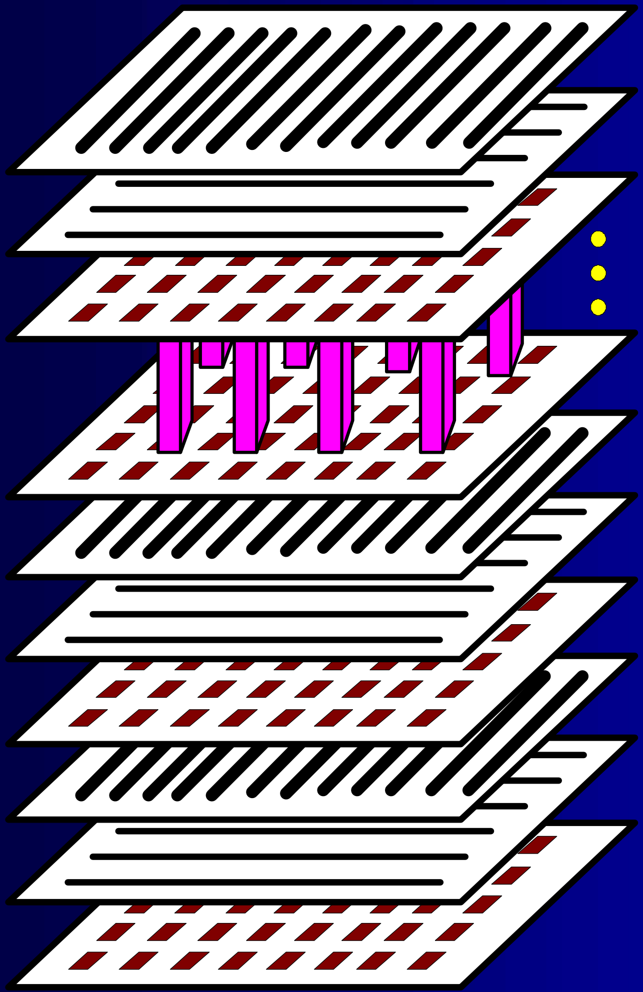
Temperature-Aware Routing in 3D ICs

Tianpei Zhang, Yong Zhan and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering
University of Minnesota

Outline

Temperature-aware 3D global routing



- Introduction to 3D ICs and global routing model
- Thermal model in 3D routing
- 3D global routing flow
- Linear programming-based thermal via and thermal wire insertion for temperature reduction
- Experimental results

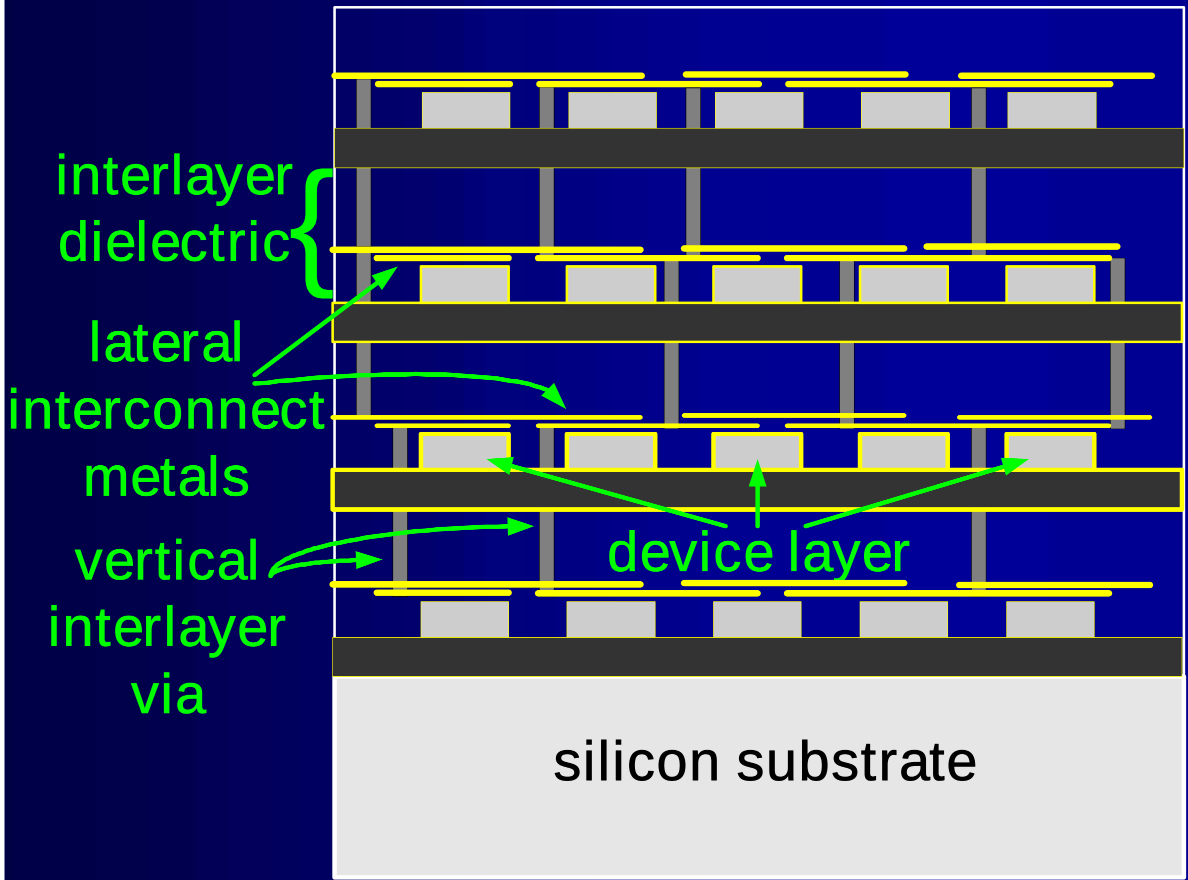
Outline

Temperature-aware 3D global routing

- ✓ Introduction to 3D ICs and global routing model
- Thermal model in 3D routing
- 3D global routing flow
- Linear programming-based thermal via and thermal wire insertion for temperature reduction
- Experimental results

3D Integrated Circuits

- 3D ICs: Stacking multiple device layers into a monolithic structure.

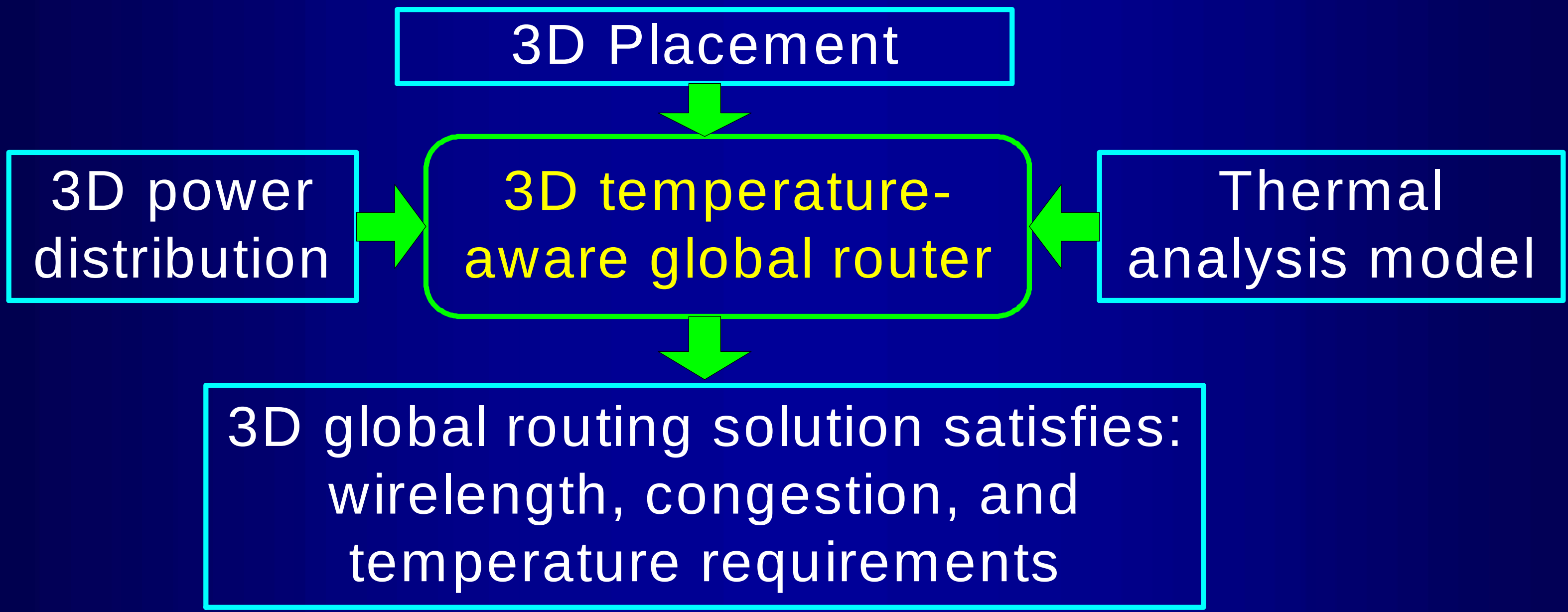


Advantages over 2D ICs:

- Better performance: wire length, timing, power, etc.
- Integrate more devices.
- Good substrate isolation, platform to integrate digital ICs, analog ICs for system on chip (SoC) application.

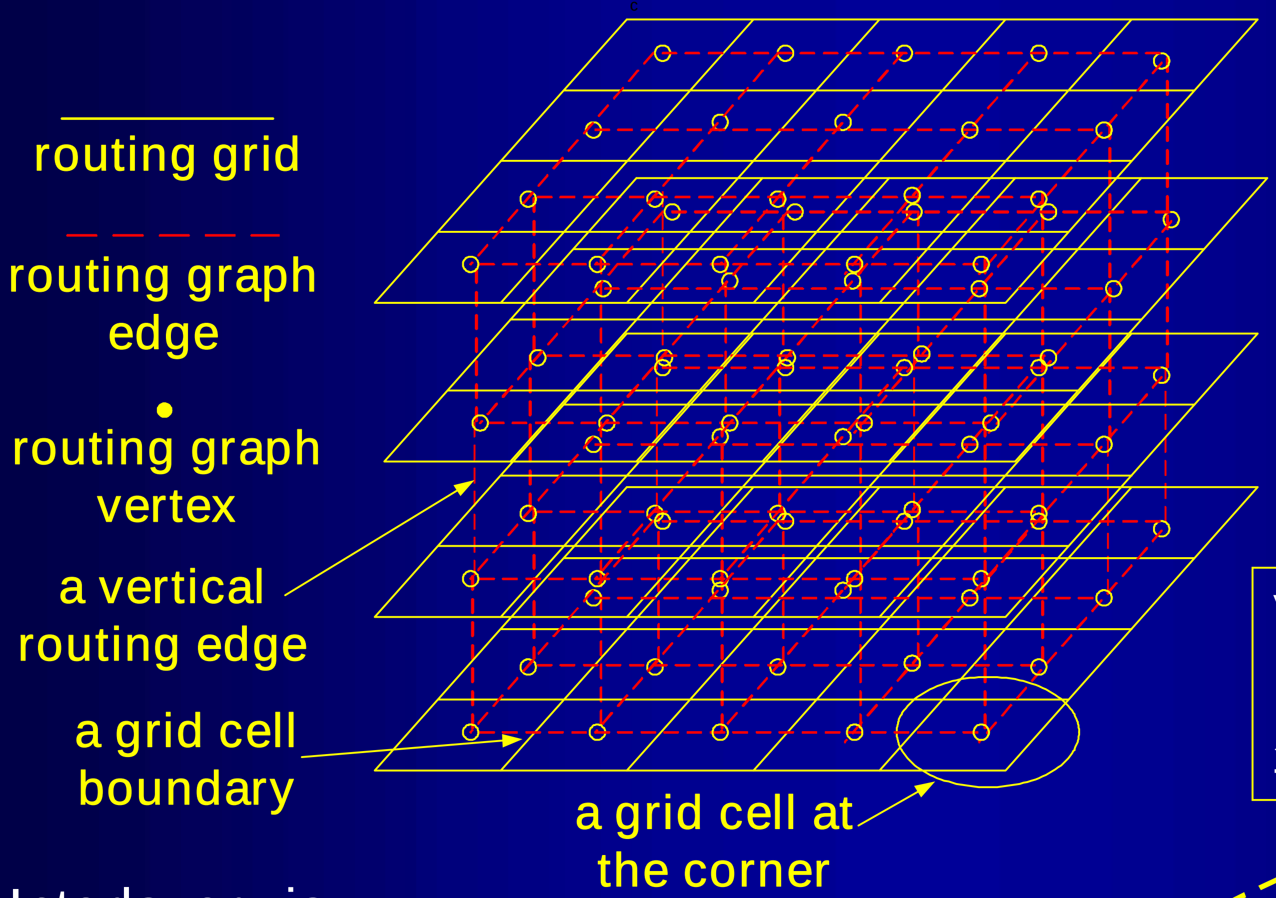
3D IC Global Routing Challenges

- 3D ICs are plagued with thermal problems:
  - High power density.
  - High thermal resistance of insulating dielectric.
- 3D global routing should be electrothermally-conscious:



- More challenges: routing in three dimensions: inherently more complex than routing in 2D ICs.

3D Global Routing Model



Routing constraints:

Intralayer edge:
wire usage
≤ edge capacity

Vertical edge:
interlayer via usage
≤ interlayer capacity

Interlayer via:

- Large (manufacturability and reliability consideration).
- Go through device layer, compete with devices for silicon area; maximum number of interlayer vias is determined by white space.

Outline

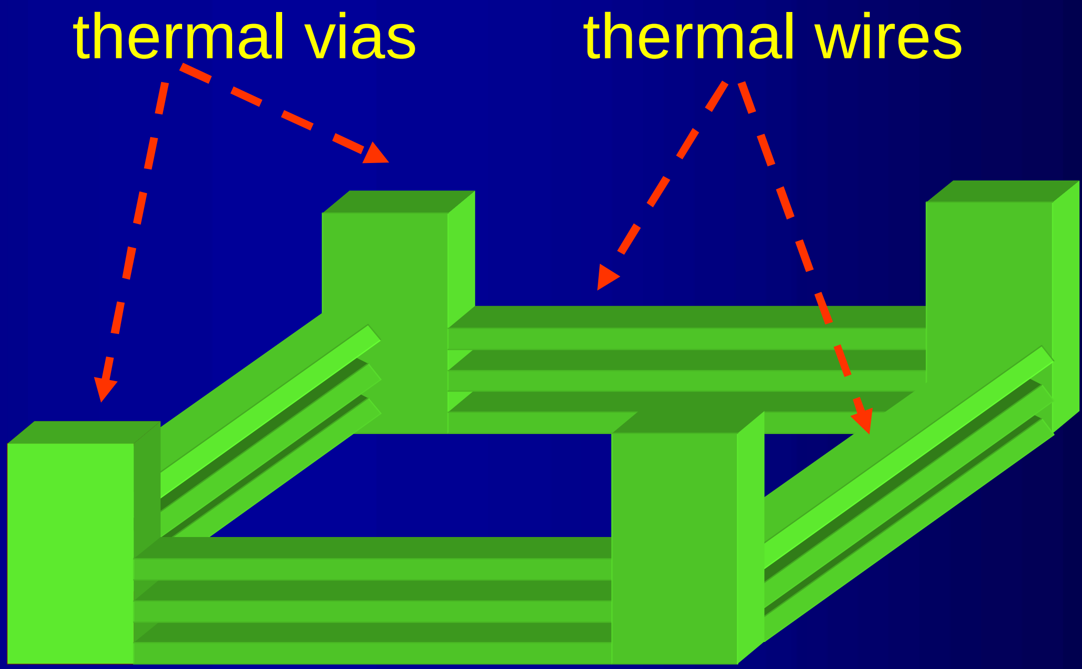
Temperature-aware 3D global routing

- Introduction to 3D ICs and global routing model
- ✓ Thermal model in 3D routing
- 3D global routing flow
- Linear programming-based thermal via and thermal wire insertion for temperature reduction
- Experimental results

Reducing Temperature in 3D Routing

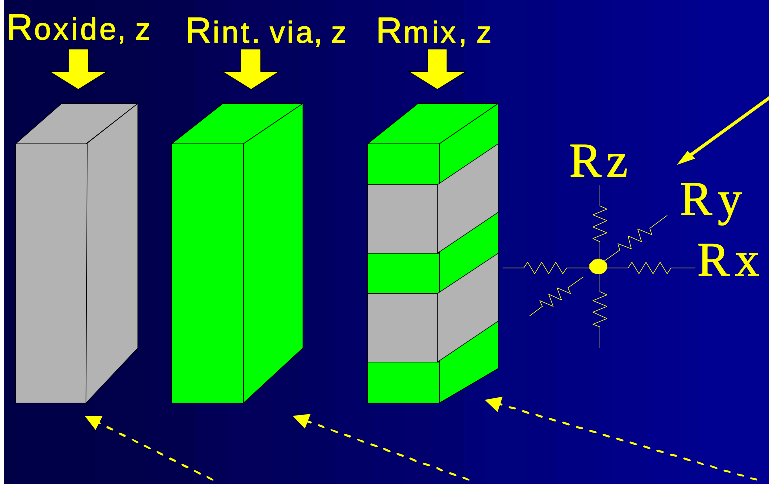
- Build good heat conduction path through dielectric:
  - Thermal vias: interlayer vias dedicated to thermal conduction.
  - Thermal wires: metal wires improve lateral heat conduction.
  - Thermal vias + thermal wires → a thermal conduction network.

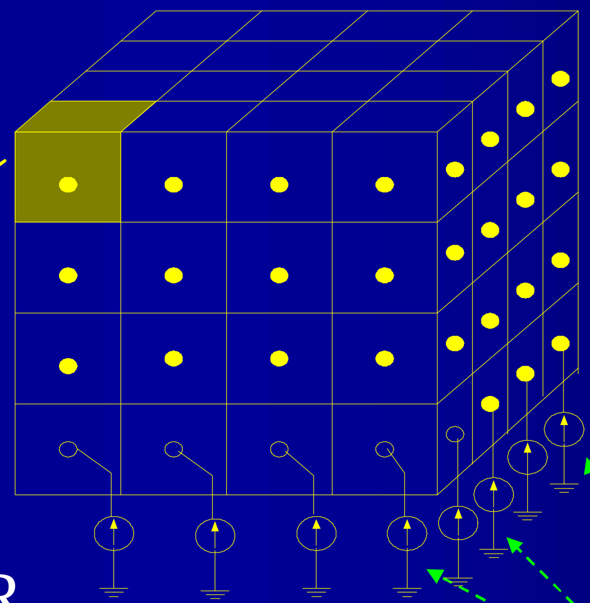
- Thermal wires:
  compete with
  lateral signal wire
  routing.
- Thermal vias:
  large, can block
  lateral signal
  routing capacity.



Thermal Analysis Model

- Thermal circuit model (C. H. Tsai, et al, IEEE Trans. CAD, 2000):
  - Ability to handle non-uniform thermal conductivity.
  - Ease of handling sensitivity information.





Temperature profile vector

$$MT = P$$

Nodal analysis coefficient matrix

Power dissipation vector

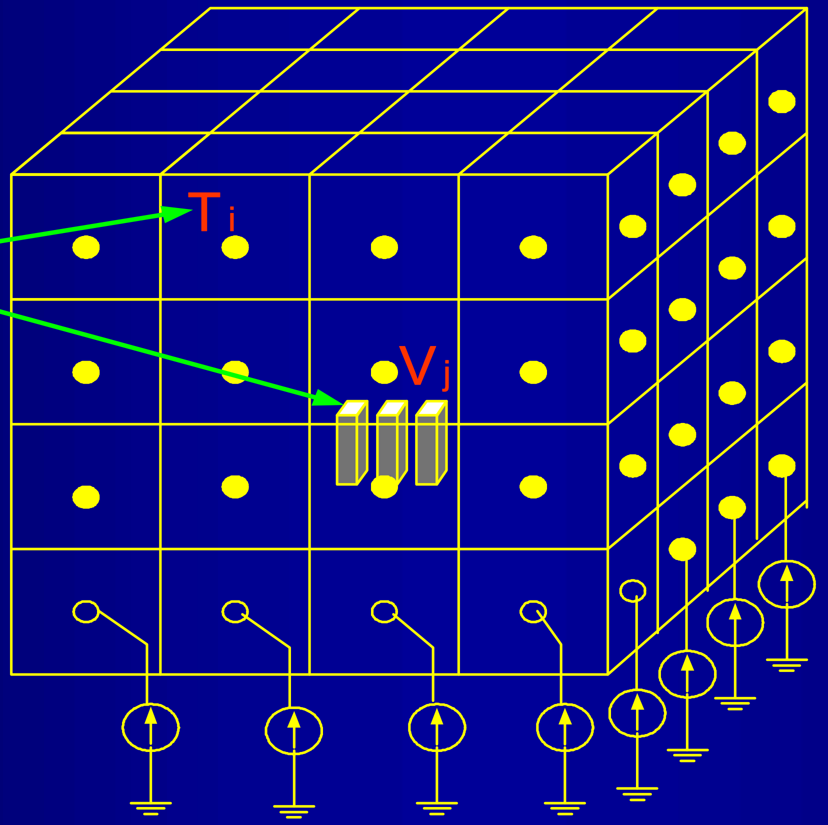
$$R_z = R_{oxide,z} \parallel R_{int.via,z} \parallel R_{mix,z}$$

$$R_x = R_{oxide,x} \parallel R_{int,x} \parallel R_{mix,x}$$

$$R_y = R_{oxide,y} \parallel R_{int,y} \parallel R_{mix,y}$$

Power generated in cells

Thermal Analysis Model



Sij
sensitivity of temperature Ti to the number of interlayer vias Vj

$$MT = P$$

Adjoint sensitivity analysis

To identify locations where thermal via and thermal wire insertions are most effective.

$$s_{ij} = \frac{\partial T_i}{\partial V_j} = \sum_{M_{kl} \in S_j} \frac{\partial T_i}{\partial M_{kl}} \frac{\partial M_{kl}}{\partial V_j}$$

$$\frac{\partial T_i}{\partial M_{kl}} = -\xi_{ik} T_l \qquad MT = P$$

Outline

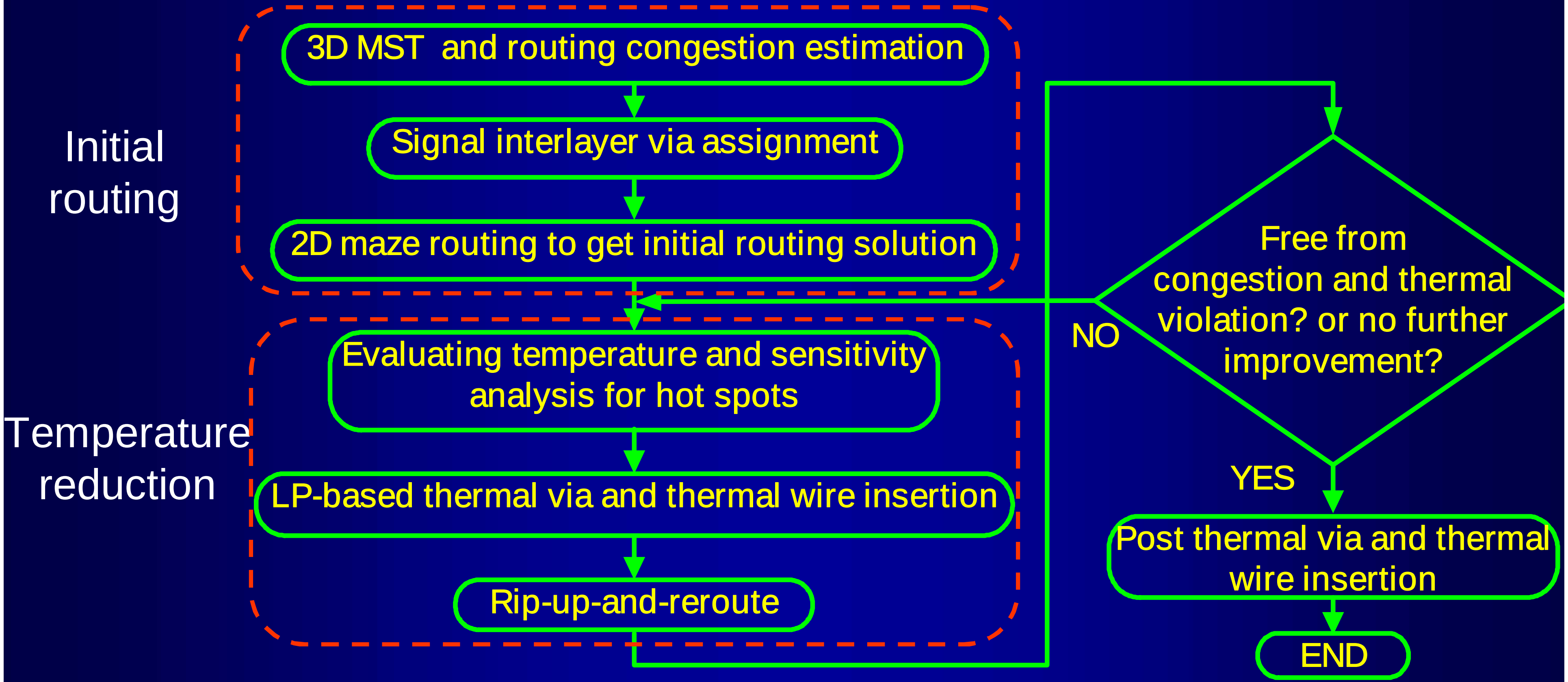
Temperature-aware 3D global routing

- Introduction to 3D ICs and global routing model
- Thermal model in 3D routing
- ✓ 3D global routing flow
- Linear programming-based thermal via and thermal wire insertion for temperature reduction
- Experimental results

Temperature-Aware 3D Global Routing Problem Formulation

- Given:
  - Placement of a 3D IC.
  - Power map of a 3D IC.
  - 3D technology parameters.
  - A specified value of maximum temperature, T_{spec}.
- Goal: A global routing solution with thermal vias and thermal wires inserted, and satisfies:
  - Routing wire capacity and interlayer via capacity.
  - Peak temperature lowered towards T_{spec}.

3D Temperature-Aware Global Routing Flow



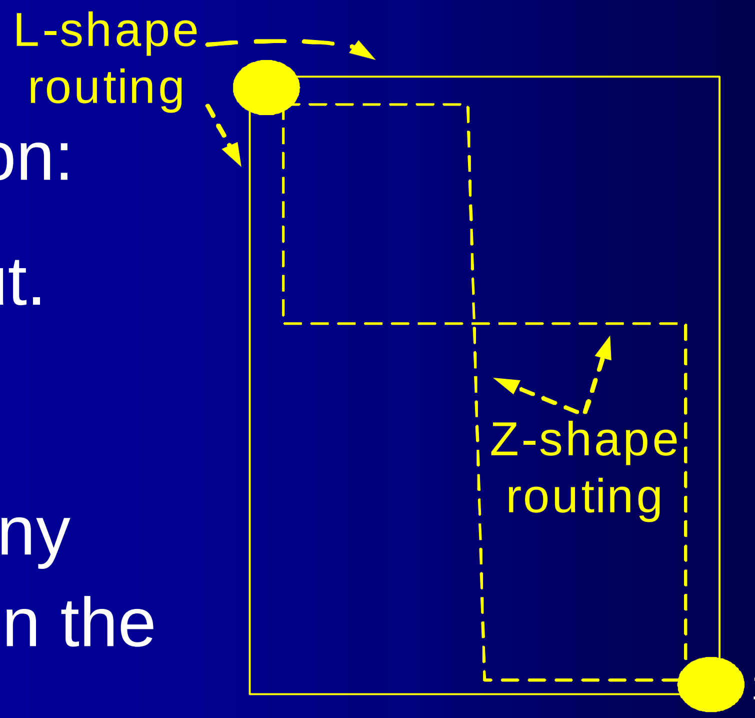
3D MST and Routing Congestion Estimation

- 3D Minimum Spanning Tree (MST): decompose all nets into a set of 2-pin nets.

Cost function: $C_{ij} = |x_i - x_j| + |y_i - y_j| + \alpha |z_i - z_j|$

large weight to discourage interlayer vias.

- Routing Congestion Estimation:
  - Take all 2-pin nets as input.
  - L/Z-shape routing model.
  - Equal probability to take any interlayer via position within the net bounding box.

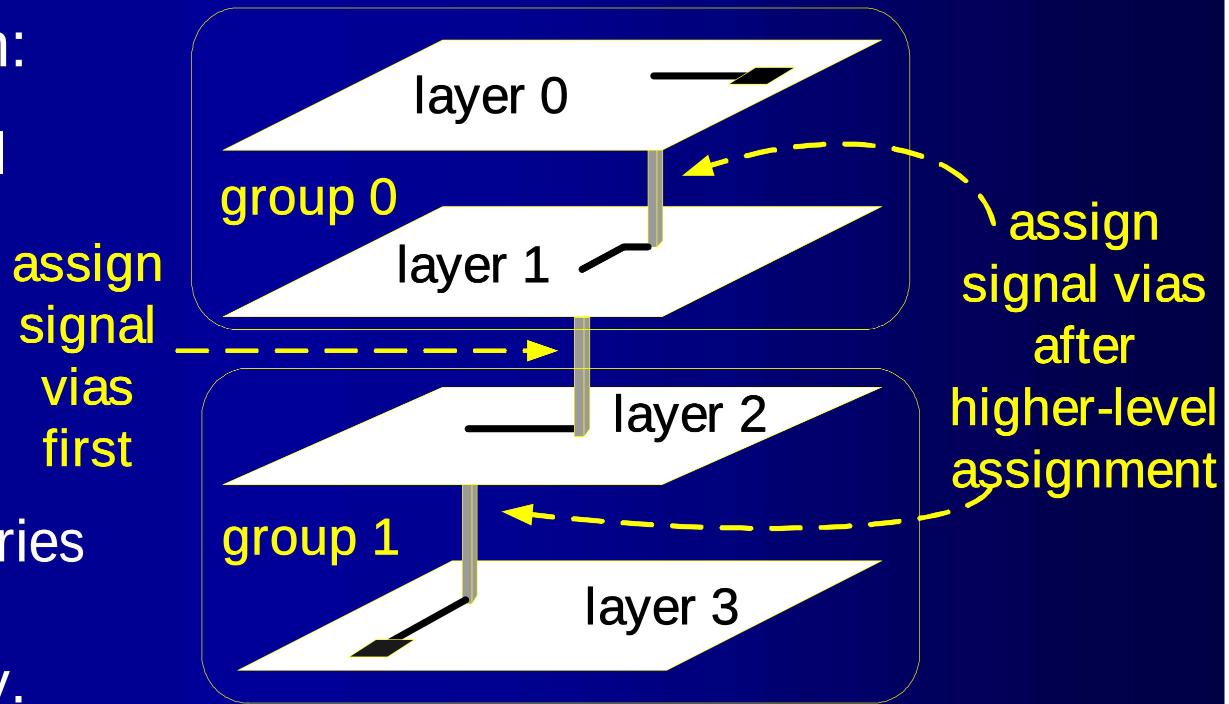


Signal Interlayer Via Assignment

- Determine the signal interlayer via positions for all interlayer nets so that a cost function is minimized.

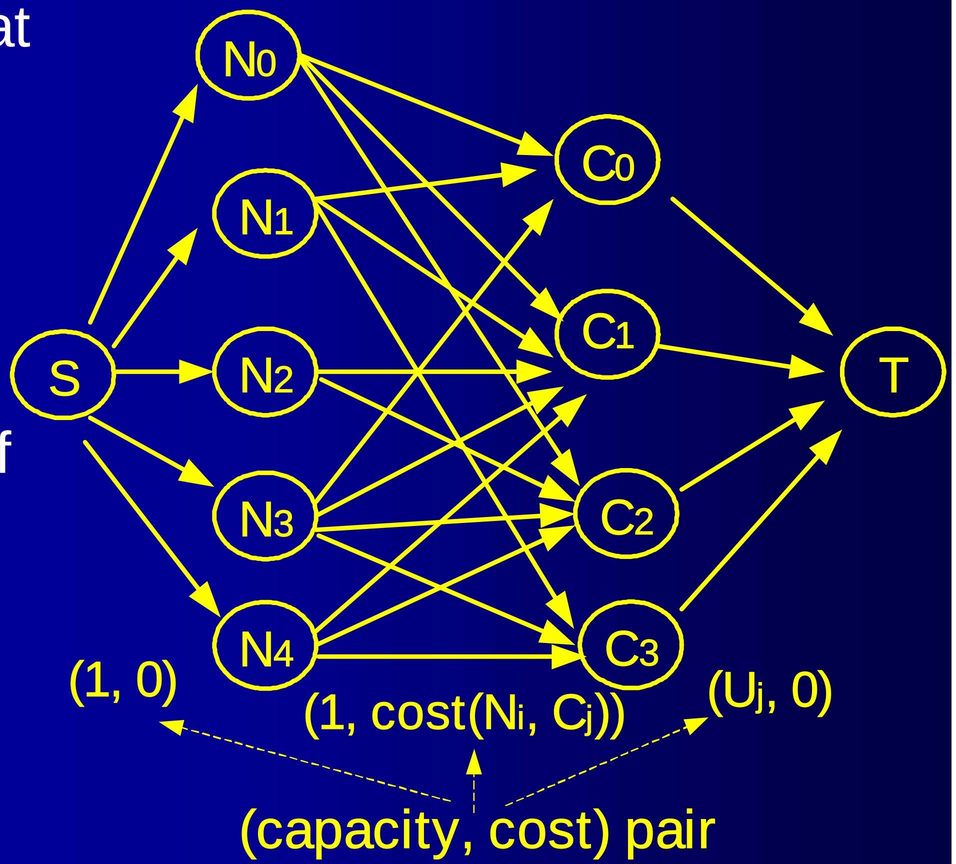
$$\text{Wire length} + \lambda \bullet \text{Congestion cost}$$

- Hierarchical approach:
  - – Recursively divide all device layers into two neighboring groups of equal size.
  - – Signal interlayer via assignment at boundaries in a top-down way following the hierarchy.



Via Assignment and 2D Maze Routing

- Signal interlayer via assignment at each boundary:
  - Each interlayer net i ↔ node Ni.
  - Each possible interlayer via position j ↔ node Cj.
  - An edge directed from Ni to Cj, if position j is within the bounding box of net i.
  - Solve the min-cost network flow problem.



- 2D Maze Routing:
  - Propagates wave-front-like expansion.
  - Includes temperature term in cost function: net seeks low-temperature path, better delay, less congestion in hot region.
  - Rip-up and reroute.

Outline

Temperature-aware 3D global routing

- Introduction to 3D ICs and global routing model
- Thermal model in 3D routing
- 3D global routing flow
- ✓ Linear programming-based thermal via and thermal wire insertion for temperature reduction
- Experimental results

Thermal Via and Thermal Wire Insertion

Initial 3D routing solution

Thermal analysis

Update temperature profile, obtain current highest temperature T_{max}

Set up target temperature $T_{targ}=T_{max}^{\mu}\bullet T_{spec}^{1-\mu}$

Perform sensitivity analysis for temperature violation spots with $T > T_{targ}$

LP-based thermal via / wire insertion

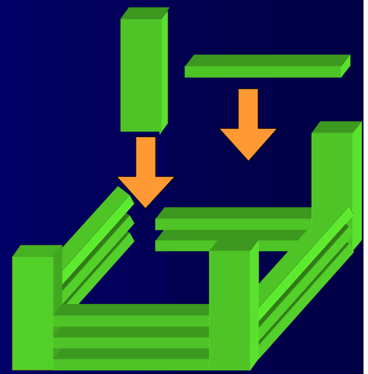
Linear programming (LP) based thermal via and thermal wire insertion

Rip-up-and-reroute for resolving overflow

Still temperature/congestion violation?

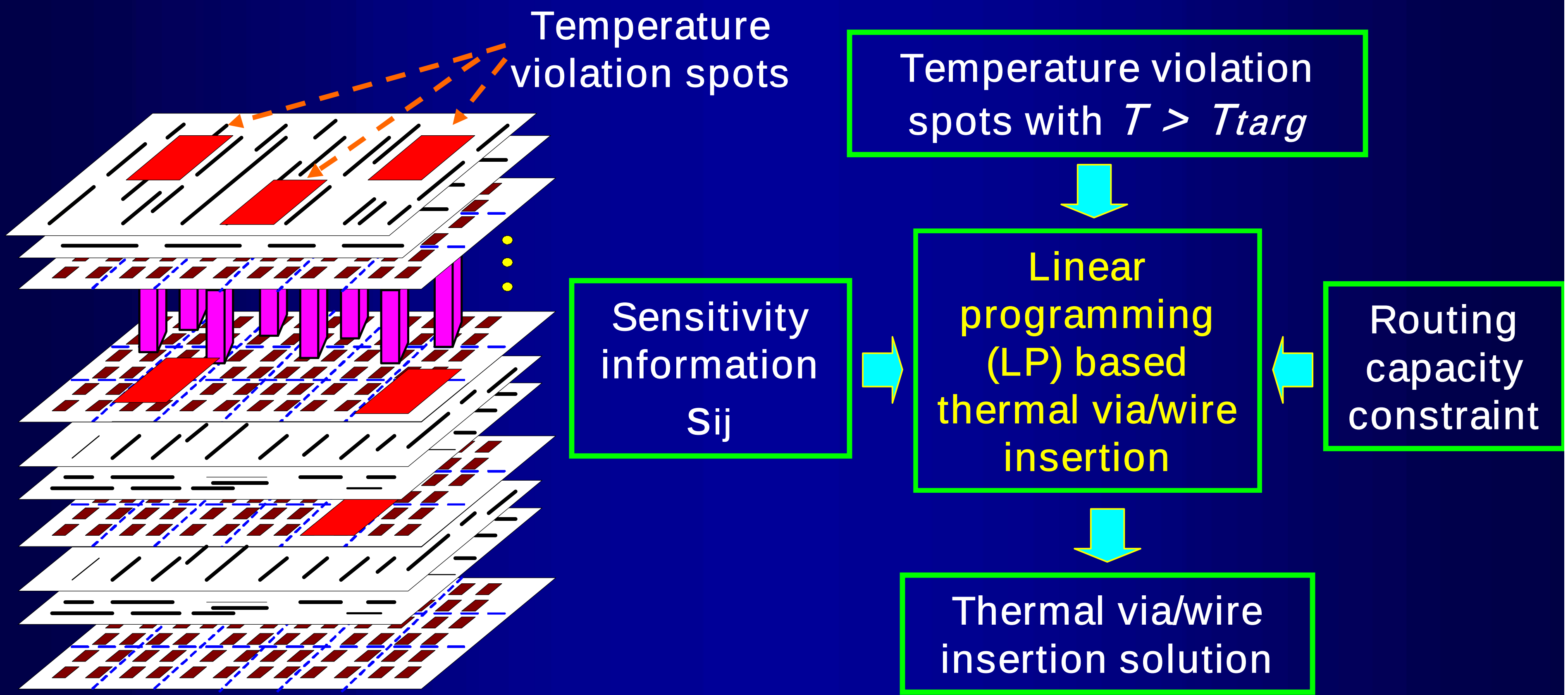
Feasible 3D routing solution

Thermal via / wire insertion



LP-based Thermal Via / Wire insertion

- In each iteration, resolve temperature violation of $T > T_{targ}$ in routing solution : linear programming (LP) based thermal via/wire insertion.



LP-based Thermal Via / Wire Insertion

- Linear programming formulation:

minimize total resource usage

relaxation variable to guarantee feasibility

$$\text{minimize} \sum_{j=1}^{p} N_{v,j} + \sum_{k=1}^{q} N_{w,k} + \Gamma \sum_{i=1}^{n} \delta_i$$

number of thermal vias and thermal wires inserted

subject to:

temperature constraint

$$\sum_{j=1}^{p} -s_{v,ij} N_{v,j} + \sum_{k=1}^{q} -s_{w,ik} N_{w,k} + \delta_i \geq \Delta T_i,$$

$$i = 1,2\ldots,n,\ \Delta T_i = T_i - T_{targ}$$

sensitivities of temperature to number of thermal vias/wires

routing capacity constraint

$$N_{v,j} \leq \min((1+\beta)R_{v,j}, U_j - V_j),\ j = 1,2,\ldots,p$$

$$N_{w,k} \leq (1+\beta)R_{w,k},\ k = 1,2\ldots,q$$

$$\delta_i \geq 0, i = 1,2,\ldots,n;\ N_{v,j} \geq 0,\ j = 1,2,\ldots,p;\ N_{w,k} \geq 0, k = 1,2,\ldots,q$$

LP-based Thermal Via / Wire Insertion

- Routing capacity constraints are relaxed by a factor β :

$$N_{v,j} \le \min((1+\beta)R_{v,j}, U_j - V_j),\ j = 1,2,\ldots,\ p$$

$$N_{w,k} \le (1+\beta)R_{w,k},\ k = 1,2\ldots,\ q$$

- Temporarily permit routing overflow, allow better temperature reduction.
- Rip-up-and-reroute will resolve the overflow.
- Optimized solution reached in a small number of iterations.

Outline

Temperature-aware 3D global routing

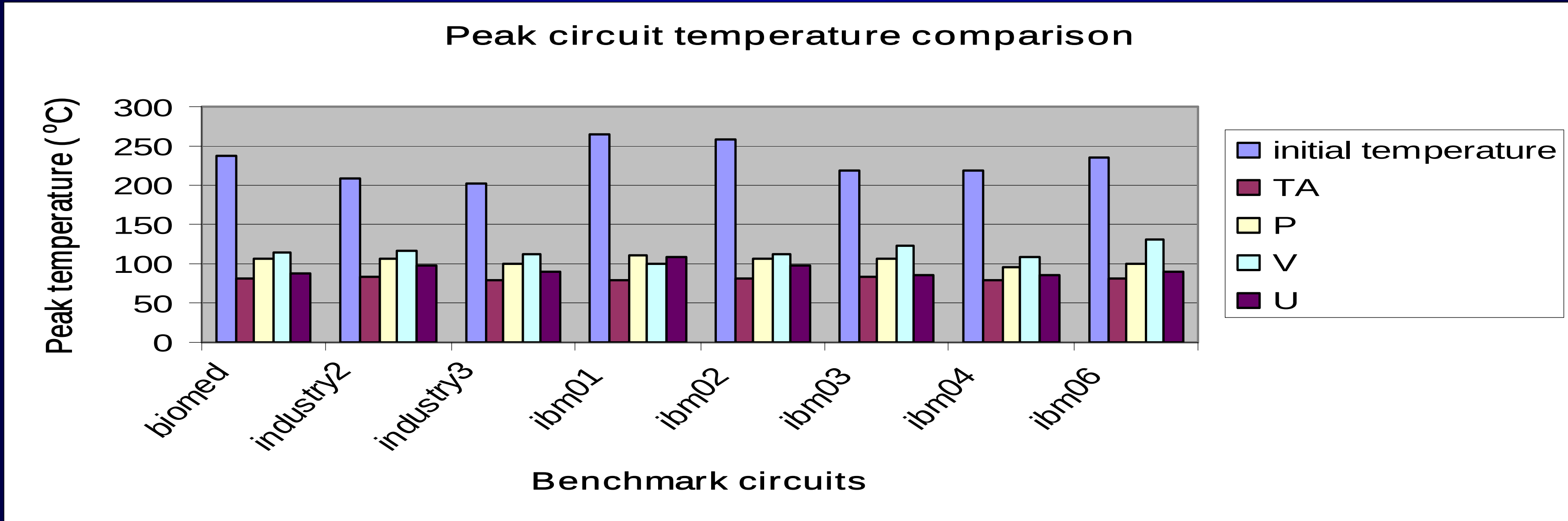
- Introduction to 3D ICs and global routing model
- Thermal model in 3D routing
- 3D global routing flow
- Linear programming-based thermal via and thermal wire insertion for temperature reduction

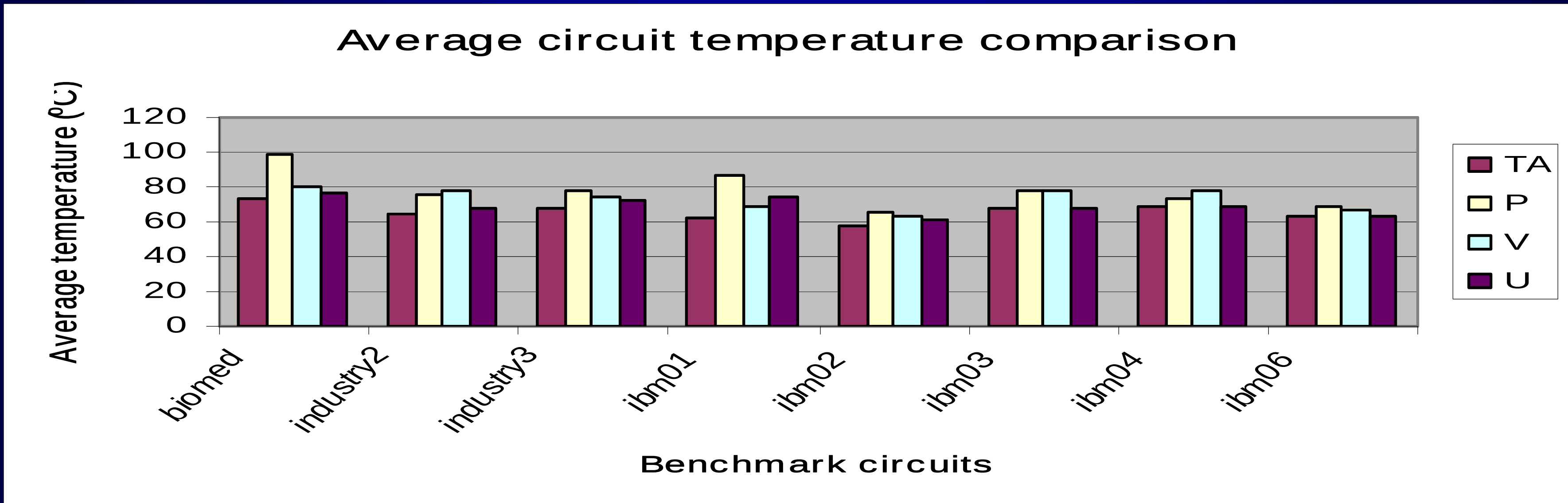
✓ Experimental results

Experimental Results

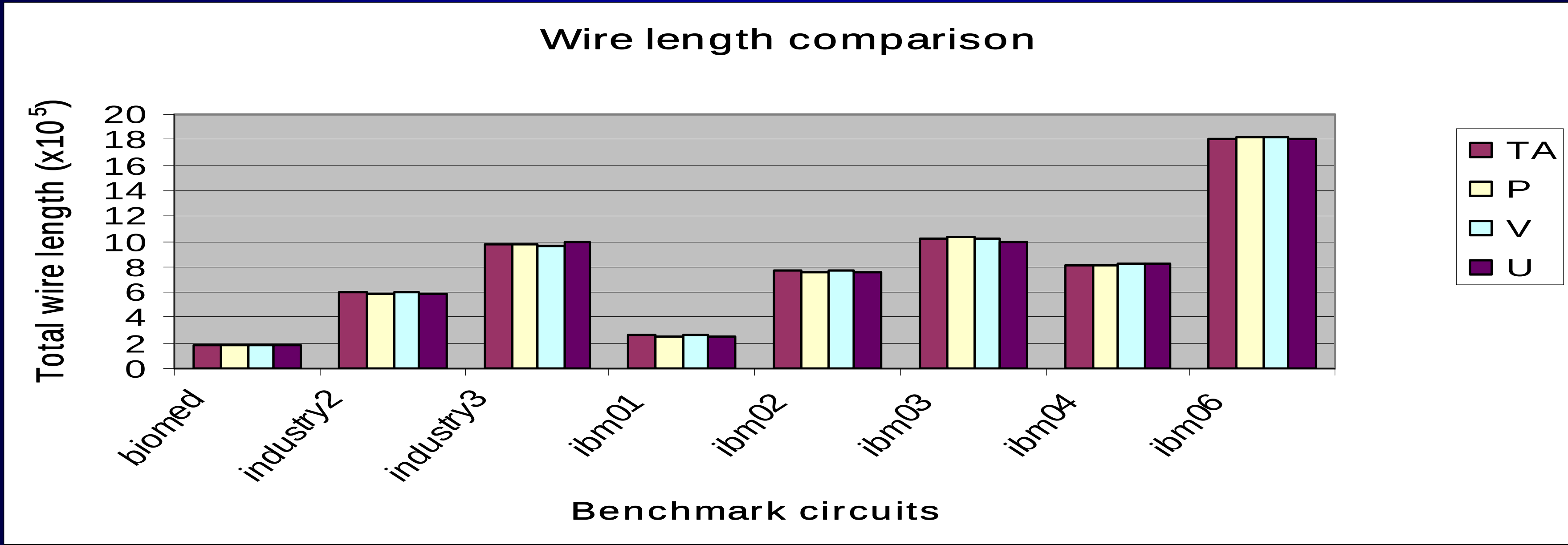
- Our algorithm: temperature-aware 3D global routing (TA) is experimented on 8 benchmarks for 0.18μm technology:
  - Chip size: 5mm x 5mm.
  - Randomly generated power density : $10W/cm^2$ – $800W/cm^2$.
  - Target temperature : 80^oC.
- Comparison algorithm 1: post (P) insertion of thermal vias and thermal wires greedily, after the same initial routing.
- Comparison algorithm 2: same framework as TA, but insertion of only thermal vias (V), no thermal wire insertion.
- Comparison algorithm 3: distribute thermal vias and thermal wires uniformly (U), with the same total amount.

Experimental Results





Experimental Results



Circuit	# routing overflows				Circuit	# of routing overflows			
	TA	P	V	U		TA	P	V	U
biomed	11	0	8	6750	ibm02	34	0	15	4262
industry2	1	0	4	6201	ibm03	11	0	1	16873
industry3	14	0	1	10021	ibm04	2	0	1	5596
ibm01	34	0	15	4262	ibm06	7	0	5	5404

# of overflows comparison

Experimental Results

- Temperature-aware 3D routing (TA) :
  - Effectively reduce temperature by appropriate allocation of thermal via and thermal wire insertion. Almost no peak temperature violation.
  - Effectively resolve congestion violation.
  - Comparable wire length performance with other approaches.
- Post-insertion approach (P):
  - No optimization of resource during routing; max peak temperature difference higher than TA approach: 30.8°C; average peak temperature difference: 22.7°C.
- Thermal via only approach (V) :
  - No thermal wires help lateral heat conduction; max peak temperature difference higher than TA approach: 50.2°C; average peak temperature difference: 32.1°C.
- Uniform distribution approach (U) :
  - No consideration of lateral routing congestion, huge overflow.

Conclusion

- Temperature-aware 3D global routing: effectively reduces peak temperatures to meet design requirements.
- Heat dissipation resources are effectively allocated, compared with other approaches:
  - Post insertion approach, 30.8°C higher peak temperature.
  - Thermal via only approach, 50.2°C higher peak temperature.
- Routing capacity constraints are satisfied with effective allocation of heat dissipation resources based on sensitivity analysis and linear programming.

Thank you!