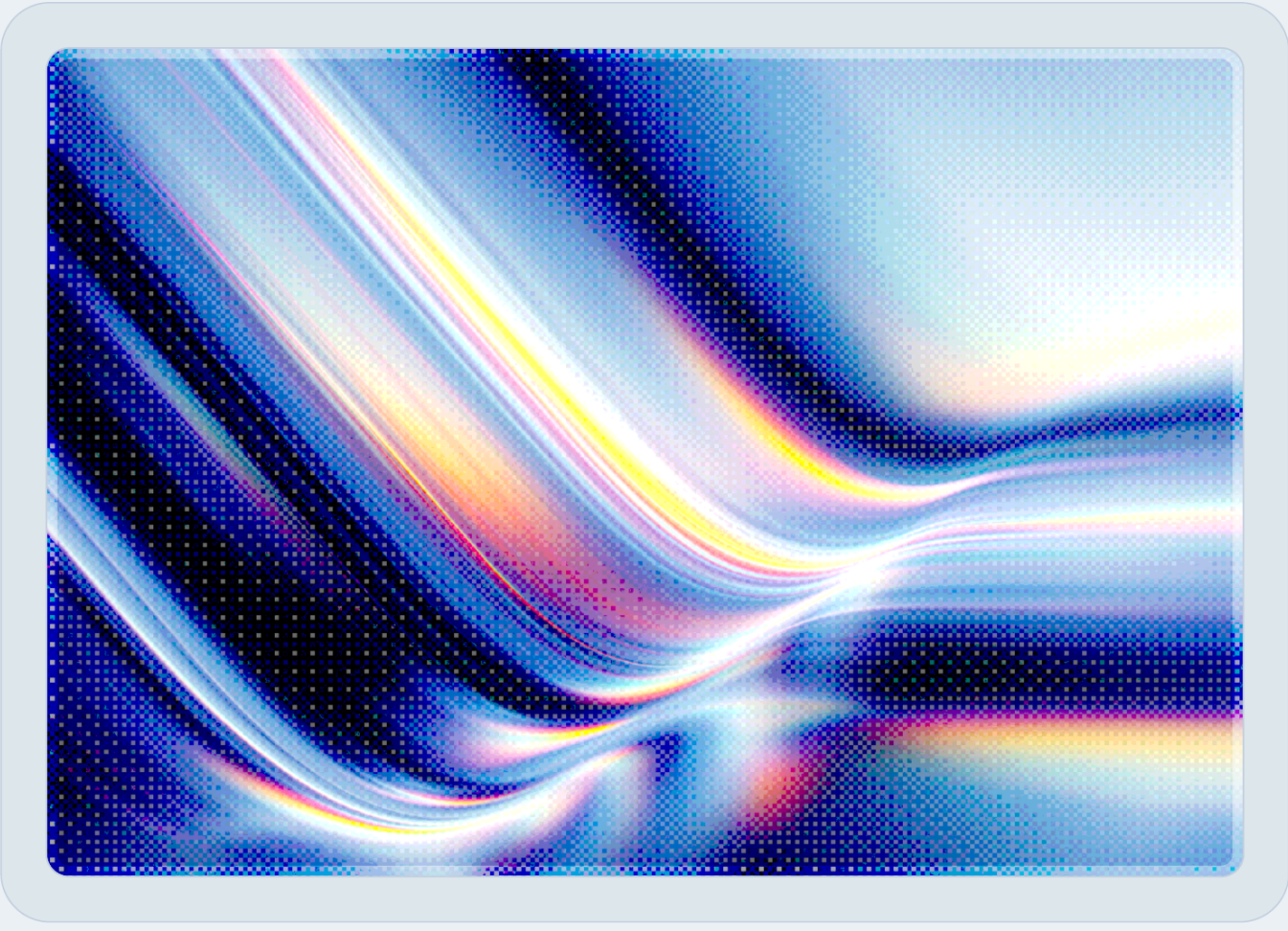
CSAI Foundation | Cloud Security Alliance

Bit2Watt: GPU Workloads as a Power Grid Attack Vector

When Legitimate Cloud Tenants Can Destabilize Electrical Infrastructure

2026-08-01

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Researchers from Zhejiang University have demonstrated Bit2Watt, a cyber-physical attack in which a cloud tenant uses only its own legitimately provisioned GPU access to induce power oscillations capable of destabilizing the electrical grid supplying a data center [1][2]. The attack requires no privilege escalation, no malware, and no exploitation of a software vulnerability in the conventional sense; it manipulates GPU compute intensity itself, which the researchers show is sufficient to generate current oscillations of 1.5 to 6 kHz, far above the frequencies utility protection systems and grid-connected inverters are designed to tolerate [2][3]. In simulation, 1,000 synchronized GPUs operating against a 1-megawatt local grid with 90 percent penetration of inverter-based distributed energy resources drove total harmonic distortion to 46.8 percent, more than three times the 13 percent threshold set by the IEC 61000-3-12 guideline, and pushed the grid's damping ratio from a stable 0.46 to an unstable -0.27 [2]. When the researchers extended the same disturbance model to a 9,241-bus simulation resembling a European transmission network, a localized perturbation equal to just 2 percent of system load cascaded through 13 stages of protective tripping and shed approximately 81 percent of total load [1][2]. The researchers also describe a secondary "Watt2Bit" effect, in which the harmonic stress the attack induces triggers thermal and overcurrent protections that shut GPU servers down, converting a power-quality attack into a denial-of-service condition, and they show the same power channel can be used covertly to exfiltrate data via electromagnetic emissions at better than 99 percent accuracy [1][2]. No vendor or cloud provider vulnerability disclosure preceded publication, and the authors state they never tested the technique against a production system; the research was, however, peer-reviewed and accepted to CHES 2026, the IACR's flagship hardware-security conference, and it builds directly on power-swing risks that Microsoft, OpenAI, and NVIDIA jointly documented in 2025 and that NERC has separately investigated following a real 2024 grid disturbance tied to data-center load behavior [4][5].

Background

Bit2Watt, formally titled "Bit2Watt: A Cyber-Physical Vulnerability Exploiting GPU Workloads Across Power and Computing Infrastructures," was authored by Zhouhao Ji, Kaikai Pan, and Wenyuan Xu of Zhejiang University and accepted to CHES 2026, the IACR's flagship hardware-security conference [1][2]. The paper's premise inverts the usual cloud-security threat model: rather than asking how an attacker might break out of a virtual machine or escalate privileges, it asks whether the computational

work a tenant is already authorized to run can itself be weaponized against physical infrastructure the tenant never touches directly. The answer hinges on a property of modern GPUs that has nothing to do with software security in the traditional sense. GPUs behave as constant-power loads, meaning their internal voltage regulators actively hold core voltage steady as workload intensity changes, forcing any change in computational demand to appear immediately as a change in current draw. A tenant who can control the timing of that computational demand can therefore control the timing and shape of the power draw itself, entirely within the bounds of a standard cloud compute agreement.

The researchers built two distinct methods for exploiting this property. The Synthetic Workload Modulation Attack, or SWMA, uses a custom CUDA kernel that toggles between a high-intensity tensor-operation mode and a near-idle mode, producing amplitude swings of up to 248 watts on a consumer-grade RTX 4090 and comparable effects on data-center-class A100 and Tesla V100 accelerators, with modulation frequencies tunable between 1.5 and 6 kHz [1][2]. The second method, the LLM Training Modulation Attack or LTMA, achieves a similar effect by adjusting ordinary training hyperparameters and inserting auxiliary operations into a real large language model training job, generating 1.2 to 3 kHz oscillations with amplitudes up to 355 watts that blend into the normal noise of a legitimate workload and are correspondingly harder to distinguish from benign training variance [1][2]. Because neither technique requires anything beyond the compute access a tenant already has, existing cloud isolation controls, which are built to prevent one tenant from touching another tenant's data or compute, do not address this class of risk at all; the attack operates entirely within the attacker's own allocated resources and produces its effect through the shared electrical infrastructure underneath.

This is not a hypothetical extension of unrelated research. In August 2025, Microsoft, OpenAI, and NVIDIA jointly published findings warning that large-scale AI training jobs already produce large, synchronized power swings as thousands of GPUs collectively transition between compute-heavy and communication-heavy phases, and that if the frequency spectrum of those swings aligns with a utility's critical frequencies, the result can be physical damage to grid infrastructure [4]. That research treated the phenomenon as an unintentional byproduct of how large training clusters behave and proposed engineering mitigations across software, hardware, and data-center infrastructure. Bit2Watt's contribution is to show that the same physical coupling can be deliberately engineered and precisely controlled by an adversarial tenant rather than arising incidentally from normal training behavior. A separate real-world event illustrates that utilities' models of aggregate data-center load behavior have gaps generally, even though it does not involve the same mechanism Bit2Watt exploits: on July 10, 2024, a 230 kV transmission line fault in the Eastern Interconnection triggered a cascading sequence that saw roughly 1,500 megawatts of data-center load disconnect from the grid nearly simultaneously as facilities switched to backup power via their own voltage-ride-through controls, an event NERC's Reliability and Security Technical Committee later characterized as demonstrating meaningful gaps in how utilities model and monitor large computational loads [5]. That disturbance stemmed from voltage-sensitive

protective tripping during a transmission fault rather than deliberately engineered harmonic resonance, so it should be read as evidence that data-center load behavior can already destabilize grid operations at scale through an unrelated mechanism, not as confirmation that Bit2Watt's specific attack has occurred.

Security Analysis

The mechanism that makes Bit2Watt effective is not a single flaw but an interaction among three separate properties of modern grid-connected power systems. The first is the negative incremental resistance behavior of GPU power supplies as constant-power loads: when voltage sags, the load's internal regulation responds by drawing more current to hold power constant, which further depresses voltage and amplifies the original disturbance through positive feedback rather than damping it [2]. The second is the control dynamics of inverter-based distributed energy resources such as rooftop solar and battery storage, which lack the physical rotational inertia of traditional generators and instead rely on layered digital control loops; the researchers show that GPU-driven modulation in the kilohertz range can excite resonant peaks in these inverters' output impedance that fall outside their control bandwidth, amplifying rather than absorbing the disturbance. The third is the erosion of grid stability margins that comes with high penetration of these same distributed resources: the researchers found that raising DER penetration from levels typical of legacy grids to 90 percent reduced the system's phase stability margin from roughly 60 degrees to 10 degrees [2]. This suggests, though the researchers tested only a specific simulated topology, that grids trending toward heavier renewable and battery integration may face growing structural exposure to this class of high-frequency manipulation.

Detection poses a distinct problem from the attack mechanism itself, because the telemetry infrastructure that both cloud operators and utilities currently rely on cannot see the disturbance as it happens. Rack-level power distribution unit counters typically sample once per second, NVIDIA's NVML telemetry interface tops out around 450 Hz, and even faster interfaces such as RAPL or server baseboard management controllers cap out near 1 kHz, all of which sit below the 1.2 to 6 kHz range where Bit2Watt's modulation lives [1][2]. The researchers' own attempt at anomaly detection using standard cloud telemetry and logistic regression performed poorly, at roughly 60 to 75 percent accuracy as shown in a figure in the paper rather than stated as an exact result in the text, while dedicated electromagnetic-interference sensing performed substantially better, an approach not deployed in production cloud environments today. The LTMA variant compounds this problem further: because it operates by adjusting parameters inside an otherwise ordinary model-training job rather than running a distinguishable synthetic kernel, it leaves no signature that current cloud workload monitoring is built to flag, and the researchers found it consistently harder to detect than the synthetic SWMA method even with dedicated instrumentation.

One of the most consequential findings is the scale-out behavior in transmission-level simulation. A localized disturbance representing just 2 percent of system load, injected into a 9,241-bus network modeled on European transmission topology, propagated through a cascading sequence of protective relay operations across 13 stages and ultimately disconnected 1,238 transmission lines, shedding approximately 81 percent of total load in the simulation [2]. This scale-out result is the basis for treating Bit2Watt as a critical-infrastructure risk rather than only a facility-level power-quality issue, because it demonstrates that a disturbance an attacker can generate using ordinary cloud compute resources at a single facility does not necessarily stay contained to that facility. The researchers are careful to note that achieving the perfect lockstep synchronization their worst-case simulations assume across genuinely independent, geographically distributed production GPUs remains an open engineering problem for an attacker, and that introducing realistic timing jitter of even 100 microseconds reduced the aggregate disturbance amplitude by roughly 20 percent. That caveat matters for calibrating urgency, but it does not eliminate the risk category; it narrows the near-term threat to well-resourced, technically sophisticated adversaries capable of coordinating large numbers of GPU instances with tight timing control, a description that could eventually fit nation-state actors or organized criminal groups, though no evidence yet indicates that existing GPU-fleet operators, including those running cryptomining and LLM-jacking operations on stolen or leased compute, possess the sub-100-microsecond timing coordination the researchers found necessary [8].

Bit2Watt also surfaces a structural governance gap that predates this specific research and is unlikely to be closed by any single technical patch. As the paper's authors put it, the compute side of a data center and the power grid side are run by different companies, monitored with different tools, and neither side's monitoring is built to watch the other. Cloud providers manage workload orchestration and tenant isolation; utilities and grid operators manage power quality and stability; and the interface between the two, the point where computational demand converts into electrical load, currently has no owner responsible for watching it as a security-relevant boundary at all.

Recommendations

Immediate Actions

Cloud and data-center operators should inventory which of their existing telemetry systems, if any, can observe power draw at sub-kilohertz resolution at the rack or facility level, since the Bit2Watt research demonstrates that standard PDU and NVML sampling rates are structurally blind to the attack's operative frequency range [1][2]. Where facility power monitoring already exists for capacity-planning or billing purposes, security and facilities teams should jointly assess whether that same data stream could

be repurposed, even coarsely, to flag statistically anomalous oscillation patterns in aggregate tenant power draw, recognizing that this is a stopgap rather than a complete detection capability given the sampling-rate gap the researchers identified.

Short-Term Mitigations

Organizations operating GPU fleets at scale, whether as cloud providers or as large enterprise AI training operators, should evaluate workload-scheduling and job-orchestration policies for their ability to detect and rate-limit unusual duty-cycle patterns, such as workloads that alternate between full utilization and near-idle states at a frequency inconsistent with normal training or inference behavior. Facility operators with significant distributed energy resource penetration should assess whether existing battery storage and harmonic filtering capacity is sized to absorb the kind of fast, high-frequency power transients the research describes, rather than only the slower transients associated with conventional load-shedding events. Utilities and grid operators serving large data-center campuses should treat this research as reinforcing, not superseding, the monitoring and communication gaps that NERC's Large Loads Working Group (formerly the Large Loads Task Force) has already identified following the July 2024 Northern Virginia disturbance described above, and should prioritize visibility into aggregate computational load behavior at the interconnection point rather than relying solely on the data center operator's own internal telemetry [5].

Strategic Considerations

The deeper implication of Bit2Watt is that the security perimeter around AI infrastructure can no longer be understood as ending at the rack or the facility wall. Organizations building AI Controls Matrix-aligned governance programs for GPU-intensive workloads should extend risk ownership to explicitly include the electrical interface between compute and grid, rather than treating power delivery as purely a facilities and engineering concern outside the scope of cybersecurity risk management. Because the attack requires no compromise of confidentiality, integrity, or conventional availability controls to execute, the standard set of cloud security controls, isolation, access management, and vulnerability patching, will not by itself close this gap; closing it requires coordination between cloud security teams, data-center facilities engineering, and the utilities that supply power, a cross-functional relationship that few organizations currently maintain on a security footing. Given that the underlying physical dynamics were independently documented by Microsoft, OpenAI, and NVIDIA before this research weaponized them, organizations should expect continued research attention to this intersection and should treat cyber-physical grid interaction as an emerging category within their AI infrastructure risk register rather than a one-off academic curiosity.

CSA Resource Alignment

Bit2Watt sits squarely within the scope of CSA's Zero Trust Guidance for Critical Infrastructure, which applies Zero Trust principles to operational technology and industrial control system environments and explicitly addresses the architectural differences between IT and OT/ICS systems that make conventional security controls insufficient on their own [6]. That guidance's five-step implementation model, defining the protect surface, mapping operational flows, building a Zero Trust architecture, creating policy, and monitoring and maintaining it, maps directly onto the gap Bit2Watt exposes: the "protect surface" for a GPU-intensive data center has historically stopped at the compute and network layer, but this research shows that the electrical interface to the grid is itself an operational flow that requires equivalent monitoring and policy attention. Organizations applying that guidance to AI data-center facilities should extend their operational-flow mapping to include power draw telemetry as a first-class signal, not an afterthought owned solely by facilities engineering.

The governance dimension of this risk also connects to CSA's AI Controls Matrix (AICM) v1.1, whose control objectives call for monitoring and managing risks that arise from AI workload behavior even when no conventional software vulnerability is present [7]. Because Bit2Watt achieves its effect without privilege escalation, malware, or a patchable defect, it is a useful case study for security teams building AICM-aligned control mappings: it demonstrates that a workload's resource-consumption pattern, not just its code or its access rights, can constitute a governable risk surface, and that vulnerability management programs scoped only to software weaknesses will not capture this category of exposure. The threat-actor capability referenced above draws on a distinct but related risk CSA has already documented in the wild: cloud accounts compromised for unauthorized AI compute use, most visibly in stolen-credential cryptomining and LLM-jacking operations, giving those same actor groups access to large GPU fleets even without the additional timing-synchronization capability Bit2Watt would require [8].

References

[1] Krishnan, R. "New Bit2Watt Attack Could Let Cloud Tenants Disrupt Power Grids Without an Exploit." The Hacker News, July 21, 2026.

[2] Ji, Z., Pan, K., and Xu, W. "Bit2Watt: A Cyber-Physical Vulnerability Exploiting GPU Workloads Across Power and Computing Infrastructures." arXiv, 2026.

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

[4] Microsoft Azure Compute Team. "Power stabilization for AI training datacenters." Microsoft Tech Community, August 2025.

[5] NERC Reliability and Security Technical Committee. "Characteristics and Risks of Emerging Large Loads." North American Electric Reliability Corporation, January 2025.

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

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

[8] Cloud Security Alliance AI Safety Initiative. "LLMjacking Evolves: Stolen AI Compute as Attack Infrastructure." Cloud Security Alliance, June 18, 2026.

