.03**	**146519**	**59.57%**	**7.43**	**22.29**
c5315	**913522**	**1.88**	**13.18**	**727853**	**20.30%**	**10.43**	**31.28**
c6288	**1455730**	**5.23**	**15.69**	**1100120**	**24.43%**	**70.56**	**352.78**
Avg.					**51.12%**		

Experimental Results of 84.1% and 99.9% yield

- The lower the yield constraints, the better the area optimization.
- All constraints (timing, power, thermal) are met.

Circuit name	84.1% yield				99.9% yield			
	area (μm^2)	Area improv.	Runtime / ite. (s)	Total runtime (s)	area (μm^2)	Area improv.	Runtime / ite. (s)	Total runtime (s)
c17	3394	52.60%	0.09	0.6	3460	51.68%	0.09	0.47
c432	27860	41.66%	1.26	7.48	29179	38.89%	0.80	2.41
c499	57758	54.56%	2.11	8.43	89148	29.86%	1.54	4.61
c880	66420	56.53%	2.91	7.82	107349	29.75%	1.54	15.41
c1355	147397	15.72%	2.11	19.03	169347	3.17%	2.29	22.9
c1908	65020	32.95%	1.56	12.48	70830	26.96%	1.38	13.57
c2670	161426	41.51%	2.93	5.85	248474	9.96%	3.47	24.32
c3540	169331	53.28%	5.57	22.27	176715	51.24%	5.23	15.70
c5315	735838	19.45%	9.24	36.95	884514	3.18%	7.16	28.65
c6288	1109090	23.81%	69.74	348.71	1291240	11.30%	82.32	411.63
Avg.		39.21%				25.60%		

Delay, Power and Temperature Performance

- Though the delay and the maximum metal temperature are increased, they all meet the given bounds.
  - Fully utilized the constraint bound to get the best optimization results.

Circuit Name	Delay (ns)			Power (mW)		Max $T_{increase}$ (□)	
	Bound	Before	**After**	Before	**After**	Before	**After**
c17	**36.82**	22.19	**32.21**	2.02	**1.35**	8.19	**10.05**
c432	**247.65**	154.59	**136.62**	22.96	**12.59**	6.97	**23.46**
c499	**186.13**	153.79	**135.32**	56.10	**35.23**	7.20	**27.20**
c880	**253.43**	208.84	**170.15**	64.03	**43.44**	7.26	**19.69**
c1355	**274.55**	203.45	**241.45**	78.56	**67.24**	7.37	**27.47**
c1908	**222.91**	161.09	**136.11**	43.32	**28.36**	7.09	**31.78**
c2670	**290.84**	176.66	**229.94**	103.81	**98.69**	8.88	**22.72**
c3540	**507.80**	308.59	**245.85**	143.14	**72.65**	8.25	**14.03**
c5315	**445.58**	313.52	**421.54**	365.65	**355.45**	7.78	**19.55**
c6288	**1333.91**	913.33	**1148.41**	661.62	**513.14**	7.35	**11.62**
Comparison		**1**	**1.13**	**1**	**0.80**	**1**	**2.72**

Outline

- Introduction
- Deterministic Algorithm
- Statistical Algorithm
- Experimental Results
- Conclusions

Conclusions

- Presented the first statistical work for area minimization under thermal and timing constraints by gate and wire sizing.
  - Obtained much better results than those of the deterministic method.
- Formulated statistical RC model by SOCPs which can be solved efficiently and effectively.
  - Used more accurate delay model (Elmore delay model)
  - Solved the problem by a two-stage approximation flow
    - Nonlinear terms are not applicable to SOCP

Thank You

Backup Slides

Temperature Distribution

- Applying the Finite Difference Method (FDM), we can divide the whole chip into *m* mesh nodes and calculate each node's temperature by

$$G_2 T_P = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,\mathrm{m}} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m,1} & g_{m,2} & \cdots & g_{m,m} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_m \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_\mathrm{m} \end{bmatrix} = P_P$$

P_i: ith mesh node's power dissipation
T_i: ith mesh node's temperature
g: power density of the heat sources (W/m3)

— Chapman, "Heat Transfer," New York: Macmillan, 1984 Vol., 4th Ed..

Temperature Dependent Delay

- An inseparable aspect of electrical power distribution and signal transmission through the interconnects
- Resistance is dependent of Temperature

$$r(x) = \rho_0 \left(1 + \beta \cdot T(x)\right)$$

  - ρ_0: the resistance per unit length at reference temperature
  - β: the temperature coefficient of resistance (1/°C)

Interconnect Temperature Calculation

- The interconnect temperature is given by

$$T_m = \Delta T_{self-heating} + T_{environment}$$

$$= I_{rms}^2 \cdot R \cdot \theta_{int} + T_{environment}$$

$$= \frac{\sigma V_{cross}^2 t_{ox} t_m}{K_{ox} l^2 R_m} \cdot \frac{x_i}{x_i + \phi t_{ox}} + T_{environment}.$$

x_i : wire width
θ_{int} : the thermal impedance of the interconnect line to the chip
σ: duty cycle
V_{cross} : cross voltage of wire
t_{ox} : the total thickness of the underlying dielectric
t_m : the thickness of the wire
K_{ox} : the thermal conductivity
l : wire length
R_m : the temperature dependent unit resistance
ψ : the heat spreading parameter

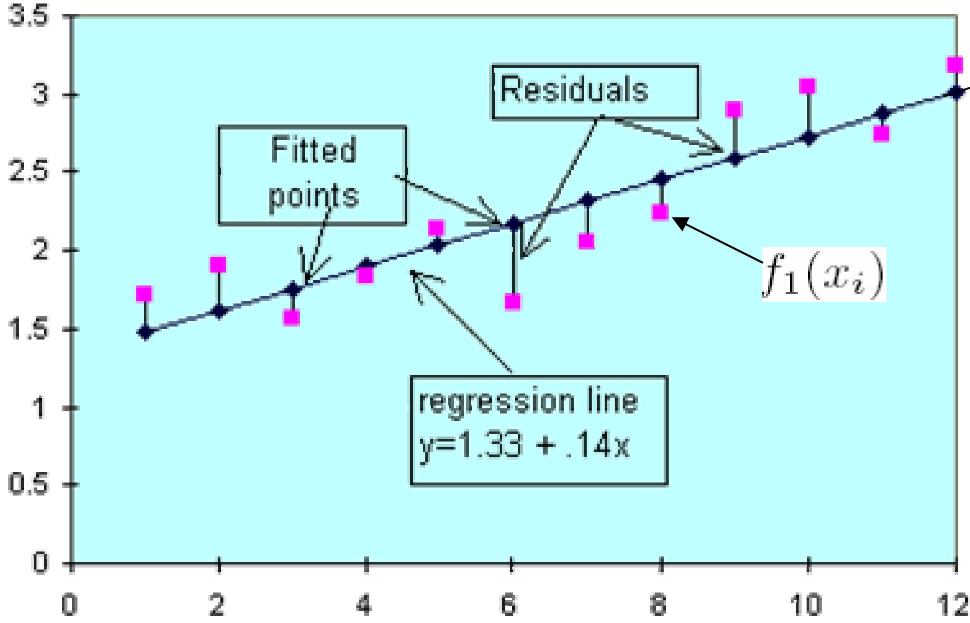
Not linear functions

⇩

Least Square Estimator

Least Square Estimator (LSE)

- Least squares solves the problem by finding the line for which the sum of the square deviations (or residuals) in the d direction (the noisy variable direction) are minimized.
  - Apply Cramer Rule to find the A_1 and A_0, which minimizes the square deviations



$y = A_1x + A_0$

Cramer Rule:

$$A_1 = \frac{\sum_{i=1}^{\mathcal{N}}(x_i - \bar{x})(f_{1i} - \bar{f}_1)}{\sum_{i=1}^{\mathcal{N}}(x_i - \bar{x})^2}$$

$$A_0 = \bar{f}_1 - A_1\bar{x},$$

$\bar{x}$ is the average of x_i

$\bar{f}_1$ is the average of $f_1(x_i)$

Approximation for Thermal Constraint

- Let N = 5 and pick five sizes of xi, we can approximate the thermal constraint by Least Square Estimator (LSE).
  - Banerjee et al., DAC 1999

$$\frac{x_i}{x_i + \phi t_{ox}} \Rightarrow A_1 \cdot x_i + A_0$$