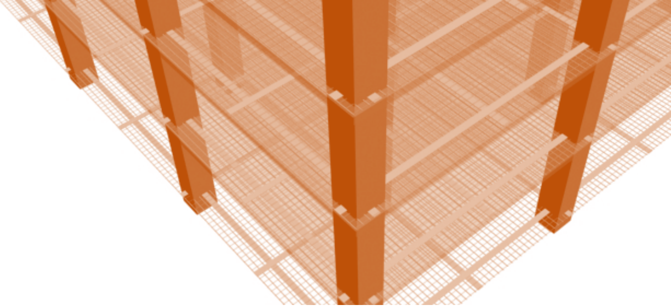
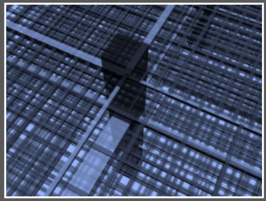
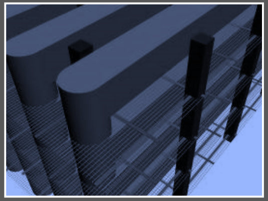
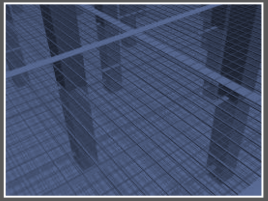
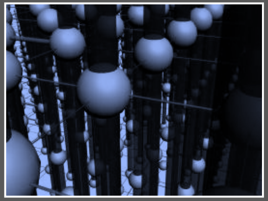


ML-Based Wire RC Prediction in Monolithic 3D ICs

Sai Surya Kiran Pentapati[1], Bon Woong Ku[2], and Sung Kyu Lim[1]

[1] Georgia Tech Computer Aided Design Lab

[2] Synopsys Inc

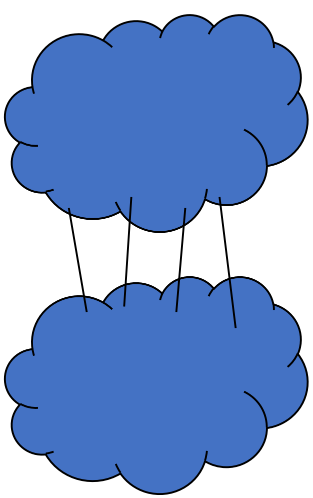
Overview

- **Motivation**
  - Various state-of-the-art flows for Monolithic 3D IC design
- **RC analysis**
  - How RCs change from a pseudo-3D (a 2D-like representation of 3D IC) stage to final 3D stage
  - Processes impacting 3D RC values
- **ML model for RC prediction in 3D ICs**
  - Implementation of the model
  - Feature space analysis
- **PPA impact of using predictive pseudo-3D RCs**

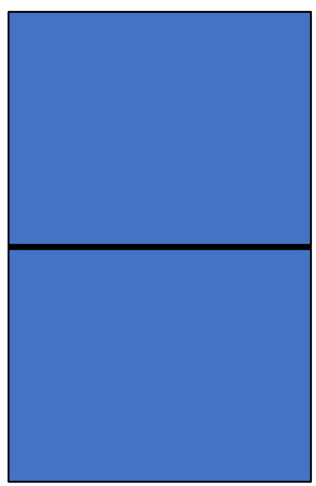
Monolithic 3D design flows

Partitioning-first flow

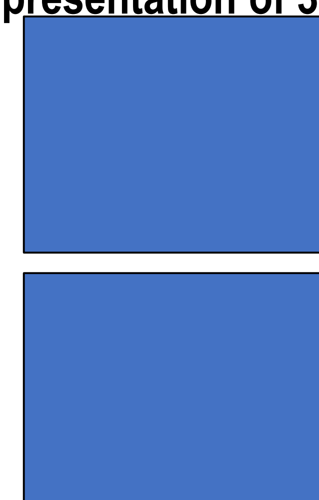
- **Partitioning-first flow:**
  - Requires an initial partitioning solution based on RTL
    - This initial partitioning has a huge impact on final PPA
    - Architecture information is required
  - Cascade-2D
  - PnR done in a modified 2:1 floorplan
    - has relatively accurate RCs in pseudo-3D stage (2D representation of 3D)



Pre-partitioned RTL



2D-like floorplan for PnR representing two 3D tiers



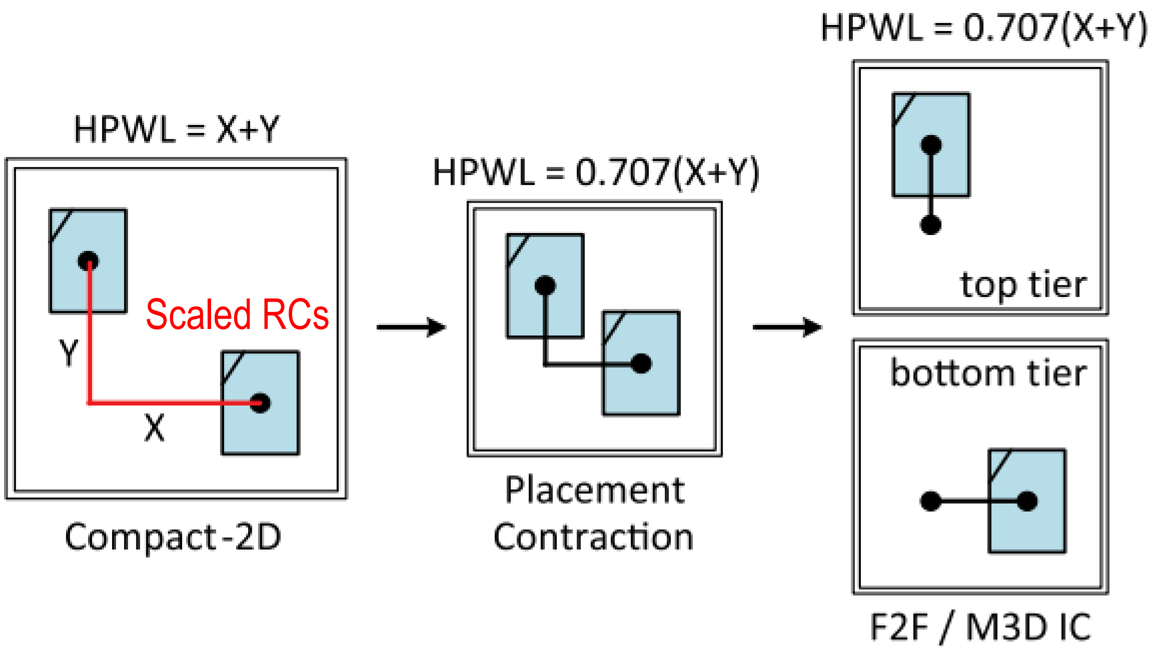
Splitting the 2D-like design into a realistic 3D

Monolithic 3D design flows

Partitioning-last flow

- **Partitioning-last flow(s):**
  - Doesn't require an initial RTL partitioning
  - PnR is first done in a pseudo-3D flow
  - Cells are partitioned into two tiers
  - Tiers are legalized, re-routed in 3D

  } Causes RC inaccuracies



Bon Woong Ku et al., "Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs," ISPD 2018, doi: 10.1145/3177540.3178244

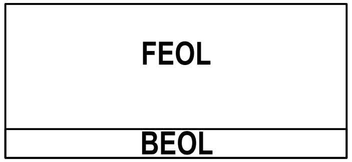
RC inaccuracies in 3D

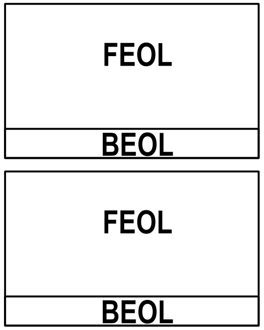
128-bit AES	Pseudo-3D (Compact-2D)	Final-3D
Metal Stack	**M1-M6**	**M1-M6–MIV–M1-M6**
Cell Pin Count	381,907	381,907
# Vias	**872,931**	**1,028,040**
Wire length (μm)	1,141,178	1,234,332
Ground Cap. (pF)	**119.34**	**140.83**
Coupling Cap. (pF)	35.20	31.71
Wire Resistance (MΩ)	**15.26**	**21.23**

3D design contains 2 metal stacks connected with an MIV (monolithic inter-tier via)

Significant RC differences b/w pseudo-3D routing to 3D routing

Cross-section views

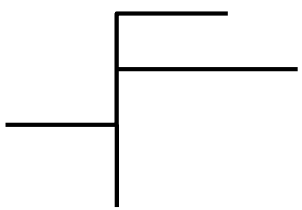




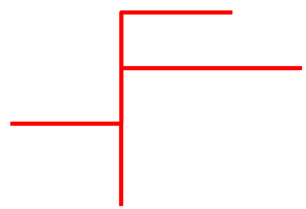
Sources of RC inaccuracies

Vias

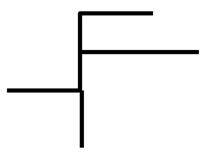
- **In Compact-2D, R and C are scaled globally by 0.707x**
- **Scenario:**
  - Let each via have v_P unit P (P is R or C parasitic); and 1µm wire w_P unit P
  - Unit parasitics remain constant b/w 2D and 3D
  - Consider how the net shown evolves at each stage



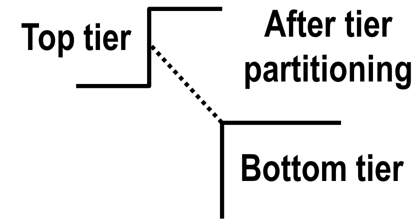
10µm long, 4vias
2D parasitic ~ $10w_p$+$4v_p$



RC Scaled parasitic ~ $7w_p$+$3v_p$
This is the value that the net is expected to have in 3D



During placement contraction
~7um long, 4 vias

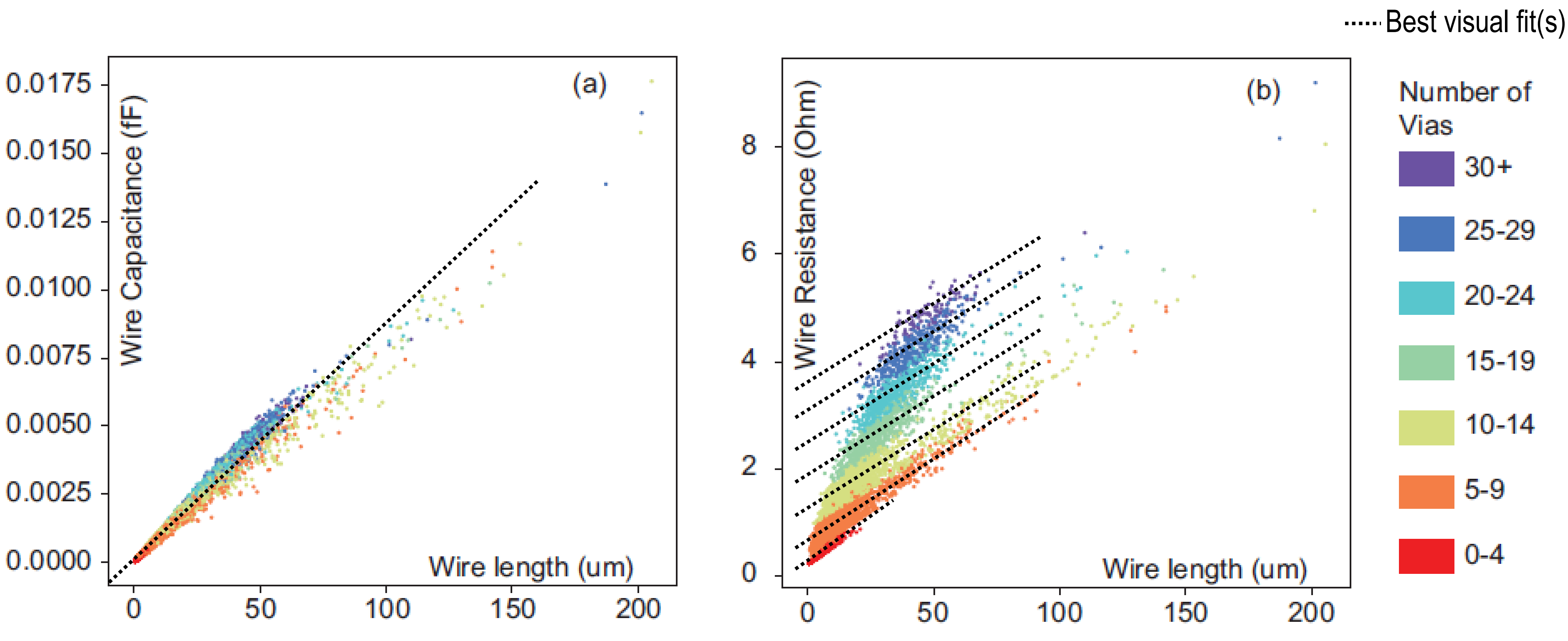


Likely wirelength > 7 um
Likely # vias > 4
Likely parasitic > $7w_p$+$3v_p$
Parasitic value estimated by pseudo-3D: $7w_p$+$3v_p$

Sources of RC inaccuracies
Tier Partitioning

- After placement contraction, cells overlap with each other
- Tier partitioning reduces but cannot eliminate overlaps completely
- The new addition of the z-axis and additional legalization increases the wirelength of the nets
- The 0.707x scaling is an optimistic value and it varies within a design

RC trends w.r.t. wirelength, via count



Scaling Error (Mismatch)

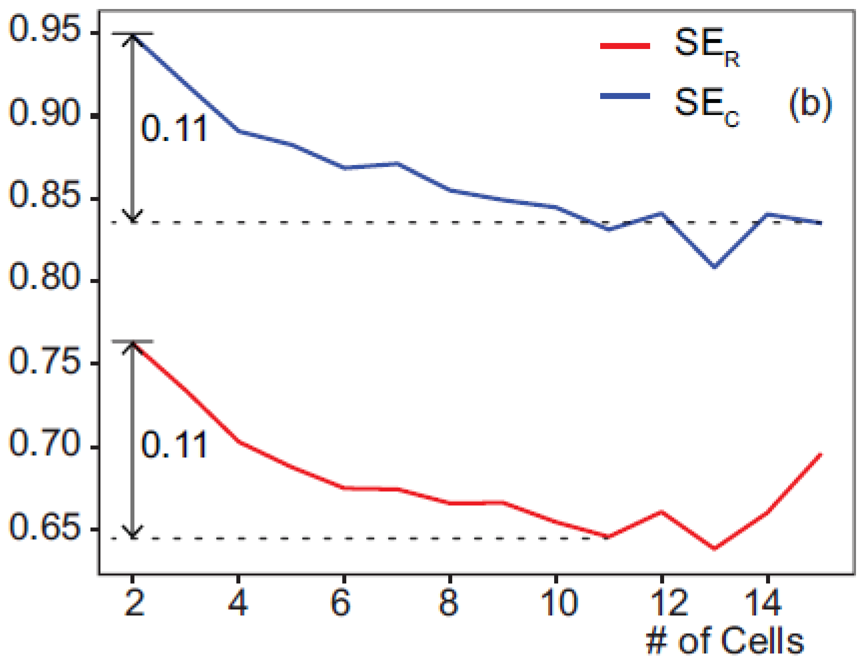
Definition and Meaning

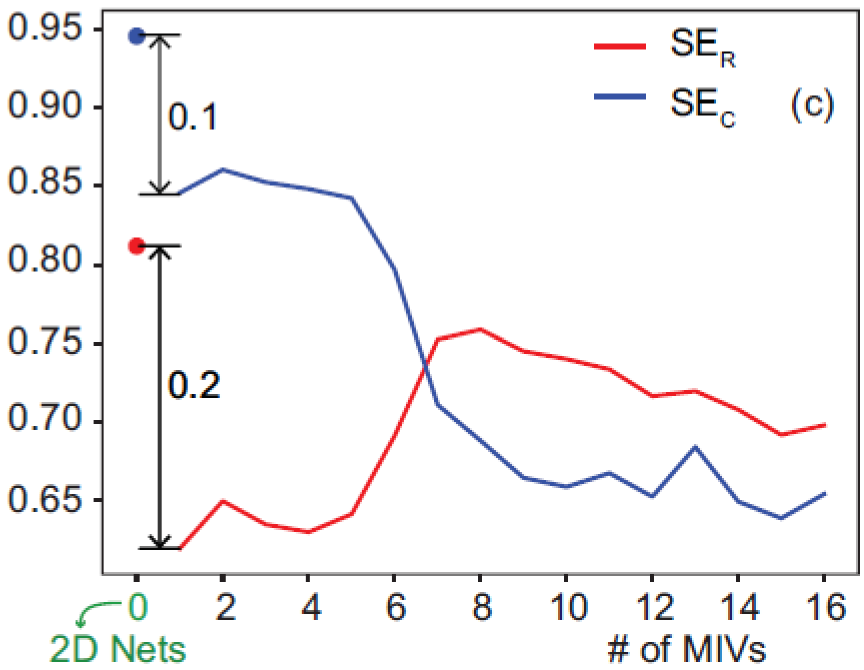
$$SE_P = \frac{P \text{ of a net in pseudo} - \text{3D}}{P \text{ of the net in final} - \text{3D}} \text{ ; } P \text{ is R or C parasitic of the net}$$

- SE_P ~ 1 means the RC scaling in pseudo-3D is adequate
  - Can happen with some nets by chance
- SE_P << 1, pseudo-3D severely underestimates the RCs
  - Expected for majority of nets
- SE_P >> 1, pseudo-3D severely overestimates the RCs
  - Only happens on a small fraction of nets

Scaling Error (Mismatch)

vs. # MIVs, #Cells

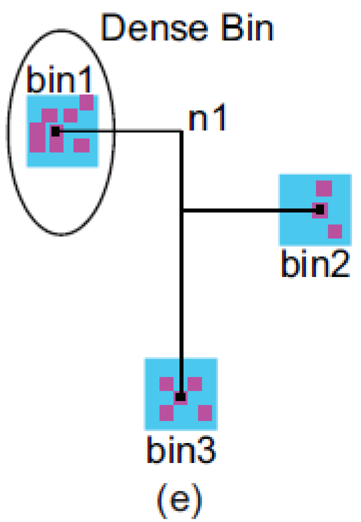


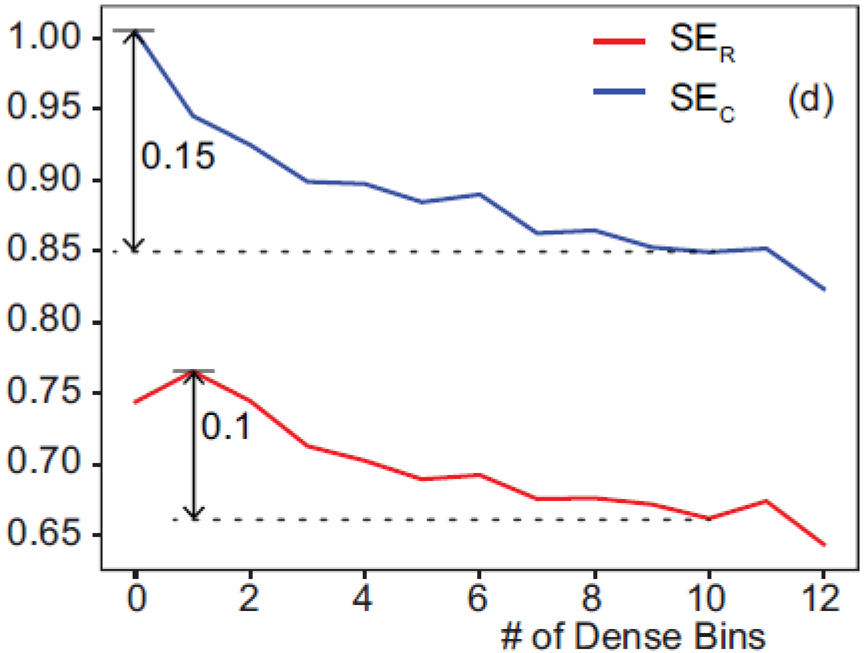


#MIVs comes from 3D routing, not a useful feature for training.
Only used for our understanding of the trends

Scaling Error (Mismatch)

vs. # Dense Bins (Regions)





A bin is defined as a small region at end-points of each net.

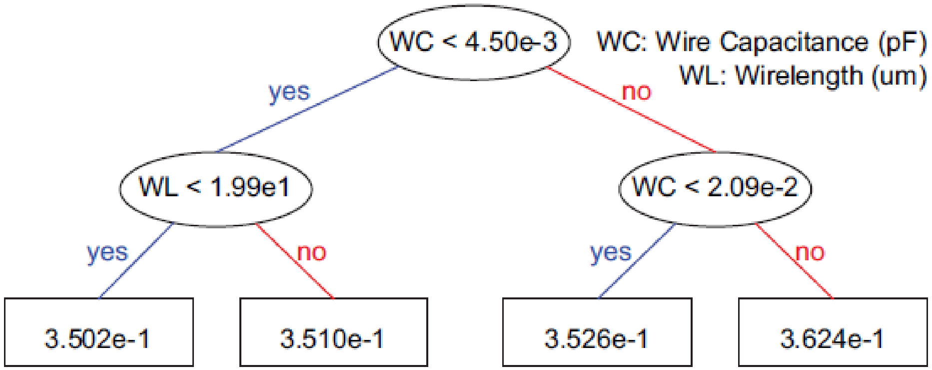
Dense bin a bin with density > threshold

Quantifies the legalization impact

Grouping nets with "Dense bins=0" has best cap. scaling factor of all the feature groupings so far
Also shows the largest range of cap. scaling error

ML model and Loss function

- **Model: XGBoost python library**
  - **Good for classification, regression**
  - **Loss function (squared error):** $\mathrm{L_P} = \sum_{nets}(P_{pred} - P_{true})^2$

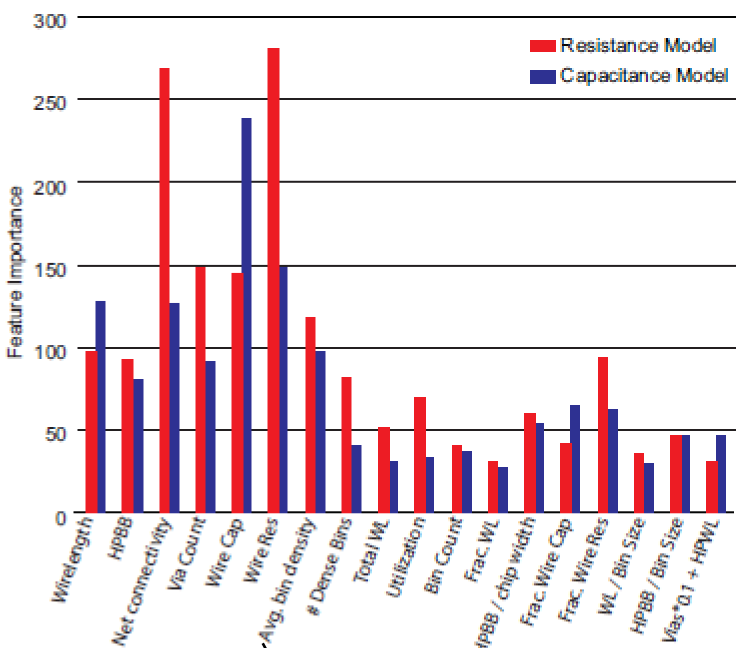


Example classification tree of a capacitance model
Leaf values are its contribution to final capacitance

Features for training

Feature Name	Significance
	Individual Net Features
Wirelength	Long, medium, short wires have varying average $SE_{R(C)}$
HPBB	Wirelength estimation from placement; not affected by congestion
Net connectivity	Number of cells connected by the net; estimates partitioning probability
Via Count	Number of vias on the net; useful for resistance calculation
Wire Cap	capacitance in pseudo-3D design, has strong effect on final capacitance
Wire Res	resistance in pseudo-3D design, has strong effect on final resistance
Average Local density	Average density of all the bins of a given net as shown in Figure 4(e); useful for deciding legalization errors
Dense Bins	Number of dense bins of each net; useful for deciding legalization errors
	Full Chip Features (Design Identifiers)
Total WL	Total routed signal wirelength, design identifier
Number of nets	Total number of nets in the design
Number of cells	Total number of cells in the design
Average Fanout	Number of cell pins/ number of nets; design connectivity information
Chip Area	Footprint of the design
Cell Area	Total standard cell area in the design
Utilization	Standard cell density
Bin Count	Number of partitioning bins to be used
	Derived Features
net WL/ total WL	Identifying global nets among various designs; detouring, cross coupling capacitance information
HPBB / $\sqrt{\text{Chip area}}$	Identifying global nets among various designs, less affected by pseudo-3D congestion
Wire Cap / Total Cap	Fractional capacitance of net w.r.t. total, importance of net within a design
Wire Res / Total Res	Fractional resistance of net w.r.t. total, importance of net within a design
net WL / Bin size	For a given wirelength, a increase in bin-size would decrease the probability of net being partitioned
HPBB / Bin size	Similar to net WL/ Bin size, but only considers placement information
HPBB + 0.1*VC	A combined HPBB, via count feature

Feature importance in training

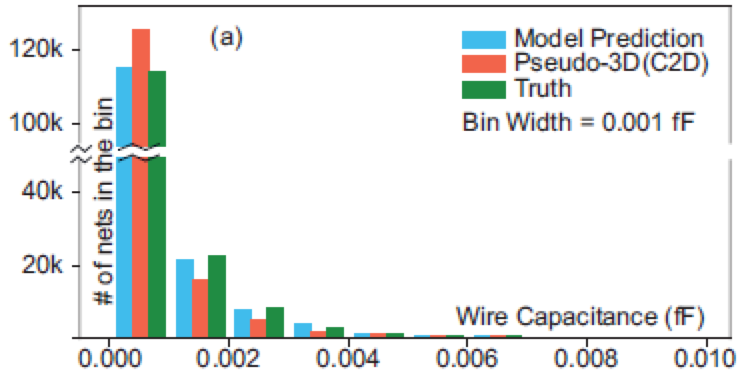


Features in order of importance

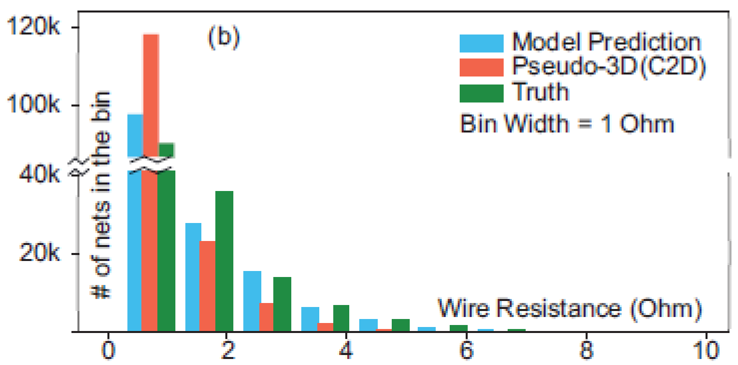
R model	C model
Wire Resistance	**Wire Capacitance**
Net connectivity	Wire Resistance
Via Count	Wirelength
Wire Capacitance	Net connectivity
Avg. bin density	**Avg. bin density**
Wirelength	Via Count

Better than #dense bins, as it is based on exact bin density
#dense bins uses a binary 1, 0 value that depends on the exact bin density

RC histograms for an unseen netlist



Shows clear improvement from using the XGBoost model for both resistance and capacitance of nets



Impact on unseen netlists

Netlist ↓	Design→		C2D-3D	C2D+ML-3D
b19 35k cells 1.25 GHz	$RMSE_{cap}$	aF	0.73	0.44
	$RMSE_{res}$	dΩ	**1.27**	**0.49**
	WNS	ns	-0.16	-0.10
	TNS	ns	**-155.8**	**-37.3**
	Power	mW	40.1	40.1
tate 150k cells 1.00 GHz	$RMSE_{cap}$	aF	1.26	0.85
	$RMSE_{res}$	dΩ	**1.42**	**0.62**
	WNS	ns	**-0.40**	**-0.45**
	TNS	ns	**-497.3**	**-76.8**
	Power	mW	132.4	131.8

For context,
average res. in training set=1.8dΩ
average cap. in training set=1.5aF

WNS is harder to control

Total power is slightly smaller even when TNS is ~6x smaller

Summary

- **Feature analysis shows a variation in RCs can be captured with ML models**

- **A general ML model for predicting RCs in 3D ICs is presented**

- **ML model limits RC mismatch before and after tier partitioning for any pseudo-3D flows; TNS improves by >6x**

- **WNS is harder to control and can be negatively impacted even when TNS is substantially small**