CSAI Foundation

Cloud Security Alliance AI Safety Initiative

GPUBreach: GDDR6 RowHammer Achieves Full System Compromise

Hardware-Level Privilege Escalation in GPU-Accelerated AI Infrastructure

Unofficial AI-assisted Research

2026-04-08

Key Takeaways

- Three independent research teams disclosed full-chain GPU-to-CPU privilege escalation attacks at IEEE Security and Privacy 2026, all exploiting RowHammer bit-flip vulnerabilities in GDDR6 memory.
- GPUBreach, the most severe of the three, demonstrates a complete attack path from GPU memory bit-flips to a CPU root shell, and critically, bypasses IOMMU protections that were widely assumed to mitigate GPU-origin attacks.
- NVIDIA's consumer RTX line carries no on-die error-correcting code (ECC) by default, leaving a wide swath of gaming, workstation, and cloud-attached GPUs unmitigated.
- A single successful bit-flip targeting model weights in GPU memory can degrade a deep neural network's inference accuracy from 80% to below 0.1% [1], creating a silent and difficult-to-detect corruption pathway in AI inference infrastructure.
- Patches are under investigation as of the time of this writing; organizations running GDDR6-equipped GPUs in multi-tenant or AI production environments should treat this as an active risk requiring immediate compensating controls.

Background

The RowHammer vulnerability class, first disclosed against consumer DRAM in 2014 [15], exploits a physical property of dynamic random-access memory: repeated, rapid activation of a DRAM memory row induces electromagnetic disturbance in adjacent rows, eventually causing bit-flips in those neighboring locations. Over the following decade, researchers demonstrated that these induced bit-flips could be leveraged to break memory isolation guarantees in operating systems, hypervisors, and browser sandboxes. Memory vendors and chipmakers responded with a mix of Target Row Refresh (TRR) mechanisms, on-die ECC, and mitigations baked into CPU memory controllers, and many practitioners assumed CPU-side DRAM mitigations had substantially reduced practical RowHammer risk.

Graphics processing units were a different matter. GDDR memory—the high-bandwidth DRAM family used in discrete GPUs—was designed to maximize throughput rather than reliability, and the mitigation investments that matured on the CPU DRAM side were not mirrored in the GPU ecosystem. Consumer GDDR6 modules do not include on-die ECC, and GPU driver stacks historically lacked the hardening that

CPU-side software received following the original RowHammer disclosures. The academic community began probing this gap in earnest in 2025 [14], and the results arriving in April 2026 confirm that GPU memory is not only vulnerable to RowHammer but can serve as a complete entry point for full system compromise.

The three papers presented at IEEE Security and Privacy 2026—GDDRHammer, GeForge, and GPUBreach—build on a predecessor, GPUHammer, disclosed at USENIX Security 2025 by the same University of Toronto team. Together, these four research papers across two venues represent an unusually concentrated cluster of independent hardware security disclosures elevating GPU RowHammer from theoretical to practically demonstrated privilege escalation. The simultaneous arrival of multiple independent teams at the same findings significantly raises the credibility of the risk and compresses the timeline between disclosure and potential exploitation.

Security Analysis

The RowHammer Attack Chain on GPUs

The foundational enabler for all three April 2026 disclosures is the work published by researchers at the University of Toronto at USENIX Security 2025 [1]. That research, GPUHammer, was the first to demonstrate practical RowHammer exploitation on discrete NVIDIA GPUs running GDDR6 memory. By reverse-engineering proprietary GDDR DRAM row mappings using timing side-channels, the researchers were able to induce up to 1,171 bit-flips on an NVIDIA RTX 3060 in controlled conditions [1]. A companion experiment on an NVIDIA RTX A6000 showed that a single bit-flip targeting the weights of a deployed neural network was sufficient to reduce model inference accuracy from approximately 80% to 0.1% [1], a result that underscores the particular danger of GPU RowHammer in AI production environments where model correctness is a mission-critical property.

The three papers presented at IEEE S&P 2026 extend this foundation into full privilege-escalation exploits. GDDRHammer, developed at UNC Chapel Hill, Georgia Tech, and MBZUAI, and GeForge, developed at Purdue University with collaborators at the University of Rochester, the University of Western Australia, and Clemson University, both demonstrate that induced GPU memory bit-flips can corrupt page table entries in a way that ultimately yields read/write access to arbitrary system memory and, from there, a root shell on the host [2][3]. Both attacks are meaningful escalations over GPUHammer, but both can be blocked by enabling IOMMU on the host system [2][3]—an important distinction.

GPUBreach, by the same University of Toronto team responsible for GPUHammer, is the most severe of the three disclosures because it does not share this IOMMU limitation [4][12][13]. GPUBreach completes the same privilege-escalation chain—GPU page table corruption via RowHammer, followed by exploitation of a memory-safety flaw in the NVIDIA kernel driver—but does so in a way that works even when the IOMMU is configured and active. The IOMMU has been the standard defensive recommendation for protecting the CPU memory space from DMA-capable peripheral devices, including GPUs. GPUBreach's IOMMU bypass eliminates the primary compensating control previously recommended for GPU DMA attacks, leaving ECC enablement and driver-level patching as the remaining mitigations for all environments running GDDR6 GPUs, including cloud-hosted GPU nodes.

Affected Hardware and the Ecosystem Breadth

Vulnerable hardware confirmed by the research includes NVIDIA Ampere-generation GPUs using GDDR6 memory, specifically the RTX 3060 and RTX A6000; the researchers note that susceptibility likely extends to GDDR6-equipped GPUs in the RTX 20-series and later based on their shared memory architecture [1][2][3][4]. The research teams collectively tested 25 GDDR6 GPU models and found that most exhibited susceptibility to bit-flip induction under experimental hammering conditions [2]. GDDR6X memory—present in select higher-tier NVIDIA consumer cards—showed greater resistance under the same attack methods, and GDDR7 memory, deployed in the GeForce RTX 50 series, includes on-die ECC that substantially raises the bar for RowHammer exploitation [2].

The enterprise and data center picture is different. NVIDIA's Hopper and Blackwell architecture GPU products (H100, H200, B100, B200) enable system-level ECC by default, a configuration that mitigates RowHammer's ability to induce stable, usable bit-flips [5]. However, this protection does not automatically extend to workloads deployed on older-generation A100 or consumer-grade GPUs in cloud environments, where ECC may be disabled for throughput reasons. Cloud providers that offer access to NVIDIA A-series or consumer-class GPUs without enforcing ECC should be considered within the vulnerable scope.

Responsible disclosure of GPUBreach was completed on November 11, 2025, with notifications to NVIDIA, Google, Amazon Web Services, and Microsoft [4]. Google acknowledged the submission with a bug bounty payment of $600 in January 2026 [4]. The relatively modest bounty likely reflects Google's assessment that the attack requires specific conditions—shared physical GDDR6 memory between attacker and target, sustained hammering durations, and co-tenancy arrangements that may not be present in Google's production GPU configurations—rather than a downgrade of the underlying vulnerability's severity. As of this writing, NVIDIA has confirmed it is investigating a fix but has not

released a patch. NVIDIA issued a formal security notice on RowHammer in July 2025 acknowledging susceptibility in GDDR6 memory and identifying ECC enablement as the primary mitigation, at the cost of approximately 6% of available VRAM capacity [5].

AI Infrastructure as a High-Value Target

The intersection of RowHammer with GPU-accelerated AI workloads introduces threat surfaces that have not previously been considered in enterprise AI security frameworks. In a shared-cloud GPU environment, a malicious tenant running RowHammer against co-located GDDR6 memory could silently corrupt the model weights of an inference workload owned by another tenant—provided the attacker shares the same physical GPU and can sustain the hammering pattern without triggering detection. Because model corruption via bit-flip produces outputs that appear structurally correct—the model continues to run and return responses—detection requires active monitoring of model output distribution or cryptographic attestation of loaded weights. To our knowledge, neither control is deployed as standard practice in most production inference environments, though comprehensive data on deployment prevalence is not yet available.

This raises a plausible threat scenario for agentic AI deployments: a system whose model weights have been silently corrupted may produce subtly incorrect outputs that accumulate across many decision cycles before becoming detectable by human reviewers, particularly if the corruption only manifests under specific input conditions. The corruption produces no log entry in standard runtime environments, and the behavioral drift may only be apparent when specific triggers arise. In environments where agentic AI systems are taking consequential actions—scheduling, procurement, code review, security operations—this class of silent corruption should be considered a high-severity risk warranting explicit treatment in AI system threat models.

The driver-level CVEs associated with this research cluster compound the hardware vulnerability. CVE-2025-33220, a use-after-free vulnerability in NVIDIA's vGPU Virtual GPU Manager, allows guest-to-host escape in multi-tenant GPU virtualization environments with a CVSS score of 7.8 [6]. CVE-2025-33218, an integer overflow in the NVIDIA GPU Display Driver for Windows (nvlddmkm.sys), enables arbitrary kernel code execution [7]. Taken together with the RowHammer-based physical attack chain, these vulnerabilities illustrate a layered attack surface that spans DRAM physics, kernel driver logic, and hypervisor boundaries.

Mitigation Landscape and Gaps

No vendor patch for GPUBreach has been released as of this writing, leaving organizations dependent on compensating controls that each carry meaningful limitations. ECC enablement at the system level is the most reliable available control for preventing usable RowHammer bit-flips, but it requires administrative action on each affected system, reduces available VRAM by approximately 6%, and may not be configurable on consumer-class GPU cards without firmware support [5]. IOMMU alone is insufficient given GPUBreach's demonstrated bypass. Memory allocation hardening at the driver level—analogous to the kernel page-table isolation introduced in response to Spectre and Meltdown—is theoretically viable but would require significant NVIDIA driver changes that are not yet released.

At the infrastructure level, organizations running multi-tenant GPU workloads should treat GPU co-location as a boundary requiring the same isolation assumptions that apply to CPU-side memory. The assumption that IOMMU provides adequate isolation has been falsified by GPUBreach, and any security architecture that relied on this assumption should be revised.

Recommendations

Immediate Actions

Organizations should begin with a comprehensive audit of all GPU deployments to identify GDDR6-equipped hardware running without ECC enabled. This audit should prioritize systems hosting AI inference workloads, multi-tenant GPU virtualization environments, and any systems where privileged access from GPU processes could reach sensitive host memory.

Where the firmware supports it, system-level ECC should be enabled on NVIDIA professional and workstation GPUs. For cloud-hosted GPU instances, organizations should verify that ECC is enforced at the provider level and confirm this directly with cloud provider documentation or support channels rather than assuming a default-enabled configuration.

Available NVIDIA driver patches should be applied promptly, and NVIDIA's security bulletin archive should be monitored for GPUBreach-specific mitigations as they become available. As of this writing, patches for the hardware-level attack remain under development, making driver-level updates for the companion CVEs (CVE-2025-33220, CVE-2025-33218) the most actionable near-term software remediation.

Short-Term Mitigations

Organizations with near-term exposure should evaluate whether co-tenancy restrictions are feasible for GPU workloads hosting sensitive models or operating in regulated environments. Migrating AI inference workloads to NVIDIA Hopper or Blackwell architecture hardware—where ECC is enabled by default and HBM3/HBM3e memory provides substantially stronger RowHammer resistance than GDDR6—reduces exposure while patches are developed.

Runtime model-integrity verification is a compensating control worth piloting for high-value inference deployments. Loading model weights with a cryptographic hash at startup and verifying periodically against a known-good baseline would detect silent corruption introduced by a RowHammer attack, though it introduces operational overhead and does not prevent the corruption event itself. Organizations with AI workloads in scope for SOC 2, ISO 42001, or NIST AI RMF compliance should document this gap and its associated residual risk in their risk registries pending patch availability.

At the network level, apply available NVIDIA security updates that address the driver-level CVEs (CVE-2025-33220, CVE-2025-33218). These patches reduce the exploitability of the software components in the privilege-escalation chain even before hardware-level mitigations are available.

Strategic Considerations

Hardware procurement strategy should incorporate GPU memory architecture as a security criterion alongside traditional performance metrics. The research record now clearly differentiates GDDR6 (vulnerable, no consumer ECC) from GDDR7 and HBM3e (significantly more resistant) as a security property, not merely a performance property. For organizations building out AI infrastructure for multi-year production use, hardware that ships with on-die ECC as a baseline should be treated as a security requirement rather than an optional feature.

Longer-term, the GPU security ecosystem needs the equivalent of the defensive investment that followed the Spectre and Meltdown disclosures on the CPU side: vendor-published threat models for GPU memory isolation, driver-level memory safety hardening comparable to kernel page-table isolation, and standardized assurance requirements for cloud providers offering shared GPU resources. CSA's STAR for AI program provides a natural venue for developing auditable assurance criteria around GPU memory security—including GDDR6 ECC status and multi-tenant GPU isolation policies—as attestation requirements for cloud AI service providers.

CSA Resource Alignment

The vulnerabilities documented in this note map directly to several domains addressed by CSA's AI and cloud security frameworks, and organizations should use these frameworks to structure their response.

CSA's AI Controls Matrix (AICM) provides the foundational control inventory for AI system security, encompassing the infrastructure, model, and operational layers that GPUBreach affects [8]. The AICM's controls on data integrity, model provenance, and infrastructure isolation are directly relevant to the silent weight-corruption risk identified in this analysis. Organizations conducting AICM-aligned assessments should add GPU memory integrity to their infrastructure security control evaluations.

The MAESTRO framework for agentic AI threat modeling provides a structured approach to reasoning about attack paths through AI agent systems [9]. The scenario in which GPU RowHammer corrupts the underlying model of an autonomous agent—causing behavioral drift without a detectable failure signal—maps to MAESTRO's coverage of integrity attacks on the AI model layer and is a scenario that MAESTRO-based threat models should explicitly include going forward.

The Cloud Controls Matrix (CCM) addresses infrastructure security controls for cloud-hosted workloads, including compute isolation and hypervisor boundary integrity [10]. The vGPU guest-to-host escape demonstrated by CVE-2025-33220 and the IOMMU bypass in GPUBreach both implicate CCM control domains around virtual machine and hypervisor security. Cloud buyers procuring GPU compute services should include GPU memory isolation assurance in their CCM-based vendor assessments.

CSA's Zero Trust guidance emphasizes the elimination of implicit trust in infrastructure components, including hardware. The assumption that co-tenant GPU workloads are isolated from one another should not be treated as an inherent property of GDDR6 hardware absent verified ECC enforcement and confirmed patch status. Organizations applying Zero Trust principles to their AI infrastructure should extend those principles explicitly to GPU memory isolation.

The STAR for AI program provides a registry-based mechanism for AI service providers to attest to their security posture [11]. As GPU memory security emerges as an auditable risk domain, STAR for AI assessments should include attestation requirements around GDDR6 ECC status, GPU driver patch currency, and multi-tenant GPU isolation policies. CSA working groups developing STAR for AI criteria should consider GPUBreach and its companion disclosures as a reference case for hardware-layer AI infrastructure assurance.

References

[1] C.S. Lin, J. Qu, G. Saileshwar. "GPUHammer: Rowhammer Attacks on GPU Memories are Practical." USENIX Security Symposium, August 2025. [arXiv preprint]

[2] Barrack AI. "GDDRHammer and GeForge: GPU Rowhammer Now Achieves Full System Compromise." Barrack AI Research, April 2026.

[3] Tom's Hardware. "New 'GeForge' and 'GDDRHammer' Attacks Can Fully Infiltrate Your System Through NVIDIA's GPU Memory." Tom's Hardware, April 2026.

[4] GPUBreach Research Team. "GPUBreach: Hardware-Level Privilege Escalation via GDDR6 RowHammer." University of Toronto, April 2026.

[5] NVIDIA. "Security Notice: Rowhammer – July 2025." NVIDIA Customer Support, July 2025.

[6] SentinelOne. "CVE-2025-33220: NVIDIA vGPU Use-After-Free Vulnerability." SentinelOne Vulnerability Database, 2026.

[7] SentinelOne. "CVE-2025-33218: NVIDIA GPU Display Driver Integer Overflow." SentinelOne Vulnerability Database, 2026.

[8] Cloud Security Alliance. "AI Controls Matrix." CSA Working Group, 2025.

[9] Cloud Security Alliance. "AI Safety Initiative: MAESTRO Framework." CSA, 2025.

[10] Cloud Security Alliance. "Cloud Controls Matrix." CSA, 2025.

[11] Cloud Security Alliance. "STAR for AI Registry." CSA, 2025.

[12] The Hacker News. "New GPUBreach Attack Enables Full CPU Privilege Escalation via GDDR6 Bit-Flips." The Hacker News, April 2026.

[13] SecurityWeek. "GPUBreach: Root Shell Access Achieved via GPU Rowhammer Attack." SecurityWeek, April 2026.

[14] CSO Online. "Alert: Nvidia GPUs Are Vulnerable to Rowhammer Attacks." CSO Online, 2025.

[15] Y. Kim, R. Daly, J. Kim, C. Fallin, J. H. Lee, D. Lee, C. Wilkerson, K. Lai, O. Mutlu. "Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors." ACM SIGARCH Computer Architecture News / ISCA 2014.

