Thermal-reliable 3D Clock-tree Synthesis Considering Nonlinear Electrical-thermal-coupled TSV Model

Yang Shang[1], Chun Zhang[1], Hao Yu[1], Chuan Seng Tan[1], Xin Zhao[2], Sung Kyu Lim[2]

[1]School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
[2]GTCAD Laboratory, Georgia Institute of Technology, Atlanta. USA

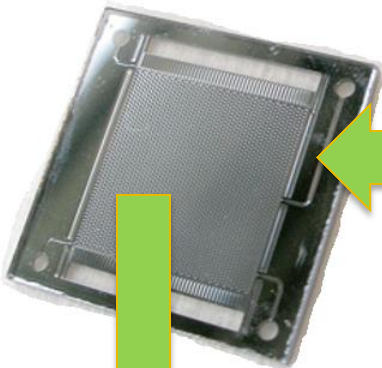
http://www.ntucmosetgp.net

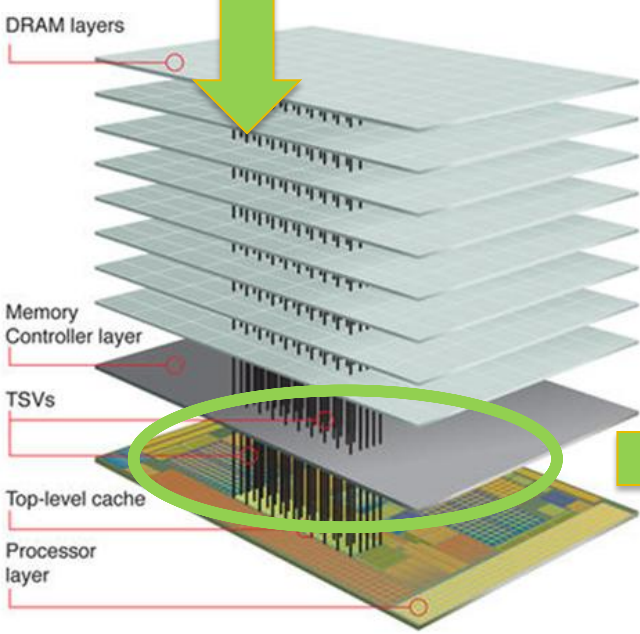
Outline

- Background and Motivation
- Electrical-thermal-coupled TSV Model
- Nonlinear Optimization of Skew Reduction
- Experimental Results
- Conclusion and Future Work

3D Server for Big-data Cloud Server

3D Processor

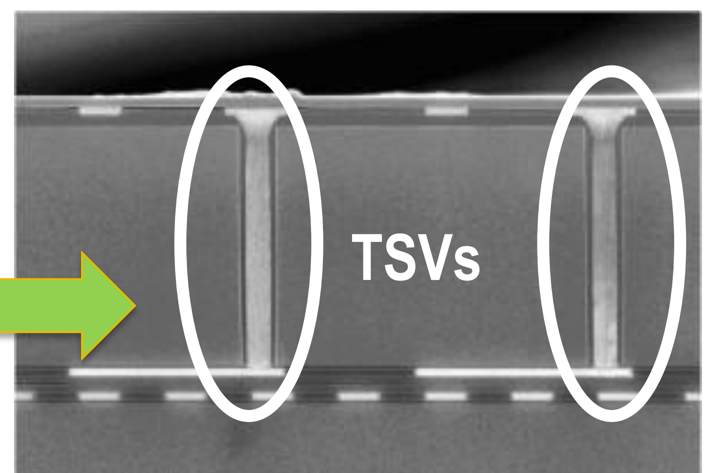




http://www.techradar.com/news/computing-components/3d-processors-memory-and-storage-explained-987509

Supercomputer

http://www.techthefuture.com/energy/climate-change-supercomputer/

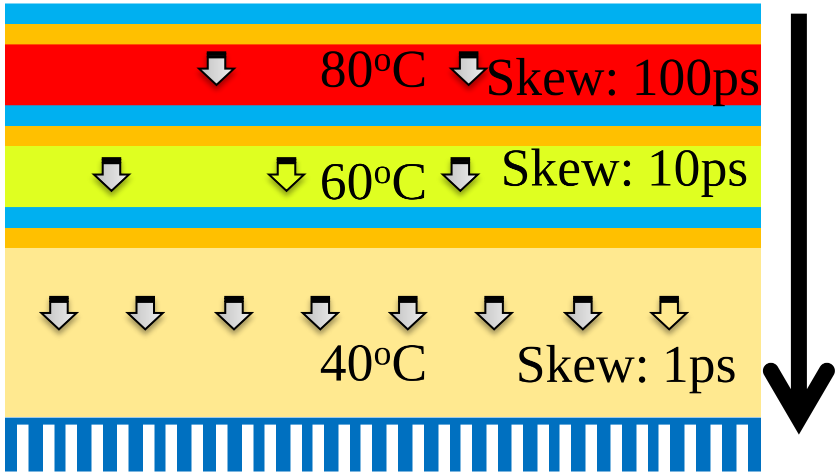


http://javier.esilicon.com/2011/01/30/thru-silicon-vias-current-state-of-the-technology/

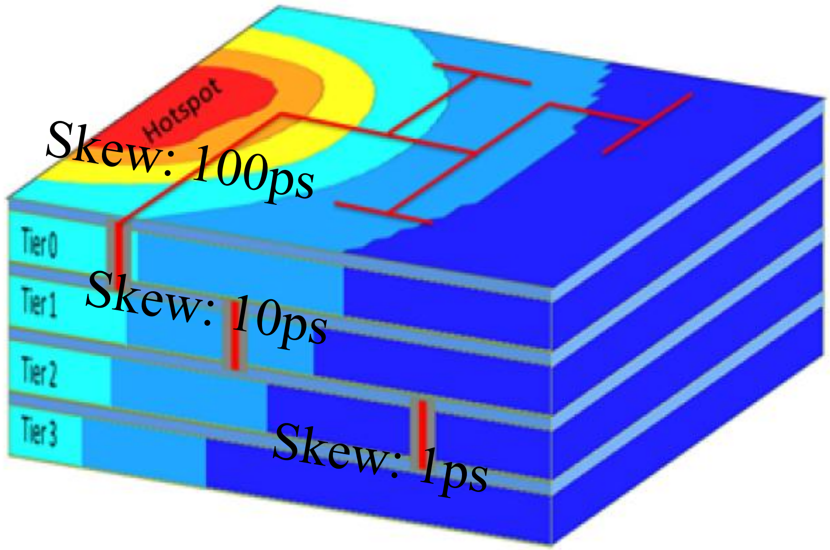
Thermal Challenges in 3D Integration

- **Thermal reliability is concerned:** large non-uniform thermal gradient as limited heat dissipation paths in 3D IC
- **3D clock-tree synthesis to balance electrical-thermal coupling induced skew**

Temperature adds up from longer heat dissipation path

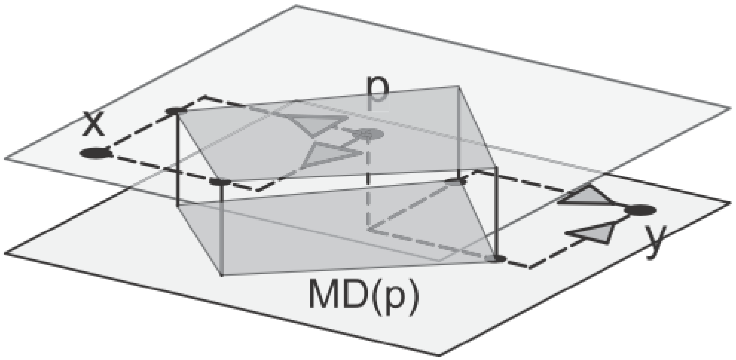


3D CLK network with skew under temperature profile

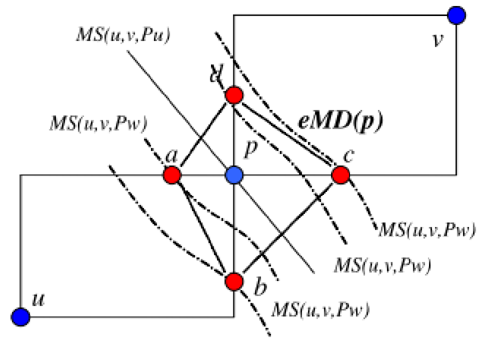


Temperature Aware Clock-tree Synthesis

- **Clock-skew reduction methods considering temperature**
  - Buffer Insertion [1]
  - Merging point adjustment [2]
  - Wire length balancing[3]
- **3D clock-skew reduction by thermal TSV insertion, which requires**
  - Accurate electrical-thermal TSV model
  - Skew balance method



[1] J. Minz, X. Zhao, and S. K. Lim, “Buffered clock tree synthesis for 3d ics under thermal variations,” in IEEE/ACM ASP-DAC, 2008.

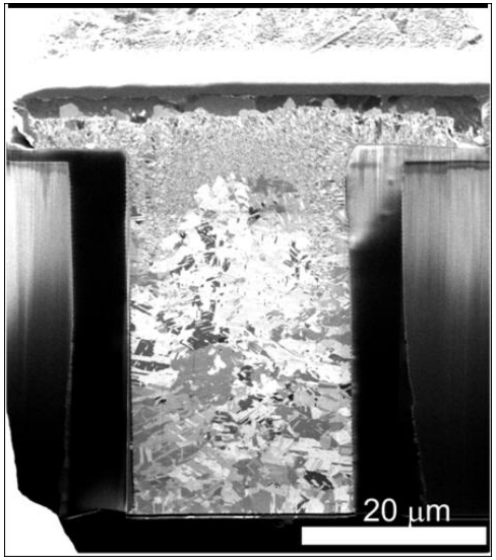


[2]M. Cho, S. Ahmed, and D. Pan, “TACO: Temperature aware clock-tree optimization,” in IEEE/ACM ICCAD, 2005.

[3]J. Cong, A. Kahng, C. Koh, and C. A. Tsao, “Bounded-skew clock and steiner routing,” ACM Trans. on Design Automation of Electronic Systems, vol. 3, no. 3, pp. 341–388, 1998.

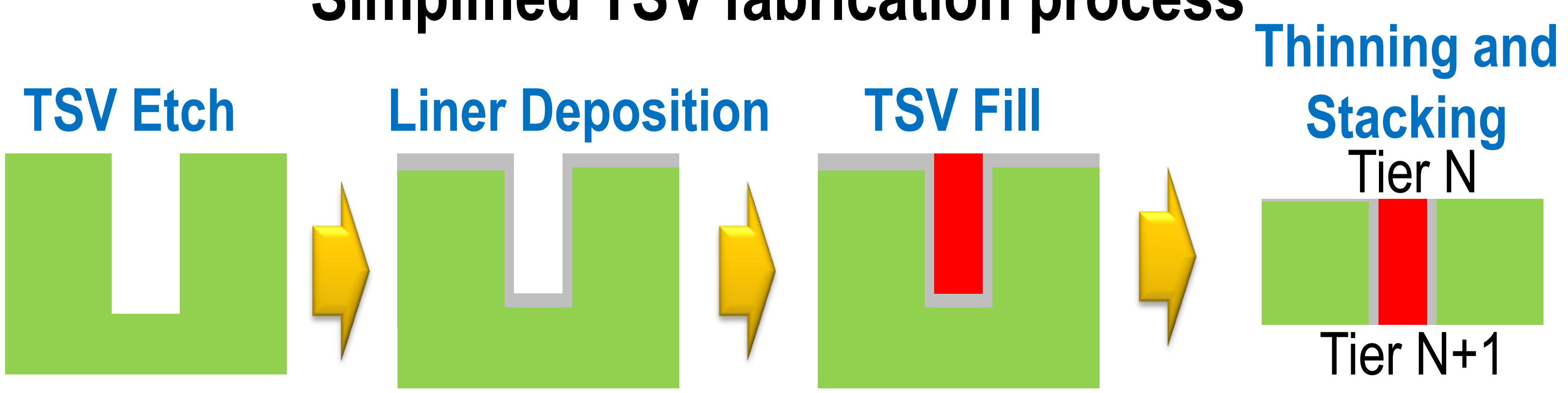
TSV Fabrication Technology

- TSV material: Al/Cu, under-bump-metal
- TSV diameter: 1~20um
- TSV height: 20~50um
- Liner material: SiO_2 or Si_3N_4
- Liner thickness: 0.2~0.5um



http://www.imec.be/ScientificReport/SR2009/HTML/1213308.html

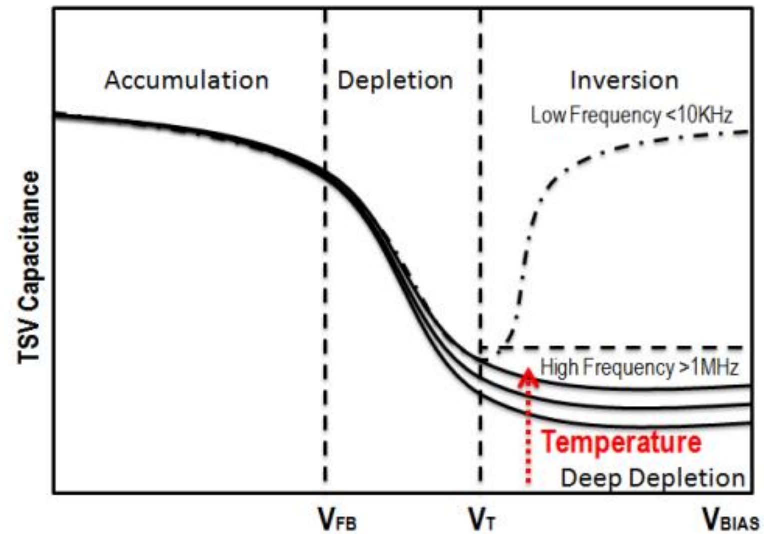
Simplified TSV fabrication process

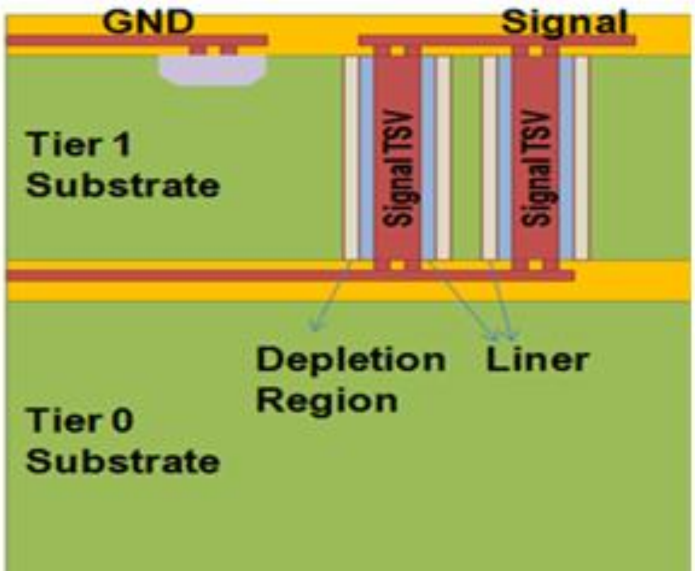


Signal TSVs

- **Signal TSVs provide electrical connection between adjacent tiers**
- **Non-linear MOS-capacitance (MOSCAP) is formed between signal TSV and substrate due to the existence of liner**
  - Low-k liner is preferred (SiO_2)
  - Depletion region exists because of the work function difference between TSV metal and silicon substrate
  - The radius of depletion region is temperature and voltage dependent

C-V curve of MOSCAP





Signal TSV Modeling

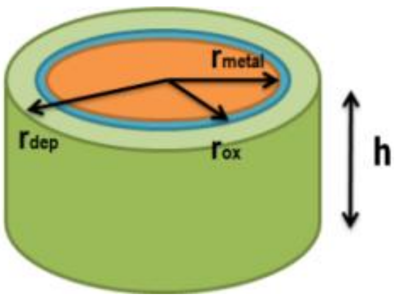
- **RC equivalent circuit of signal TSVs**

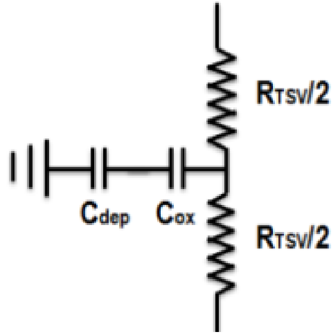
$$\begin{cases} \frac{1}{C_T} = \frac{1}{C_{ox}} + \frac{1}{C_{dep}}; R_T = \frac{\rho h}{\pi r_{metal}^2} \\ C_{ox} = \frac{2\pi\varepsilon_{ox} h}{ln(\frac{r_{ox}}{r_{metal}})} \quad C_{dep} = \frac{2\pi\varepsilon_{si} h}{ln(\frac{r_{dep}}{r_{ox}})} \end{cases}$$

r_{ox} is the outer radius of liner,

r_{metal} is the radius of TSV,

r_{dep} is the outer radius of depletion region

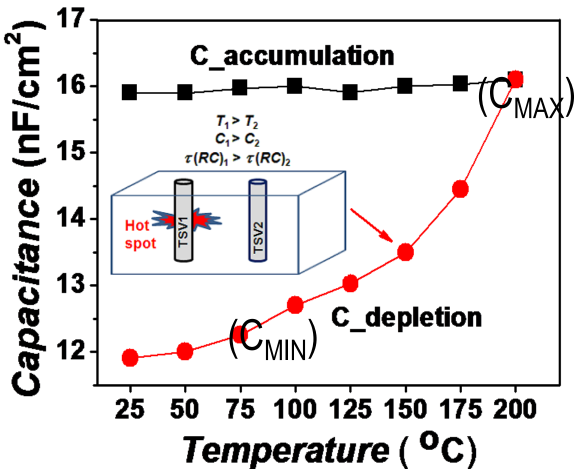




- **TSV model considering non-linear temperature effect**

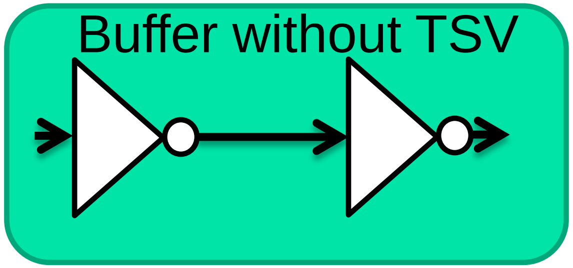
$$\begin{cases} R_T = R_0(1 + \alpha(T - T_0)) \\ C_T = C_0 + \beta_1 T + \beta_2 T^2 \end{cases}$$

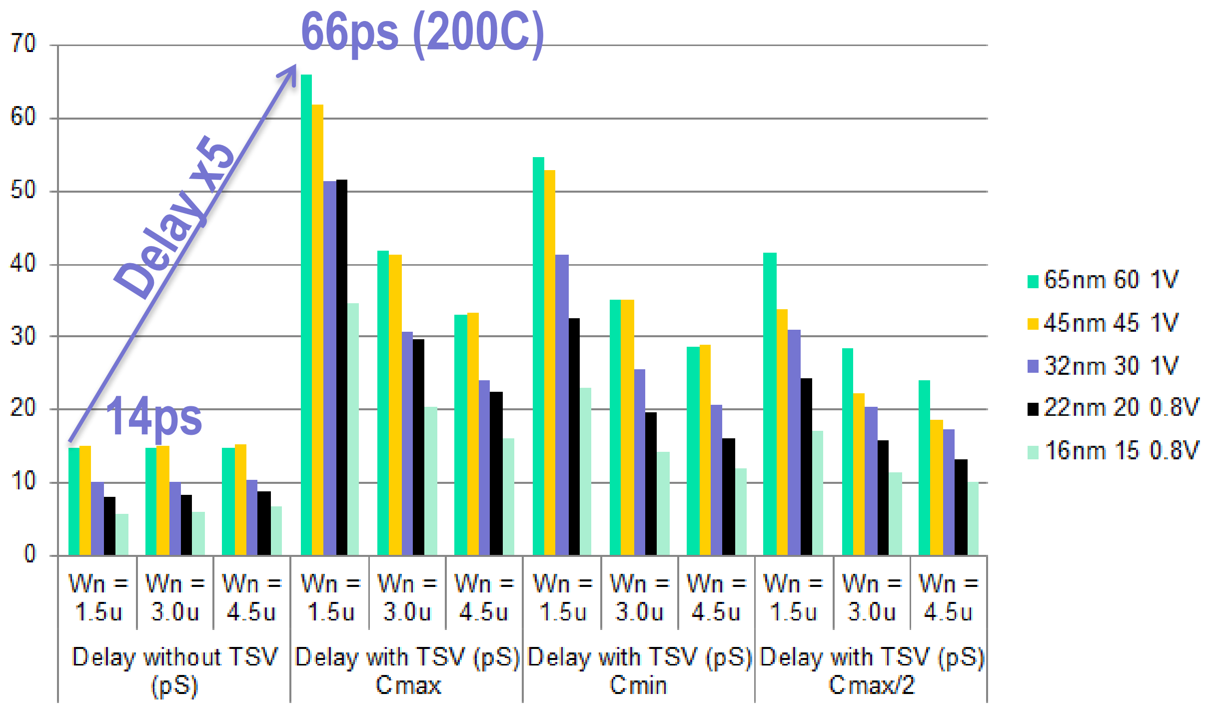
α and β_1, β_2 are the temperature coefficients

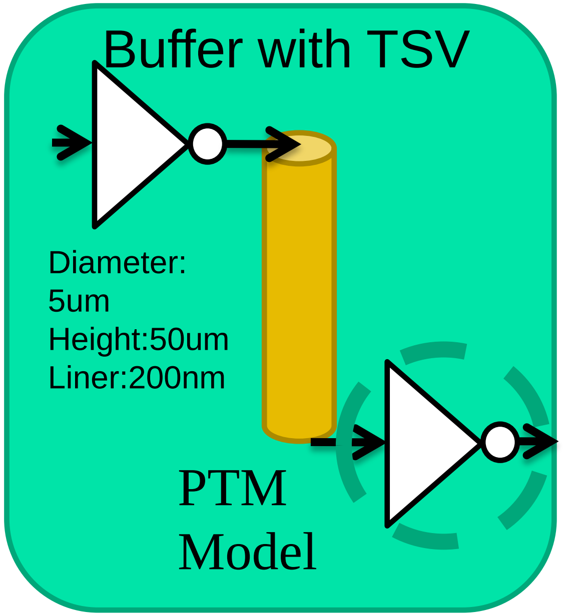


Signal TSV Delay Modeling

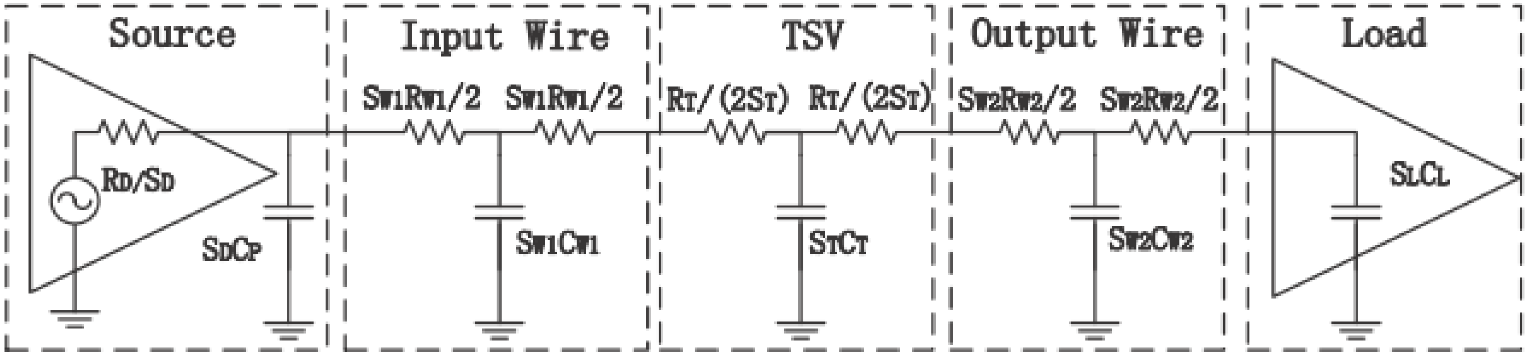
- **Signal TSVs introduce noticed delay**
  - Process and capacitance dependent
- **Non-linear function with temperature**







Nonlinear Electrical-thermal-coupled Signal TSV Delay Model



- **Scalable signal TSV delay model**

$$\tau = R_{in}\alpha\beta_2 T^3 + R_{in}[(1-\alpha T_0)\beta_2 + \alpha\beta_1]T^2$$
$$+[\alpha(\tau_0 + R_{in}C_0) + (1-\alpha T_0)R_{in}\beta_1]T$$
$$+(1-\alpha T_0)(R_{in}C_0 + \tau_0)$$

where R_{in} is the total resistance looking from C_T to the input and τ_0 is the delay of circuit without C_T

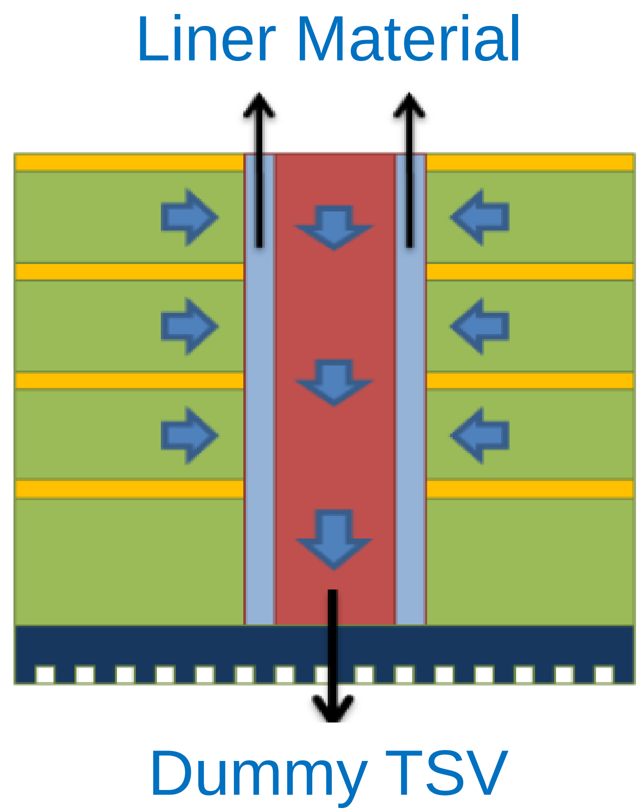
- **Non-linear function of in delay model with temperature**

$$D_{TSV} = k_0 + k_1 T + k_2 T^2 + k_3 T^3$$

where k0, k1, k2, k3 are the temperature coefficient in different order

Dummy TSVs

- **Dummy TSVs provide additional heat dissipation paths to heat-sink as TSV metal (Cu, 400W/m · K) has much larger thermal conductivity than liner (SiO_2, 1.2W/m · K)**
  - Chip temperature can be reduced by adding thermal TSVs
  - High thermal conductivity liner material is preferred (Si_3N_4, 30W/m · K)

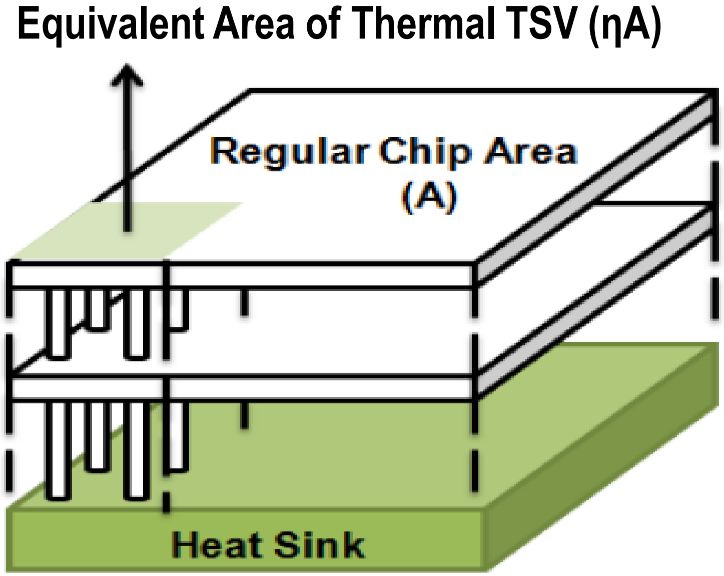


Dummy TSV Model

- **Dummy TSV density (η) is the ratio equivalent area of thermal TSV over total chip area(A)**
- **Total thermal conductivity between chip and heat sink**

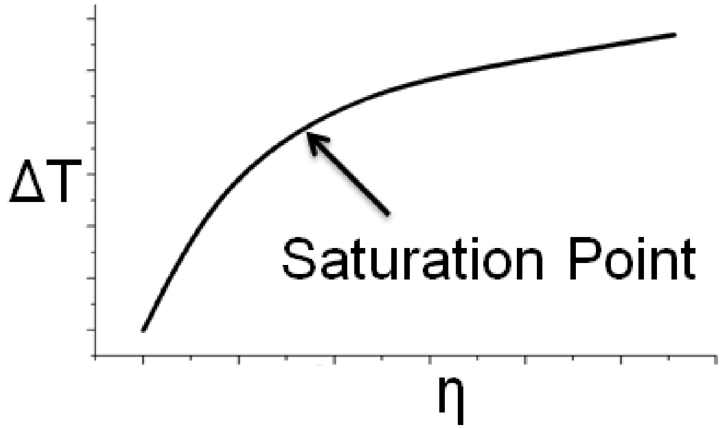
$$\sigma_{Total} = \eta \cdot \sigma_{TSV} + (1-\eta)\sigma_0$$

σ_{TSV} and σ_0 are the chip thermal conductivity with and without TSV, respectively



Thermal Gradient Reduction by Dummy TSV Insertion

- **Temperature reduction saturation effect due to dummy TSV insertion**



$$\Delta T = T_0 - T_{TSV} = \frac{P \cdot l}{A\sigma_0} \cdot \frac{\eta}{\frac{\sigma_0}{\sigma_{TSV}-\sigma_0} + \eta}$$

  - P/A is the power density,
  - l is the equivalent length of thermal dissipa
  - ΔT increases linearly with η as η is smaller than $\sigma_0 /(\sigma_{TSV} - \sigma_0)$,
  - ΔT is less sensitive to η when η is approaching or larger than $\sigma_0 /(\sigma_{TSV} - \sigma_0)$,
- **η is limited by the chip area overhead and saturation of temperature reduction effect**
- **Balanced thermal gradient will balance clock skew**

Nonlinear OPT of 3D CLK Skew by Dummy TSV Insertion: Problem Formulation I

- **Objective function**: minimize the variance of skew in all clock branches

$$min: f(\mathbf{D}) = \frac{1}{C-1}\sum_{k=1}^{C}(D_k - \overline{D})^2$$

- **Parameters**: dummy TSV insertion density

$$\begin{aligned} min: f(\mathbf{x}) &= \frac{1}{C-1}\sum_{k=1}^{C}(\hat{c}_k^2 + 2\hat{c}_k\hat{\mathbf{f}}_k^T\mathbf{x} \\ &+ \mathbf{x}^T(\hat{\mathbf{f}}_k\hat{\mathbf{f}}_k^T + \hat{c}_k\hat{H}_k)\mathbf{x} \\ &+ \hat{\mathbf{f}}_k^T\mathbf{x}^T\mathbf{x}\hat{H}_k\mathbf{x} + \frac{1}{4}\mathbf{x}^T\hat{H}_k\mathbf{x}\mathbf{x}^T\hat{H}_k\mathbf{x}) \end{aligned}$$

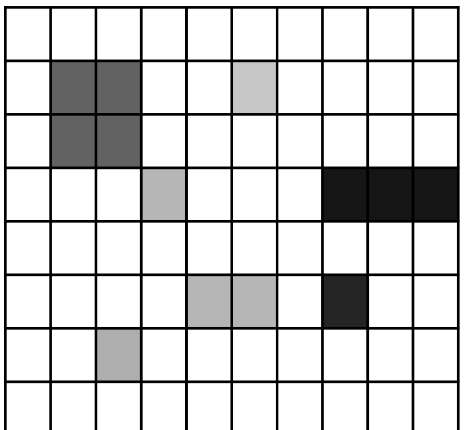
- **Constraints**:

$$\mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub}$$

  - lb is the min thermal TSV density ddetermined by the foundry process limitation
  - ub is the max thermal TSV density determined by temperature reduction sensitivity function as well as the maximum allowed chip overhead

Nonlinear OPT of 3D CLK Skew by Dummy TSV Insertion: Problem Formulation II

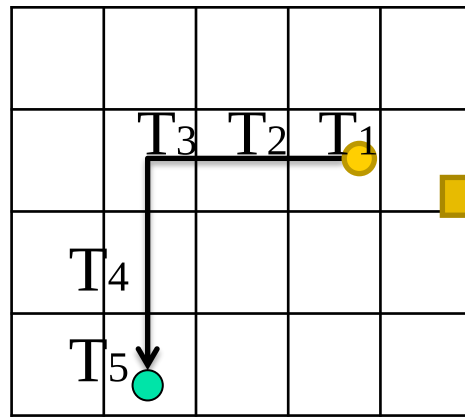
Chip Gridding (M x N)



Clock Tree Benchmarks

Temperature Profile from Hotspot

Benchmark Circuit RC Extraction



Parent Node

Child Node

Nonlinear Delay (Dk)

Each grid is available for the thermal TSV insertion with density of x_i, i=1є[1,MxN]

$$D_k = \sum_{i \in C_k} \tau_i .$$

is the summation of delay of passing through all grids

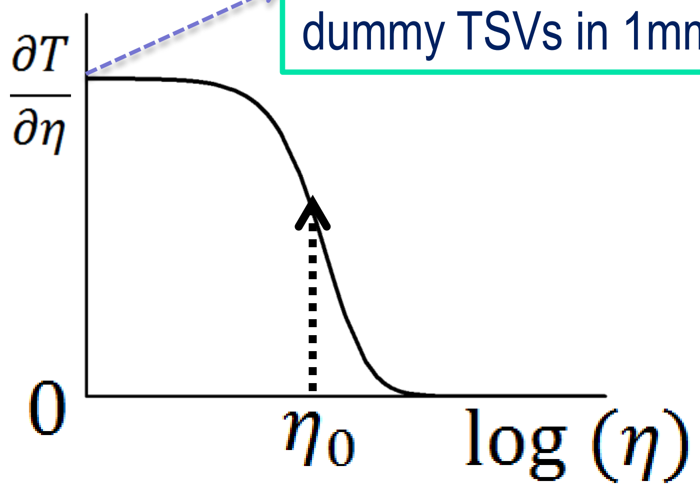
Dummy TSV Insertion Sensitivity

- **Sensitivity of temperature reduction w.r.t. dummy TSV density**

$$\frac{\partial T}{\partial \eta} = \frac{\Delta T_0 \cdot \eta_0}{(\eta_0 + \eta)^2}$$

  - $\Delta T_0 = \frac{P \cdot l}{A \cdot \sigma_0}$ is relative chip temperature without dummy TSV insertion,
  - $\eta_0 = \frac{\sigma_0}{\sigma_{TSV} - \sigma_0}$ is threshold density of dummy TSV.
  - Constant sensitivity is achieved when $\eta \ll \eta_0$.
  - Sensitivity is approaching zero when $\eta \gg \eta_0$

Typical: 2~10 K for 100 dummy TSVs in $1mm^2$



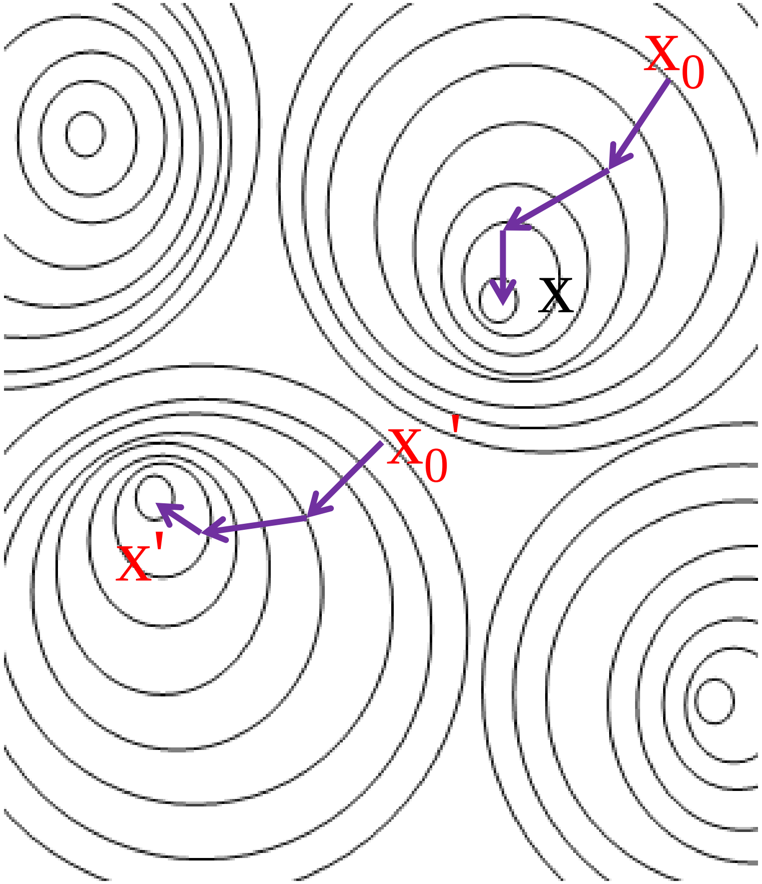
Nonlinear OPT of 3D CLK Skew by Dummy TSV Insertion: Conjugate-gradient Solution

- Original problem is relaxed with Lagrange penalty factor to remove the inequality constraint

$$min: f^*(\mathbf{x}) = f(\mathbf{x}) + \lambda \cdot h^2(\mathbf{x})$$

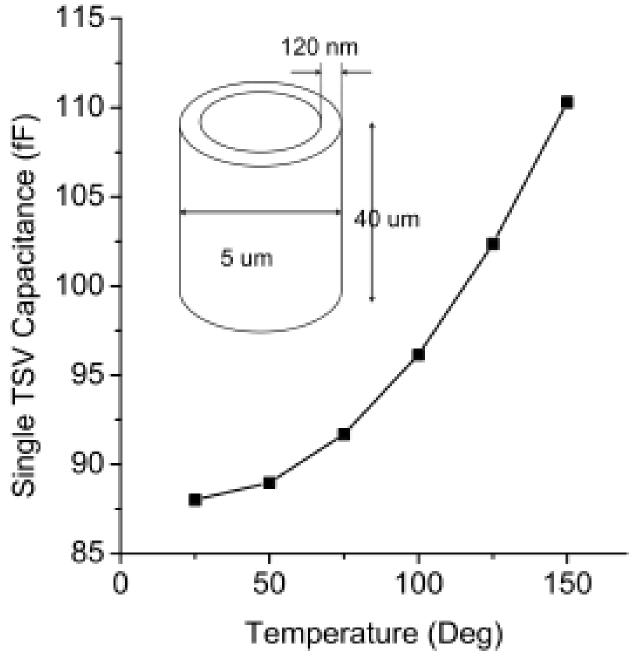
where $h(\mathbf{x}) = \begin{cases} 0, & \mathbf{lb} \le \mathbf{x} \le \mathbf{ub} \\ \rho \gg 0, & \text{otherwise} \end{cases}$

- Conjugate gradient method iteratively searches along the gradient drop reduction to find the x which minimizes $f^*(x)$.
- Problem is solved multiple times with properly selected x_0 to avoid local minimum

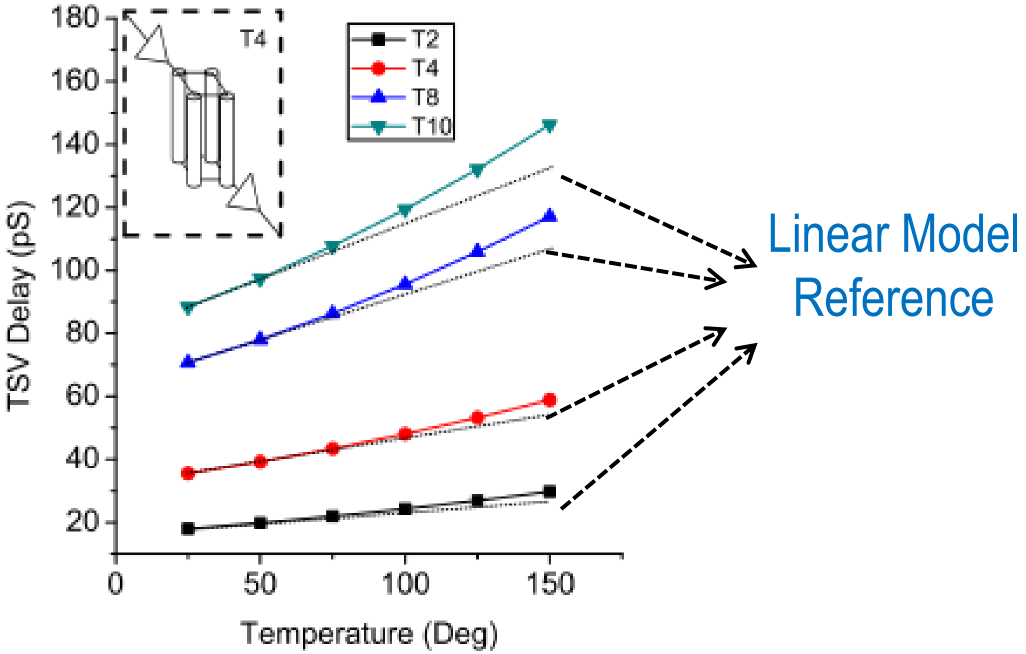


Nonlinear Electrical-thermal Coupling by Signal TSVs

Nonlinear temperature-dependent TSV capacitance



Nonlinear temperature-dependent TSV delay

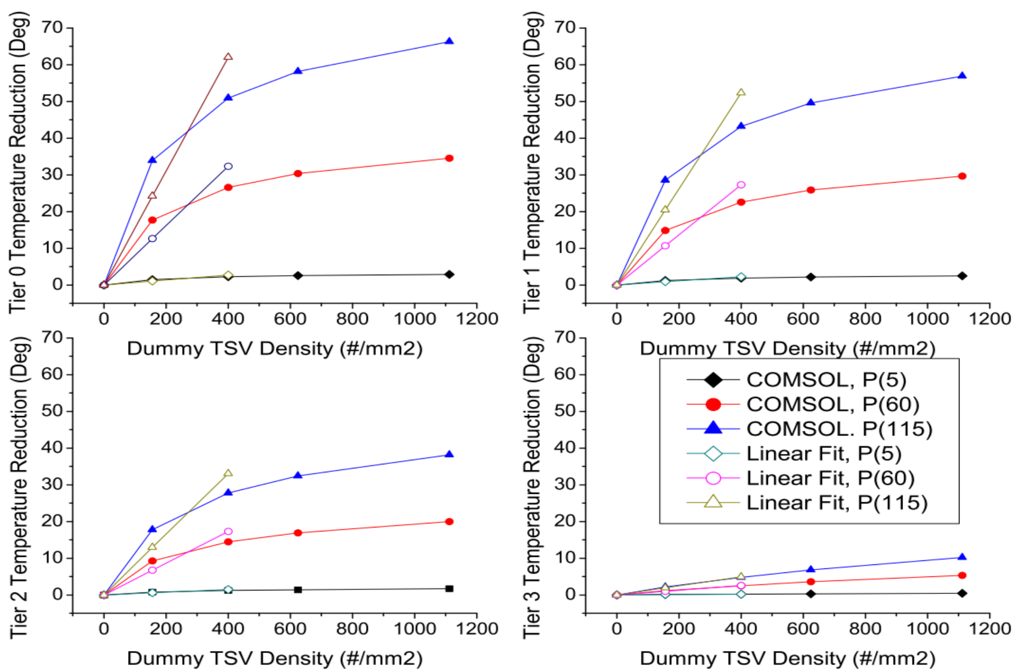


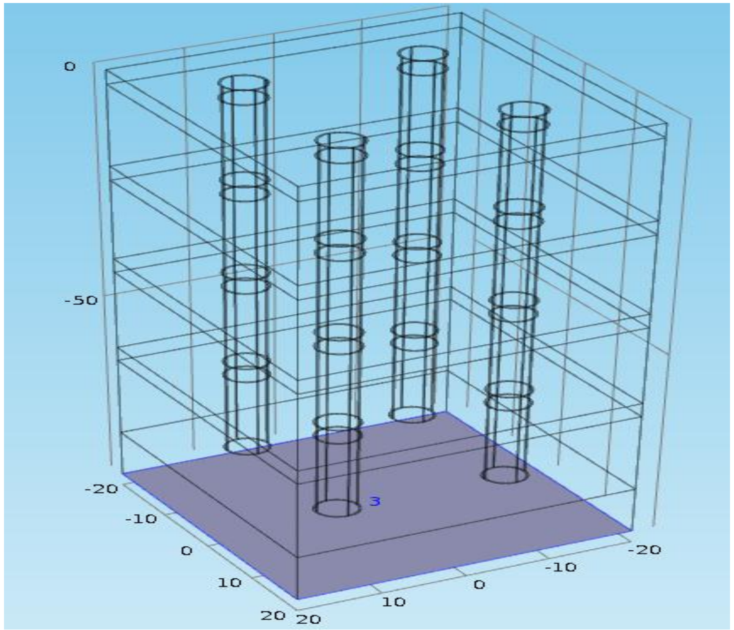
- Measurement result [4] shows non-linear TSV capacitance
- Delay introduced by non-linear TSV model shows large difference from linear model at high temperature

[4] G. Katti and et al., "Temperature dependent electrical characteristics of through-si-via (tsv) interconnections," in IITC, 2010.

Temperature Reduction by Dummy TSVs

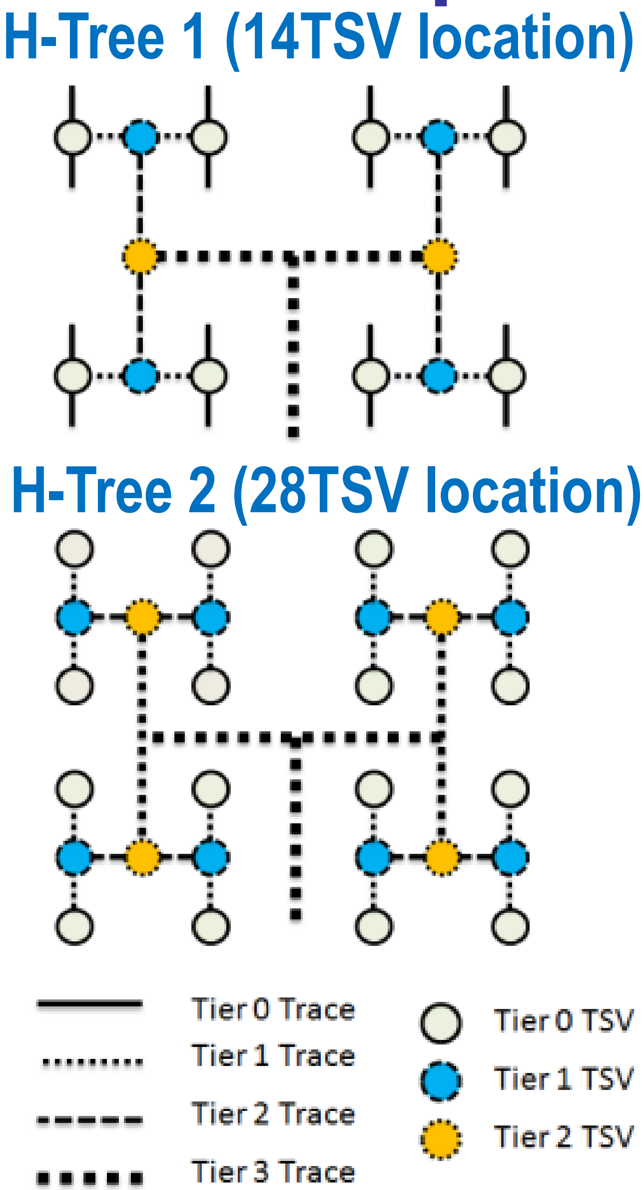
- COMSOL multi-physics simulator for thermal analysis
- A 4-tier 3D-IC test case : Tier thickness: 40μm; Thermal TSV has a diameter of 15μm; Heat sink conductance: 1.24×10^5 W/(K · m^2)
- Chip temperature is linearly reduced by increasing thermal TSV density





Dummy TSV Insertion for 3D Clock Skew Reduction: Set-up I

- Two generalized 4-Tier 3D IC H-Trees designed
- IBM clock-tree benchmarks r1-r5 [5] synthesized to 4-tier 3D clock-tree [6]
- 3D-IC chip divided into 64x64 grids for dummyTSV insertion

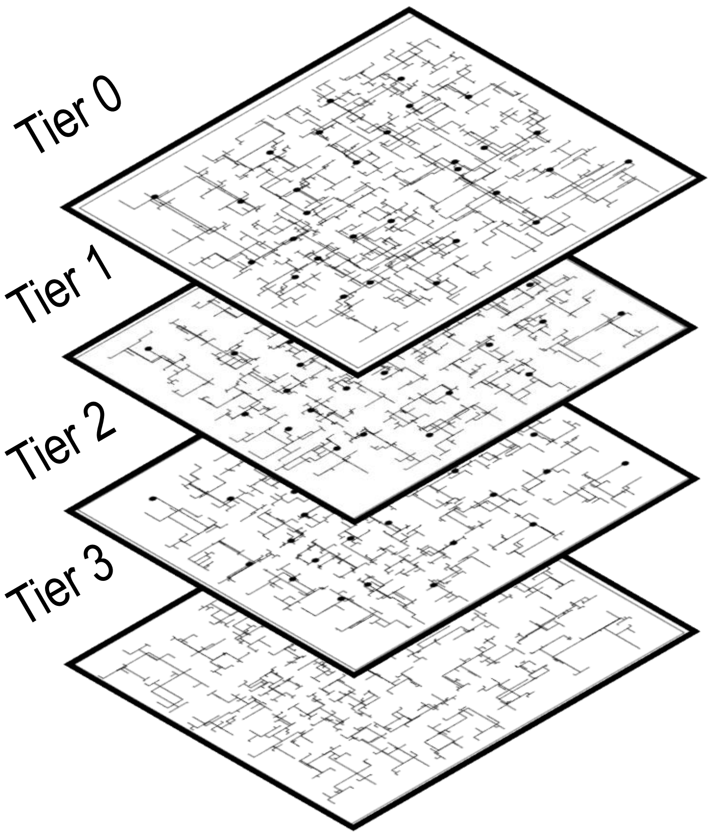


[5] "Ibm clock tree benchmarks," http://vlsicad.ucsd.edu/GSRC/bookshelf/Slots/BST/.
[6] X. Zhao, J. Minz, and S. K. Lim, "Low-power and reliable clock network design for through-silicon via (tsv) based 3d ics," IEEE Trans. on Components, Packaging, and Manufacturing Technology, vol. 1, no. 2, pp. 247 –259, feb 2011.

Dummy TSV Insertion for 3D Clock Skew Reduction: Set-up II

- HotSpot [7] is used to extract the temperature distribution at each location
- Temperature distribution is calculated from the average input of all SPEC2000 benchmarks [8]
- Maximal thermal TSV density is limited to be lower than 7% of the local grid area.
- Different signal TSV-bundles T2,T4, T8 and T10 are deployed with number of 2, 4, 8 and 10 TSVs

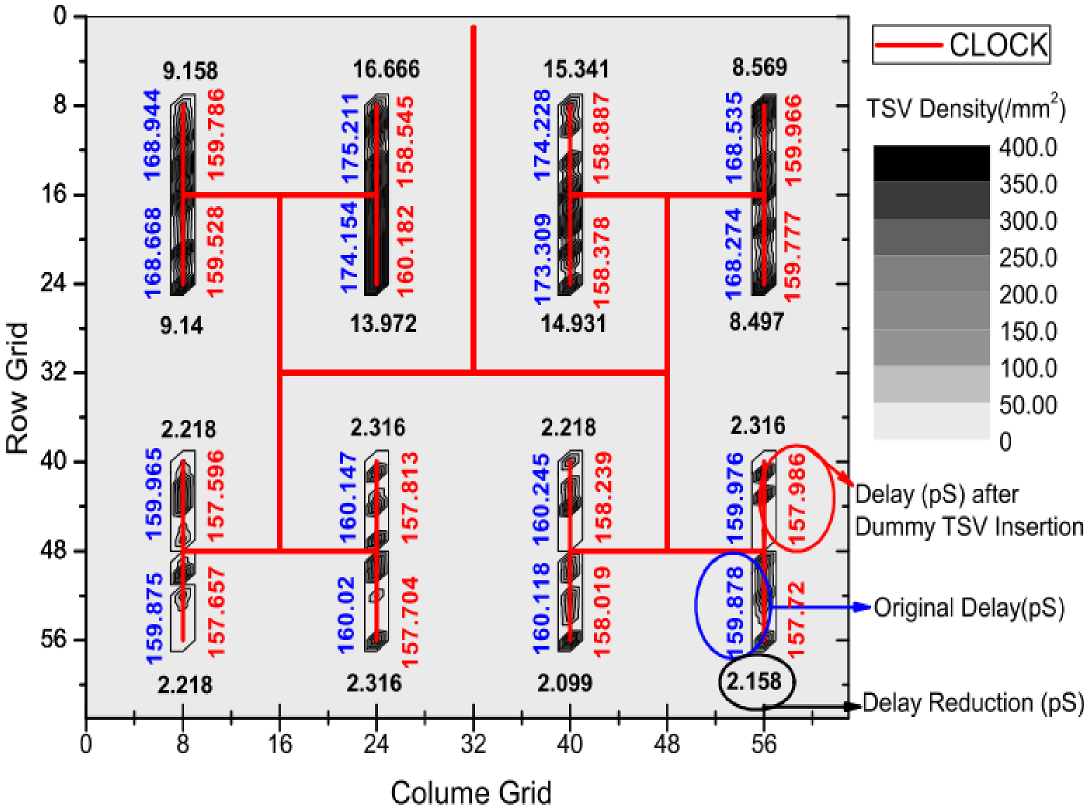
3D Clock Tree of r5



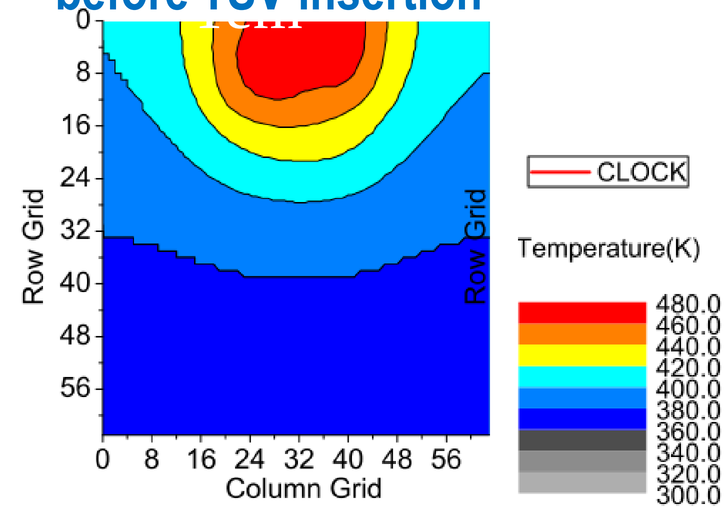
[7] "Hotspot: http://lava.cs.virginia.edu/hotspot/."
[8] http://www.spec.org/cpu2000/

Dummy TSV Insertion for 3D Clock Skew Reduction: Result I

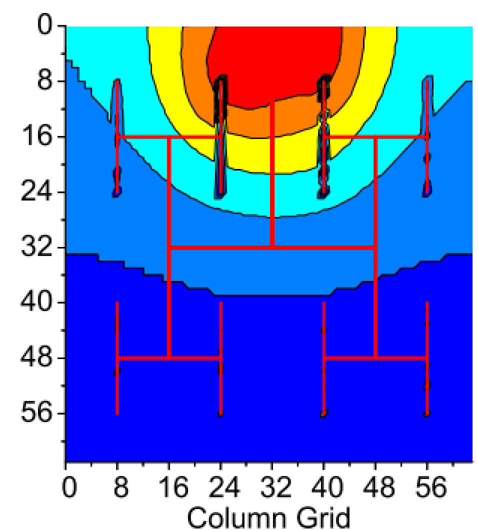
H-Tree 1 clock skew comparison before and after TSV insertion



Temperature profile before TSV insertion

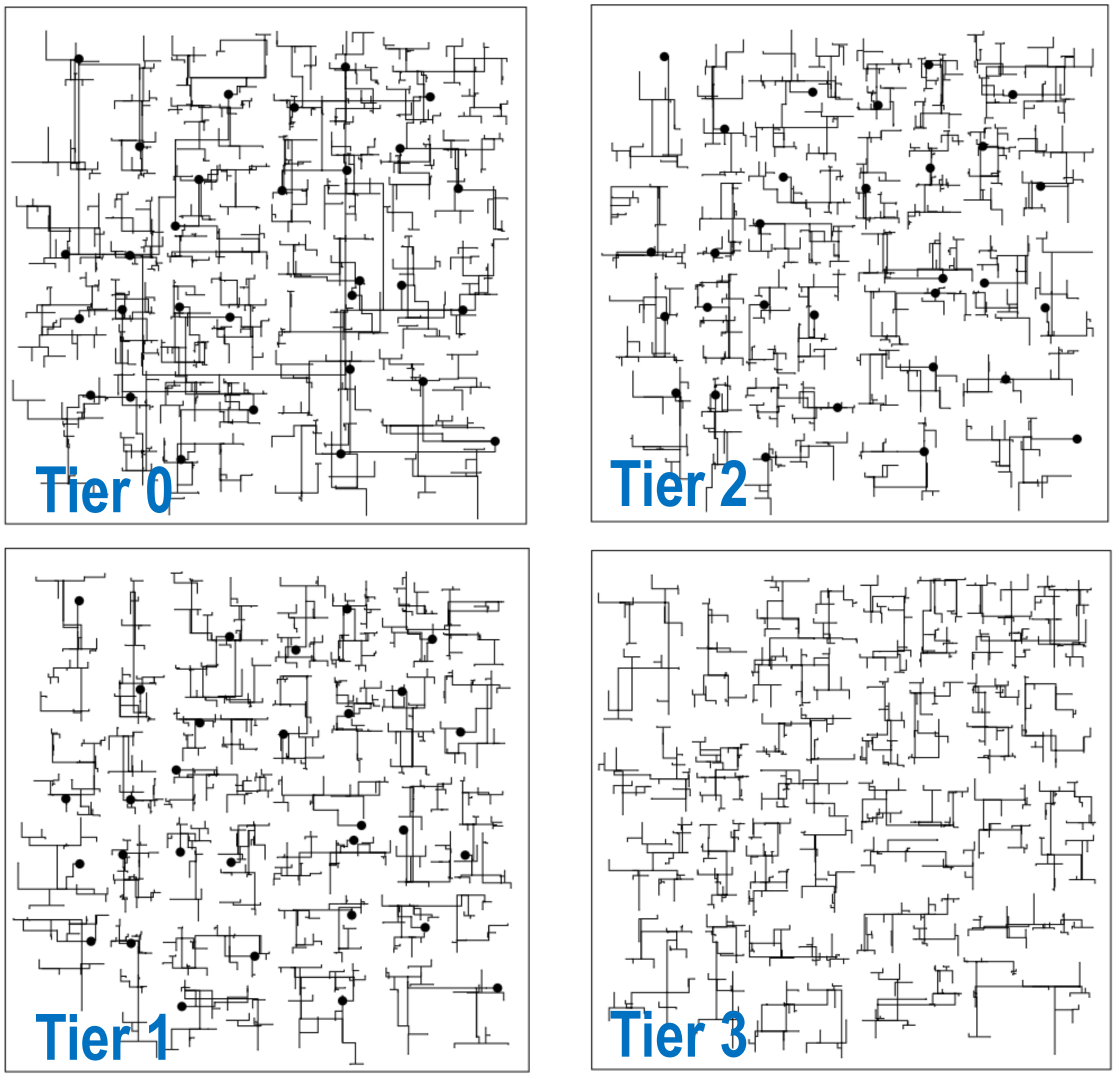


After TSV insertion



Dummy TSV Insertion for 3D Clock Skew Reduction: Result II

Synthesized 3D Clock Tree of Benchmark r5



Dummy TSV Insertion for 3D Clock Skew Reduction: Result III

htree1	(14 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	15.34	10.02	34.70%	14.29	2.59	83.10%	57.95
T4	26.44	8.67	67.20%	14.19	4.48	83.10%	57.9
T8	47.42	12.1	74.50%	14.58	8.14	82.80%	58.81
T10	58.42	15.1	74.20%	15.35	10.19	82.60%	58.84
Mean	-	-	62.60%	14.6	-	82.90%	58.38
htree2	(28 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	23.48	8.69	63.00%	13.98	3.57	84.80%	56.95
T4	43.97	12.4	71.80%	14.02	5.38	87.80%	57.12
T8	82.76	16.18	80.40%	13.92	9.35	88.70%	58.87
T10	103.1	17.69	82.80%	13.93	11.44	88.90%	57.58
Mean	-	-	74.50%	13.96	-	87.50%	57.63
r1	(45 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	30.5	18.4	39.70%	41.5	15.34	49.70%	106.6
T4	61.87	36.29	41.30%	29.3	27.5	55.60%	158.4
T8	121.1	71.39	41.60%	31.9	57.1	52.80%	170.5
T10	152.7	91.1	40.30%	37.2	74.26	51.40%	108.1
Mean	-	-	40.70%	35	-	52.40%	135.9
r2	(60 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	35.13	23.6	32.80%	134	20.2	42.50%	389
T4	69.75	48.31	30.70%	102.8	37.9	45.70%	393.8
T8	134.7	94.67	29.70%	106.5	74	45.10%	705.8
T10	169	119.3	29.40%	139.3	93.82	44.50%	325.6
Mean	-	-	30.70%	120.7	-	44.50%	453.6

r3	(75 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	32.36	20.92	38.40%	220.7	19.5	39.70%	749.5
T4	64.8	41.02	36.70%	170.8	34.3	47.10%	451
T8	125.6	80.29	36.10%	177.8	66.9	46.70%	745.4
T10	157.7	100.7	36.20%	231	85.8	45.60%	436.7
Mean	-	-	36.90%	200.1	-	44.80%	595.7
r4	(90 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	31.68	18.63	41.20%	211.8	17.63	44.00%	890.6
T4	64.57	38.3	40.70%	232.2	30.1	53.40%	557.8
T8	126.8	75.38	40.60%	233.4	69.8	45.00%	707.5
T10	159.8	93.1	41.70%	327.6	80	50.10%	564.7
Mean	-	-	41.10%	251.2	-	48.10%	680.2
r5	(90 Signal TSV locations)						
Type	Orig	Lin	Impr%	Time(s)	Nonlin	Impr%	Time(s)
T2	35	21.5	38.60%	665.6	19.9	42.60%	1963
T4	68.4	39	43.00%	695	33.62	50.90%	1716
T8	131	77.9	40.50%	725.2	65.2	50.20%	1694
T10	164.1	97	40.90%	938.5	80.13	51.20%	1750
Mean	-	-	40.80%	756.1	-	48.70%	1781
Overall			**46.8%**			**58.4%**	

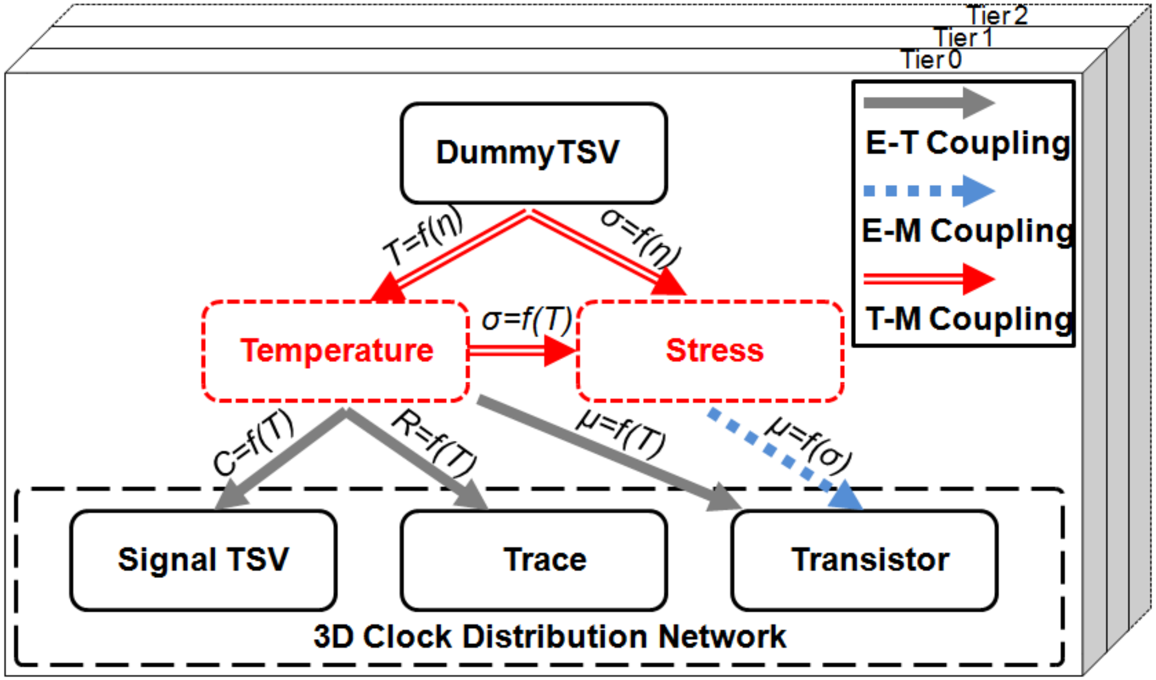
- **58.4% clock-skew reduction**
- **11.6% higher clock-skew from non-linear model than linear model**

Conclusions

- Physics-based electrical-thermal models for both signal and (dummy) thermal TSVs are provided with the consideration of nonlinear temperature dependence: liner is important to form a nonlinear CAP for delay
- One nonlinear programming problem is formulated to reduce clock-skew by dummy TSVs insertion for the thermal-reliable 3D clock-tree synthesis
- Under realistic nonlinear TSV models, insertion of dummy TSV can effectively reduce the clock-skew by 58.4% on average, which is also 11.6% higher clock-skew reduction on average than using the linear model

Electrical-Thermal-Stress-Coupling to 3D Clock

- Stress needs to be considered in 3D clock distribution network of surrounding transistors
- Insertion of dummy TSVs reduce the chip temperature and improve the transistor carrier mobility



Experiment Results with Electrical-Thermal-Stress-Coupling

Type	Htree1 (14 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr	Time
T2	15.39	9.54	38.01%	16.98	2.45	84.08%	76.15
T4	26.62	8.07	69.68%	16.85	3.92	85.27%	77.11
T8	47.41	11.38	76.00%	17.05	7.03	85.17%	77.19
T10	58.64	14.8	74.76%	17.22	10.02	82.91%	77.36
			64.61%	17.025		84.36%	76.9525

Type	Htree2 (28 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr%	Time
T2	23.55	8.15	65.39%	17.41	3.31	85.94%	79.02
T4	44.13	11.61	73.69%	17.51	4.81	89.10%	79.59
T8	82	14.6	82.20%	17.4	9	89.02%	79.98
T10	103.7	15.5	85.05%	17.49	9.51	90.83%	79.32
			76.58%	17.4525		88.72%	79.4775

Type	r1 (45 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr%	Time
T2	30.5	17.4796	42.69	41.5	14.4612	52.59	127.2951
T4	61.87	32.6635	47.21	29.3	25.5463	58.71	164.0101
T8	121.1	66.137	45.39	31.9	53.1844	56.08	182.709
T10	152.7	83.557	45.28	37.2	69.1715	54.70	197.4859
Mean	-	-	45.14	34.975	-	55.52	167.875

Type	r2 (60 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr%	Time
T2	35.13	23.481	33.16	134	19.3329	44.97	475.3568
T4	69.75	46.2756	33.66	102.8	34.1747	51.00	485.1885
T8	134.7	86.8894	35.49	106.5	71.0108	47.28	490.5759
T10	169	118.8894	29.65	139.3	88.0831	47.88	513.4299
Mean	-	-	32.99	120.65	-	47.78	491.1378

Type	r3 (75 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr%	Time
T2	32.36	20.1915	37.60	220.7	19.2129	40.63	751.53
T4	64.8	39.969	38.32	170.8	31.1747	51.89	682.8994
T8	125.6	79.8619	36.42	177.8	60.707	51.67	658.8006
T10	157.7	95.4677	39.46	231	82.7985	47.50	708.3861
Mean	-	-	37.95	200.075	-	47.92	700.404

Type	r4 (90 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr%	Time
T2	31.68	17.9695	43.28	211.8	16.5271	47.83	912.3927
T4	64.57	36.2087	43.92	232.2	26.1332	59.53	844.4599
T8	126.8	70.6158	44.31	233.4	64.4995	49.13	898.2112
T10	159.8	88.811	44.42	327.6	76.8253	51.92	909.2504
Mean	-	-	43.98	251.25	-	52.10	891.0786

Type	r5 (90 Signal TSVs)						
	orig	Lin	Impr%	Time	Non Lin	Impr%	Time
T2	35	21	40.00	665.6	18.9	46.00	1992.597
T4	68.4	38.4	43.86	695	28.5	58.33	1843.766
T8	131	74.6	43.05	725.2	63.8	51.30	1705.817
T10	164.1	93.4	43.08	938.5	79.434	51.59	1934.662
Mean	-	-	42.50	756.075	-	51.81	1869.21
Overall	-	-	49.10%		-	61.30%	

- **61.3% clock-skew reduction**
- **12.2% higher clock-skew from non-linear model than linear model**

Thank You!

Please send comments to haoyu@ntu.edu.sg

http://www.ntucmosetgp.net