CSAI Foundation

Cloud Security Alliance AI Safety Initiative

GPUBreach: GDDR6 RowHammer Achieves Full CPU Privilege Escalation

New Research Demonstrates Root Shell via NVIDIA Driver Exploit Chain, With No Patch for Consumer GPUs

Unofficial AI-assisted Research

2026-04-07

Key Takeaways

- GPUBreach, disclosed April 6–7, 2026 by researchers at the University of Toronto, demonstrates for the first time that RowHammer bit-flips in GDDR6 GPU memory can be chained into full CPU-level privilege escalation — producing a root shell — by exploiting memory-safety bugs in the NVIDIA kernel driver, even when an IOMMU is enabled [1][2].
- The attack is part of a coordinated April 2026 research disclosure spanning three independent teams: GPUBreach (University of Toronto), GDDRHammer (UNC Chapel Hill / Georgia Tech / MBZUAI), and GeForge (Purdue / Clemson / University of Rochester / University of Western Australia). All three papers have been accepted to IEEE S&P 2026 [1][3][4].
- NVIDIA has not issued a new security bulletin or assigned CVEs to the driver vulnerabilities exploited in GPUBreach. NVIDIA's existing guidance redirects to a July 2025 Rowhammer security notice that characterizes DRAM bit-flip attacks as an industry-wide hardware issue rather than a driver-specific vulnerability [5].
- Consumer NVIDIA GPUs currently have no vendor-issued patch and limited hardware-level protection: ECC memory — the most commonly cited mitigation — is largely absent from consumer graphics cards, and prior research has shown it bypassable under multi-bit flip scenarios [6].
- Organizations running shared GPU infrastructure for AI and machine learning workloads face elevated exposure, as multi-tenant deployments create conditions — unprivileged CUDA process co-residence — that GPUBreach requires to succeed [1].

Background

The RowHammer Paradigm and GPU Memory

RowHammer is a class of hardware attack that exploits a physical property of dynamic random-access memory (DRAM): repeatedly accessing ("hammering") a memory row causes electrical interference in adjacent rows sufficient to induce unintended bit-flips. First publicly demonstrated against DRAM by Kim et al. at Carnegie Mellon University in 2014, RowHammer attacks have since evolved into practical

privilege escalation, sandbox escape, and cryptographic key recovery techniques against CPU DRAM [7]. The attack works because DRAM cells are physically close together, and the capacitive coupling between rows increases as manufacturing processes shrink cell dimensions [7].

For most of that decade, GPU memory was assumed to be a separate problem domain. Graphics cards use GDDR-series memory rather than DDR, and the differences in access patterns, refresh rates, and physical organization led many researchers to treat GPU DRAM as a different attack surface. That assumption was challenged by GPUHammer, published at USENIX Security 2025 by the same University of Toronto team, which demonstrated that GDDR6 memory in NVIDIA Ampere-generation GPUs is vulnerable to RowHammer bit-flips — including 1,171 bit-flips on an RTX 3060 and 202 on an RTX A6000 [8]. GPUHammer established the primitive — reliable bit-flip induction on GDDR6 — but did not achieve privilege escalation or cross-boundary exploitation.

GPUBreach is the direct successor to GPUHammer and answers the question GPUHammer left open: once an attacker can reliably induce bit-flips in GDDR6, what can they do with them?

The April 2026 Disclosure Window

GPUBreach was not disclosed in isolation. Three independent research groups, apparently working in parallel without prior coordination, submitted GPU RowHammer exploitation papers to IEEE S&P 2026, and all three were accepted [2]. The coordinated public disclosure — April 6–7, 2026 — marks the first instance of three independent GPU RowHammer exploitation papers being accepted simultaneously to a top security venue, and together they substantially advance an open question in the field: whether GPU DRAM bit-flips could be chained into host-level system compromise [1][3][4][9].

GDDRHammer, from a collaboration between UNC Chapel Hill, Georgia Tech, and Mohamed bin Zayed University of Artificial Intelligence, exploits NVIDIA's memory allocator co-locating page tables and user data in GDDR6. By inducing bit-flips in aperture bits within page table entries, GDDRHammer can redirect GPU virtual addresses to CPU physical memory via PCIe BAR1, achieving arbitrary DMA reads and writes and ultimately a root shell [3]. GeForge, from a team at Purdue, Clemson, the University of Rochester, and the University of Western Australia, takes a different angle: it targets the page directory (PD0) rather than last-level page tables, forging entirely new page table mappings using non-uniform RowHammer patterns and timing side-channels, but requires the IOMMU to be disabled to achieve full exploitation [4][11].

GPUBreach's distinguishing characteristic is that it achieves full CPU privilege escalation with the IOMMU enabled — the configuration recommended for hardened enterprise Linux and Windows systems and typical of major cloud provider virtualization stacks.

Security Analysis

How GPUBreach Works

GPUBreach constructs a four-stage exploit chain. In the first stage, the attack targets the memory layout of the NVIDIA driver's GPU page tables. The University of Toronto researchers reverse-engineered NVIDIA driver internals to identify where contiguous 2 MB page-table regions are allocated in GDDR6 memory, then used allocation primitives within CUDA's Unified Virtual Memory (UVM) subsystem to position 64 KB and 4 KB memory frames in RowHammer-accessible positions. A timing side-channel monitors eviction events to detect when new page-table regions materialize in vulnerable DRAM rows [1].

In the second stage, the attacker hammers adjacent DRAM rows to induce bit-flips in page table entries. A corrupted PTE, pointing to an attacker-controlled physical address rather than its legitimate target, grants the unprivileged CUDA process arbitrary GPU memory read/write access. This provides the attacker with a powerful intra-GPU primitive but does not yet cross the hardware boundary [1].

The third and fourth stages are where GPUBreach diverges from its concurrent counterparts. Rather than attempting to issue unauthorized DMA through the IOMMU — the strategy that GDDRHammer relies on and GeForge requires disabling the IOMMU to bypass — GPUBreach uses its arbitrary GPU memory access to corrupt trusted driver-state buffers that reside within memory regions already authorized for IOMMU-permitted transfers. These corrupted buffers trigger memory-safety bugs in the NVIDIA kernel driver itself, which runs in ring-0 (kernel space). Exploiting these bugs produces arbitrary kernel writes, which the researchers chain into a root shell on the host CPU [1].

The attack is significant precisely because it navigates around the IOMMU. Modern OS deployments treat IOMMU enforcement as a meaningful boundary between device and host memory. GPUBreach demonstrates that this boundary provides less isolation than commonly assumed when the driver handling the device contains exploitable memory-safety bugs.

Affected Hardware and Resistance Profile

The primary test platform for GPUBreach is the NVIDIA RTX A6000 (Ampere architecture, GDDR6), on which the full exploit chain was demonstrated. The researchers assessed the broader GDDR6 GPU population and observed that resistance to the underlying RowHammer bit-flip primitive varies considerably. Testing found zero bit-flips on the RTX 3080 (GDDR6X), the RTX 4060 and 4060 Ti (Ada Lovelace, GDDR6), and the RTX 6000 Ada (Ada Lovelace, GDDR6), as well as the RTX 5050 (Blackwell,

GDDR7). The researchers also assessed that NVIDIA's data center AI accelerators using High Bandwidth Memory — the A100, H100, and H200 — are likely resistant because HBM includes on-die ECC enabled by default [1][2].

These results suggest that the Ampere generation of NVIDIA GDDR6 GPUs carries the highest risk profile, while Ada Lovelace and newer architectures may have mitigated the bit-flip primitive at the memory hardware level. However, the researchers caution that these assessments reflect testing on specific cards and should not be interpreted as architectural guarantees across the full GDDR6 product line.

From an enterprise AI infrastructure standpoint, the affected hardware class — Ampere-generation NVIDIA GPUs — has been deployed across cloud provider GPU instances and on-premises AI clusters beginning with its 2020 commercial introduction. The RTX A6000, specifically, has been a common choice for workstation-class AI inference deployments.

Vendor Response and Disclosure Timeline

The research team disclosed GPUBreach to NVIDIA, Google, AWS, and Microsoft on November 11, 2025 — approximately five months before public disclosure. NVIDIA did not issue a new security bulletin or assign CVEs to the driver memory-safety bugs exploited in the attack chain. Instead, NVIDIA directed the research team and the public to a July 2025 security notice issued in response to GPUHammer, which classifies DRAM bit-flip attacks as an industry-wide hardware issue and recommends enabling System-Level ECC on supported GPUs [5]. Google acknowledged the disclosure and awarded a $600 bug bounty. No acknowledgments or bounties from NVIDIA, AWS, or Microsoft were publicly reported as of April 7, 2026 [2].

The full paper PDF and GitHub exploit artifact are expected to be released on approximately April 13, 2026, ahead of the IEEE S&P 2026 presentation scheduled for May 18–20, 2026 in San Francisco [1]. Because full technical disclosure — including precise identification of the NVIDIA driver memory-safety bugs — has not yet occurred at the time of this writing, the complete scope of affected driver versions and the feasibility of targeted patches remains unclear.

Mitigation Landscape and Its Limits

ECC memory is the most widely cited mitigation for RowHammer attacks. By detecting and correcting single-bit errors, ECC can interrupt the exploit chain before a corrupted PTE becomes exploitable. However, ECC has three significant limitations in this context. First, it is absent from most consumer NVIDIA GPUs; the RTX A6000 is a workstation product that supports ECC, but mass-market consumer gaming cards in the GeForce lineup generally do not. Second, System-Level ECC must be explicitly

enabled on workstation GPUs — it is not on by default in all configurations. Third, prior research including ECCploit and ECC.fail has demonstrated that multi-bit flip attacks can bypass ECC correction in some configurations, suggesting that ECC alone should not be treated as a complete defense against determined adversaries [6][12].

The IOMMU, widely deployed in enterprise Linux and Windows systems, provides no protection against GPUBreach specifically, because the attack chains through the driver rather than attempting unauthorized DMA. Disabling Unified Virtual Memory (CUDA UVM) impedes the memory massage primitives required for the attack but is operationally disruptive for AI and machine learning workloads that depend on UVM for memory management. As of April 7, 2026, no software patch targeting the underlying NVIDIA driver memory-safety bugs has been announced.

Recommendations

Immediate Actions

Organizations running NVIDIA Ampere-generation GPUs in multi-tenant or shared compute environments should assess their exposure promptly. The core risk condition — an unprivileged CUDA process co-resident with a victim process on a shared GPU — is precisely the configuration used in GPU-as-a-service cloud instances and many AI training cluster deployments. Security teams should verify whether workloads running on affected hardware classes are adequately isolated, either through dedicated GPU assignment or hypervisor-level GPU partitioning that prevents CUDA process co-residence. Security operations centers should be prepared for disclosure of full technical artifacts on approximately April 13, 2026, and should monitor the National Vulnerability Database and NVIDIA's Product Security page for CVE assignments to the driver vulnerabilities involved [13].

For GPU instances in public cloud environments — AWS, Google Cloud, Microsoft Azure — verify with the provider whether their NVIDIA driver versions on Ampere-based instances have been patched or whether additional tenant isolation controls have been applied. Cloud providers have had advance notice since November 2025, and some may have implemented driver-level mitigations or infrastructure controls not yet publicly documented.

Short-Term Mitigations

Where operationally feasible, enabling System-Level ECC on NVIDIA workstation GPUs that support it reduces the probability of successful bit-flip induction, though it does not eliminate risk under multi-bit scenarios. Organizations deploying NVIDIA RTX A6000 or similar Ampere workstation GPUs should confirm ECC status through `nvidia-smi --query-gpu=ecc.mode.current --format=csv` and enable it if not already active, accepting the associated VRAM reduction (ECC reduces usable memory by approximately 6% per NVIDIA's product specifications for Ampere workstation GPUs). Organizations should also evaluate whether CUDA Unified Virtual Memory is required for their workloads; disabling UVM via driver configuration or workload migration eliminates the memory massage primitives GPUBreach requires, at the cost of compatibility with some ML frameworks.

Firmware and driver update cadence should be reviewed. NVIDIA publishes security bulletins on a quarterly cadence, along with driver updates; organizations should subscribe to NVIDIA's security notification service and monitor the NVIDIA Product Security page for updates relevant to affected GPU models [13]. Ensuring that GPU driver versions remain current minimizes exposure to known driver-level vulnerabilities and positions affected systems to receive patches promptly when NVIDIA addresses the memory-safety bugs exploited in this attack chain.

Strategic Considerations

GPUBreach and the concurrent GDDRHammer and GeForge disclosures collectively signal that GPU memory should be treated as an attack surface rather than assumed-safe infrastructure. The security industry has spent a decade hardening CPU DRAM against RowHammer, developing mitigations such as Target Row Refresh (TRR), LPDDR5 Refresh Management (RFM), and probabilistic adjacent-row activation (pTRR). GPU DRAM has not undergone equivalent hardening, and the three concurrent April 2026 papers demonstrate that the primitive — reliable bit-flip induction in GDDR6 — is achievable on production hardware today.

For organizations building AI infrastructure strategy, this has architectural implications. GPU-accelerated AI workloads increasingly run in shared, multi-tenant environments — cloud GPU instances, Kubernetes GPU node pools, and shared HPC clusters. In many such deployments, process isolation relies primarily on software-level controls. GPUBreach challenges this posture, demonstrating that hardware-level memory interference can undermine software isolation boundaries when the kernel driver handling the GPU contains exploitable bugs. Organizations planning multi-tenant AI infrastructure deployments should factor hardware isolation — dedicated GPU assignment per tenant, or GPU models with architectural RowHammer resistance — into their threat model alongside traditional software controls.

The long-term trajectory of GPU security research suggests that RowHammer-style attacks against GDDR and future GPU memory generations will continue to mature, following the decade-long progression that CPU DRAM attacks have already demonstrated.

CSA Resource Alignment

GPUBreach intersects with several CSA frameworks and research programs in ways that practitioners should connect explicitly.

The CSA AI Controls Matrix (AICM) addresses infrastructure security controls for AI systems, including requirements for workload isolation, patch management, and vulnerability response under the Technology Management and Infrastructure Security domain. GPU hardware vulnerabilities that bypass software isolation controls — as GPUBreach does — represent an attack vector not fully addressed by current AI controls frameworks, which generally assume hardware integrity. CSA's ongoing work to extend AICM for agentic and infrastructure-layer AI security should treat GPU memory attack surfaces as a distinct control objective category, separate from the software-level prompt injection and model integrity threats that current AICM coverage emphasizes.

The MAESTRO threat modeling framework for agentic AI is directly relevant to the deployment context where GPUBreach poses the greatest risk. AI agents executing on shared GPU infrastructure — a configuration increasingly common in cloud-native agentic deployments — are potentially subject to co-resident attacks by other GPU users. MAESTRO Layer 6 (Infrastructure & Compute) addresses the physical and virtualized compute resources underlying agentic systems; GPUBreach exemplifies the category of hardware-layer attack that this layer must account for. Threat models for agentic AI systems deployed on shared GPU infrastructure should now include GPU memory interference as a lateral movement vector alongside traditional hypervisor escape and container breakout scenarios.

The Cloud Controls Matrix (CCM) provides mappings to relevant control domains. IVS-01 and IVS-03 address infrastructure and virtualization security, including requirements for isolation between tenants in shared compute environments. TVM-06 covers patch and vulnerability management for cloud infrastructure components, which applies directly to NVIDIA driver patch cadence. Organizations using the STAR (Security Trust Assurance and Risk) self-assessment program to evaluate their AI infrastructure posture should explicitly address GPU driver update policies and hardware isolation architecture in their TVM and IVS control responses.

CSA's Zero Trust guidance is applicable to the architectural response: rather than relying on perimeter-based assumptions about GPU memory integrity, organizations should adopt a verify-continuously posture that includes monitoring for anomalous GPU memory access patterns, driver-level integrity attestation, and hardware-level isolation as a defense-in-depth layer rather than a secondary control.

References

[1] C. S. Lin, Y. Yan, G. Ding, J. Qu, J. Zhu, D. Lie, and G. Saileshwar (University of Toronto). "GPUBreach: GPU Rowhammer Achieves Root Shell." IEEE Symposium on Security and Privacy 2026 (accepted), April 2026. *(Note: PDF expected to become available approximately April 13, 2026.)*

[2] B. Toulas. "New GPUBreach Attack Enables System Takeover via GPU Rowhammer." BleepingComputer, April 6, 2026.

[3] Y. Hu, N. Brown, Y. Chen, J. Bakita, T. Chen, D. Genkin, and A. Kwong (UNC Chapel Hill, Georgia Tech, MBZUAI). "GDDRHammer: Cross-Component Rowhammer Attacks from Modern GPUs." IEEE Symposium on Security and Privacy 2026 (accepted), April 2026.

[4] J. Wan, Y. Guo, Z. Zhang, Z. Li, D. Tian, and Z. Zhang (Purdue, Clemson, University of Rochester, University of Western Australia, HydroX AI). "GeForge: Hammering GDDR Memory to Forge GPU Page Tables." IEEE Symposium on Security and Privacy 2026 (accepted), April 2026.

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

[6] L. Cojocar, K. Razavi, C. Giuffrida, and H. Bos. "Exploiting Correcting Codes: On the Effectiveness of ECC Memory Against Rowhammer Attacks." IEEE Symposium on Security and Privacy, May 2019.

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

[8] G. Saileshwar, C. S. Lin, Y. Yan, G. Ding, J. Qu, J. Zhu, and D. Lie (University of Toronto). "GPUHammer: Rowhammer Attacks on NVIDIA GDDR6 GPUs." USENIX Security Symposium 2025, July 2025.

[9] The Hacker News. "New GPUBreach Attack Enables Full CPU Privilege Escalation via GDDR6 Bit-Flips." The Hacker News, April 7, 2026.

[10] CyberInsider. "New GPUBreach Attack Achieves Root Access via Memory Corruption." CyberInsider, April 2026.

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

[12] N. Kamadan, W. Wang, S. van Schaik, C. Garman, D. Genkin, and Y. Yarom (Georgia Tech, University of Michigan, Purdue, Ruhr University Bochum). "ECC.fail: Mounting Rowhammer Attacks on DDR4 Servers with ECC Memory." USENIX Security Symposium 2025, July 2025.

[13] NVIDIA. "Product Security." NVIDIA, accessed April 2026.