Information leakage through shared GPU hardware in data centers

Shared GPU infrastructure in cloud data centers leaks information through at least a dozen proven, reproducible attack channels spanning memory residue, cache side channels, interconnect contention, and container escapes. Between 2018 and early 2026, researchers demonstrated that co-located attackers can eavesdrop on LLM conversations, steal entire neural network architectures, corrupt model weights via rowhammer, and escape container boundaries to access other tenants' data — often with fewer than 10 lines of attack code. NVIDIA's Multi-Instance GPU (MIG), the industry's primary hardware isolation mechanism, has been bypassed via unpartitioned TLBs. (ACM Digital Library) Confidential computing on H100/Blackwell GPUs represents the most promising defense, but adoption remains early-stage and the technology has known limitations including reliance on proprietary firmware trust chains. For any organization running sensitive AI workloads on shared infrastructure, the threat surface is significantly larger than commonly understood.

GPU memory is not your friend: residue and local memory attacks

The most immediately dangerous class of GPU vulnerability involves **memory that persists between tenants**. GPUs, unlike CPUs, were not designed with multi-tenant security in mind — their memory management optimizes for throughput, not isolation.

LeftoverLocals (CVE-2023-4969) is the landmark finding. Discovered by Tyler Sorensen and Heidy Khlaaf at Trail of Bits and disclosed in January 2024, it demonstrated that GPU local memory (on-chip scratchpad) is not zeroed between kernel executions on AMD, Apple, Qualcomm, and Imagination GPUs. An attacker needs only ~10 lines of OpenCL code to read data left by previous processes. On an AMD Radeon RX 7900 XT running llama.cpp with a 7B model, the attack recovers **~181 MB per LLM query** (Trail of Bits) — enough to reconstruct full conversational responses in real time. NVIDIA GPUs were confirmed *not* affected. (Trail of Bits) AMD released a "secure compute" mode (disabled by default due to performance impact) in May 2024 (Leftoverlocals) and continues updating mitigations as of mid-2025.

Earlier work established the pattern. Roberto Di Pietro et al. showed in 2013 that NVIDIA GPU global memory retains data between process context switches ("CUDA Leaks"). (ResearchGate) Clémentine Maurice et al. at EURECOM confirmed in 2014 that this persists even in virtualized environments — memory is only incidentally cleared as a side effect of ECC, not for security. (ResearchGate) (Eurecom) Zhe Zhou et al. (2016) demonstrated recovery of raw rendering data, cryptographic keys, and model weights from GPU global memory. (Virascience) (ResearchGate)

No peer-reviewed cold boot attack on GPU VRAM (GDDR6 or HBM) has been demonstrated, though the principle of data remanence applies. The practical barriers are significant: GDDR6 is soldered to GPU boards, HBM is integrated into the package, and refresh rates are fast. Software-based memory residue attacks (LeftoverLocals, CUDA Leaks) are the functional GPU equivalent and far more practical.

MIG isolation is weaker than NVIDIA claims

NVIDIA MIG (Multi-Instance GPU), available on A100 and H100, is the primary hardware isolation mechanism marketed for multi-tenant GPU sharing. It partitions a GPU into up to 7 instances with dedicated

memory and compute resources. (NVIDIA +2) **Two independent research efforts have demonstrated that MIG isolation is incomplete.**

TunneLs for Bootlegging (Zhenkai Zhang et al., ACM CCS 2023) reverse-engineered NVIDIA's GPU TLB hierarchy and discovered that **MIG does not partition the last-level TLB**, which remains shared across all MIG instances. The researchers constructed a covert channel achieving **31 Kbps at 99.8% accuracy** across MIG isolation boundaries (Fan-yao) on A100 GPUs in a commercial cloud environment, and demonstrated application fingerprinting during DNN training. This is the first proven breach of MIG's isolation guarantees. (ACM Digital Library)

Veiled Pathways (Penn State, disclosed to NVIDIA February 2024) identified additional unpartitioned resources: NVDEC/NVENC/NVJPG hardware engines, dynamic voltage/frequency scaling (DVFS) observability, and PCIe bandwidth — all shared across MIG instances. These enable further covert and side channels that bypass both MIG and MPS isolation. (Psu)

A third paper, "**Behind Bars**", analyzed H100 MIG cache partitioning using memory barriers, finding additional side-channel leakage. AMD confirmed their MI3XX GPUs are not affected by this specific attack. (AMD)

Multi-GPU interconnects are a rich attack surface

Modern AI training relies on NVLink to connect GPUs at hundreds of GB/s. This high-bandwidth interconnect creates novel attack channels that operate *across GPUs* — an attacker on one GPU can observe or influence workloads on entirely separate GPUs.

NVBleed (Yicheng Zhang et al., arXiv March 2025) reverse-engineered NVLink operations and identified two leakage sources: timing variations from contention and user-accessible NVLink performance counters. (arXiv) (arXiv) The attack achieves a **covert channel exceeding 70 Kbps** (arXiv) and application fingerprinting with **F1 scores up to 97.78%** across 18 HPC/deep learning applications. (ResearchGate +2) Critically, NVBleed was demonstrated **cross-VM on Google Cloud Platform** — even VM boundaries do not prevent NVLink information leakage. (arXiv) (arXiv)

Spy in the GPU-box (Dutta et al., ISCA 2023) (arXiv) reverse-engineered the L2 cache hierarchy in NVIDIA DGX multi-GPU systems and built a Prime+Probe covert channel across GPUs via NVLink achieving **~4 MB/s** bandwidth. (ACM Digital Library) (arXiv) The side channel recovered the number of neurons in a co-located victim's neural network. (arXiv) (ACM Digital Library) **SideLink** (Baddour et al., CARDIS 2025) achieved 93% application fingerprinting accuracy on DGX A100 via NVLink contention. (Springer)

PCIe bus snooping enables lossless model theft

The PCIe bus connecting CPUs and GPUs carries unencrypted data in most current deployments, creating a devastating attack surface for AI model theft.

Hermes Attack (Zhu et al., USENIX Security 2021) demonstrated that by snooping the PCIe bus between CPU and GPU, an attacker can **reconstruct complete DNN models — architecture, hyperparameters, and all weights — with lossless inference accuracy**. (USENIX) (arXiv) Tested on NVIDIA GT 730, GTX 1080 Ti, and RTX 2080 Ti, the attack recovers MNIST, VGG, and ResNet models identically. (arXiv) This requires physical

access or a compromised PCIe device, but PCIe 5.0/6.0 IDE (Integrity and Data Encryption) — the hardware mitigation — is not yet widely deployed.

INVISIPROBE (Tan et al., IEEE S&P 2022) exploits shared PCIe bandwidth between devices connected via the same PCIe switch. When a victim GPU and an attacker's RDMA NIC share a switch, the attacker measures I/O delay variations to infer GPU workload identity with **96.3% accuracy** — including which ML model is running. ResearchGate ResearchGate **LockedDown** (Hartono et al., EuroS&P 2022) demonstrated the first cross-VM covert channel through virtualized GPU infrastructure (vGPU) via PCIe contention. Ucf

RDMA networks in AI clusters enable remote cache attacks

AI training clusters universally use RDMA (Remote Direct Memory Access) over InfiniBand or RoCE for inter-node communication. This introduces remote attack vectors that do not require co-location on the same physical machine.

NetCAT (Kurth et al., IEEE S&P 2020, CVE-2019-11184) exploited Intel DDIO (Data-Direct I/O), which gives NICs direct last-level cache access. The Hacker News vusec Using RDMA from a remote machine, an attacker performs Prime+Probe on the LLC to observe victim network activity, Andrea Fortuna demonstrating SSH keystroke timing extraction. DDIO is enabled by default on all Intel Xeon processors since 2012. The Hacker News

ReDMArk (Rothenberger et al., USENIX Security 2021) demonstrated that InfiniBand RDMA security mechanisms are fundamentally insufficient. arXiv Attacks include packet injection via impersonation, unauthorized memory access, and denial of service USENIX through discovered vulnerabilities in remote key generation algorithms in Mellanox NICs. This potentially enables unauthorized access to GPU memory regions via GPUDirect RDMA, threatening distributed AI training. **NeVerMore** (Taranov et al., CCS 2022) extended these attacks to NVMe-oF storage ACM Digital Library used in AI clusters. arXiv

GPUHammer proves rowhammer works on discrete GPU memory

GPUHammer (Chris S. Lin, Joyce Qu, Gururaj Saileshwar, University of Toronto; USENIX Security 2025) is the **first successful rowhammer attack on discrete GPU GDDR6 memory**. arXiv The researchers reverse-engineered proprietary GDDR6 DRAM row mappings on NVIDIA RTX A6000, developed GPU-specific memory access optimizations, and bypassed Target Row Refresh mitigations, arXiv achieving **8 bit-flips across 4 DRAM banks**. arXiv

The AI implications are severe. A single bit-flip in an FP16 weight's exponent MSB degraded ImageNet DNN model accuracy **from 80% to 0.1%** GPUHammer — the researchers dubbed this "Terminal Brain Damage." In time-shared GPU setups, an attacker positions victim data into vulnerable DRAM rows via memory massaging. GPUHammer NVIDIA acknowledged the attack in a July 2025 security notice and recommends enabling ECC, which mitigates all observed single-bit flips NVIDIA but introduces **~10% inference slowdown and 6.25% memory reduction**. GPUHammer ECC is enabled by default on Hopper and Blackwell data center GPUs; NVIDIA consumer GPUs typically lack it. HBM memory (used in A100/H100) showed no bit-flips in testing. GPUHammer

The earlier **GLitch** attack (Frigo et al., IEEE S&P 2018) used WebGL to perform rowhammer from GPU *to CPU memory*, demonstrating GPUs can be weaponized as attack tools against host DRAM.

CPU side channels compound the GPU threat surface

Every GPU server also has CPUs processing data for AI workloads, and the 2023–2025 period saw an acceleration of new CPU microarchitectural vulnerabilities:

- **Downfall** (CVE-2022-40982, Daniel Moghimi/Google, USENIX Security 2023): Leaks data from AVX-512 gather instructions on Intel 6th–11th Gen Core and Xeon Scalable 1st–4th Gen. (Wikipedia) AVX-512 is extensively used in CPU-side AI inference. Mitigation costs **up to 50% performance** in vectorized workloads. (The Hacker News)
- **Inception** (CVE-2023-20569, ETH Zurich, USENIX Security 2023): Affects all AMD Zen 1–4 processors including every EPYC generation used in AI data centers. (PCWorld)
- **Training Solo** (CVE-2024-28956, CVE-2025-24495, VUSec, May 2025): Self-training Spectre v2 attacks breaking Intel eIBRS protection. Leaks kernel memory at **17 KB/s** on all Intel server CPUs. (The Hacker News)
- **Branch Privilege Injection** (CVE-2024-45332, ETH Zurich COMSEC, May 2025): Restores full Spectre-BTI attack capability on all Intel CPUs since 9th Gen via branch predictor race conditions. (Phoronix +2)
- **Transient Scheduler Attacks** (CVE-2024-36350 and others, Microsoft, July 2025): Exploits timing leaks in AMD Zen 3 (EPYC Milan) and Zen 4 (EPYC Genoa) scheduler logic. (Wikipedia)
- **VMScape** (CVE-2025-40300, ETH Zurich, September 2025): Cross-VM Spectre-BTI attack exploiting incomplete branch predictor isolation in QEMU/KVM on AMD Zen 1–5 and Intel Coffee Lake. (Wikipedia)

For AI-specific exploitation, **Cache Telepathy** (Mengjia Yan et al., USENIX Security 2020) used CPU cache side channels to extract DNN architectures, reducing VGG-16's search space from **5.4×10^{12} to 16 candidate architectures**. (USENIX) **MoEcho** (Ding et al., ACM CCS 2025) introduced four novel side channels targeting Mixture-of-Experts LLMs (DeepSeek, Mixtral), achieving **99.8% prompt inference accuracy** and **92.8% response reconstruction** through CPU cache occupancy and GPU TLB eviction channels. (ResearchGate +2)

Container escapes are the most immediate practical threat

While side channels require sophistication, **NVIDIA Container Toolkit vulnerabilities have repeatedly enabled trivial container escapes** affecting the backbone of all cloud AI services:

- **CVE-2024-0132** (CVSS 9.0, Wiz Research, September 2024): TOCTOU vulnerability enabling full host root access from a malicious container. (Wiz) (Wiz) Over **35% of GPU-enabled cloud environments** were at risk. The initial patch was incomplete — the bypass was tracked as CVE-2025-23359. (Wiz)
- **CVE-2025-23266 "NVIDIAScape"** (CVSS 9.0, Wiz Research, Pwn2Own Berlin May 2025): Container escape via OCI hooks misconfiguration, **exploitable with a 3-line Dockerfile**. (Wiz) Affects all NVIDIA Container Toolkit versions up to v1.17.7.

Wiz Research's 2024 "Isolation or Hallucination?" program demonstrated practical cross-tenant attacks on Wiz **Replicate** (accessed shared Redis server serving multiple customers), Wiz **SAP AI Core** (full cross-tenant access to AWS S3 buckets and EFS instances), SC Media and **Hugging Face** (container escape to other customers' models). TechTarget Their key finding: AI service providers are inherently more vulnerable because allowing users to run AI models is equivalent to allowing arbitrary code execution. The Hacker News

NVIDIA vGPU software shows a recurring pattern of high-severity vulnerabilities enabling guest-to-host escalation, including CVE-2025-33220 (use-after-free), CVE-2025-23283 (stack buffer overflow), and multiple 2024–2025 CVEs discovered by Microsoft's Offensive Research team.

Network timing reveals LLM outputs even through encryption

A newer class of attacks targets AI inference through pure network observation, requiring no co-location whatsoever.

Wiretapping LLMs (IACR ePrint 2025/167) demonstrated that streaming APIs over HTTPS leak token information through inter-packet timing patterns created by speculative decoding. Language identification reaches **77–83% accuracy** across multilingual models, with certain languages (Chinese, Russian, Korean, Arabic) at **near 100%**. The attack works against commercial services including GPT-4o.

PromptPeek (NDSS 2025) exploits KV-cache sharing in multi-tenant LLM serving systems (vLLM, SGLang). An adversary crafts requests to determine whether KV-cache was reused from a victim's cached data, leaking prompt prefixes through timing. NDSS Symposium **"The Early Bird Catches the Leak"** (2024) further identifies timing side channels from shared KV caches and GPU memory allocations in production LLM serving systems. Semantic Scholar

Timing Channels in Adaptive Neural Networks (Akinsanya et al., NDSS 2024) showed that early-exit networks create measurable timing differences exploitable over the public internet, revealing sensitive attributes of user inputs. NDSS Symposium

Confidential computing: promising but nascent and not yet proven

NVIDIA's H100 (Hopper) introduced the **first commercial GPU confidential computing**, arxiv arXiv featuring an on-die hardware root of trust, NVIDIA Developer AES-GCM 256 encryption for CPU-GPU transfers, ACM PCIe/NVLink firewalls, measured boot, and attestation. Introl Introl Performance overhead is **below 5% for most LLM queries** and approaches zero for large models. Microsoft Community Hub arXiv Azure offers GA availability (NCC H100 v5 VMs with AMD SEV-SNP), NVIDIA Blog and Google Cloud supports H100 CC with Intel TDX on A3 instances. Google

NVIDIA Blackwell advances the architecture significantly with Introl **TEE-I/O (TDISP)** for hardware-encrypted direct GPU-VM communication (eliminating software bounce buffers), multi-GPU confidential computing with encrypted NVLink, arXiv and near-native performance. NVIDIA The announced **Vera Rubin NVL72** promises rack-scale confidential computing across 72 GPUs. NVIDIA

However, significant limitations and attacks exist:

- **TEE.Fail** (Georgia Tech & Purdue, October 2025): Physical memory-bus interposition attack on DDR5 systems (requiring <$1,000 equipment) that can **forge TDX attestations and fake NVIDIA GPU**

attestation reports. Demonstrated extraction of ECDH private keys from enclaves.

- **RMPocalypse** (CVE-2025-0033, ETH Zurich, 2025): A single 8-byte write during AMD Secure Processor RMP initialization compromises **all integrity guarantees of SEV-SNP** with 100% success rate.
- **CVE-2024-56161** (Google, February 2025, CVSS 7.2): Improper signature verification in AMD CPU ROM microcode patch loader allows loading malicious microcode, destroying SEV-SNP confidentiality guarantees. The Hacker News
- **Heracles** (ETH Zurich, 2025): Chosen plaintext attack on SEV-SNP leaking 16-character passwords in **~6.5 seconds**. Heracles-attack
- **Fundamental trust issue**: GPU firmware is proprietary and closed-source. The attestation chain relies on NVIDIA's Remote Attestation Service. Independent verification is impossible. arXiv As one security CEO noted, "attestation demands trust with the cloud provider, which in many ways beats the purpose of confidential computing." InfoQ

Research prototypes (Graviton/OSDI 2018, ACM Other conferences HIX/ASPLOS 2019, ResearchGate HETEE/S&P 2020, arXiv StrongBox/USENIX Security 2022, GEVisor/SoCC 2023) Semantic Scholar explore alternative GPU TEE designs, but none are production-ready. ARM CCA shows promise for integrated GPU security (CAGE/NDSS 2024), but commercial hardware is not yet available.

Supply chain and firmware: the unauditable trust base

GPU firmware represents a particularly opaque attack surface. The **JellyFish** GPU rootkit PoC (2015) demonstrated that GPU malware can snoop host memory via DMA and persist across warm reboots. SecurityWeek In August 2021, a toolkit for hiding malicious code in GPU memory was sold on a hacker forum, supporting AMD, NVIDIA, and Intel GPUs via OpenCL 2.0+. Bleeping Computer

NVIDIA mitigates firmware concerns in CC mode through encrypted and signed firmware, secure boot chains, on-die root of trust, and firmware revocation. NVIDIA Developer Blackwell adds AES-CBC 128-bit encrypted secure flash. Uvation However, **the GPU firmware stack is entirely closed-source** — it is part of the Trusted Computing Base but cannot be independently audited. arXiv A compromised firmware update could theoretically undermine all confidential computing guarantees. NVIDIA publishes quarterly security bulletins addressing driver vulnerabilities, NVIDIA with a consistent cadence of high-severity CVEs in display drivers, vGPU software, and CUDA toolkit components. NVIDIA Security Online

What cloud providers claim versus what researchers demonstrate

Mechanism	Provider claim	Research reality
MIG partitioning	"Full isolation of the entire GPU memory system" (NVIDIA) Scaleway	TLB shared across instances; covert channels demonstrated at 31 Kbps (CCS 2023) Fan-yao
Container isolation	"Namespace and cgroup isolation" (Kubernetes GPU providers)	Trivially escapable with 3-line Dockerfile (CVE-2025-23266) Wiz

Mechanism	Provider claim	Research reality
vGPU isolation	"IOMMU-backed hardware isolation" (NVIDIA GRID)	Recurring guest-to-host escape CVEs in vGPU Manager (quarterly pattern) Liquid Web
Memory scrubbing	"GPU memory cleared between tenants" (most CSPs)	Race conditions allow reading data before wipe completes; not all GPUs zero local memory (LeftoverLocals) Liquid Web
ECC for integrity	"Error correction prevents bit-flip attacks" (NVIDIA)	Effective but 10% performance penalty; not available on consumer GPUs GPUHammer
Dedicated instances	"Full physical isolation" (AWS, Azure)	Eliminates cross-tenant risk but at significant cost premium
Confidential computing	"Hardware-encrypted, attested GPU TEE" (NVIDIA H100/Blackwell)	TEE.Fail forges attestation; underlying CPU TEEs have multiple bypasses; proprietary firmware unauditable

Conclusion

The attack surface of shared GPU infrastructure is **broader, deeper, and more actively exploited in research** than the cloud industry's security messaging suggests. Five categories of threat stand out as most critical for organizations running sensitive AI workloads:

1. **Memory residue attacks** (LeftoverLocals) enable real-time eavesdropping on LLM sessions with trivial exploit code Wiz — and AMD GPUs remain only partially patched.
2. **MIG isolation bypasses** undermine the primary hardware mechanism cloud providers recommend for multi-tenant GPU sharing.
3. **Container infrastructure vulnerabilities** (CVE-2024-0132, CVE-2025-23266) provide trivial cross-tenant compromise paths affecting the entire cloud AI ecosystem. Introl Wiz
4. **Interconnect side channels** (NVBleed, Spy in the GPU-box) demonstrate that multi-GPU AI training configurations leak information across VM boundaries even on major cloud platforms. arXiv +2
5. **CPU microarchitectural attacks** continue accelerating, with Training Solo, BPI, TSA, and VMScape all disclosed in 2025, each affecting processors standard in AI data centers. Wikipedia

Confidential computing on H100/Blackwell GPUs is the industry's convergence point for long-term defense, but it requires trusting NVIDIA's proprietary firmware, arXiv faces proven attestation forgery attacks (TEE.Fail), and sits atop CPU TEEs (AMD SEV-SNP, Intel TDX) with their own growing vulnerability lists. For truly sensitive workloads today, **single-tenant dedicated GPU instances remain the only configuration that eliminates cross-tenant risk** NVIDIA — at a cost that makes multi-tenant sharing economically compelling and therefore the norm. The gap between security research findings and production cloud security posture represents one of the most significant unaddressed risks in the AI infrastructure stack.