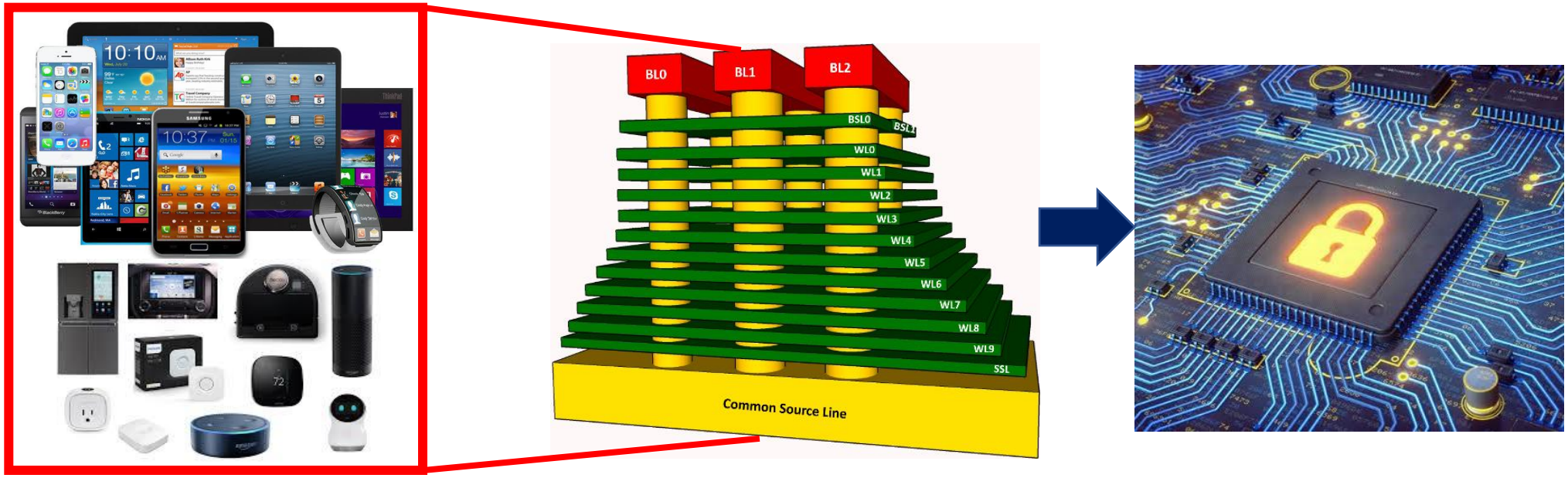


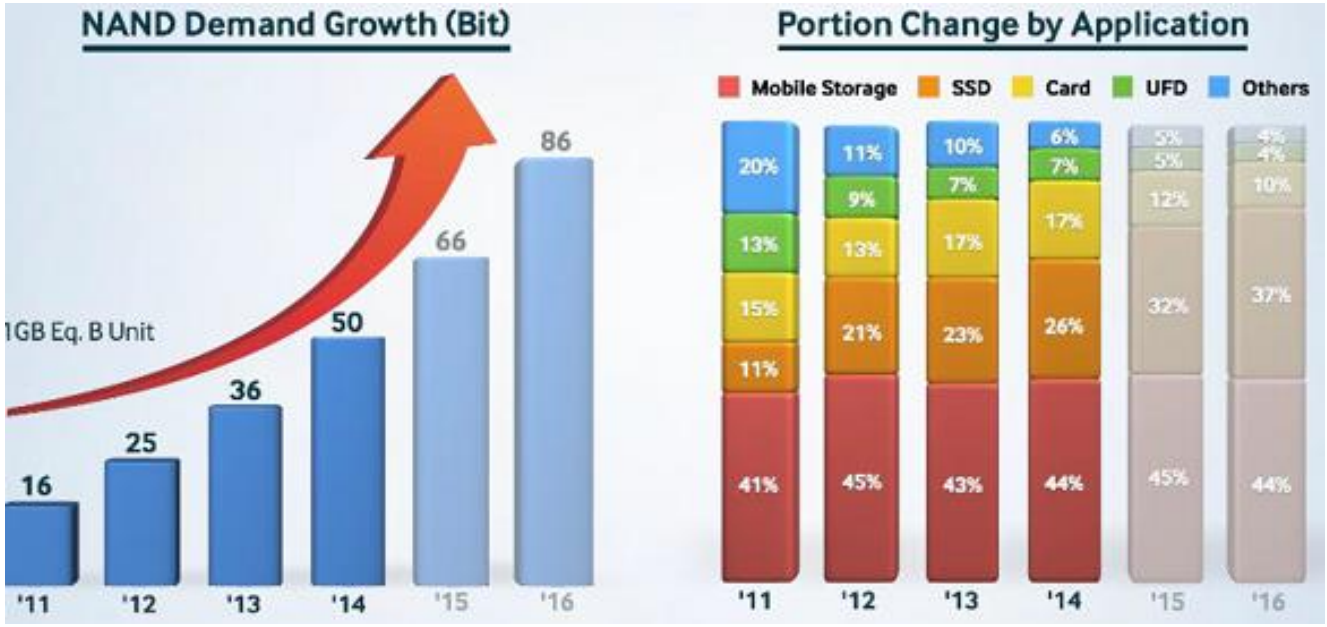
Compact Modeling and Unconventional Applications of 3D-NAND Flash memory

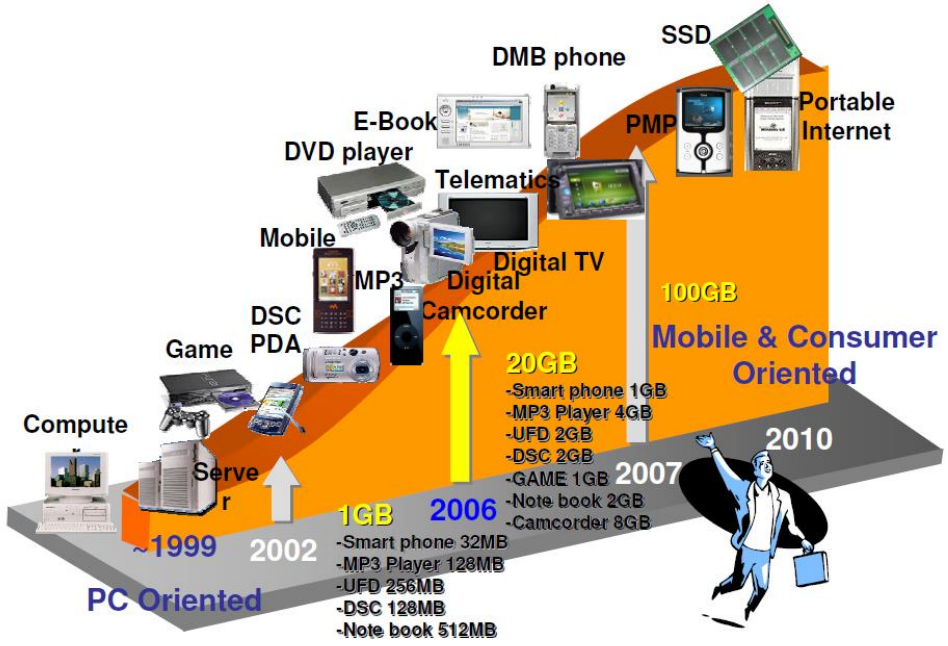
Dr. Shubham Sahay,
Assistant Professor,
Department of Electrical Engineering,
IIT Kanpur

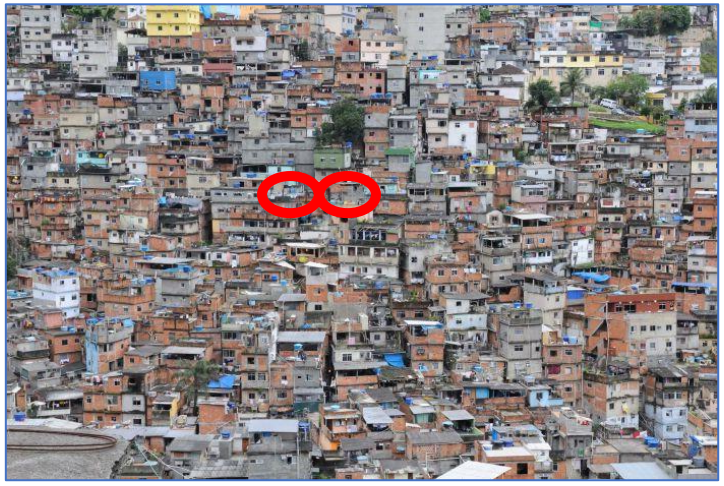
Contents

- Big Data and Flash memory
- 3D NAND flash memory
- Compact modeling Approach
  - Parasitic capacitance extraction
  - Model parameter extraction
- Hardware security primitives
- PUF implementation
- Conclusion

NAND FLASH MEMORY: THE NEXT GOD?



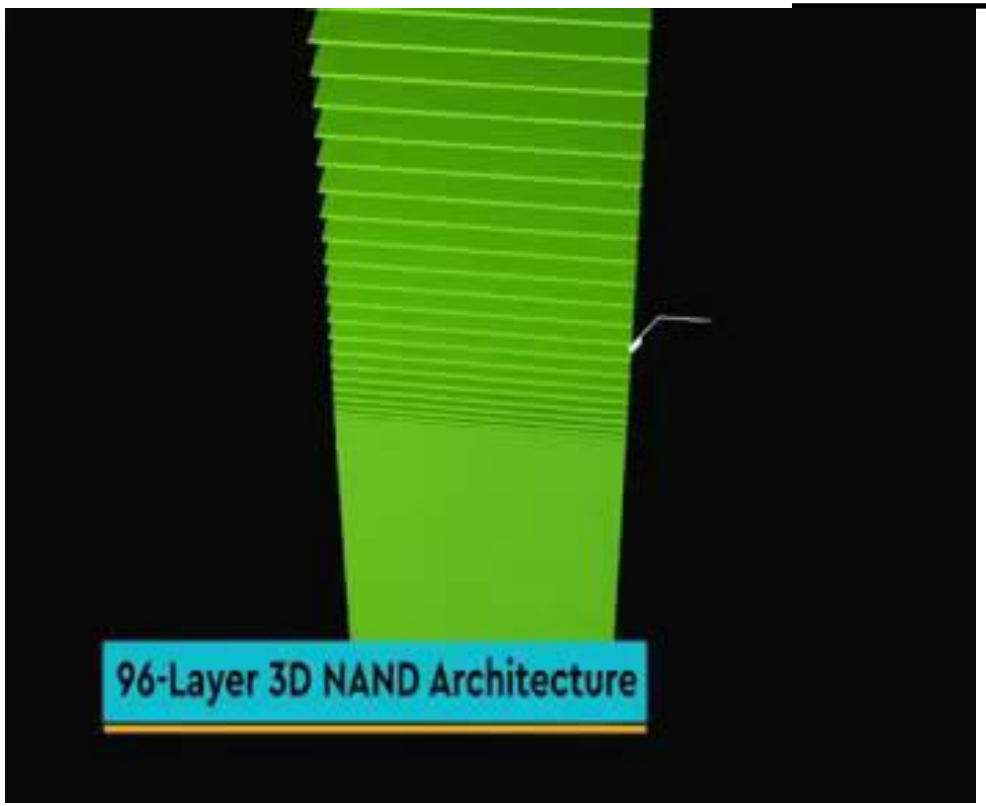


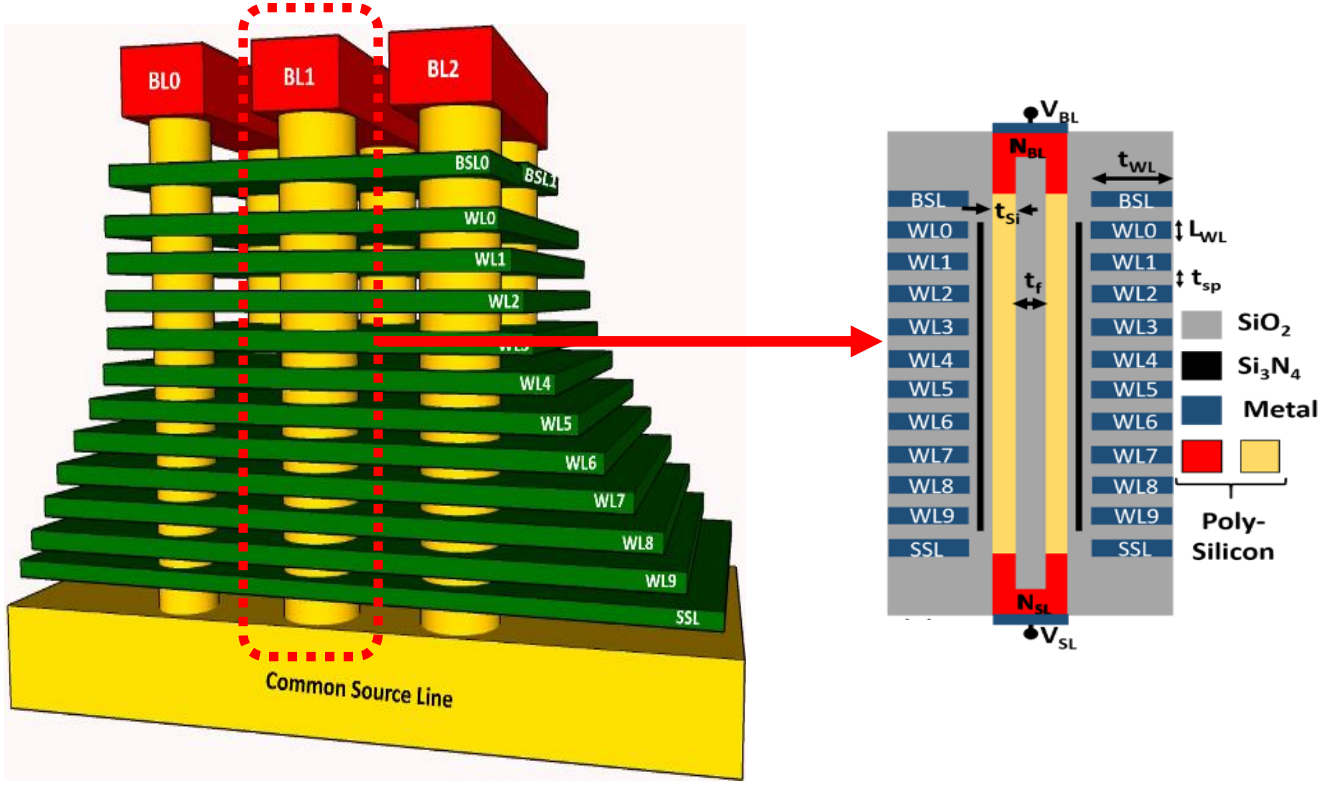


Large neighboring cell disturbance

Cost effective and less cell disturbance

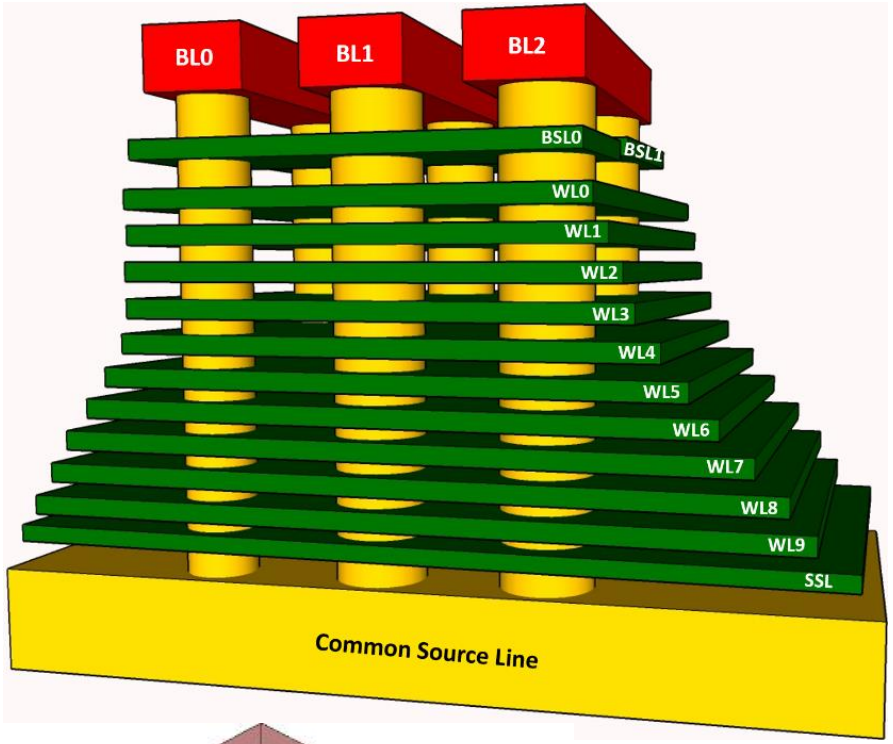
3D NAND FLASH

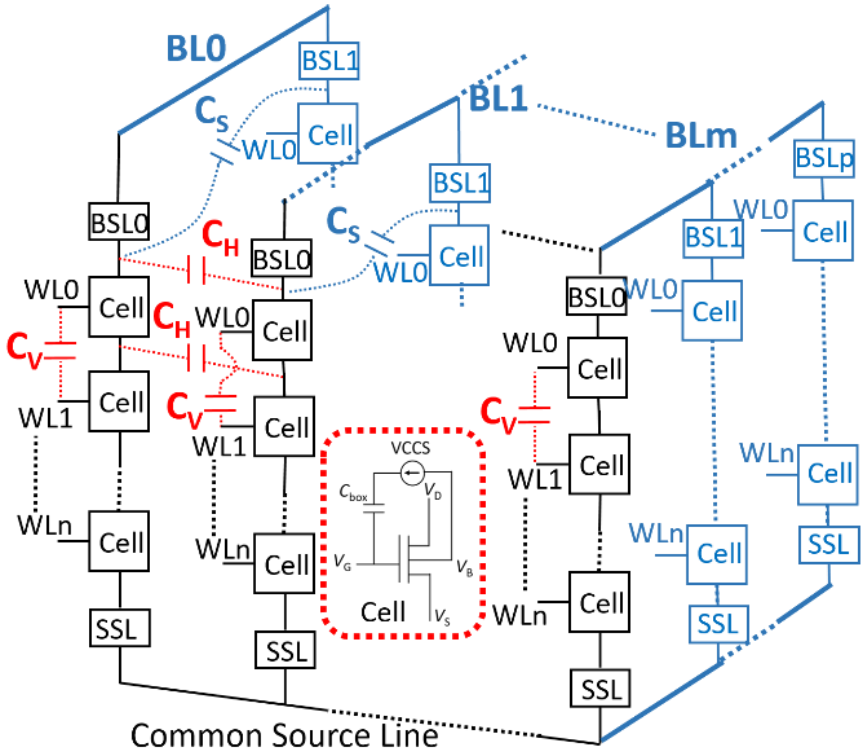


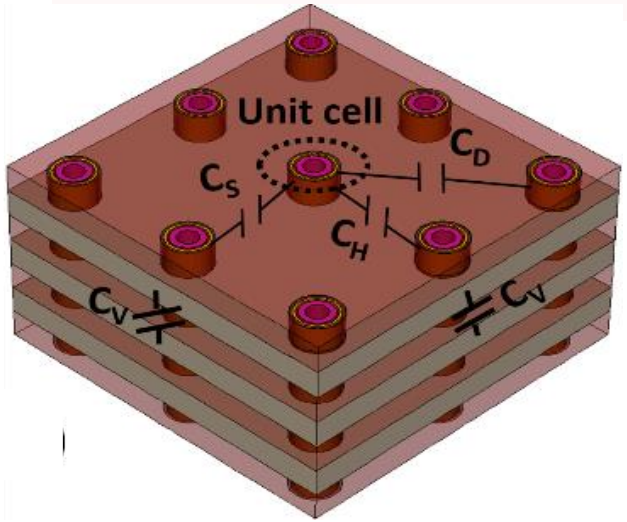


- Punch and plug process: etch holes in stacked metal and insulators, deposit ONO and polysilicon.
- Ultra-high density, Multi-bit capability (QLC), Industries already ≥ 176 layers
- Rarely used for circuit or system level applications till now.
- Limitation: No methodology to extract system level/circuit performance estimates.
- Given the wide range of application and integration with the cyber-physical systems, it becomes essential to explore its potential for unconventional applications.

COMPACT MODEL OF 3D-NAND FLASH







3D-NAND Behavioral Compact Model

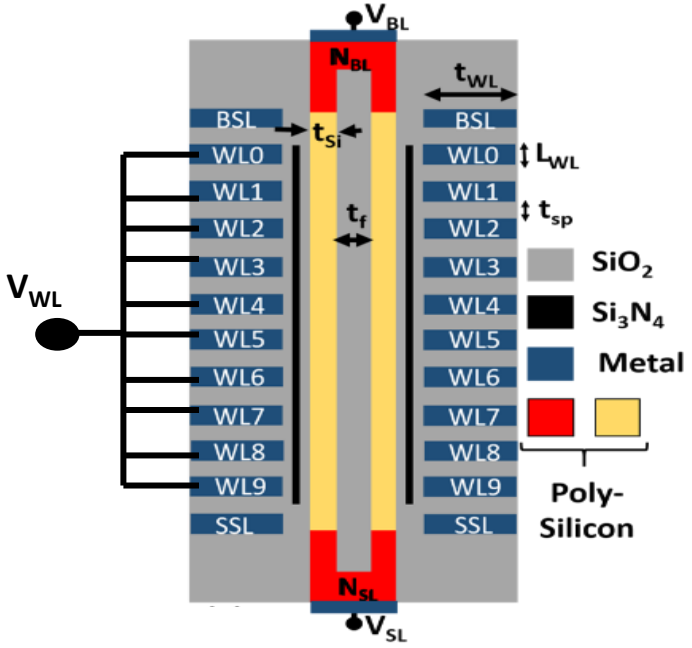
- Individual cells: GAANW (BSIM-CMG) with VCCS and blocking oxide capacitor

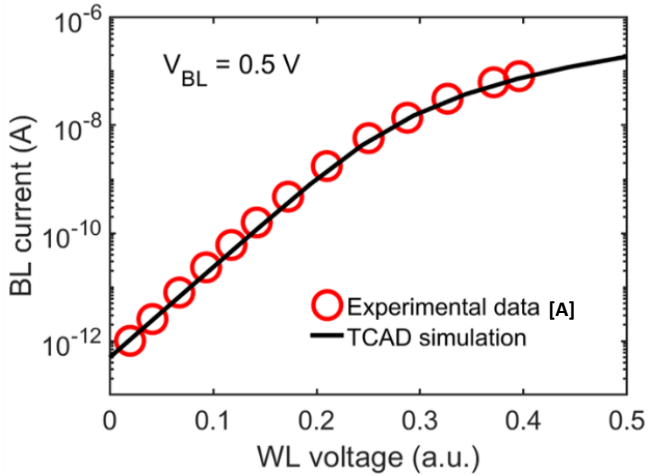
Parasitic Capacitance extraction:

- Parasitic capacitances extracted via mixed-mode TCAD simulations
- Central string in 3×3 string array used for extraction.

S. Sahay and D. Strukov., A behavioral compact model for static characteristics of 3D NAND flash memory, *IEEE EDL*, vol. 40, 2019.

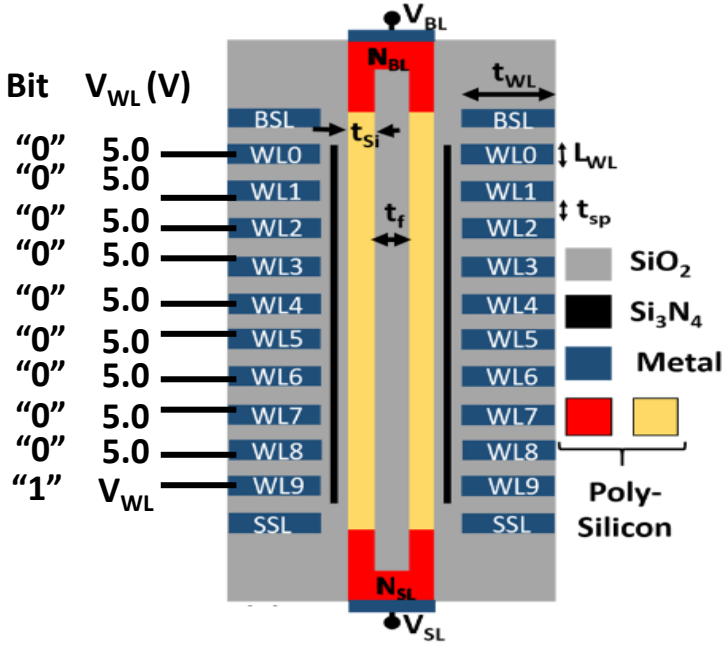
Model Parameter Extraction (Static)



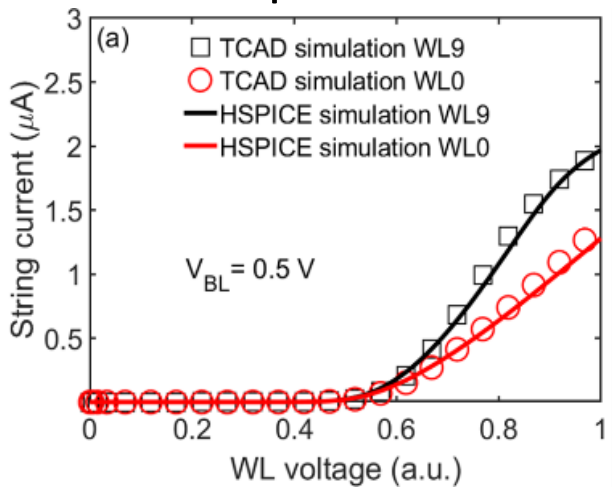


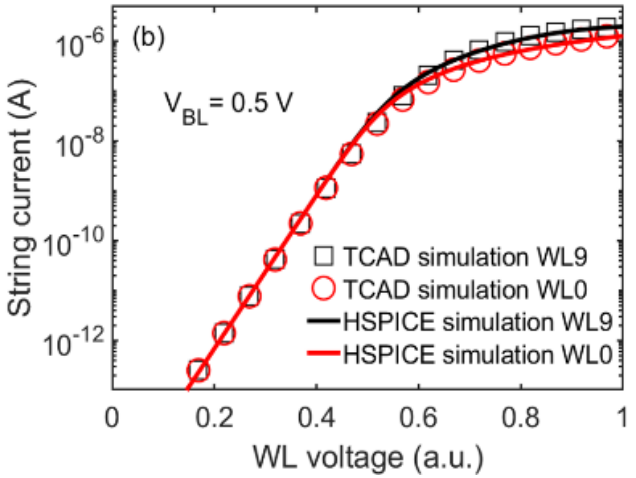
- Cell behavior at different WL layer.
- No experimental data for cells at different WL locations.
- Only experimental data in literature used.
- String current characteristics with 10 WL reproduced in TCAD with all cells active.

[A] D. Resnati et. al., IEEE TED, vol. 65, no. 8, Aug. 2018.

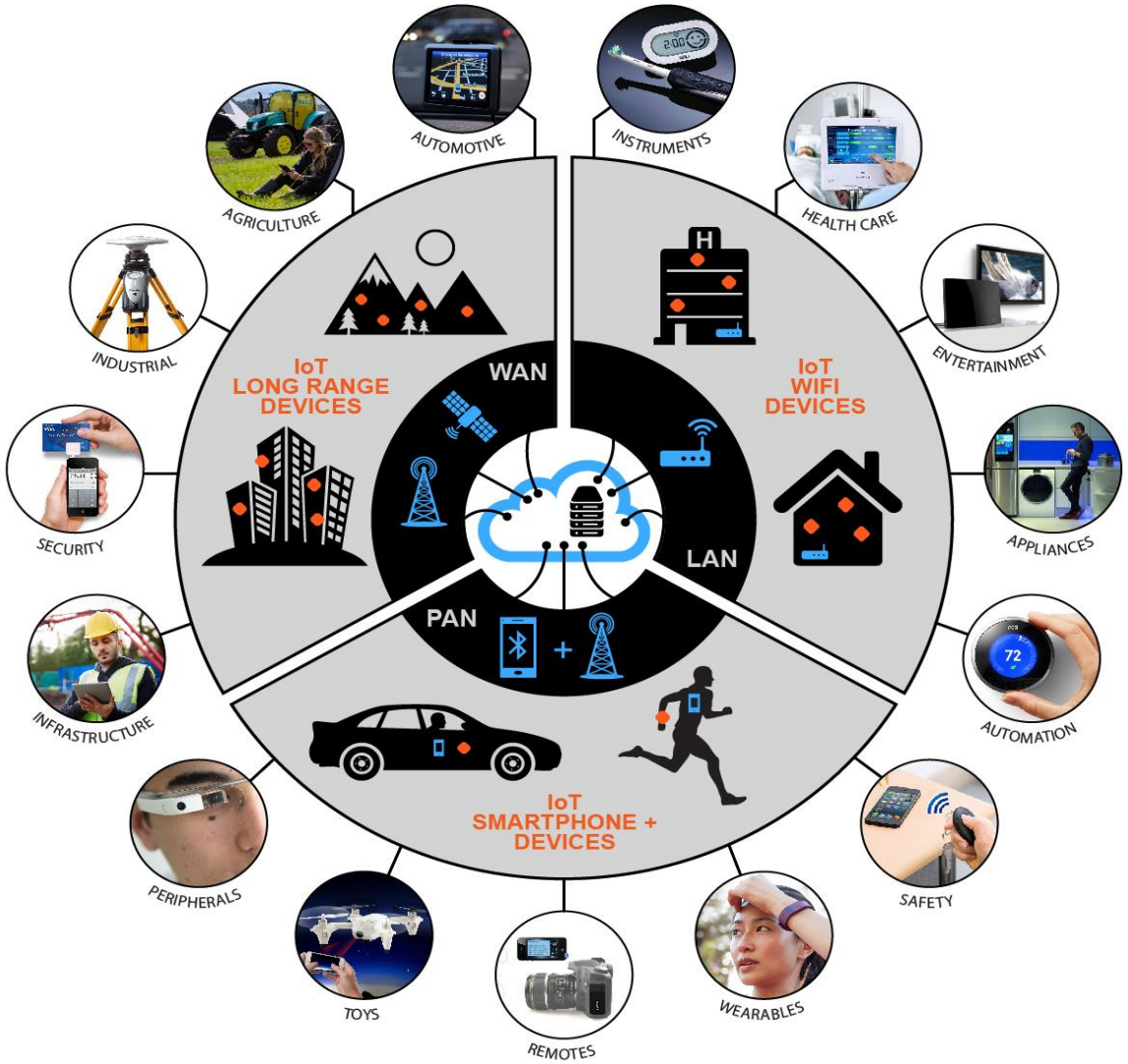


- Unselected cells: pass mode: $V_{WL} = V_{read} = 5.0$ V (bit "0")
- Selected cells: ramp V_{WL}.
- BSIM CMG parameters tuned to reproduce TCAD data.





INTERNET OF THINGS (IOT) ECOSYSTEM



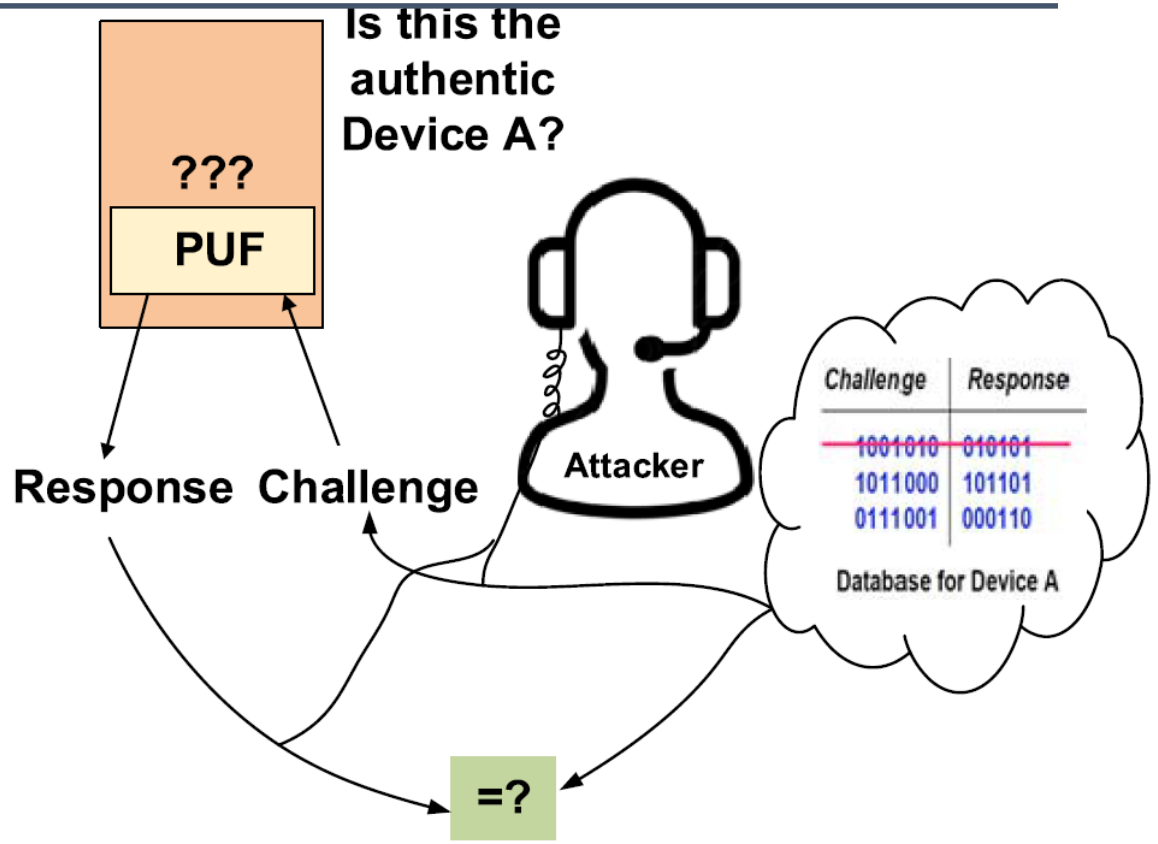
- Each device: cyber-physical system.
- Each device is a potential threat and source of adversary attack.

- ❑ Unauthorized access
- ❑ Data/Identity theft
- ❑ Electronic counterfeiting
- ❑ Reverse engineering
- ❑ Side channel attacks
- ❑ Intellectual property (IP) piracy

- Embed security module in each device itself.
- Hardware security primitives such as Physical unclonable function.

Image source: https://www.rooksecurity.com/building-security-into-iot-development/

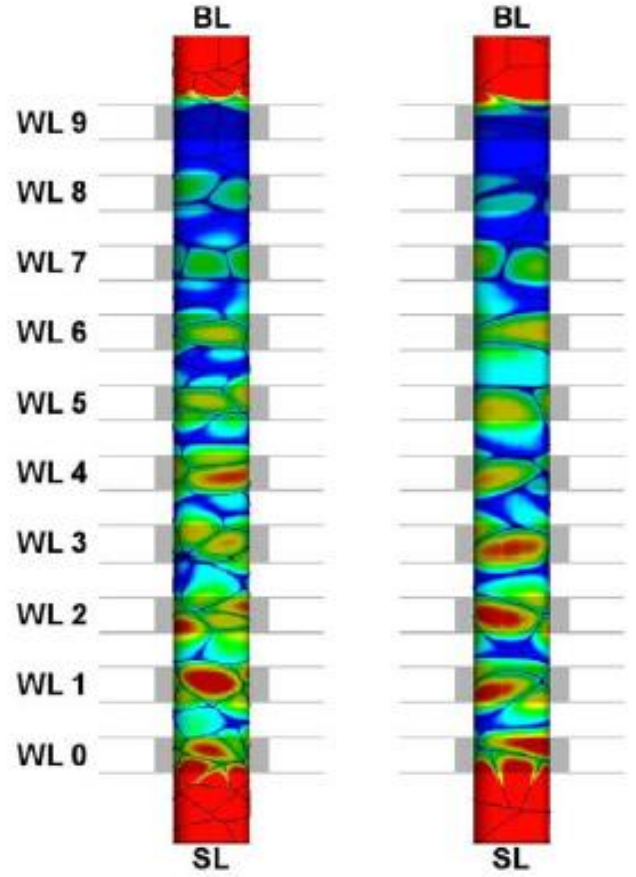
PHYSICAL UNCLONABLE FUNCTION



- PUFs exploit inherent randomness in physical systems.
- Generates unique response when excited with particular input challenge.
- Unique Challenge-response pairs forms basis for identification and authentication.
- Strong PUF: large CRP set, used for device authentication.
- Weak PUF: limited CRP set, used for cryptographic key generation.

INDIVIDUAL CELL VARIABILITY

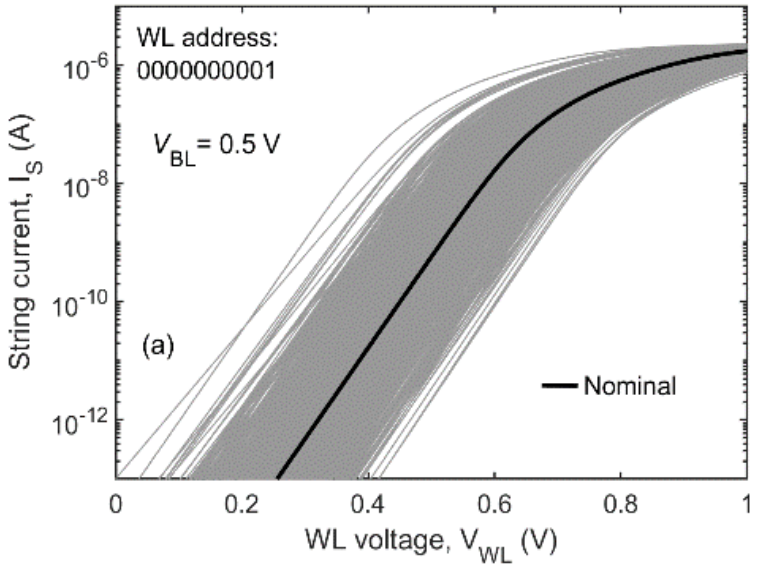
- Process non-uniformity: cell length, width, oxide thickness, etc.
- Random dopant fluctuation.
- Punch and plug process to etch high-aspect ratio trenches
  - Line edge roughness
  - Tapered pipes with different taper angle
- Polysilicon channel
  - Grain boundaries: location and size.
  - Traps: Concentration and location in the energy landscape.
- Significant variability: ECC for memory applications.
- Cell library for Monte-Carlo simulations to characterize the variability in the string current characteristics.

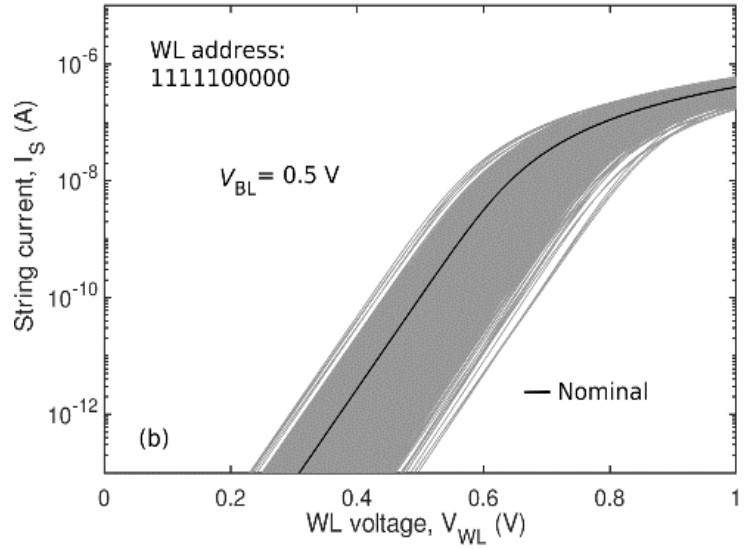


source: D. Resnati, *et al* "Temperature activation of the string current and its variability in 3-D NAND Flash arrays," in *IEDM Tech. Dig.*, Jan. 2017, pp. 4.7.1–4.7.4,

VARIABILITY IN STRING CURRENT CHARACTERISTICS

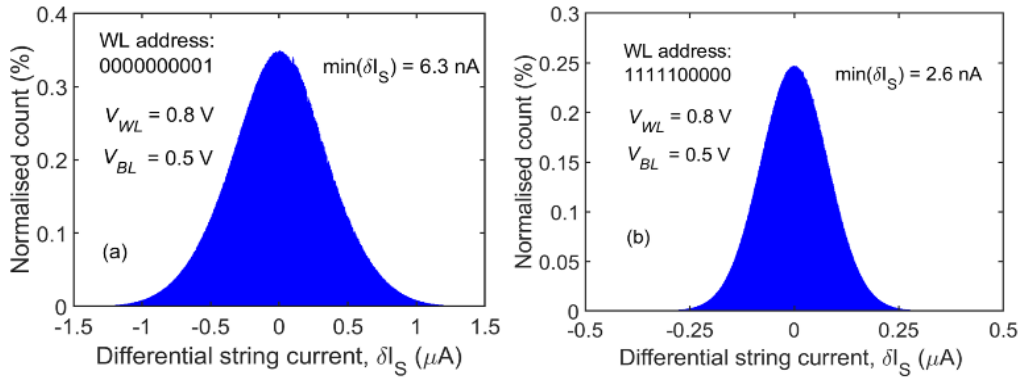
- The intrinsic string current characteristics for 10,000 bit strings was obtained using Monte-Carlo simulations.
- Intrinsic: all cells assumed to be in the same erase-state.
- Threshold voltage and sub-threshold swing varies drastically.
- The characteristics cannot be reproduced even with the aid of partial programming.
- Variation is larger for strings with only one active cell in the WL address.
- Each 3D NAND flash array will have unique variation in string current.
- Even if adversary characterizes one array, it is difficult to reproduce/predict the characteristics for other arrays.

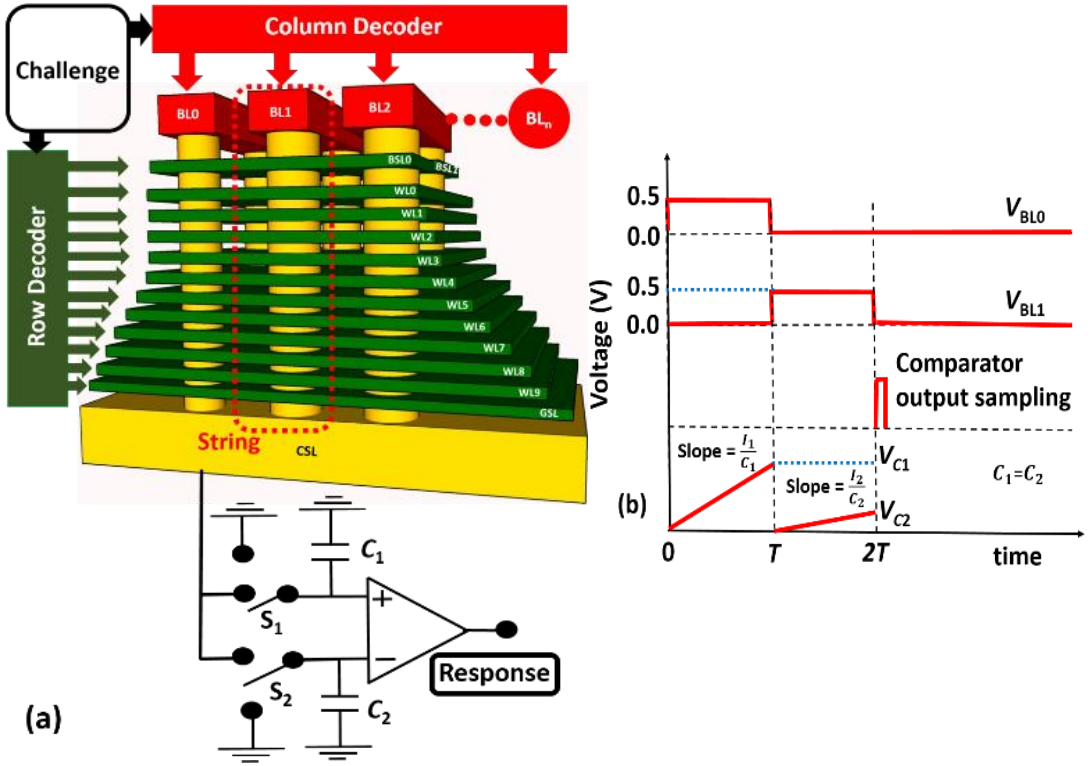




PUF DESIGN

- Intrinsic string current for different strings was sampled at a particular WL voltage of the active cells (V_{WL} = 0.8 V).
- The pair-wise differential string current follows a Gaussian distribution.
- PUF circuit includes a comparator and row and column decoders.
- Bit line (BL) addresses of two different strings is concatenated and fed as input challenge to the PUF for same WL address.
- The string current of the two strings is compared to obtain the response bit.
- String current comparison takes two cycles.





S. Sahay *et. al.*, Hardware Security Primitive Exploiting Intrinsic Variability in Analog Behavior of 3-D NAND Flash Memory Array, *IEEE TED*, vol. 66, no. 5, pp. 2158-2164, May 2019.

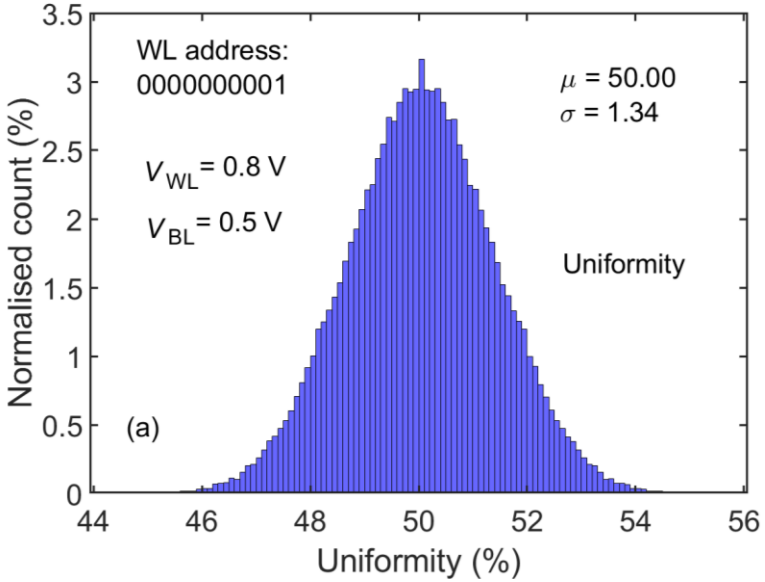
UNIFORMITY

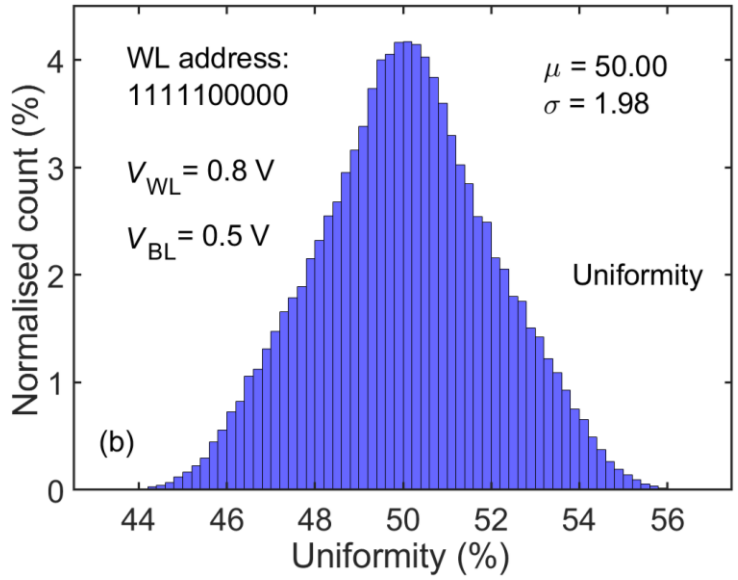
- Uniformity (UF): distribution of bit '0' and '1' in the response bit-stream and is defined as:

$$UF(X_i) = \frac{\sum_{j=1}^{n} x_{i,j}}{n}$$

where $x_{i,j}$ is the j^{th} binary bit in the n-bit response stream of the i^{th} response packet X_i.

- 49,995,000 response bits corresponding to each WL address were collected.
- Single response bits were grouped into packets of 1000 bits each.
- Uniformity exhibits a close to the ideal 50% value.
- Bits '0' and '1' are balanced in the output response bit stream (PUF is unbiased).



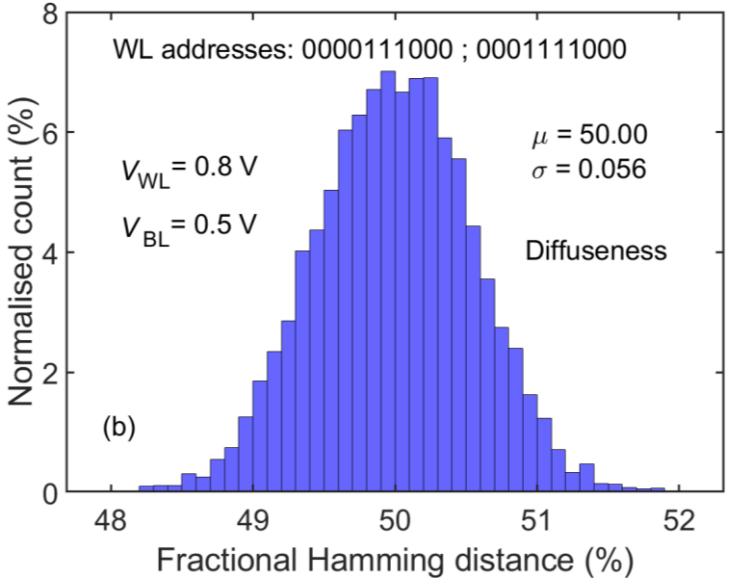


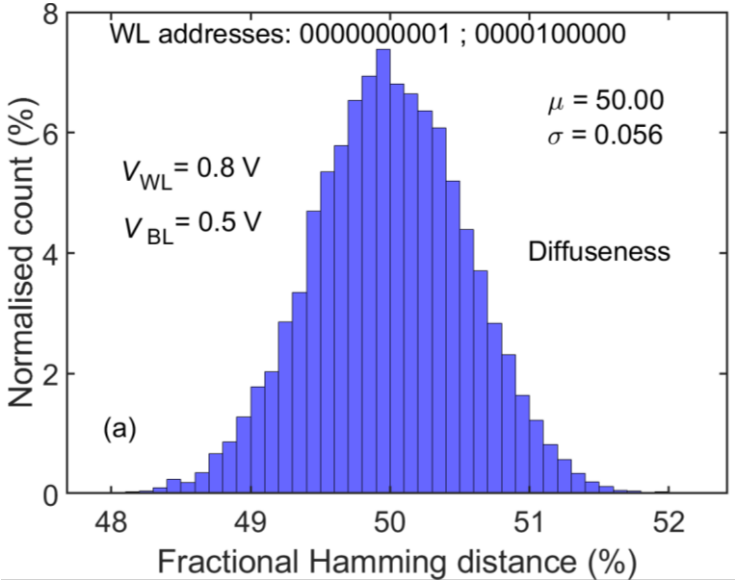
DIFFUSENESS

- Diffuseness (DF): Similarity in the output response bits corresponding to different input challenges to the same PUF instance.
- Diffuseness can be defined as the hamming distance (HD) of the pairwise response pairs given by:

$$DF = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} HD(X_i, X_j)$$

- Different input challenges: different WL address for the bit strings (keeping the BL address same).
- Output response bit stream was grouped into packets of 10,000 bits.
- DF of 50%: response bits of the proposed PUF for different challenges are mutually exclusive.



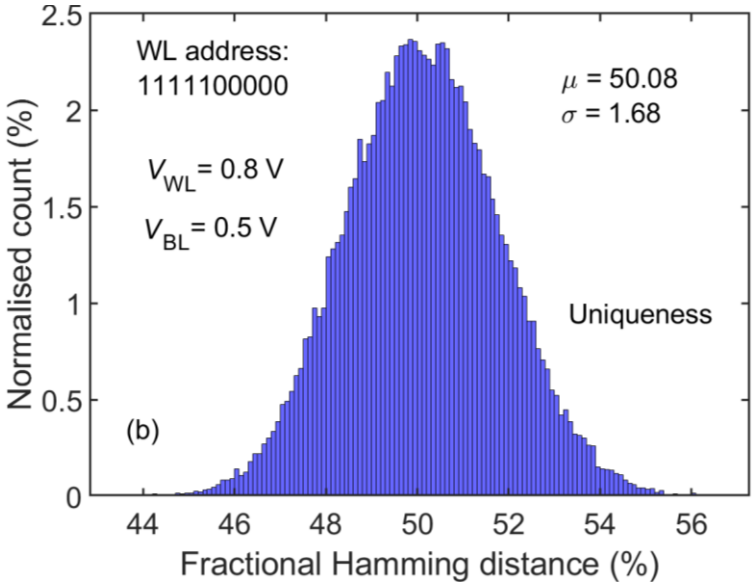


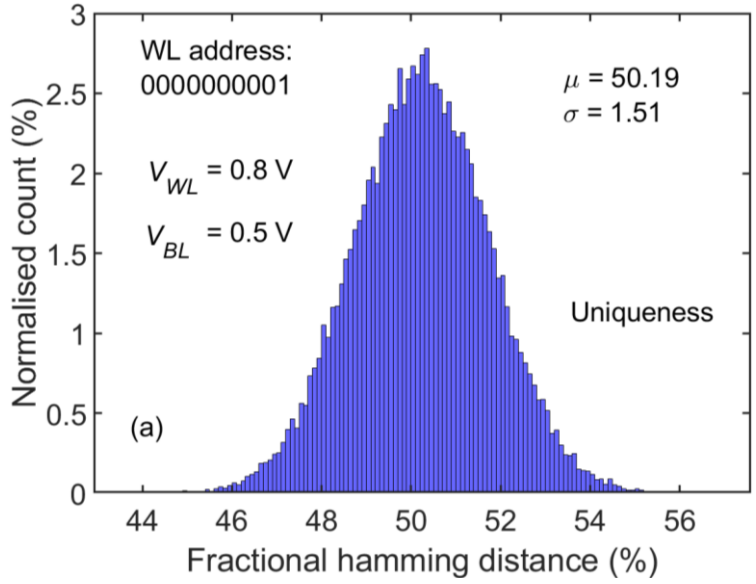
UNIQUENESS

- Uniqueness (UQ): Similarity in the output response of two different PUF instances when excited with same input challenge.
- Uniqueness for '*m*' PUF instances can be given as:

$$UQ = \frac{2}{m(m-1)} \sum_{k=1}^{m} \sum_{l=k+1}^{m} HD(X_i^k, X_i^l)$$

- Different PUF instances were generated by changing the initial random seed for the Monte-Carlo simulations.
- Response bits were grouped in packets of 1000 bits.
- 3D NAND PUF exhibits 50% uniqueness: response packets corresponding to same challenges are mutually exclusive for different PUF instances.



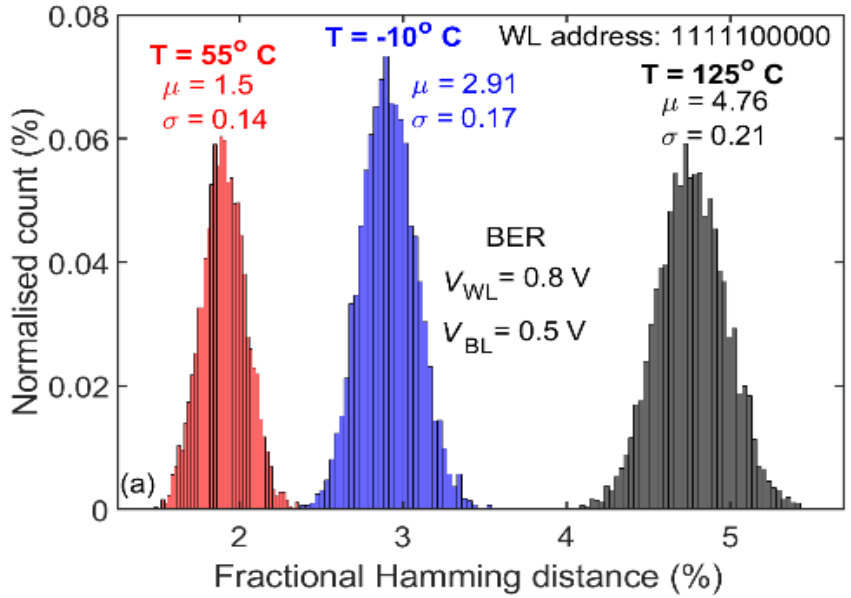


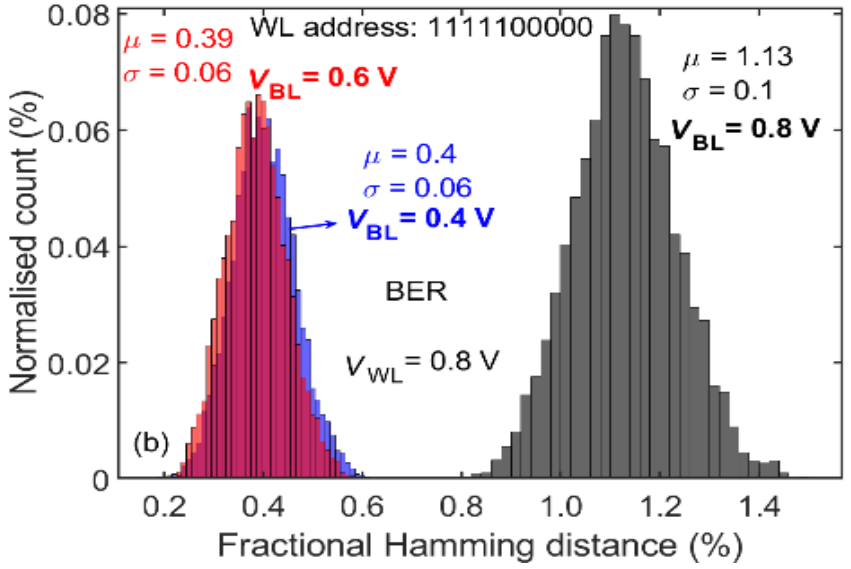
RELIABILITY

- Reliability: reproducibility of same response bits over different measurement instances or different operating conditions.
- Harsh conditions: extreme temperature ranges or fluctuations in the supply voltage.
- Bit error rate is given by:

$$BER = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{HD(X_{i,ref},X_{i,t})}{n}$$

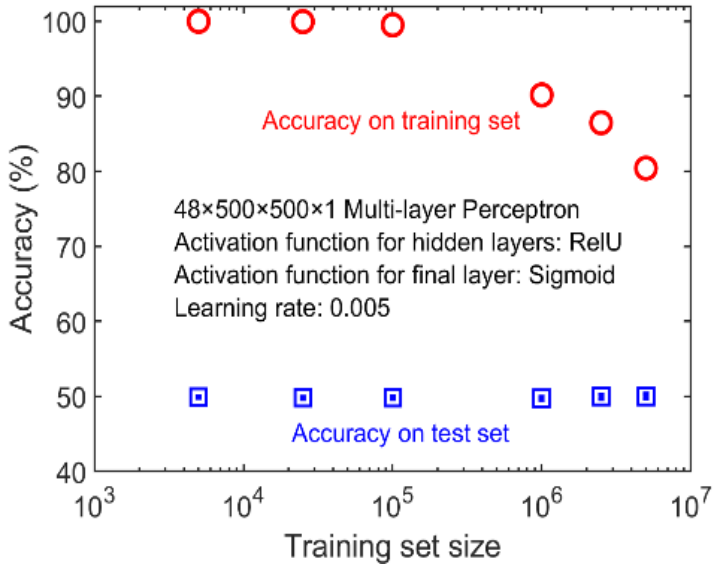
- Results of the Monte-Carlo simulations do not depend on the sampling instances (no time dependence).
- Reliability analyzed only by varying temperature and fluctuations in the BL voltage, independent of the time. (lack of experimental data on drift, stress, etc.)
- BER is ~4.26% at 125 °C and 1.13% for a fluctuation in the BL voltage by 0.3 V (V_{BL} = 0.8 V).





MACHINE LEARNING ATTACKS

- A 48×500×500×1 multi-layer perceptron classifier to test the machine learning and modelling attack resilience.
- Assumption: adversary has access to a limited set of CRPs.
- Given the known CRP set, can we predict the response bits for the unknown challenges?
- 14 bit BL address concatenated with the 10 bit WL address for two bit strings (yielding a 48 bit input) fed as the input.
- For a particular WL address, a subset of the total (49,995,000) CRPs was chosen as the training sample.
- Classifier was then tested on the remaining set of CRPs for that WL address.
- This near-ideal unpredictability (50% accuracy) on training set indicates resilience to the machine learning attacks.



CONCLUSION

- A behavioral compact model was developed for 3D NAND flash memory utilizing industry-standard BSIM CMG model for the first time.
- For the first time, an ultra-dense, ML-attack resilient and strong PUF based on the analog behavior i.e. stochastic string current of 3D NAND flash memory array was demonstrated.
- Considering a random read time (t_R) of ~50 µs, for V_{WL} = 0.8 V and V_{BL} = 0.5 V, the energy consumed while generating one output bit is estimated to vary between ~773 fJ – 10.6 pJ.
- Although the thermal noise voltage floor is significantly low, the flicker noise is significantly higher for 3D NAND flash owing to the polysilicon channel.
- However, due to lack of experimental data on the distribution of the flicker noise spectral density within the different cells in bit string, its impact on PUF performance was not analyzed.
- Similarly, the analysis of 3D NAND PUF with actual variability values of 3D NAND flash, drift, P/E stress is important.

Experimental data is the bottleneck…

Thanks!!

Questions/Feedbacks/Suggestions!!!!