GPUs, the next frontier for security professionals

Ted Gould
SCALE 23x
March 7, 2026

WHOAMI

- **Lead Engineer @ Stealthium.io**
  - Building out an end point detection system for GPUs
  - Coordinating a group of security researchers
  - Doing all the leftover engineering tasks
- **Previously**
  - Axiom: built out a sophisticated cloud data platform
  - Canonical: application confinement on Ubuntu Phone

STEALTHIUM

A large and GROWING problem

Scaling at the Speed of AI

The fleet sizes required for modern LLM training have shifted the goalposts for infrastructure teams.

- **xAI**: 1,000,000 GPUs targeted by end of 2026.
- **CoreWeave**: 250,000 GPUs across 32 datacenters.
- **Hyperscalers**: Microsoft and AWS fleets reaching the hundreds of thousands.

Economics of the Attack Surface

In a traditional CPU environment, a compromise is about a pivot. In a GPU fleet, a compromise is a direct financial drain.

Hardware Cost: $30k–$40k per H100 unit.

Revenue Potential: $5–$12 per GPU-hour.

The Prize: A 2,048 GPU cluster represents a $70M+ investment.

Historical Prescident

1865: The Steam Boom

Economic Acceleration vs. Material Science

- Boilers were pushed beyond physical limits to outpace rivals.
- Safety valves were frequently screwed down to increase pressure.

The Sultana Disaster

1,000+ Lives Lost

The deadliest maritime disaster in U.S. history resulted from uninspected, overtaxed steam boilers.

STEALTHIUM

Correction: Standardized Safety

From Disaster to Operational Cornerstone

- **1880**: American Society of Mechanical Engineers (ASME) founded.
- **1884**: First uniform Boiler Testing Code issued.

1931: Early Aviation

"Daring" vs. "Safe"

- Aircraft were built of wood and fabric; engines failed often.
- The "Knute Rockne" crash revealed fatal structural failure in wing glue.

Correction: Proactive Data

The "Safety-First" Transformation

- Shift from reactive crash investigation to proactive Safety Management Systems (SMS).
- Mandatory all-metal designs and public accident reports.

2016: The IoT Explosion

Connectivity Without Encryption

- Manufacturers prioritized time-to-market over security design.
- Many devices had default passwords and unencrypted communication.
- The Mirai Botnet hijacked millions of insecure cameras and routers.

Almost 98%

Percentage of IoT device traffic that remains unencrypted.

STEALTHIUM

Over 50%

Percentage of IoT devices vulnerable to high-severity attack.

STEALTHIUM

We are watching the last 10 years of cloud security lessons being forgotten in the rush to scale.

Who's Responsibility?

Fig. .1: No one knows

STEALTHIUM

Understanding CPU and GPU Relationship

CPU

- Has the core Operating System
- Security software that is watching processes
- Workload Isolation (containers, VMs, etc.)

GPU

- Has built in Firmware
- Can communicate on PCI bus to CPU
- Isolation largely CPU defined

STEALTHIUM

Example: CoffeeLoader

CPU ⟶ GPU

- An attacker sends an encrypted payload to the GPU.
- Loads the decryption algorithm

CPU ⟵ GPU

- Attack can no be put into CPU memory without checking

Memory Isolation

From Spectre to LeftoverLocals

- **CPU Ancestor**: **Spectre/Meltdown (2018)** – Speculative execution broke kernel/user isolation.
- **GPU Reality**: **LeftoverLocals (CVE-2023-4969)** – VRAM local memory is often not cleared between kernel launches.

Memory Isolation

From Spectre to LeftoverLocals

- **Consequence**: ~181 MB leaked per LLM query; attackers can reconstruct interactive responses.
- **The Mitigation**:
  - Mandatory driver-level zero-initialization (Scrubbing).
  - Utilize **Multi-Instance GPU (MIG)** for hardware-enforced partitioning.
  - Avoid sharing GPUs between workload types

Privilege & Escape

From Ring 3 to Container Breakouts

- **CPU Ancestor**: **Ring 3 to Ring 0 Escapes** – Exploit kernel vulnerabilities for hardware control.
- **GPU Reality**: **NVIDIAEscape (CVE-2024-0132)** – Critical flaws in the Container Toolkit allowed host compromise via GPU hooks.

Privilege & Escape

From Ring 3 to Container Breakouts

- **Consequence**: Tenant A escapes their container; steals Tenant B’s model weights directly from host memory.
- **The Mitigation**:
  - **Hypervisor-Level Monitoring**: Deploy runtime observability to track GPUs.
  - **Least Privilege**: Strip unnecessary driver capabilities from guest environments.

Supply Chain & Persistence

From SolarWinds to Firmware Rootkits

- **Supply Chain**: Attackers insert **Poisoned CUDA Kernels** on Hugging Face or use **CoffeeLoader** (malware that uses GPU-accelerated decryption to hide from CPU-centric EDR).
- **Persistence (CVE-2024-0146)**: Guest VMs exploiting vGPU software to write to host hardware firmware.

Supply Chain & Persistence

From SolarWinds to Firmware Rootkits

- **The "Forever Compromise"**: The attacker owns the GPU card itself, surviving reboots and OS wipes.
- **The Mitigation**:
  - **Runtime Computation Audits**: Monitor for unauthorized modifications to model computation graphs.
  - **Attestation**: Cryptographic verification of firmware states at runtime using Root of Trust (RoT).

Resource Exhaustion

Denial of Service (DoS)

- **CPU World**: Overwhelming system memory or I/O interfaces to cause "thrashing" and service outages.
- **GPU Gang Scheduling Death Spiral**: Distributed training synchronization failures cause GPUs to sit idle while "burning money".

Resource Exhaustion

- **Cost**: "Idle" GPUs can add up to millions of dollars in cost
- **The Mitigation**:
  - **Topology-Aware Scheduling**: Use schedulers (like SUNK) that understand NVLink topology and enforce runtime resource quotas.

Intellectual Property

- **Database Scraping**: Exfiltrating proprietary data or code to replicate business logic.
- **Model Extraction & Distillation**: Querying a model repeatedly to replicate its functionality at less than 1% of the original training cost.

Intellectual Property

- **Economic Impact**:
  - $500,000,000 is the estimated training cost for a frontier model like Llama 3.
  - Models trained on propriety data may include it
- **The Mitigation**:
  - **GPU Abuse Detection**: Implement rate limiting and identify patterns consistent with distillation probing at the GPU layer.
  - **Memory Behavior Analysis**: Watch GPU memory accesses and frequencies to track validity

STEALTHIUM

The Risk to Business

- **Intellectual Property**: Exfiltrating a model can replicate functionality at <1% of the $500M training cost and steal internal IP.
- **Operational Waste**: Unmonitored cryptomining is a silent tax on AI budgets.
- **Compliance Wall**: GDPR/HIPAA requires proof of isolation, not just "trust."

Strategy for Resilience

Actionable Steps for 2026

1. **Visibility**: Treat GPUs as first-class citizens in the security stack (not black boxes).
2. **Verified Scrubbing**: Don't rely on cloud defaults; implement automated zeroing between shifts.
3. **GPU Abuse Detection**: Identify patterns consistent with model distillation or "non-computational" behavior.
4. **Auditability**: Generate tamper-evident trails of every GPU operation.

Any Questions???

https://www.gould.cx/ted/presentations/
ted@gould.cx
@ted@gould.cx

STEALTHIUM