Rethinking PUF Design for Scalable Edge AI: A Position on Balancing ML-Attack Resistance and Real-World Deployment

Gaoxiang Li, Yu Zhuang

Department of Computer Science, Texas Tech University, TX 79409, USA
gaoli@ttu.edu, yu.zhuang@ttu.edu

Abstract

Generative and embedded AI are rapidly migrating from centralized cloud infrastructures to resource-constrained edge devices. While this shift promises reduced latency and improved data privacy, it also creates challenging security and scalability trade-offs. Physical Unclonable Functions (PUFs) are widely touted as low-overhead hardware security primitives suitable for edge and IoT scenarios, yet most existing work emphasizes resistance to machine learning (ML) attacks at the expense of authorized modelability—the ability for trusted entities to accurately model PUF behavior without storing massive Challenge-Response Pair (CRP) databases. This position paper argues that "authorized modelability" should become one of the first-class design objectives for future PUFs. We review existing insights and propose guidelines aimed at balancing ML-attack resistance with the practical requirements of large-scale deployment, thereby addressing a critical yet underexplored aspect of hardware authentication for edge AI.

Introduction

The Drive for Edge Security

The Internet of Things (IoT) continues to expand at a rapid rate, with billions of devices deployed across smart homes, factories, and urban environments (Evans 2011; van der Meulen Gartner, Newsroom, Press Releases, 2017). Many of these devices handle sensitive data or perform mission-critical tasks, yet they must operate within tight power and memory constraints. Physical Unclonable Functions (PUFs) have emerged as a promising hardware security solution in these contexts (Gassend et al. 2002; Herder et al. 2014; Yu et al. 2016). By exploiting device-specific manufacturing variations, PUFs generate unclonable challenge-response behaviors without requiring cryptographic key storage in non-volatile memory (Pappu et al. 2002; Suh and Devadas 2007).

PUFs and the Rise of Machine Learning Attacks

Since the concept of PUFs was introduced, multiple variants (e.g., Arbiter PUFs, Ring Oscillator PUFs) have evolved to prevent increasingly advanced *machine-learning (ML)* attacks (Rührmair and Holcomb 2014; Rostami et al. 2014; Chen et al. 2022). Attackers armed with ML can intercept a subset of challenge-response pairs (CRPs) and train models—often neural networks or even generative frameworks—to replicate PUF behavior (Rührmair et al. 2010, 2013a; Wisiol 2022). In response, PUF developers introduced intricate obfuscation techniques and dynamic transformations (Zhang et al. 2021; Dominguez and Rezaei 2024; Liu et al. 2023) to complicate unauthorized modeling.

Although these methods bolster adversarial resistance, they often overlook or downplay a parallel concern: trusted partners—manufacturers or authentication servers—may also need to model the PUF's behavior to avoid reliance on large-scale CRP databases.

Scalability Challenges

In small-scale deployments, maintaining a dedicated repository of Challenge-Response Pairs for each device on the server is feasible. However, as IoT and edge networks expand to thousands or even millions of devices, this approach becomes untenable (Lim et al. 2005; Zhang and Shen 2021). The overhead of curating and updating massive CRP databases on the server undermines the scalability of PUF technologies. An alternative is to store a "soft model" of each PUF on the server, generating predicted responses on demand (Maes and Verbauwhede 2010; Zalivaka, Ivaniuk, and Chang 2019). This method drastically reduces storage overhead and streamlines provisioning—but hinges on the ability to accurately model the PUF during enrollment (Tun and Mambo 2024).

Complication of ML Attack Resistance on Authorized Modeling

Recent work on PUF has introduced innovative strategies to counter ML-based cloning—ranging from dynamic feedback loops and complex non-linear transformations to "noisy" challenge-response mappings (Chen et al. 2022; Wisiol 2022; Deb Paul, Dasgupta, and Bhunia 2024). While these techniques effectively complicate adversarial modeling, they can also hinder trusted parties from building authorized models if no additional access privileges are granted (Xi 2019; Wisiol 2022). In such cases, some advanced PUF architectures become "unmodelable" not only for attackers but possibly also for legitimate stakeholders.

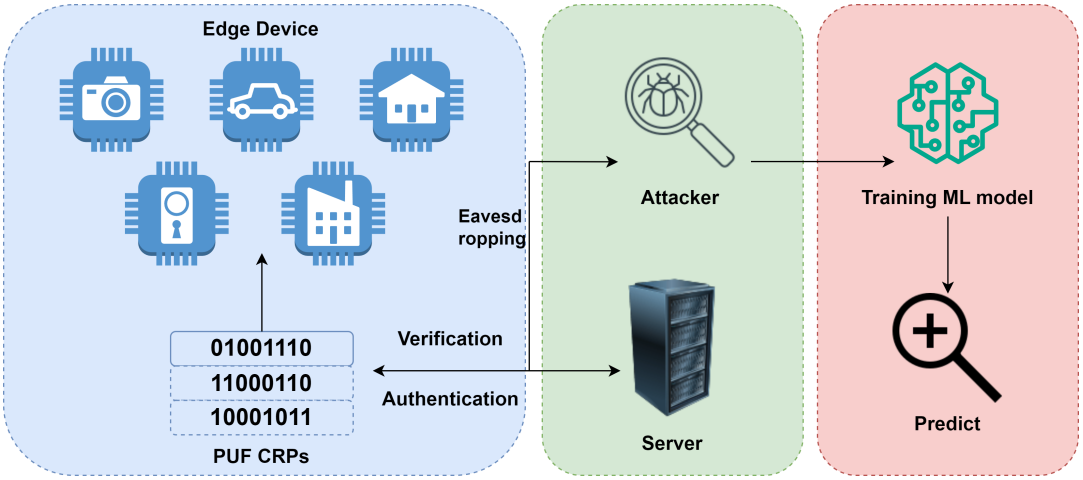

Figure 1: Illustration of machine learning-based modeling attacks on PUFs in an edge-device authentication scenario. The left section represents edge devices that utilize PUF-generated CRPs for authentication. The server verifies the responses from these devices. However, an attacker can eavesdrop on the communication, as shown in the middle section, and collect CRPs. The right section demonstrates how the attacker leverages these intercepted CRPs to train a machine learning model, enabling the prediction of future responses.

Authorized Modelability: An Underexplored Priority and Our Position

Some studies on PUF designs or PUF-based protocols have acknowledged the need for authorized modeling in their works, using approaches of granting trusted partners temporary or limited access to internal PUF states (e.g., through fuse-based enrollment or secure debug modes) (Yu et al. 2016; Babaei and Schiele 2019). But to the best of our knowledge, no work prioritizes authorized modelability as highly as adversarial ML resistance, and less encouragingly, quite some recent PUF design papers give no consideration of this issue.

Security against ML attacks is surely of paramount priority, but we argue that authorized modelability is of equal, or at least, close importance. Not only does large-scale IoT deployment necessitate authorized modeling as elaborated in the paragraph of Scalability Challenges, but for any scale IoT system, large or small, the CRP database approach mandated by the lack of authorized modelability will enable replay attacks (Yu et al. 2016) unless each device is equipped with sufficient non-volatile memory (NVM) for all challenges needed for the entire operational lifespan. Replay attacks would enable adversarial parties to send instructions to devices by masquerading as their legitimate partners, leading to another security issue deserving equal or similar attention as ML attacks.

Thus, our position is that authorized modelability is an underexplored design priority but of equal importance with ML-attack resistance.

Toward Authorized Modelability: Enabling Technical Approaches

With our position stated, we therefore, challenge the community in the exploration of enabling techniques for authorized modelability, to (i) go deeper via assessing existing approaches to enable the selection of best matches for different PUFs and devices, and (ii) go broader by exploring new techniques to better fit current or future IoT devices.

Existing Approaches for Authorized Modelability

Fuse-Based Time-Limited Calibration: Past research explores hardware fuses (e.g., eFuses, anti-fuses) that grant device manufacturers a transient, "white-box" view during production (Yu et al. 2016; Babaei and Schiele 2019). In this calibration phase, sufficient Challenge-Response Pairs or partial internal signals can be collected to build a server-side "soft model." Following enrollment, device manufacturers or owners lock the fuse, relegating attackers to a black-box interface.

Open Questions: How can fuse circuitry be engineered to prevent attackers from exploiting residual debug access? Could advanced tamper-detection or cryptographic protocols lock fuse interfaces irreversibly without risking false lockouts? Answering these questions requires both hardware-level reliability studies and robust security analyses.

Selective Obfuscation or Cryptographic Wrappers: Some PUFs employ cryptographic overlays—e.g., hashed or encrypted outputs—to thwart black-box modeling (Maes and Verbauwhede 2010; Yu et al. 2016). By selectively disabling these wrappers in a secure enrollment phase, the manufacturer can observe relatively transparent responses. Yet, ensuring that this "lower security" mode never reactivates post-deployment presents an ongoing challenge.

Open Questions: Could dynamic cryptographic toggles track device state transitions, ensuring that an authorized debugging session cannot be retriggered in the field? Do we risk openning up new side channels when toggling between

normal and debug modes?

Configured Challenge Spaces: Another line of work proposes configurations of the challenge space for different levels of obfuscation (Zhuang, Li, and Mursi 2022; Xu et al. 2023). In practice, "flagged" challenges could yield near-raw PUF responses during manufacturer calibration, while standard challenges remain fully obfuscated. This design requires robust authentication of the calibrating party and tamper-evident transitions back to obfuscated operation.

Open Questions: How can we securely bind "flagged" challenge subsets to authorized calibration? What cryptographic protocols are necessary to prevent reusing these subsets as an attack vector? Could partial or dynamic revocation of flagged challenges further secure the device after the enrollment phase?

New Technical Approaches

Leveraging Side-Channel Information: Side-channel leakage is typically studied as a vulnerability exploitable by adversaries, and has also been considered difficult or even practically infeasible for attackers to implement (Xu and Burleson 2014; Zalivaka, Ivaniuk, and Chang 2019) for requiring expensive specialized instruments, well-timed accurate measurements, and physical proximity (Mahmoud et al. 2013; Rührmair et al. 2013b; Wei et al. 2014; Becker and Kumar 2014; Becker 2015; Delvaux and Verbauwhede 2013; Gao et al. 2023). However, we can't help speculating that side-channel can possibly offer a new approach to authorized modeling, where trusted entities, under controlled conditions, are allowed to exploit side-channel data to model PUF behavior. For example, accurate measurements of timing or power consumption with high-precision instruments and privileged access in a secure environment might reveal subtle and reproducible patterns in a PUF's behavior that are not easily accessible to remote attackers who lack physical proximity or and special access privilege. However, the following investigations are needed:

- to understand what side-channel information, power consumption, infrared measures, electromagnetic traces, or others, offer stable and reproducible data suitable for model building for legitimate parties while inexpensively configurable to defeat adversarial access; or
- to design ephemeral side-channel interfaces to provide one-time legitimate access for model building during the enrollment in a secure environment while guaranteeing no post-deployment access to any party.

Leveraging PUF Reliability Information: Another new approach to authorized modeling, we observe, can come from reliability-based modeling attacks, in which, each challenge used in an attack is repeatedly fed to the PUF to obtain multiple responses for every challenge, and variations of the responses due to minuscule environmental changes provide attackers more information that can enable ML cloning of a PUF that conventional ML attacks cannot break. Since reliability-based modeling attacks can be thwarted by lockdown schemes (Yu et al. 2016), we are optimistic that they can offer a mechanism for authorized modeling by allowing trusted parties to repeatedly query the PUF with each challenge during enrollment but installing a lockdown scheme afterward. Of course, investigations are needed to figure out implementation details and potential risks.

New Approaches Sorely Needed: Besides the two aforelisted potential approaches, we believe that substantially more work is needed to look for new and different technical approaches to enable authorized modelability. How to ensure adversary resistance while addressing the need of authorized modelability is a challenging issue that may call for investigations from theoretical cybersecurity, PUF circuit development, protocol design, and a lot more areas.

Conclusion

As ML techniques become increasingly central to both attacking and defending Physical Unclonable Functions, the importance of authorized modelability is not to be ignored. While substantial efforts have been made to prevent adversarial ML-based cloning, inadequate attention, in our view, has been given to how trusted partners can reliably model PUF behavior. In large-scale edge-IoT deployments, where efficient device authentication is critical, neglecting authorized modelability undermines the very benefits that make PUFs attractive.

In this position paper, we argue that authorized modelability should be elevated to a design priority equal to adversarial attack resistance. We have listed some existing techniques and also pointed to some directions that may offer new approaches. But all of these need further investigations to identify strengths and weakness as well as conditions under which they are effective or risk-prone. By addressing the needs and challenges of authorized modelability, we believe PUF technologies can better support secure and scalable edge deployments.

Acknowledgments

This work was supported in part by the National Science Foundation under grant No. 2103563.

References

Babaei, A.; and Schiele, G. 2019. Physical unclonable functions in the internet of things: State of the art and open challenges. *Sensors*, 19(14): 3208.

Becker, G. T. 2015. The gap between promise and reality: On the insecurity of XOR arbiter PUFs. In *International Workshop on Cryptographic Hardware and Embedded Systems*, 535–555. Springer.

Becker, G. T.; and Kumar, R. 2014. Active and passive side-channel attacks on delay based PUF designs. *Cryptology ePrint Archive*.

Chen, Z.; Lee, W.; Hong, Q.; Gu, C.; Guan, Z.; Ding, L.; and Zhang, J. 2022. A Lightweight and Machine-Learning-Resistant PUF Using Obfuscation-Feedback-Shift-Register. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(11): 4543–4547.

Deb Paul, S.; Dasgupta, A.; and Bhunia, S. 2024. FDPUF: Frequency-Domain PUF for Robust Authentication of Edge

Devices. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(11): 3479–3490.

Delvaux, J.; and Verbauwhede, I. 2013. Side channel modeling attacks on 65nm arbiter PUFs exploiting CMOS device noise. In *2013 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, 137–142.

Dominguez, M.; and Rezaei, A. 2024. CycPUF: Cyclic Physical Unclonable Function. In *2024 Design, Automation Test in Europe Conference Exhibition (DATE)*, 1–6.

Evans, D. 2011. The Internet of Things How the Next Evolution of the Internet Is Changing Everything. *Cisco White Paper*, 2–3.

Gao, Y.; Yao, J.; Pang, L.; Yang, W.; Fu, A.; Al-Sarawi, S. F.; and Abbott, D. 2023. MLMSA: Multi-Label Multi-Side-Channel-Information enabled Deep Learning Attacks on APUF Variants. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

Gassend, B.; Clarke, D.; Van Dijk, M.; and Devadas, S. 2002. Controlled physical random functions. In *18th Annual Computer Security Applications Conference, 2002. Proceedings.*, 149–160. IEEE.

Herder, C.; Yu, M.-D.; Koushanfar, F.; and Devadas, S. 2014. Physical unclonable functions and applications: A tutorial. *Proceedings of the IEEE*, 102(8): 1126–1141.

Lim, D.; Lee, J. W.; Gassend, B.; Suh, G. E.; Van Dijk, M.; and Devadas, S. 2005. Extracting secret keys from integrated circuits. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 13(10): 1200–1205.

Liu, Y.; Li, J.; Qu, T.; and Dai, Z. 2023. CBDC-PUF: A Novel Physical Unclonable Function Design Framework Utilizing Configurable Butterfly Delay Chain Against Modeling Attack. *ACM Trans. Des. Autom. Electron. Syst.*, 28(5).

Maes, R.; and Verbauwhede, I. 2010. Physically unclonable functions: A study on the state of the art and future research directions. *Towards Hardware-Intrinsic Security: Foundations and Practice*, 3–37.

Mahmoud, A.; Rührmair, U.; Majzoobi, M.; and Koushanfar, F. 2013. Combined modeling and side channel attacks on strong PUFs. *Cryptology ePrint Archive*.

Pappu, R.; Recht, B.; Taylor, J.; and Gershenfeld, N. 2002. Physical one-way functions. *Science*, 297(5589): 2026–2030.

Rostami, M.; Majzoobi, M.; Koushanfar, F.; Wallach, D. S.; and Devadas, S. 2014. Robust and reverse-engineering resilient PUF authentication and key-exchange by substring matching. *IEEE Transactions on Emerging Topics in Computing*, 2(1): 37–49.

Rührmair, U.; and Holcomb, D. E. 2014. PUFs at a glance. In *Proceedings of the conference on Design, Automation & Test in Europe*, 347. European Design and Automation Association.

Rührmair, U.; Sehnke, F.; Sölter, J.; Dror, G.; Devadas, S.; and Schmidhuber, J. 2010. Modeling attacks on physical unclonable functions. In *Proceedings of the 17th ACM conference on Computer and communications security*, 237–249.

Rührmair, U.; Sölter, J.; Sehnke, F.; Xu, X.; Mahmoud, A.; Stoyanova, V.; Dror, G.; Schmidhuber, J.; Burleson, W.; and Devadas, S. 2013a. PUF modeling attacks on simulated and silicon data. *IEEE transactions on information forensics and security*, 8(11): 1876–1891.

Rührmair, U.; Xu, X.; Sölter, J.; Mahmoud, A.; Koushanfar, F.; and Burleson, W. 2013b. Power and timing side channels for PUFs and their efficient exploitation. *Cryptology ePrint Archive*.

Suh, G. E.; and Devadas, S. 2007. Physical unclonable functions for device authentication and secret key generation. In *2007 44th ACM/IEEE Design Automation Conference*, 9–14. IEEE.

Tun, N. W.; and Mambo, M. 2024. Secure PUF-based authentication systems. *Sensors*, 24(16): 5295.

van der Meulen, R. Gartner, Newsroom, Press Releases, 2017. Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016.

Wei, S.; Wendt, J. B.; Nahapetian, A.; and Potkonjak, M. 2014. Reverse engineering and prevention techniques for physical unclonable functions using side channels. In *Proceedings of the 51st Annual Design Automation Conference*, 1–6.

Wisiol, N. 2022. Towards attack resilient delay-based strong PUFs. In *2022 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 5–8. IEEE.

Xi, X. 2019. *Modeling attack resistant strong physical unclonable functions: design and applications*. Ph.D. thesis.

Xu, C.; Zhang, L.; Law, M.-K.; Zhao, X.; Mak, P.-I.; and Martins, R. P. 2023. Modeling Attack Resistant Strong PUF Exploiting Stagewise Obfuscated Interconnections With Improved Reliability. *IEEE Internet of Things Journal*.

Xu, X.; and Burleson, W. 2014. Hybrid side-channel/machine-learning attacks on PUFs: A new threat? In *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 1–6. IEEE.

Yu, M.-D.; Hiller, M.; Delvaux, J.; Sowell, R.; Devadas, S.; and Verbauwhede, I. 2016. A lockdown technique to prevent machine learning on PUFs for lightweight authentication. *IEEE Transactions on Multi-Scale Computing Systems*, 2(3): 146–159.

Zalivaka, S. S.; Ivaniuk, A. A.; and Chang, C.-H. 2019. Reliable and Modeling Attack Resistant Authentication of Arbiter PUF in FPGA Implementation With Trinary Quadruple Response. *IEEE Transactions on Information Forensics and Security*, 14(4): 1109–1123.

Zhang, J.; and Shen, C. 2021. Set-Based Obfuscation for Strong PUFs Against Machine Learning Attacks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(1): 288–300.

Zhang, J.; Shen, C.; Guo, Z.; Wu, Q.; and Chang, W. 2021. CT PUF: Configurable tristate PUF against machine learning attacks for IoT security. *IEEE Internet of Things Journal*, 9(16): 14452–14462.

Zhuang, Y.; Li, G.; and Mursi, K. T. 2022. A Permutation Challenge Input Interface for Arbiter PUF Variants Against

Machine Learning Attacks. In *2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 418–421.