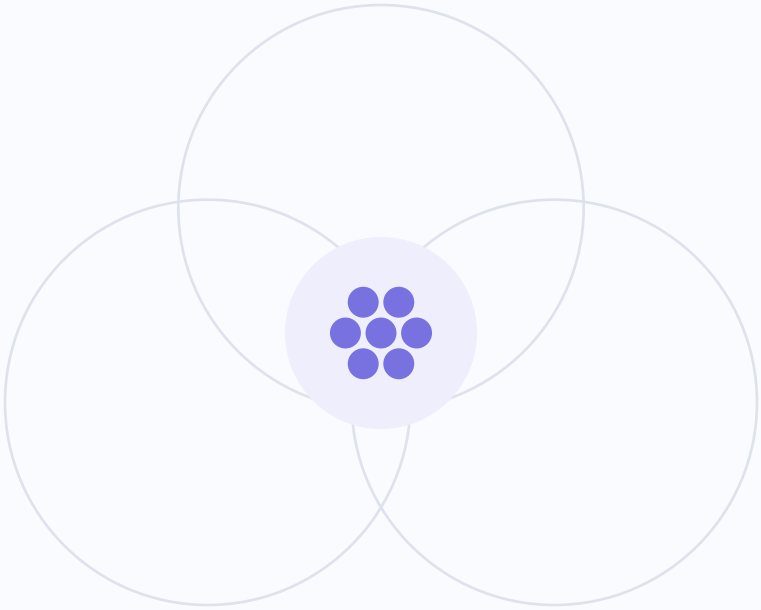
Quantx

RESEARCH PAPER / ENVIRONMENT 02

The footprint of distributed AI.

Energy, water, carbon and the value of reuse.



Can existing devices reduce the footprint of useful AI work? A framework for testing operating and lifecycle effects.

Analytical working paper. No measured Quantx environmental reduction.
Public methods edition. Not peer reviewed.

01 / READING GUIDE

Reuse is the opportunity. Displacement is the test.

A lower footprint depends on the useful work delivered, the operating resources added, and the infrastructure genuinely avoided.

Existing hardware can provide additional useful work

The strongest case uses devices people already own to replace some otherwise-required dedicated capacity, without inducing enough extra operation, wear, or purchases to outweigh the benefit.

Energy, carbon, water, and construction are separate accounts

A lower electricity bill does not prove fewer emissions. Air-cooled endpoints still have electricity-related and hardware-related water impacts. Moving a workload does not prove a building was canceled.

Efficient data centers are a demanding baseline

Google reports 0.24 Wh for a median Gemini Apps text prompt in a 2025 production study [1]. A Microsoft-led study reports 0.31 Wh under its frontier-scale deployment assumptions [2]. Neither figure is a measured baseline for the Quantx example.

Evidence status

Published context	This paper	Still required
External studies and provider disclosures.	Explicit equations and numerical sensitivities.	Matched-workload Quantx measurements and a defensible displacement case.

Reading path: boundaries p. 3; electricity p. 4; carbon and overhead p. 5; construction pp. 6-7; water p. 8; rebound and evaluation p. 9; sources p. 10.

02 / COMPARISON BOUNDARY

The same useful output. The whole affected system.

Use a fixed workload, acceptance threshold, completion requirement, and service period in both alternatives. Do not compare a small-model answer with a different-quality frontier-model answer as equivalent work.

Boundary	Count in the comparison
Existing operating facility	Device/server electricity, facility overhead, communications, supporting services, retry and fallback work.
Otherwise-required new capacity	Add attributable equipment, buildings, cooling and electrical systems, construction, transport and end of life.
Participating device fleet	Additional wall electricity, induced awake time and cooling, communications, model distribution, wear, early replacement and new purchases caused by participation.

Use a consequential question for an avoidance claim

Ask what changes because the mesh serves the work. Existing-device manufacturing is generally sunk in that question unless participation changes purchase or replacement behavior. New Quantx Go hardware, if purchased because of the service, is an induced burden and belongs in the account.

Attribution allocates an existing footprint to a service. Consequential analysis asks what actually happens differently. Both can be useful, but an allocated share of a building is not automatically a building avoided. Deferring construction also needs a stated time horizon.

The model reports accepted tasks, n, over a common period. A is attempt-equivalent effort per accepted result; z is the hosted fallback fraction. Each cost or energy term must have a declared, non-overlapping boundary. [4,6,7,9,10]

03 / OPERATIONAL ELECTRICITY

Measure the increment at the wall.

Power telemetry for one chip is not the full operating boundary. Include conversion losses and the actual state a device would otherwise have occupied.

$$E_{dev} = (1 / 3{,}600) \sum_i \int [P_i(t) - P_{0i}(t)]\, dt \quad (1)$$

P and the no-mesh counterfactual P_o are wall watts; t is seconds; E is Wh. Include time the service keeps a device awake when it would otherwise sleep.

$$e_M = (E_{dev} + E_{net} + E_{svc} + E_{dist} + E_{fb}) / n \quad (2)$$

Wh per accepted task includes devices, networking, shared services, model distribution, and fallback. Place validation, failed work, and cooling in the appropriate non-overlapping term.

$$e_M \approx A\, P\, t / 3{,}600 + e_o + z\, e_C \quad (3)$$

A planning approximation. P is added wall power; t attempt time; e_o non-worker overhead; e_C comparable hosted Wh per result. Savings require e_M < e_C at matched requirements.

Assumed case	Inputs	Calculated Wh/task
Low-power example	P=5 W; t=20 s; A=1.15; e_o=.025 Wh	.081
High-power example	P=150 W; t=30 s; A=1.50; e_o=.310 Wh	2.209

Both examples assume z=.10 and e_C=.24 Wh. The .24 Wh value is an illustrative anchor informed by [1], not a measured equal-workload baseline. The examples demonstrate sensitivity, not Quantx performance.

04 / CARBON AND OVERHEAD

Less energy can still mean more carbon.

Electricity quantity and electricity supply both matter. Keep geographic, temporal, and accounting boundaries consistent.

$$G_{op} = \sum_i E_i g_i \quad (4)$$

E is kWh and g is kg CO2e/kWh. Use factors appropriate to the question: historical average, marginal change, and contractual market-based accounting are not interchangeable. [5,6]

Illustrative electricity	Assumed grid factor	Emissions per 1 M tasks
.10 Wh/task	.40 kg CO2e/kWh	40 kg CO2e
.20 Wh/task	.10 kg CO2e/kWh	20 kg CO2e

In this illustration, halving electricity doubles emissions because the grid factor is four times larger. These factors are not assigned to Quantx or a provider.

Facility overhead has a ceiling

PUE is total facility energy divided by IT energy. At PUE 1.09 [3], overhead is 1 – 1/1.09 = 8.26% of facility electricity. Removing that overhead alone cannot yield a 50% saving against this specific efficient reference. Google’s published efficiency reporting provides context, not a universal data-center profile.

Household and idle-time effects still count

If air conditioning removes 100 Wh of extra device heat at a coefficient of performance of 3, it uses about 33 Wh more electricity. Useful-heating credits require an actual displaced heating source. Separately, 100,000 devices kept awake at 5 extra watts for 8 hours consume 4 MWh/day before useful-task allocation.

A fleet median energy value must not be multiplied into a total as though it were the fleet mean. Long tasks, reasoning, concurrency, and the mix of requests change aggregate demand. [1,2]

05 / LIFECYCLE MODEL

Count what is built. Count what is avoided.

An operating-energy disadvantage can be offset by genuine avoided equipment or construction. The benefit disappears if the avoided-capacity assumption is wrong.

$$\Delta G = O_C + f_a B - (O_M + I_M) \quad (5)$$

Positive ΔG is a benefit. O_C and O_M are central and mesh operating impact; B is the relevant new-capacity embodied impact; f_a is the share genuinely avoided; I_M is induced mesh lifecycle impact.

$$f_a B > (O_M - O_C) + I_M \quad (6)$$

The break-even condition. Use consistent kg or tonnes of CO2e and a common output and service period. No building-avoidance credit follows from moving one request.

Term	Illustrative value	Interpretation
O_C	500 t CO2e	Central operating impact.
B	100 t CO2e	Referenced otherwise-required new capacity.
O_M	520 t CO2e	Mesh operating impact.
I_M	5 t CO2e	Induced hardware and other mesh lifecycle impact.

Keep the total comparison scope fixed

The full central alternative is $O_C + B = 600$ tonnes. The mesh alternative is $O_M + I_M + (1 - f_a)B$: any referenced new capacity that is not avoided remains in the mesh-side comparison. Subtracting these totals gives Equation (5).

A fixed-scope comparison prevents a reduction percentage from changing simply because its denominator changes. Values are explanatory assumptions, not estimates of an actual facility. [6,7,10]

06 / AVOIDED-CAPACITY SENSITIVITY

Construction can change the conclusion.

The working illustration breaks even when 25% of the referenced capacity is genuinely avoided. Greater operating or induced burdens raise that threshold.

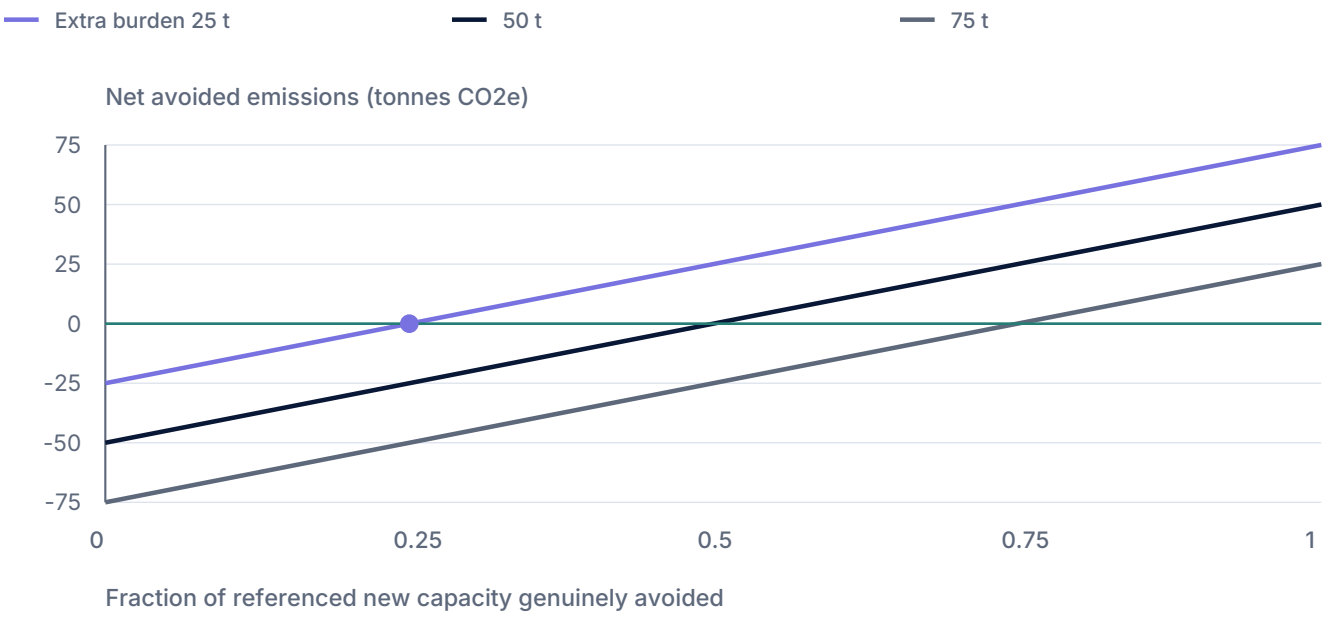


Figure 1. Analytical sensitivity from Equation (5), with B=100 tonnes. Each line varies the combined extra mesh operating and induced burden. The dot marks working-case break-even. No project observations are plotted.

Avoided share	Central total	Mesh total	Net outcome
0%	600 t	625 t	4.2% higher impact
25%	600 t	600 t	Break-even
100%	600 t	525 t	12.5% lower impact

All percentages use the same 600-tonne central reference and the same accepted output. The mesh total includes any of the 100 tonnes of referenced capacity not avoided. On an operating-plus-induced basis alone, 525 versus 500 tonnes is 5% higher; that is a different boundary.

What must be justified: a specific alternative capacity plan, the share actually displaced, its equipment and facility scope, utilization, and the time horizon. Reusing existing devices does not itself prove displacement.

07 / WATER

No cooling tower does not mean no water.

Include direct cooling, the water consumed in energy supply, and the manufacturing or construction induced by the chosen alternative.

$$W = W_{direct} + \sum_i E_i w_i + W_{induced} \quad (7)$$

W is liters consumed; E is kWh and w is the compatible water-consumption factor in L/kWh. Compare like boundaries and do not mix consumption with withdrawal.

Where sector water consumption occurred

IEA historical sector estimate, 2023 [4]



Approximate share	Boundary
About two-thirds	Primary energy supply and electricity generation.
About one-quarter	Direct cooling.
Remainder	Semiconductor and microchip manufacturing.

These are approximate historical sector shares, not a distribution for every facility or participating device. Moving work to an air-cooled endpoint can change direct cooling water while retaining electricity-related consumption and induced hardware impacts.

Compare the facility actually displaced

For a WUE defined per IT kWh, direct cooling water equals WUE × IT energy, not total facility electricity. Microsoft describes closed-loop designs intended to avoid ongoing cooling-water consumption, with energy tradeoffs [8]. A zero-evaporation design requires a different baseline from an evaporatively cooled site.

Report consumption separately from withdrawals, identify geography and watershed stress, and state factors and uncertainty. A sum of liters does not capture all local consequences. No Quantx water-reduction percentage is established here.

08 / DEMAND AND VALIDATION

Efficiency per task. Impact at scale.

An efficient task does not guarantee lower total impact when demand grows. Report both the unit result and the total system effect.

$$k\,r < 1 \tag{8}$$

r is per-accepted-task impact relative to the baseline; k is the task-volume multiplier. Total operational impact falls only when their product is below one, under the same boundary.

0.5×

per-task impact

Illustrative efficiency gain

3×

task volume

Illustrative demand growth

1.5×

total impact

50% higher overall

A testable environmental pilot

Fix the useful output, quality threshold, timing requirements, and service period. Measure full wall electricity, added awake time, failures, verification, model downloads, networking, shared services, fallback, and any household cooling. Use representative devices and days. [1,2,9]

For lifecycle claims, document the otherwise-required capacity plan and its credible displacement or deferral. Add purchases and accelerated replacements caused by participation. Use compatible carbon and water factors, separate attribution from marginal effects, and report low- and high-displacement sensitivities.

Conclusion

The environmental thesis is strongest when useful work on existing devices genuinely avoids additional dedicated capacity. The equations establish the conditions, not a measured Quantx reduction. Quality, power, grid supply, water stress, device lifetime, construction displacement, and new demand determine the result.

Public analytical edition. No new Quantx hardware measurements, environmental savings claim, or proprietary implementation method is presented. The companion systems paper addresses capacity and operator economics.

09 / SOURCES

Environmental evidence

Research, historical factors, and provider disclosures. Sources have different scopes; none is treated as a universal profile for Quantx or all data centers.

[1] C. Elsworth et al. Measuring the Environmental Impact of Delivering AI at Google Scale. Provider study, 2025.
Read source ↗

[2] F. Oviedo et al. Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling. Joule, 2026.
Read source ↗

[3] Google. Data center efficiency. PUE reporting and 2025 trailing-period context; checked 5 November 2026.
Read source ↗

[4] IEA. Energy and AI. 2025, Box 5.4 and Figure 5.27.
Read source ↗

[5] US EPA. eGRID Summary Data: historical regional electricity emissions.
Read source ↗

[6] Google Cloud. Carbon Footprint Reporting Methodology. Retrieved November 5, 2026.
Read source ↗

[7] Microsoft. Datacenter Cooling Lifecycle Study Overview. Research-originator account, May 2025.
Read source ↗

[8] Microsoft. Sustainable by Design: Next-Generation Datacenters Consume Zero Water for Cooling. December 2024.
Read source ↗

[9] S. Laskaridis et al. MELTing Point: Mobile Evaluation of Language Transformers. 2024.
Read source ↗

[10] Using life cycle assessment to drive innovation for sustainable cool clouds. Nature, 2025.
Read source ↗