Deep Geeks

PRESENTED BY WEKA

The AI Energy Metrics Cheatsheet

What to measure, what to fix — straight from the teams at Nebius and WEKA building efficient AI infrastructure at scale.

SECTION 01

Core Energy & Efficiency Metrics

Goodput vs. Throughput

Throughput measures total system output — how busy your GPUs are. Goodput measures useful output — how much of that activity produces the tokens and completions your users actually need.

```
GOODPUT = USEFUL OUTPUT/ TOTAL SYSTEM OUTPUT
```

WHY IT MATTERS

A system at 95% GPU utilization can still have low goodput. This metric separates efficient infrastructure from expensive idling.

Tokens Per Watt (TPW)

Measures how much useful AI work — in tokens — is produced for every watt of power consumed. The defining efficiency benchmark for the inference era.

```
TPW = TOKENS GENERATED / WATTS CONSUMED
```

WHY IT MATTERS

Inefficient inference can consume a household's daily energy per session. Efficient memory tiering reduces that by 80%+.

Tokens Per GPU

How many tokens a single GPU produces in a given time window. Useful for comparing fleet productivity across configs and providers.

```
TOKENS/GPU = TOTAL TOKENS / ACTIVE GPUS
```

WHY IT MATTERS

Exposes over-provisioning. If you're scaling GPUs for memory (not compute), this number drops dramatically.

Power Usage Effectiveness (PUE)

Industry-standard metric for data center energy efficiency. Ratio of total facility energy to energy delivered to IT equipment.

```
PUE = TOTAL FACILITY ENERGY / IT EQUIPMENT ENERGY
```

WHY IT MATTERS

Global avg ~1.58. Best-in-class hits 1.1. But PUE alone doesn't tell you how efficiently energy becomes AI output.

Energy Per Token (Wh/token)

The inverse of TPW — how much energy in watt-hours is consumed to generate one token. Key for cost modeling and sustainability reporting.

```
ENERGY/TOKEN = WATT-HOURS / TOKENS GENERATED
```

WHY IT MATTERS

Optimized frontier inference: median 0.31 Wh per query. Combined interventions can deliver 8–20× reductions.

Energy Per Workload

Total energy for a complete AI task — training run, fine-tuning, or batch inference — including all supporting infrastructure overhead.

```
ENERGY/WORKLOAD = TOTAL ENERGY / COMPLETED WORKLOADS
```

WHY IT MATTERS

The metric for enterprise budgeting & sustainability reporting. Captures the full stack, not just compute.

> "You can have a very busy system that's not really producing a lot of useful output."
>
> **Val Bercovici**
> Chief AI Officer, WEKA

SECTION 02

Seven Tips for Building Efficient Infrastructure

01 Autoscaling for GPU Workloads

GPU autoscaling is fundamentally harder than CPU autoscaling. Model weights are orders of magnitude larger than prior binaries or artifacts, and KV cache warming can take significant time before inference is production-ready. Getting this right is non-trivial — but it's critical to profitability and gross margin. Over-provisioned idle GPUs drain power; under-provisioned clusters drop requests. The right software matches cluster size to real-time demand without sacrificing warm-up time.

02 Memory Tiering

Inference is a memory problem — not a storage problem. But storage competitors often try to confuse customers by equating storage-class tiers with memory-class tiers. Only memory-class tiers add real value at inference time by decoupling GPU memory from GPUs. This lets you provision memory where you need it without over-provisioning expensive, idle GPUs just for the memory they bring along. The efficiency gains are 75–90% — but only if you're using actual memory-class architecture, not relabeled storage.

03 Drive Endurance & Token Routing (DWPD)

Drive-writes per day (DWPD) is a critical-path element for inference storage. AI inference generates intense, sustained write patterns that destroy consumer-grade SSDs in months. Vendors who lack the maturity to understand high-level token routing and low-level token scheduling will fail here — with bricked SSDs at customer sites. AI builders (and soon enterprise IT) are focused on this because token costs will force domain-specific models, making storage endurance and write-path efficiency non negotiable.

04 Post-Training & Model Optimization

Tune model setup for the specific hardware hosting it: faster completion, fewer failures, less power draw. This matters more now than ever — token costs are forcing the industry toward domain-specific models, meaning more organizations are running post-training optimization more frequently across more model variants. AI builders and soon enterprise IT are focused here because efficient post-training directly determines inference economics. Model architecture choices define runtime efficiency — and with domain-specific proliferation, those choices multiply.

05 Cold Start, Warm Cache & Checkpointing

For production inference, cold start model loading and always-warm KV cache are critical — users can't wait minutes for a model to become responsive. For training, fast checkpoint and restart remain essential. Beware async checkpoint workarounds: they're a complexity trap that limits recovery point objectives (RPOs) to coarser-grained recovery than researchers actually need. The energy cost of restarting a failed training run from a stale checkpoint — re-doing hours of compute — dwarfs the cost of proper checkpoint infrastructure.

06 Closed-Loop Cooling

Traditional data center cooling relies on evaporative systems that consume millions of gallons of water annually — a growing liability as water scarcity becomes a regulatory and reputational risk. Closed-loop liquid cooling with dry coolers eliminates water intake entirely. Nexus designs servers to operate reliably at up to 45°C, which fundamentally changes the cooling equation: less delta-T required means less energy spent on heat rejection. The result is a PUE of 1.15 — nearly half the overhead of the 1.58 global average — with zero water consumption and no evaporative components to maintain or replace.

07 Heat Recovery

Traditional data center cooling relies on evaporative systems that consume millions of gallons of water annually — a growing liability as water scarcity becomes a regulatory and reputational risk. Closed-loop liquid cooling with dry coolers eliminates water intake entirely. Nexus designs servers to operate reliably at up to 45°C, which fundamentally changes the cooling equation: less delta-T required means less energy spent on heat rejection. The result is a PUE of 1.15 — nearly half the overhead of the 1.58 global average — with zero water consumption and no evaporative components to maintain or replace.

> "Software plays an important role in orchestrating all of that hardware and infrastructure...Idling is actually a killer to efficiency."
>
> **Dasha Mukhortova**
> Head of Sustainability, Nebius

EPISODE 01 · NOW STREAMING

Watch the Full Conversation

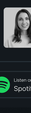
Deep Geeks

PRESENTED BY WEKA

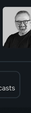
Can AI Innovation Survive Its Own Energy Appetite?

Dr. Serena Huang F100 Strategic Advisor

SPECIAL GUESTS:



Daria Mukhortova
Head of Sustainability, Nebius



Val Bercovici
Chief AI Officer, WEKA

Listen on Spotify

Listen on Apple Podcasts