東南大學
SOUTHEAST UNIVERSITY

Pre-Routing Path Delay Estimation Based on Transformer and Residual Framework

Tai Yang, Guoqing He, Peng Cao

National ASIC System Engineering Technology Research Center, Southeast University, Nanjing, China

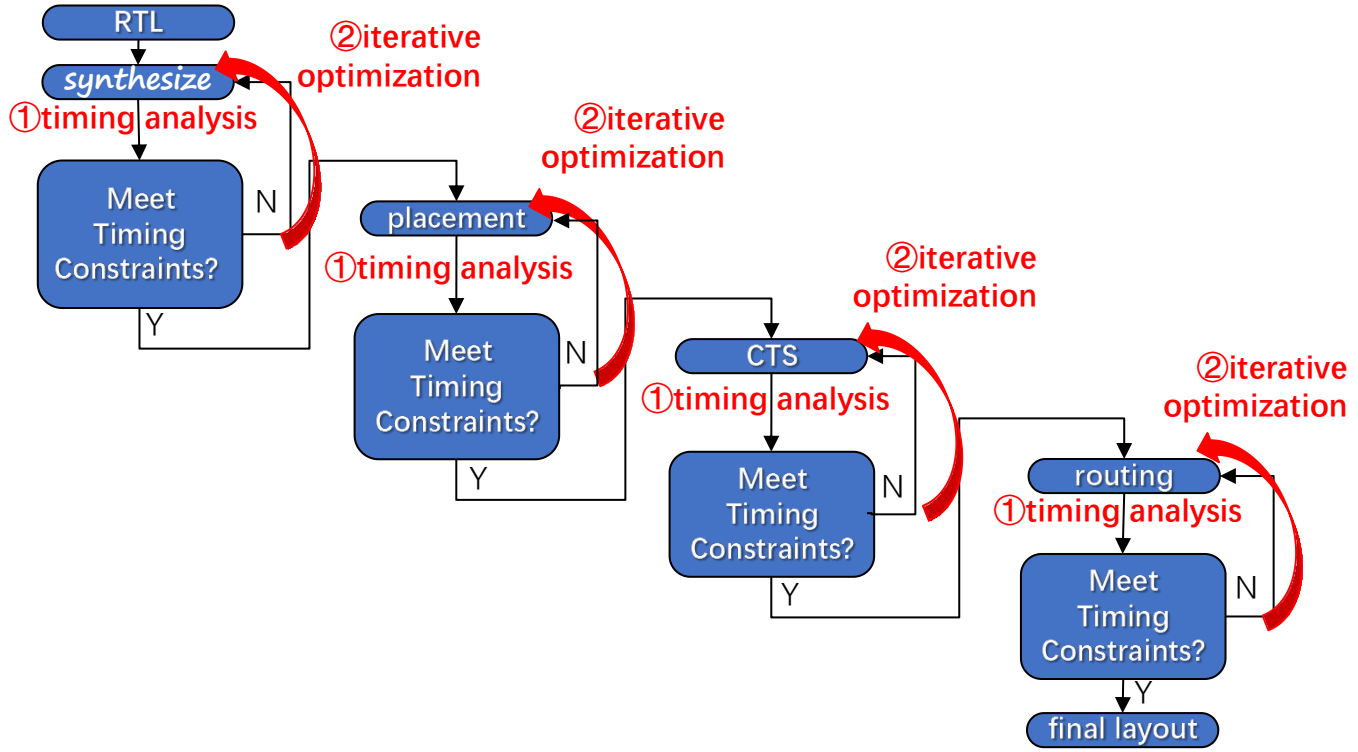
caopeng@seu.edu.cn

ASP-DAC2022

OUTLINE

- Background
- Related Work
- Pre-Routing Path Delay Framework
- Results
- Conclusion

Background

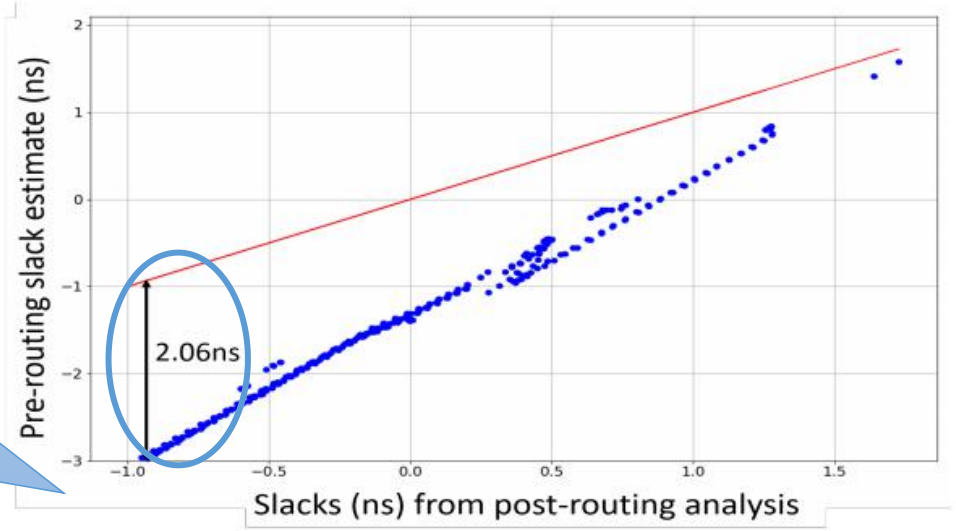


IC Backend Flow

- Problems: As the design flow gets closer to tape-out, the updated circuit timing faces nonnegligible mismatch between each stage of design flow, posing severe challenges for circuit optimization.

Background

The worst case slack is over-estimated by more than 2ns. Such pessimism causes over-design that wastes power, area and optimization time.

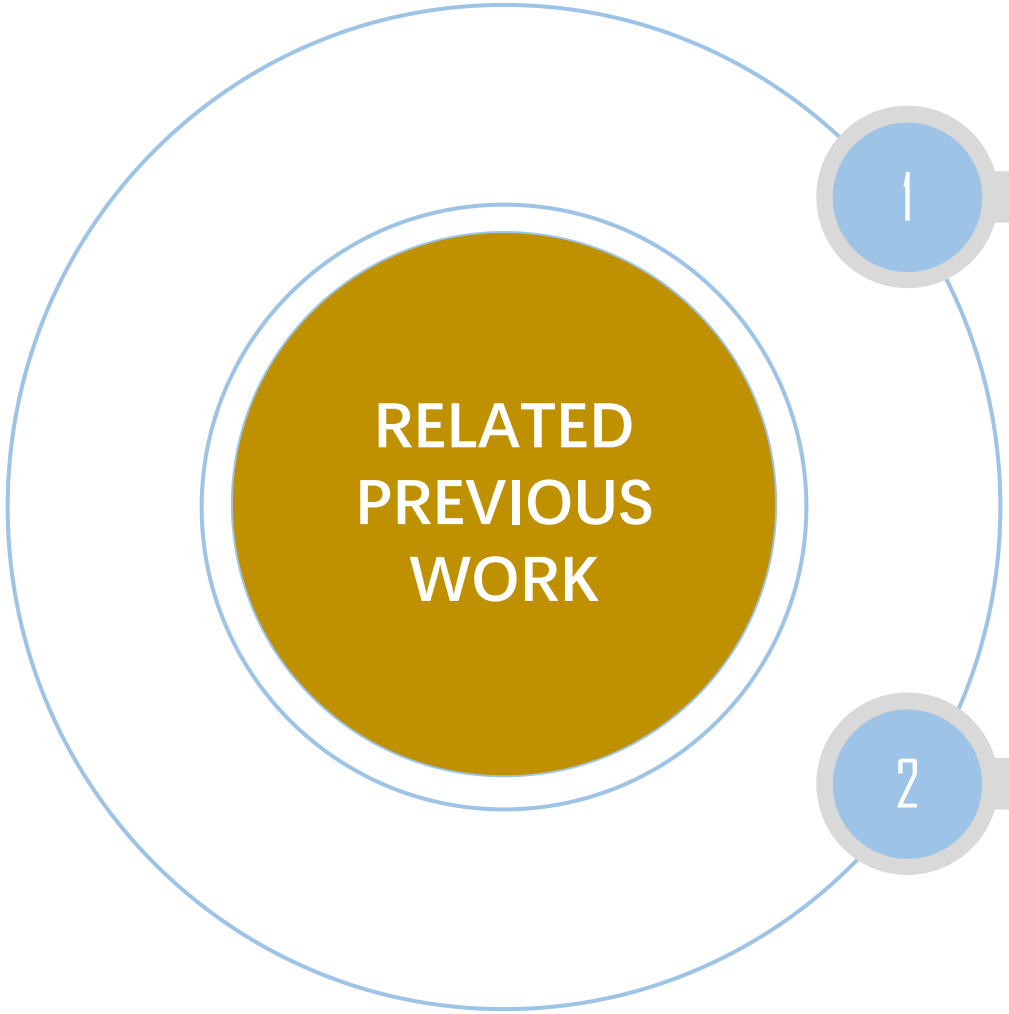


The pessimism of a commercial tool pre-routing timing estimate.[DAC'19]

- This work: pre- and post-routing timing correlation

[DAC'19]E. C. Barboza, N. Shukla, Y. Chen, J. Hu, "Machine learning-based pre-routing timing prediction with reduced pessimism," in 2019 56th ACM/IEEE Design Automation Conference (DAC), pp. 1-6, IEEE, 2019.

Related work



1. Fast pre-routing timing estimation based on traditional mathematical model

2. Timing analysis based on machine learning method

Related work

Fast pre-routing timing estimation based on traditional mathematical model

Ref.	Affiliation	Title	Focus
10'SLIP	Synopsys, Brown University	Fast, accurate a priori routing delay estimation	Post-routing delay estimation
04'DAC	University of California	Pre-layout wire length and congestion estimation	Wire length and congestion estimation
06'ICECC	Syracuse University	Pre-layout estimation of interconnect lengths for digital integrated circuits	Pre-layout interconnect lengths estimation
00'SLIP	University of Toronto	Pre-layout estimation of individual wire lengths	Individual wire lengths estimating during the technology mapping phase of logic synthesis

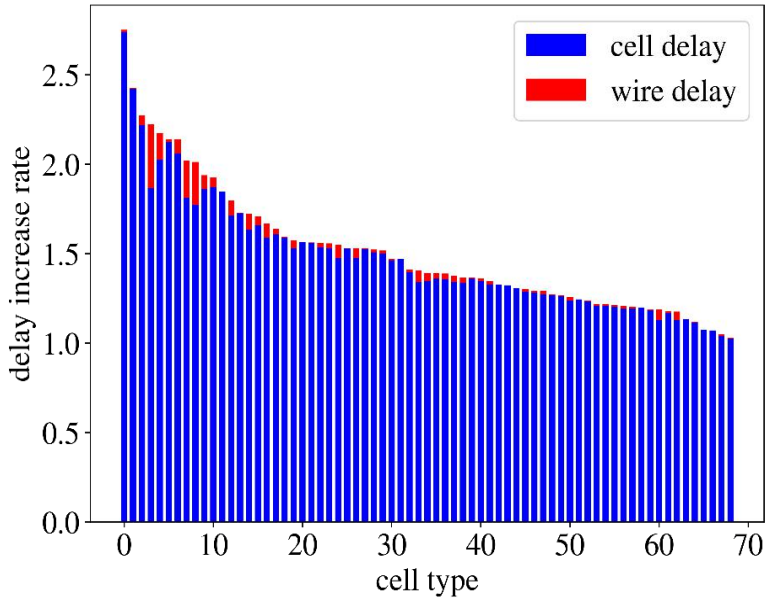
Focus：wire length or wire delay estimation

Related work

Fast pre-routing timing estimation based on traditional mathematical model

Ref.	Affiliation	Title	Focus
10'SLIP	Synopsys, Brown University	Fast, accurate a priori routing delay estimation	Post-routing delay estimation
04'DAC	University of California	Pre-layout wire length and congestion estimation	Wire length and congestion estimation
06'ICECC	Syracuse University	Pre-layout estimation of interconnect lengths for digital integrated circuits	Pre-layout interconnect lengths estimation
00'SLIP	University of Toronto	Pre-layout estimation of individual wire lengths	Individual wire lengths estimating during the technology mapping phase of logic synthesis

Focus：wire length or wire delay estimation



Average increase ratio of net delays between routing and placement stages for all types of cells

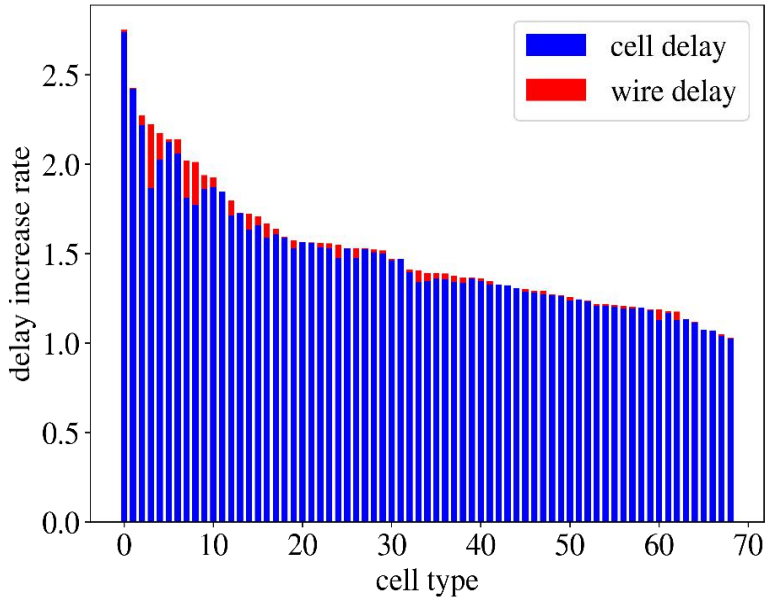
The impact of routing to the cell delay is much more significant than that of wire delay.

Related work

Fast pre-routing timing estimation based on traditional mathematical model

Ref.	Affiliation	Title	Focus
10'SLIP	Synopsys, Brown University	Fast, accurate a priori routing delay estimation	Post-routing delay estimation
04'DAC	University of California	Pre-layout wire length and congestion estimation	Wire length and congestion estimation
06'ICECC	Syracuse University	Pre-layout estimation of interconnect lengths for digital integrated circuits	Pre-layout interconnect lengths estimation
00'SLIP	University of Toronto	Pre-layout estimation of individual wire lengths	Individual wire lengths estimating during the technology mapping phase of logic synthesis

Focus：wire length or wire delay estimation



Average increase ratio of net delays between routing and placement stages for all types of cells

The impact of routing to the cell delay is much more significant than that of wire delay.

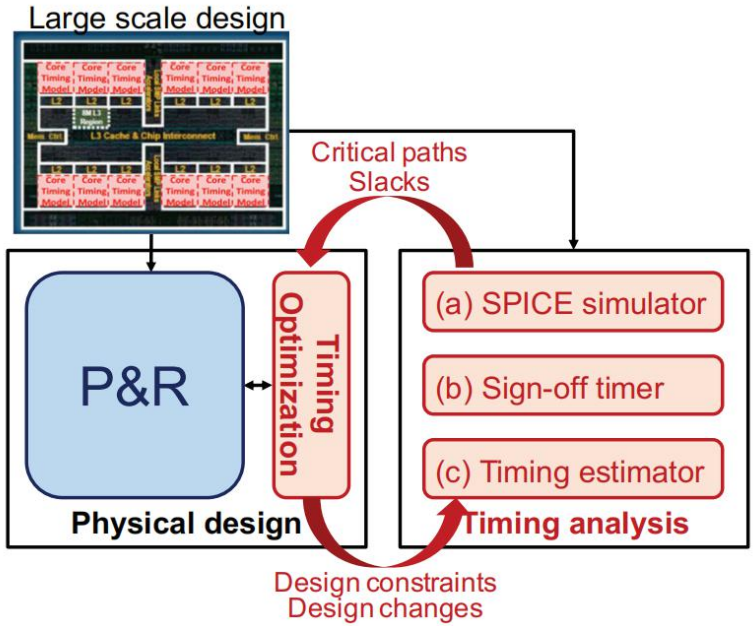
Pre-routing timing estimation requirement:
(1)fast and accurate (2)pay more attention to cell delay or path delay

Related work

SOUTHEAST UNIVERSITY 東南大學

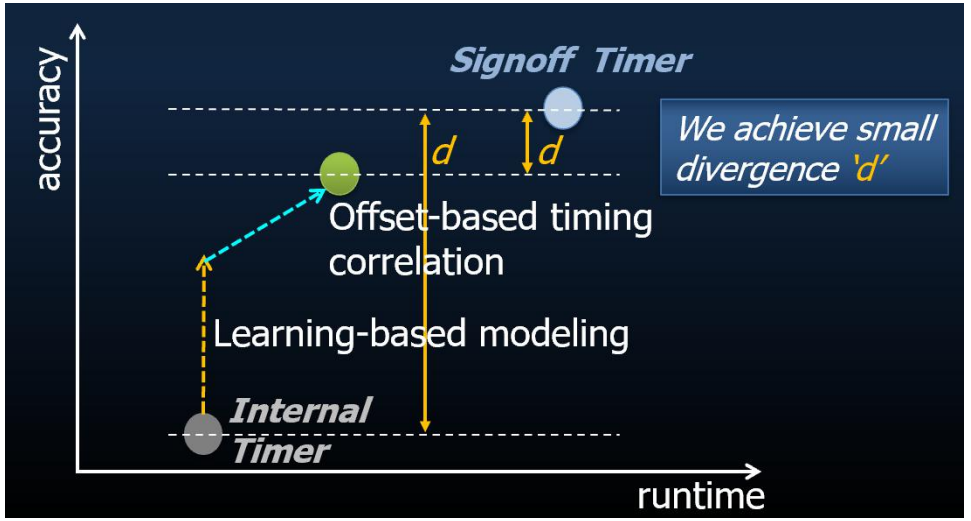
In recent years, the learning-based methods have been extended in the application of timing analysis

- Application: A fast and accurate timing estimator which can highly correlate with a sign-off timer to shorten turn-around time
- ML models: RF, Lasso, XGBoost



[DAC'20, H. H. Cheng, Fast and accurate wire timing estimation on tree and non-tree net structures]

- Application: Wire delay/slew models for internal incremental STA to delay the deviation in endpoint slack from a STA tool.
- ML models: Least squares regression



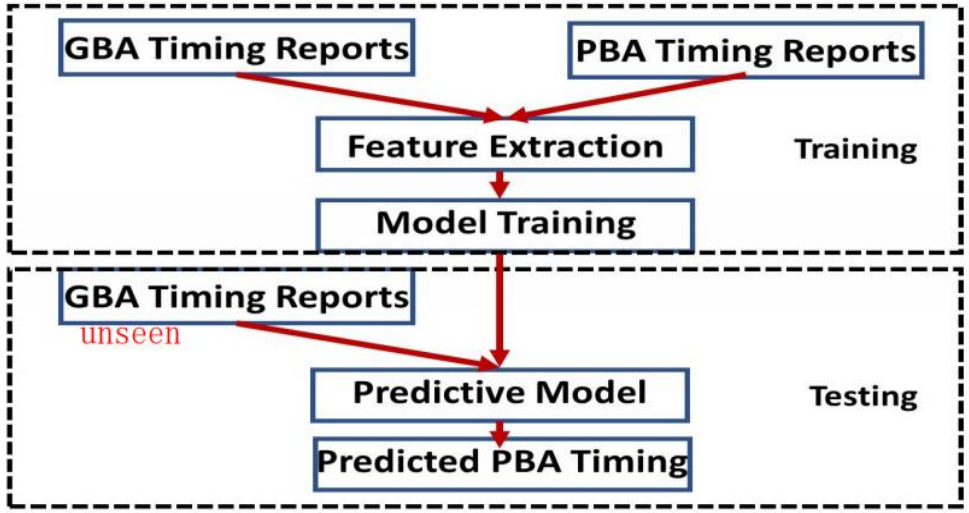
[SLIP'13, A. B. Kahng, Learning-based approximation of interconnect delay and slew in signoff timing tools]

Related work

SOUTHEAST UNIVERSITY 東南大學

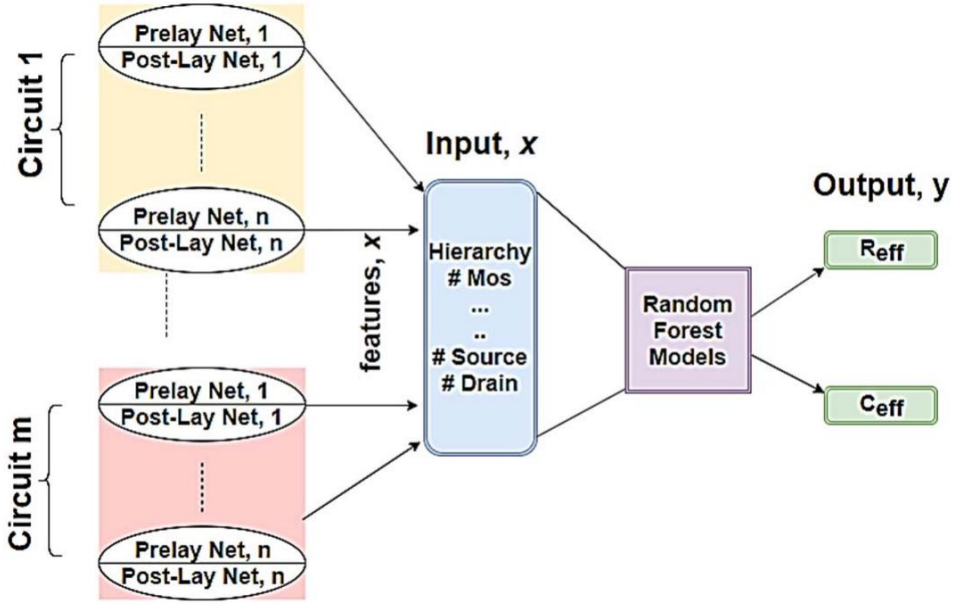
In recent years, the learning-based methods
have been extended in the application of timing analysis

- Application：Predict path-based slack from graph-based timing analysis
- ML models：RF



[ICCD'18, A. B. Kahng, Using machine learning to predict path-based slack from graph-based timing analysis]

- Application：MLParest provides an accurate estimate of expected post-layout interconnect parasitics in the pre-layout design phase
- ML models：RF



[DAC'20, B. Shook, MLParest: Machine learning based parasitic estimation for custom circuit design]

Pre-Routing Path Delay Framework

SOUTHEAST UNIVERSITY

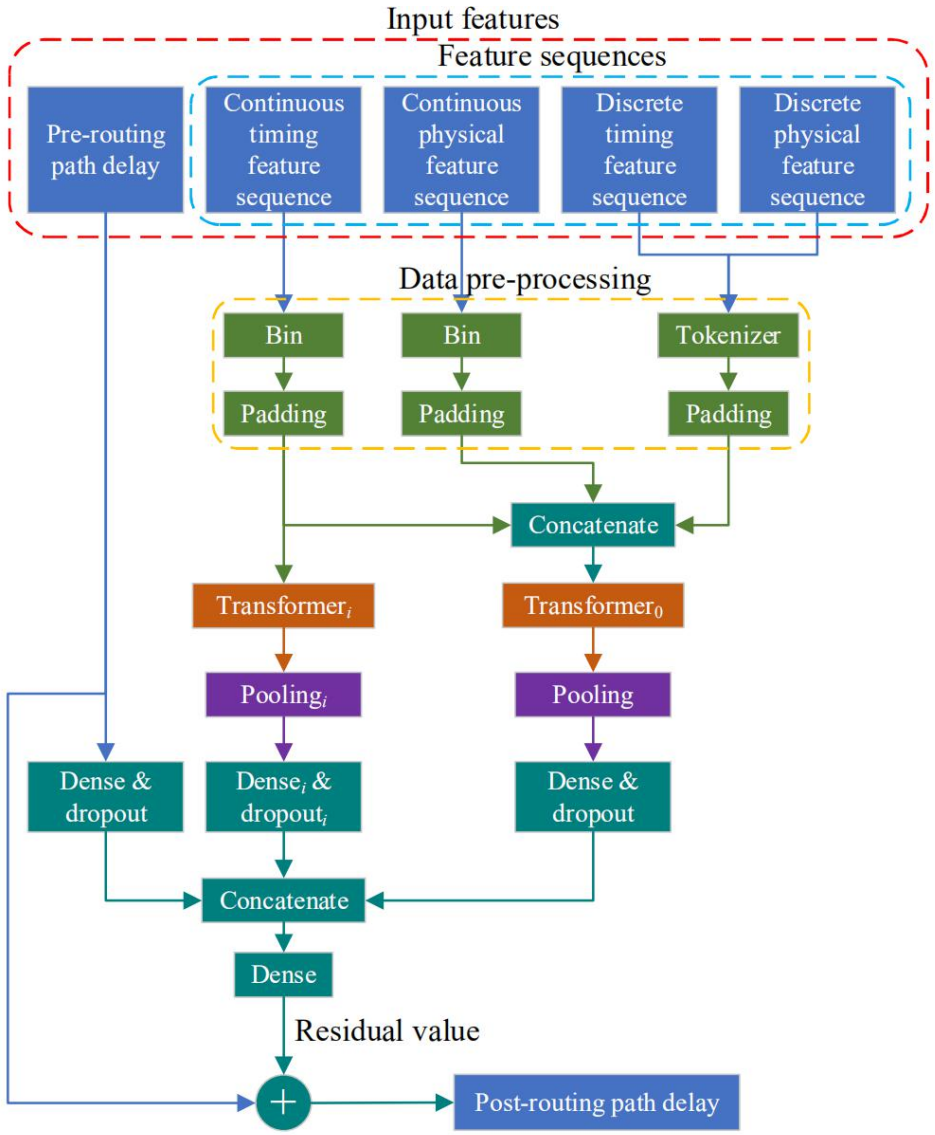
Problems:

- Neglect of the delay correlation along the path
- Prediction error accumulation and computational complexity increase

An efficient and accurate pre-routing path delay prediction framework is proposed in this work by employing transformer network and residual model.

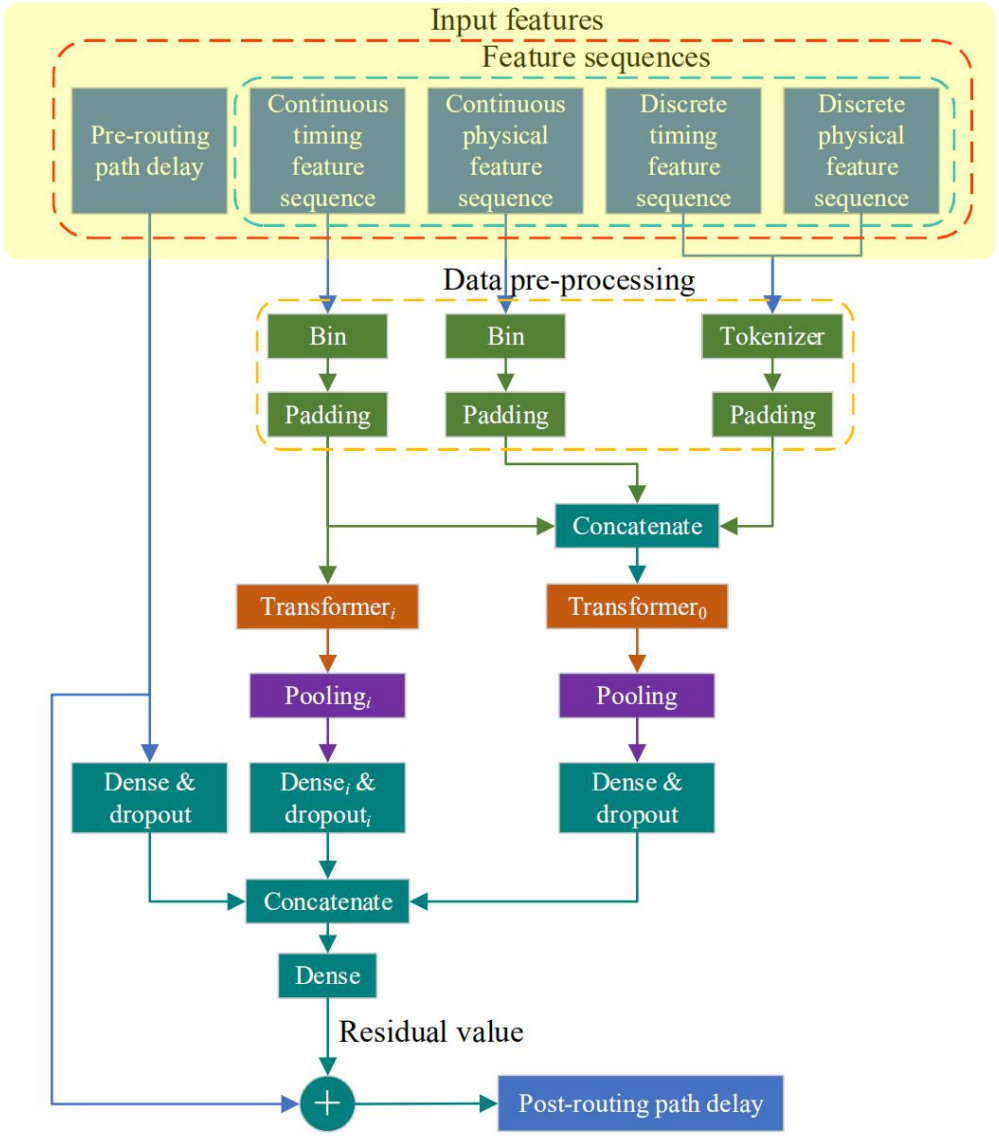
- Sequence features at placement stage
- Transformer network: exploits the correlations through circuit path
- Residual model: calibrate the mismatch between the pre- and post-routing path delay
- Without additional computation

Pre-Routing Path Delay Framework



Overview of the prediction

Framework: feature selection



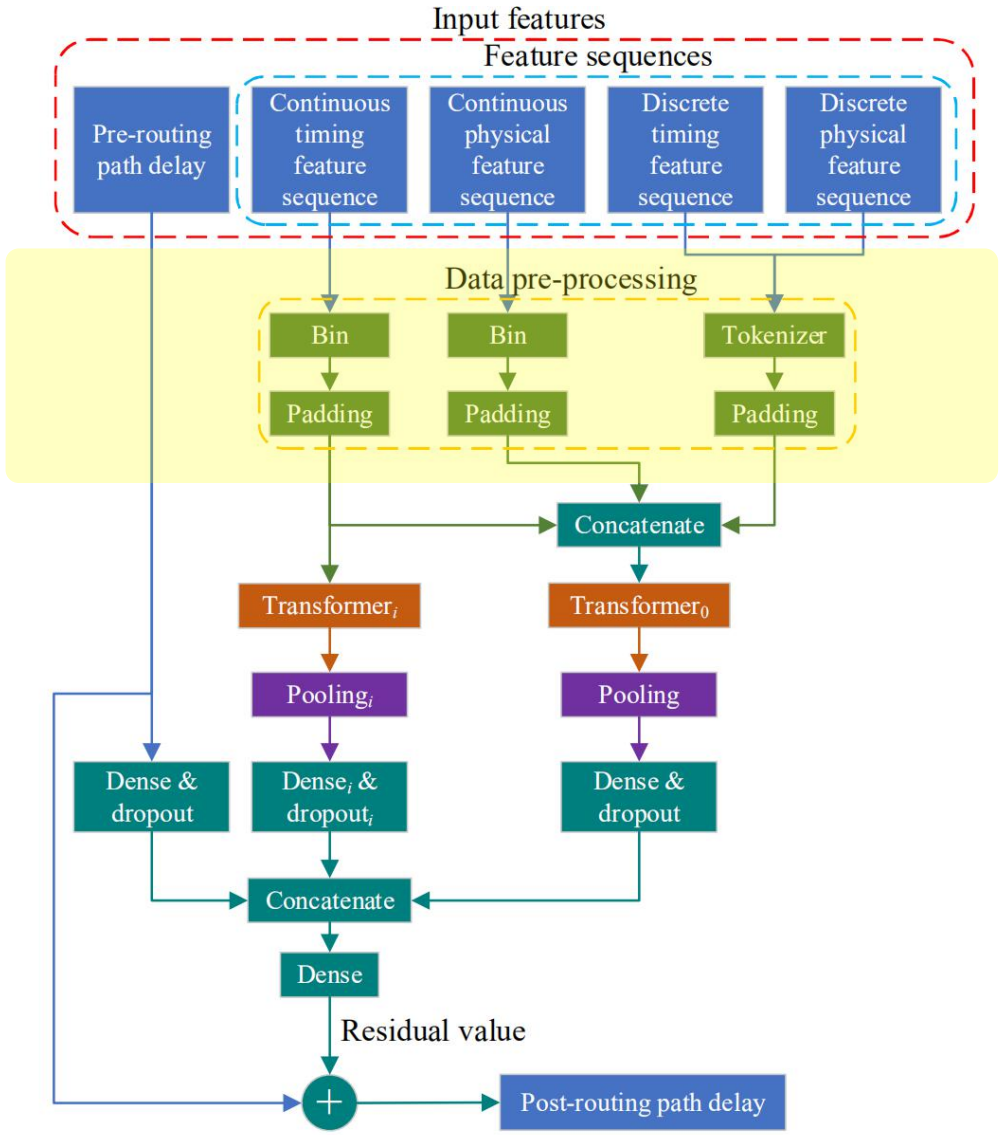
Overview of the prediction

Feature selection and data pre-process

Features	Continuous variable	Discrete variable
Physical sequence	Pin cap, pin location	Cell type
Timing sequence	Input/output transition time, cell delay	Signal polarity
Timing scalar	Pre-routing path delay	

Sequence:
the representation of path characteristics

Framework: data pre-process

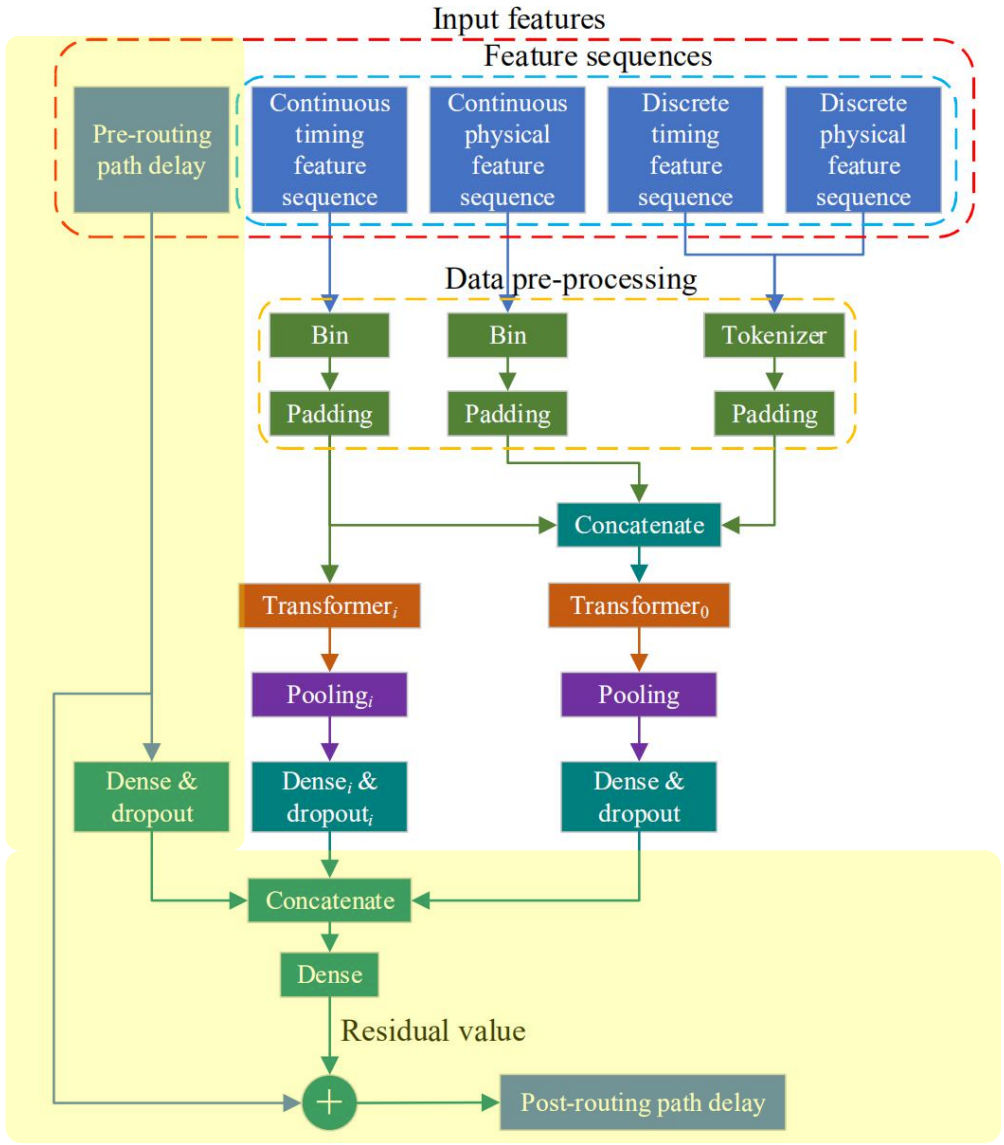


Overview of the prediction

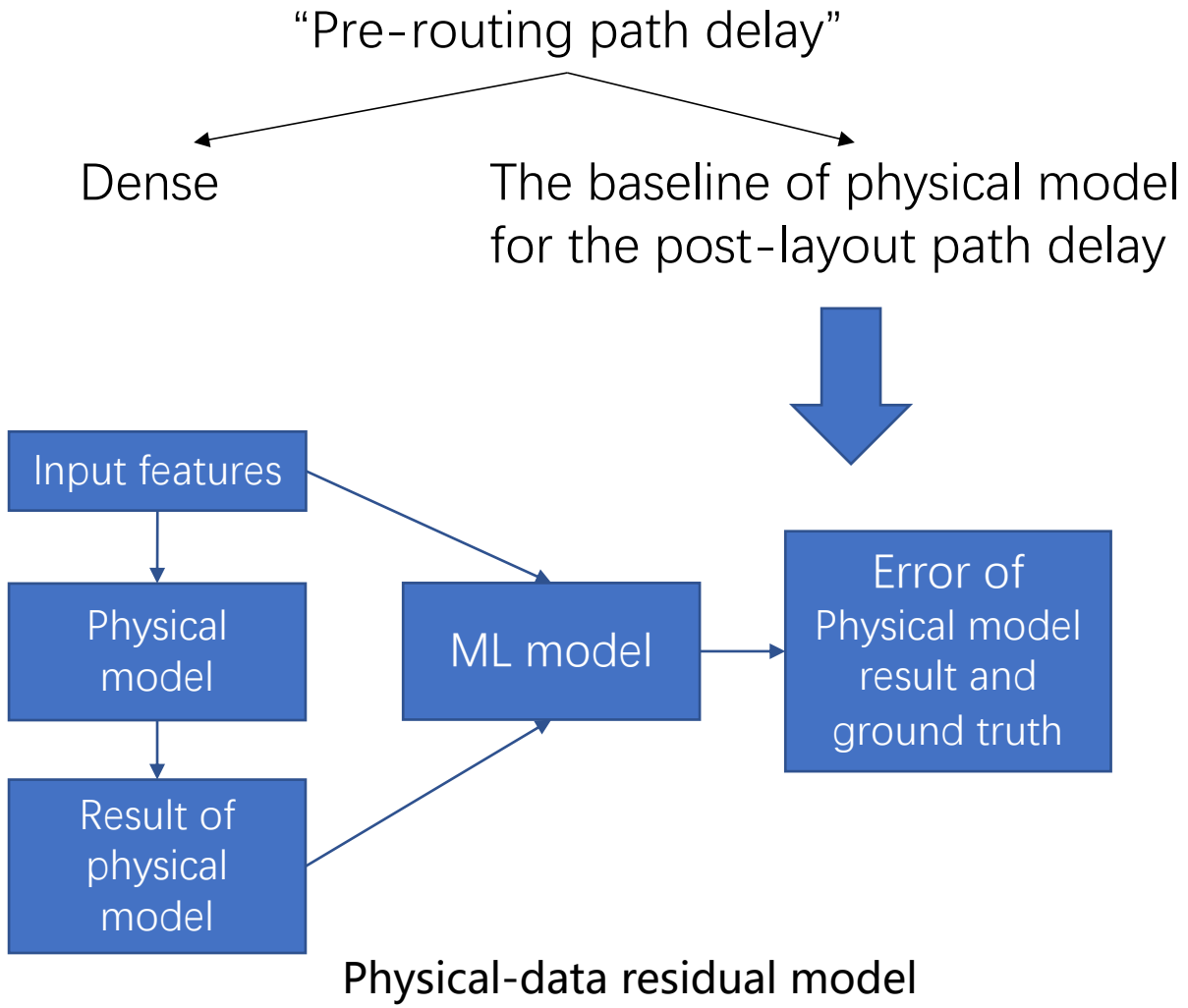
Feature selection and data pre-process

Features	Continuous variable	Discrete variable
Physical sequence	Pin cap, pin location	Cell type
Timing sequence	Input/output transition time, cell delay	Signal polarity
data pre-process	Bin +padding	Tokenizer +padding

Framework: “pre-routing path delay”

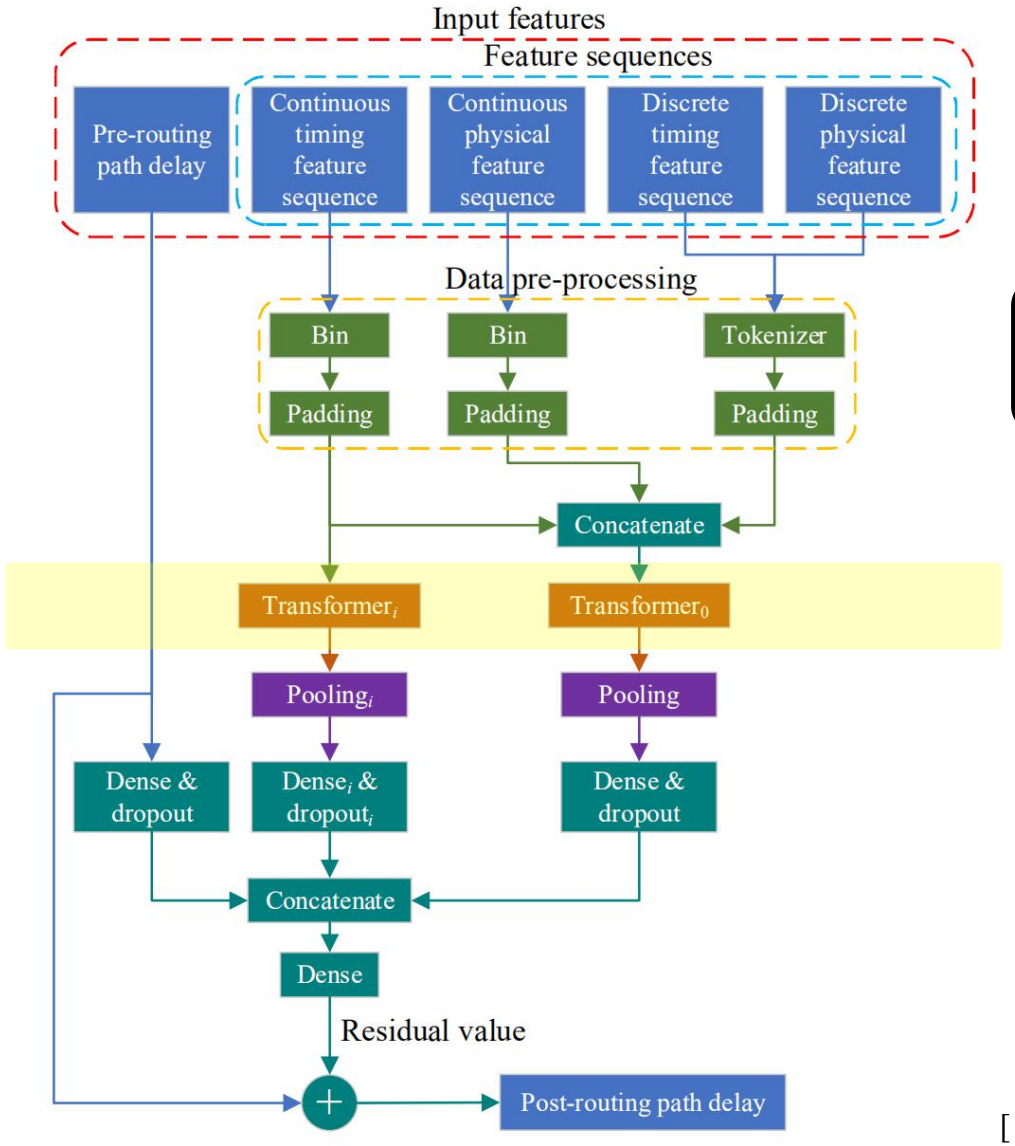


模型框架

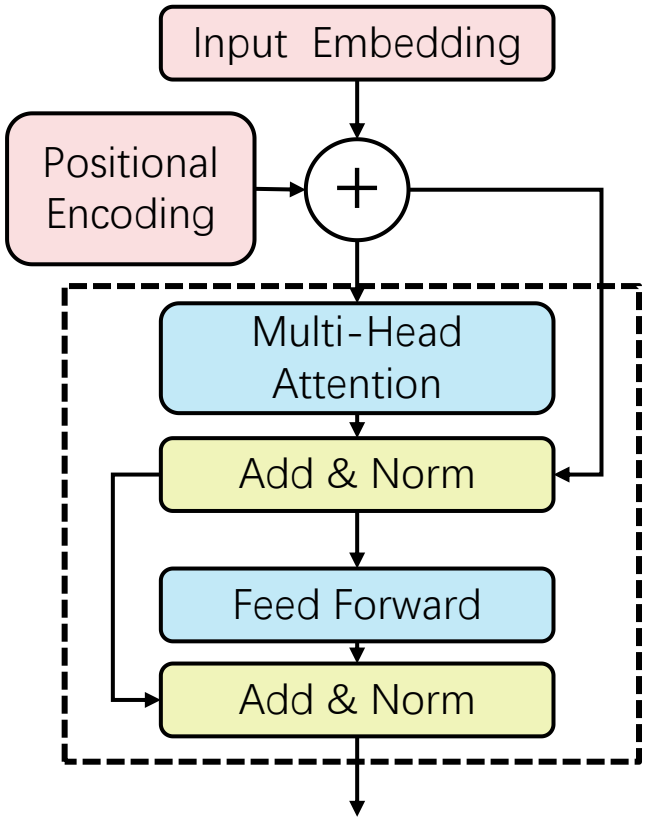


Physical-data residual model

Framework: transformer encoder



Overview of the prediction



Transformer encoder[17'NIPS]

Input embedding and position encoding: consider cell positions information in a data path

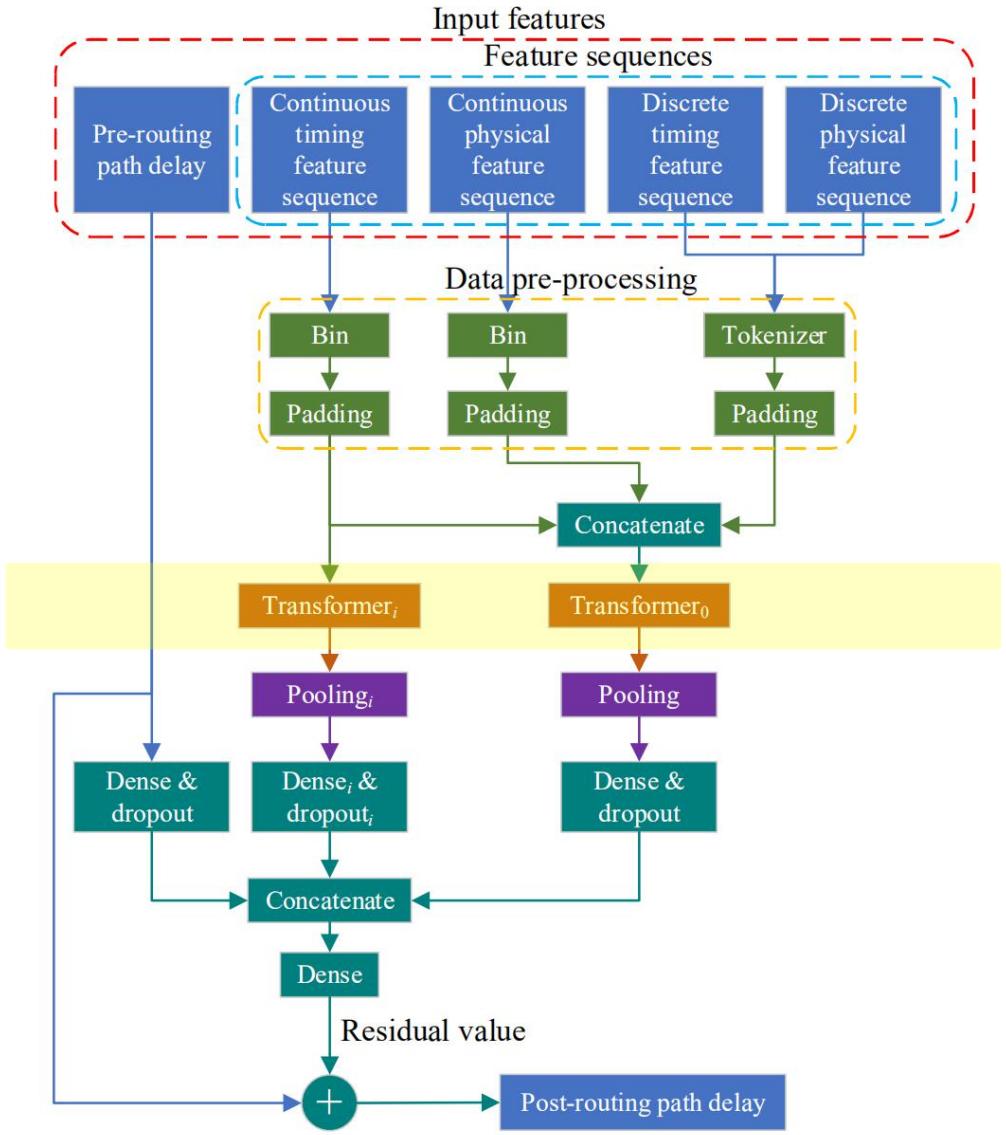
Multi-head attention: most important

Feed Forward: performed in parallel

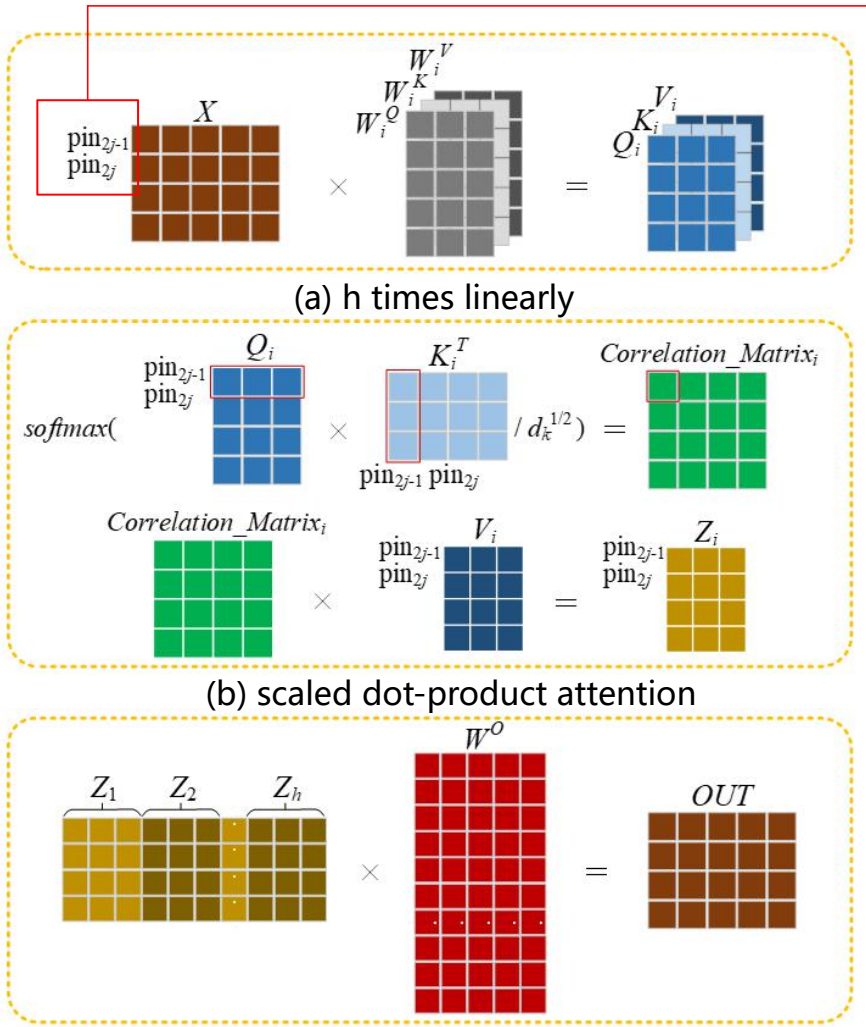
Add and normalization: solved the problem of vanishing and exploding gradients

[17'NIPS] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998-6008, 2017.

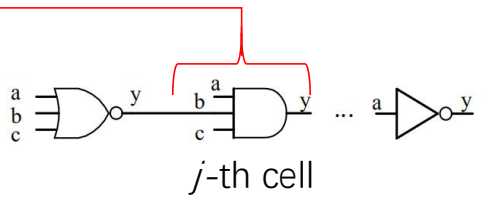
Framework: attention mechanism



Overview of the prediction



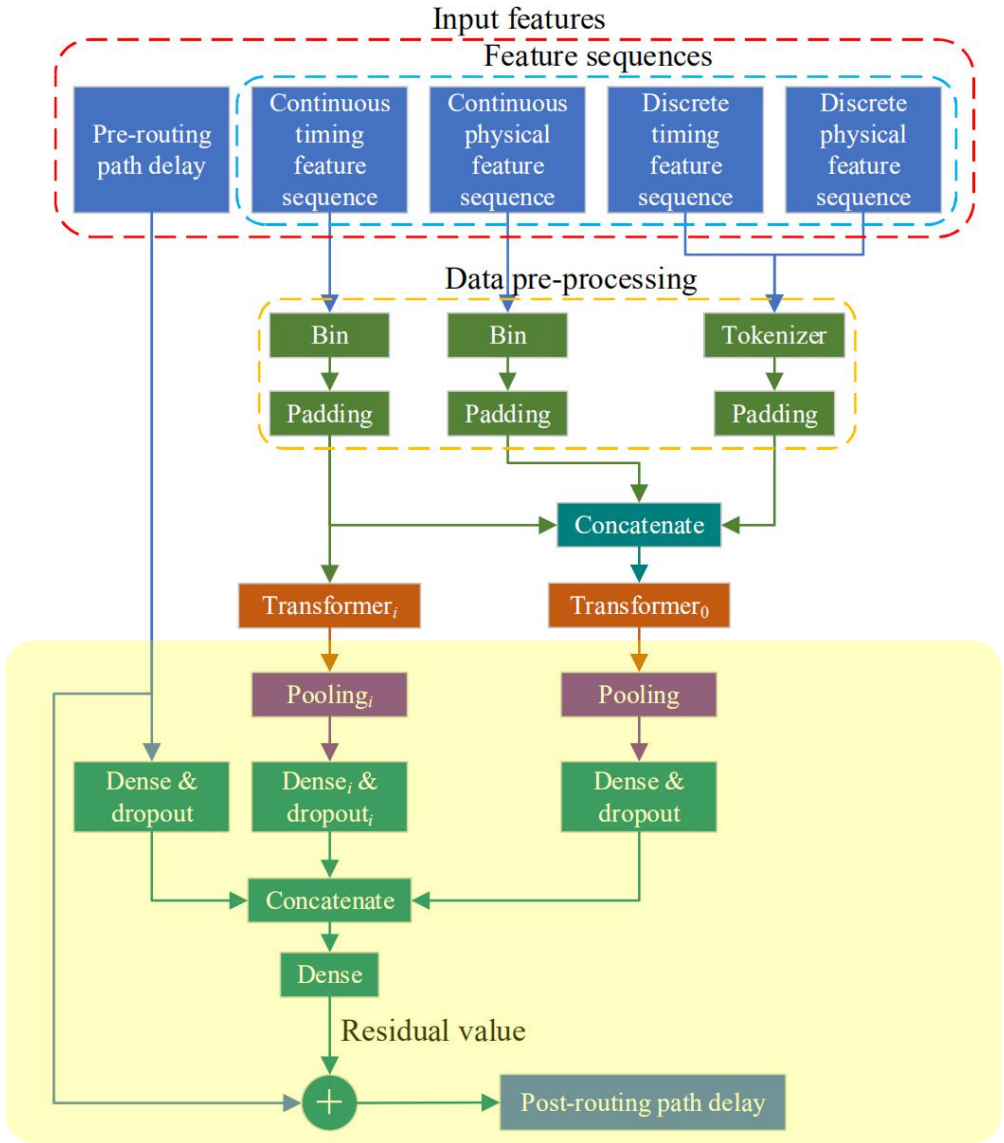
h-head self-attention mechanism



Why transformer ?

- Position information in the sequence
- Performed in parallel
- Exploits the correlations of the timing and physical information through circuit path by its multi-head self-attention mechanism

Framework: data dimension reduction



Overview of the prediction

- Dimension reduction and data concatenation
- Predict the residual value and add it to the pre-routing path delay

Results

Experiment setup：

- Framework implementation：
  Python， keras
- TSMC 28nm technology
- Circuits： 5 circuits

  3 seen circuits， randomly divide training and test sets

  2 unseen circuits ， all of them are test sets

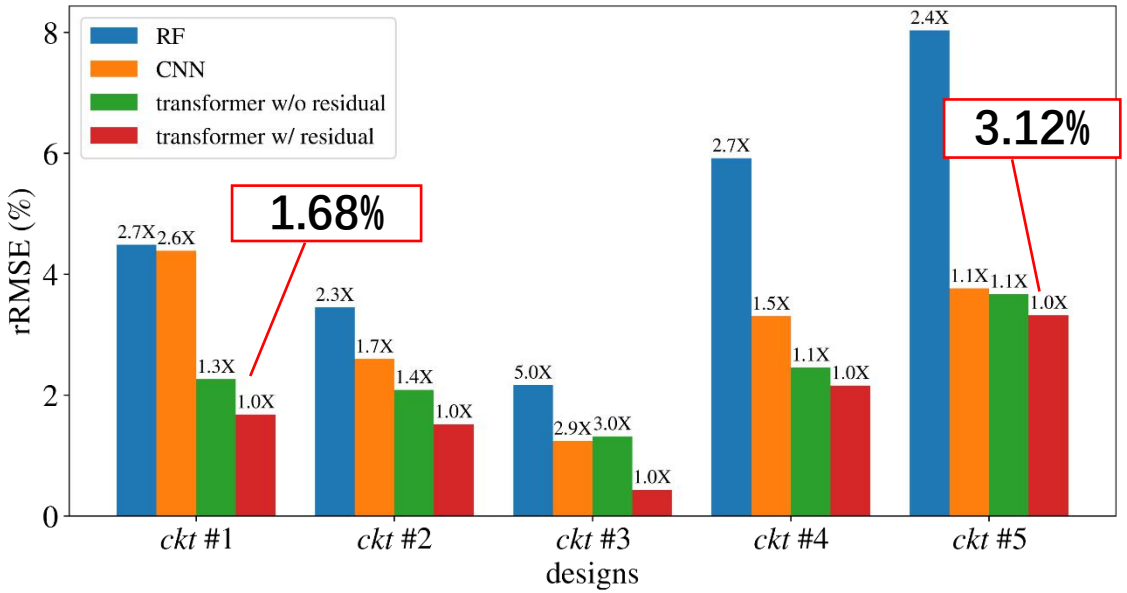
Circuit Statistics

Circuits	# Train	# Test	#cell	#net	category
ckt #1	40791	17483	10154	18892	seen
ckt #2	93786	40194	234391	340004	seen
ckt #3	16099	6900	37958	51175	seen
ckt #4	0	16998	6667	9072	unseen
ckt #5	0	23785	11830	15170	unseen
Total	150676	105360	301000	434312	

Results

SOUTHEAST UNIVERSITY

Accuracy Comparison:



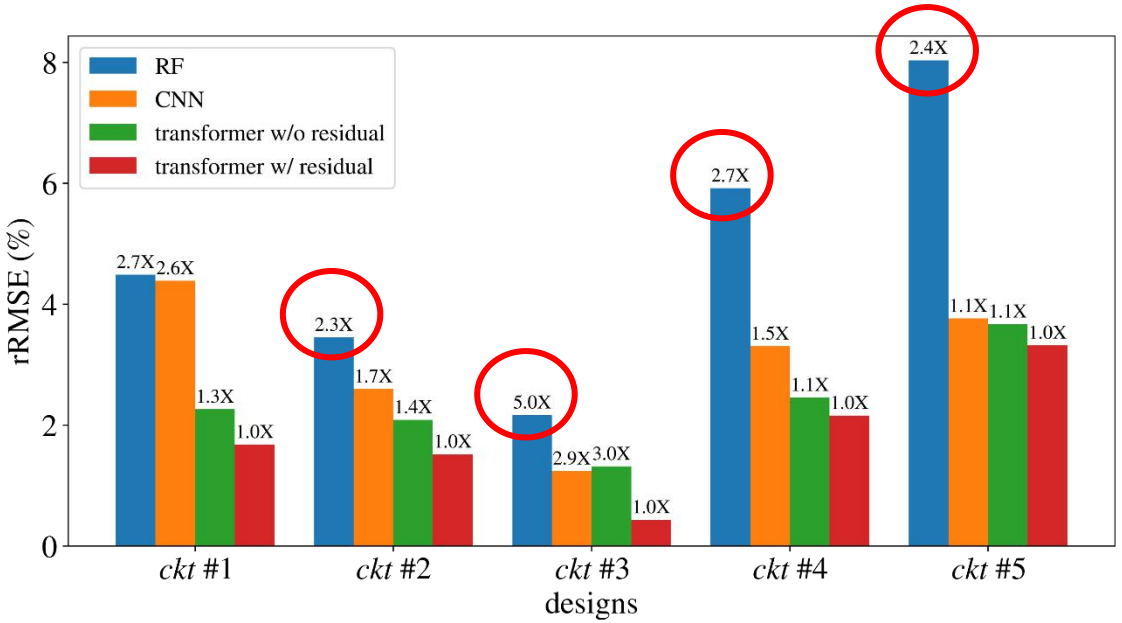
Error distribution of different models on seen and unseen designs.

	Seen ckt	Unseen ckt
rRMSE	<1.68%	<3.12%
Compared with RF, reduced by	2.3X~5X	2.4X~2.7X
Compared with CNN, reduced by	1.7X~2.9X	1.1X~1.5X
Residual model benefits	>30%	>10%

Results

SOUTHEAST UNIVERSITY

Accuracy Comparison:



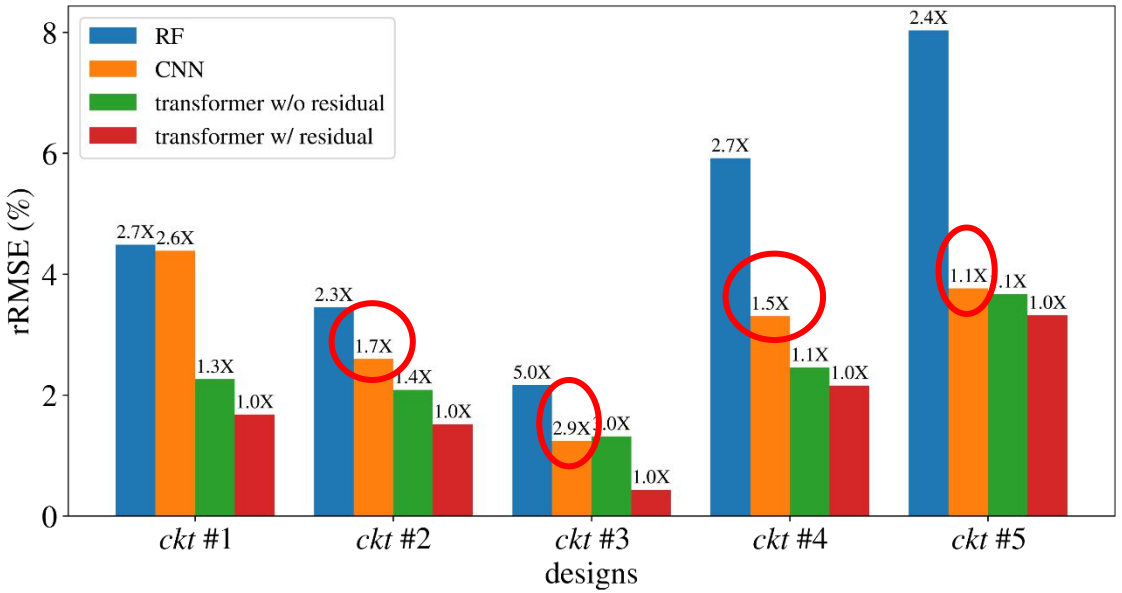
Error distribution of different models on seen and unseen designs.

	Seen ckt	Unseen ckt
rRMSE	<1.68%	<3.12%
Compared with RF, reduced by	2.3X~5X	2.4X~2.7X
Compared with CNN, reduced by	1.7X~2.9X	1.1X~1.5X
Residual model benefits	>30%	>10%

Results

SOUTHEAST UNIVERSITY

Accuracy Comparison:

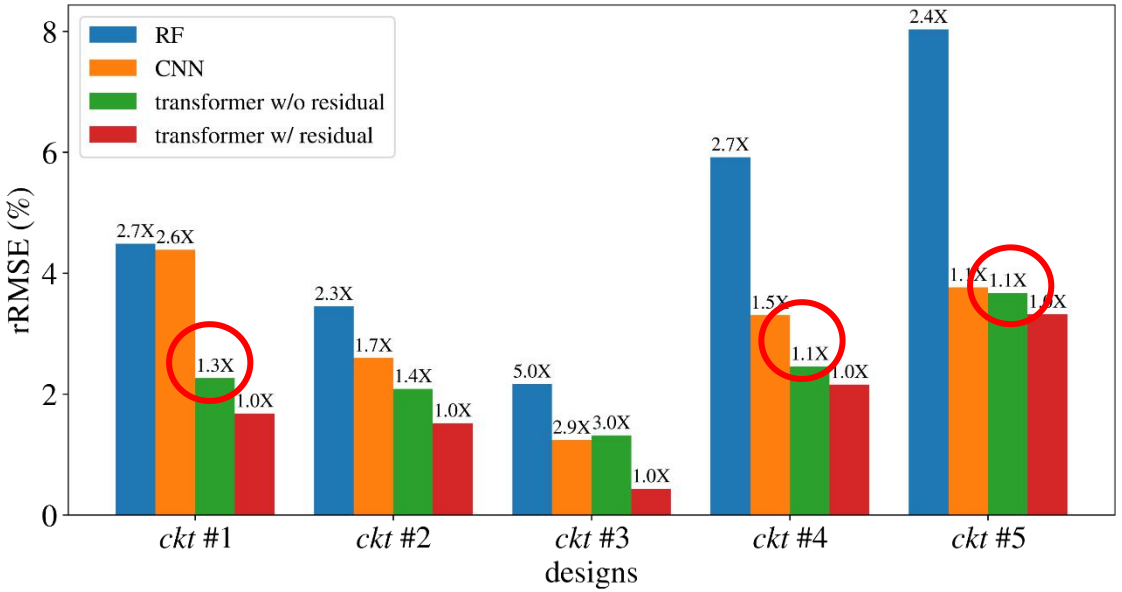


Error distribution of different models on seen and unseen designs.

	Seen ckt	Unseen ckt
rRMSE	<1.68%	<3.12%
Compared with RF, reduced by	2.3X~5X	2.4X~2.7X
Compared with CNN, reduced by	1.7X~2.9X	1.1X~1.5X
Residual model benefits	>30%	>10%

Results

Accuracy Comparison:

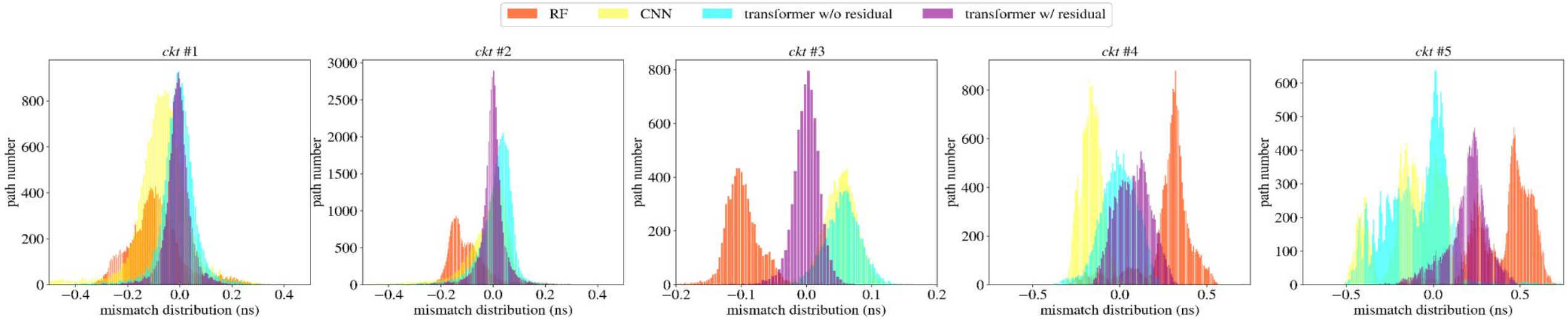


Error distribution of different models on seen and unseen designs.

	Seen ckt	Unseen ckt
rRMSE	<1.68%	<3.12%
Compared with RF, reduced by	2.3X~5X	2.4X~2.7X
Compared with CNN, reduced by	1.7X~2.9X	1.1X~1.5X
Residual model benefits	>30%	>10%

Results

Accuracy Comparison:



Mismatch distribution of different models.

Results

SOUTHEAST UNIVERSITY 東南大學

Runtime Comparison:

Runtime analysis

Model		Prediction Runtime (s)				
		ckt #1	ckt #2	ckt #3	ckt #4	ckt #5
Traditional IC flow	CTS	1378	31896	109	193	620
	Routing	1818	411968	143	254	816
	STA (PT)	655	1608	276	680	951
	Total	3851	445472	528	1127	2387
RF		11.2	26.5	10.6	15.7	28.4
CNN		1.28	2.68	0.57	1.22	1.66
This work		1.02 (3775X)	2.16 (206237X)	0.40 (1320X)	1.02 (1105X)	1.38 (1730X)

Conclusion

SOUTHEAST UNIVERSITY 東南大學

An efficient and accurate pre-routing path delay prediction framework is proposed in this work by employing transformer network and residual model.

- Transformer network: exploits the correlations of the timing and physical information through circuit path by its multi-head self-attention mechanism
- Residual model: calibrate the mismatch between the pre- and post-routing path delay
- More accurate and less runtime

東南大學
SOUTHEAST UNIVERSITY

THANKS
Q&A

Tai Yang, Guoqing He, Peng Cao

National ASIC System Engineering Technology Research Center, Southeast University, Nanjing, China

caopeng@seu.edu.cn

ASP-DAC2022