INTELLIGENCE BRIEFING
Security Command Center
TLP:CLEAR
2026-04-07 06:06 UTC

GPUBreach: GDDR6 Rowhammer Attack Chains GPU Memory Corruption to Full System Compromise

SECURITY ANALYSIS | CRITICAL | CVSS 9.5

SCC Item ID	SCC-STY-2026-0049
Type	Security Analysis
Severity	CRITICAL
CVSS Base Score	9.5
Affected Products	NVIDIA RTX A6000 (GDDR6) and other NVIDIA GPUs without ECC memory; NVIDIA GPU drivers (all platforms); cloud environments including Google Cloud, AWS, and Microsoft Azure
Published	2026-04-06T17:44:15
Discovery Source	Rss

Executive Summary

Researchers at the University of Toronto have demonstrated GPUBreach, a hardware-level attack that exploits Rowhammer-style bit-flips in GDDR6 GPU memory to corrupt page tables, ultimately enabling full system compromise including root-level privilege escalation on the host. The attack bypasses IOMMU protections and affects consumer and prosumer NVIDIA GPUs without ECC memory, a category that includes hardware widely deployed in AI/ML workloads and GPU-accelerated cloud instances across Google Cloud, AWS, and Microsoft Azure. Researchers have indicated that full technical details and reproduction scripts are scheduled for public release on April 13, 2026, creating a narrow window for organizations to assess exposure before weaponization becomes accessible to a broader threat population.

Technical Analysis

GPUBreach chains two distinct exploitation primitives into a single end-to-end compromise path. The first stage induces Rowhammer-style bit-flips in GDDR6 memory, a technique previously demonstrated against DRAM but now extended to GPU memory, to corrupt GPU page table entries. Successful page table corruption yields arbitrary GPU memory read/write access, a powerful primitive that researchers then pivoted through NVIDIA driver vulnerabilities to escape the GPU context entirely. The second stage traverses the GPU driver interface to achieve privilege escalation on the host OS. This path explicitly sidesteps IOMMU protections: IOMMU typically defends against direct DMA-based escapes, but GPU privilege escalation through driver interface exploitation does not rely on direct memory access, making IOMMU an ineffective compensating control for this

vulnerability. The attack maps to CWE-119 (improper memory operations), CWE-787 (out-of-bounds write), CWE-125 (out-of-bounds read), and CWE-1232 (inadequate physical memory protection). MITRE ATT&CK coverage spans T1082 (System Information Discovery), T1611 (Escape to Host), T1212 (Exploitation for Credential Access), T1068 (Exploitation for Privilege Escalation), T1055 (Process Injection), and T1203 (Exploitation for Client Execution). The hardware mitigation gap is the most consequential finding: consumer and prosumer GPUs without ECC memory have no complete hardware-level fix available, meaning software patching alone cannot fully close the exposure. Cloud multi-tenancy amplifies the risk - GPU-accelerated instances sharing physical hardware across tenants are a plausible lateral movement surface if the attack is adapted to that context. As of this report, no CVE has been assigned and no NVIDIA patch has been confirmed. The underlying academic research has not yet been published in a peer-reviewed venue or independently replicated. Organizations should treat this as credible pre-publication research pending full technical disclosure and vendor confirmation.

Action Checklist

1. Step 1: Assess exposure - inventory all NVIDIA GPU deployments, specifically identifying RTX A6000 and other GDDR6-based GPUs operating without ECC memory; include GPU-accelerated cloud instances on Google Cloud, AWS, and Azure in scope
2. Step 2: Review controls - verify GPU driver versions across all endpoints and servers; check NVIDIA's latest security advisories for available mitigations; prioritize systems where GPUs are accessible to unprivileged users or shared workloads
3. Step 3: Update threat model - add GPU memory corruption via Rowhammer as an attack vector in your hardware threat register; flag AI/ML pipeline infrastructure and GPU-accelerated cloud instances as elevated-risk assets pending patch availability; note that IOMMU is not a compensating control for this attack path
4. Step 4: Communicate findings - brief leadership on exposure in AI/ML and cloud GPU infrastructure with emphasis on the hardware mitigation gap for non-ECC GPUs; frame as a pre-disclosure window requiring proactive inventory, not a wait-for-patch situation
5. Step 5: Monitor developments - track NVIDIA security advisories at https://nvidia.custhelp.com for driver advisories; subscribe to Google Cloud, AWS, and Microsoft Azure security bulletins; search CVE databases for GPUBreach or GPU Rowhammer references; monitor academic preprint repositories (arXiv) for independent replication. Watch for the April 13, 2026 scheduled technical disclosure.

IR / Forensic Enrichment

Triage Priority	URGENT
Escalation Criteria	Escalate to immediate priority and activate IR plan upon any of the following: CVE assignment with CVSS >= 9.0 confirmed by NVD, public release of PoC exploit code for GDDR6 Rowhammer page-table corruption, CISA KEV listing, cloud provider (AWS, GCP, Azure) issuing a customer security advisory for GPU instance types, or detection of anomalous privilege escalation events on systems hosting non-ECC GDDR6 GPUs in shared or multi-tenant workload configurations.

Recovery Notes	Because GPUBreach operates at the hardware memory layer, software-only recovery (reimaging the OS) is insufficient if bit-flip corruption has reached host page tables — a full power cycle is required to clear GDDR6 DRAM state, followed by OS reinstallation from verified clean media and driver reinstallation from NVIDIA's post-advisory patched release. Post-recovery, enable ECC memory where hardware supports it ('nvidia-smi --ecc-config=1') and reboot to activate, then verify with 'nvidia-smi --query-gpu=ecc.mode.current --format=csv,noheader'. Monitor affected systems for 30 days post-recovery using kernel crash logs ('/var/log/kern.log' on Linux, System Event Log on Windows) and NVIDIA driver error logs for any recurrence of memory corruption indicators such as ECC single-bit or double-bit error counts via 'nvidia-smi --query-gpu=ecc.errors.corrected.volatile.total,ecc.errors.uncorrected.volatile.total --format=csv'.
Forensic Artifacts	NVIDIA driver kernel error logs: On Linux, '/var/log/kern.log' and 'dmesg' output filtered for 'nvidia' and 'Xid' error codes — specifically Xid 13 (Graphics Engine Exception), Xid 31 (GPU memory page fault), and Xid 79 (GPU memory ECC double-bit error) which are the driver-visible symptoms of GDDR6 memory corruption events consistent with GPUBreach bit-flip propagation | Host kernel page fault and crash artifacts: Linux kernel oops/panic logs in '/var/log/kern.log' or '/var/crash/' and Windows minidump files in 'C:\Windows\Minidump\' — GPUBreach page-table corruption triggering privilege escalation would produce kernel-mode access violations or unexpected privilege transitions visible in crash dump analysis using crash (Linux) or WinDbg (Windows) | GPU memory ECC error counters: Output of 'nvidia-smi --query-gpu=ecc.errors.corrected.volatile.total,ecc.errors.uncorrected.volatile.total,ecc.errors.corrected.aggregate.total --format=csv' timestamped at regular intervals — anomalous spikes in corrected or uncorrected ECC errors on systems where ECC is enabled are a primary hardware-level indicator of Rowhammer-style bit-flip activity against GDDR6 memory | Unprivileged process GPU memory access records: On Linux, '/proc/[pid]/maps' and '/proc/[pid]/smaps' for processes with open file descriptors to '/dev/nvidia*', combined with 'auditd' records for open/mmap syscalls against NVIDIA device nodes — these identify which user-space processes had GPU memory access during a suspect window and whether access was granted to non-root accounts | Cloud provider instance metadata and GPU passthrough configuration logs: For AWS, CloudTrail logs filtered for 'RunInstances' events with GPU instance types (p3, p4d, g4dn, g5) and any 'ModifyInstanceAttribute' events; for GCP, Cloud Audit Logs for Compute Engine GPU instance creation and GPU driver installation events — these establish whether GPU-accelerated instances were deployed with shared workload configurations that would increase GPUBreach exposure

Per-Action IR Details

Step 1: Assess exposure — inventory all NVIDIA GPU deployments, specifically identifying RTX A6000 and other GDDR6-based GPUs operating without ECC memory; include GPU-accelerated cloud instances on Google Cloud, AWS, and Azure in scope

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: Asset identification and risk-based prioritization of vulnerable systems before an incident occurs

Controls: NIST IR-4 (Incident Handling) — establishes the requirement to maintain preparation activities including asset scoping, NIST RA-3 (Risk Assessment) — requires identification of threats and vulnerabilities specific to organizational systems, CIS 1.1 (Establish and Maintain Detailed Enterprise Asset Inventory) — mandates accurate hardware inventory including GPU model and memory type, CIS 7.1 (Establish and Maintain a Vulnerability Management Process) — requires scoping vulnerable assets as the first step in vulnerability management

Compensating: Run 'nvidia-smi --query-gpu=name,memory.total,ecc.mode.current --format=csv,noheader' on all Linux/Windows hosts with NVIDIA drivers installed to enumerate GPU model and ECC status in one command; pipe output to a CSV for fleet-wide aggregation. For cloud instances, use AWS CLI 'aws ec2 describe-instances --filters Name=instance-type,Values=p3.*,p4.*,g4dn.*,g5.*' and equivalent GCP/Azure CLI commands to enumerate GPU-accelerated instance types. Cross-reference GPU model names against NVIDIA's GDDR6 product list — RTX A6000, RTX 3000/4000-series, and A-series datacenter GPUs without ECC are highest priority.

Evidence: Before inventorying, capture a baseline snapshot of current GPU driver versions and ECC status: run 'nvidia-smi -q | grep -E "Product Name|ECC Mode|Driver Version"' and preserve output with timestamp. On Windows, query 'Get-WmiObject Win32_VideoController | Select Name, DriverVersion, DriverDate' and export to CSV. This establishes a pre-remediation baseline and documents which systems were non-ECC at time of discovery — critical for any regulatory disclosure timeline.

Step 2: Review controls — verify GPU driver versions across all endpoints and servers; cross-reference against NVIDIA's January 2026 Security Bulletin (a_id/5747) for any driver-level mitigations already released; prioritize systems where GPUs are accessible to unprivileged users or shared workloads

NIST Phase: Detection Analysis

Reference: NIST 800-61r3 §3.2 — Detection and Analysis: Evaluate existing controls against the known attack vector and determine exposure scope prior to active exploitation

Controls: NIST SI-2 (Flaw Remediation) — requires identifying and reporting system flaws and testing updates for effectiveness before deployment, NIST SI-5 (Security Alerts, Advisories, and Directives) — mandates receiving and acting on vendor security advisories such as NVIDIA Security Bulletin a_id/5747, NIST CM-6 (Configuration Settings) — requires verifying configuration of GPU driver hardening options where vendor mitigations exist, CIS 7.2 (Establish and Maintain a Remediation Process) — requires risk-based remediation prioritization; shared/multi-tenant GPU workloads represent highest risk tier, CIS 2.2 (Ensure Authorized Software is Currently Supported) — GPU drivers running below the NVIDIA January 2026 bulletin baseline are out-of-supported-mitigation status

Compensating: Query installed NVIDIA driver versions across Windows hosts with: 'Get-ItemProperty HKLM:\SOFTWARE\Microsoft\Windows\CurrentVersion\Uninstall* | Where-Object {$_.DisplayName -like "*NVIDIA*"} | Select DisplayName, DisplayVersion'. On Linux: 'cat /proc/driver/nvidia/version' or 'modinfo nvidia | grep version'. Compare output against NVIDIA Security Bulletin a_id/5747 minimum mitigated version. Flag systems running drivers older than the bulletin's remediated release. For shared-workload prioritization, check whether the GPU is accessible to non-root users on Linux via 'ls -la /dev/nvidia*' — world-readable device nodes indicate unprivileged access exposure.

Evidence: Capture the full driver version string and device permissions before any updates: on Linux, record 'dmesg | grep -i nvidia' output (kernel module load events), 'lsmod | grep nvidia', and '/dev/nvidia* permissions'. On Windows, export the NVIDIA entry from 'HKLM\SYSTEM\CurrentControlSet\Services\nvlddmkm' registry key including ImagePath and DisplayVersion. These artifacts establish the driver state at time of assessment and are required if a later incident investigation needs to determine whether the system was running a vulnerable driver during a suspect time window.

Step 3: Update threat model — add GPU memory corruption via Rowhammer as an attack vector in your hardware threat register; flag AI/ML pipeline infrastructure and GPU-accelerated cloud instances as elevated-risk assets pending patch availability; note that IOMMU is not a compensating control for this attack path

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: Maintaining and updating threat models and risk registers to reflect newly discovered attack surfaces is a core preparatory activity

Controls: NIST RA-3 (Risk Assessment) — requires updating risk assessments when new threat information emerges, specifically that IOMMU bypass is a confirmed characteristic of GPUBreach, NIST PM-16 (Threat Awareness Program) — requires incorporating threat intelligence — here, the University of Toronto GPUBreach research — into organizational threat awareness, NIST IR-4 (Incident Handling) — requires that incident handling capability account for the full scope of threat vectors, including hardware-layer attacks bypassing software controls, CIS 7.1 (Establish and Maintain a Vulnerability Management Process) — threat register updates for novel hardware attack classes must feed

back into the vulnerability management process

Compensating: Document the IOMMU bypass characteristic explicitly in your hardware threat register with a note that VFIO/IOMMU isolation — commonly used in cloud GPU passthrough and AI/ML container environments — does not mitigate GPUBreach bit-flip propagation to host page tables. If using a threat modeling tool like OWASP Threat Dragon or a simple risk register spreadsheet, add a new threat entry: Asset = 'GDDR6 GPU (non-ECC)', Attack Vector = 'Physical/Local — Rowhammer bit-flip via unprivileged GPU memory access', Existing Control = 'IOMMU', Control Effectiveness = 'INEFFECTIVE per GPUBreach research', Residual Risk = 'CRITICAL'. Flag all Kubernetes GPU node pools, ML training clusters, and cloud GPU instances as elevated-risk in your CMDB or asset tracker.

Evidence: Before updating the threat model, collect and archive the University of Toronto GPUBreach research paper (pre-print or April 2026 full release when available) and NVIDIA's advisory response as primary source evidence supporting the threat register entry. Document the specific claim that IOMMU protections are bypassed — this is the critical finding that invalidates a common assumed compensating control and must be traceable to a primary source in your risk register.

Step 4: Communicate findings — brief leadership on exposure in AI/ML and cloud GPU infrastructure with emphasis on the hardware mitigation gap for non-ECC GPUs; frame as a pre-disclosure window requiring proactive inventory, not a wait-for-patch situation

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: Establishing communication plans and ensuring leadership awareness of exposure prior to exploitation is a preparation-phase obligation; also maps to RS.MA-01 (IR plan execution with relevant stakeholders)

Controls: NIST IR-6 (Incident Reporting) — requires timely reporting of suspected vulnerabilities and incidents to organizational leadership and IR capability owners, NIST IR-8 (Incident Response Plan) — the IR plan must include pre-incident communication procedures for high-severity hardware vulnerabilities with no immediate patch, NIST PM-12 (Insider Threat Program) — AI/ML infrastructure with shared GPU access represents an elevated insider and tenant threat surface that leadership must be made aware of, CIS 7.1 (Establish and Maintain a Vulnerability Management Process) — leadership communication is required when a vulnerability is rated critical and no vendor patch is yet available

Compensating: Prepare a one-page executive summary that quantifies business exposure: number of non-ECC GDDR6 GPU systems identified in Step 1, whether any are shared/multi-tenant (AI/ML training clusters, cloud GPU instances), and the specific gap — hardware-level Rowhammer on GDDR6 cannot be patched via software alone for non-ECC hardware. Use the CVSS 9.5 score and the IOMMU bypass finding as the two key technical facts leadership needs. Include a three-option decision matrix: (a) accept risk pending NVIDIA advisory, (b) isolate shared-GPU workloads to single-tenant use, (c) prioritize ECC GPU procurement for highest-risk systems.

Evidence: Attach the asset inventory output from Steps 1 and 2 as supporting evidence in the leadership brief — specifically the list of non-ECC GPU systems and their workload types (shared AI/ML, single-tenant, cloud-hosted). Document the date and recipients of the briefing and retain this record to establish that leadership was informed during the pre-disclosure window; this record is relevant to any subsequent regulatory notification timeline if exploitation is later confirmed.

Step 5: Monitor developments — track for the April 13, 2026 full technical release and any NVIDIA driver advisories issued in response; watch for CVE assignment, cloud provider advisories from Google, AWS, and Microsoft Azure, and independent researcher replication that would confirm exploitability

NIST Phase: Detection Analysis

Reference: NIST 800-61r3 §3.2 — Detection and Analysis: Continuous monitoring of threat intelligence sources and vendor advisories to detect escalation from theoretical to actively exploitable is a core detection-phase activity; maps to DE.AE-07 (CTI integration into adverse event analysis)

Controls: NIST SI-5 (Security Alerts, Advisories, and Directives) — mandates monitoring of external organizations (NVIDIA, CISA, cloud providers) for security advisories and acting within defined timeframes, NIST RA-3 (Risk Assessment) — risk level must be re-evaluated upon CVE assignment, full technical release, and independent replication, each of which changes the exploitability score, NIST IR-5 (Incident Monitoring) — requires tracking the

status of known threats including changes in severity, public exploit availability, and vendor response, CIS 7.1 (Establish and Maintain a Vulnerability Management Process) — the vulnerability management process must include a monitoring cadence triggered by the April 13, 2026 full disclosure date and any interim NVIDIA bulletin updates

Compensating: Set up three free monitoring channels requiring no SIEM: (1) Subscribe to NVIDIA Security Bulletins via RSS or email at nvidia.com/en-us/security — specifically watch for updates to a_id/5747 or new bulletins referencing GDDR6, GPU memory, or Rowhammer; (2) Monitor NVD (nvd.nist.gov) for CVE assignment using a keyword watch on 'GDDR6' or 'GPU Rowhammer' — NVD provides email alerts for new CVE entries; (3) Set a Google Scholar or arXiv alert for 'GPUBreach' and 'GDDR6 Rowhammer' to detect independent replication. On April 13, 2026, immediately review the full technical paper for PoC code, specific trigger conditions, and minimum unprivileged access requirements, then re-triage all non-ECC GDDR6 systems accordingly.

Evidence: Maintain a dated threat intelligence log for GPUBreach tracking each advisory, CVE assignment, cloud provider bulletin, and researcher publication with the date received and the specific new information it contained. If a CVE is assigned and CVSS changes, re-document the updated score against your asset inventory — this log establishes the timeline of organizational awareness and response, which is required for any post-incident regulatory review or forensic reconstruction of whether response was timely relative to public exploit availability.

Detection Guidance

No confirmed IOCs or observed exploitation have been reported as of this disclosure. Detection opportunities are currently limited to behavioral and anomaly-based approaches. For GPU-accelerated systems: monitor for unexpected privilege escalation events (sudo, SYSTEM-level process spawning) originating from processes with GPU driver access; audit logs for anomalous NVIDIA driver interactions, particularly unusual ioctl calls or driver interface access from unprivileged processes. For cloud environments: review GPU instance activity logs for unexpected inter-tenant behavior or privilege changes; coordinate with cloud provider security teams for any GPU hypervisor-layer telemetry they can expose. Hunt for T1611 (Escape to Host) patterns, processes executing outside expected container or VM boundaries on GPU nodes. Flag any host compromise events on systems running NVIDIA drivers for retroactive review once the full technical details publish on April 13, 2026. Organizations should also audit whether ECC memory is enabled or available on deployed NVIDIA hardware, as this status is directly relevant to hardware-layer exposure.

Framework Mappings

MITRE-ATTACK

- **T1082** — System Information Discovery
- **T1611** — Escape to Host
- **T1212** — Exploitation for Credential Access
- **T1068** — Exploitation for Privilege Escalation
- **T1055** — Process Injection
- **T1203** — Exploitation for Client Execution

NIST-800-53R5

- **AC-6** — Least Privilege
- **SC-7** — Boundary Protection
- **SI-2** — Flaw Remediation
- **SI-3** — Malicious Code Protection

- **SI-4** — System Monitoring
- **SI-16** — Memory Protection
- **SI-10** — Information Input Validation

OWASP-TOP10-2021

- **A03:2021** — Injection

CIS-V8

- **16.10** — Apply Secure Design Principles in Application Architectures
- **5.4** — Restrict Administrator Privileges to Dedicated Administrator Accounts

ISO-27001-2022

- **A.5.23** — Information security for use of cloud services

SOC2-TSC

- **CC6.3** — Authorizes, modifies, or removes access

MITRE ATT&CK Mapping

Technique ID	Technique Name	Tactic
T1082	System Information Discovery	Discovery
T1611	Escape to Host	Privilege-Escalation
T1212	Exploitation for Credential Access	Credential-Access
T1068	Exploitation for Privilege Escalation	Privilege-Escalation
T1055	Process Injection	Defense-Evasion
T1203	Exploitation for Client Execution	Execution

Sources

Source	URL	Tier
Security News	https://www.bleepingcomputer.com/news/security/new-gpubreach-attack...	T3
GPUHammer: 1,171 Bit Flips Expose Nvidia GPU Flaw [2026]	https://tech-insider.org/gpuhammer-nvidia-gpu-rowhammer-gddr6-vulne...	T3
Nvidia drivers are affected by a security vulnerability, update asap	https://www.reddit.com/r/linux_gaming/comments/1givvlt/nvidia_drive...	T3

Source	URL	Tier
Security Bulletin: NVIDIA GPU Display Drivers - January 2026	https://nvidia.custhelp.com/app/answers/detail/a_id/5747/~/security...	T3
NVIDIA GPU Driver Vulnerability Opens Door to Elevated Privileges	https://www.linkedin.com/pulse/nvidia-gpu-driver-vulnerability-open...	T3

Generated 2026-04-07 06:06 UTC by TJS Security Command Center