Tutorial on Congestion Estimation

Taraneh Taghavi   **Foad Dabiri**

Ani Nahapetian   **Majid Sarrafzadeh**

Embedded and Reconfigurable Systems Lab
Computer Science Department
University of California, Los Angeles

ER
UCLA

Outline

- Motivation
- WL prediction
  - Individual net
  - Total interconnect length
- Congestion prediction
- Concurrent wirelength and congestion prediction
- Density utilization alleviation
- Statistical timing analysis
- Integrated congestion and delay prediction
- Conclusion

Motivation

- To achieve more efficient designs => combining front-end floorplanning and physical placement.
- The total wirelength affects three major parameters:
  - Chip size
  - Clock frequency
  - Power dissipation.
- Interconnect length is a measure of the quality of the placement
- accurate predictions of circuit properties are necessary to limit the vast search space

ER
UCLA

Motivation

New wirelength and congestion prediction methods are needed:

- Increase in size and sophistication of circuits

- Decrease in physical feature size

- Presence of IP blocks
  - Soft IP
  - Hard IP

- Accurate prediction help improve
  - floorplanning
  - Placement
  - routing

ER
UCLA

Integrated WL and Cong. Prediction

- Minimizing the total routed wirelength a fundamental goal
- In the presence of IP blocks, alleviating congestion after placement may result in an abrupt increase in wirelength
  - Congestion needs to be estimated early enough
- Internal routing demand is well-correlated with the length of interconnects
  - Congestion estimation needs the information of wirelength
- Detour around the congested area with no IP blocks.
  - Wirelength estimation may need to know the congestion map of the circuit
- Congestion and wirelength estimation are dependent

ER
UCLA

Outline

- Motivation
- WL prediction
  - Individual net
  - Total interconnect length
- Congestion prediction
- Concurrent wirelength and congestion prediction
- Density utilization alleviation
- Statistical timing analysis
- Integrated congestion and delay prediction
- Conclusion

WL Estimation: Individual Net

## Half Perimeter Bounding Box

- Optimal solution for 2-pin and 3-pin nets

- Lower bound for nets with higher degree.

- However, it can significantly underestimate wirelength for higher-degree nets.

- Net weighting technique to scale up the HPWL estimation.
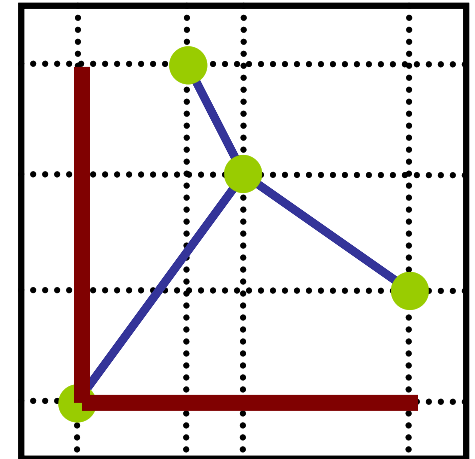
WL Estimation: Individual Net

Minimum Spanning Tree

- A spanning tree $T = \{V_T, E_T\}$ of a graph is a sub-graph T of G, where T is a tree (no cycles) and such that $V_T = V$ and $E_T \subset E$.

- MST = a minimum weight spanning tree over a weighted graph

- Can produce good wirelength estimation in reasonable amount of time.

- Best Time complexity = $O(n \log(n))$

- Simple $O(n^2)$ time implementation of Prim's algorithm is usually used

ER
UCLA
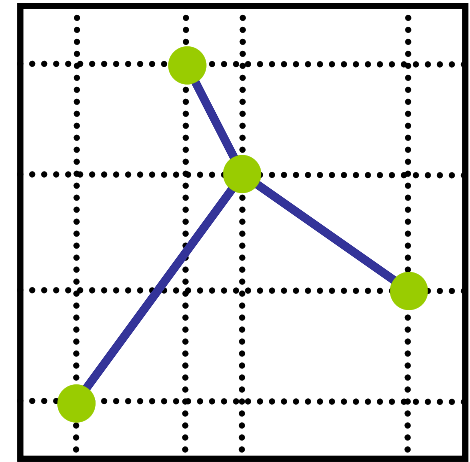
WL Estimation: Individual Net

Rectilinear Steiner Minimal Tree

- Given a set *P* of *n* points, find a set *S* of Steiner points such that *MST* (*P* ∪ *S*) has the minimum cost.

- Optimal RSMT can always be constructed based on the Hanan grid.

  - the length of an edge in the Hanan grid is equal to distance

- RSMT is NP complete

- Good heuristic: Iterative 1-steiner RSMT

Steiner point

ER
UCLA
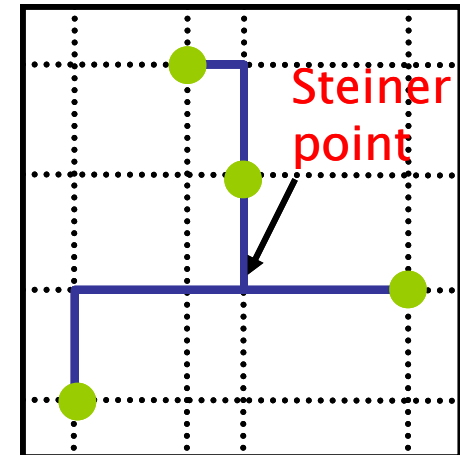
Rent's Rule

- **Base for most of WL estimation methods**

- **Empirical power law, $T=AC^P$**

- **A= Rent coefficient**

- **P= Rent exponent**
  - Higher rent exponent → Higher complexity

- **Rent exponent range : $0 \leq P \leq 1$**
  - Regular arch. (RAM) : $P = 0.5$
  - Complex arch. (VLSI circuits) : $P = 0.75$
  - Random logic : $P = 1$
  - Long chain of logic : $P = 0$

ER
UCLA

Rent's Rule

- **T** : Terminals per partition
- **C** : Cells per partition
  - *T/C Terminals per cell*
    - $\Delta T = (T/C)\, \Delta C$
  - Small ΔT, ΔC
    - *$dT = (T/C)\, dC$*
  - *Results in $T = AC$*
    - *A = Avg. no. of Terminals*

- Placement optimization level
  - *T/C Terminals per cell*
    - $\Delta T = P(T/C)\, \Delta C$
  - Small ΔT, ΔC
    - *$dT = P(T/C)\, dC$*
  - *Results in $T = AC^P$*

Pessimistic

$P = 0.66$

Ignores optimization level
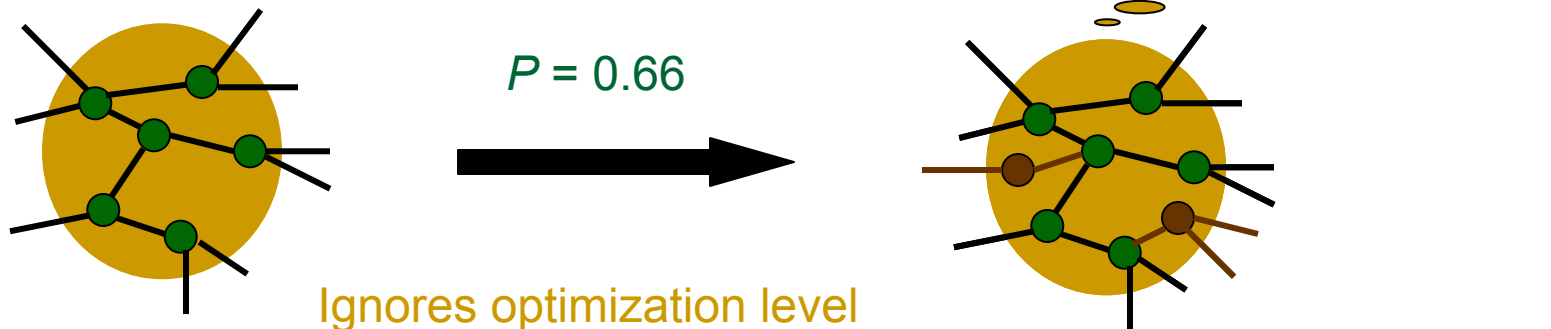
Rent's Rule Extraction
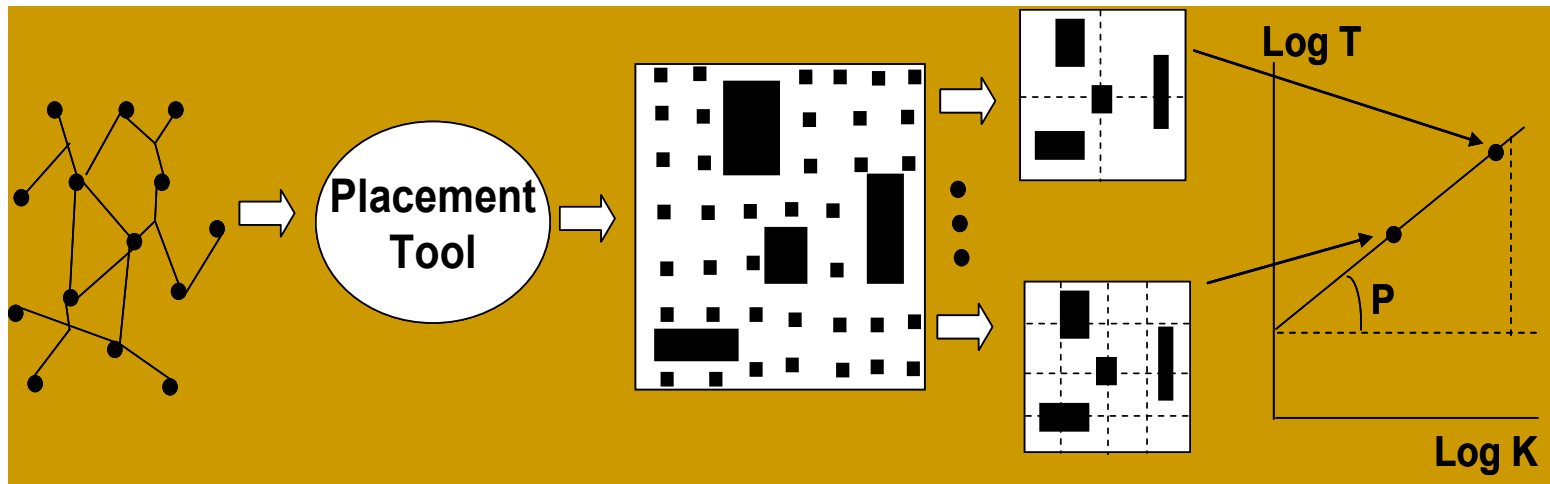
- Partitioning
- Coarse placement
- Log-Log plot of external terminals vs. avg. no. of cells
- Rent exponent: slope of the regression line

$$\log(T) = \log(A) + P\log(K)$$

Donath WL Estimation Method

- **Hierarchical estimation**
  - Partition into 4 sub-circuits.
  - At each level of hierarchy,
    - n_h: Number of interconnections
    - L_h: Average length of interconnections
  - Partitioning satisfies the Rent's rule.
- **Drawback:**
  - Ignores IP Blocks
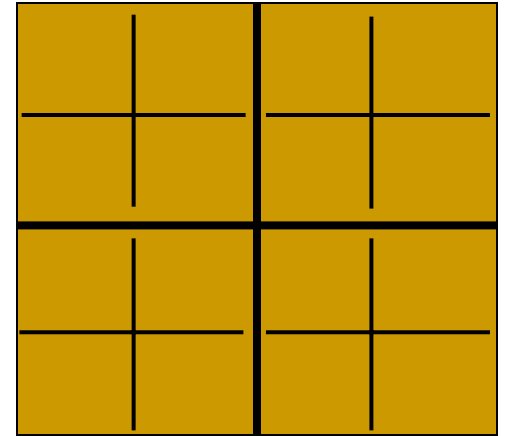
Davis WL Estimation Method

- ## Derivation of wirelength distribution

- ## Recursively applying Rent's rule

  - Stochastic WL distribution of a single gate

  - Remove it from the system

  - Repeat process for other gates

  - Superimpose WL distributions

- ## Drawbacks

  - Complex equations

  - Ignores IP Blocks

Block A   **Block B**

Block C
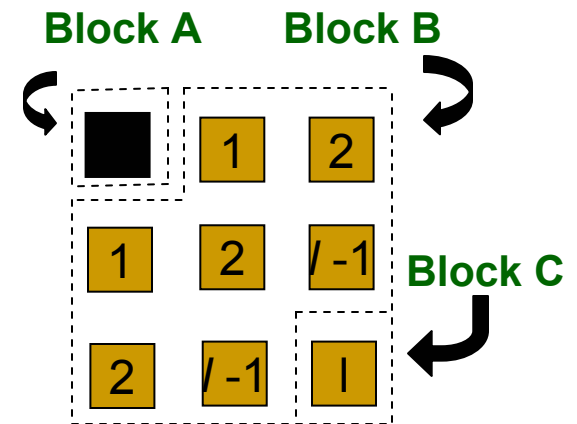
Non-Uniform Probability Distribution

Why Overestimate in Donath Method?

- Uniform probability distribution
- Optimal placement behavior
  - Keep connected cells closer to each other
  - More short wires than long wires
- Solution: Non-uniform probability distribution
  - Advantage: Accurate estimation
  - Disadvantage: Complex equations

ER
UCLA

Non-Uniform Probability Distribution

- Donath hierarchical scheme
    - Treat each level of hierarchy independently
    - Uniform probability distribution for terminals
- Optimal placement behavior
    - More terminals at the border
- Define:
    - $q(l)$: Occupying probability

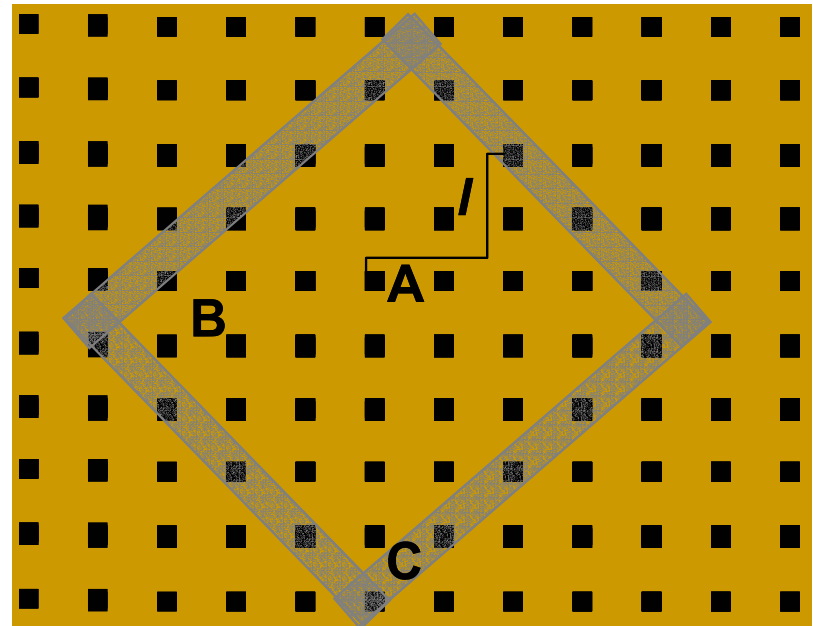$$q'(l) \approx l^{-(4-2P)}$$
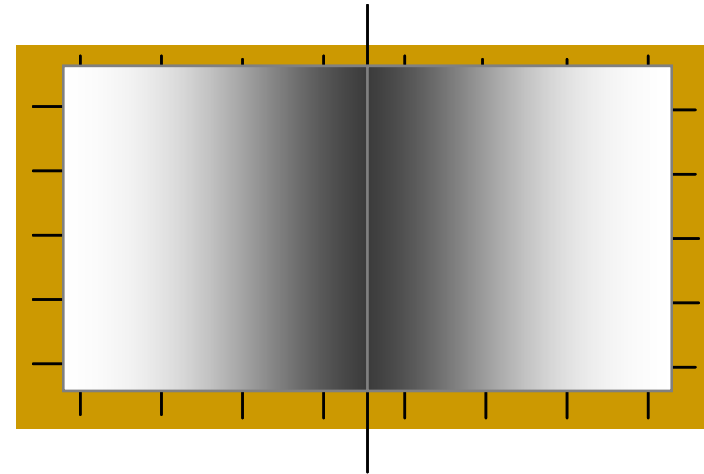
Non-Uniform Probability Distribution

- ## Conservation of terminals
  - Internal terminals
  - External terminals

$$T_{A \to C} = T_{AB} + T_{BC} - T_B - T_{ABC}$$

- ## Applying Rent's rule

$$T_{A \to C} = t \left[(1 + C_B)^P + (C_B + C_C)^P - C_B^P - (1 + C_B + C_C)^P \right]$$

$$n_{A \to C} = \alpha T_{A \to C}$$

Non-Uniform Probability Distribution

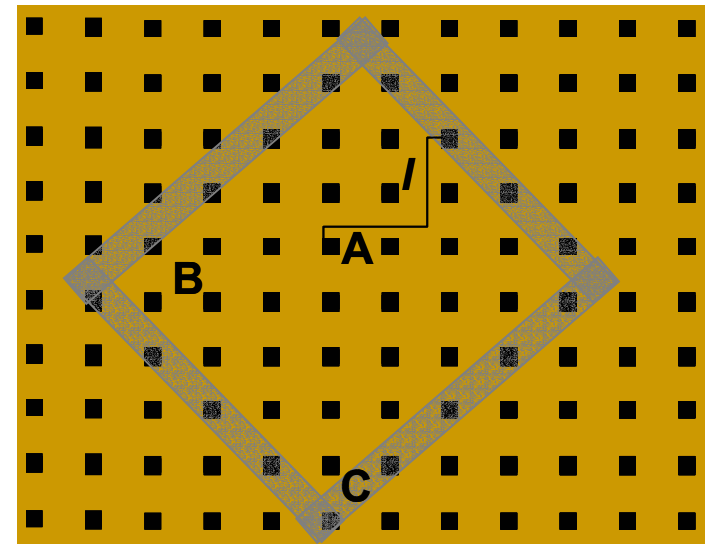- Number of terminals in C_C, C_B: $\quad C_B = \sum_{l=1}^{l-1} 4l = 2l(l-1), \quad C_C = 4l$

$$n(l) = \alpha t \left[\left(1 + 2l(l-1)\right)^P + \left(2l(l-1) + 4l\right)^P - \left(2l(l-1)\right)^P - \left(1 + 2l(l-1) + 4l\right)^P \right]$$

$$q(l) = \frac{1}{2l} \left[\left(1 + 2l(l-1)\right)^P + \left(2l(l-1) + 4l\right)^P - \left(2l(l-1)\right)^P - \left(1 + 2l(l-1) + 4l\right)^P \right]$$

$$q'(l) \approx P(1-P) 2^{-(1-P)} l^{-(4-2P)}$$

- For $l \gg 1$: $\quad q'(l) \approx l^{-(4-2P)}$

Stroobandt WL Estimation Method

- **Modification of Donath method**
  - Non-uniform probability distribution
- **Interconnection length distribution**
  - Structural distribution
  - Occupying probability
- **More accurate results than Donath**

ER
UCLA

Cheng WL Estimation

Considering IP Block Effect

- Flat WL estimation approach
- Total wirelength
  - Redistribution
  - Detour
- Based on geometrical characteristics
- Provide guidelines for revising floorplanning/global placement before detailed placement and routing

Taraneh WL Estimation

Considering IP Block Effect

- Hierarchical approach
    - Consider complexity of circuit using Rent exponent
    - Consider geometrical characteristics
- Deploy effect of different placement alg. in Rent exponent extraction
- Good for large-scale circuits

ER
UCLA

Methodology

- Physically partitioning into 4 sub-circuits

- Continue till no. of cells $\leq \beta$

- At each level of hierarchy

  - n_h: Avg. no. of interconnections
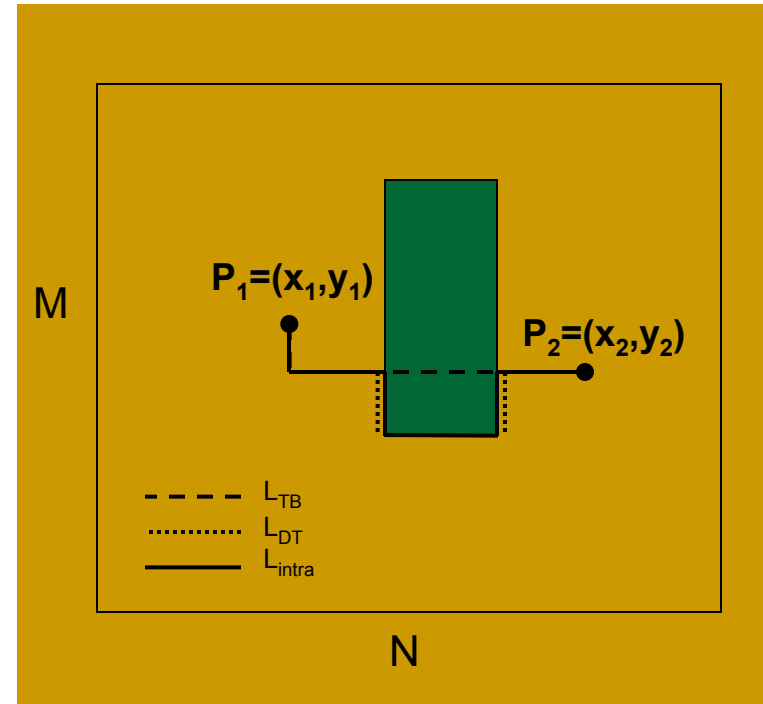
  - L_h: Average length of interconnections
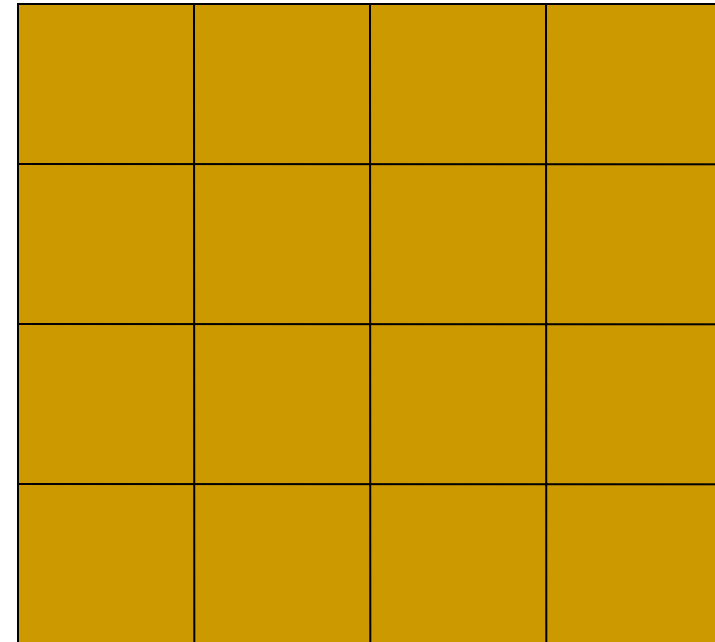
Total wirelength: $\quad L_{tot} = \sum_{h=0}^{H} n_h L_h$

Average Length of Interconnection

- L_{TB} : WL, transparent blockage

- L_{DT}^{v} : Vertical detour

- L_{DT}^{h} : Horizontal detour

- Average WL

$$\overline{L}_{intra} = \overline{L}_{TB} + \overline{L}_{DT}^{h} + \overline{L}_{DT}^{v}$$

Average Length of Interconnection

- ## Horizontally, vertically, diagonally adjacent bins
    - Transparent block
    - Horizontal detour
    - Vertical detour

- ## Detour happens if two terminals are on different sides of IP Block
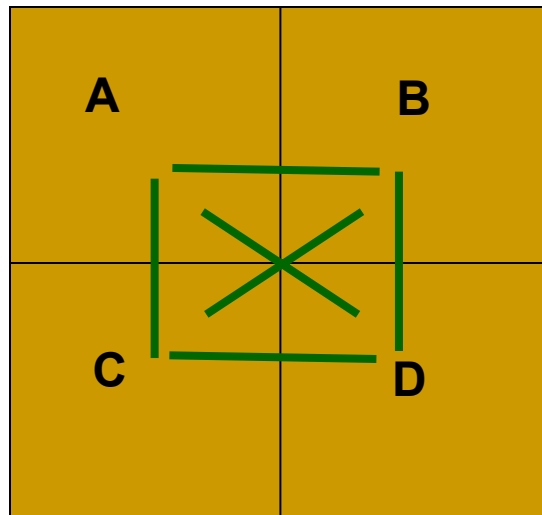   - Probabilistic nature
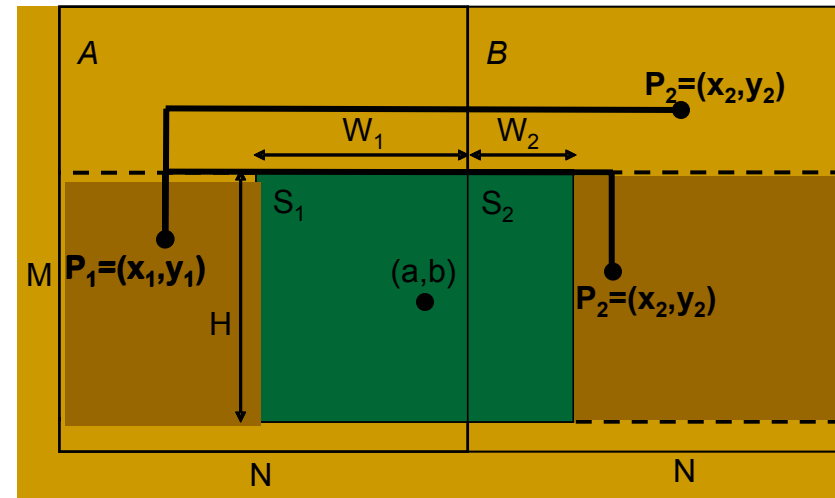
Average Estimated WL per Level

$$\overline{L}_{\text{inter},l} = \frac{1}{6}\left(\begin{array}{l} \delta\left(L_{\text{inter}}^{h}(A,B) + L_{\text{inter}}^{h}(C,D) + L_{\text{inter}}^{v}(A,C) + L_{\text{inter}}^{h}(B,D)\right) \\ + (1-\delta)\left(L_{\text{inter}}^{d}(A,D) + L_{\text{inter}}^{d}(B,C)\right) \end{array}\right)$$

δ captures optimization behavior of placement algorithms

Open Research Problems

- Study the effect of IP blocks on Rent exponent

- Considering the effect of congestion around IP blocks and slivers

- Extending WL estimation for timing driven placement methods

Outline

- **Motivation**
- **WL prediction**
  - Individual net
  - Total interconnect length
- **Congestion prediction**
- **Concurrent wirelength and congestion prediction**
- **Density utilization alleviation**
- **Statistical timing analysis**
- **Integrated congestion and delay prediction**
- **Conclusion**

ER
UCLA

Routing Congestion Prediction

- Several post-placement algorithms presented to analyze congestion before routing.

- Use routing estimation model to predict routing congestion

- Derive the mathematical equations to estimate the congestion using a normal distribution approximation

- Use a probabilistic approach to estimate the congestion
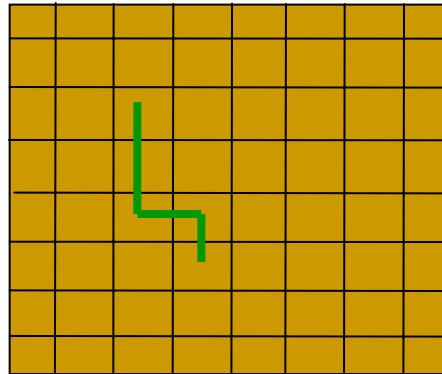
Congestion Prediction

Concurrent Congestion and Wirelength Estimation

- High-utilization benchmarks
  - Congestion is very likely
  - Detour b/c of congestion

- WL estimation needs congestion map
  - Modeling congested areas as artificial IP blocks

- Regional congestion estimation
  - Needs wirelength estimation for internal routing demand

- Concurrent congestion & wirelength estimation
  - Hierarchical bottom-up approach

ER
UCLA

Congestion Prediction

Regional Congestion Estimation

- ## Routing demand: $D(r) = ID(r) + ED(r)$

  - ❏ Internal routing demand *ID(r)*

  - ❏ External routing demand *ED(r)*

- ## Internal routing demand

  - ❏ Equal to WL
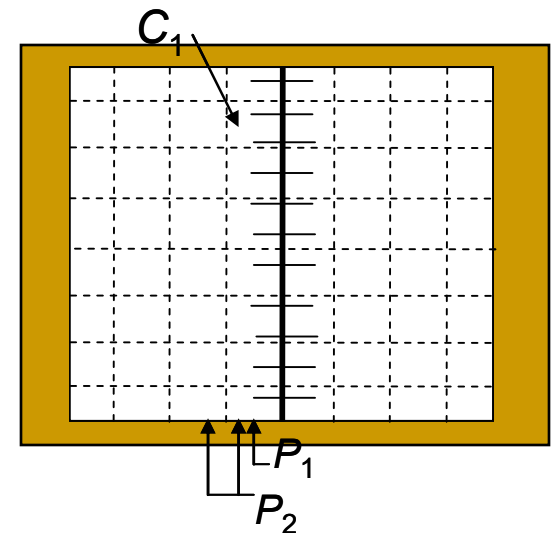
Congestion Prediction

Regional Congestion Estimation

- **External Routing Demand**

  - The probability of occurring a wire of length l, $P_l = l^{-(4-2P)}$

  - If NB = # of bins ➡️ $\sum_{i=2}^{2\sqrt{NB}} P_i + P_1 = 1$

  - And, we have ➡️ $\dfrac{p_{i+1}}{p_i} = \dfrac{(i+1)^{-(4-2P)}}{(i)^{-(4-2P)}}$

  - Calculate $P_1, \ldots, P_{2\sqrt{NB}}$

ER
UCLA

Congestion Prediction

Regional Congestion Estimation

- **External Routing Demand**
  - Calculate $P_1, \ldots, P_{2\sqrt{NB}}$

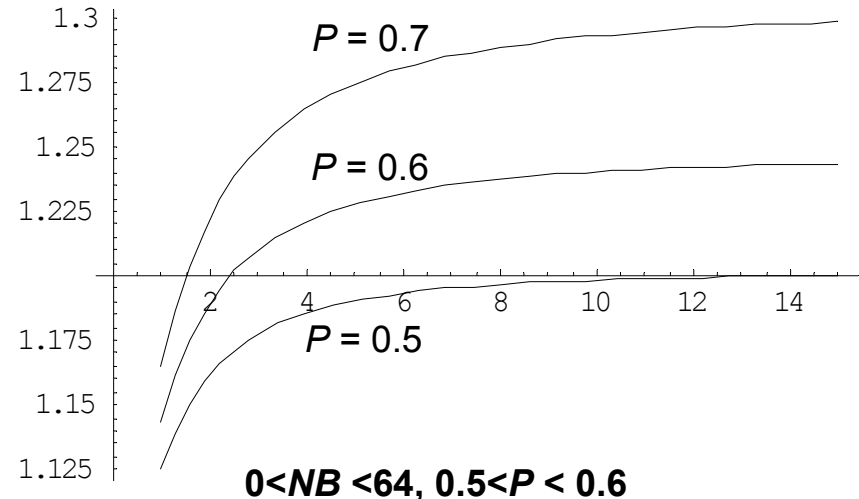 $$\sum_{i=1}^{2\sqrt{NB}} P_i = 1$$

  $$P_2 \sum_{i=1}^{2\sqrt{NB}} (i+1)^{-(4-2P)} + P_1 = 1$$

  - Riemann Zeta function:

  $$\zeta(x) = \frac{1}{\Gamma(x)} \sum_{0}^{\infty} \frac{u^{x-1}}{e^u - 1} du$$

  $$\sum_{i=0}^{2\sqrt{NB}-1} (i+1)^{-(4-2P)} = Zeta(4-2P) - Zeta[4-2P, 2\sqrt{NB}+1]$$

  - Total external demand: $ED = \sum_{i=1}^{2\sqrt{NB}} C_i \times P_i$

P = 0.7

P = 0.6

P = 0.5

$0 < NB < 64,\ 0.5 < P < 0.6$

ER
UCLA

Outline

- Motivation
- WL prediction
  - Individual net
  - Total interconnect length
- Congestion prediction
- Concurrent wirelength and congestion prediction
- Density utilization alleviation
- Statistical timing analysis
- Integrated congestion and delay prediction
- Conclusion

Concurrent WL and Cong. Prediction

- **Bottom-up hierarchical analysis**

- **Congestion estimation**
  - ❑ Analyze external routing demand
  - ❑ Analyze internal routing demand using WL of previous level
  - ❑ Obtain congestion map of this level

- **WL Estimation**
  - ❑ Obtain avg. no. of interconnect per level
  - ❑ Analyze avg. length of interconnect
  - ❑ Model congested areas (known from previous level) as artificial IP blocks

ER
UCLA

Open Research Problems

- Considering the effect of vias in the estimation

- Extending congestion estimation for 3D placement

- Extending the concurrent estimation methods for congestion and wirelength on the other estimation methods

Outline

- Motivation
- WL prediction
  - Individual net
  - Total interconnect length
- Congestion prediction
- Concurrent wirelength and congestion prediction
- Density utilization alleviation
- Statistical timing analysis
- Integrated congestion and delay prediction
- Conclusion

Density Utilization Alleviation

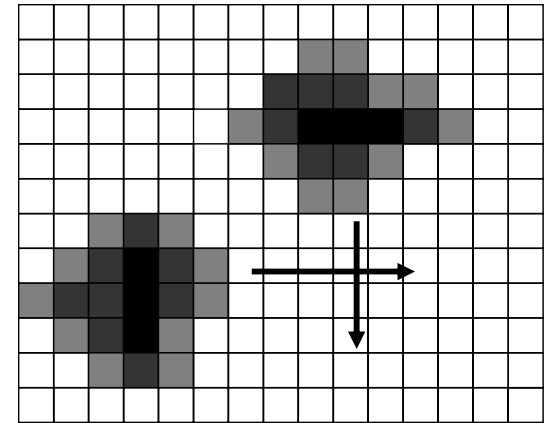- Need a global view of density (congestion) map
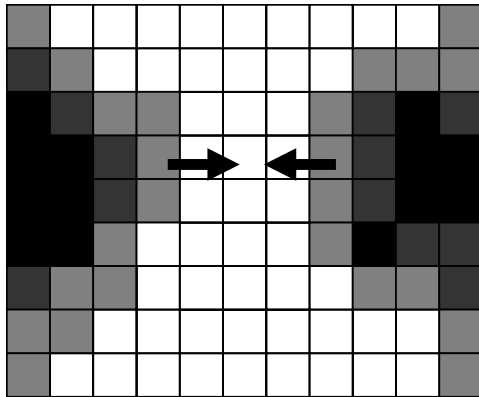
Local density alleviation
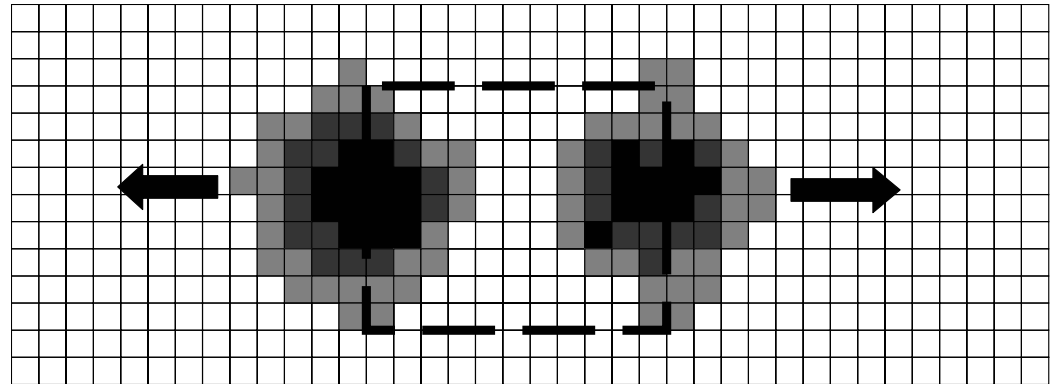
Global density alleviation

Density Utilization Alleviation

- **Flow algorithms**
  - Local view
  - Ignore WL optimization
- **Solution**
  - White space allocation
  - Linear/quadratic
  - Greedy

The interfering flow in the crossing of two arrows

Density Utilization Alleviation

- ## Cell Redistribution
  - Form working area round highly-congested bins
  - Redistribute cells in working (by using min-cut partitioning)

- ## Cell Migration
  - Move cells out of highly-congested bins
  - Accept move if reduces congestion
  - Different Criteria to accept a move

Open Research Problems

- Incorporating the issues like
  - Thermal placement and hot spot distribution in congestion alleviation algorithms.
- Considering the white space needed around the big IP blocks in the density alleviation algorithms
 - Guarantee routability in later phases.

Outline

- Motivation
- WL prediction
  - Individual net
  - Total interconnect length
- Congestion prediction
- Concurrent wirelength and congestion prediction
- Density utilization alleviation
- Statistical timing analysis
- Integrated congestion and delay prediction
- Conclusion

Statistical Timing Analysis

- **Key parameters in device variations**
  - Length L,
  - Threshold voltage V_t
  - Oxide thickness t_{ox}

- **Process parameter variations normally distributed as random variables**

- **Gate delay, d, will have a probability distribution function: Gaussian distribution with parameters μ and σ.**

ER
UCLA

Variation Sources
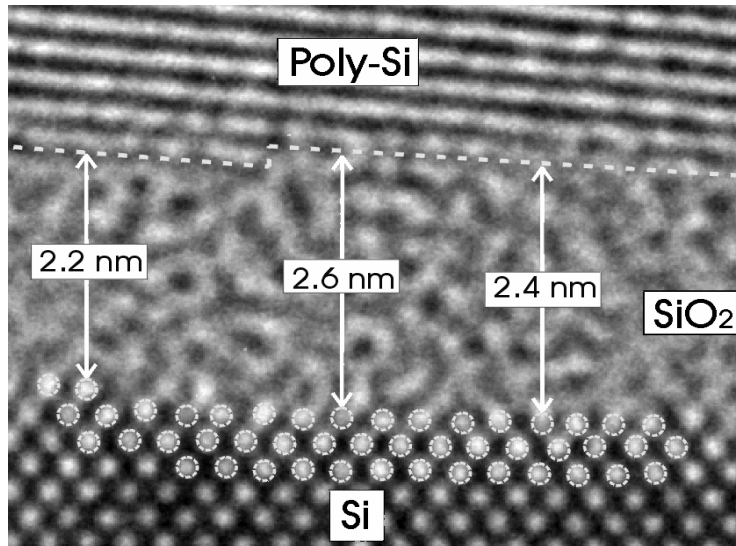
t_{ox}
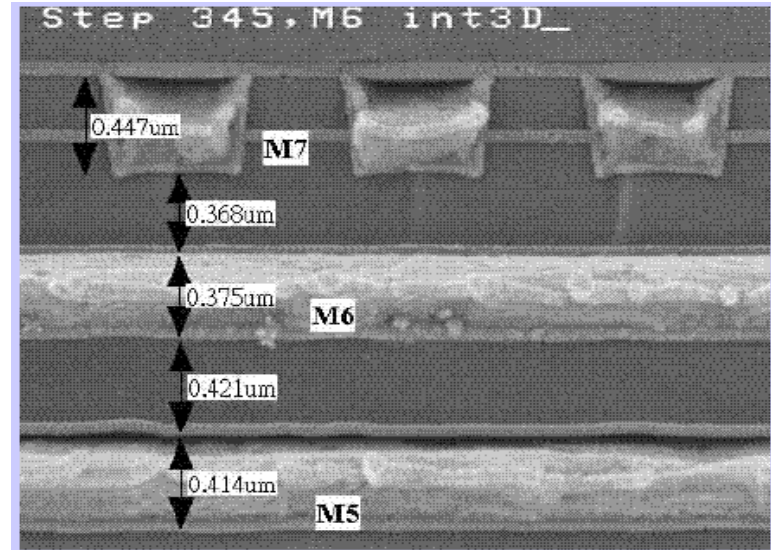
Wire Thickness

**Target=0.375um**

• Figures are courtesy of IBM, Intel and TSMC

Gate Delay Distribution

- Gate Delay: *almost* worst case delay

$$prob\left(d > D_{gate}\right) < 0.05$$

- Delay distribution of a circuit using two operators: Max and Sum

- Re-convergent paths produce correlations between delay

- No set of distributions which is closed under both sum and max

- Quantization can be applied to work with discrete distributions with a polynomial associated:

$$G_d\left(x\right) = \sum_{i=1} P_i x^{d_i}$$

ER
UCLA
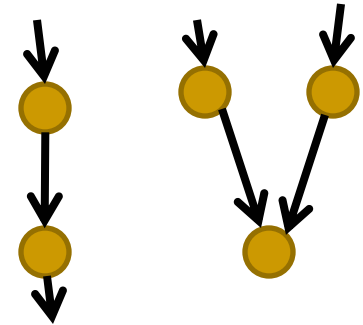
MAX operation: inherently non-linear
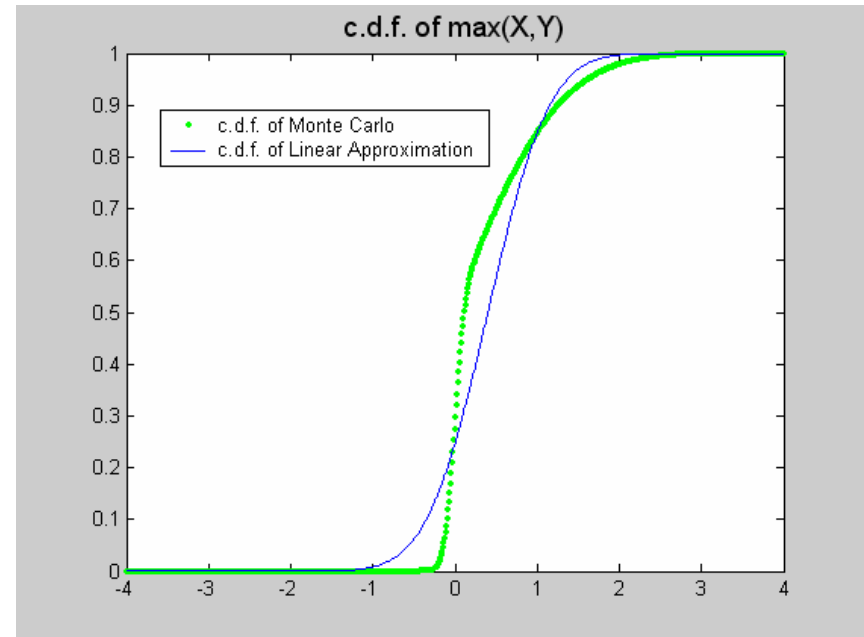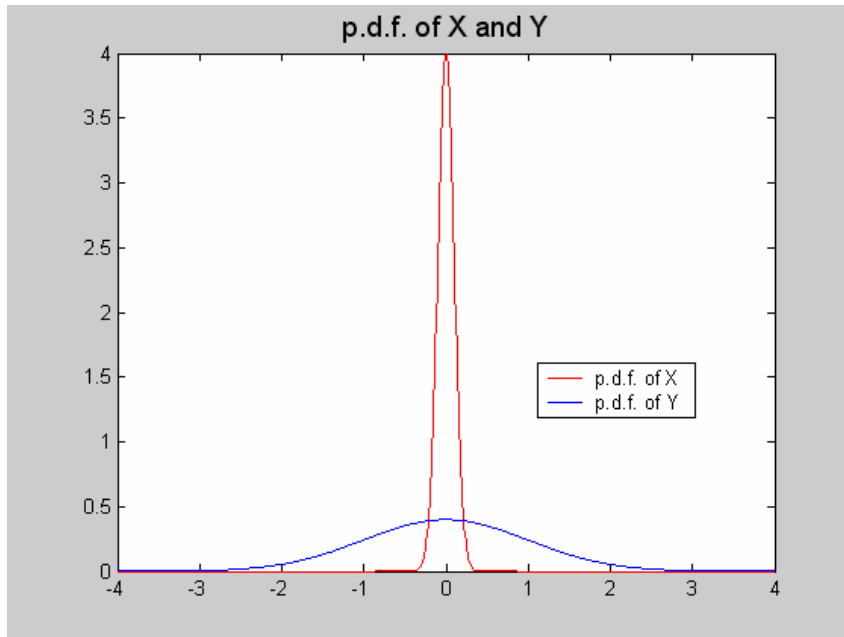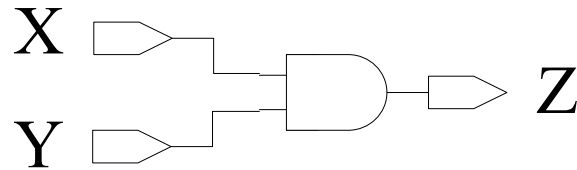
Timing Model

- **Canonical timing model for gate/wire delay**
  - Assumption: Gate/wire delay is linearly dependent on the variation sources
  - Problem: Non-linear dependency of delay on process parameters
- **Linear MAX approximation for circuit delay**
  - Problem: Significant error could occur in some cases

Open Problems

- Obtain a probabilistic model nets with re-convergent fan-outs

  - Half perimeter approximation no longer accurate because of the branches in a net

- Analyze delay distribution for nets during placement.

- Incorporating delay distribution into placement algorithm in timing driven placement

Outline

- **Motivation**
- **WL prediction**
  - ❑ Individual net
  - ❑ Total interconnect length
- **Congestion prediction**
- **Concurrent wirelength and congestion prediction**
- **Density utilization alleviation**
- **Statistical timing analysis**
- **Integrated congestion and delay prediction**
- **Conclusion**

Integrated Cong. & Delay Estimation

- Hierarchical CAD flows to consider wirelength, congestion, thermal issues and power early in design flow
  - Pure wirelength minimization without congestion estimation methods may lead to un-routable designs
  - Wirelength minimization along with congestion removal but ignoring thermal issues may lead to designs with several hot-spots.
- All these predictions should be highly integrated
  - Goal is to degrade the harmful side effects of optimizing each parameter individually

Integrated Cong. & Delay Estimation
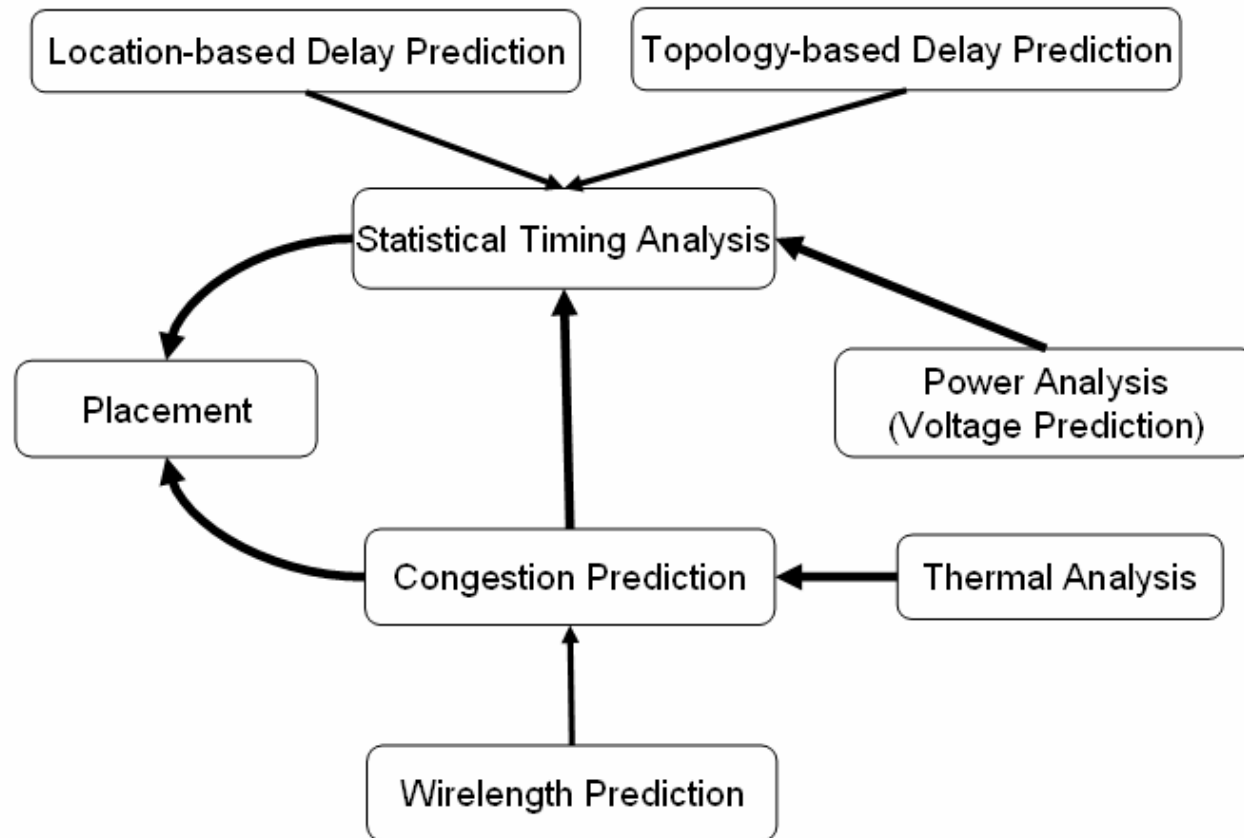
Conclusion

- **Integrated prediction methods**
- **WL estimation**
  - Individual net length
  - Total interconnect length
- **Congestion Estimation**
  - Pre-Placement
  - Post-Placement
- **Timing Analysis**
- **Integrated Estimation of WL, Congestion, and Delay**

Thank You!