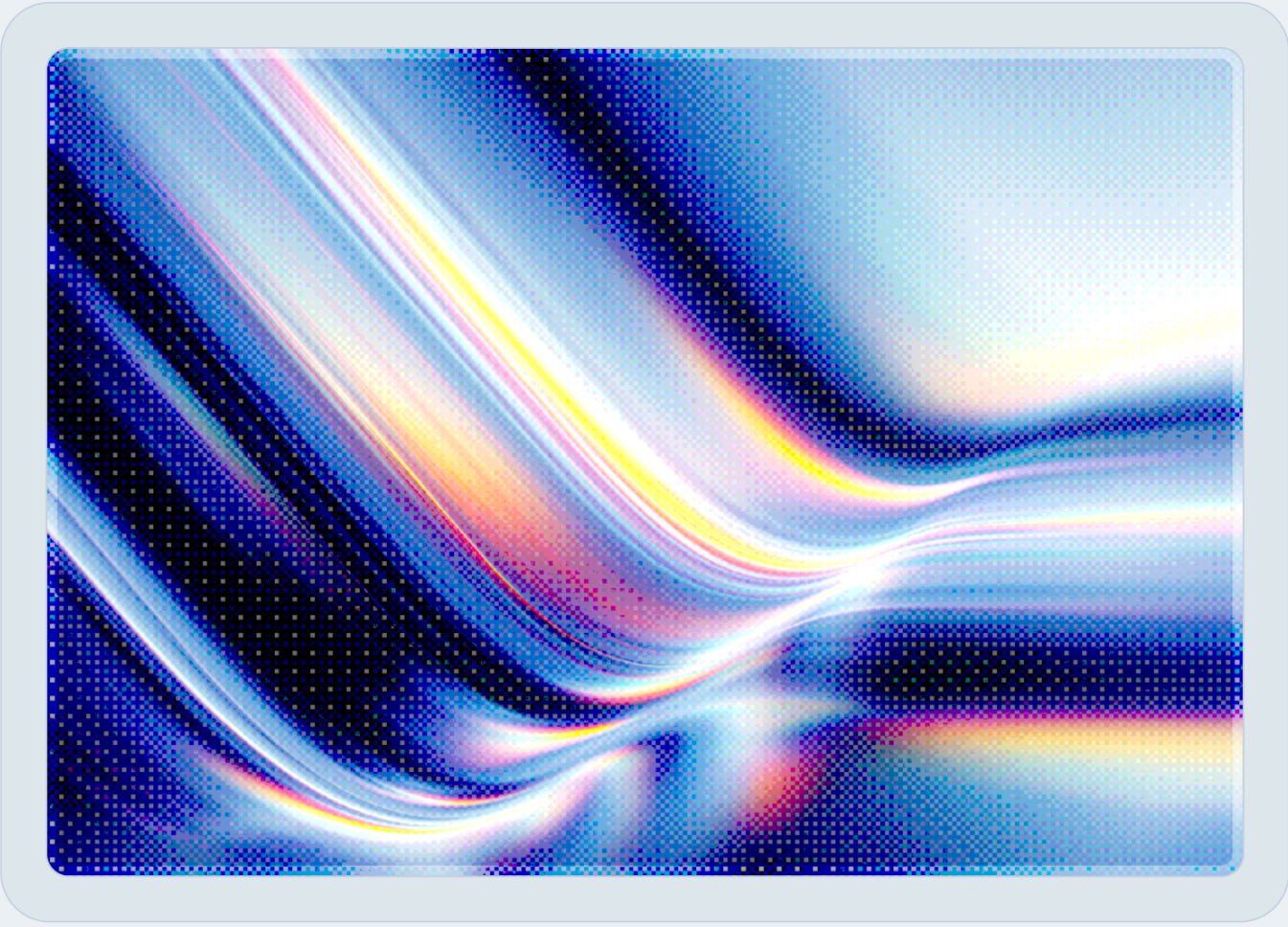
CSAI Foundation | Cloud Security Alliance

Bit2Watt: When a Cloud GPU Tenant Attacks the Power Grid

2026-07-21

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Researchers Zhouhao Ji, Kaikai Pan, and Wenyuan Xu of Zhejiang University have disclosed Bit2Watt, a cyber-physical vulnerability class in which a fully authorized cloud tenant can destabilize the power grid supplying a GPU data center simply by scheduling ordinary-looking compute jobs in carefully timed patterns [1][2]. The attack requires no credential theft, no network intrusion, and no violation of cloud provider terms of service; it works entirely through the electrical side effects of legitimate GPU workloads, which makes it largely invisible to the security monitoring tools that cloud providers and tenants rely on today [1].

In the paper's worst-case modeling, synchronizing the power draw of 1,000 GPUs against a 1-megawatt local grid segment with heavy penetration of solar and battery-based distributed energy resources drove current total harmonic distortion to 46.8 percent — far beyond the limits set by the IEC 61000-3-12 power-quality standard — and pushed the system's damping ratio to -0.27, a value indicating the grid segment had entered a self-sustaining oscillatory instability rather than a merely noisy one [1][2]. The researchers also describe a companion technique, Watt2Bit, in which the same power-modulation channel is used in reverse: as a covert data-exfiltration path that encodes information in electrical fluctuations, and as a mechanism by which grid-side protective equipment tripping in response to the disturbance can trigger denial-of-service conditions across the compute cluster itself [1][3].

Bit2Watt is best understood as a new category of tenant-isolation failure: cloud providers have historically modeled tenant isolation in terms of compute, memory, network, and storage boundaries, but Bit2Watt demonstrates that the electrical domain — the shared power infrastructure underneath every rack in a multi-tenant GPU data center — is also a boundary that a malicious tenant can cross. Because the attack rides entirely on legitimate GPU scheduling APIs, existing cloud workload monitoring, which watches for anomalous API calls, unusual network egress, or policy violations, would have no obvious signal to flag; the malicious workload looks, from the scheduler's perspective, like an unusually well-synchronized but otherwise unremarkable batch job.

Background

Prior industry research indicates that modern AI data centers are already operating close to the limits of what local power infrastructure was designed to tolerate [1]. Research from Microsoft, Nvidia, and OpenAI documented that large, near-instantaneous power swings during GPU cluster synchronization — the moments when tens of thousands of accelerators transition together between compute-intensive and idle phases during distributed training — can stress power delivery hardware to the point of physical damage, and Meta's published account of training Llama 3 separately noted that simultaneous power draw across its GPU fleet produced instantaneous fluctuations on the order of tens of megawatts [1]. Until now, this phenomenon was treated primarily as an unintentional engineering hazard: a byproduct of synchronized training jobs that data center operators needed to buffer against with batteries, capacitor banks, and careful power provisioning, not as something an adversary could deliberately weaponize.

Bit2Watt reframes that hazard as an attack surface. The paper, accepted at the IACR's Conference on Cryptographic Hardware and Embedded Systems (CHES) 2026 and posted as a preprint in July 2026, demonstrates that a cloud tenant with no special privileges beyond a normal GPU rental — the kind of access anyone can purchase from a public cloud marketplace — can deliberately craft workloads whose power-consumption pattern is optimized to destabilize the electrical infrastructure serving the facility, rather than simply being a side effect of legitimate computation [1][2][3]. The researchers validated the underlying power-modulation capability on commercially available accelerators, including Nvidia's V100, A100, and RTX 4090 GPUs, showing that carefully scheduled compute-idle transitions can produce power modulation at frequencies exceeding 6,000 hertz — several orders of magnitude faster than the few-hertz cycling seen in conventional grid loads such as air conditioning compressors [1][3].

The researchers describe two complementary attack techniques for producing this modulation. Synthetic Workload Modulation Attack (SWMA) toggles CUDA kernels between compute-intensive and near-idle states on a controlled schedule, which is the more direct and higher-frequency of the two methods. LLM Training Modulation Attack (LTMA) instead embeds the same modulation logic inside the structure of an actual deep-learning training job, adjusting batch sizes and hyperparameters so that the resulting power oscillation is lower in frequency but higher in amplitude, and — critically — resembles the power-draw noise that training jobs already produce for entirely benign reasons [3]. LTMA's design goal is explicitly evasive: an attacker does not need to run an obviously anomalous job to achieve grid impact, only a training workload tuned to look statistically ordinary while still driving harmful modulation.

Attack component	Mechanism	Effect
SWMA (Synthetic Workload Modulation	CUDA kernels cycle between compute-intensive and idle states on	High-frequency power modulation (>6 kHz), directly

Attack component	Mechanism	Effect
Attack)	a controlled schedule	detectable if monitored for
LTMA (LLM Training Modulation Attack)	Batch size and hyperparameter manipulation embedded in real training jobs	Lower-frequency, higher-amplitude modulation disguised as normal training noise
Watt2Bit (side channel)	Frequency-shift keying encoded in power draw	Covert data exfiltration; researchers recovered a 50-bit test sequence
Watt2Bit (feedback DoS)	Grid disturbance stresses UPS, PDUs, and protective relays, which shed compute load	Denial of service across the affected GPU cluster

Table 1: Bit2Watt attack components and their downstream effects, as described in the source research [1][2][3].

Security Analysis

Why This Bypasses Existing Controls

A key reason Bit2Watt is difficult to defend against is that it does not require the attacker to do anything a cloud provider's acceptable-use policy would flag. A tenant running SWMA or LTMA workloads is, from the perspective of a GPU scheduler, hypervisor, or cloud security monitoring stack, indistinguishable from any other customer running a compute-intensive or training job — because it is one. There is no unauthorized API call, no privilege escalation, no anomalous network traffic, and no data exfiltration signature in the conventional sense for SWMA. The attack instead operates entirely in a layer that cloud-native security tooling was never built to observe: the physical power draw of the underlying hardware and its propagation through the facility's electrical distribution system.

This matters because it inverts what has typically been assumed in cloud threat modeling — namely, that multi-tenant isolation failures manifest as data leakage or resource contention that shows up in logs, metrics, or billing anomalies. Bit2Watt demonstrates an isolation failure that manifests instead as physical stress on shared infrastructure that sits entirely outside the software stack — power distribution units, uninterruptible power supplies, and, where the facility draws on photovoltaic or battery-based

distributed energy resources, the grid-facing inverters and control loops that manage those resources. This kind of equipment is not typically instrumented by security teams in current practice, based on the researchers' framing of the gap between OT/facilities operations and IT security monitoring; closing that gap is itself one of the paper's central recommendations.

The Scale Problem: Aggregation Across Tenants

The paper's central technical contribution, in CSA's assessment, is not that a single GPU can modulate power — that has been documented as an engineering hazard for years — but that an attacker can achieve grid-destabilizing effects by synchronizing many independently rented GPU instances, potentially across multiple tenancies or even multiple providers sharing local grid infrastructure. The researchers note that the same market and technical conditions that make cloud GPU capacity convenient for legitimate customers — inexpensive, on-demand rental, topology-aware schedulers that tend to co-locate workloads for latency and cost reasons, and cloud time-synchronization services that keep distributed jobs aligned to sub-millisecond precision — also make it easier for an attacker to achieve the spatial and temporal coordination Bit2Watt requires [3]. In effect, the infrastructure features that cloud providers built to make distributed AI workloads efficient are the same features that let a coordinated attacker aggregate many small, individually unremarkable GPU rentals into a single large-scale power-modulation event.

Under the paper's worst-case scenario — 1,000 synchronized GPUs targeting a 1-megawatt local grid segment with roughly 90 percent penetration of distributed energy resources such as rooftop solar and battery storage — the resulting current total harmonic distortion reached 46.8 percent [1][2], which the researchers note is enough to waste nearly half of the delivered electrical current as unusable distortion and to generate roughly 20 percent excess heat in affected equipment [1]. The same scenario drove the segment's damping ratio to -0.27, meaning the electrical system's natural oscillations were no longer decaying over time but instead growing — a hallmark of the kind of resonant instability that, left unchecked, leads to equipment trips rather than a return to steady state [1][2].

These worst-case parameters are useful for establishing an upper bound on the vulnerability's severity, but they also mark the outer edge of what is realistic for most GPU-dense facilities today. Ninety percent distributed-energy-resource penetration on a single 1-megawatt grid segment is well above what typical hyperscale data centers see in practice, since most large cloud campuses draw from dedicated utility feeds or larger grid segments with substantially lower renewable penetration than the paper's worst case assumes. Achieving the necessary synchronization would also require an attacker to acquire and coordinate roughly 1,000 independent GPU rental instances — potentially spanning multiple tenancies or providers — without the acquisition pattern itself triggering billing, quota, or scheduling anomalies that a cloud provider might notice independent of any power-quality monitoring. Neither

constraint eliminates the underlying vulnerability, which the researchers demonstrate is real at the physical layer regardless of how difficult a full-scale exploitation campaign would be to mount; they do, however, mean that the 46.8 percent THD and -0.27 damping ratio figures describe a worst-case ceiling rather than a typical or easily reachable outcome, and organizations assessing their exposure should weigh the likelihood of encountering DER penetration and tenant-coordination conditions anywhere near this severe.

Cascading Failure and the Watt2Bit Feedback Loop

Once local protective relays and circuit protections detect the resulting distortion or instability, they are designed to trip and shed load, which is the correct and expected engineering response to protect equipment from damage. The paper's concern is what happens next at scale: if an attacker triggers this protective response simultaneously across enough distributed infrastructure, the researchers model that cascading protective actions could produce blackouts affecting more than 80 percent of a large-scale power system, a scenario driven not by a single point of failure but by the interaction of many independently correct local safety responses to a coordinated stimulus [1].

This same feedback dynamic underlies what the researchers call Watt2Bit: the electrical and thermal stress that Bit2Watt induces does not stay confined to the grid side of the interaction. When protective equipment inside the data center — UPS units, power distribution units, and switching power supplies — responds to the disturbance by shedding computing load, it produces a denial-of-service effect on the very GPU cluster whose workload caused the disturbance, and potentially on co-located tenants sharing the same power infrastructure who had nothing to do with the attack [3]. Separately, the researchers demonstrated that the same power-modulation channel used offensively in Bit2Watt can be used as a covert side channel: by encoding data using frequency-shift keying within the power draw pattern, they successfully recovered a 50-bit test sequence, indicating that a compromised or malicious workload inside a data center could, at least in principle, use power-line fluctuations as a low-bandwidth exfiltration channel that falls outside what conventional network-based data-loss-prevention controls monitor.

Recommendations

Immediate Actions

Cloud providers and colocation operators running GPU-dense, multi-tenant facilities should engage their facilities and electrical engineering teams to determine whether current power-quality and protective-relay monitoring has any visibility into workload-correlated power modulation, since this is a new instrumentation gap rather than a known one. Security teams should not assume that facilities engineering already treats this as a security-relevant signal; the paper's central point is that these two disciplines have not historically needed to talk to each other, and Bit2Watt is precisely the kind of finding that requires them to start. Organizations operating high-density GPU capacity behind distributed energy resources — solar arrays, battery storage, or microgrids — should specifically assess their exposure, since the paper's worst-case modeling assumed high DER penetration as a contributing factor to instability.

Short-Term Mitigations

Cloud providers should evaluate whether GPU scheduling telemetry can be correlated with facility-level power-quality metrics to detect synchronized, high-frequency power modulation across tenant workloads, treating sustained modulation frequencies well above normal training-job cycling (the researchers' SWMA benchmark exceeded 6 kHz, far above typical workload transition rates) as an anomaly worth investigating even when no other security control is triggered [1][3]. Facilities with significant distributed energy resource penetration should review local energy buffering capacity — batteries and capacitor banks sized to absorb demand spikes — since the researchers identify buffering as one of the more immediately deployable mitigations against both accidental and adversarial power modulation. Providers should also examine whether topology-aware scheduling, which co-locates workloads for performance reasons, could be adjusted to avoid concentrating unusually large numbers of independent tenants' GPUs behind the same local power segment without a countervailing security review.

Strategic Considerations

Bit2Watt is a concrete instance of a broader pattern the AI infrastructure buildout is producing: in CSA's assessment, the physical scale-up of GPU capacity is outpacing the threat models applied to it, and cyber-physical interactions that were previously treated as pure engineering-reliability problems now need adversarial threat modeling applied to them. Cloud providers should extend their tenant-isolation

threat models beyond compute, network, and storage boundaries to explicitly include shared electrical infrastructure, and should treat power-quality monitoring as a security telemetry source rather than solely a facilities-operations concern. Over the longer term, this points toward a need for cross-disciplinary incident response capability that spans security operations and electrical/facilities engineering, since a Bit2Watt-style event would otherwise be diagnosed as a power-quality incident with no obvious adversary, delaying recognition that it was a deliberate attack.

CSA Resource Alignment

Bit2Watt sits at the intersection of two areas where the Cloud Security Alliance has already published guidance: the security of operational technology and critical-infrastructure-adjacent systems, and the control objectives that govern multi-tenant cloud and AI infrastructure.

CSA's Zero Trust Guidance for Critical Infrastructure is the most directly applicable prior work, even though it was written before Bit2Watt was disclosed [4]. The guidance's core argument — that Operational Technology and Industrial Control System environments require security models built around continuous verification of the "protect surface" rather than a hardened perimeter, precisely because OT/IT convergence keeps expanding the boundary between digital and physical systems — describes exactly the boundary Bit2Watt exploits [4]. A GPU data center's power distribution and DER control loops are, functionally, OT systems sitting immediately downstream of an IT-managed compute layer, and the guidance's five-step protect-surface methodology (define the protect surface, map operational flows, build the architecture, create policy, monitor and maintain) offers a starting framework for cloud and colocation operators who need to bring facilities engineering into a security review process it was previously excluded from [4].

CSA's more recent "Zero Trust Guidance for Achieving Operational Resilience", published in April 2026, extends this line of guidance to the specific failure mode Bit2Watt describes: cascading disruption that propagates through shared infrastructure and third-party dependencies rather than through a single point of compromise [7]. Where the 2024 critical-infrastructure guidance establishes the protect-surface methodology for defining and monitoring OT/IT boundaries, the operational-resilience guidance speaks more directly to Bit2Watt's central finding — that a coordinated stimulus can trigger a wave of independently correct local protective responses that combine into a large-scale outage. Cloud providers assessing exposure to Bit2Watt-style scenarios should treat the two publications as complementary: one for defining the protect surface around shared electrical infrastructure, the other for planning how to contain and recover from a cascading event once that protect surface has already been breached.

CSA's AI Controls Matrix (AICM v1.1), the 18-domain framework that builds on and extends the Cloud Controls Matrix for AI-specific control objectives, is the relevant control baseline for the tenant-isolation and infrastructure-resilience dimensions of this finding [5]. AICM's coverage of infrastructure and virtualization security already requires cloud providers to document and test isolation boundaries between tenants; Bit2Watt is evidence that those isolation boundaries must be defined to include shared physical power infrastructure, not only compute, storage, and network resources, when providers assess and attest to their control environment [5]. Providers evaluating their AICM control implementation for GPU-dense environments should treat this research as a prompt to explicitly scope power-domain isolation into that assessment rather than assuming it is implicitly covered by existing infrastructure controls.

Finally, CSA's "Securing the Cloud – Taking Back the Attacker's Mindset" briefing, which examines AI-powered cloud threats through an offensive-security lens, reinforces the broader methodological point this research illustrates [6]: effective validation of cloud security postures increasingly requires researchers and defenders to think adversarially about interactions that fall outside conventional software attack surfaces, including, as Bit2Watt shows, the electrical and mechanical systems that cloud software ultimately depends on.

References

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

[2] Zhouhao Ji, Kaikai Pan, Wenyuan Xu. "Bit2Watt: A Cyber-Physical Vulnerability Exploiting GPU Workloads Across Power and Computing Infrastructures." arXiv preprint / IACR Transactions on Cryptographic Hardware and Embedded Systems (CHES 2026), July 2026.

[3] GBHackers. "Bit2Watt Attack Turns AI Data Centers Into Cyber-Physical Threats to Local Power Grids." GBHackers, July 2026.

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

[5] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, June 22, 2026.

[6] Cloud Security Alliance. "Securing the Cloud – Taking Back the Attacker's Mindset." Cloud Security Alliance, 2024.

[7] Cloud Security Alliance. "Zero Trust Guidance for Achieving Operational Resilience." Cloud Security Alliance, April 2026.