WARDEN: Warranting Robustness Against Deception in Next-Generation Systems

Hazar Yueksel [1] Ramon Bertran [1] Alper Buyuktosunoglu [1]

1 Problem Statement

Power-management system design. The complexity of effective power-management system design, which involves engineering numerous components and hardware-software interfaces across the computing stack, creates an attack space for adversaries (Tang et al., 2018; Vega et al., 2017). The components in today's computing systems demonstrate highly optimized power-efficient designs, which may, for example, include heterogeneous architectures and per-core frequency/voltage islands, and very fine-grained software control of the frequency and voltage settings that are designed to be extremely sensitive to maximize performance. Unfortunately, the dynamic voltage frequency scaling interfaces to power-management hardware can be abused by malicious actors to induce hardware/software faults, infer confidential data, and rewrite these data.

Power oversubscription in data centers. In addition to the complexity introduced by an effective power-management system design, power oversubscription has become a trend in data centers in order to locate more servers on the existing power infrastructure of a data center than it can support when they all operate at the maximum possible power consumption at the same time (Xu et al., 2014). The reason why power oversubscription is possible is that the power consumption of the servers in a data center rarely reaches its maximum. Nevertheless, the possibility that the power consumption of these server peaks remains, which results in a risk of power outages because maximum power consumption produces the overloading of electrical circuits and then triggers the trip of circuit breakers. This can constitute an attack vector since malicious actors can induce power outages by making all servers within a rack reach their maximum power consumption simultaneously.

Power contention. Furthermore, power needs to be treated as a valuable shared resource for systems especially with the emergence of power capping mechanisms and techniques. For example, executing a program with relatively high power consumption on power capped systems can cause "power contention" (Sasaki et al., 2016). Power contention forces the power management system to throttle the system and as a result degrades its performance. This can be a serious concern from both performance and power perspectives; for instance, a victim server with a reasonable power cap for typical applications can observe unexpected performance degradation and wasted power consumption when such intentional power hogs are executed.

Attack-pattern recognition by ML and ECCs. Malicious users of a data center can reverse engineer power-management functions to exploit several power-management design issues. Despite hardware-enforced isolation, all three key security properties can be violated, namely confidentiality, integrity, and availability. Designing effective defenses against malicious actors for a robust and secure system thus requires engineering strong attacks. We propose an attack-pattern recognition system which is powered by machine learning (ML) and which consists of using error-correcting codes (ECCs) in order to detect the malicious workloads, thereby conferring robustness and security to power-management system design.

2 An ML Approach to Attack-Pattern Recognition

Creating images from workload properties. Representing certain properties, such as the estimated power consumption, of subsets of servers in a data center as multidimensional input features of a machine learning architecture helps to answer the question as to whether malicious activities, such as simultaneously occurring power peaks in a data center attributable to malicious actors, can be recognized as attack patterns. Such multidimensional representations result in an input space that bears the information about the location of the servers within a data center. In one setting, a subset of servers in a data center can be treated as a pixel in a two-dimensional (2-D) image according to the location of these servers within the data center, and a chosen property of that subset, such as the average estimated power consumption of the servers in that subset, represents the

[1] IBM Research, Yorktown Heights, NY, USA. Correspondence to: Hazar Yueksel <Hazar.Yueksel@ibm.com>.

Proceedings of the MLSys Workshop on Secure and Resilient Autonomy, Austin, TX, USA, 2020. Copyright 2020 by the authors.

value of each pixel in that 2-D image. Another property of a subset of servers, such as their total storage use, can be used to create another 2-D image.

Neural architecture search. Bundling as many such 2-D images as relevant for the considered threat model by using different properties of the subsets of servers, a set of 2-D input images can be input to a machine learning architecture for attack-pattern recognition. Given the large number of such properties relevant to the comprehensive threat model considered, a neural architecture search is conducted to ensure the successful learning of the attack patterns by using training data obtained from experiments, which are enhanced with manually designed attacks. Under the considered threat model, the number of dimensions of the feature space is reduced by identifying the features most relevant to detecting attack patterns in order to ensure the proper training of the neural network. Once a neural architecture that can recognize attack patterns has been found, the robustness and security of the power-management system is improved significantly by denying service to users who submitted workloads that are recognized with high probability to be malicious.

3 Using ECCs for Protecting Confidential Data

Compromised confidentiality of system security. Malicious actors can develop novel attacks that target numerous independently designed system components in order to breach system security. Fault attacks are more dangerous than side-channel attacks that only violate data integrity because fault attacks can rewrite data, thereby compromising both confidentiality and integrity of system security. For example, both CLKscrew and Rowhammer are microarchitectural fault attacks that actively cause errors during computations by stretching the operating limits of target devices (Tang et al., 2018). In conjunction with the proposed machine learning approach, we use error-correcting codes to detect maliciously inferred and then rewritten confidential data.

Low-complexity codes for confidential data. Using error-correcting codes reduces the attack surface that a malicious actor can exploit by limiting the number of allowed legal codewords, whose representation depends on other chosen codewords when memory is introduced to the encoding process. As the area use and latency of an encoder and decoder are restricted in many systems, we use low-complexity, low-latency codes, such as four-dimensional five-level pulse-amplitude modulation trellis-coded modulation, that jointly encode and modulate information without bandwidth expansion (Ungerboeck, 1982; Hatamian et al., 1998; IEEE, 2015).

Concatenated codes for coding gains. Concatenated to this modulation and encoding scheme, we use a simple Reed–Solomon code without violating the area use and latency constraints of the power-management system to achieve significant coding gains (Reed & Solomon, 1960; Cideciyan et al., 2013). The attack surface is then significantly reduced because all codewords in a local storage unit are dependent on each other. Even when the encoding scheme is known to an attacker, the attacker cannot, without the confidentiality breach being detected, rewrite a portion of confidential data unless all codewords in the local storage are rewritten.

References

Cideciyan, R. D., Gustlin, M., Li, M. P., Wang, J., and Wang, Z. Next generation backplane and copper cable challenges. *IEEE Communications Magazine*, 51(12): 130–136, 2013.

Hatamian, M., Agazzi, O. E., Creigh, J., Samueli, H., Castellano, A. J., Kruse, D., Madisetti, A., Yousefi, N., Bult, K., Pai, P., Wakayama, M., McConnell, M. M., and Colombatto, M. Design considerations for Gigabit Ethernet 1000Base-T twisted pair transceivers. In *IEEE Custom Integrated Circuits Conference*, 1998.

IEEE. Physical coding sublayer, physical medium attachment (PMA) sublayer and baseband medium, type 1000BASE-T. *IEEE Standard 802.3ab*, 2015.

Reed, I. S. and Solomon, G. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304, 1960.

Sasaki, H., Buyuktosunoglu, A., Vega, A., and Bose, P. Characterization and mitigation of power contention across multiprogrammed workloads. In *IEEE International Symposium on Workload Characterization (IISWC)*, pp. 1–10. IEEE, 2016.

Tang, A., Sethumadhavan, S., and Stolfo, S. Motivating security-aware energy management. *IEEE Micro*, 38(3): 98–106, 2018.

Ungerboeck, G. Channel coding with multilevel/phase signals. *IEEE Transactions on Information Theory*, 28(1): 55–67, 1982.

Vega, A., Buyuktosunoglu, A., and Bose, P. Secure swarm intelligence: A new approach to many-core power management. In *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 1–6. IEEE, 2017.

Xu, Z., Wang, H., Xu, Z., and Wang, X. Power attack: An increasing threat to data centers. In *Network and Distributed System Security Symposium (NDSS)*, 2014.