MARCH 13, 2026

AI, GPUs, and the systemic energy blindness of digital infrastructure

AI requires not only computing power, but also energy consumption, grid load and social pressure. Yet many data centers lack fundamental insight into the energy effectiveness of GPU environments. This article argues that AI energy blindness is not a temporary measurement problem, but a systemic risk that requires new metrics, standardization, and governance responsibility.

AI is not only changing workloads, but also responsibilities

The rise of AI applications has fundamentally changed data centers in a short period of time. Where traditional enterprise IT has been dominated by CPU-based workloads with relatively predictable power behaviors, modern data centers are increasingly driven by GPU/TPU Clusters[1] with extremely high power densities.

This shift is often presented as a scale problem: "more power, more cooling, larger connections". In reality, it is an **administrative and systemic problem**. AI infrastructure not only consumes more energy, but also largely withdraws from existing measurement and accountability structures.

The key question is not: *can the grid handle this?*
The key question is: *do we actually know what AI infrastructure does to energy — and why?*

GPU energy is fundamentally different from CPU energy

GPUs are not faster CPUs. They have a completely different energetic behavior:

- **Extremely high peak power** per node
- **Highly variable load** depending on workload phase
- **Non-linear relationship** between load, temperature and energy
- **Direct reliance on cooling** for performance

Yet, GPU environments are often monitored with tools and KPIs designed for CPU-driven data centers. This leads to a fundamental distortion of reality.

What is visible:

- Uptime
- throughput
- FLOPS

What remains largely invisible:

- Idle Power
- Thermal Throttling
- Energy loss due to waiting times
- energies for AI-output

As long as these aspects are not measured, they cannot be controlled.

Idle is not a fault, but a structural property

One of the biggest misconceptions surrounding AI infrastructure is that idle time equals inefficient management. In reality, idle is a **structural feature** of GPU environments.

Causes include:

- Batch-based training
- Waiting times between runs
- data-preprocessing buiten de GPU
- Conservative provisioning due to scarcity

However, GPUs consume significant amounts of energy even when idle. Without insight into idle consumption, an illusion of efficiency is created: the hardware "does nothing", but the energy meter continues to run.
This creates a fundamental mismatch between IT logic and energy reality.

Performance optimization masks energy loss

AI teams are judged on:

- Training time
- Latency
- Model quality

Energy is rarely an explicit control variable. This creates optimizations that are logical from a performance perspective, but energetically destructive. Examples:

- Higher clocks to meet deadlines
- Extra nodes "for security"
- Aggressive cooling to prevent throttling

Without an energy context, these choices are rationally rewarded, while the systemic impact remains out of the picture.

The Dieselgate parallel: not fraud, but measurement blindness

The comparison with Dieselgate is often emotionally charged. This is not about fraud, but about **measurement blindness**. Systems are optimized within the framework of what is visible and measurable. What falls outside that framework does not count.
In AI infrastructure, this leads to:

- Optimization for benchmarks
- Accountability without context
- Scaling up without efficiency gains

The risk is not only technical, but social: AI is seen as an uncontrollable energy guzzler, which evokes policy-related reflexes that can inhibit innovation.

AI energy as a public precondition

AI energy use touches:

- grid congestion
- Climate goals
- Geopolitical dependencies
- Public acceptance of digital growth

Therefore, AI energy cannot be treated exclusively as an internal optimization issue. Transparency, comparability and verifiability are public preconditions.

The **Save Energy Foundation** therefore approaches AI energy not as a competition parameter, but as a system issue. By developing and securing independent metrics for GPU and AI efficiency, a foundation is laid for policy, standardization and fair comparison.

From black box to controllable system

Without integrated data, AI remains a black box. By measuring IT performance, energy consumption and cooling in conjunction — as Zirrow does — it becomes visible:

- where energy is actually used
- where structural losses occur
- Which design choices cause inefficiencies

Only then can AI infrastructure be steered towards effectiveness per kilowatt hour consumed.

AI calls for energy sovereignty

AI innovation without energy transparency is not progress, but dependency. Anyone who takes AI seriously must consider energy insight to be just as fundamental as model architecture.