Machine Security

SL5 Novel Recommendations

NOVEMBER 2025

Guy

Lisa Thiergart

Yoav Tzfati

Peter Wagstaff

Luis Cosio

Philip Reiner

Overview

Executive Summary

Capable adversaries have proven time and again they can do the seemingly impossible, often by exploiting vulnerabilities easily overlooked in highly complex systems or by invalidating trusted assumptions. This is especially clear in Machine Security, with examples including Spectre/Meltdown and Bullrun [1]–[4]. Most of security is not about outsmarting the adversary, but minimizing attack vectors and simplifying systems [5]–[8].

The scope of Machine Security is the trust granted by devices to higher-level logic. Due to the specific use-case we are addressing, there are key points we can leverage into novel yet accessible solutions, deferring trust from complex to simple systems and incorporating existing technologies into process design. **We highlight five priority recommendations below**, including several existing industry trends we believe should be accelerated and adopted for SL5.

Top 5 Recommendations

1. Push for AI Accelerators to have stronger security features.
2. Employ verified/measured boot on devices with privileged access in a way trusted individuals can reliably verify.
3. Use simple devices to defer authorization away from complex systems like personal computers.
4. Push for tamper-proof and tamper-evident enclosures that can envelop an entire sensitive scope.
5. Design processes with planned adversarial injections.

Push for AI Accelerators to have stronger security features

Explanation of Recommendation

Labs should push chip designers to include critical security features in the next generation of AI accelerators. Requested features should include an integrated root-of-trust, support for interconnect encryption, and comprehensive tamper-proofing. Given lengthy development cycles, these requests need to be made well before expected availability [9].

Research Reasoning

AI accelerators have historically lacked critical security features, leaving model weights and sensitive data vulnerable to extraction. One example is earlier generations of accelerators (e.g., TPU v3 and v4) that typically lack native link-layer encryption for ICI. For instance, TPU v4 relies on Optical Circuit Switches (OCS) to create "air gapped network isolation" between customers rather than encrypting the data on the interconnects [26]. Consequently, model weights and other sensitive data are transmitted between accelerators in plaintext. These could potentially be intercepted directly from the cables in data centers without any need to interact with the chip itself. The data exists unencrypted not only in the cables, but in the networking components on the chips. This weakness undermines existing security features on the chips. See more about this in the network security memo.

The most important security feature needed on accelerators is an integrated root of trust that enables measured boot capabilities. This root of trust must be the first component to activate at power-on and should perform a measured boot that is attested from a protected key. The tamper-proofing standards should likely meet FIPS 140-3 Level 4 requirements, though currently existing products are only FIPS 140-2 validated, which will expire in September 2026 [11], [12].

An unavoidable challenge is that model weights and sensitive data must exist unencrypted inside the actual processing units during computation [13], [14]. This means that even with a secure root of trust, the compute cores themselves become attractive targets for adversaries. Therefore, comprehensive tamper-proofing must extend beyond just the root of trust to cover all contexts where sensitive data exists in plaintext form. Without this holistic approach to hardware security, adversaries could potentially bypass cryptographic protections by targeting the physical components where data must be processed in its unencrypted state.

By implementing these features at the hardware level, organizations can establish a foundational security layer that protects against both physical and logical attacks on AI infrastructure [10].

Cost-Benefit

- **Costs:** Increased chip design complexity and development time; Higher manufacturing costs due to additional security modules; Potential performance overhead from encryption/decryption operations; Committing to only decrypting model weights at a low level may block some higher-level optimizations.
- **Benefits:** Protection against physical cable-tapping attacks; Cryptographic verification of hardware integrity; This is crucial for trusting other mechanisms like code signing; Foundation for building higher-level security architectures; Competitive advantage as security becomes a key differentiator.

- **Alternatives:** Rely on host-side TEEs/DPUs to encrypt interconnect traffic and gate access, combined with strict key management and physical controls, instead of adding new accelerator features.

Employ verified/measured boot on all devices with privileged access

Explanation of Recommendation

Labs should implement cryptographically verified boot processes on every device with access to sensitive information, including AI accelerators, authorization servers, and employee workstations. A cryptographically robust and tamper-proof attestation process should be included that prevent spoofing verification results. For architectural reference, see [16]–[18].

Research Reasoning

Many devices across the infrastructure stack have privileged access to sensitive information including model weights, algorithmic secrets, and other critical data. These devices exist at different network locations, and many are responsible for authorization and authentication decisions. The distributed nature of these systems creates multiple potential entry points that adversaries could exploit if any single device is compromised.

Consider an authorization server where employee devices connect to receive authentication tokens or keys for accessing model weights (directly or indirectly): This server requires measured boot capabilities implemented in a tested way, with cryptographically solid attestation [15], [19], [20]. Without proper verification, an adversary could potentially modify the authorization server's software to grant themselves access or to log legitimate credentials for later use.

The attestation mechanism itself must be tamper-proof to prevent adversaries who gain physical access from copying private keys and creating their own falsely "verified" attestations. This multi-layered approach to boot verification creates defense in depth. Even if an adversary compromises one aspect of the system, they cannot easily bypass the entire verification chain. The measured boot process should create an immutable record of exactly what software is running, allowing security teams to detect any unauthorized modifications before sensitive operations begin.

Cost-Benefit

- **Costs:** Implementation complexity across diverse device types; Performance impact during boot sequences; Need for secure key management infrastructure; Training requirements for IT staff; Potential compatibility issues with legacy systems; Ongoing maintenance of attestation servers.
- **Benefits:** Detection of rootkits and firmware-level malware; Prevention of unauthorized software modifications; Centralized visibility into device integrity status; Compliance with zero-trust architecture principles; Reduced insider threat potential; Faster incident response through integrity verification.
- **Alternatives:** Use golden-image allowlists, signed packages, and frequent integrity scans with EDR/SIEM monitoring as an interim control, accepting residual firmware risk until measured boot with remote attestation is feasible.

Use simple devices to defer authorization away from complex systems

Explanation of Recommendation

We recommend that AI labs ensure critical authorization decisions do not rely solely on complex, general-purpose computers that present large attack surfaces. Instead, labs should deploy dedicated, simple-hardware devices that handle cryptographic signing and user confirmation for sensitive operations. These devices, potentially as simple as a single die with basic display, signing and input capabilities, should create a buffer between the complex computing environment and the final authorization step, ensuring that compromising a workstation alone cannot result in unauthorized access to sensitive systems.

Research Reasoning

The proposal is to use a physical hardware device ideally consisting of a single die with integrated trusted components capable of cryptographic signing, text display, and basic communication with a computer via simple interfaces like serial or USB. While complex systems will inevitably have exploitable flaws, it is plausible that such a device could not be compromised.

By requiring both the computer and a separate simple device for authorization, the system ensures that compromising the computer alone is insufficient to execute authorization-required actions. The simple device displays information to the user, who then accepts or declines via a physical button, after which the device signs the request and returns it. This creates a physical, verifiable action that cannot be simulated by malware on the compromised computer.

This principle extends beyond individual authorization to other sensitive contexts, such as code review stations where the approved code's integrity must be guaranteed. The simplicity of these devices also enables potential hardware verification. If manufactured using appropriate processes rather than cutting-edge nodes like 5nm, the devices could be scanned and verified using techniques such as **infrared imaging** (IR) or **scanning electron microscopy (SEM)**, with the resulting netlists compared to the original design [22], [23]. This level of hardware transparency is practically impossible with complex general-purpose computers but becomes feasible with purpose-built simple devices.

Cost-Benefit

- **Costs:** Development of custom hardware devices; Manufacturing and distribution costs; Integration with existing authorization workflows; User training for new authorization process; Potential workflow disruption during transition; Physical device management and replacement.
- **Benefits:** Dramatic reduction in attack surface for critical operations; Physical verification of user intent; Immunity to software-based attacks on workstations; Potential for hardware verification and auditing; Clear separation of concerns in security architecture; Enhanced protection against insider threats; Visible security measure that reinforces security culture.
- **Alternatives:** Require multi-party approval (e.g., 2-of-3) using standard FIDO2 security keys [21] or HSM-backed approvals on existing workstations, instead of introducing a new dedicated authorization device. This requires in-depth review of each action.

Support tamper-proof and tamper-evident enclosures that can envelop entire sensitive scopes

Explanation of Recommendation

Labs should use tamper-proof enclosures that extend beyond individual components to protect entire operational contexts: from single chips to complete server racks. These physical security measures must meet stringent certification standards and should be scaled appropriately to the scope of the sensitive operations being protected.

Research Reasoning

Security cannot rely solely on cryptographic protections when sensitive data must exist in plaintext during processing. Instead, we can rely on physical tamper-proofing protections. The current state of tamper-proof technology reveals significant limitations. Only four devices have achieved FIPS 140-2 security level 4 certification, all of which are Hardware Security Modules (HSMs), with three manufactured by IBM [25], though many are likely in the pipeline. While HSMs provide a critical security foundation, they represent only one component in a larger system where the actual computational work occurs elsewhere.

If the HSM is not the component performing the main computational work, then the processing units themselves require protection. An adversary who can access the sensitive context outside the HSM could potentially manipulate what gets signed by the HSM, compromising the entire security model. This highlights a fundamental challenge: cryptographic protections are insufficient when attackers can access the physical components where data must be processed in plaintext.

The scope of tamper-proofing may need to extend far beyond individual components. When interconnects lack encryption, protecting a single chip becomes insufficient. The entire rack might need a tamper-proof enclosure. If network protocols like InfiniBand also lack encryption, the theoretical requirement could extend to tamper-proofing entire floors, which while seemingly infeasible, underscores the critical importance of implementing encrypted interconnects in AI accelerators. This escalating scope of physical security requirements creates a strong economic argument for building security features directly into the hardware rather than attempting to retrofit physical protections around insecure components.

Questions of maintainability and resettability of enclosures also have important security implications to be considered for specific solutions separately. Tamper-proof enclosures that need to be opened for maintenance need the ability to be reset, and may require a way to distinguish between legitimate access and tampering. New tamper-proof technologies may be needed to meet these needs.

Cost-Benefit

- **Costs:** Significant hardware investment for certified enclosures; Custom engineering for different deployment scales; Increased data center space requirements; Complex maintenance procedures requiring security protocols; Higher operational costs for physical security monitoring; Potential cooling and accessibility challenges.
- **Benefits:** Protection against physical attacks and insider threats; Compliance with highest security certification standards; Defense against side-channel attacks;

Visible deterrent to potential adversaries; Complementary layer to cryptographic protections; Preservation of data confidentiality even during processing; Potential for creating secure computation zones.

- **Alternatives:** Layer physical security (cages, cameras, access logs, tamper seals) and encrypting all links and storage to reduce plaintext exposure, avoiding full rack/floor tamper-proof enclosures.

Design processes with planned adversarial injections

Explanation of Recommendation

Labs should implement automated systems that regularly inject benign but realistic adversarial actions into their processes, creating a constant state of vigilance and ensuring that detection and response mechanisms remain sharp. Labs should leverage AI capabilities to craft adversarial content to more realistically gauge the state of security [24].

Research Reasoning

Security procedures often include tripwires designed to detect malicious activity, but these are rarely triggered in practice. This infrequency creates a dangerous situation where security infrastructure becomes worn down by disuse. When real attacks occur, teams may not know how to respond effectively because they lack practical experience with their response procedures.

The proposed solution extends the concept of red teaming by creating automated, continuous adversarial injections. Unlike traditional red teaming, which provides point-in-time assessments, this approach ensures constant readiness. Just as red teams often receive advantages to test layered defenses, these planned injections can simulate various levels of compromise to ensure that security remains effective even when individual controls fail.

Critical to this approach is maintaining clear distinction between planned injections and real attacks. Organizations must design these systems carefully to prevent adversaries from hiding actual attacks among the planned injections. A practical example involves code review processes where language models could automatically insert malicious code to test whether reviewers catch it. This would quickly identify reviewers who tend to "gloss over" code reviews, allowing for targeted training before real malicious code slips through.

Cost-Benefit

- **Costs:** Development of injection systems and automation; Computational resources for generating realistic scenarios; Time spent responding to planned injections; Risk of desensitization if not properly managed; Complexity in distinguishing planned vs. real incidents; Ongoing maintenance and scenario updates.

- **Benefits:** Continuous validation of security controls; Regular practice for incident response teams; Early identification of process weaknesses; Metrics-driven security improvements; Reduced response time during real incidents; Enhanced security awareness across teams; Proactive discovery of detection gaps; Cost-effective alternative to continuous red teaming.

- **Alternatives:** Run scheduled red-team exercises, tabletop drills, and a bug-bounty program with periodic synthetic log replays to validate detections, rather than continuous adversarial injections in production.

Bibliography

[1] P. Kocher et al., "Spectre Attacks: Exploiting Speculative Execution," 2019. [Online]. Available: https://spectreattack.com/spectre.pdf

[2] M. Lipp et al., "Meltdown," 2018. [Online]. Available: https://meltdownattack.com/meltdown.pdf

[3] G. Greenwald, E. MacAskill, and L. Poitras, "Revealed: how US and UK spy agencies defeat internet privacy and security," The Guardian, Sep. 5, 2013. [Online]. Available: https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security

[4] N. Perlroth, J. Larson, and S. Shane, "N.S.A. Foils Much Internet Encryption," The New York Times, Sep. 5, 2013. [Online]. Available: https://www.nytimes.com/2013/09/06/us/nsa-foils-much-internet-encryption.html

[5] J. H. Saltzer and M. D. Schroeder, "The Protection of Information in Computer Systems," 1975. [Online]. Available: https://web.mit.edu/Saltzer/www/publications/protection/

[6] NIST, "Security and Privacy Controls for Information Systems and Organizations," SP 800-53 Rev. 5, 2020. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

[7] NIST, "Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems," SP 800-160 Vol. 1 Rev. 1, 2022. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v1r1.pdf

[8] OWASP, "Attack Surface Analysis Cheat Sheet." [Online]. Available: https://cheatsheetseries.owasp.org/cheatsheets/Attack_Surface_Analysis_Cheat_Sheet.html

[9] Semiconductor Industry Association, "Chip Design and R&D," Jan. 2025. [Online]. Available: https://www.semiconductors.org/policies/chip-design/.

[10] V. Ivanov et al., "SAGE: Software-Based Attestation for GPU Execution," in 2023 USENIX Annual Technical Conference (USENIX ATC 23), 2023. [Online]. Available: https://www.usenix.org/conference/atc23/presentation/ivanov

[11] NIST, "FIPS 140-3 Transition Effort," 2024. [Online]. Available: https://csrc.nist.gov/projects/cryptographic-module-validation-program/fips-140-3-transition-effort

[12] NIST, "Security Requirements for Cryptographic Modules," FIPS PUB 140-3, 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.140-3.pdf

[13] Confidential Computing Consortium, "Confidential Computing: Hardware-based Trusted Execution for Applications and Data," Whitepaper, 2021. [Online]. Available: https://confidentialcomputing.io/wp-content/uploads/sites/6/2021/03/confidentialcomputingprimer.pdf

[14] T. Hunt et al., "Telekine: Secure Computing with Cloud GPUs," in 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20), 2020. [Online]. Available: https://www.usenix.org/conference/nsdi20/presentation/hunt

[15] Microsoft, "Device Health Attestation," Documentation. [Online]. Available: https://learn.microsoft.com/windows/security/operating-system-security/device-health-attestation

[16] IETF, "Remote ATtestation procedureS (RATS) Architecture," RFC 9334, 2023. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9334

[17] Trusted Computing Group, "DICE Attestation Architecture," 2023. [Online]. Available: https://trustedcomputinggroup.org/wp-content/uploads/DICE-Attestation-Architecture-r23-final.pdf

[18] Trusted Computing Group, "TPM 2.0 Library Specification." [Online]. Available: https://trustedcomputinggroup.org/resource/tpm-library-specification/

[19] Trusted Computing Group, "PC Client Platform Firmware Profile Specification." [Online]. Available: https://trustedcomputinggroup.org/resource/pc-client-platform-firmware-profile-specification/

[20] Microsoft, "Windows Defender System Guard: Measured Boot." [Online]. Available: https://learn.microsoft.com/windows/security/information-protection/windows-defender-system-guard/measure-boot

[21] W3C, "Web Authentication: An API for accessing Public Key Credentials Level 2," 2021. [Online]. Available: https://www.w3.org/TR/webauthn-2/

[22] R. Torrance and D. James, "The State-of-the-Art in IC Reverse Engineering," IEEE, 2011. [Online]. Available: https://www.iacr.org/archive/ches2009/57470361/57470361.pdf

[23] Visual 6502, "Visual 6502 Project." [Online]. Available: http://www.visual6502.org/

[24] NIST, "Guide to Test, Training, and Exercise Programs for IT Plans and Capabilities," SP 800-84, 2006. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-84/final

[25] NIST, "CMVP Validated Modules Search: IBM, FIPS 140-2, Level 4, Active," 2024. [Online]. Available: https://csrc.nist.gov/projects/cryptographic-module-validation-program/validated-modules/search?SearchMode=Advanced&Vendor=IBM&Standard=140-2&CertificateStatus=Active&ValidationYear=0&OverallLevel=4

[26] N. P. Jouppi et al., "TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings," arXiv preprint arXiv:2304.01433, p. 4, Apr. 2023.