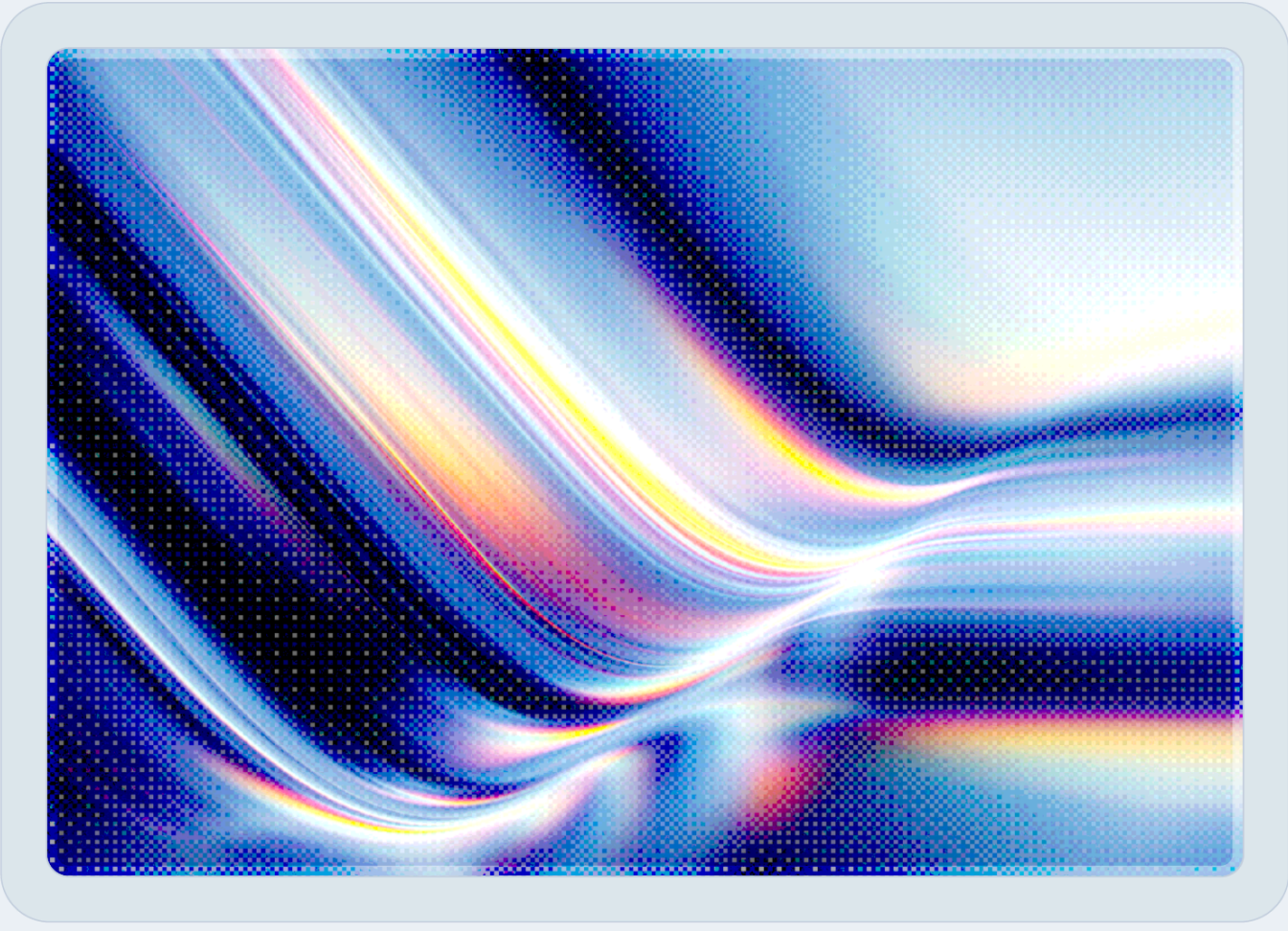
CSAI Foundation | Cloud Security Alliance

# Bit2Watt: GPU Workloads as a Cyber-Physical Grid Attack

## Legitimate Cloud Tenants Can Destabilize Power Infrastructure Without Any Exploit

2026-07-22

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

## Key Takeaways

Researchers at Zhejiang University have described Bit2Watt, a cyber-physical attack in which a cloud tenant with entirely legitimate GPU access can destabilize the power grid serving a data center simply by shaping how it uses its own compute allocation [1][3]. No exploit, malware, or stolen credential is required; the attacker's workload never leaves the boundaries of what a cloud provider's acceptable-use policy already permits, which is one reason the technique is difficult to detect with conventional security tooling that assumes malicious activity looks anomalous at the network or process layer. In simulation, synchronized power modulation across roughly 1,000 GPUs drawn from a 1-megawatt facility with heavy reliance on distributed energy resources pushed current harmonic distortion to 46.8%, well past the 13% threshold typically used as a stability guideline, and drove the local grid's damping ratio negative, a condition associated with self-sustaining oscillation rather than the passive decay a healthy grid exhibits [1][2]. The same underlying mechanism doubles as a covert channel, encoding data in power-draw frequency and recovering it error-free from nearby electromagnetic emissions, a companion technique the researchers call Watt2Bit [2]. The attack's practical reach is bounded today by a hard coordination requirement — modulating thousands of physically co-located GPUs in tight lockstep — that the authors themselves describe as an open problem, but the underlying physics connects directly to concerns Microsoft, OpenAI, and NVIDIA raised in August 2025 about synchronized AI training loads interacting badly with grid resonance [4].

## Background

Modern GPU accelerators draw power in a way that tracks their computational state almost instantaneously: a chip idling between kernel launches consumes a small fraction of the power it draws mid-computation, and the transition between the two happens in microseconds. That responsiveness is normally an efficiency feature. Bit2Watt, published as a preprint and accepted to CHES 2026 — the International Association for Cryptologic Research's hardware security conference — reframes it as an attack surface by showing that a tenant who controls enough GPUs inside a single facility can turn ordinary compute scheduling into a deliberate, high-frequency power signal that propagates out through the data center's electrical system and into the connected grid [1][2][3].

The researchers, Zhouhao Ji, Kaikai Pan, and Wenyuan Xu of Zhejiang University, describe two ways to generate this signal. The first, which the reporting on their work refers to as the Synthetic Workload Method, uses custom CUDA kernels that toggle deliberately between high-intensity computation and idle states, producing power oscillations in the 1.5–6 kHz range — far above the few hertz typical of household electrical loads [1]. The second, the LLM Training Method, achieves a similar effect by adjusting hyperparameters within an otherwise ordinary large language model training run, which both increases the achievable amplitude and blends the resulting oscillation into the power-consumption noise that large training jobs already produce, making it substantially harder to distinguish from benign activity [1]. Both techniques operate, in the authors' words, "entirely within the cyber layer as a legal tenant," meaning the traffic a security team would review looks like nothing more than a customer running compute-intensive workloads it has every right to run [3].

What turns this from a curiosity into a grid-stability concern is where the power lands. Data centers increasingly co-locate with, or draw substantially from, distributed energy resources such as rooftop solar and battery storage, which use power electronics to interface with the grid rather than the large rotating generators that give traditional grids their inertia and natural damping [1][3]. Grids with high penetration of these inverter-based resources are inherently more sensitive to fast, high-frequency disturbances, which is the kind of signal a synchronized GPU cluster can produce on demand.

## Security Analysis

The Zhejiang University team modeled a worst-case scenario using roughly 1,000 GPUs — spanning data-center accelerators such as the A100 and Tesla V100 as well as consumer cards like the RTX 4090 — operating within a 1-megawatt facility at 90% distributed energy resource penetration [1][2][3]. Under synchronized modulation, the simulation produced current total harmonic distortion of 46.8%, more than three times the 13% level commonly used as a power-quality guideline, and a damping ratio of -0.27 [1][2]. A negative damping ratio indicates that oscillations in the system grow over time rather than fade, which is the signature of an unstable rather than merely noisy grid [1][2]. Extending the simulation to a European-style transmission network, the researchers observed cascading effects that shed on the order of 80% of connected load, and separately noted that the same modulation can trip thermal or overcurrent protection circuits, causing an involuntary shutdown of the GPU servers themselves — a denial-of-service outcome distinct from, but layered on top of, the grid-stability risk [1][2].

The attack's covert-channel variant, Watt2Bit, repurposes the same power-modulation capability for data exfiltration rather than grid disruption. By encoding a logical one as a roughly 2 kHz oscillation and a logical zero as roughly 200 Hz, the technique transmits bits that can be recovered from electromagnetic emissions near the hardware with zero errors in the researchers' testing, without any network egress that

a data-loss-prevention system would ordinarily inspect [2]. Because both variants rely on manipulating a physical property (power draw) rather than any data-plane behavior, they sit outside the visibility of the network- and process-focused controls most cloud security stacks are built around, though this note is not aware of a survey establishing that gap across the industry.

Detecting the attack in production is complicated by a mismatch between how fast the signal moves and how fast existing telemetry samples it. The table below summarizes the gap the researchers identified between common power-monitoring interfaces and the frequencies the attack actually uses.

| Telemetry Source | Typical Sampling Rate | Sufficient for Bit2Watt Detection? |
|---|---|---|
| Rack PDU counters | ~1 Hz (once per second) | No |
| NVIDIA NVML | ~450 Hz | No |
| RAPL / BMC interfaces | ~1 kHz | Marginal |
| Bit2Watt modulation frequency | 1.5–6+ kHz | — |

*Source: [1]. The "Marginal" assessment for RAPL/BMC interfaces reflects this note's own synthesis of the sampling-rate gap rather than a labeled verdict from the researchers.*

A research-grade detector built from this conventional telemetry performed poorly against the attack, and the researchers concluded that reliable detection would require dedicated GPU-level power profiling combined with electromagnetic-interference sensing purpose-built for this frequency range — instrumentation that is not part of standard cloud telemetry stacks today, as far as the researchers could determine [1].

Among the mitigating factors the researchers identify, coordination difficulty appears to be the most limiting today: achieving the modeled worst case requires synchronizing more than a thousand physically clustered GPUs in near-perfect lockstep, and introducing timing jitter with a standard deviation of just 100 microseconds reduced the aggregate signal amplitude by roughly 20%. The researchers explicitly describe achieving reliable real-world synchronization at scale as an open problem rather than a solved one [1]. That caveat matters: Bit2Watt as demonstrated is a controlled-simulation finding from an academic hardware-security venue, not a technique observed in an active incident, and organizations should weigh it as a credible emerging risk rather than an imminent, proven exploitation path. This note nonetheless assesses the broader trend as concerning: in August 2025, nearly sixty researchers across

Microsoft, OpenAI, and NVIDIA co-authored a paper warning that synchronized power swings from large AI training jobs can, when their frequency coincides with a utility's critical resonance frequencies, cause physical damage to grid infrastructure even without any adversarial intent behind the training run [4]. Bit2Watt takes that same physical coupling and asks what happens when the synchronization is deliberate rather than incidental. The Microsoft/OpenAI/NVIDIA authors called for power-aware training algorithms and standardized grid-facing telemetry to address the unintentional case; those same mitigations, read alongside Bit2Watt, would also blunt a deliberate version of the same physical coupling [4].

# Recommendations

## Immediate Actions

Cloud and data-center operators running large GPU fleets, particularly those co-located with or drawing significantly from distributed energy resources, should inventory which facilities have the aggregate GPU density and DER penetration profile the Bit2Watt research flags as high-risk, and should confirm with their utility or grid operator what harmonic distortion and damping margins that connection point actually carries today. Security and facilities teams should also treat sustained, unusual power-draw oscillation patterns — even from workloads that are contractually and technically legitimate — as worth a facilities-side look, since the entire premise of this attack is that it will not surface in conventional security telemetry.

## Short-Term Mitigations

Operators should evaluate local energy buffering, such as battery or supercapacitor systems, and harmonic filtering at the facility power interconnect, both of which the original researchers identify as direct countermeasures to the specific oscillation frequencies the attack exploits [1][2]. On the compute side, anomaly detection tuned to GPU utilization and training-job scheduling patterns — rather than network behavior alone — can help surface synthetic workloads that toggle between idle and full-load states in ways ordinary training or inference traffic does not. Cloud providers should also begin closing the telemetry gap identified above, since PDU- and NVML-level sampling — common in production facilities — runs one to four orders of magnitude too slow to observe the attack's actual modulation frequency.

### Strategic Considerations

The durable fix here is architectural rather than tactical: cross-layer defenses that jointly monitor the compute and power domains, so that a facility's grid-interconnect telemetry and its GPU scheduling telemetry can be correlated in something close to real time. The researchers behind Bit2Watt describe this integration as future work rather than a solved problem, and the Microsoft/OpenAI/NVIDIA paper independently calls for the same thing from the opposite direction — asynchronous, power-aware training algorithms and standardized communication channels between AI operators and grid operators for resonance specifications, ramp rates, and load signaling [1][4]. Organizations planning new GPU capacity in grids with high DER penetration should treat this coordination as a design requirement for the interconnection agreement itself, not an afterthought layered on after commissioning.

## CSA Resource Alignment

Bit2Watt sits at the intersection of two domains CSA has already published guidance on separately, and this note reads its core recommendation — treating compute and power telemetry as a single monitored system rather than two disconnected ones — as an extension of that existing work rather than a wholly new problem.

CSA's Zero Trust Guidance for Critical Infrastructure [5] is the most directly applicable prior artifact. It was written for exactly the environment Bit2Watt threatens: operational technology and industrial control systems where legacy assumptions about implicit trust in "authorized" activity break down. Its five-step Zero Trust implementation model — define the protect surface, map operational flows, build the architecture, create policy, then monitor and maintain — offers a plausible starting structure for the Bit2Watt problem once a data center's power interconnect and GPU scheduling fabric are understood as a protect surface in their own right, one where a "legal tenant" is not automatically a trusted one for purposes of physical-layer impact.

More broadly, the underlying detection and control gap — legitimate-looking workload behavior producing an unmonitored physical-layer effect — falls within control domains covering threat and vulnerability management and AI-specific risk in CSA's AI Controls Matrix (AICM) v1.1 [6], which organizations can use as a control baseline while more Bit2Watt-specific guidance develops. Notably, a confirmed Bit2Watt-style incident at a European data center with grid interconnection would sit at the same intersection of AI-specific risk and critical-infrastructure regulatory obligation addressed by the EU AI Act and NIS2, underscoring that grid-facing AI compute is increasingly a dual-regulated surface and not purely a technical one.

# References

[1] The Hacker News. "New Bit2Watt Attack Could Let Cloud Tenants Disrupt Power Grids Without an Exploit." The Hacker News, July 2026.

[2] The Register. "Malicious cloud customers can bring down the power grid." The Register, July 20, 2026.

[3] Ji, Zhouhao, Kaikai Pan, and Wenyuan Xu. "Bit2Watt: A Cyber-Physical Vulnerability Exploiting GPU Workloads Across Power and Computing Infrastructures." arXiv preprint, accepted to CHES 2026.

[4] Microsoft, OpenAI, and NVIDIA. "Power Stabilization for AI Training Datacenters." arXiv preprint, August 2025.

[5] Cloud Security Alliance. "Zero Trust Guidance for Critical Infrastructure." Cloud Security Alliance, 2024.

[6] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.