Testing for AI
AI for Testing

2024 中央大學演講

李建模

Self Introduction

- **1993 BSEE, NTU. 2002 PhD, Stanford.**
- **Research Interests**
  - VLSI testing & diagnosis, AI application to EDA
  - Testing AI chips, Testing quantum circuits
- **Youtube open course "VLSI Testing"**
  - 94 videos, 800K views
  - Course material download http://cc.ee.ntu.edu.tw/~cmli

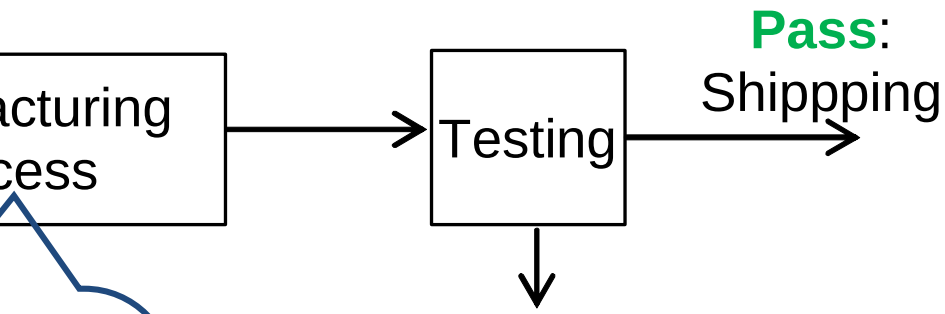
Outline

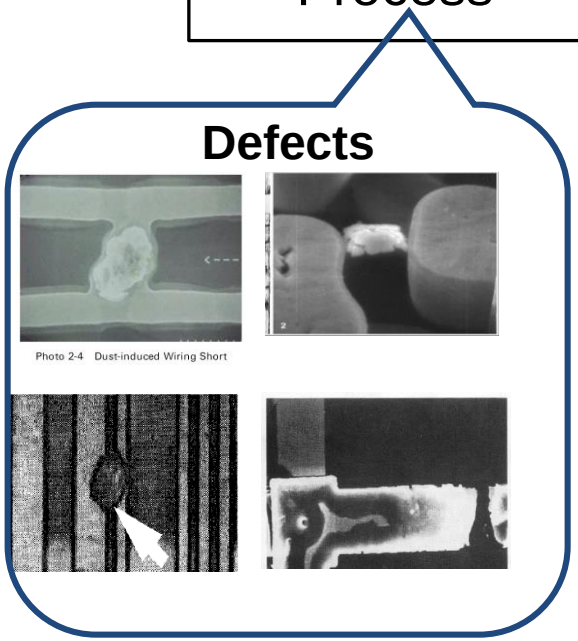
- **Introduction**
- **AI for Testing**
  - IR Drop
  - Thermal
  - V_{min}
- **Testing for AI**
  - Application-oriented Test
  - Manufacture-oriented Test
- **Conclusion**

Generated by ChatGPT

What is Testing?

- ***Testing* is process of determining whether a piece of hardware**
  - Functioning correctly (**PASS**) or defective (**FAIL**)
- **Why do we need to test Integrated Circuit (IC)?**
  - Because *defects* occur in manufacturing process





Testing is a Decision

Four Possible Outcomes

- **True pass and true reject are correct decision**
- ***Test escapes* = defective chips that pass test**
  - also known as (aka.) *under-testing*
- ***Yield loss* = good chips that fail the tests**
  - aka. *overkill*, *over-testing*
- **Goal of good testing: reduce both test escape and yield loss**
  - Trade off between test cost and test quality
    - Quality test reduces test escape but increases yield loss
    - Low cost test reduces yield loss but increase test escape

	Good IC	Defective IC
Pass tests	True PASS	Test Escapes (less is better)
Fail tests	Yield Loss (less is better)	True Reject

Quiz !

Q: Which of following is NOT IC testing?

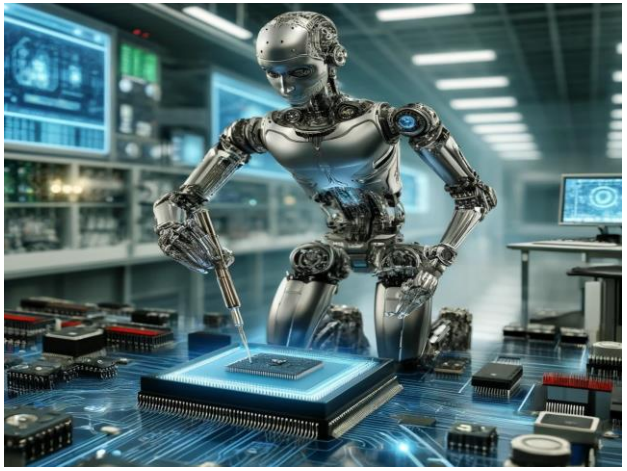
A: Run SPICE simulation on amplifier design to check if output is correctly amplified

B: Apply analog signal to an ADC IC and check if output is correctly digitized

C: Apply two numbers to an adder IC and check if output number is correctly added

Why AI is Useful for Testing?

- **1. AI is good at decision making**
  - Same job as testing
- **2. AI needs a lot of data**
  - Testing generates a lot of labeled data
- **3. AI is very fast**
  - Save test time= save test cost
- **4. AI also good at generating images**



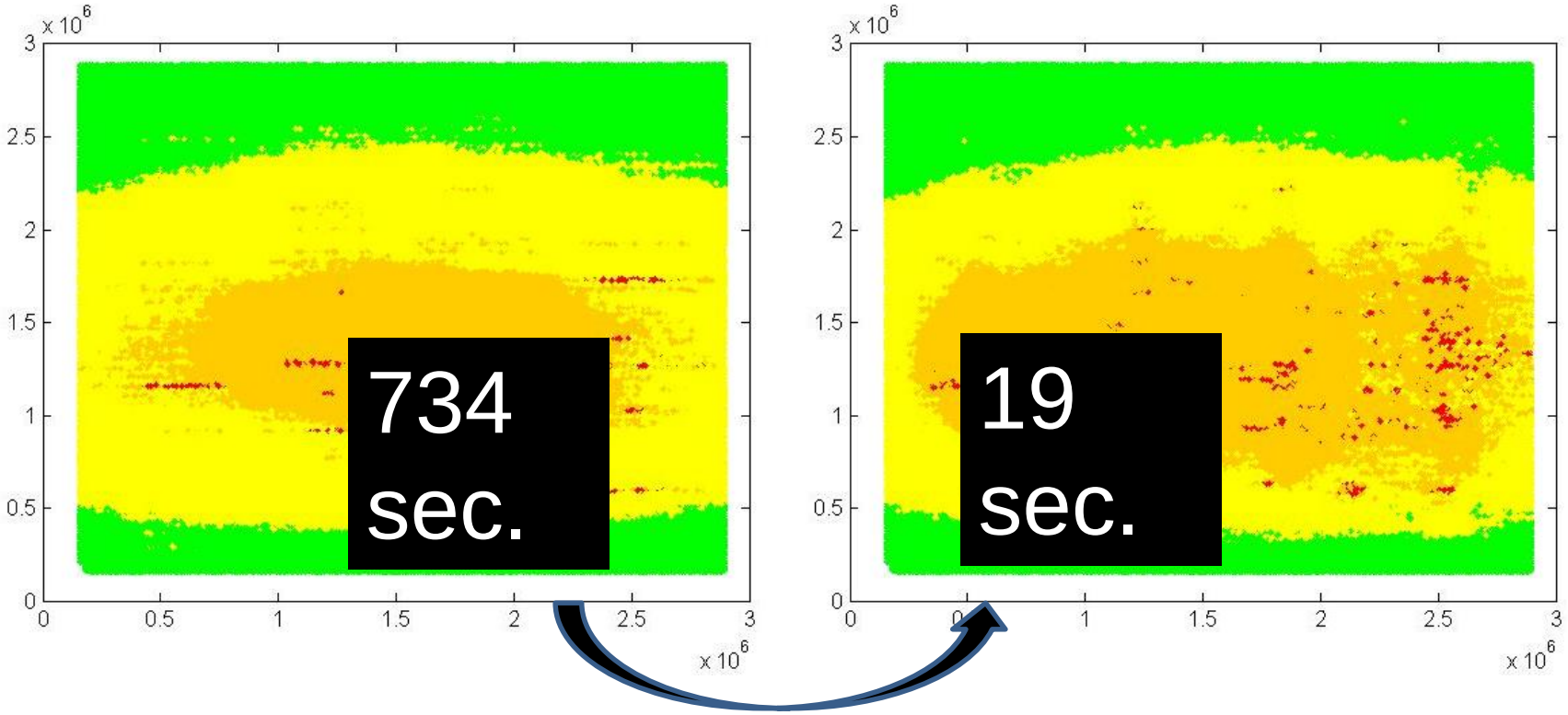
Generated by ChatGPT

Which Is AI generated?

https://www.whichfaceisreal.com

Which Is AI generated?

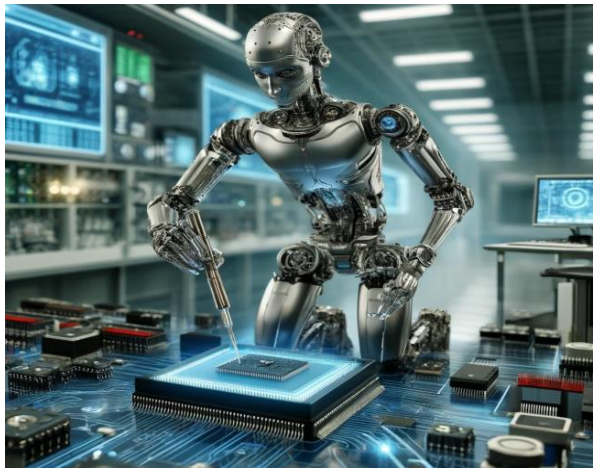
IR drop map



~40x speed up!

AI for Testing, Testing for AI

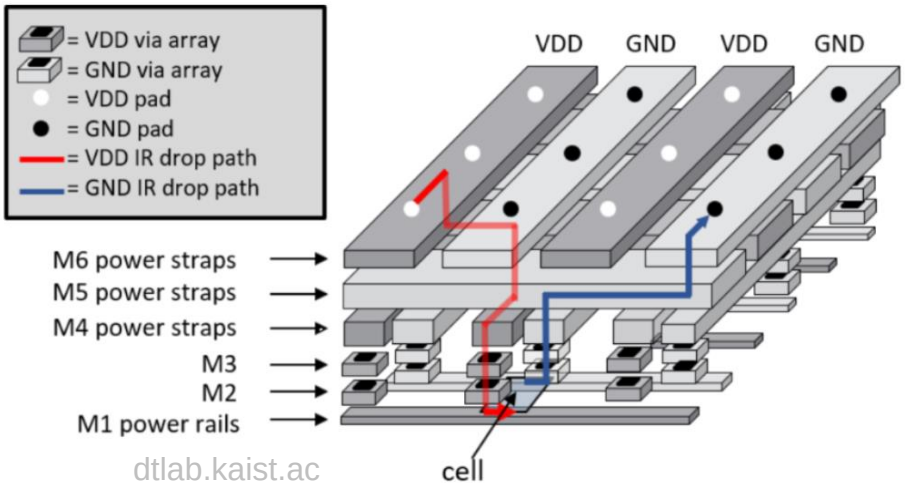
- Introduction
- **AI for Testing**
  - IR Drop [Fan ICCAD 18]
  - Thermal
  - V_{min}
- Testing for AI
  - Application-oriented Test
  - Manufacture-oriented Test
- Conclusion

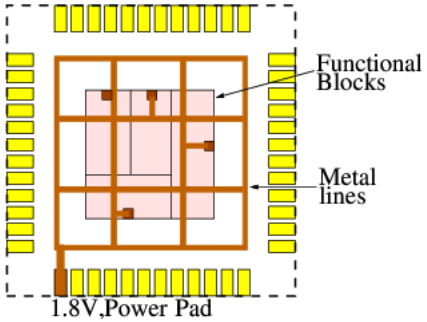


[Fan ICCAD 18] Yen-Chun Fang, Heng-Yi Lin, Min-Yan Su, James C.M. Li, Eric Jia-Wei Fang, “Machine-learning-based Dynamic IR Drop Prediction for ECO,” IEEE Int’l Conf. on CAD, 2018.

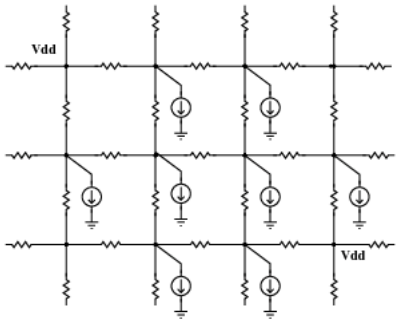
IR Drop

- **IR drop is important issue in VLSI testing**
  - Test power is typically higher than normal power
  - Lead to **over-testing**
    - Timing failure





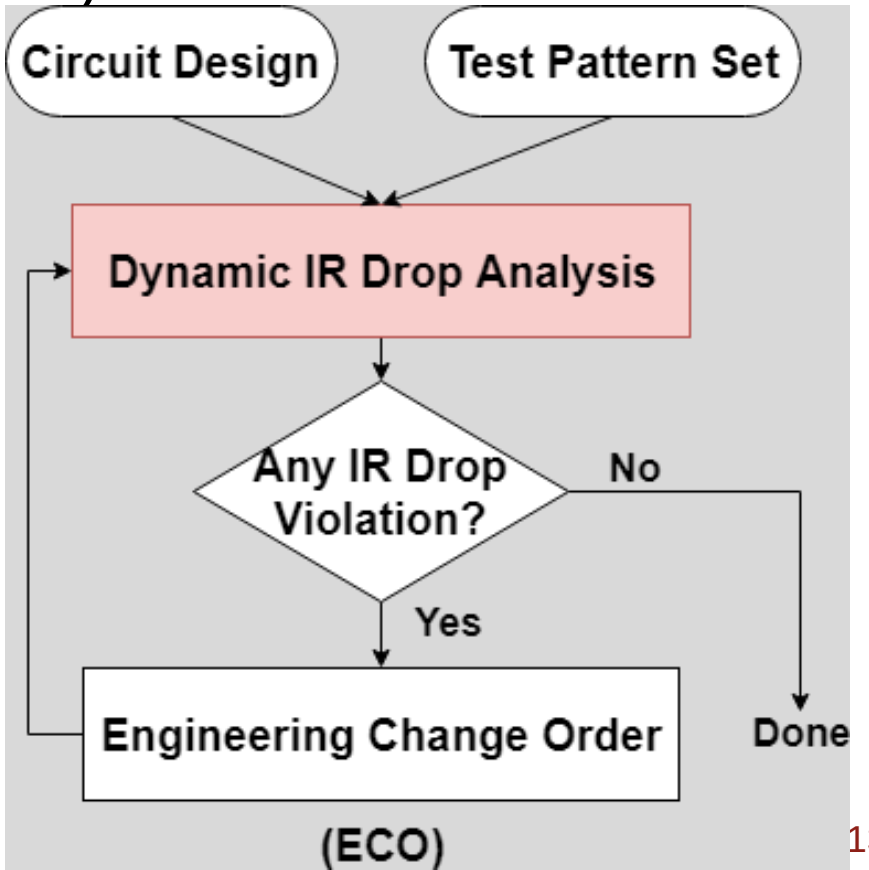
(a)



(b)

Traditional IR Drop Signoff Flow

- **IR drop analysis is needed in sign off**
- **Traditional IR drop analysis very slow**
  - 8hr for 5M cells design(2018)
- **What is worst**
  - Need many iterations



Machine-learning-based Dynamic IR Drop Prediction for ECO [Fan ICCAD 18]

- **Goal**
  - Reduce simulation time during ECO iterations
- **Contributions**
  - IR drop prediction using machine learning for ECO
    - Propose 17 kinds of feature extractions
    - Propose regional model
- **Key Results**
  - Mean absolute error is 3mV in high IR drop region
  - Runtime less than 2 minutes on 100K cell instances

[Fna ICCAD 18] Yen-Chun Fang, Heng-Yi Lin, Min-Yan Su, James C.M. Li, Eric Jia-Wei Fang, “Machine-learning-based Dynamic IR Drop Prediction for ECO,” IEEE Int’l Conf. on CAD, 2018.

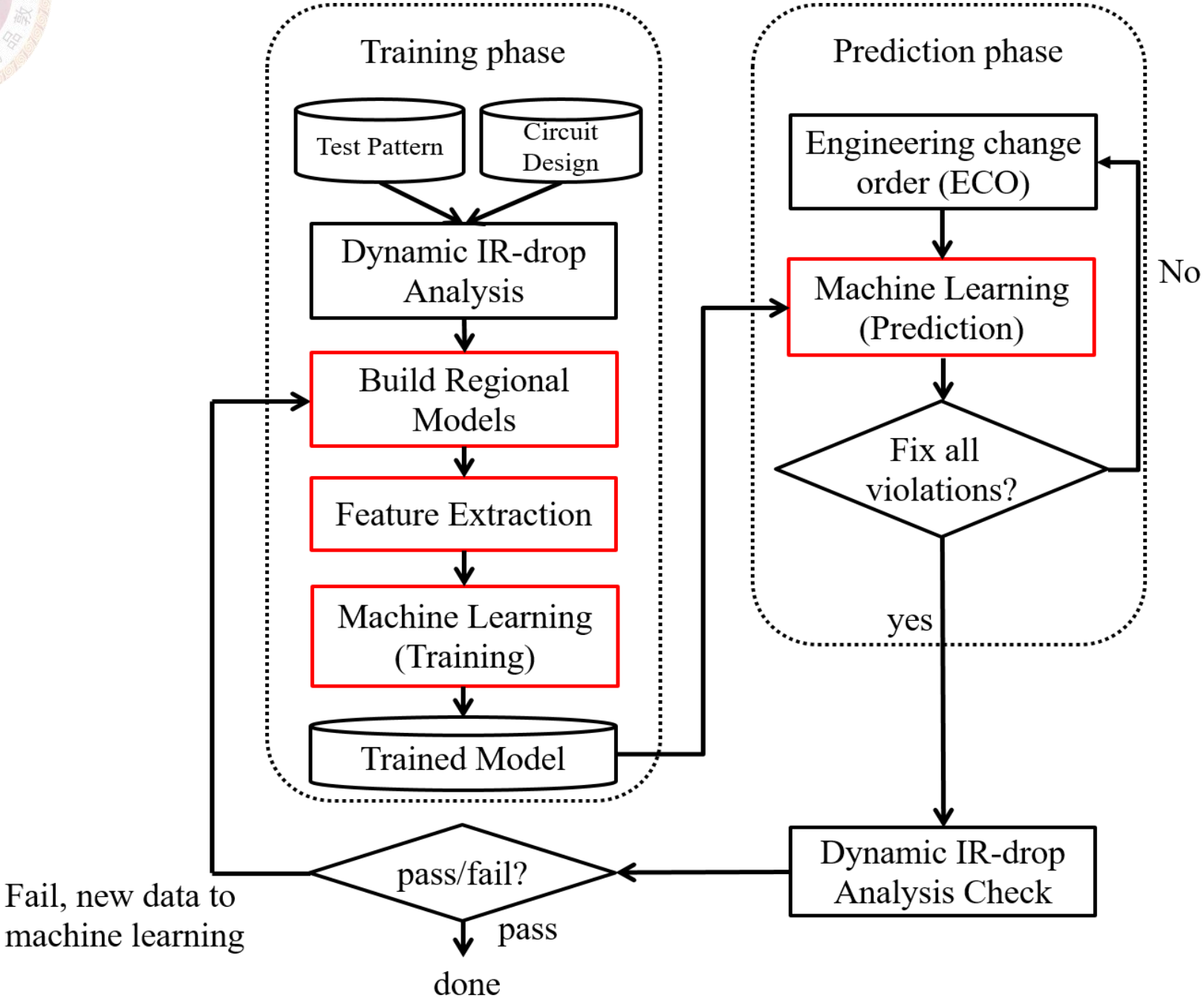
Machine Learning Techniques

- **Predict IR drop by linear model [Yamato 12]**
  - Predict actual IR drop value
  - Too many models
  - May not be good enough for complex design
- **Dynamic IR drop prediction [Lin 18]**
  - Feature extraction
  - More features should be considered

[Yamato 12] Yamato, Yuta, et al. "A fast and accurate per-cell dynamic IR-drop estimation method for at-speed scan test pattern validation." *Proc. Int'l Test Conf.*, IEEE, 2012. p. 1-8.
[Lin 18] Shih-Yao Lin, et al. "IR Drop Prediction of ECO-Revised Circuits Using Machine Learning" *VLSI Test Symp.*, 2018.

Outline

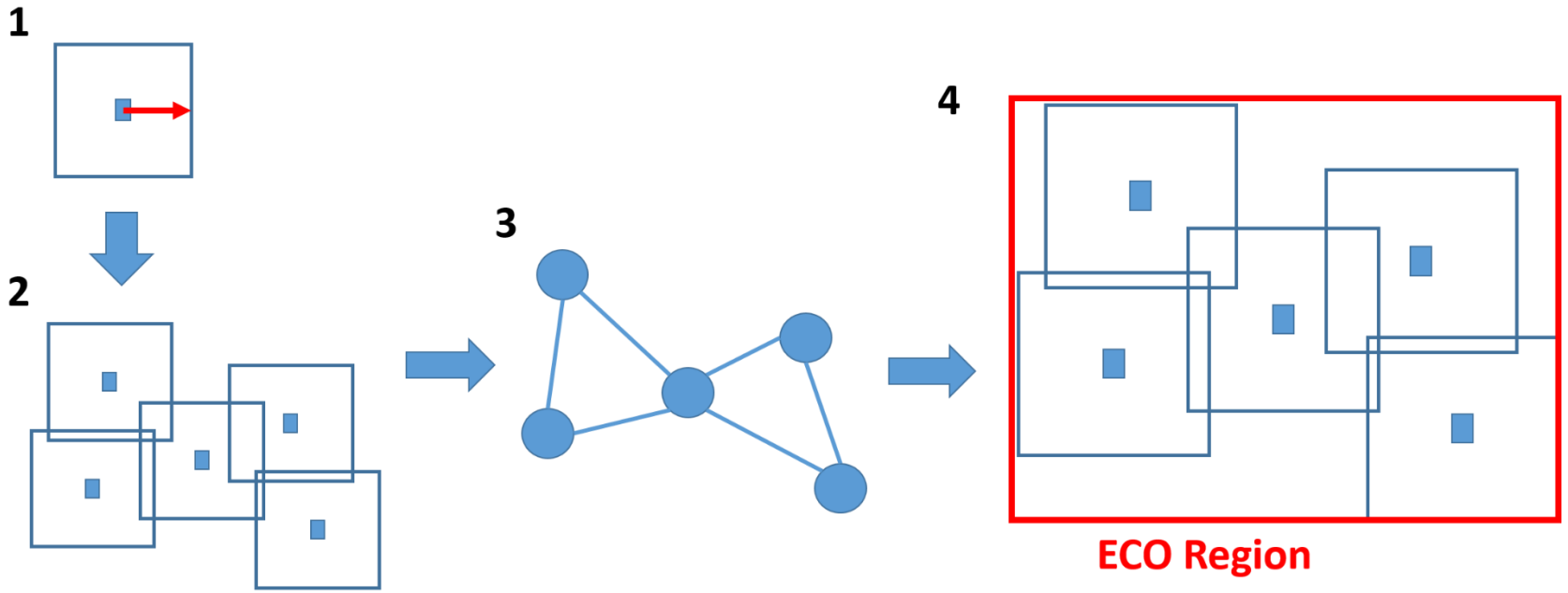
- Introduction
- Past Research
- Proposed Techniques
  - Regional model
  - Feature extraction
  - Machine learning model
- Experimental Results
- Conclusion


Training phase
Test Pattern
Circuit Design
Dynamic IR-drop Analysis
Build Regional Models
Feature Extraction
Machine Learning (Training)
Trained Model
Prediction phase
Engineering change order (ECO)
Machine Learning (Prediction)
Fix all violations?
No
yes
Dynamic IR-drop Analysis Check
pass/fail?
Fail, new data to machine learning
pass
done

Regional Model

Define ECO region to build machine learning model

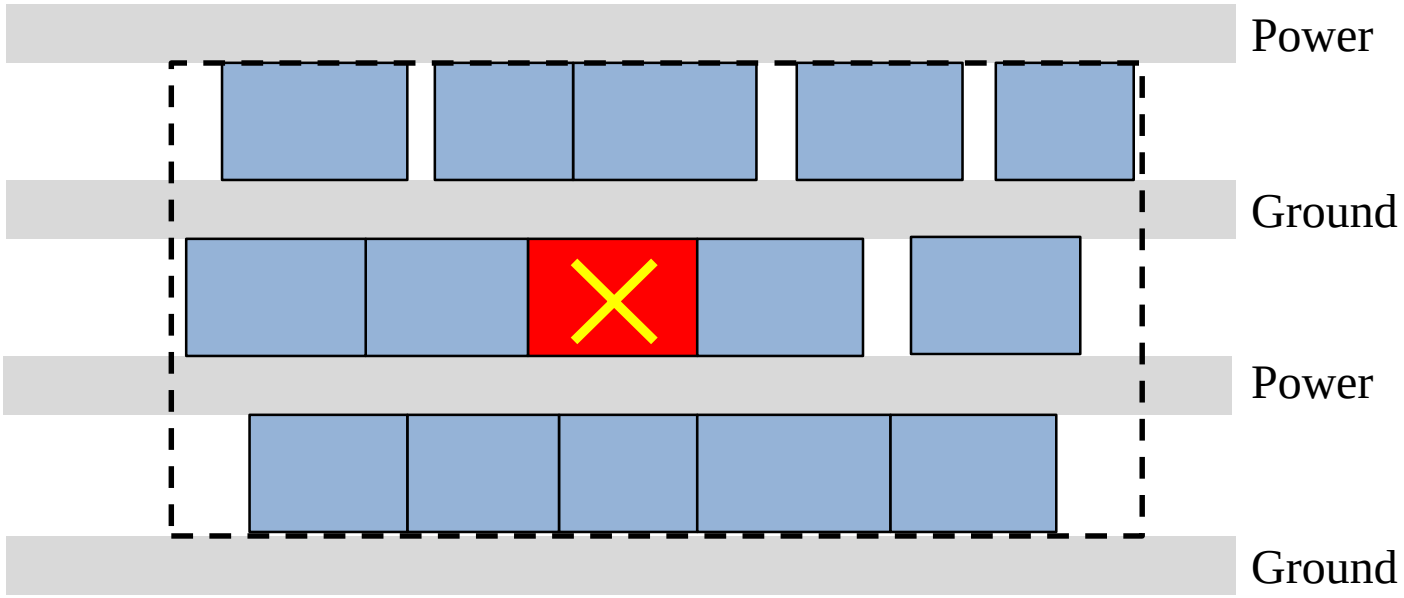
1. Define maximal cell move distance
2. Cover every IR drop violation by a square
3. Build a graph to describe adjacency
4. Build ECO region for connected component



Feature Extraction

- **Target cell features**
  - Power Features
  - Physical Features
  - Timing Features
- **Neighbor cell features**
  - Density Maps

Blue squares are neighbor cells



Outline

- Introduction
- Past Research
- **Proposed Techniques**
  - Regional model
  - **Feature extraction**
  - Machine learning model
- Experimental Results
- Conclusion

Power Features

- **Total power (P_{total})**
  - Switching power + Leakage power + Internal power
- **Toggle rate (*TR*)**

$$\text{Toggle rate} = \frac{\text{number of toggles}}{\text{number of clock cycles}}$$

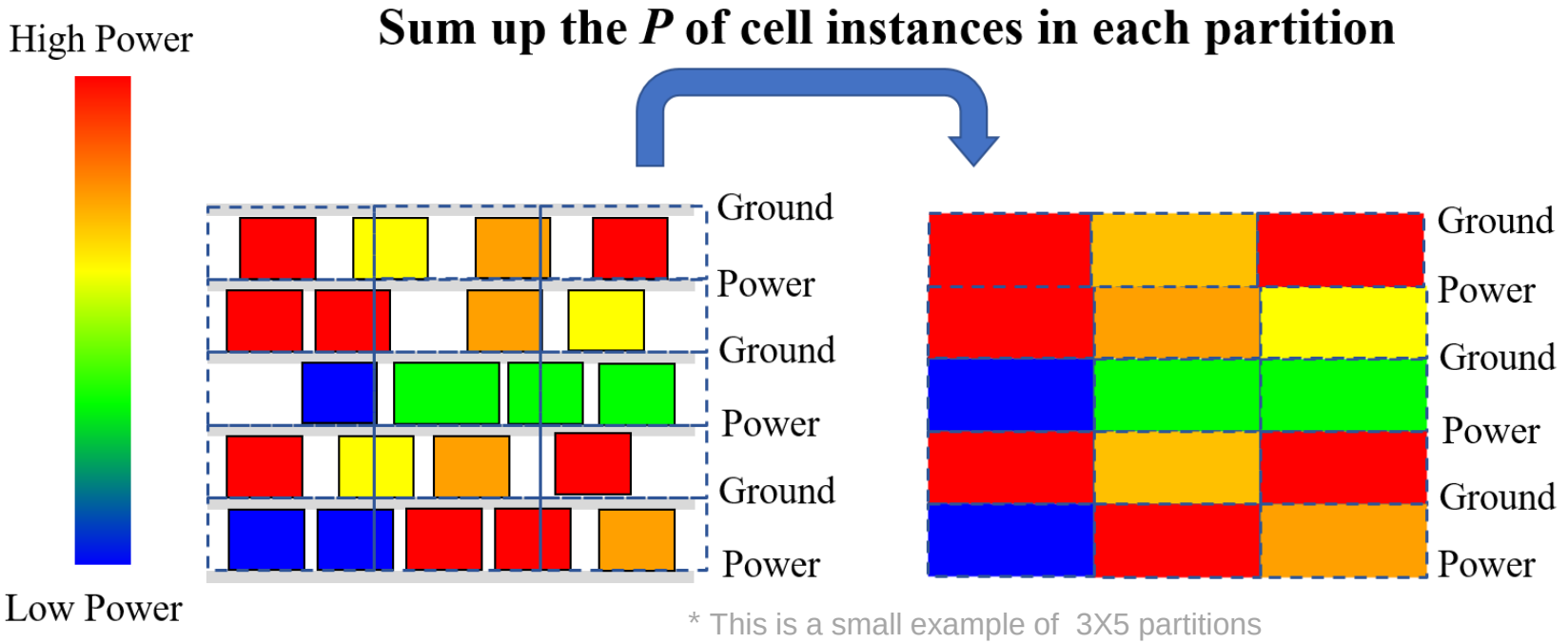
- **Cell type index**
  - e.g. 1:BUF, 2:INV, 3:AND
- **Load capacitance**
- **Peak current (I_{peak}), Avg. current (I_{avg})**

Physical & Timing Features

- **Physical location**
  - X,Y coordinate
- **Path resistance (*R*)**
  - The resistance from power bump to target instance
- **Timing window**
  - Max/min rising/falling time from STA report

Neighbor Features

- **Transform circuit layout to density map**

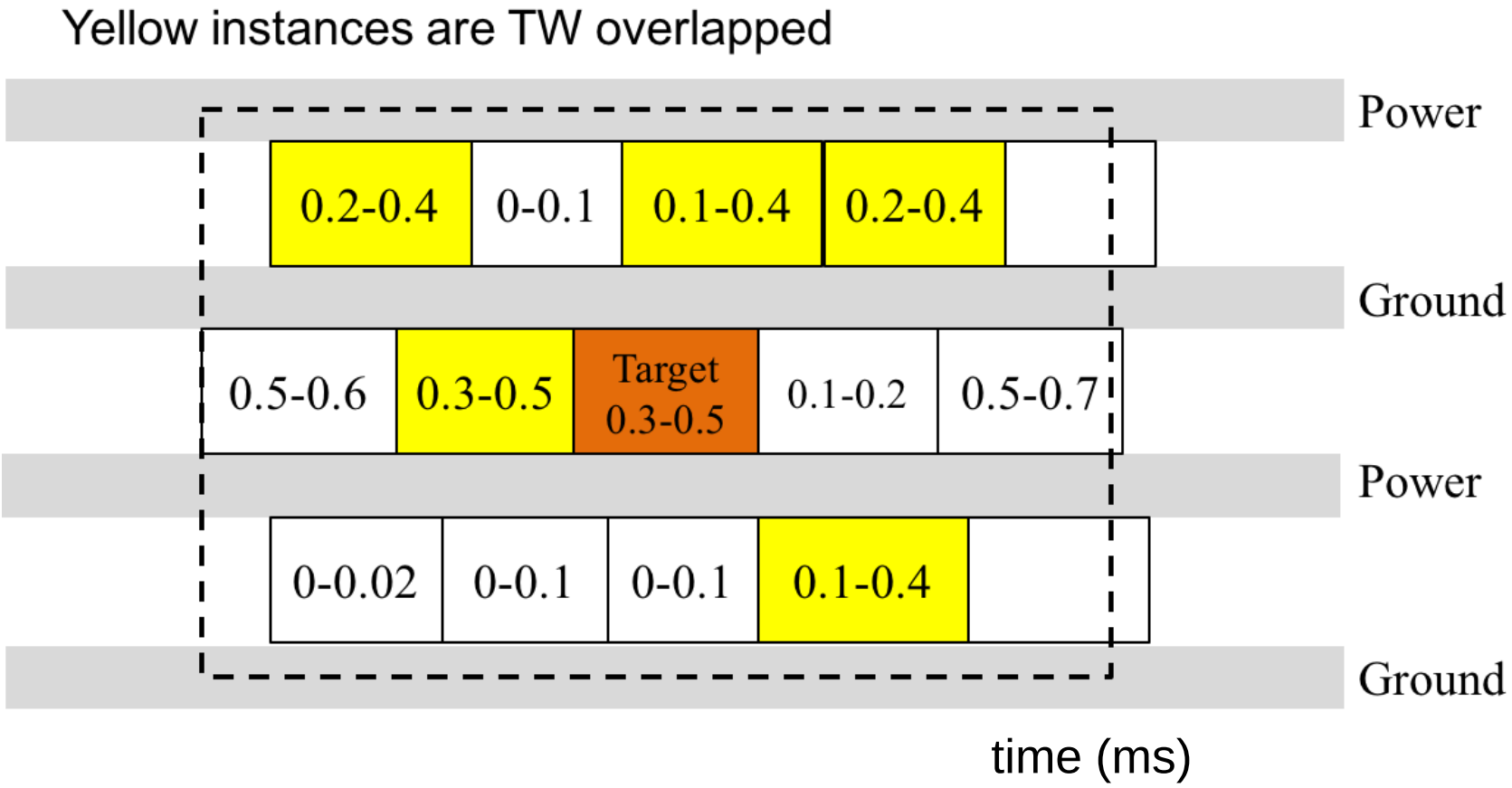


* This is a small example of 3X5 partitions

- 21X11 partitions
- Partition: 2μmX1row height

Neighbor Features

- 4 density maps for P_{total}, TR, I_{avg}, I_{peak}

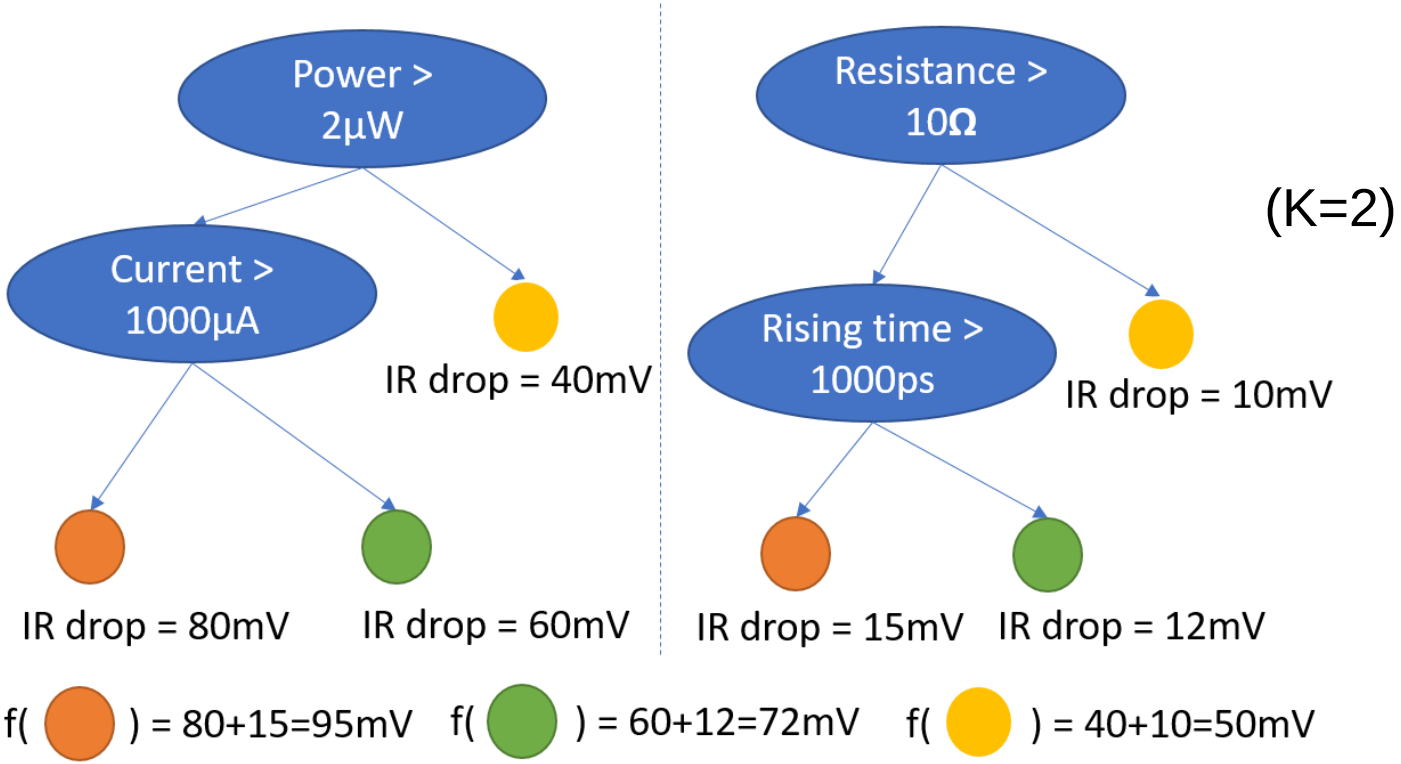


Outline

- Introduction
- Past Research
- **Proposed Techniques**
  - Regional model
  - Feature extraction
  - **Machine learning model**
- Experimental Results
- Conclusion

XGBoost [Chen 16]

- Ensembles *K* additive functions
- 3-fold cross-validation to find the optimal K



[Chen 16] T. Chen et, al "Xgboost: A scalable tree boosting system." *ACM Int'l Conf. on Knowledge Discovery and Data Mining*

Outline

- Introduction
- Past Research
- Proposed Techniques
- **Experimental Results**
- Conclusion

Benchmarks

- Two industry designs

Circuit Design	Design 1	Design 2
VCD	Unavailable	Available
Num. of cell inst.	500K	5,000K
Num. of ECO region	1	15
V_{DD} (V)	0.85	0.8
Mean IR drop (mV)	17.9	10.4
Max IR drop (mV)	44.9	152.8

Evaluate Metrics

Normalized Root Mean Square Error (NRMSE)

- $NRMSE = \frac{RMSE}{mean(\hat{y})} \times 100\%$ $\quad RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$

Correlation coefficient (*CC*)

0~1 higher means better coorelation

- $CC = \frac{\sum_{i=1}^{N}[y_i - mean(y)][\hat{y}_i - mean(\hat{y})]}{\sqrt{\sum_{i=1}^{N}[y_i - mean(y)]^2 \ \sum_{i=1}^{N}[\hat{y}_i - mean(\hat{y})]^2}}$

Mean Absolute Error (MAE)

unit: mV larger means worse prediction

- $\mathrm{MAE} = \frac{\sum_{i=1}^{N}\|\hat{y}_i - y_i\|}{N}$

Max Error (MaxE)

unit: mV larger means worse prediction

- $\mathrm{MaxE} = \max(\hat{y}_i - y_i), \ i = 1 \text{ to } N$

$\hat{y}_i$: golden IR drop
y_i : predicted IR drop
N : number of data points

IR Drop Prediction for Design 1

- **Only 1 ECO Region in Design 1**
- **70% cell instances for training**
- **30% cell instances for testing**
- **XGBoost (XGB) outperform CNN (convolutional neural network)**

Size (μm x μm)	1,200 X 1,000	
Model	XGB	CNN
MAE (mV)	**0.54**	0.77
MaxE (mV)	**13.5**	24.2
CC	**0.99**	0.98
NRMSE	**5.1%**	6.5%

IR Drop Prediction for Design 2

- **Regional model better than global model**
  - Global model: Training time: 4hr
  - Regional model: Training time: 1.5hr

Region	Global Model		Region 1		Region 2		Other Regions	
Size	1,583x2,615		822x114		311x349		-	
Mean IR Drop (mV)	23.4		25.5		24.3		21.5	
Max IR Drop (mV)	141.6		141.6		105.1		137	
Model	XGB	CNN	XGB	CNN	XGB	CNN	XGB	CNN
MAE (mV)	4.3	5.5	**3.0**	3.6	**2.7**	3.1	2.5	3.6
MaxE (mV)	91.9	110	62.8	58.8	**48.3**	58.9	56.8	74.6
CC	0.94	0.92	**0.97**	0.96	**0.97**	0.97	0.97	0.95
NRMSE(%)	36.2	44.1	**22.2**	25	**23.5**	26.0	24.0	30.6

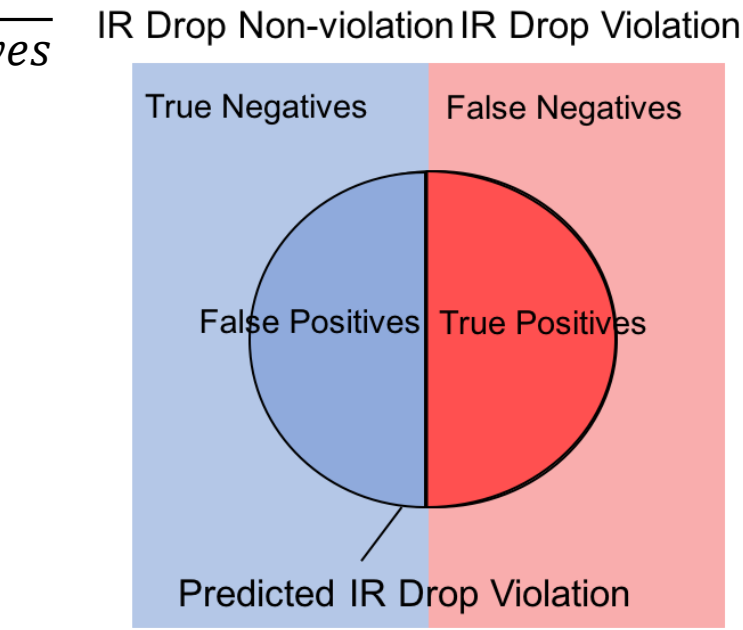
Precision & Recall

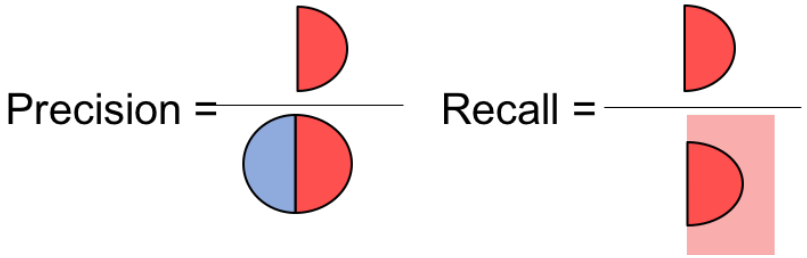
Precision

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall

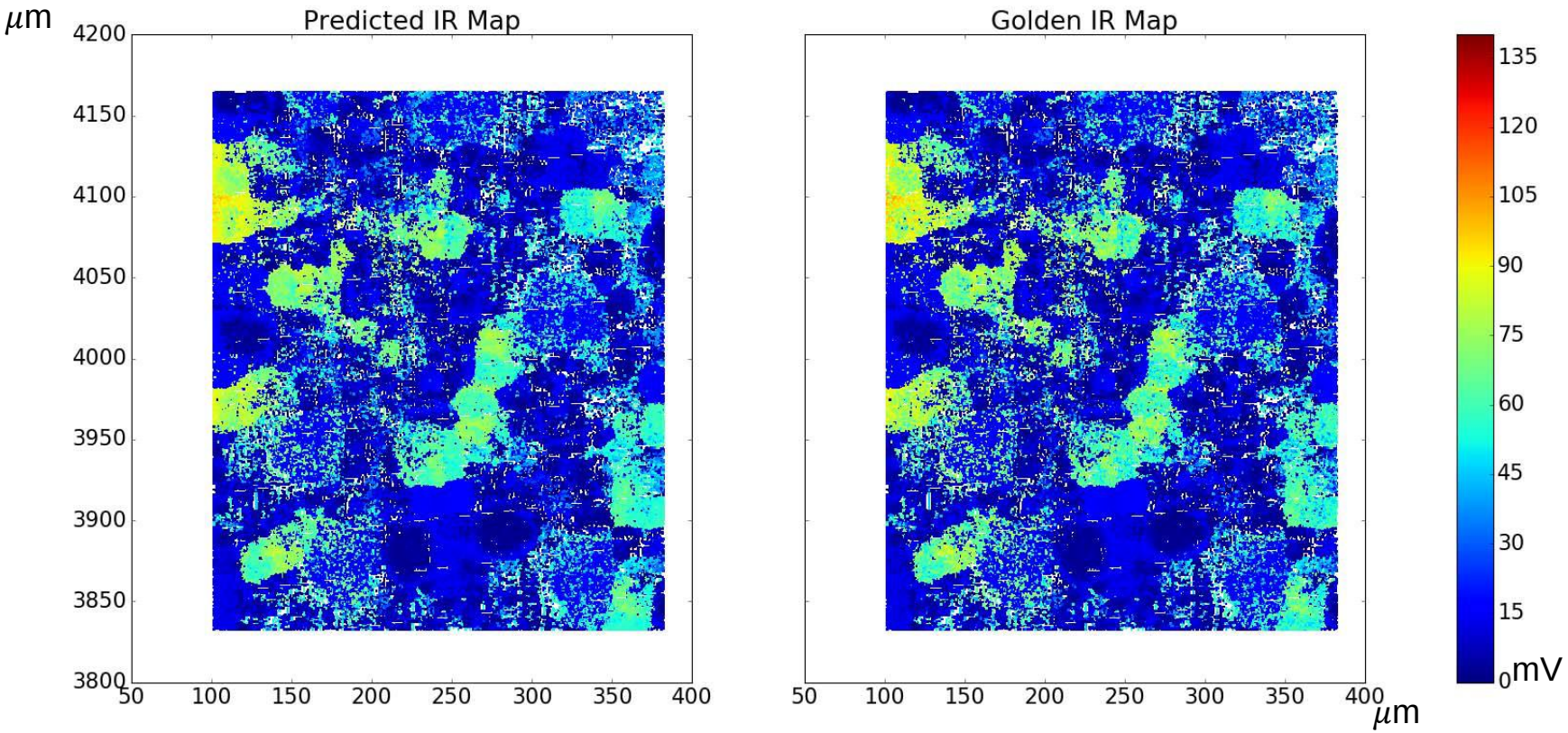
$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$





IR Drop Map Comparison

- **Similar result on Region2 of Design 2**



Runtime Comparison

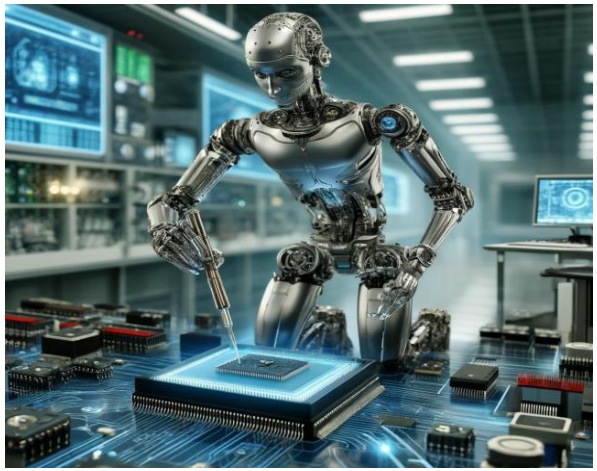
Circuit Design	Design 1	Design 2
Parsing Data	33s	11m19s
Build Regional Model	-	3m
Feature Extraction	2m24s (500K instances)	3m14s (650K instances)
Training	16m40s	1h3m
Total training time	20m	1h21m
Feature Extraction of 100K instances	30s	40s
Prediction of 100K instances	41s	44s
Total prediction time	**1m11s** (100K instances)	**1m24s** (100K instances)
Commercial tool	**42m** (500k instances)	**8h** (5,000k instances)

Conclusion

- **Machine-learning-based IR drop predictor**
  - Consider timing, power and physical information
  - Density map to encode neighbor cell's feature
  - Regional model better than global model
- **Prediction CC=0.97 and Mean Absolute Error=3mV**
- **Prediction time > 7 times faster than commercial tool**
- **All materials available on 教育部ATP課程教材資料庫**

AI for Testing, Testing for AI

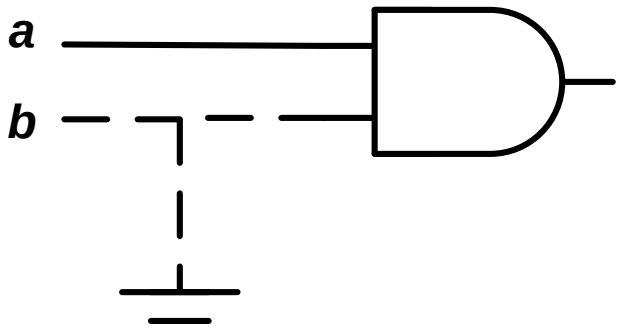
- Introduction
- **AI for Testing**
  - IR Drop [Fang ICCAD 18]
  - Thermal [Liang ITC 23]
  - V_{min}
- Testing for AI
  - Application-oriented Test
  - Manufacture-oriented Test
- Conclusion



[Liang ITC 23] Z.-J. Liang et al “High-Speed, Low-Storage Power and Thermal Predictions for ATPG Test Patterns”, IEEE Int’l Test Conf. 2023

Fault Models and Test Generation

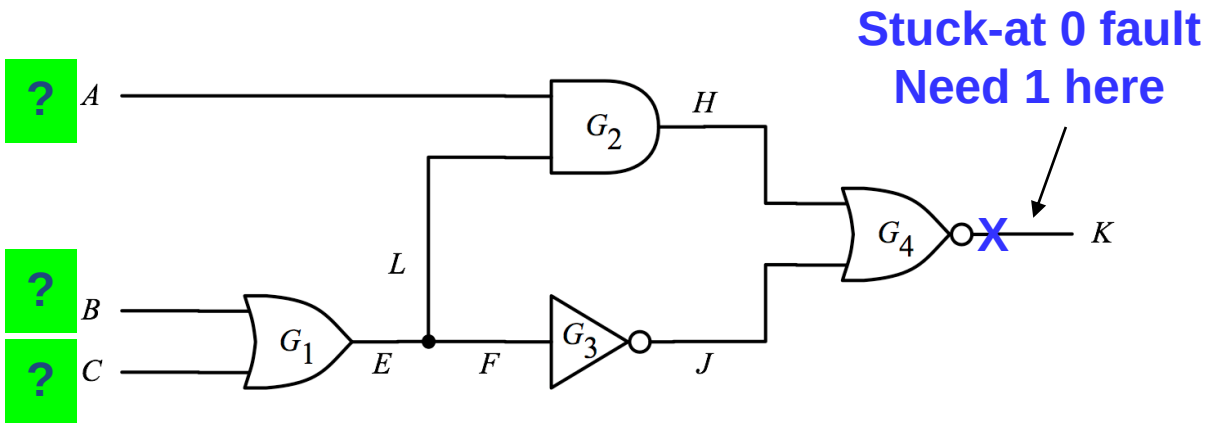
- ***Fault model***
  - High-level representation of defects
- ***Test patterns,* also known as (aka): *Test Vectors***
  - Input Boolean values for specific fault
    - Expected output often included (but not required)
  - Example: b stuck-at zero fault
    - Test pattern a=1, b=1
- ***Test Length* = Number of test patterns**
  - Example: ab= {11, 00}, TL=2
- **ATPG (automatic test pattern generator)**



Shorter Test Length = Lower Test Cost

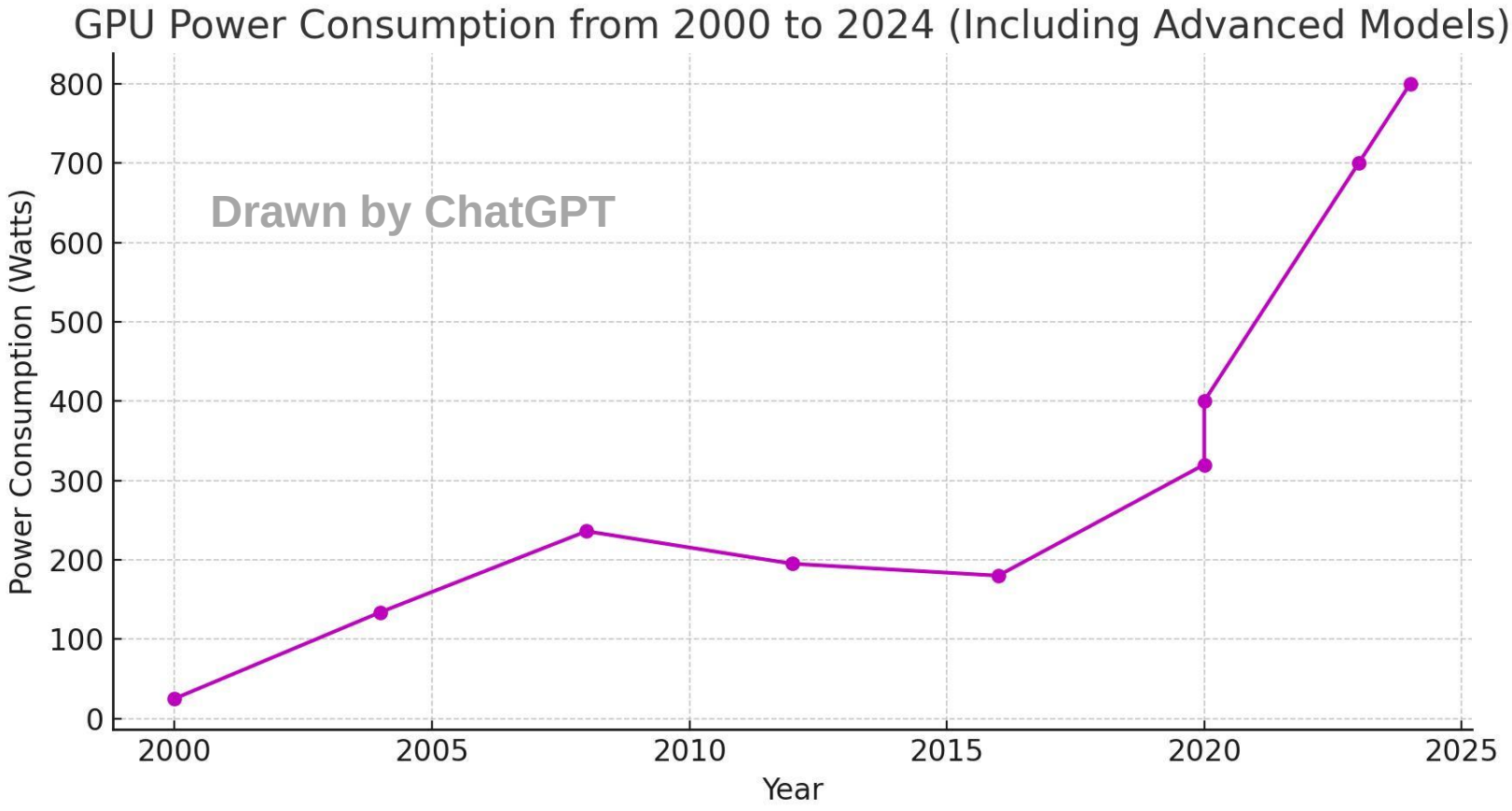
Quiz

- **Your manger asks you to generate a test pattern for output stuck-at zero fault. What is your answer? (Maybe more than one answer)**



- **Test power is typically much higher than normal power**
  - because ATPG test patterns are random

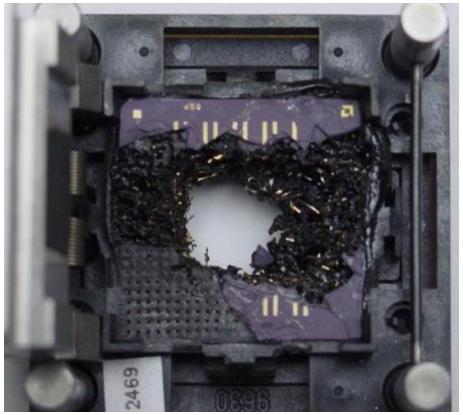
nVidia GPU Power Growth



Motivation

- **Thermal issue is important for testing**
  - High test power cause thermal damage [Intel 21]
- **Cannot finish analysis to ensure thermal safety**
  - Too many shift cycles: long runtime, large disk storage

Analysis	Cost	Design
Gate-level simulation	**15hour 13GB**	**206k cells** design 500 patterns 281 shift cycles per pattern
Power analysis	**2 days**	
Thermal analysis	**2 days**	**4mm x 4mm** design 10μm x 10μm tile 0.57s test time

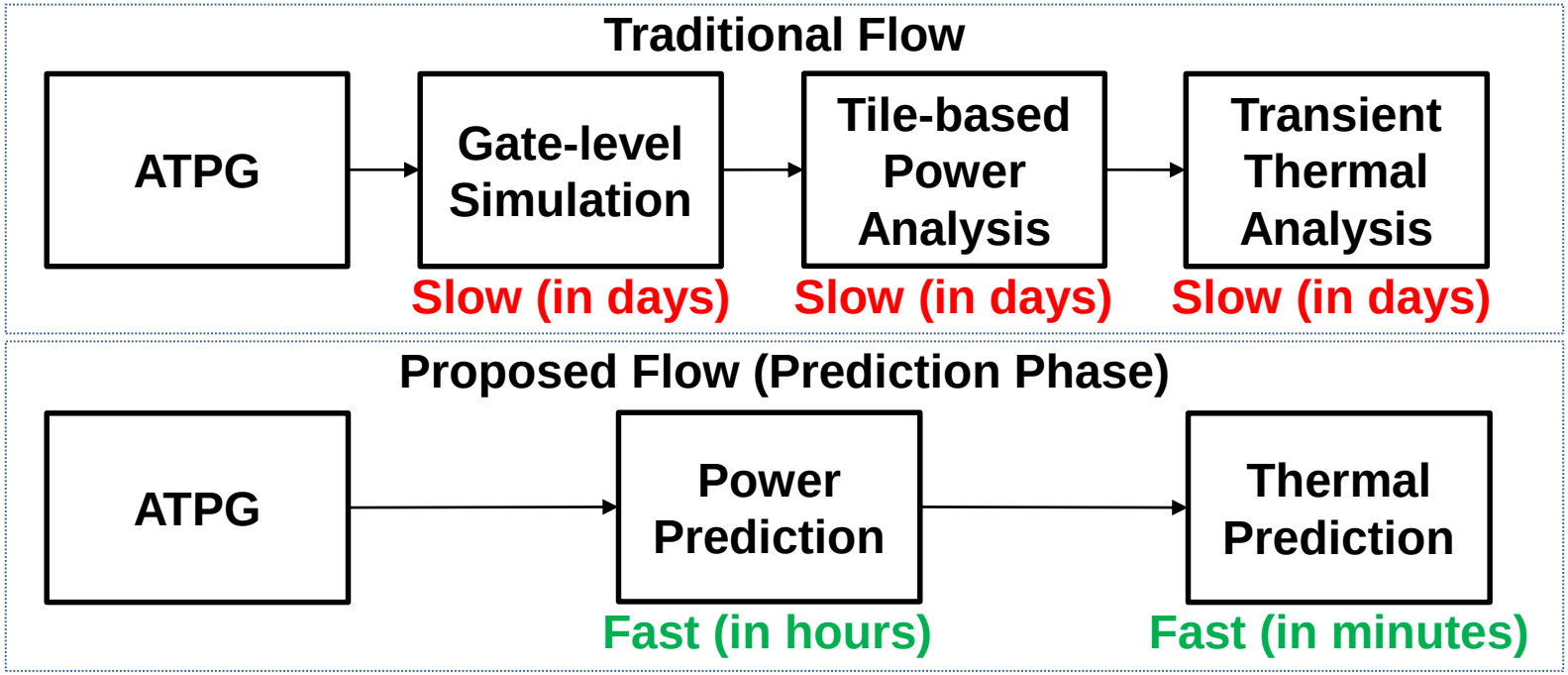


Burned chips during testing [Miller 01]

[Intel 21] Too Hot to Test workshop, Intel, 2021, https://youtu.be/0gPSbZqbXUg
[Miller 01] M. Miller, "Next generation burn-in & test systems for Athlon microprocessors: Hybrid burn-in", Proc. Burn-in and Test Socket Workshop Session 5, 2001.

High-Speed, Low-Storage Power and Thermal Predictions for ATPG Test Patterns [Liang ITC 23]

- **Enable thermal analysis for ATPG test patterns**
  - Solve runtime and disk storage issue by predictions

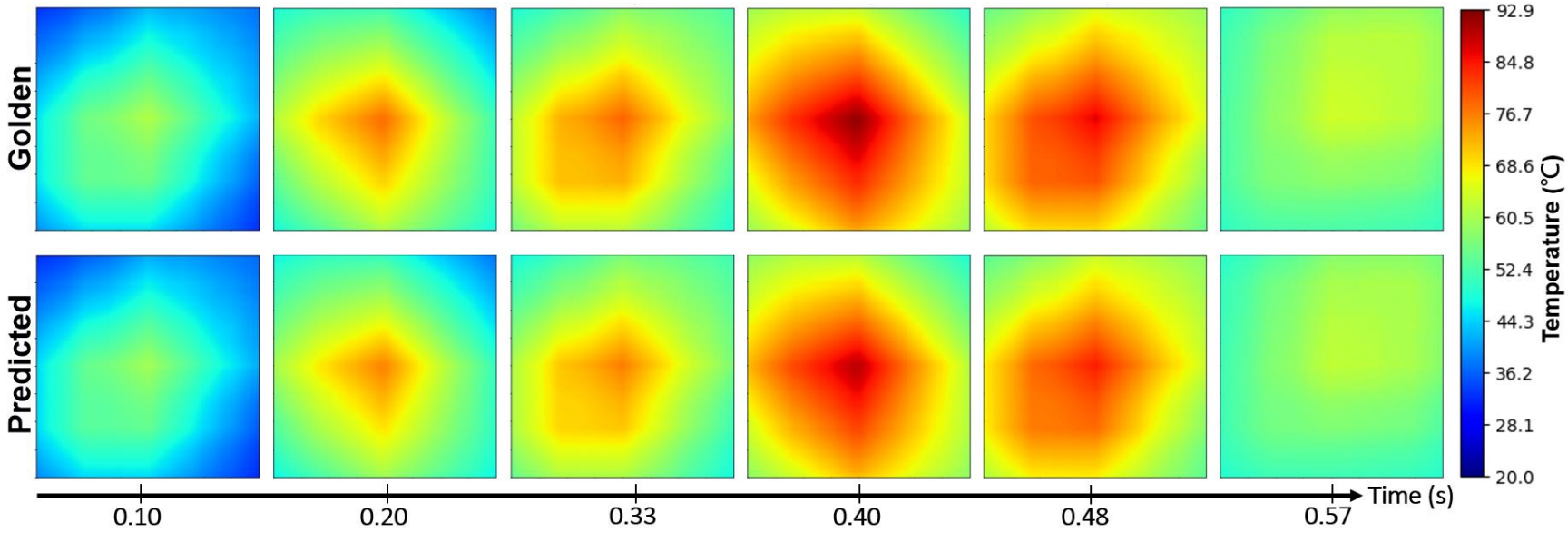


ATPG: automatic test pattern generation

[Liang ITC 23] Z.-J. Liang et al "High-Speed, Low-Storage Power and Thermal Predictions for ATPG Test Patterns", IEEE Int'l Test Conf. 2023

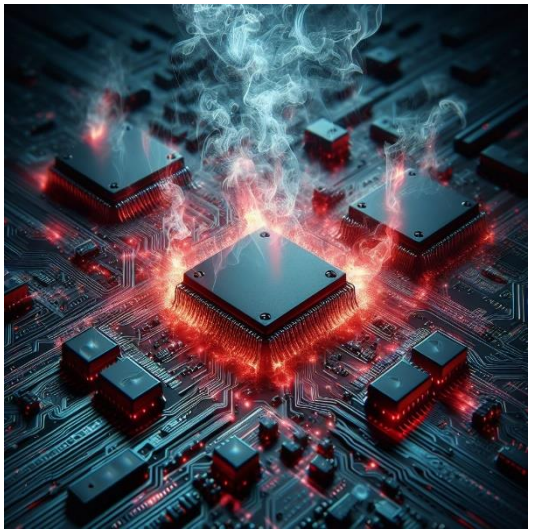
Key Results

- **Good accuracy, high speed, and low storage**
  - Mean absolute percentage error 8% (dynamic power)
  - Mean absolute error 1.2℃ (temperature)
  - 75X runtime speedup, 118X disk storage reduction



Outline

- Introduction
- **Background**
- Proposed Technique
- Experimental Results
- Conclusion



Generated by Copilot

Previous Works
Power Prediction

- **PRIMAL: Power Inference using Machine Learning [Zhou 19]**
  - Predict total power from register values
  - Thermal analysis needs tile-based power (power map)
- **Machine Learning-based Prediction of Test Power [Dhotre 19]**
  - Predict gate-based power from test patterns
  - Gate-based is inefficient compared to tile based

Existing techniques not work for thermal analysis

[Zhou 19] Y. Zhou et al., "PRIMAL: Power Inference using Machine Learning," DAC, 2019
[Dhotre 19] H. Dhotre et al., "Machine Learning-based Prediction of Test Power," ETS, 2019

Previous Works
Thermal Prediction

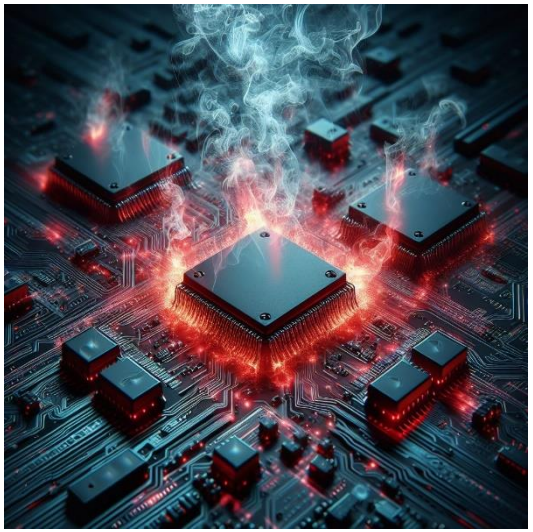
- **Full-Chip Thermal Map Estimation for Commercial Multi-Core CPUs with Generative Adversarial Learning [Jin 20]**
  - Predict thermal maps from performance metrics
  - Performance metrics is not available during testing
- **Thermal and IR Drop Analysis Using Convolutional Encoder-Decoder Networks [Chhabria 21]**
  - Predict thermal maps from power maps
  - Need power maps and thermal maps for training

Existing techniques not work for ATPG test patterns

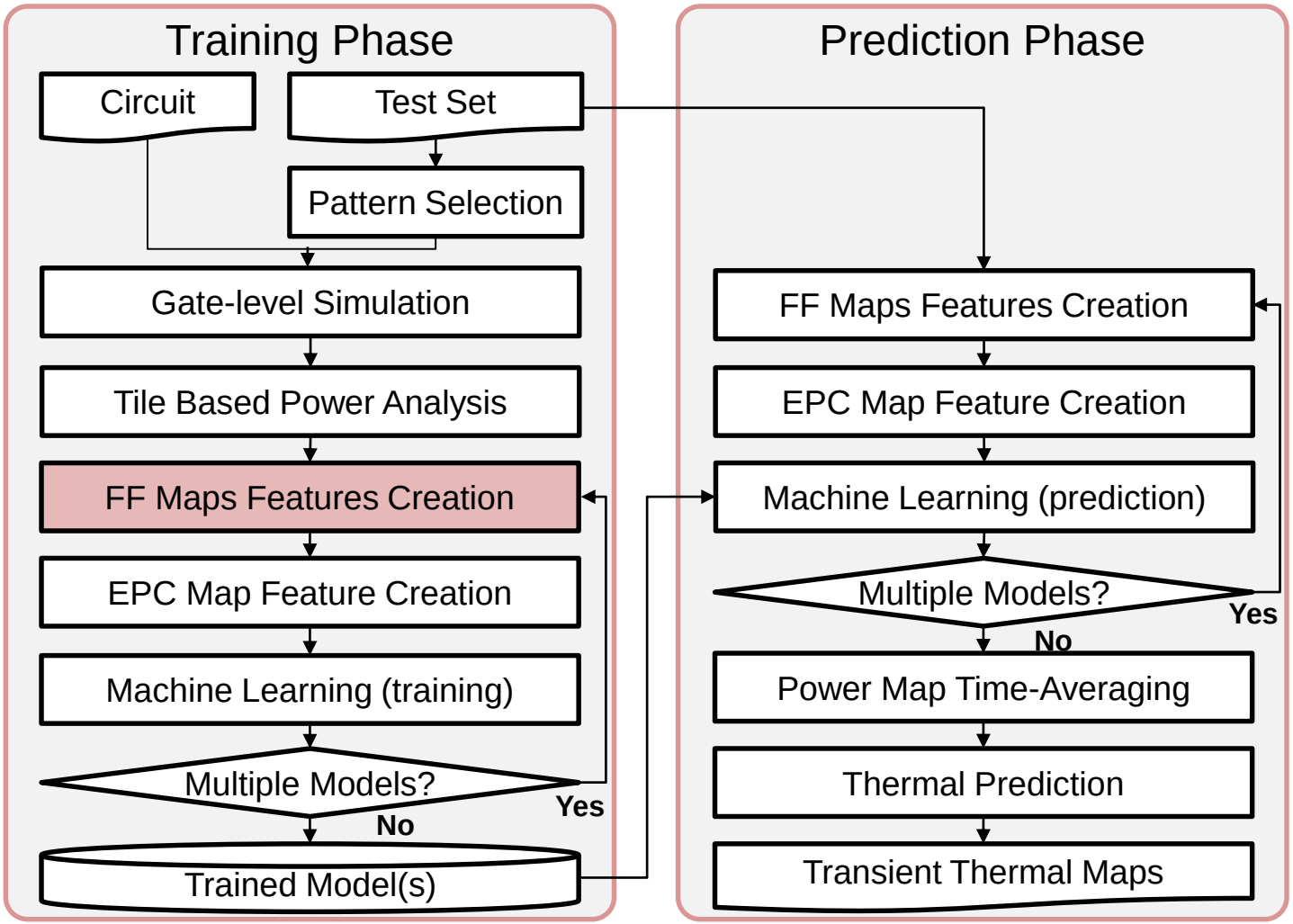
[Jin 20] W. Jin et. Al., "Full-Chip Thermal Map Estimation for Commercial Multi-Core CPUs with Generative Adversarial Learning," ICCAD, 2020
[Chhabria 21] V. A. Chhabria et al., "Thermal and IR Drop Analysis Using Convolutional Encoder-Decoder Networks," ASP-DAC, 2021

Outline

- Introduction
- Background
- **Proposed Technique**
- Experimental Results
- Conclusion



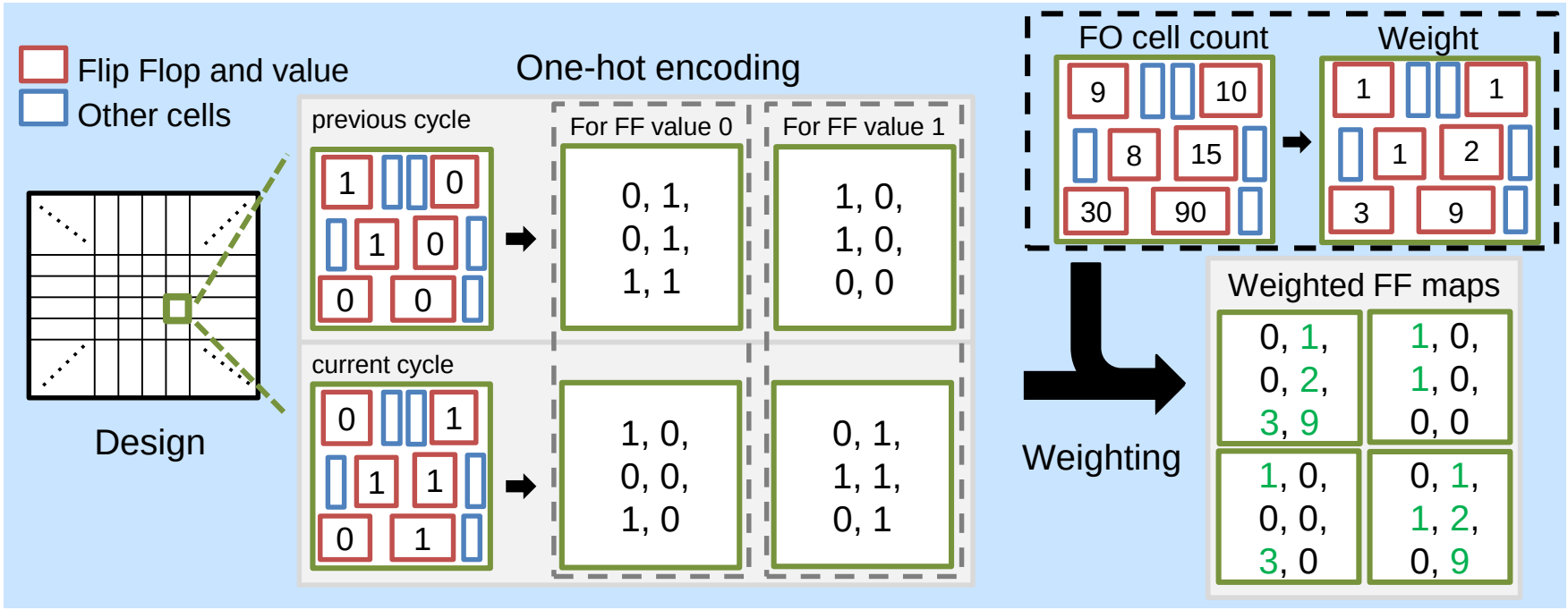
Overall Flow



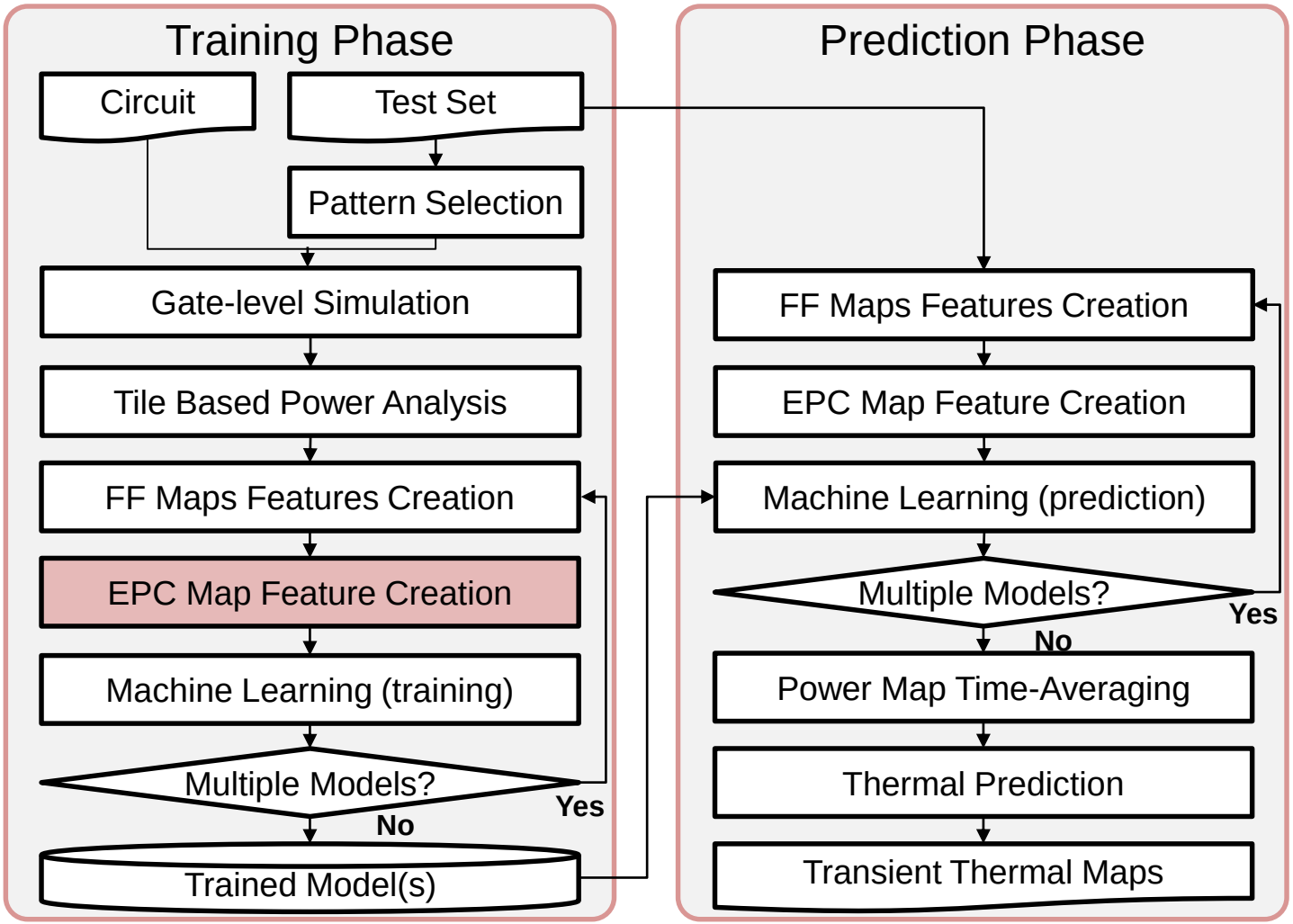
Flip Flop (FF) Maps

Encode and weight FF values for two cycles

Represent FF values and their importance



Overall Flow



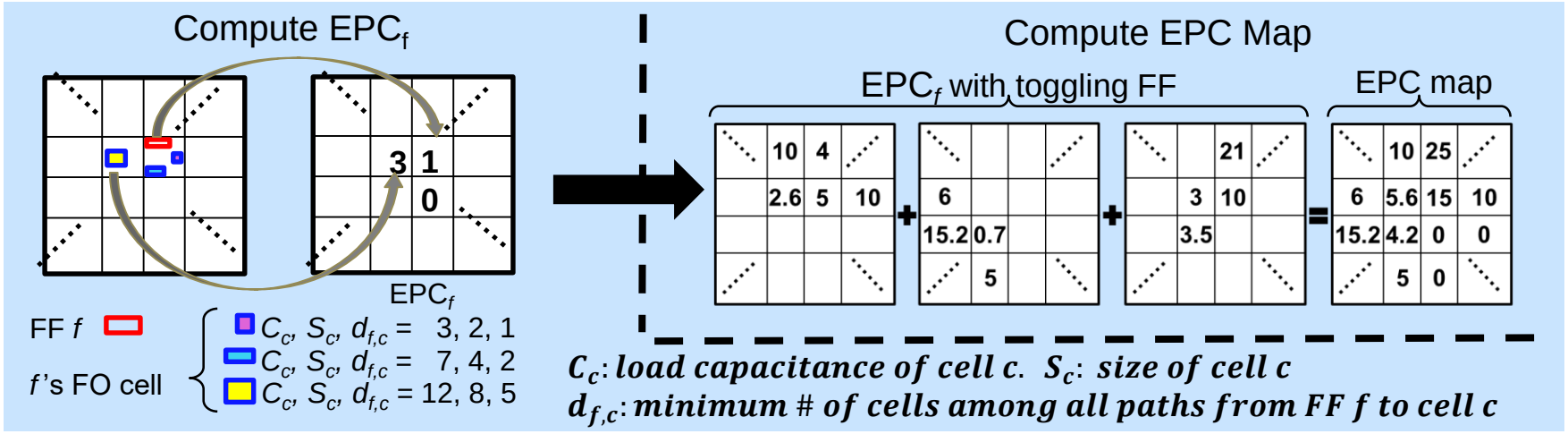
Estimated Power Contribution (EPC) Maps

One EPC map for one cycle

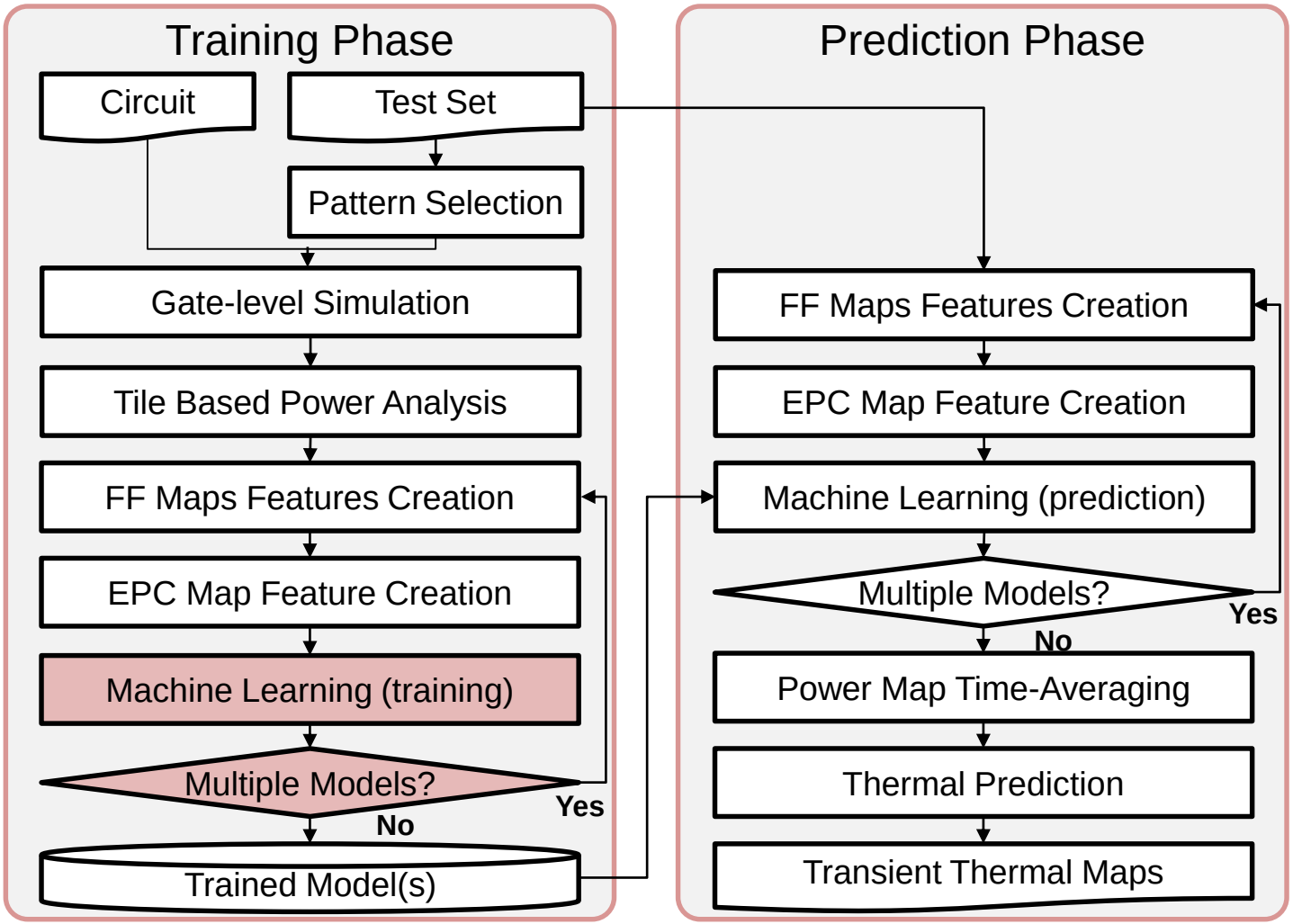
Represent **combinational power** w/o gate-level simulation

EPC_f : FF *f* 's estimated power contribution for all tiles

Compose by EPC_f

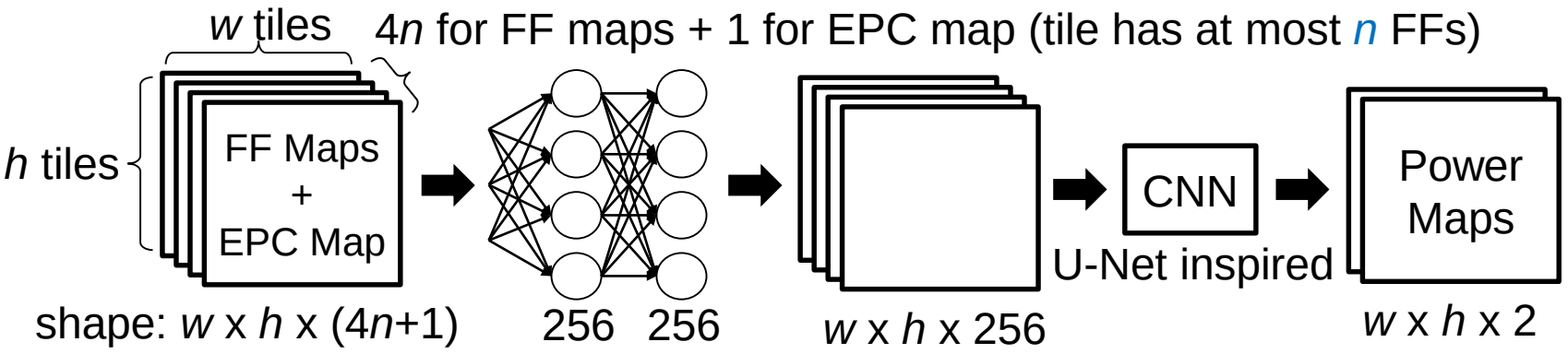


Overall Flow



Model Architecture

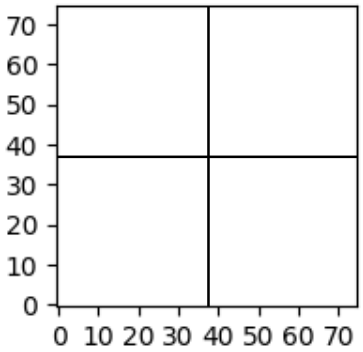
- **Use CNN inspired by U-Net [Ronneberger 15]**
  - View cycle-based power maps as images
  - Convolutional Neural Network (CNN) is suitable
  - Two output channels: one for static, one for dynamic
- **Use multiple models**
  - Improve accuracy since CNN cannot represent circuitry



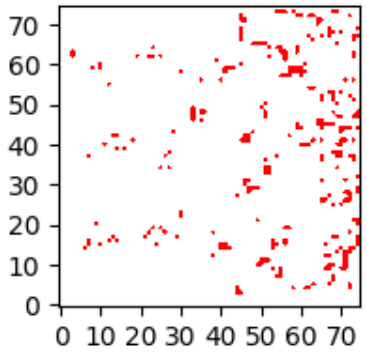
[Ronneberger 15] Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation." MICCA. Springer, 2015.

Multiple Models
By-Region or By-Error

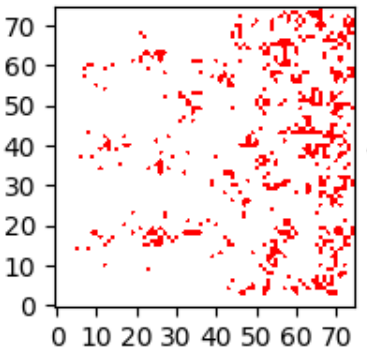
- **By-region (total 4 models)**
  - Assume close tiles have similar power behaviors
  - 2 x 2 regions. One region, one model
- **By-error (total 4 models)**
  - Assume high error tiles need more ML parameters
  - Train 1 model first. Then 3 models for high error tiles



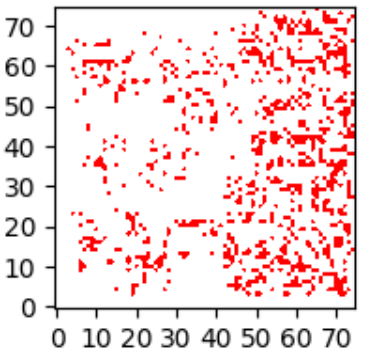
2 by 2 regions



Highest 5%



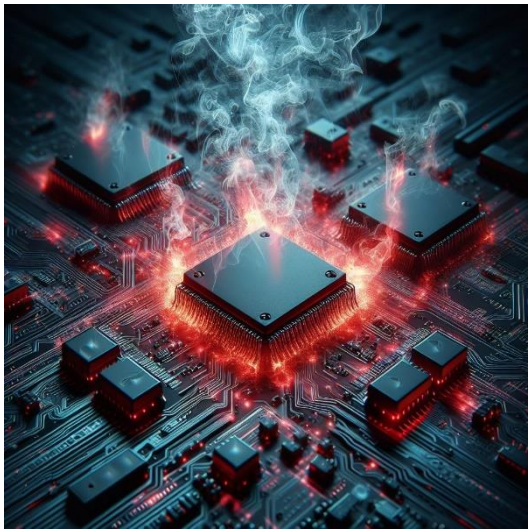
Highest 5% to 15%



Highest 15% to 30%

Outline

- Introduction
- Background
- Proposed Technique
- **Experimental Results**
  - Power prediction
  - Thermal prediction
  - Runtime and storage comparison
- Conclusion



Power Prediction Setup

- **Three single-core designs**
  - Predict shift cycles only
  - 10μm x 10μm tile, 100MHz test application speed
  - VCS (gate-level simulation), Redhawk (power analysis)
  - Mean absolute error to train ML models

Design	Tiles	# Cells	# FFs	Chain Length	For Training		For Validation	
					# Bins	# Ptns	# Bins	# Ptns
MIPS	75x75	206k	22k	281	300	100	300	118
MEMC	105x105	306k	59k	297	300	64	300	57
leon3mp	144x144	823k	109k	273	3,000	74	100	90

Power Prediction Results
Single-core Designs

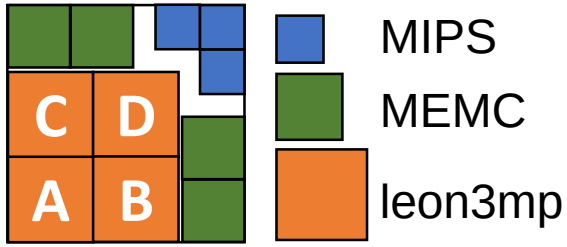
- **Overall: low MAPE, high CC**
  - Ignore unimportant tiles (power less than 1nW)
- **High power bins: low MaxPE**
  - 10 equal-width bins, show 3 highest power bins only
  - Do not worry about missing thermal hot spots

	Static (Overall)		Dynamic (Overall)		Dynamic (bin8)	Dynamic (bin9)	Dynamic (bin10)
	MAPE(%)	CC	MAPE(%)	CC	MaxPE(%)	MaxPE(%)	MaxPE(%)
MIPS	1.14	**0.997**	7.37	**0.996**	6.77	3.12	**2.03**
MEMC	2.18	**0.986**	4.71	**0.991**	17.65	4.91	**1.93**
leon3mp	1.35	**0.992**	3.64	**0.997**	3.02	0.79	**3.63**

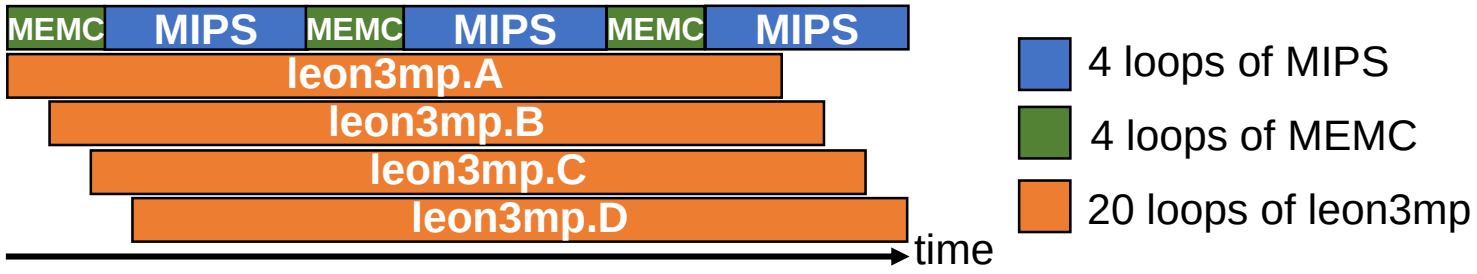
Mean Absolute Percentage Error (MAPE)
Max Percentage Error (MaxPE)

Thermal Prediction Setup

- **Create multi-core design**
  - 4mm x 4mm, 5.13M cells
  - Single-core designs are too small

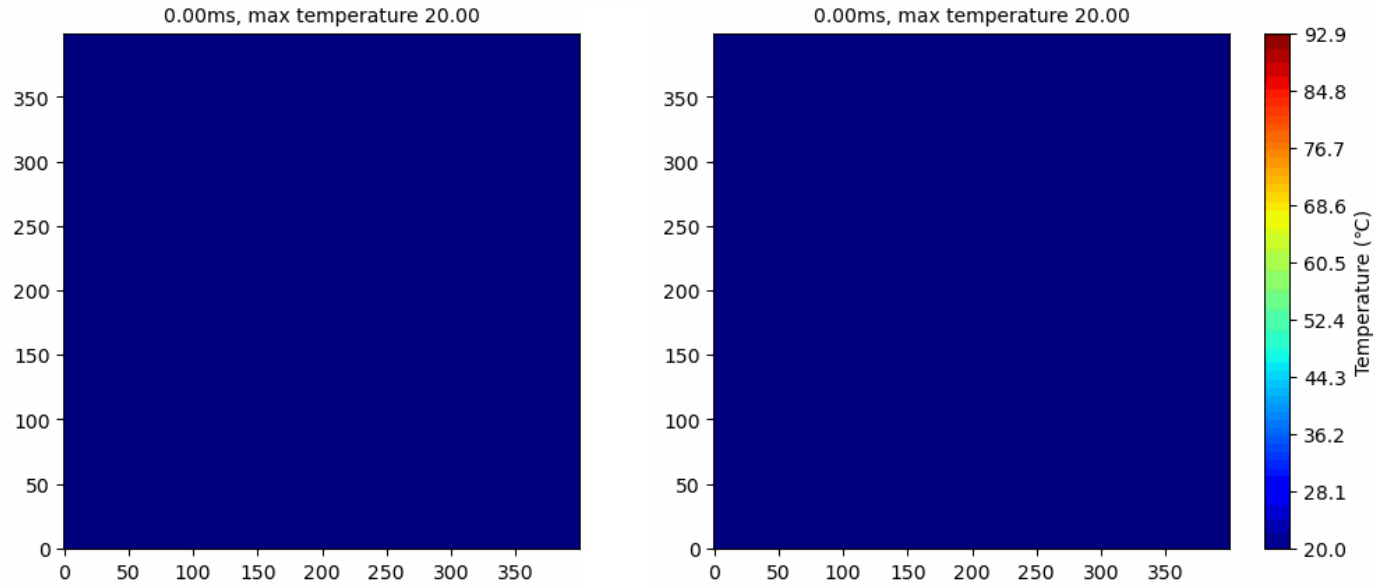


- **Use Ansys Fast Transient Thermal Solver**
  - Initial temperature 20℃
- **0.57s test time, 300MHz at-speed**
  - A loop: repeatedly apply each pattern 300 times



Thermal Maps Comparison

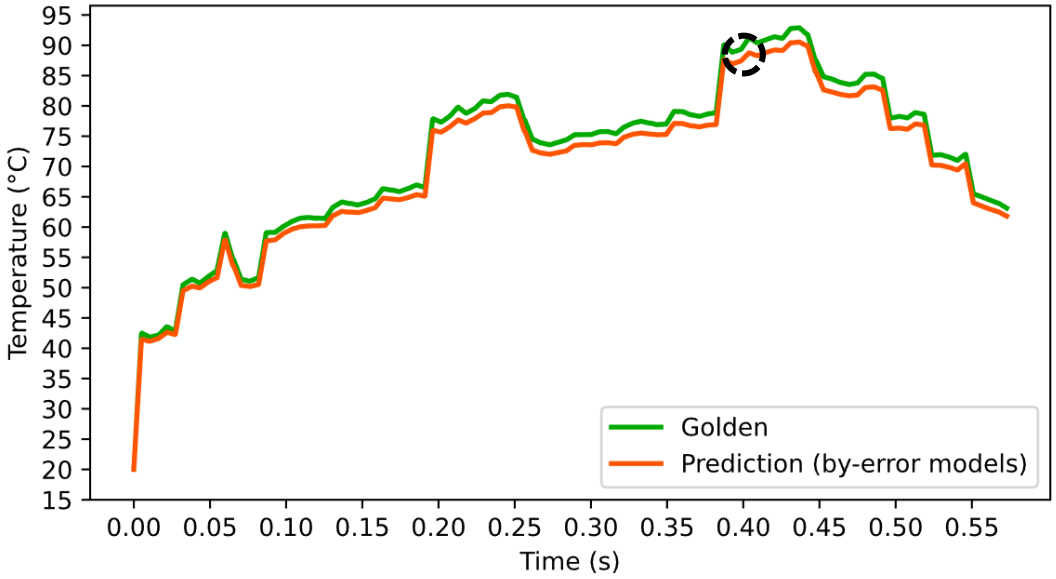
- **Good accuracy**
  - MAE: 1.17℃, MaxE: 2.48℃, CC: 0.999

	Golden	Prediction
Power Maps Source	Ansys redhawk	Power prediction
Thermal Solver	Ansys Fast Transient Thermal Solver	
Thermal Maps		

Power prediction: by-error models with weighted FF maps and EPC maps

Transient Temperature of Hottest Tile

- **Predicted waveforms very close to golden one**
  - Error does not accumulate significantly with time
- **Error less than 2.47℃**
  - Max error happens at 0.4s



Runtime and Storage Comparison

75X Runtime speedup

Power prediction

Design	Traditional	Our	Speedup
MIPS	477.1	11.5	41X
MEMC	813.7	13.8	59X
leon3mp	1541.4	13.3	116X

Thermal prediction

Design	Traditional	Our	Speedup
Multi-core	2910.9	38.7	**75X**

Unit: second per pattern
Traditional thermal solver: [Huang 06]

118X Storage reduction

Disk storage

Design	Traditional	Our	Reduction
MIPS	13.15	0.09	146X
MEMC	5.35	0.09	59X
Leon3mp	109.28	0.9	121X
Multi-core	127.78	1.08	**118X**

Unit: GB

Conclusion

- **Predict power maps by machine learning**
  - Propose FF maps, EPC maps, and multiple models
  - MAPE: less than 8% (all tiles) and 2% (high power tiles)

- **Enable thermal analysis of long test patterns**
  - Combine power prediction with Ansys Thermal Solver
  - MAE: less than 1.2°C (5M cells, 0.57s testing)
  - 75X runtime speedup, 118X disk storage reduction
  - Ensure thermal safety during testing

Quiz !

Q: Which of following about this research is NOT correct?

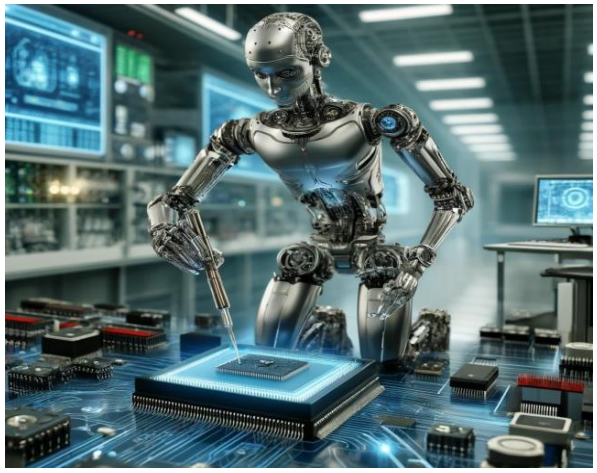
A. AI predictor do not need any domain knowledge from engineers

B: AI predictor still needs domain knowledge to extract useful features

C: AI predictor can be much faster than traditional analyzer with little acceptable error

AI for Testing, Testing for AI

- Introduction
- **AI for Testing**
  - IR Drop
  - Thermal
  - V_{min} [Kuo ITC 21]
- Testing for AI
  - Application-oriented Test
  - Manufacture-oriented Test
- Conclusion



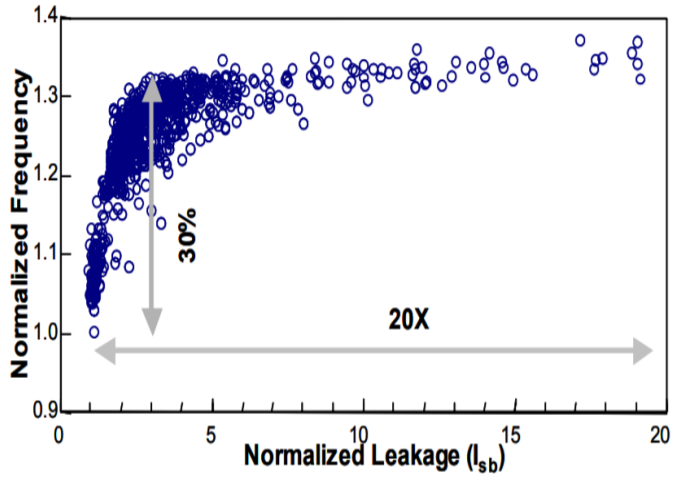
[Kuo ITC 21] Y.T. Kuo and et. al "Minimum Operating Voltage Prediction in Production Test Using Accumulative Learning," Int'l Test Conf. 2021

Production Test Floor

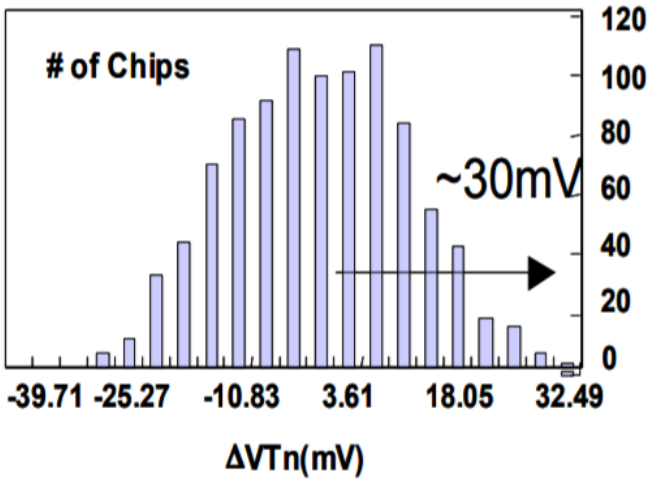
Watch video
17:28~18:10

Chip Performance Variation

- As feature size shrinks, chip performance variation increases [Borkar 05]
- 3 root causes of chip performance variation [Borkar 03]
  - PVT values : Process, Voltage, Temperature



Leakage and frequency variation [Borkar 03]



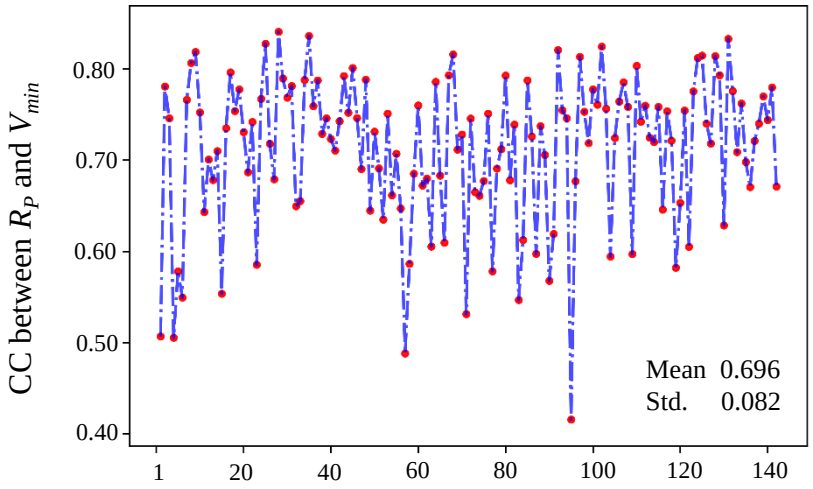
Threshold voltage variation [Borkar 03]

Chip Performance Measurement

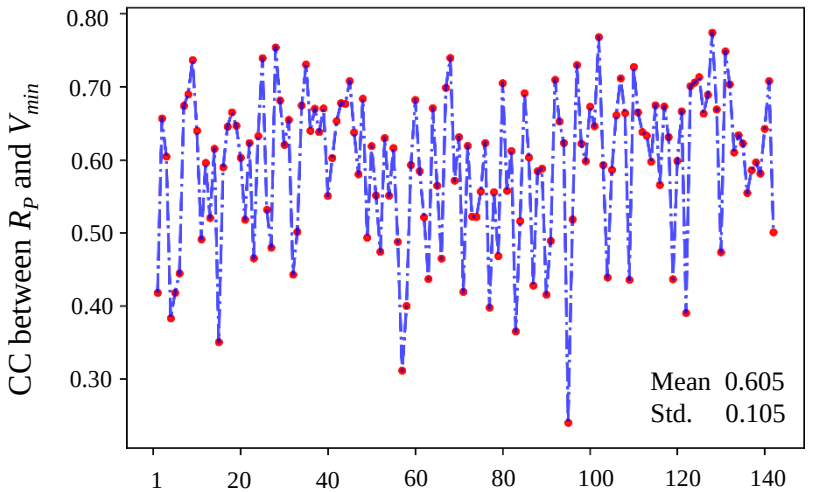
- **Two important metrics to evaluate chip performance**
  - **Minimum operating voltage (V_{min})**
  - **Maximum operating frequency (F_{max})**
- **Use ATE to test two important metrics**
  - Expensive
- **Use AI-based framework to predict chip performance**
  - On-chip sensor: path slack sensors, ring oscillators, thermal sensors
  - Structure test measurements
  - Wafer electrical test measurements

Prediction Problems of Past Works

- **Insufficient prediction accuracy**
- **Impact of lot-to-lot variations to accuracy**
  - These works are unsuitable for industrial test flow



(A) 142 CCs on Design A



(B) 142 CCs on Design B

V_{min} Prediction in Production Test Using Accumulative Learning [Kuo 21]

Motivation

- Enhance chip performance prediction accuracy
  - Accurate V_{min} means longer battery life
- Reduce impact of lot-to-lot variations and test time

Goal

- Find new features to improve chip performance prediction accuracy
- Invent a new methodology which can reduce the impact of lot-to-lot variations and test time

[Kuo 21] Y.T. Kuo and et. al "Minimum Operating Voltage Prediction in Production Test Using Accumulative Learning," Int'l Test Conf. 2021

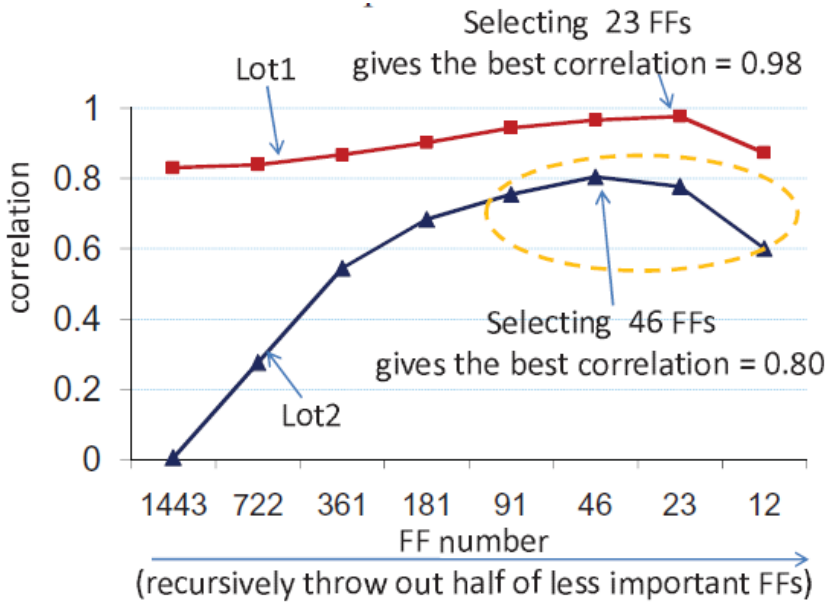
Outline

- Introduction
- Previous Works
- Proposed Techniques
- Experimental Results
- Conclusion & Future Work

Generated by Copilot

Selecting the Most Relevant Structural F_{max} for System F_{max} Correlation [Chen 10]

- Proposed a reliable functional F_{max} predictor given structural F_{max}
- One model per lot (Per-lot model)
- Consider the impact of lot-to-lot variations
- Unsuitable for production test



[Chen 10] Janine Chen, Li-C. Wang, et al. "Selecting the Most Relevant Structural Fmax for System Fmax Correlation." VLSI Test Symposium, 2010.

DVFS Binning Using Machine Learning Techniques [Chang 18]

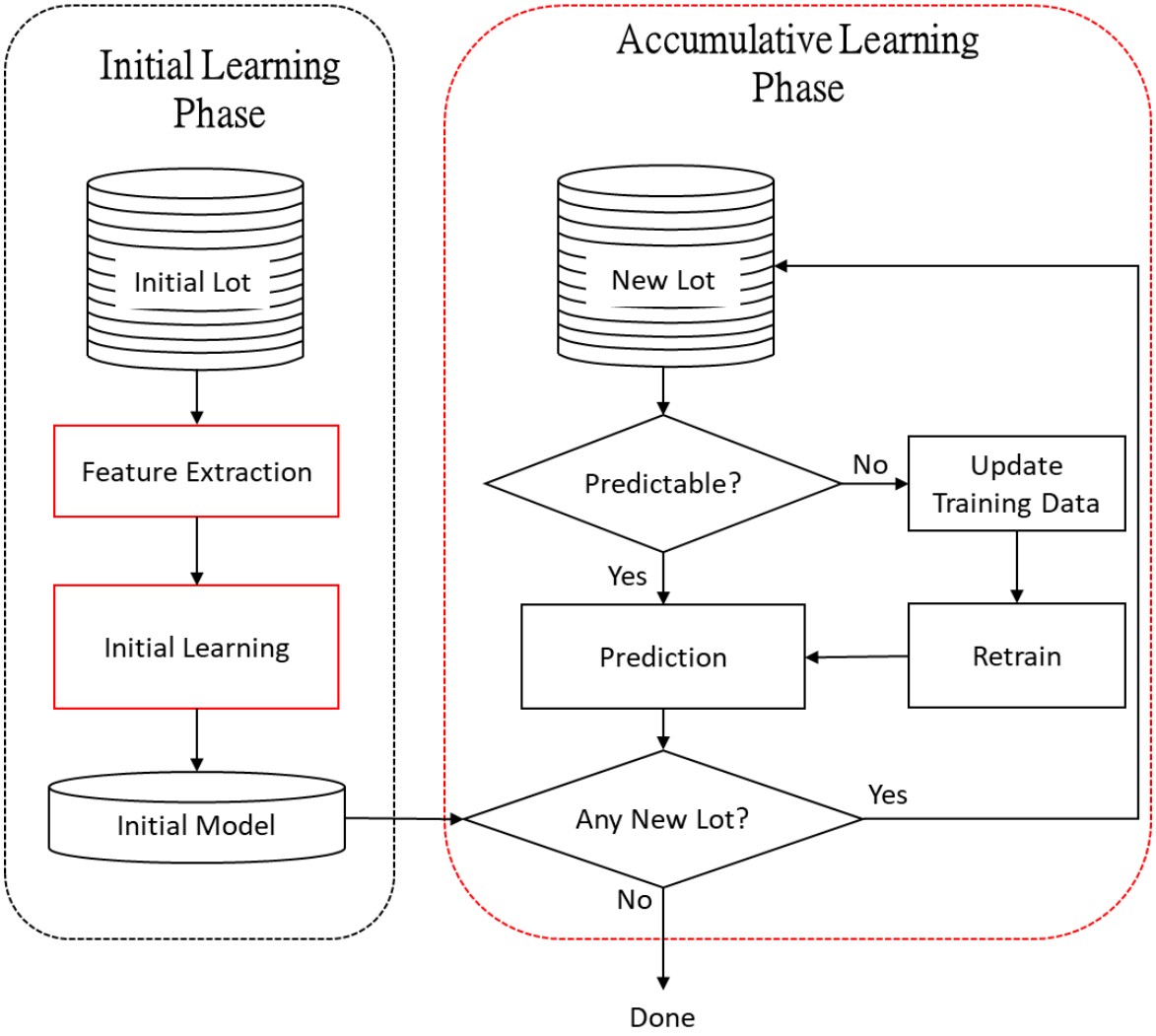
- **Consider many different features**
  - Freq. of on-chip ROs
  - CPU functional test result
  - Chip-power measurements
  - Thermal-sensor measurements
- **75.1%~85.9% accuracy on different DVFS policies**
  - Binary classification (Pass/ Fail)
- **Prediction without fail-to-pass case (test escape)**
  - Their error bound is 4σ
- **Test cost too large**

[Chang 18] Chang, K. W., Huang, C. Y., Mu, S. P., Huang, J. M., Chen, S. H., & Chao, M. C. T. (2018, August). DVFS Binning Using Machine-Learning Techniques" IEEE International Test Conference in Asia.

Outline

- Introduction
- Previous Works
- **Proposed Techniques**
  - **Feature extraction**
  - Initial training
  - Accumulative learning
- Experimental Results
- Conclusion & Future Work

Overall Flow



Feature Extraction

- Three important features related to V_{min}

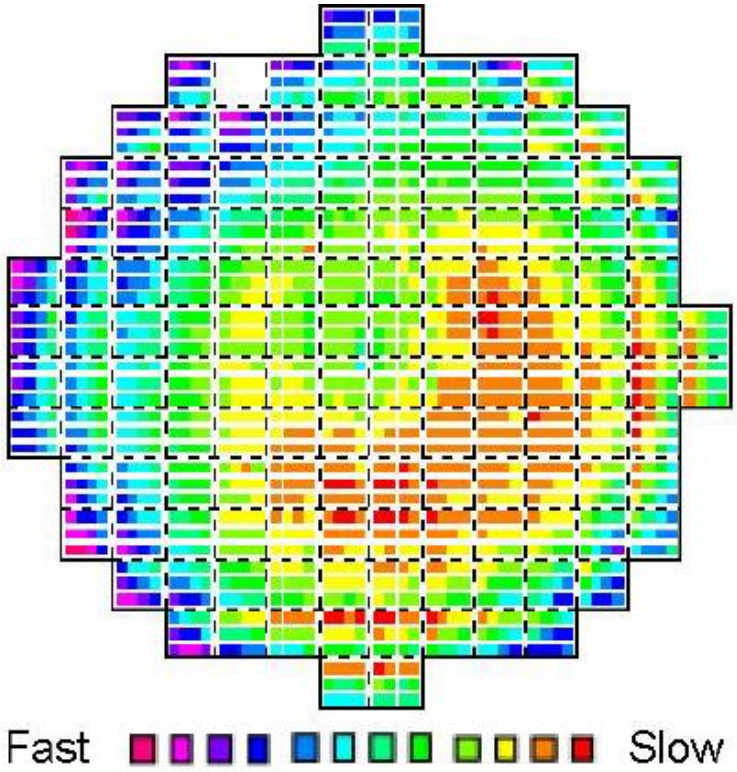
Features	Explanation
$Count_{(Cell\ Type)}$	The count of ring oscillators of a certain cell type e.g. $Count_{NAND}$, $Count_{NOR}$, etc.
I_{on}	A set of on-state current of transistors
Flatness	The flatness of a wafer

On-state Current

- **Related to some attributes of transistors [Sedra 98]**
  - Channel length, oxide thickness, and threshold voltage
- **Direct current (DC) measurement**
  - Regardless of alternating current (AC)

Flatness of a Wafer

- Related to the results of chemical-mechanical polishing [Gattiker 08]



Delay Variation on The Wafer Map [Gattiker 08]

Accuracy Comparison about Features

- F_R, F_{RI}, F_{RF}, and F_{RIF} are four feature set candidates
- Per-lot model (30% training, 70% test)

F_R: Only RO
F_{RI}: RO + I_{on}
F_{RF}: RO + *Flatness*
F_{RIF}: All

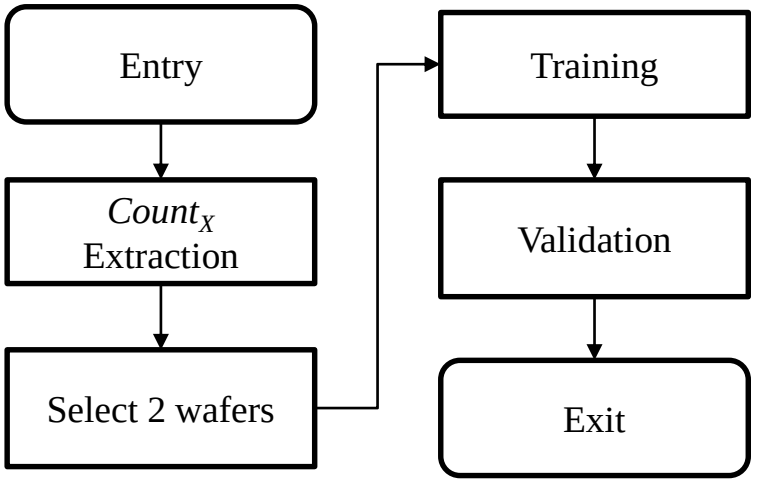
Train (30%)	Test (70%)	Data Amount	Design A				Design B			
			F_R	F_{RI}	F_{RF}	F_{RIF}	F_R	F_{RI}	F_{RF}	F_{RIF}
Lot A	Lot A	9472	93.2%	93.7%	93.5%	93.9%	91.0%	91.2%	91.9%	92.3%
Lot B	Lot B	8287	94.2%	94.4%	94.5%	94.8%	91.4%	91.6%	92.0%	91.9%
Lot C	Lot C	9608	93.8%	94.1%	95.0%	94.9%	90.6%	92.0%	92.3%	92.5%
Lot D	Lot D	9368	93.0%	93.8%	93.9%	94.2%	90.7%	91.0%	91.6%	91.4%
Lot E	Lot E	7458	93.2%	93.5%	95.0%	94.9%	89.5%	90.6%	92.4%	92.7%
Average			**93.5%**	**93.9%**	**94.4%**	**94.5%**	**90.6%**	**91.3%**	**92.0%**	**92.2%**

Outline

- Introduction
- Previous Works
- **Proposed Techniques**
  - Feature extraction
  - **Initial training**
  - **Accumulative learning**
- Experimental Results
- Conclusion & Future Work

Initial Training

- **$Count_X$ = count of ring oscillator of a certain cell type x**
  - Highly correlated with V_{min}
  - Low variation between different lots

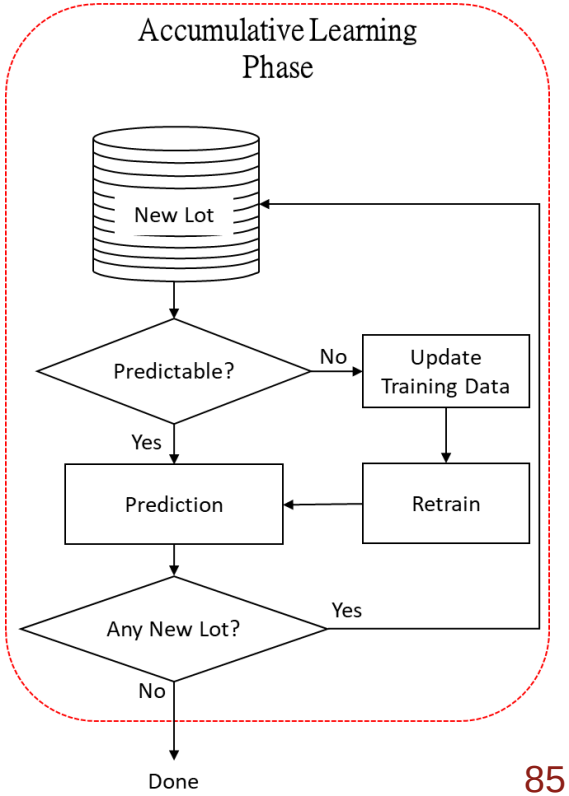


- **Select two training wafers**
  - Sizes of range of $Count_x$ are highest among all wafers in the initial lot
  - The range of $Count_x$ of two wafers:

$$R = (\min Count_x, \max Count_x)$$

Accumulative Learning

- **Use *R* to determine whether our model can predict new lot or not**
  - Range of $Count_x$ of two wafers of new lot is within *R* (Predictable)
  - Otherwise, it is not predictable.
- **How to retrain a new model**
  - Add the new data of two wafers into the training data
  - *R* is updated by expanding the range of $Count_x$



Outline

- Introduction
- Previous Works
- Proposed Techniques
- **Experimental Results**
- Conclusion & Future Work

Experimental Setup

- **Designs for experiments**
  - **1,195,095 chips** from **142 good yield lots**
  - Two 7nm designs on each chip
  - All experimental data are package test data
- **Use *scikit-learn* linear model [Pedregosa 11]**
- **Accuracy calculation method**

$$|Predicted\ V_{min} - V_{min}| \leq Res$$

$$\text{Accuracy} = \frac{C_{pass}}{C_{all}} \times 100\%$$

Res: Voltage resolution of ATE
C_{pass}: Number of chips of correct predicted V_{min}
C_{all}: Number of chips of all test data

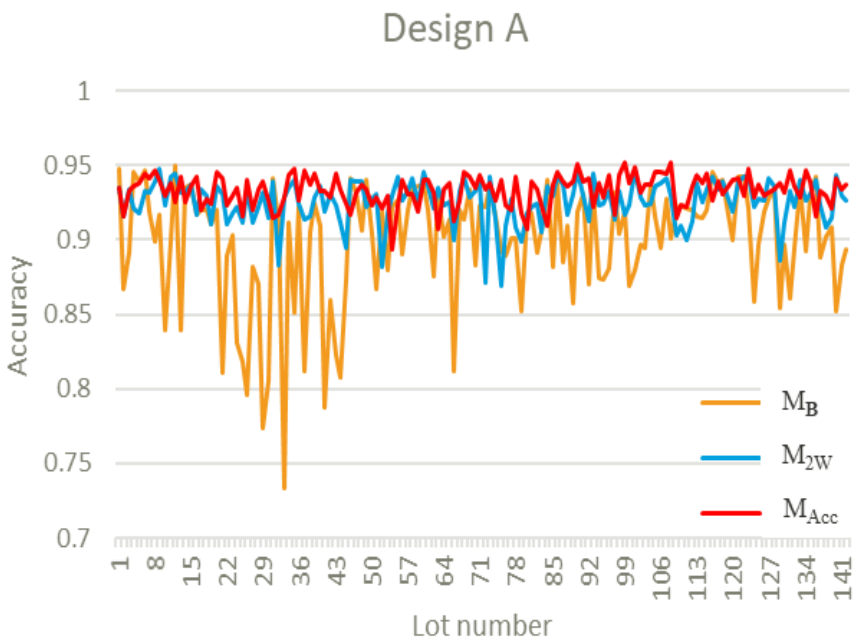
Comparison of Four Models

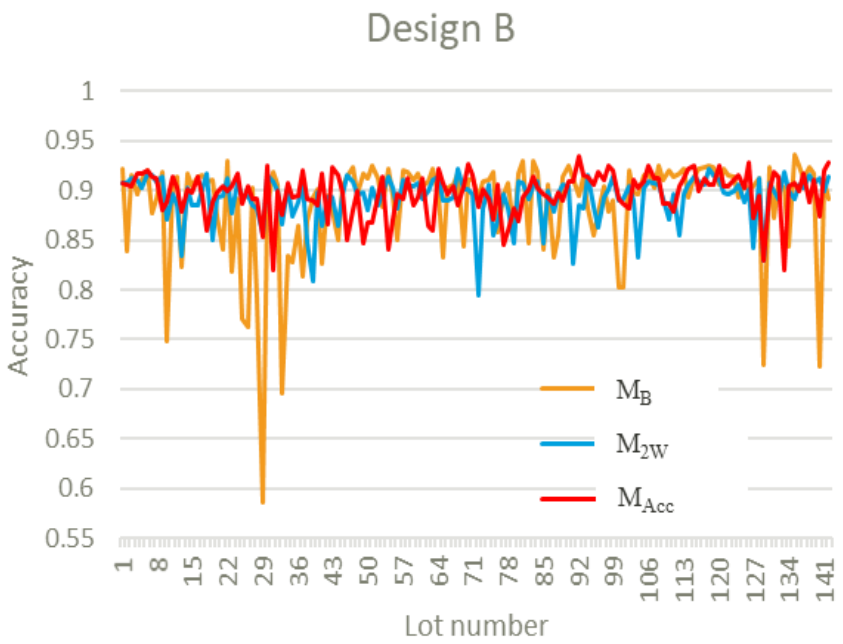
- M_A and M_B are one-lot models
- M_{2W} is a per-lot model selecting two wafers per lot
- M_{acc} is accumulative learning model
  - Best results

Model	Design A				Design B			
	Worst	Best	Std.	Avg.	Worst	Best	Std.	Avg.
M_A	77.7%	94.9%	3.9%	87.5%	50.2%	90.1%	9.4%	71.2%
M_B	73.3%	95.0%	4.0%	89.8%	58.6%	93.6%	5.1%	88.6%
M_{2W}	87.0%	94.8%	1.5%	92.6%	79.5%	92.2%	2.3%	89.2%
M_{Acc}	**89.3%**	**95.1%**	**1.0%**	**93.4%**	**82.0%**	**93.4%**	**2.2%**	**89.8%**

Accuracies of 141 Lots of 3 Models

- **Proposed method remains good accuracy**





Test Time Comparison

- **Reduce 75 % test time**
  - T_F is the test time of all features
  - T_{vmin} is the test time of testing V_{min}
  - L_s is the number of selected test lots
  - L_A is the number of all lots

$$\text{Test Time Ratio} = \frac{T_F \times 25 + T_{vmin} \times \mathbf{2}}{T_F \times 25 + T_{vmin} \times 25} \times \frac{\boldsymbol{L_S}}{L_A}$$

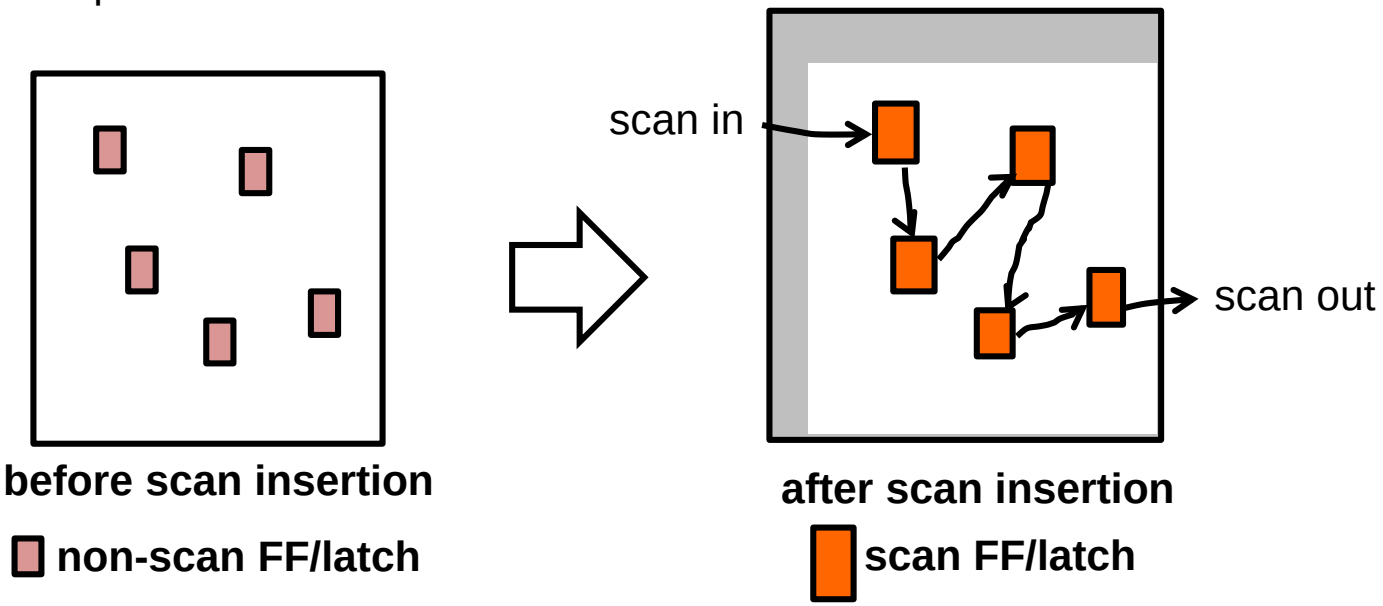
Test method	Test Time Ratio
ALL	1.00
2W	0.34
ACC	0.25

Conclusion

- Two new features *(I_{on} and Flatness)* enhance prediction accuracy
- Our methodology reduces impact of lot-to-lot variations
- Almost 90% accuracy on 1.2M chips
- Our methodology saves 75% test time
- Need different test flows for
  - initial learning and accumulative learning

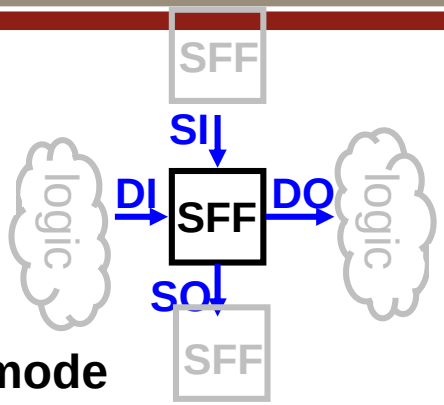
Design for Testability: Scan Chains

- **Scan: connect (internal) FF/latches as shift register**
  - Control and observe FF/latches in test mode
  - Remain original function in normal mode
- **Proposed in early 1970's [Williams 73][Eichelberger 77]**
  - Most important DFT for synchronous digital circuits
- ***Scan chain insertion*: aka. *DFT insertion*, *DFT synthesis***
  - 1. Replace FF/latch 2. Stitch FF/latch into a chain

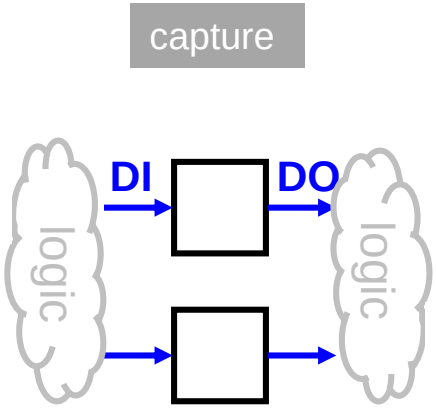


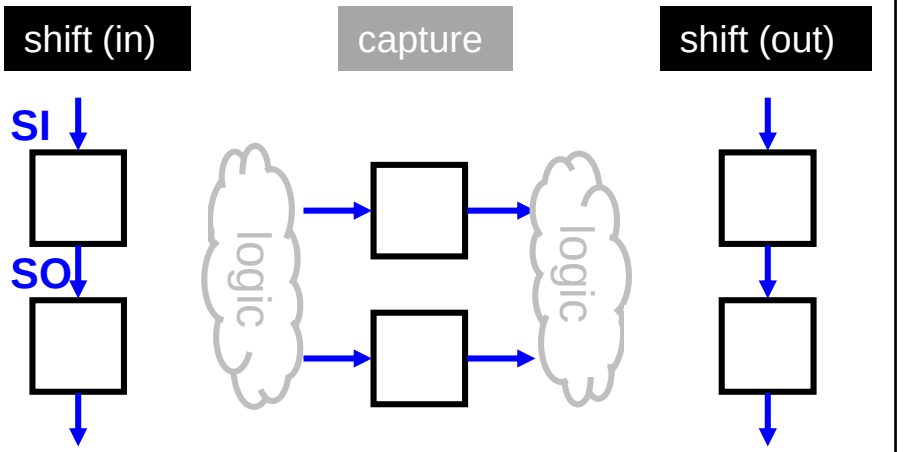
Scan Flip-Flop (SFF)

- **SFF has four main pins:**
  - Scan Chain: *Scan Input* (SI), *Scan output* (SO)
  - Logic: *Data Input* (DI), *Data Output* (DO)
- **SFF has two functions: *shift* and *capture***
- **Circuit has two operation modes: *Normal* mode and *Test* mode**



Normal Mode	Test Mode
capture	shift (in) → capture → shift (out)



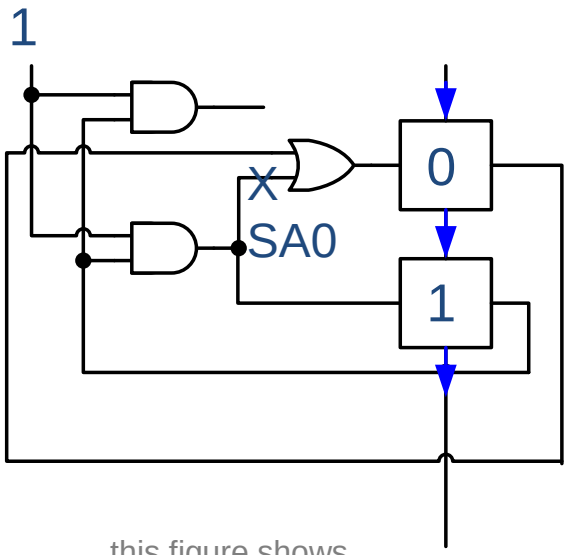


Quiz!

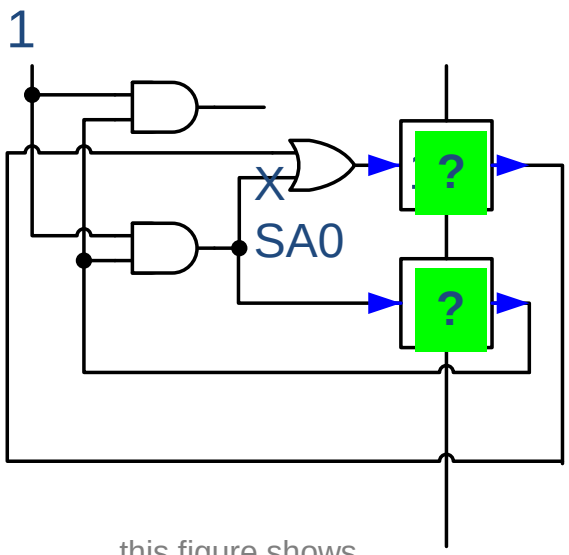
Shift (in)
2 cycles

Capture
1 cycle

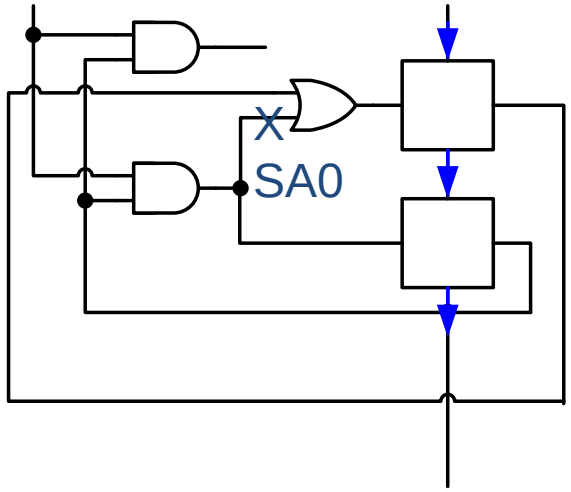
Shift (out)
2 cycles



this figure shows
state of SFF after 2 cycles



this figure shows
state of SFF after capture cycle



observed 1/0
at 2nd cycle

Fault Detected in Test Mode

AI for Testing, Testing for AI

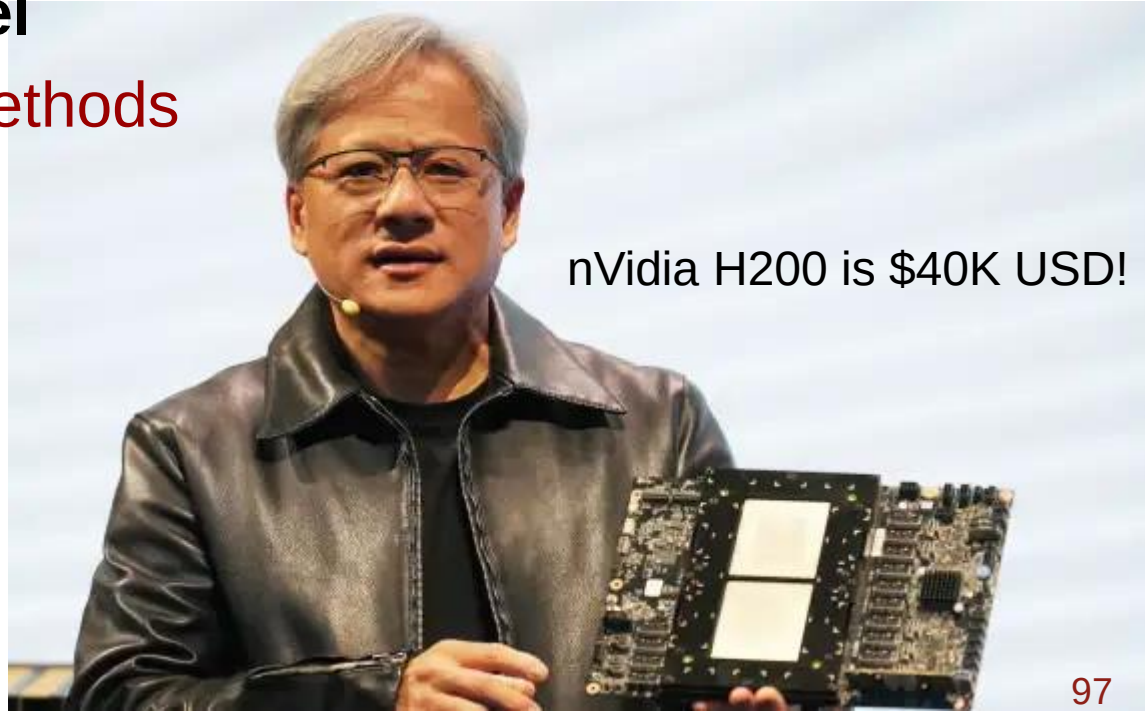
- Introduction
- AI for Testing
  - IR Drop
  - Thermal
  - V_{min}
- **Testing for AI**
  - Testing Neuromorphic Chips [Tseng ICCAD 21]
- Conclusion

MOST 科技部
Ministry of Science and Technology

[Tseng ICCAD 21] Hsiao-Yin Tseng, I-Wei Chiu, Mu-Ting Wu, and James C.M. Li, "Machine Learning-Based Test Pattern Generation for Neuromorphic Chips," ICCAD 2021

Why Testing is Important for AI Chip?

- **1. AI chips are everywhere**
  - Some of them are expensive
- **2. AI chips used for safety critical mission**
  - Self-driving Car
- **3. AI chips are novel**
  - Traditional test methods not applicable

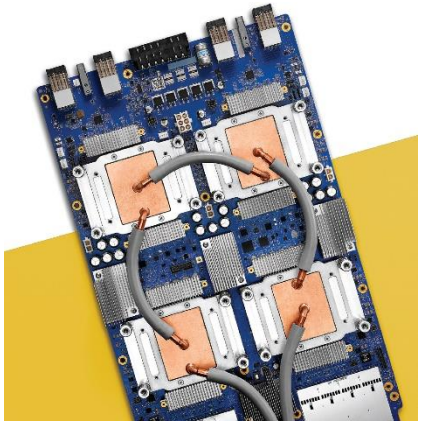


nVidia H200 is $40K USD!

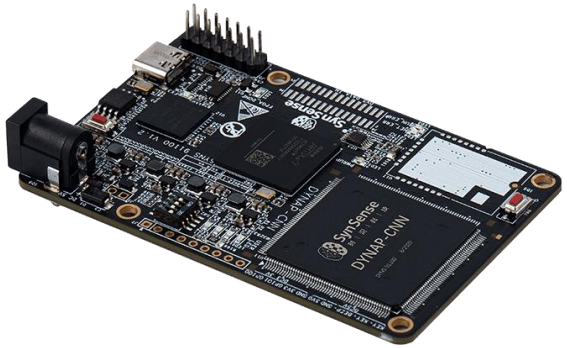
Introduction

- **Conventional AI chip consumes too much power!**
- **Many neuromorphic chips have been proposed**
  - Mimic biological neural operation to reduce power
  - Asynchronous, ReRAM, analog, mix-signal devices

Google TPUv3 (digital chip)
450W [Jouppi 21]



SynSence Dynap-CNN (neuromorphic)
10mW [Ivanov 21]



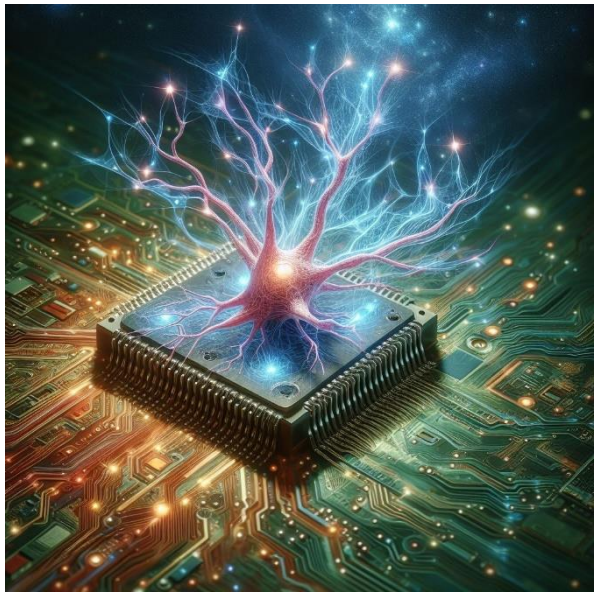
Machine Learning-Based Test Pattern Generation for Neuromorphic Chips [Tseng 21]

- **Test issues**
  - Traditional Scan DFT not applicable to neuromorphic chips
  - No functional ATPG method for neuromorphic chips
- **Traditional solution**
  - Functional testing with ML dataset
    - Test length is quite long
    - Fault coverage is low
- **Goal**
  - Propose an ATPG method for neuromorphic chips
    - Reduce test length
    - Improve fault coverage

[Tseng ICCAD 21] Hsiao-Yin Tseng, I-Wei Chiu, Mu-Ting Wu, and James C.M. Li, "Machine Learning-Based Test Pattern Generation for Neuromorphic Chips," IEEE ICCAD 2021

Outline

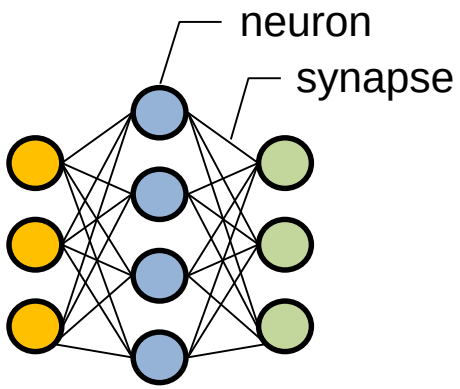
- Introduction
- **Background**
  - Spiking neural network (SNN)
  - Past research
    - SNN fault model
    - Testing of neuromorphic chips
- Proposed ATPG Technique
- Experimental Results
- Conclusion

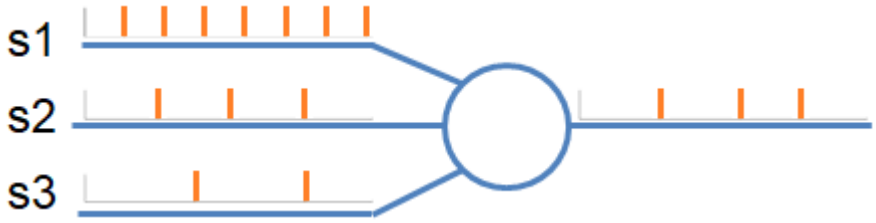


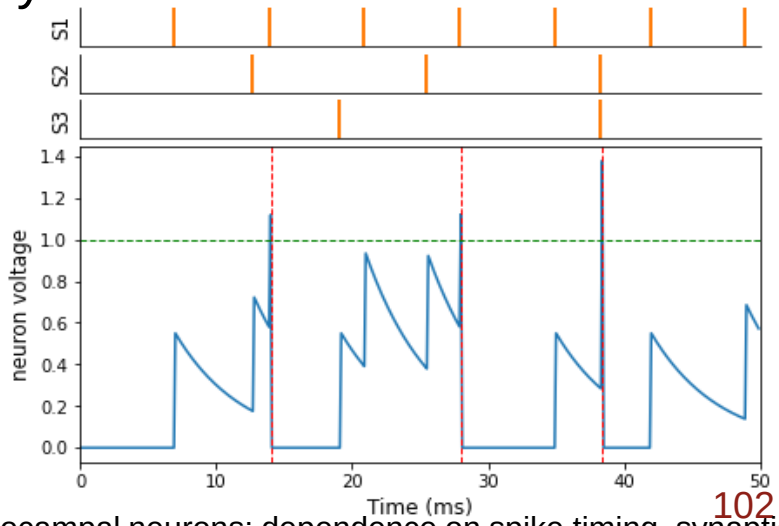
Generated by Copilot

Spiking Neural Network (SNN) [Bi 98]

- **Components**
  - Neuron
    - With membrane potential
  - Synapse
    - With weights
- **spike density represents information**
  - Dense spike represents high intensity



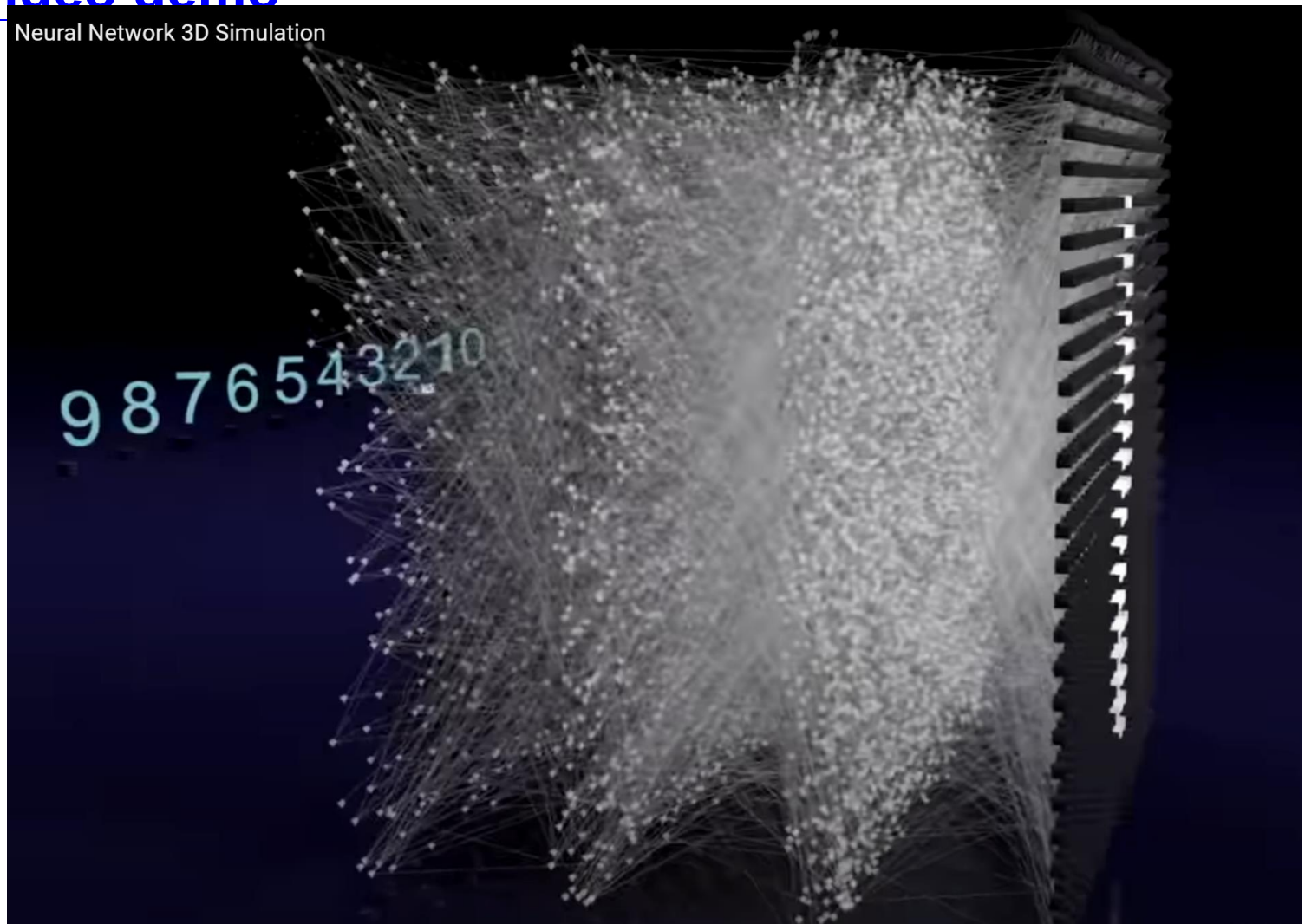




[Bi 98] G-q Bi, and Mu-ming Poo. "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type." *Journal of neuroscience* 18.24 (1998)

Spiking Neuro Network

- Video demo

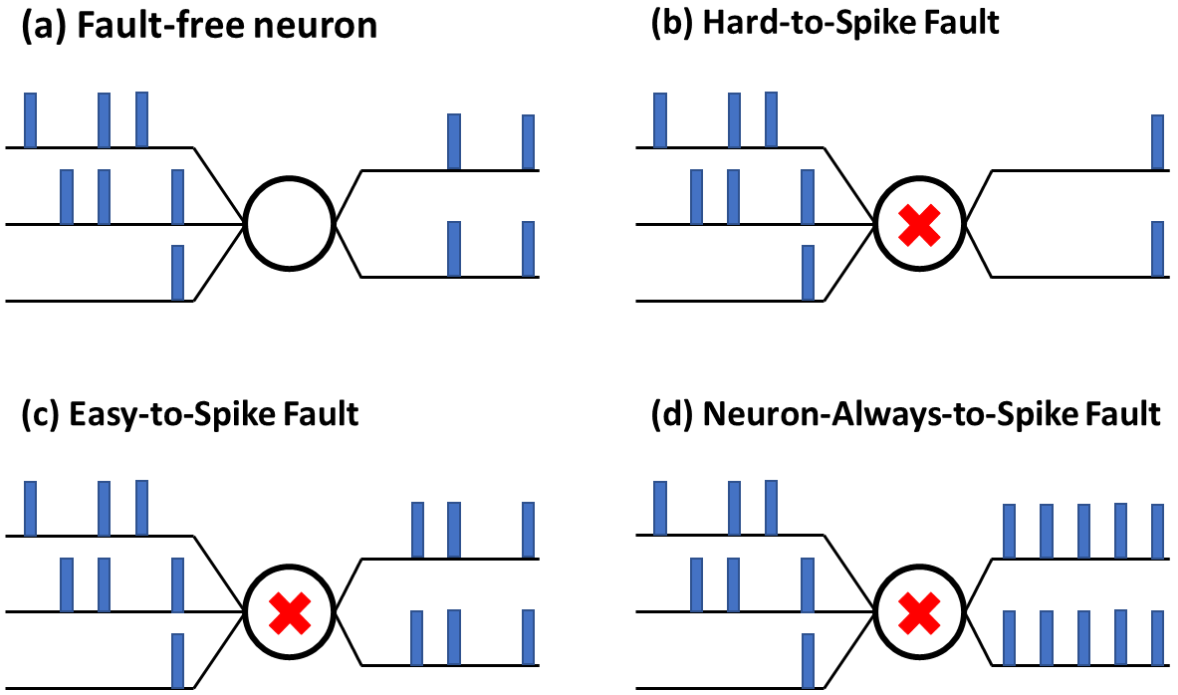


SNN Fault Models

[Hsieh 21]

New fault models needed for SNN

- Hard-to-Spike Fault (HSF): Faulty neuron is harder to spike
- Easy-to-Spike Fault (ESF): Faulty neuron is easier to spike
- Neuron Always Spike Fault (NASF): Faulty neuron always emits spike



[Hsieh 21] Y.-Z. Hsieh, et al. “Fault Modeling and Testing of Spiking Neural Network.” ITC-Asia, 2021

Testing of Neuromorphic Chips

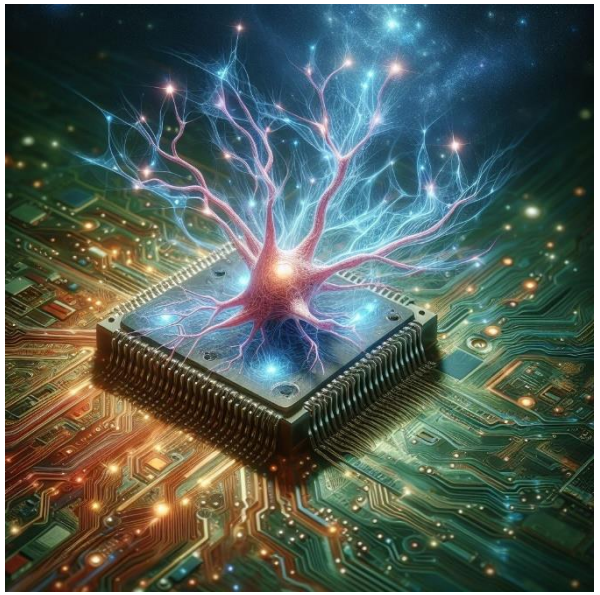
- **Self-testing Analog Spiking Neuron Circuit** [El-Sayed 19]
  - Only for specified analog hardware structure
- **Testing of Neuromorphic Circuits: Structural vs Functional** [Gebregiorgis 19]
  - Compare efficiency of structural and functional test
  - Use ML dataset as test patterns makes test length long

[El-Sayed 19] S. A. El-Sayed, et al. “Self-Testing Analog Spiking Neuron Circuit,” in 2019 16th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)

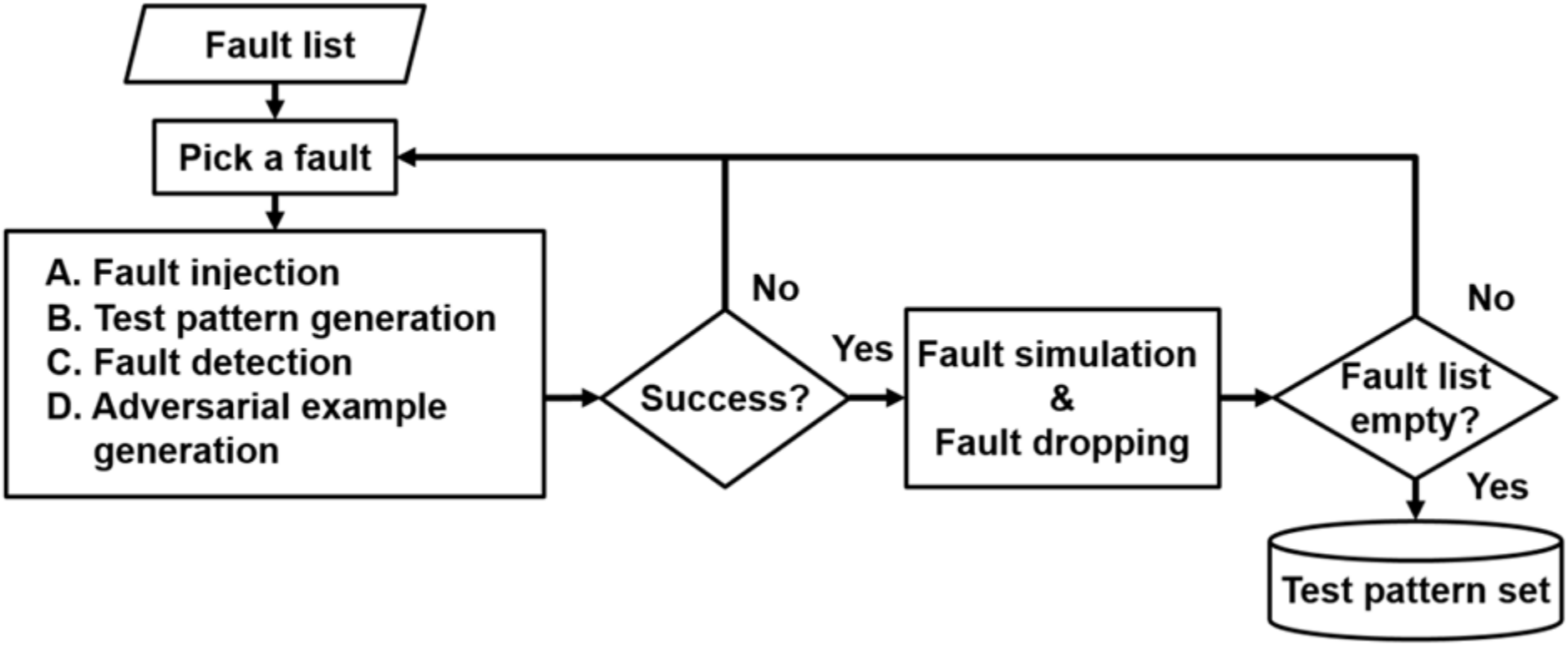
[Gebregiorgis 19] A. Gebregiorgis and M. B. Tahoori, "Testing of Neuromorphic Circuits: Structural vs Functional," 2019 IEEE International Test Conference (ITC), Washington, DC, USA, 2019, pp. 1-10

Outline

- Introduction
- Background
- **Proposed ATPG Technique**
- Experimental Results
- Conclusion

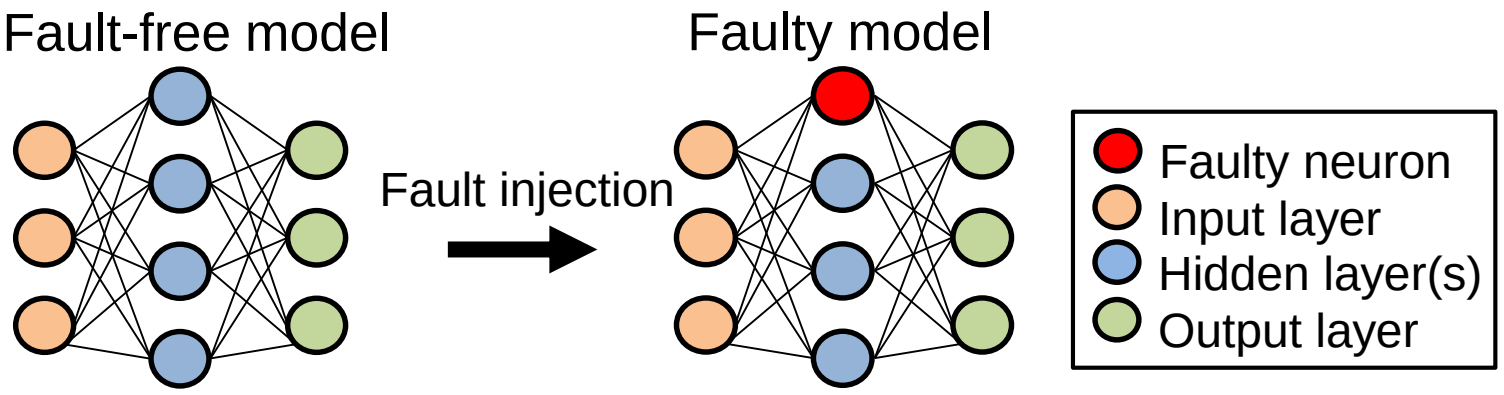


ATPG Flow

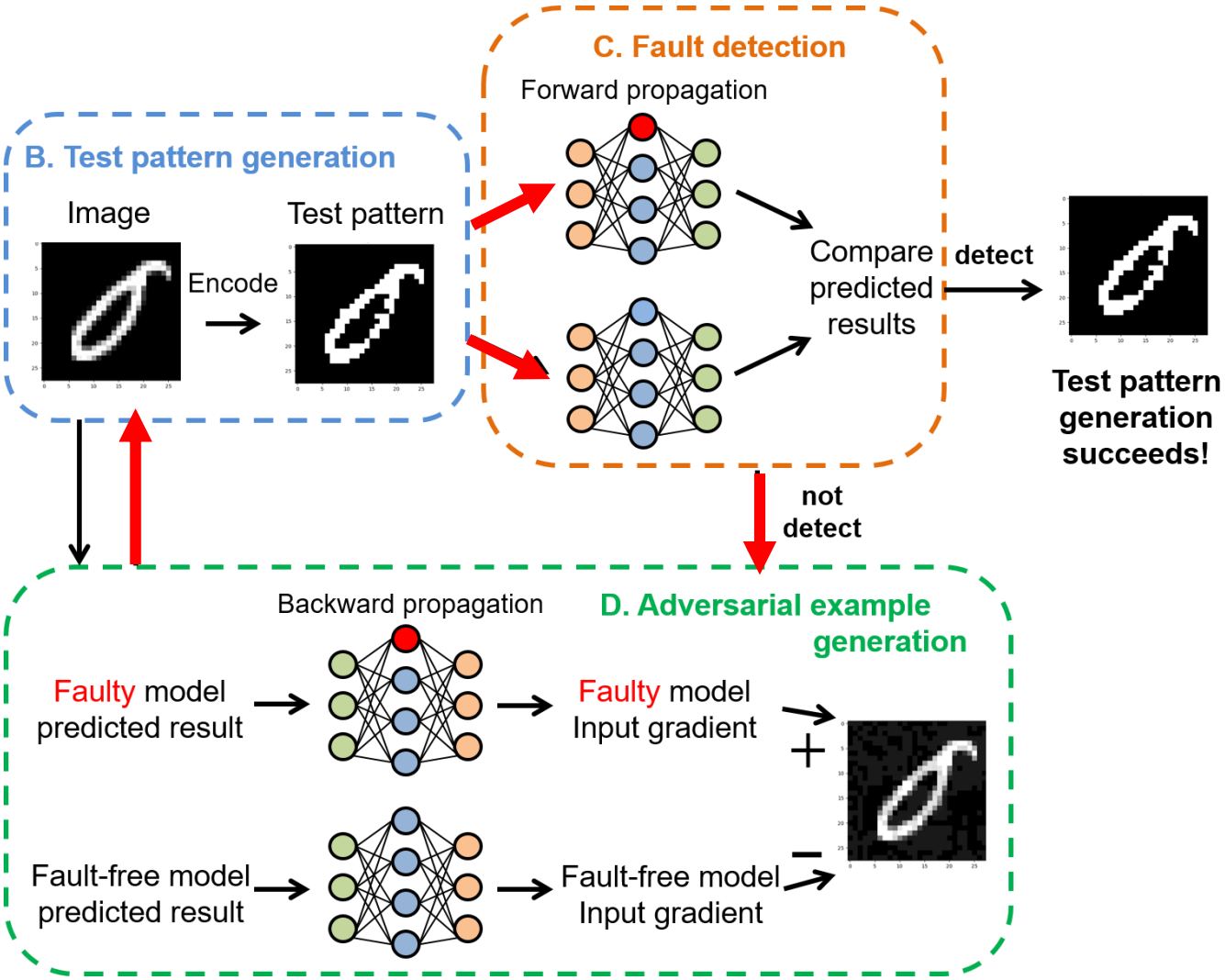

Fault list
Pick a fault
A. Fault injection
B. Test pattern generation
C. Fault detection
D. Adversarial example generation
Success?
No
Yes
Fault simulation & Fault dropping
Fault list empty?
No
Yes
Test pattern set

A. Fault Injection

- **Inject the target fault into the fault-free model**
  - HSF, ESF: modify changes of membrane potential when neuron receives spikes
  - NASF, SASF: neuron/synapse always outputs a spike
  - SWF: change synapse weight to a stuck value

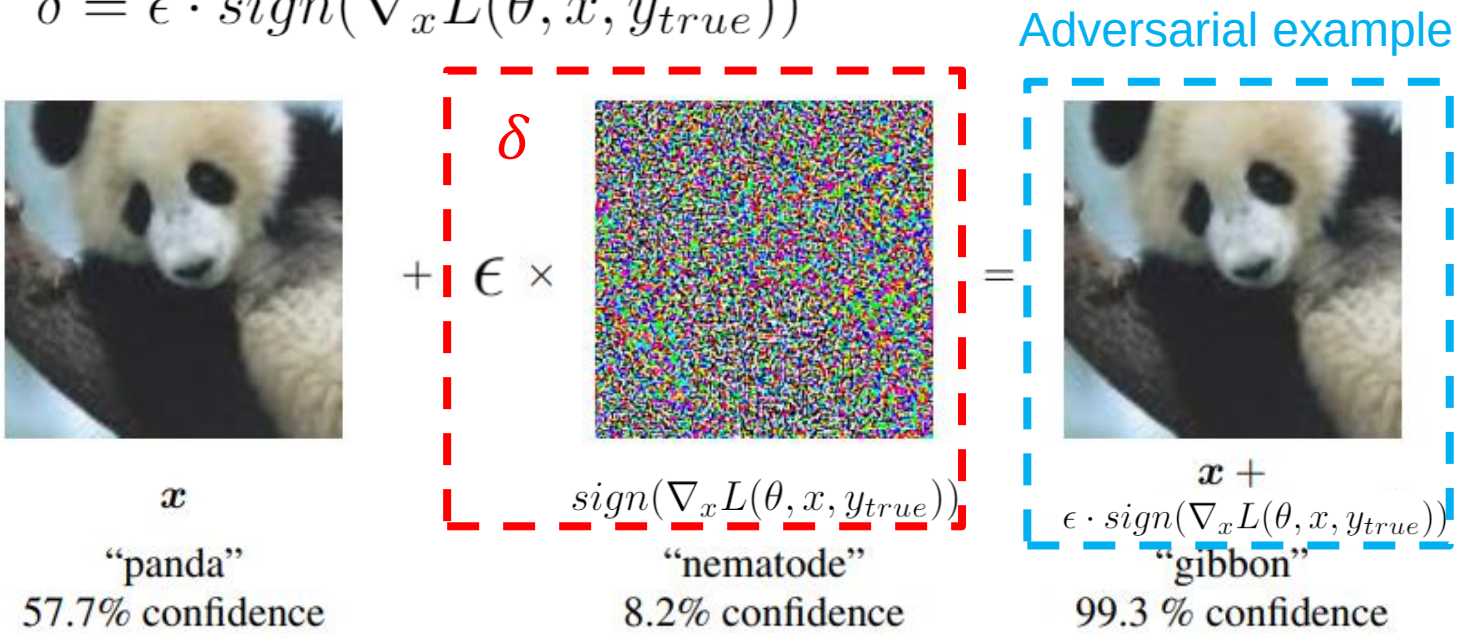


Flow of B, C, D


C. Fault detection
Forward propagation
B. Test pattern generation
Image
Test pattern
Encode
Compare predicted results
detect
Test pattern generation succeeds!
not detect
Backward propagation
D. Adversarial example generation
Faulty model predicted result
Faulty model Input gradient
Fault-free model predicted result
Fault-free model Input gradient

Adversarial Attack

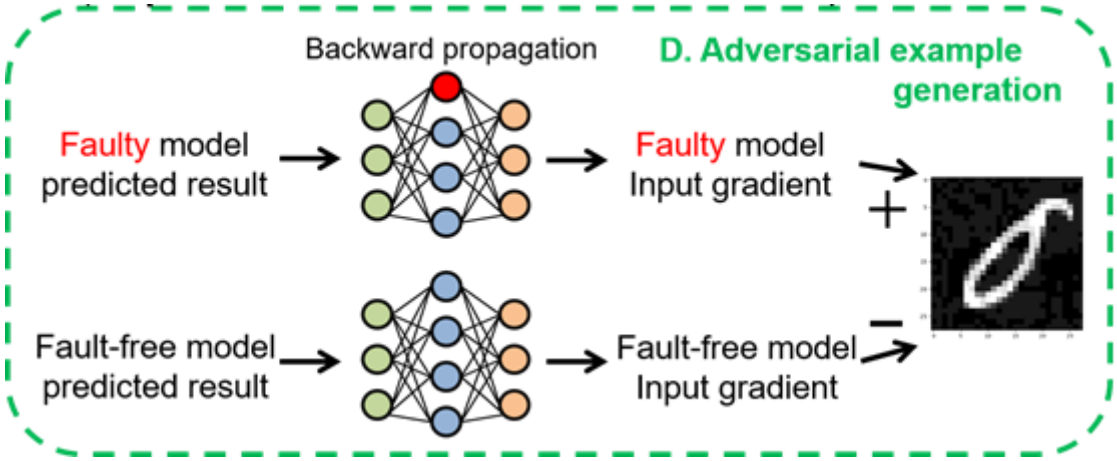
- **Inputs formed by applying small but worst-case perturbations (δ) induce incorrect answer**
- **Fast Gradient Sign Method** [Goodfellow 15]
  - $\delta = \epsilon \cdot sign(\nabla_x L(\theta, x, y_{true}))$



[Goodfellow 15] I. Goodfellow, J. Shlens & C. Szegedy, "Explaining and Harnessing Adversarial Examples." In International Conference on Learning Representations (ICLR), 2015.

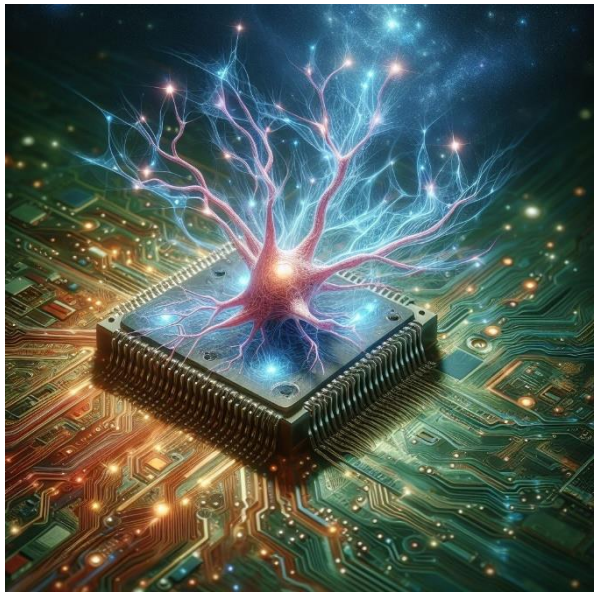
D. Adversarial Example Generation

- **Derive δ from both faulty and fault-free model**
  - $\delta = \epsilon \cdot sign(\nabla_x L(\theta_{faulty}, x, y_{true}) - \nabla_x L(\theta_{free}, x, y_{true}))$
  - + $\nabla_x L(\theta_{faulty}, x, y_{true})$: faulty model input gradient
    - make faulty model predict wrong
  - − $\nabla_x L(\theta_{free}, x, y_{true})$: fault-free model input gradient
    - make fault-free model predict correct



Outline

- Introduction
- Background
- Proposed ATPG Technique
- **Experimental Results**
- Conclusion



Experiment Setup

- **Package: BP-for-SpikingNN** [Wu 18]
  - Supervised learning
  - GPU accelerated
- **Model**
  - Trained on MNIST [LeCun 98]
  - Repeat inference 10,000 times

Model	Network Structure	Number of Neurons	Mean Accuracy
3-layer model	Input-128-10	922	97.05 %
4-layer model	Input-256-32-10	1,082	98.04 %

[Wu 18] Wu, Y., Deng, L., Li, G., Zhu, J., Xie, Y., & Shi, L. (2019). Direct Training for Spiking Neural Networks: Faster, Larger, Better. *Proceedings of the AAAI Conference on Artificial Intelligence, 33(01)*
[LeCun 98] Y. LeCun and C. Cortes, "The MNIST Database of Handwritten Digits," 1998

Definition

- **Traditional functional testing**
  - Repeat faulty simulation 30 times
  - Fault detected if $mean_{faulty} < mean_{fault-free} - 0.05$
- **Important faults (IFs)**
  - Faults that have significant impact on inference accuracy
  - Neural network is inherently fault-tolerant
- $Test\ coverage = \frac{number\ of\ detected\ IFs}{number\ of\ IFs}$

ATPG Results for Other Faults

- 100% test coverage
- Reduce test length by 566x to 8,824x

	3-layer-model (# IFs: 176, # faults: 922)					4-layer-model (# IFs: 147, # faults: 1,082)				
	HSF	ESF	NASF	SASF	SWF	HSF	ESF	NASF	SASF	SWF
Test length Impr.	2,400x	4,762x	8,824x	3,750x	566x	1,987x	2,158x	8,571x	5,000x	2,000x
Test coverage	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

AI for Testing, Testing for AI

- **Introduction**
- **AI for Testing**
  - IR Drop
  - Thermal
  - V_{min}
- **Testing for AI**
  - Testing spiking neural network chips
- **Conclusion**

今日重要結論

- AI 非常適合應用在測試上
  - AI擅長快速決定(和測試一樣)
  - AI需要大量有標註資料（測試可以提供）
- 測試AI晶片很重要
  - AI晶片應用於高度安全需要
  - 傳統測試技術不一定可用

  （沒有scan，可能有多種組態，可能有多種答案）
- 運用AI於測試上的正確觀念
  - 需要domain knowledge(各位不會失業!)
  - 需要大量資料
  - 需要檢視其正確性

Thank You

IR Drop 課程相關資料可以於教育部課程資料庫下載