

Wanted: A New Efficiency Metric for AI Data Centers

Power usage effectiveness, or PUE, is the go-to sustainability metric for data centers. But it says nothing about the power loss between the PoL converter and the AI processor itself.

[Power usage effectiveness](#) (PUE) has become the default way we discuss the overall efficiency of [data centers](#). Regulators cite it, sustainability reports lead with it, and operators compete to push it closer to the theoretical ideal of 1.0.

It's a useful number, but it's also an incomplete one. PUE measures everything that happens to power before it reaches the server — [cooling](#), [lighting](#), and [power distribution](#) — and stops the moment power crosses the server's threshold. What happens after that point, as electricity moves through the server itself and down to the processor, sits entirely outside the metric. A facility can report an excellent PUE while still losing a meaningful share of its power [inside the server](#) invisibly.

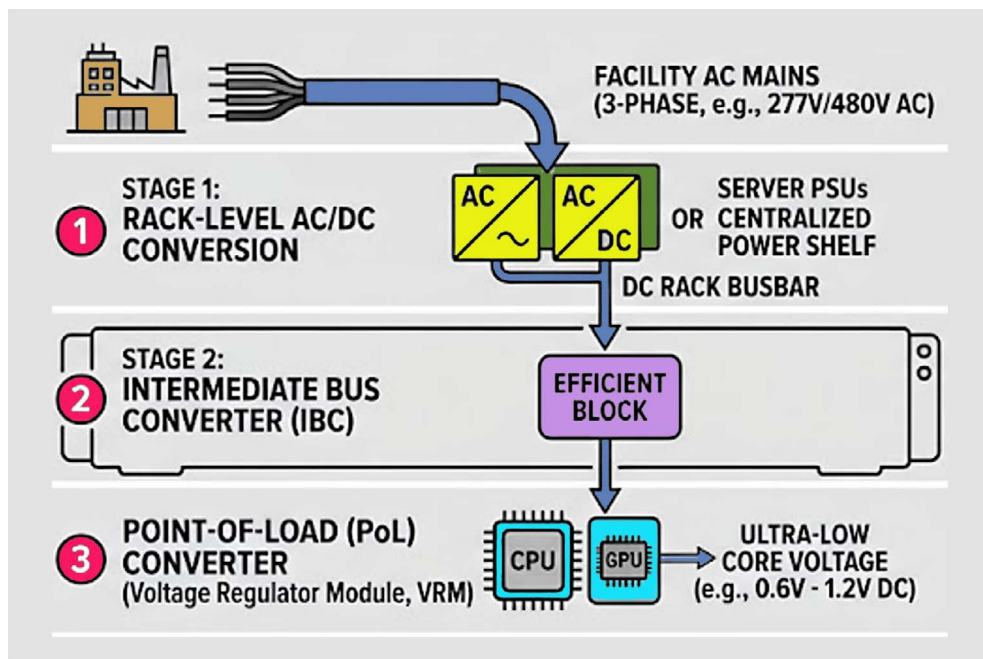
As AI workloads push power density higher than anything

the data center industry has designed around before, that blind spot matters. The next real gains in AI data center efficiency will not only come from better cooling or smarter facility design. They will also come from the last few inches of [the power delivery network \(PDN\)](#), inside the server itself.

What PUE Actually Measures, and Where It Stops

PUE, as a metric, was designed for a different era of computing. It compares a facility's total energy draw to the energy that ends up reaching IT equipment, capturing overhead like cooling, lighting, and power distribution outside the server.

That was the right lens to look through when server power



1. A simplified power delivery path from the grid to the AI processor, showing the AC-DC, DC-DC, and point-of-load conversion stages. (Credit: Lotus Microsystems)

draw was relatively modest and relatively uniform across a facility. It gave companies a clean, comparable number to track [facility-level waste](#), and it has worked well for years.

What PUE was never designed to do is look inside the server. It has no visibility into how efficiently power is converted once it arrives: from the AC feed, through [the server's power supply](#), down through the [intermediate voltage converter \(IBC\) stages](#), to the specific voltage a processor needs (*Fig. 1*).

That entire path is treated as a black box. A data center can have an industry-leading PUE and still have no visibility into how much power its servers are wasting as heat before it reaches a chip.

The Hidden Losses Inside the Server

Getting power from the wall to an AI processor isn't a single step. It's a chain of conversions, and every stage in that chain suffers some losses. Power typically comes in as AC and is converted to a high-voltage DC bus. Then it's stepped down through one or more [DC-DC conversion](#) stages. Finally, power is regulated by a [point of load \(POL\)](#) converter, which is the last stage before the processor itself, where voltage needs to be delivered precisely and hold steady under load (*Fig. 2*).

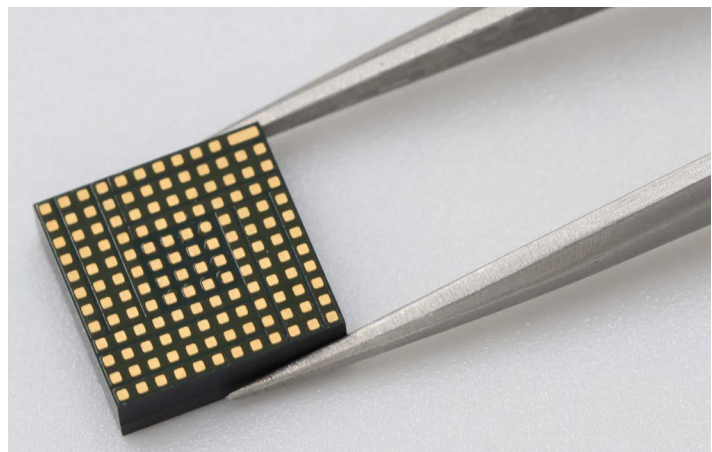
Each of those stages has its own efficiency curve, and each one leaks some power as heat rather than delivering it to compute. None of that is visible in a facility-level metric. It shows up as heat which then must be cooled, as power draw that doesn't correspond to any useful work, and as the cost of spending all of that power to spin [cooling fans](#) or pump [liquid coolant](#). Anyone monitoring PUE alone has no way to see any of it.

Why This Matters for AI Workloads Specifically

This isn't a new problem, but the urgency surrounding it is increasing. [AI accelerators](#) draw far more current, at far more aggressive and less predictable load transients, than the CPU-dominated racks that PUE was developed around. [Power electronics](#) need to respond faster and hold tighter tolerances, and do both at power levels that would have been unusual only a few years ago.

The physics don't change with scale, but the stakes do. [Five years ago](#), a typical data center rack drew roughly 5 to 8 kW. Today, AI facilities are being designed with power-per-rack specifications of 15 to 50 kW per rack, with [GPU-dense configurations](#) reaching more than 100 kW. A conversion inefficiency that was a rounding error at modest power draw represents real, measurable waste once it's multiplied across AI-era power levels and multiplied again across a fleet of racks.

The same percentage loss now means significantly more wasted energy and more heat that must be removed. It also



2. Lotus Microsystems' vStrata is a vertical power module intended for use with ultra-high-current AI accelerators. (Credit: Lotus Microsystems)

leads to higher electricity costs, simply because the baseline power draw is so much higher. PUE doesn't move to reflect any of this, because it was never measuring this part of the path in the first place.

Toward a More Complete Efficiency Measure

None of this is an argument against PUE. It still does what it was built to do. But PUE alone can no longer stand in for the full picture of AI data center efficiency. A more complete measure needs to extend past the server door and account for what happens at each conversion stage inside it. This includes the point-of-load regulation closest to the processor, where a large share of the remaining losses tend to concentrate.

That doesn't necessarily mean replacing PUE. It could mean pairing it with a conversion-stage efficiency measure or asking equipment vendors to report power delivery efficiency the way they already report other performance specs. What matters is that the data-center industry starts treating server-internal power delivery as something to be measured and compared, rather than an assumed constant.

What This Means for Data Center Operators

An operator relying on PUE alone is seeing part of the cost and emissions picture, not the whole one. The energy lost inside the server still shows up on the power bill, still generates heat that needs to be cooled, and still carries an emissions footprint. It's just not attributed to anything, because there's no metric currently assigned to catch it. At the scale AI infrastructure is being built out today, and against the grid demand that scale is already creating, that's a gap worth closing.

Operators and architects don't need to wait for an indus-

try-standard metric to start asking better questions. Ask vendors how power delivery efficiency is measured stage by stage, not just at the server's rated input. Look at point-of-load conversion specifically, since that's where AI-era current density concentrates the remaining losses. PUE will keep telling you how efficient your facility is. It's also time to ask about the efficiency of your servers, because increasingly, that's where the real losses are hiding.

With over 20 years of experience in the semiconductor industry, Hans Hasselby-Andersen currently serves as CEO at Lotus Microsystems, where he is advancing ultra-compact, high-efficiency power solutions for next-generation compute. Hans has built a diverse career as a founder, CEO, board member and chairman, and investor. He founded Merus Audio, a fabless semiconductor company, which he led through its acquisition by Infineon in 2018.

Throughout his career, Hans has held key management positions at technology leaders, including Texas Instruments, Infineon, Toccata Technology, and Nokia. He holds an M.Sc. in Electrical Engineering and an Executive MBA.

Captions: