

Heat-sink-constrained siting of distributed AI infrastructure: Energy, carbon and water impacts in Los Angeles

Jonathan Bily

Cloud Wash Laundry Corporation, 44370 Cella, Teruel, Spain

Corresponding author: Jonathan Bily, Cloud Wash Laundry Corporation, 44370 Cella, Teruel,
Spain | jonathan@cloudwashlaundry.com | ORCID: 0009-0004-3526-7271

Abstract

Growth in artificial-intelligence infrastructure is increasing electricity, cooling and siting pressures, while most data-centre heat-recovery studies begin with a fixed location and then search for a heat user. This study treats commercial heat demand as a siting and capacity constraint for distributed liquid-cooled AI modules. A deterministic hourly model places 6,160 B300-equivalent accelerators (11.17 MW rated IT) across a synthetic Los Angeles screening module of 120 representative car-wash and laundromat hosts and couples process-heat demand to liquid-loop capture, four-hour storage, photovoltaic-thermal rejection, optional adiabatic trim, dry rejection, representative-year weather, water use and operational greenhouse-gas accounting. If all host demand is temperature-compatible, 63.49 GWh yr⁻¹ of IT electricity yields 50.38 GWh yr⁻¹ of useful heat (79.4%), equivalent to 59.28 GWh yr⁻¹ of input to the modeled 85%-efficient gas heaters. When only 75%, 50% or 25% of demand is compatible, useful heat falls to 63.7%, 42.5% and 21.2% of IT electricity, and the same-supply emissions reduction at 43 gCO₂e kWh⁻¹ falls from 77.6% to 62.4%, 41.7% and 21.0%. Hollow-core fibre reduces maximum one-way propagation from 244 to 163 μs and enlarges the host catchment reachable within a fixed latency budget 2.24-fold; in the least favourable screened case, networking could draw 152 W per accelerator-equivalent before cancelling the carbon benefit. A prospectively specified fibre-spool testbed and three-site field pilot are proposed to test whether heat-sink-sited metropolitan clusters match co-located clusters on time-to-result and energy per token. Temperature-compatible heat-sink

availability is the binding thermal constraint; workload-level network performance is the open experimental question.

Keywords: Data centers; Waste heat recovery; Liquid cooling; Urban energy systems; Artificial intelligence; Hollow-core fibre; Water-energy-carbon nexus

Nomenclature

Symbol or abbreviation	Definition
f_T	Temperature-compatible fraction of total host process-heat demand (–)
S	Number of sites participating in one collective operation (–)
τ	One-way optical propagation budget or delay (μs)
v	Fibre propagation delay per routed kilometre ($\mu\text{s km}^{-1}$)
CV	Coefficient of variation of a replicated metric (–)
GHG	Greenhouse gas
HCF	Hollow-core fibre
IT	Information technology (accelerator and server) electricity
ITD	Initial temperature difference between the 45 °C loop and ambient air (K)
NCCL	NVIDIA Collective Communications Library
PV/T	Photovoltaic-thermal
SMF	Standard single-mode fibre
TOST	Two one-sided tests (equivalence testing)
WUE	Water usage effectiveness (L kWh^{-1} of IT electricity)

1. Introduction

Artificial-intelligence infrastructure is increasingly constrained not only by accelerator supply but also by power delivery, cooling, land, water and community acceptance. These burdens become spatially concentrated when large numbers of accelerators are placed in a small number of mega-scale campuses. Recent work has quantified the energy, water and carbon footprint of AI servers and documented environmental and socioeconomic pressures in communities that host concentrated data-centre development [1,2].

Liquid cooling creates an opportunity to treat data centres as thermal-energy resources rather than only electrical loads. Direct heat reuse has been demonstrated experimentally, and recent studies have examined waste-heat recovery, cooling optimization and data centres as heat-active urban energy prosumers [3–7]. Most of this literature nevertheless begins with a data centre whose location and capacity are already chosen and then asks where its heat can be used. That order of decisions can leave the thermal sink as an after-the-fact constraint. Recent work in this journal illustrates the pattern: double-layer multi-objective optimization of combined cooling and waste-heat recovery systems for data centres [8], and dual-evaporating-temperature steam-generation heat pumps that upgrade waste heat from air–liquid hybrid-cooled data centres [9]. Such studies improve the conversion and delivery of heat from a facility whose size and location are given. The present study addresses the decision that precedes them: it uses distributed commercial process-heat sinks as the constraint that sets where AI capacity is located and how much each site hosts, and couples that constraint to the metropolitan network that the resulting distributed layout requires.

The complementary question is whether useful heat demand can help determine where compute is placed and how much compute a site should host. That design principle has historically been limited by the need to co-locate tightly coupled computing hardware. The networking boundary is now changing: NVIDIA describes Spectrum-XGS as a scale-across extension of Spectrum-X for geographically separated facilities and reports NCCL collective tests at 10 km, while current B300 reference architectures provide high-bandwidth east-west interfaces on a one-NIC-per-GPU basis [10–13]. These vendor results do not demonstrate application-level parity across the longer metropolitan distances considered here, but they make geographic co-location a testable engineering constraint rather than an immutable assumption.

This study therefore evaluates heat-sink-constrained siting as an urban energy-system architecture. We construct a Los Angeles screening model in which 120 synthetic car-wash and laundromat hosts, placed geometrically within a 35-km metropolitan radius rather than drawn from a business inventory, define candidate thermal sinks and the accelerator-equivalent capacity assigned to each

host is sized to approximate its process-heat demand. The model resolves hourly IT power, process-heat demand, liquid-loop capture, thermal storage, photovoltaic-thermal (PV/T) rejection, optional adiabatic trim, dry heat rejection, auxiliary electricity, direct cooling water and operational greenhouse-gas emissions under representative weather.

The contribution is fourfold. First, the study reverses conventional heat-reuse siting by making useful host heat demand a primary capacity constraint. Second, it quantifies how workload intensity, heat-demand magnitude, temperature grade, climate and cooling choice alter energy, carbon and water outcomes in one coupled model. Third, it places the thermal design inside an explicit metropolitan network geometry while keeping network-performance claims appropriately bounded: propagation and endpoint provisioning are screened, but application-level distributed-training efficiency is not assumed. Fourth, it converts that boundary condition into a falsifiable experimental programme: screens of hollow-core fibre propagation, conservative propagation bounds for collective communication and the network electricity that would cancel the carbon benefit set the arms and pass criteria of a prospectively specified experiment comparing heat-sink-sited metropolitan clusters with co-located clusters. The purpose is not to argue that all AI workloads should be geographically distributed; it is to quantify the energy-system consequences that become available where workload-level networking performance is acceptable, and to specify how that condition can be tested.

2. Methods

2.1 System boundary and facility classes

The model represents six host classes: small, medium and large car washes (75, 125 and 300 services d^{-1}) and small, medium and large laundromats (200, 400 and 600 loads d^{-1}). Twenty instances of each class yield 120 sites. Central design capacities are 9, 15, 36, 41, 83 and 124 B300-equivalent accelerators per site, respectively, totaling 6,160 accelerator-equivalents and 11.165 MW of rated IT input. The hosts are synthetic representative facilities: their locations are generated

geometrically (Section 2.7) and their throughputs are scenario values, not an inventory of Los Angeles businesses.

The accelerator counts are continuous-equivalent thermal-sizing variables at per-accelerator resolution rather than procurement bills of materials. Rated input uses 14.5 kW per eight B300 GPUs (1.8125 kW per accelerator-equivalent) [12]. A physical deployment would use whole systems, power caps and workload allocation to approximate these thermal targets. Sensitivity cases multiply central compute capacity by 0.75, 1.25 and 1.50. Facility-class sizing is summarized in Supplementary Fig. S1 and Supplementary Note 2.

2.2 IT power and heat-following dispatch

Hourly IT power is expressed as a fraction of installed rated input. The central profile averages 64.9%, varies between 45% and 80%, and is modestly reduced on weekends. The screening model scales input power linearly with this fraction and does not include a separate idle-power floor. A capped proportional allocator distributes requested network load across sites. In the heat-following case, allocation weights combine current process-heat demand, available thermal-storage headroom and a small minimum weight; comparison cases allocate according to rated IT capacity alone.

Flat 64.9%, 80%, 90% and 95% cases test progressively more sustained compute loads. The 90% and 95% cases are intended as training-like power screens rather than measured B300 training traces. Annual IT-to-useful-heat conversion is defined as the sum of useful host heat divided by annual IT electricity, while heat-demand coverage is useful host heat divided by annual modeled process-heat demand.

2.3 Process-heat demand

Central process-heat inputs are 2.88 kWh_{th} per car-wash service and 4.96 kWh_{th} per laundromat load. These are reproducible scenario values rather than measured Los Angeles fleet averages. Car-wash lifecycle analysis likewise identifies operational energy as a material footprint contributor

[14]. Commercial laundry literature indicates that water heating and drying are major thermal loads, but usable waste-heat substitution depends on equipment and temperature grade [15–18].

The laundromat coefficient is therefore varied from 2.48 to 6.20 kWh_{th} per load. To avoid assuming that a ~45 °C liquid loop can displace every higher-temperature water-heating or dryer duty, the model also introduces a temperature-compatible fraction f_T of total host heat demand. Sensitivities use $f_T = 0.25, 0.50, 0.75$ and 1.00. The $f_T = 1.00$ case is retained as an optimistic upper-bound screen for comparability; lower values restrict useful heat to preheat and other temperature-compatible service. Detailed deployment must measure the compatible fraction at each host.

2.4 Heat capture, storage and rejection

The central liquid-loop heat-capture fraction is 85%, with 80% and 95% sensitivities. Captured heat is dispatched sequentially to concurrent process demand, a four-hour thermal store, PV/T rejection, optional adiabatic trim and dry residual rejection. The uncaptured 15% of IT heat is explicitly assigned to an air-side dry path. PV/T rejection area is parameterized at 2.2 m² per kW of rated IT input.

The free-rejection screen combines stack-assisted convection, wind-assisted flow and long-wave radiative exchange at a 45 °C rejection surface, reduced by absorbed solar gain [19–21]. Fan electricity uses a screening thermal-to-electric rejection ratio of 22.2 and liquid-loop pumping electricity is 1% of IT input. These coefficients are engineering-screen assumptions rather than vendor performance guarantees.

In the screening case, dry rejection is available up to 1.2 times rated IT input in every hour. A dry-cooler sensitivity instead limits that capacity in proportion to the initial temperature difference (ITD) between the 45 °C loop and ambient air, reaching 1.2 times rated IT input at design ITDs of 10, 15 or 20 K, and reports any heat that cannot be rejected without water. Fan electricity is also varied from 22.2 to 15.0 and 11.1 units of heat rejected per unit of electricity, applied identically to the integrated and conventional architectures.

Because a dry cooler competes with the PV/T array (2.2 m^2 per kW of rated IT) for the same roof and yard area at small hosts, a hybrid-cooler sensitivity sizes the dry coil as a fraction of rated IT input (1.2 to 0.1) at a 15-K design ITD. At fixed airflow the coil capacity scales linearly with the difference between the 45°C loop and the inlet air. When dry capacity is insufficient, adiabatic pre-cooling with 85% effectiveness lowers the inlet air toward the wet-bulb temperature, computed from dry-bulb temperature and relative humidity [22], and the additional heat rejected consumes 1.6 L of water per kWh_{th} . Spray water is drawn first from recycled host process water, limited hour by hour to 4 L per kWh_{th} of host demand without storage, with municipal water as top-up. For each site, the model also records the dry-coil size that would avoid spray in every hour.

2.5 Representative weather and direct water accounting

Weather is the California CZ2025 DEER representative EPW series for Los Angeles International Airport (WMO 722950), the representative station for Climate Zone 6. CEDARS identifies CZ2025 as a typical meteorological year based on 2000–2020 observations [23]. The 8,760 records used in the model have a mean dry-bulb temperature of 17.3°C , a $5.0\text{--}37.8^\circ\text{C}$ range, mean relative humidity of 70.2%, mean wind speed of 3.20 m s^{-1} and mean global horizontal irradiance of 226 W m^{-2} . Because the coastal airport year is the mildest in the metropolitan area, the model is rerun with the CZ2025 years for Fullerton (CZ08, WMO 722976; intermediate) and Burbank–Glendale–Pasadena (CZ09, WMO 722880; inland), which reach 42.2 and 43.9°C and exceed 35°C for 46 and 71 h, against 37.8°C and 6 h at Los Angeles International Airport [23].

Adiabatic trim is available during warm, irradiated hours and is capped at 20% of site-rated IT heat, with 1.6 L of water per kWh_{th} rejected. A separate quantity screen assumes up to 4 L of recoverable host process water per kWh_{th} demand after treatment. Direct cooling water usage effectiveness follows the Green Grid convention of site water divided by IT electricity [24]. Upstream water embodied in electricity generation, water chemistry, treatment losses and blowdown are outside this direct-WUE boundary.

2.6 Operational greenhouse-gas accounting

The primary carbon comparison holds electricity lifecycle intensity identical between the conventional and integrated architectures. Both cases deliver the same IT electricity and host process-heat demand. The conventional case rejects IT heat through a dry system and supplies all process heat with an 85%-efficient natural-gas heater. The integrated case uses recovered compute heat first and natural gas only for residual heat demand. Natural-gas combustion factors are combined with an upstream lifecycle factor [25–27].

Electricity intensity is swept from 13 to 486 gCO_{2e} kWh⁻¹, with 43 gCO_{2e} kWh⁻¹ used as the PV-like central value. Each temperature-compatible fraction is also evaluated at the central intensity. A separate transition case compares gas-like electricity at 486 gCO_{2e} kWh⁻¹ plus gas process heat against PV-like electricity plus heat reuse. The carbon calculation is operational; embodied emissions in accelerators, buildings, optical networking, PV/T equipment and storage are excluded.

2.7 Metropolitan network boundary condition

Synthetic sites are deterministically placed within a 35-km metropolitan radius using a low-discrepancy radial pattern. A 1.40 route factor converts straight-line distance to routed fibre distance, and optical propagation in standard single-mode fibre (SMF) is represented conservatively as 5 μs km⁻¹ (a group index of 1.47 gives 4.90 μs km⁻¹). Endpoint provisioning is screened at one 800-Gb s⁻¹ interface per B300-equivalent accelerator, consistent with B300 reference interface density [12,13]. The resulting aggregate installed endpoint line rate is a hardware-interface quantity, not metropolitan backbone or bisection bandwidth.

Spectrum-XGS documentation provides evidence that geographically separated facilities can participate in single-job AI training and inference, but the published 10-km tests do not validate the longer distances in this model [10,11]. The model therefore calculates geometry, propagation and endpoint capacity only. Switch and optical power, backbone oversubscription, fibre construction

cost and embodied carbon, collective-communication timing, tokens per second and scaling efficiency are outside the boundary and require separate engineering or experimental validation. Supplementary Note 1 and Supplementary Fig. S2 provide the corresponding network-boundary detail.

2.8 Fibre-type, collective-latency and network-overhead screens

Three further screens translate the network boundary condition into experimental design quantities. First, the same site geometry is evaluated with hollow-core fibre (HCF), in which light is guided in air at close to its vacuum speed. Double-nested antiresonant nodeless HCF has reached 0.091 dB km⁻¹ at 1,550 nm with approximately 30% lower latency than standard fibre [28], and HCF is being manufactured at scale for cloud networks [29]. HCF propagation is represented as 3.34 μs km⁻¹ with the same 1.40 route factor. The straight-line catchment radius reachable within a one-way propagation budget τ is $\tau/(1.40 v)$, where v is the per-kilometre delay, so the reachable host area scales as v^{-2} .

Second, a conservative worst-path propagation bound is computed for one all-reduce across S sites by charging every sequential step the worst site-to-site path through the metropolitan core: $2(S - 1)$ steps for a ring algorithm and $2\lceil \log_2 S \rceil$ steps for a tree algorithm. Because every step is charged the longest path, the result is an upper bound on the propagation term for these algorithms, not a lower bound on an optimized collective. Serialization, switching, protocol and congestion delays are excluded. The bound is not a prediction of collective completion time; it identifies the regimes in which propagation could become material.

Third, the incremental electricity of a metropolitan network (optics, switching and amplification) that would cancel the same-supply operational greenhouse-gas benefit is calculated as the avoided emissions divided by the electricity intensity, for each f_T and electricity intensity. This break-even quantity is an upper bound against which measured network electricity is compared in the proposed experiment (Section 5).

2.9 Scenarios and reproducibility

The study is a deterministic engineering screening analysis rather than an inferential statistical study. Central results are accompanied by predefined sensitivities for workload profile, heat-following allocation, compute capacity, liquid-loop capture, laundromat heat demand, temperature-compatible heat fraction, electricity carbon intensity, residual cooling technology, dry-cooler temperature limit, hybrid-cooler size, fan electricity, climate and fibre type. The complete Python model, machine-readable inputs and outputs, public weather-file identifier and figure-generation source are supplied as Supplementary Data; Supplementary Tables S1–S12 summarize the principal numerical outputs. Generative AI tools (OpenAI ChatGPT and Anthropic Claude) were used for code review and consistency checking during model development; all numerical results were generated by the supplied deterministic Python code and reviewed by the author.

3. Results

3.1 Central energy balance and useful heat recovery

The central upper-bound compatibility case consumes 63.49 GWh yr⁻¹ of IT electricity and represents 53.96 GWh yr⁻¹ of total host process-heat demand. At 85% liquid-loop capture and $f_T = 1.00$, 53.97 GWh yr⁻¹ of IT heat is captured and 50.38 GWh yr⁻¹ is delivered as useful process heat. Captured-heat utilization is therefore 93.4%, useful heat equals 79.4% of annual IT electricity, and recovered compute heat supplies 93.4% of modeled host thermal demand.

Under the modeled baseline of 85%-efficient natural-gas heaters, the same 50.38 GWh of useful heat corresponds to approximately 59.28 GWh yr⁻¹ of gas-equivalent heater input. Real hosts may use other heating technologies, so this is a modeled baseline rather than an inventory of fuel use. The integrated architecture also uses 1.07 GWh yr⁻¹ of modeled pumping and heat-rejection auxiliary electricity, compared with 2.86 GWh yr⁻¹ for conventional dry rejection of the same IT heat, a reduction of about 1.79 GWh yr⁻¹. The remaining 9.52 GWh yr⁻¹ of IT heat that is not

captured by the liquid loop is explicitly rejected through the air-side dry path. Figure 1 summarizes the modeled energy and heat-flow architecture. Supplementary Note 4 gives the corresponding derived energy quantities.

These high central utilization values are a consequence of heat-sink-matched capacity rather than evidence that arbitrary compute density can be accommodated. Increasing accelerator capacity by 25% raises host heat-demand coverage to 100% but lowers IT-to-useful-heat conversion to 68.0%; a 50% increase lowers conversion to 56.7% and increases direct evaporative cooling water to 4,505 m³ yr⁻¹. Supplementary Note 3 discusses the heat-sink ceiling and sustained-workload interpretation.

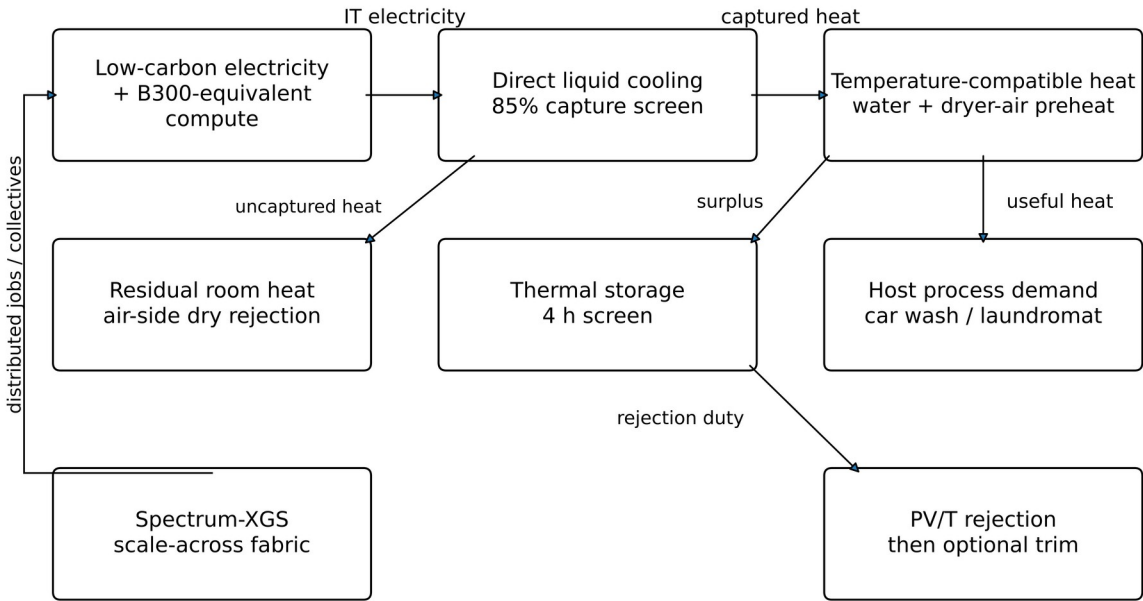


Fig. 1. Heat-sink-constrained distributed AI architecture. Captured liquid-loop heat is dispatched first to host process demand and thermal storage, with PV/T, adiabatic and dry rejection available for residual heat. Uncaptured room heat is explicitly rejected through an air-side dry path.

3.2 Workload intensity limits the thermal benefit

Removing heat-following placement while retaining the central temporal power profile changes IT-to-useful-heat conversion only slightly, from 79.4% to 78.9%. A flat 64.9% power fraction

yields 74.5%. At a flat 80% power fraction, useful heat equals 68.3% of IT electricity with heat-following allocation and 67.6% without it.

Training-like sustained loads materially reduce the thermal fraction. At 90% power, IT-to-useful-heat conversion is 61.0–61.3%, and at 95% it is 58.0–58.1%. Direct evaporative water simultaneously increases to 1,133–1,168 m³ yr⁻¹ at 90% and 1,774–1,961 m³ yr⁻¹ at 95%. The thermal advantage therefore persists at high load, but surplus heat rejection becomes increasingly important as compute intensity moves away from the host thermal profile. The workload sensitivity is summarized in Fig. 2.

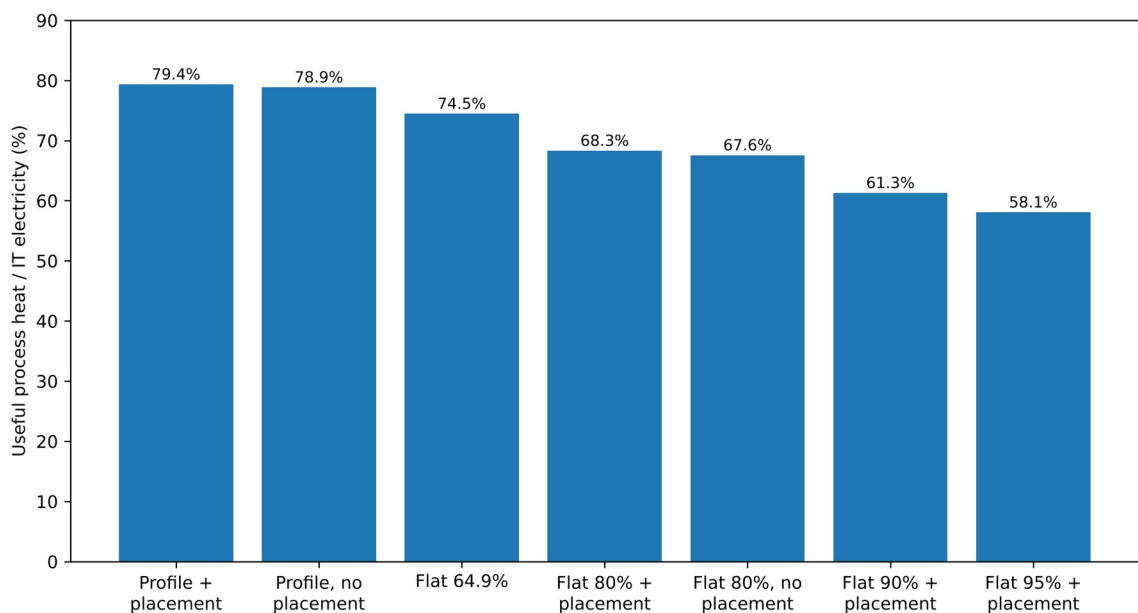


Fig. 2. Sensitivity of useful process heat to IT power profile and placement. Flat 90% and 95% cases represent sustained training-like power screens rather than measured accelerator traces.

3.3 Host heat magnitude and temperature grade are first-order sizing variables

Because the central laundromat coefficient is not a measured Los Angeles mean, its uncertainty is treated as a headline sensitivity. Across 2.48–6.20 kWh_{th} per load, IT-to-useful-heat conversion spans 50.8–84.2%, heat-demand coverage spans 82.5–100%, and direct evaporative cooling spans 8–5,978 m³ yr⁻¹. Temperature grade is independently decisive. When only 75%, 50% or 25% of total host heat demand is compatible with the ~45 °C loop or preheat service, useful heat is 40.47,

26.98 and 13.49 GWh yr⁻¹, corresponding to 63.7%, 42.5% and 21.2% of IT electricity. The $f_T = 1.00$ case is therefore an upper bound, not an assertion that all dryer and hot-water demand can be supplied directly at loop temperature. Figure 3 shows both sensitivities. Field measurement of load magnitude and temperature grade is essential before selecting site capacity. Supplementary Note 5 and Supplementary Table S6 report the temperature-grade sensitivity explicitly.

Lower compatibility also leaves more captured heat to reject. With evaporative trim available, direct cooling water rises to 735, 8,418 and 10,792 m³ yr⁻¹ at $f_T = 0.75, 0.50$ and 0.25, whereas dry-only residual rejection keeps direct cooling water at zero with unchanged useful heat.

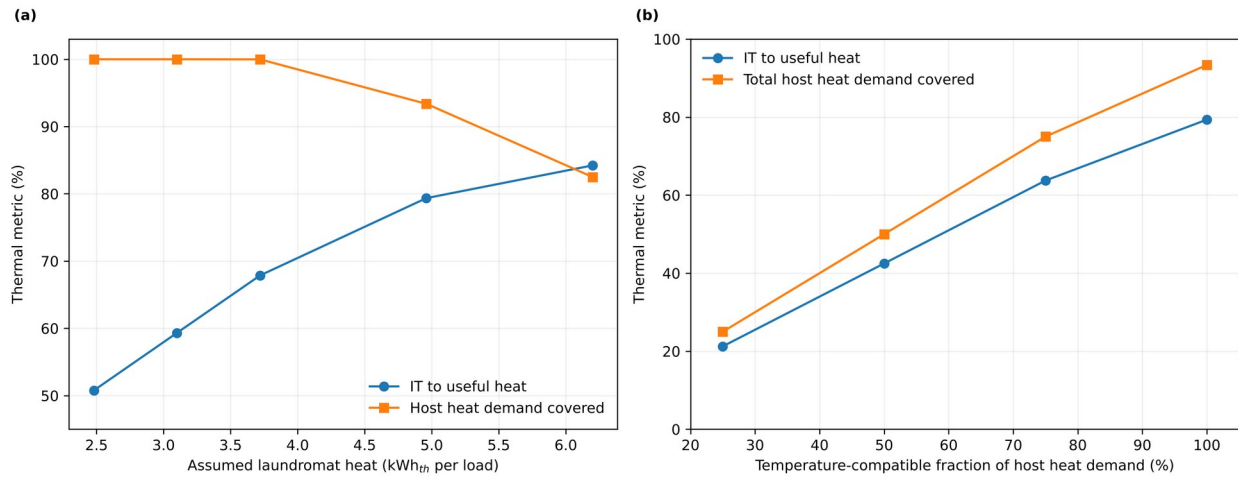


Fig. 3. Host heat-demand and temperature-grade sensitivity. (a) Sensitivity to the assumed laundromat process-heat coefficient. (b) Sensitivity to the fraction of total host thermal demand that is compatible with the ~45 °C liquid loop or preheat service.

3.4 Same-supply carbon benefits are robust but grid and temperature dependent

Holding electricity supply identical between architectures isolates the effect of useful heat recovery and different auxiliary loads. At 43 gCO₂e kWh⁻¹ electricity, modeled operational emissions fall from 16,372 to 3,672 tCO₂e yr⁻¹, a 77.6% reduction. The reduction is 63.5% at 100 gCO₂e kWh⁻¹, 43.4% at 250 gCO₂e kWh⁻¹ and 29.5% at 486 gCO₂e kWh⁻¹. The separate transition from gas-like electricity plus gas process heat to PV-like electricity plus heat reuse produces a 92.0% reduction, but most of that incremental benefit belongs to electricity decarbonization and is therefore not attributed to the heat-reuse architecture itself. These carbon values correspond to the

$f_T = 1.00$ upper-bound case. Lower temperature compatibility reduces avoided gas heating: at 43 gCO₂e kWh⁻¹, the same-supply reduction is 62.4%, 41.7% and 21.0% at $f_T = 0.75$, 0.50 and 0.25 (Supplementary Table S6). Figure 4 summarizes the same-supply sensitivity.

Percentage carbon reduction is not an adequate capacity objective on its own. A compute-oversizing case can displace more gas heat and improve the percentage reduction under a chosen denominator while simultaneously lowering useful-heat conversion and increasing cooling demand. Thermal utilization, host demand coverage and residual rejection are therefore treated as primary sizing criteria.

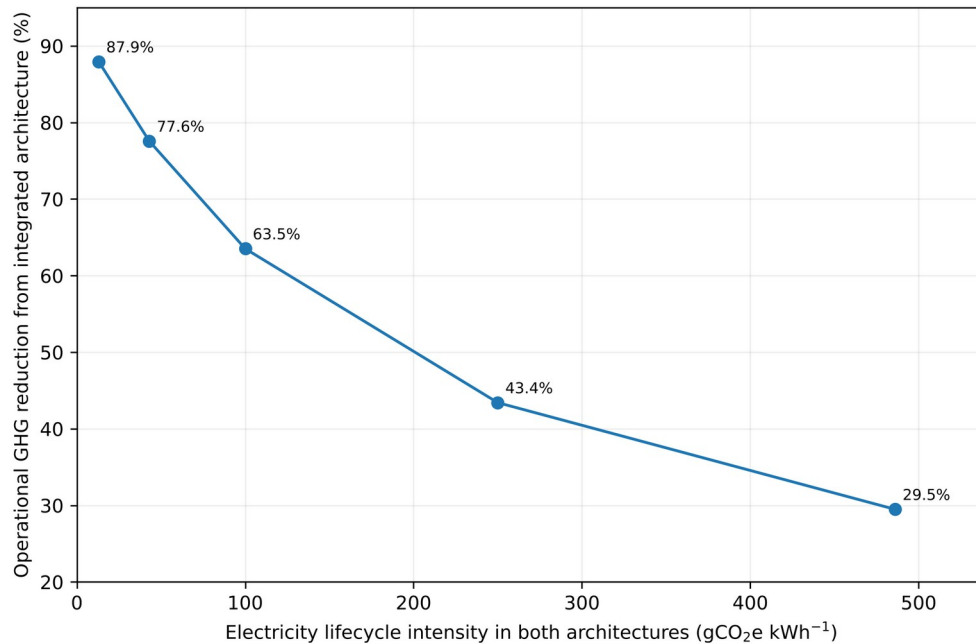


Fig. 4. Same-supply operational greenhouse-gas reduction. Electricity lifecycle intensity is held identical between architectures to isolate useful heat recovery and auxiliary-load differences.

3.5 Direct cooling-water use is small centrally but technology dependent

With representative Los Angeles weather, the central architecture uses 74 m³ yr⁻¹ of direct evaporative trim water, equivalent to 0.00116 L kWh⁻¹ of IT electricity. A dry-only residual-rejection case uses no evaporative trim water for the captured liquid-loop heat, but Section 3.6 shows that avoiding water in this way requires dry-cooler capacity that competes with the PV/T

array for site area. At flat 80% IT power, evaporative use rises to 612–641 m³ yr⁻¹; the 90% and 95% results increase further as described above. Figure 5 summarizes the direct-water sensitivity. The quantity screen also indicates that treated host-process water could be sufficient to cover the central trim-water requirement. This is not a water-quality finding: treatment recovery, chemistry, blowdown, health requirements and local permitting remain site-specific, and upstream electricity-generation water is outside the direct-water boundary.

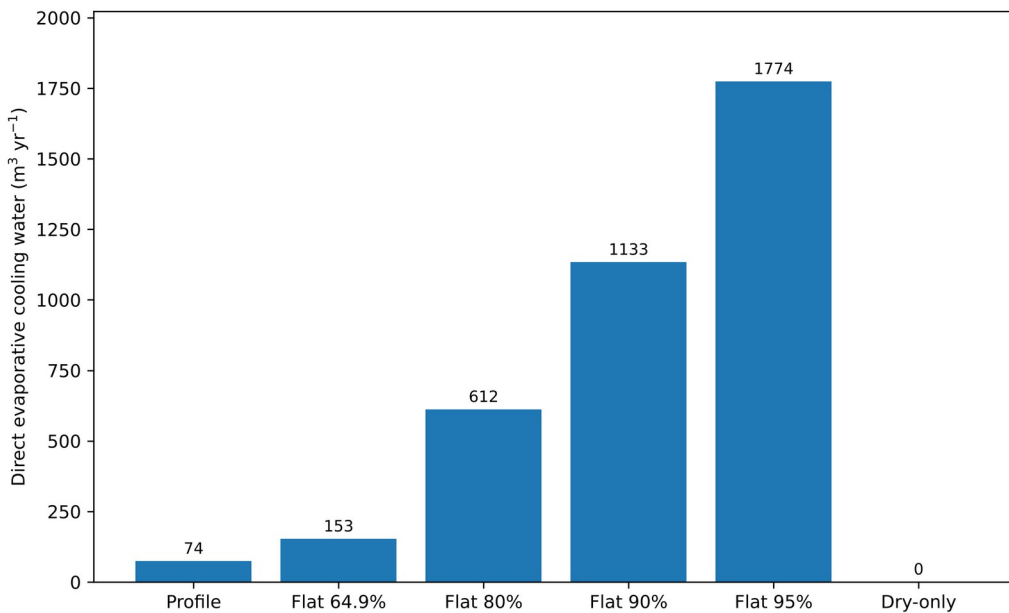


Fig. 5. Direct evaporative cooling-water sensitivity. Sustained 90% and 95% power illustrate the penalty from surplus heat; the dry-only case is a model boundary for captured liquid-loop heat whose site-area cost is evaluated in Section 3.6.

3.6 Climate changes the thermal results little; cooler footprint sets the water requirement

Replacing the coastal airport year with the intermediate Fullerton and inland Burbank years leaves IT-to-useful-heat conversion at 79.4% for $f_T = 1.00$, because host demand and heat capture do not depend on weather in the model and the residual heat to be rejected is small. Direct evaporative water with trim is 99 m³ yr⁻¹ at Fullerton and 74 m³ yr⁻¹ at Burbank, against 74 m³ yr⁻¹ at the coastal airport; at $f_T = 0.50$ it is 8,722 and 8,575 m³ yr⁻¹, against 8,418 m³ yr⁻¹ (Supplementary Table S10).

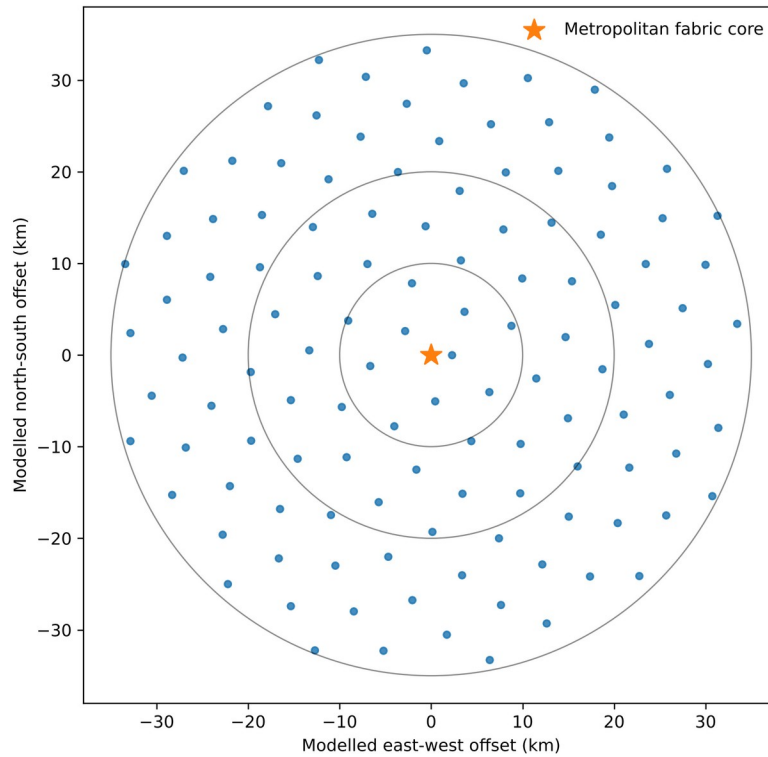
Within the modeled captured-liquid-loop rejection path, limiting dry capacity by the loop-to-ambient temperature difference leaves evaporative trim at zero at $f_T = 1.00$ in all three climates for design ITDs of 10–20 K. At lower compatibility more heat must be rejected and the limit binds in hot hours: at $f_T = 0.25$ in the inland climate, 12–52 MWh yr⁻¹ (13–43 h) cannot be rejected dry, at most 0.1% of captured heat. This sensitivity does not establish whole-facility zero-water cooling, because the 9.52 GWh yr⁻¹ uncaptured room-heat path is not temperature-limited in the model and in practice needs rejection at lower temperature than the 45 °C loop. Doubling fan electricity per unit of rejected heat raises integrated auxiliaries from 1.07 to 1.50 GWh yr⁻¹ but raises the conventional dry baseline more, from 2.86 to 5.72 GWh yr⁻¹, so the auxiliary saving widens to 4.22 GWh yr⁻¹ (Supplementary Table S11); the simplified fan model therefore does not favour the integrated architecture. These remain screening results: fan power is not resolved hour by hour against ambient temperature.

Zero-water operation is also costly in site area. PV/T, at 2.2 m² per kW of rated IT, is already the main residual-rejection surface in the model, and a dry cooler competes for the same roof and yard. Sizing a dry coil to avoid spray in every hour would require, on a capacity-weighted basis, 0.14–0.16 times rated IT at $f_T = 1.00$, but 0.47, 1.37 and 3.51 times at $f_T = 0.50$ and 0.87, 2.45 and 6.24 times at $f_T = 0.25$ in the coastal, intermediate and inland climates. A hybrid cooler with a dry coil of 0.6 times rated IT rejects all captured heat in every hour, climate and compatibility case, using at most 217 m³ yr⁻¹ of spray water (inland, $f_T = 0.25$), 30 m³ yr⁻¹ at $f_T = 0.50$ and none at $f_T = 1.00$ (Supplementary Table S12). Recycled host process water covers the spray in every hour, so no municipal top-up is required, although that result rests on the assumed 4 L per kWh_{th} of recoverable host water. Shrinking the coil to 0.4 times rated IT leaves 0.3 MWh yr⁻¹ unrejected in 11 h inland at $f_T = 0.25$. A compact hybrid cooler on recycled host water is therefore the recommended rejection design: inland, it needs about one-tenth of the dry-coil capacity that zero-water operation would require at low temperature compatibility, leaving site area for PV/T.

3.7 Metropolitan geometry is plausible as a siting screen, not a performance proof

The 120 synthetic sites lie within a 35-km straight-line radius and uses a 1.40 route factor. One-way routed propagation to the logical metropolitan core is 173 μs at the median, 238 μs at the 95th percentile and 244 μs at the maximum. At one 800-Gb s^{-1} endpoint per accelerator-equivalent, aggregate installed endpoint line rate is 4.93 Pb s^{-1} . These values describe physical geometry and interface provisioning only. The site geometry and propagation distribution are shown in Fig. 6 and Supplementary Fig. S2.

Because NVIDIA's published Spectrum-XGS example is at 10 km while routed core distances in this screen reach approximately 49 km, application-level training performance remains an unresolved boundary condition. The energy-system results should therefore be interpreted conditionally: they quantify what heat-sink-constrained siting can deliver if relevant workloads tolerate the required metropolitan communication pattern.



120 sites; 6,160 B300-equivalent accelerators; 4.93 Pb/s aggregate installed endpoint line rate.
Route-adjusted one-way propagation: 173 μ s median; 244 μ s maximum. Application-level training efficiency is not inferred.

Fig. 6. Metropolitan scale-across siting screen. The 120 synthetic host sites lie within a 35-km radius. Fibre distances use a 1.40 route factor; endpoint line rate is not interpreted as metropolitan backbone throughput.

3.8 Hollow-core fibre enlarges the reachable heat-sink catchment

With HCF, one-way propagation to the core is 116 μ s at the median, 159 μ s at the 95th percentile and 163 μ s at the maximum, and the worst site-to-site path through the core falls from 488 to 326 μ s (Fig. 7a). The more consequential effect for heat-sink siting is geographic. For the same one-way budget, HCF extends the reachable straight-line radius by a factor of 1.50 and the reachable area by a factor of 2.24; at a 250- μ s budget, the catchment grows from a 35.7-km radius (4,007 km²) to a 53.5-km radius (8,980 km²). Because temperature-compatible heat demand is the binding thermal resource, a fibre that more than doubles the candidate-host area within a given latency budget acts directly on the constraint that limits capacity. The factor assumes uniform host density; realized gains depend on the spatial distribution of suitable hosts (Supplementary Table S7).

Conservative worst-path propagation bounds for one all-reduce are below 1 ms between two sites but grow linearly with site count for ring algorithms, reaching 116 ms in SMF and 78 ms in HCF across all 120 sites, whereas tree algorithms remain at 6.8 and 4.6 ms (Fig. 7b; Supplementary Table S8). A flat ring spanning every site would therefore expose tens of milliseconds of propagation per synchronization, while hierarchical, tree-based or infrequent synchronization keeps the bound at a few milliseconds. Because every step is charged the longest site-to-site path, an optimized collective would expose less propagation; switching, serialization and congestion are excluded. The bounds inform experimental design rather than predict collective time.

The carbon benefit leaves substantial room for network electricity (Fig. 7c; Supplementary Table S9). In the central case ($f_T = 1.00$, 43 gCO₂e kWh⁻¹), incremental network electricity would have to exceed 465% of IT electricity, about 5.5 kW per accelerator-equivalent, before cancelling the avoided emissions. In the least favourable screened case ($f_T = 0.25$, 486 gCO₂e kWh⁻¹), the break-even is 12.9% of IT electricity, or 152 W per accelerator-equivalent. Measured metropolitan network electricity is therefore treated as a pass criterion of the proposed experiment rather than an assumption.

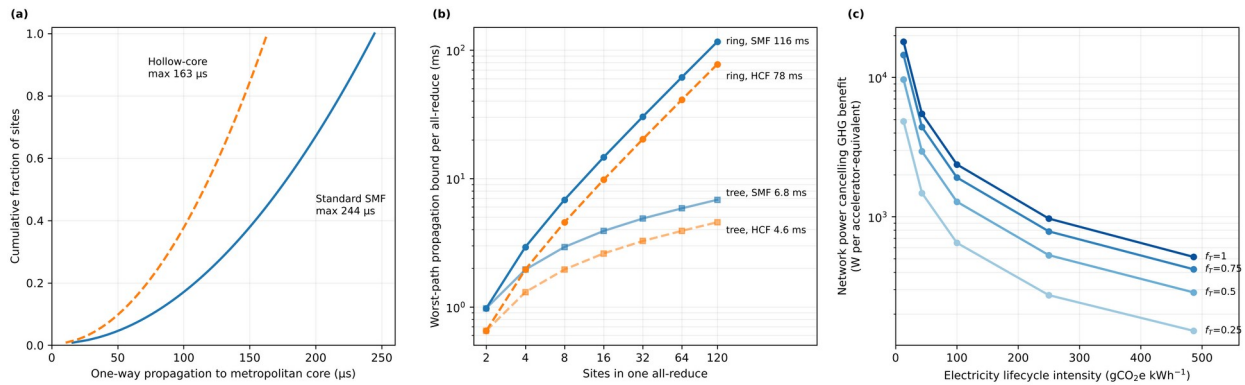


Fig. 7. Fibre and network screens that set the proposed experiment. (a) Cumulative one-way propagation from the 120 sites to the metropolitan core in standard single-mode fibre (SMF, 5.0 μs km⁻¹) and hollow-core fibre (HCF, 3.34 μs km⁻¹). (b) Conservative worst-path propagation bound of one all-reduce across S sites for ring (circles) and tree (squares) algorithms, charging every step the worst site-to-site path through the core; serialization, switching and congestion are excluded. (c) Incremental metropolitan-network power that would cancel the same-supply operational GHG benefit, by electricity intensity and temperature-compatible fraction f_T .

4. Discussion

The main energy-systems insight is that useful heat demand can be treated as a design input rather than a destination for whatever waste heat remains after a data centre is sized. In the upper-bound compatibility case, this reversal of the siting problem allows most captured IT heat to displace useful process energy, while the oversizing and temperature-grade sensitivities show two distinct ceilings: the magnitude of host heat demand and the fraction that can accept $\sim 45^\circ\text{C}$ heat or preheat. The result complements recent work that models data centres as urban energy prosumers [7] by shifting attention from a fixed data-centre campus to a network of smaller commercial thermal sinks.

This distinction matters for deployment because data-centre heat reuse is frequently limited by spatial and temporal mismatch between a large heat source and available heat demand. Distributed modules can reduce that mismatch when host facilities already require hot water or process heat. However, distribution does not eliminate infrastructure burdens; it redistributes them. Smaller sites introduce their own grid-connection, noise, permitting, security, serviceability and community-acceptance constraints. Site area is one of them: PV/T arrays, heat-rejection equipment and host operations compete for the same roof and yard, which favours compact hybrid coolers supplied with recycled host process water over zero-water dry cooling. A real deployment study should therefore combine a verified host inventory with measured thermal loads, property-level electrical capacity, executable fibre routes and local environmental constraints rather than infer social benefit from smaller site size alone.

The workload sensitivities also clarify where the concept is most attractive. The central result depends on a moderate time-varying power profile that can be allocated partly in response to heat demand. Sustained 90–95% operation still yields substantial useful heat, but conversion falls toward 58–61% and residual rejection rises. The architecture is therefore not a claim that thermal matching can absorb unlimited frontier-training capacity. Workload flexibility, inference, fine-tuning or mixtures of jobs that permit spatial and temporal allocation may be more compatible with

host heat profiles than a continuously saturated synchronous workload. Heat-following allocation itself changes central conversion by only 0.5 percentage points (79.4% versus 78.9%); the dominant thermal lever is sizing compute to the magnitude and temperature grade of the available heat sink.

The networking assumption is intentionally separated from the thermal result. Commercial scale-across networking provides a credible reason to test spatial separation, and the model shows that fibre propagation within a metropolitan radius is small in absolute time. Yet tightly synchronous AI training is sensitive not only to propagation but also to collective algorithms, topology, congestion, jitter and exposed communication. Demonstrating acceptable performance over the modeled distances requires identical co-located and multi-site benchmarks reporting collective time, step time, accelerator occupancy and application throughput; Section 5 specifies that experiment.

The fibre screens sharpen the question of whether campus-scale co-location remains necessary. A frontier model has already been trained on accelerators spread across multiple data centres [30], and low-communication training methods report quality comparable to fully synchronous training while communicating 400–500 times less between sites at the scales tested [31,32]; a 10-billion-parameter model has been trained across three continents with such a method [33]. If these methods hold at metropolitan distances on current accelerators, campus co-location becomes an option chosen for cost or operations rather than a technical requirement, and heat sinks, grid capacity and water can govern siting. If they do not, heat-sink-sited capacity is restricted to inference, fine-tuning and loosely coupled workloads. Either outcome is decision-relevant for energy planning, which is why the experiment is specified with pass criteria fixed before data collection.

The principal limitations are therefore clear. The 120 facilities are a synthetic metropolitan design module, not an inventory of existing or contracted Los Angeles businesses, and each climate year is applied to the whole module. Process-heat coefficients and temperature compatibility require field measurements; the $f_T = 1.00$ central case is explicitly an optimistic upper bound rather than a temperature-grade validation. The power model does not represent an idle floor. In the central case

dry rejection has no approach-temperature limit; the dry-cooler and hybrid-cooler sensitivities bound that simplification, but the air-side path for uncaptured room heat remains unlimited, so whole-facility zero-water cooling is not established. Cooler and PV/T footprints are expressed as capacity and area ratios rather than site layouts, recycled-water availability is an assumed 4 L per kWh_{th} of host demand, and fan power is not resolved against hourly ambient temperature. The carbon analysis is operational rather than full life cycle. The network model omits measured switch and optical electricity (bounded instead by the break-even screen), backbone oversubscription, construction carbon and capital cost, and the collective screen is a conservative propagation-only bound. Economic feasibility is also outside the current boundary; site-specific capital costs, tariffs, gas prices, maintenance, fibre charges and value of recovered heat should be incorporated before investment decisions.

Despite these limitations, the coupled sensitivities identify a practical decision rule: size distributed compute against verifiable useful heat demand and then test whether the target workload can use the resulting network topology. Where both conditions are satisfied, AI infrastructure can function as a source of process energy rather than solely as an electrical load and cooling burden.

5. Proposed experiment: a falsifiable test of campus-scale co-location

5.1 Objective and hypotheses

The experiment asks whether accelerators distributed across heat-sink sites at metropolitan distances can deliver the same time-to-result and energy per unit of work as the same accelerators co-located in one hall. Four hypotheses are specified in advance (Table 1); the protocol will be deposited in a public registry with a time-stamped record before any data are collected. H1 (collectives): for bandwidth-bound messages, cross-site collective completion time at routed site-to-site distances up to 100 km is within 5% of the co-located value, and for latency-bound messages the excess equals the propagation term within measurement error, so that HCF reduces it in proportion to its lower delay. H2 (training): time to a fixed validation loss is equivalent to the co-

located baseline within $\pm 5\%$ for at least one training algorithm at 100 km, the worst site-to-site path of the screening module. H3 (inference): throughput and 99th-percentile time-to-first-token for requests dispatched to heat-demanding sites are equivalent to the co-located baseline within $\pm 5\%$. H4 (energy): measured network electricity per accelerator-equivalent is below the break-even allowance for the measured f_T and local electricity intensity (Fig. 7c), and net energy per token after crediting metered useful heat is lower than for a co-located, dry-cooled baseline.

Table 1. Hypotheses, primary metrics and pass criteria specified before data collection.

Hypothesis	Primary metric	Pass criterion	Implication if not passed
H1 Collectives	Completion time versus message size	Bandwidth-bound: within $\pm 5\%$ of co-located at site-to-site distances ≤ 100 km; latency-bound excess equal to propagation term	Protocol or switching overhead is material; topology or HCF required
H2 Training	Time to target validation loss	Equivalent within $\pm 5\%$ for at least one algorithm at 100 km	Training remains co-located; heat-sink sites host other workloads
H3 Inference	Tokens per second; 99th-percentile time-to-first-token	Equivalent within $\pm 5\%$	Concept not supported on the tested hardware
H4 Energy	Network W per accelerator-equivalent; net energy per token	Below break-even at measured f_T (Fig. 7c); net energy per token below co-located dry-cooled baseline	Network electricity or heat credit insufficient

5.2 Phase 1: fibre-spool testbed

Phase 1 isolates distance and fibre type in a single laboratory. Sixty-four B300-class accelerators (eight eight-GPU systems) are partitioned into 2, 4 or 8 logical sites joined through fibre spools of 0 (co-located reference), 10, 25, 50 and 100 km, using the same scale-across switches and optics in every arm. Because sites exchange data through the metropolitan core, the 50-km level represents the longest site-to-core spoke (49 km) and the 100-km level the worst site-to-site path through the core (98 km) in the Los Angeles screen. Each length is run in SMF and in HCF. Where HCF spools cannot be obtained, an SMF spool shortened by the measured ratio of group delays (approximately 0.68) reproduces the HCF propagation term, and a short HCF segment confirms the equivalence of

latency and jitter. Hardware, software versions, datasets, random seeds and power caps are identical across arms, so differences are attributable only to distance, fibre type, site count and training algorithm (Table 2).

Table 2. Experimental factors and levels.

Factor	Levels
Routed distance	0 (co-located reference), 10, 25, 50 (longest site-to-core spoke) and 100 km (worst site-to-site path)
Fibre type	Standard single-mode fibre; hollow-core fibre (or delay-equivalent SMF)
Logical sites	2, 4 and 8 (Phase 1); three host sites (Phase 2)
Training algorithm	A1 hierarchical synchronous data parallel; A2 DiLoCo; A3 Streaming DiLoCo
Workload	W1 NCCL collectives (8 B–8 GiB); W2 pretraining, 1B and 8B parameters; W3 inference, ~70B parameters
Replication	Replicates per arm from pilot CV, minimum five; equivalence margin $\pm 5\%$ (TOST, $\alpha = 0.05$, 90% power)

Three workloads are run. W1 comprises NCCL all-reduce, all-gather and reduce-scatter benchmarks from 8 B to 8 GiB. W2 pretrains 1-billion and 8-billion-parameter decoder-only transformers on a fixed public corpus to a fixed validation-loss target. W3 serves an open-weight model of about 70 billion parameters under Poisson request arrivals at three load levels. Training uses three algorithms: A1, hierarchical synchronous data parallelism (sharded within a site, all-reduce across sites every step); A2, DiLoCo [31]; and A3, Streaming DiLoCo [32], with inner-step counts fixed in the protocol. Measured quantities are collective completion time and bus bandwidth; step time and exposed-communication fraction; model FLOPs utilization; time to target loss; tokens per second; median and 99th-percentile time-to-first-token; and sub-metered electricity of accelerators, hosts, switches and optics, from which energy per token is computed.

Replication is set from measured variability. Three co-located pilot runs per workload estimate the run-to-run coefficient of variation (CV) of each primary metric. For two one-sided tests at $\alpha = 0.05$ against a $\pm 5\%$ margin (δ) with 90% power and no true difference, the number of replicates per arm is $n = 2(z_{1-\alpha} + z_{1-\beta/2})^2(CV/\delta)^2$, rounded up, with a minimum of five. With $\alpha = 0.05$ and $\beta = 0.10$ both

quantiles equal $z_{0.95} = 1.645$; the $\beta/2$ term applies because, when the true difference is zero, both one-sided tests must reject [34,35]. At a CV of 2% the minimum of five governs and gives more than 97% power; at 3% and 4%, eight and fourteen replicates are required. The margin, target loss, primary metrics, sample-size rule and analysis code are fixed before data collection, and all arms are reported regardless of outcome. Measured per-hop latency, bandwidth and jitter calibrate a network simulation for site and accelerator counts beyond the testbed, and simulated extrapolations are reported separately from measurements.

5.3 Phase 2: three-site field pilot

Phase 2 transfers the configurations that pass Phase 1 to three host sites in the study metropolitan area, one car wash and two laundromats, connected to a metropolitan core by leased dark fibre, with 16–32 accelerator-equivalents per site. Over 12 months, the pilot measures what the screening model assumes: hourly process-heat demand; supply and return temperatures and flow, which give useful heat and the temperature-compatible fraction f_T ; electricity at accelerator, cooling and network level; and direct water use. Workloads W1–W3 are rerun quarterly against a co-located reference, and the measured f_T , heat demand and network electricity replace the corresponding screening inputs.

5.4 Interpretation

The outcomes map directly onto siting decisions. If A1 passes at 100 km, co-location is not required for synchronous training at the tested scale. If only A2 or A3 passes, metropolitan heat-sink clusters are viable for training that uses low-communication methods, while campus co-location retains an advantage for fully synchronous training. If only W3 passes, heat-sink sites are suited to inference and fine-tuning, which need not leave the metropolitan area. If W3 fails, the concept is not supported on the tested hardware. Because HCF changes outcomes only where the propagation term is material, H1 also determines whether HCF is required or merely advantageous;

its larger effect may be the 2.24-fold expansion of reachable heat sinks rather than faster collectives.

6. Conclusions

This study evaluated heat-sink-constrained siting of distributed AI infrastructure as an urban energy-system strategy. In a synthetic 120-host Los Angeles screening module, 6,160 B300-equivalent accelerators consume 63.49 GWh yr⁻¹ of IT electricity. Under the optimistic $f_T = 1.00$ temperature-compatibility bound, 50.38 GWh yr⁻¹ becomes useful process heat (79.4% of IT electricity), equivalent to approximately 59.28 GWh yr⁻¹ of input to the modeled 85%-efficient gas heaters, while modeled cooling and pumping auxiliaries fall by about 1.79 GWh yr⁻¹ relative to conventional dry rejection.

The result is bounded by both the magnitude and temperature grade of the thermal sink. If only 75%, 50% or 25% of host demand is temperature-compatible, IT-to-useful-heat conversion falls to 63.7%, 42.5% and 21.2%, and the same-supply emissions reduction at 43 gCO₂e kWh⁻¹ falls to 62.4%, 41.7% and 21.0%, respectively. Sustained 90–95% compute loads reduce upper-bound useful-heat conversion to approximately 58–61%, and oversizing beyond host demand rapidly increases rejection duty and water use. In the $f_T = 1.00$ case, same-supply operational emissions reductions range from 77.6% at 43 gCO₂e kWh⁻¹ to 29.5% at 486 gCO₂e kWh⁻¹. Central direct evaporative water is 74 m³ yr⁻¹. Modeled evaporative trim falls to zero in the dry-only captured-loop sensitivity, but avoiding water entirely at low temperature compatibility would require dry coils of up to 6.2 times rated IT inland; a hybrid cooler at 0.6 times rated IT instead rejects all captured heat with at most 217 m³ yr⁻¹ of recycled host water. Coastal, intermediate and inland Los Angeles weather leave useful heat unchanged.

Hollow-core fibre reduces maximum one-way propagation from 244 to 163 μ s and enlarges the host catchment within a fixed latency budget 2.24-fold, and even in the least favourable screened case the metropolitan network could draw 152 W per accelerator-equivalent before cancelling the

operational carbon benefit. The architecture should therefore be interpreted as a conditional integration pathway, not a replacement for all mega-scale data centres. Its energy value depends on measured, temperature-compatible heat sinks and on workload-level validation of metropolitan scale-across networking. Section 5 specifies the experiment that decides the second condition: a fibre-spool testbed and a three-site field pilot with prospectively specified equivalence margins, to be paired with site-specific techno-economic analysis.

Data availability

Machine-readable central results, workload/sizing, laundromat heat, temperature-grade, carbon, facility, network-site, fibre latency-budget, collective-latency, network-overhead break-even, climate and dry-cooler, fan-power, hybrid-cooler and weather-source data are supplied as Supplementary Data with this submission. The weather series are the public California CZ2025 Los Angeles International Airport, Fullerton Municipal Airport and Burbank–Glendale–Pasadena Airport EPW files identified in Section 2.5.

Code availability

The complete Python model and figure-generation source required to reproduce the analyses are supplied with the submission. The model accepts local EPW files or retrieves the identified public CZ2025 Los Angeles, Fullerton and Burbank EPW files from a pinned repository commit.

CRedit authorship contribution statement

Jonathan Bily: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing, Project administration.

Funding

This research received no external funding.

Declaration of competing interest

The author is affiliated with Cloud Wash Laundry Corporation and has a pending patent application related to distributed compute systems that recover GPU waste heat at laundromats, car washes and other commercial heat-use sites. This financial and intellectual-property interest is directly related to the concept evaluated in this work and should be considered when interpreting the study. The quantitative model, assumptions and sensitivity outputs are supplied for independent scrutiny.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used OpenAI ChatGPT and Anthropic Claude (Claude Opus 5.5) in order to assist with drafting, code review, document preparation and checking that reported values match the model outputs. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- [1] Xiao, T., Fuso Nerini, F., Matthews, H. D., Tavoni, M., You, F. Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. *Nature Sustainability* 8 (2025) 1541–1553. <https://doi.org/10.1038/s41893-025-01681-y>
- [2] Ngata, W. M., Bashir, N., Westerlaken, M., Lioté, L., Chandio, Y., Olivetti, E. The Cloud Next Door: Investigating the Environmental and Socioeconomic Strain of Datacenters on Local Communities. *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '25)* (2025) 769–774. <https://doi.org/10.1145/3715335.3736324>
- [3] Zimmermann, S., et al. Aquasar: a hot water cooled data center with direct energy reuse. *Energy* 43 (2012) 237–245.
- [4] Yuan, X., et al. Waste heat recoveries in data centers: a review. *Renewable and Sustainable Energy Reviews* 188 (2023) 113777.
- [5] Gheni, M., Kerskes, H., Stergiaropoulos, K. Operational analysis of the cooling system in a direct liquid-cooled data center: a measurement and simulation study on the impact of supply water temperature. *Applied Energy* 403 (2026) 127061. <https://doi.org/10.1016/j.apenergy.2025.127061>

- [6] Stahlhut, M., Sekulla, M., Großöhme, P., Auweter, A., Urbaneck, T. Data Centers With Direct Liquid-Cooled Servers: Experimental Analysis and Modeling. *Energy Science & Engineering* 13 (2025) 3605–3619. <https://doi.org/10.1002/ese3.70116>
- [7] Ravi, S. S., et al. From servers to services: modeling data centers as heat-active urban energy prosumers. *Applied Energy* 402 (2026) 127049. <https://doi.org/10.1016/j.apenergy.2025.127049>
- [8] Chen, S., Zhang, Q., Hu, M., Zhai, J., Li, J., Guo, Z. Double-layer multi-objective optimization for combined cooling and waste heat recovery systems in data centers. *Energy Conversion and Management* 356 (2026) 121349. <https://doi.org/10.1016/j.enconman.2026.121349>
- [9] Zhang, C., Li, M., Chen, X., Dang, C., Li, X., Han, Z. Dual evaporating temperature steam generation heat pump system for waste heat recovery of air–liquid hybrid cooling data center. *Energy Conversion and Management* 353 (2026) 121199. <https://doi.org/10.1016/j.enconman.2026.121199>
- [10] NVIDIA. NVIDIA Introduces Spectrum-XGS Ethernet to Connect Distributed Data Centers Into Giga-Scale AI Super-Factories. 22 August 2025.
- [11] NVIDIA. How to Connect Distributed Data Centers Into Large AI Factories with Scale-Across Networking. 9 September 2025.
- [12] NVIDIA. Introduction to NVIDIA DGX B300 Systems. DGX B300 User Guide (accessed 2026).
- [13] NVIDIA. NVIDIA HGX AI Factory Enterprise Reference Architecture: Components and networking physical topologies. <https://docs.nvidia.com/enterprise-reference-architectures/hgx-ai-factory/latest/components.html> (accessed 28 September 2026).
- [14] Maciejewska, K., Reizer, M. Evaluating the Impacts of Different Car Washing Systems on Carbon Footprint: Insights from Poland. *Sustainability* 17 (2025) 1384. <https://doi.org/10.3390/su17041384>
- [15] U.S. Department of Energy. Energy Savings Potential and RD&D Opportunities for Commercial Building Appliances. Building Technologies Office (2014).
- [16] U.S. Department of Energy. Energy Savings Potential and RD&D Opportunities for Commercial Building Appliances: 2015 Update. Building Technologies Office (2015).
- [17] Choi, Y., Chen, Y., Ozaki, A., Lee, H., Arima, Y. Advancing sustainable laundry practices utilizing waste heat from commercial laundry facilities. *Applied Thermal Engineering* 247 (2024) 123136. <https://doi.org/10.1016/j.applthermaleng.2024.123136>
- [18] Williamson, N., Bansal, P. K. Experimental investigation of the energy efficiency of gas commercial laundry dryers. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering* 218 (2004) 143–152. <https://doi.org/10.1243/0954408041323421>
- [19] Hnayno, M., Chehade, A., Klabi, H., Polidori, G., Maalouf, C. Experimental investigation of an optimized indirect free cooling system including a dry cooler equipped with evaporative cooling pads for data center. *Energy Reports* 9 (2023) 460–469. <https://doi.org/10.1016/j.egy.2023.12.005>
- [20] Zondag, H. A. Flat-plate PV-Thermal collectors and systems: a review. *Renewable and Sustainable Energy Reviews* 12 (2008) 891–959.
- [21] Raman, A. P., Anoma, M. A., Zhu, L., Rephaeli, E., Fan, S. Passive radiative cooling below ambient air temperature under direct sunlight. *Nature* 515 (2014) 540–544.

- [22] Stull, R. Wet-bulb temperature from relative humidity and air temperature. *Journal of Applied Meteorology and Climatology* 50 (2011) 2267–2269. <https://doi.org/10.1175/JAMC-D-11-0143.1>
- [23] California Public Utilities Commission, CEDARS. CZ2025 DEER Weather Files: representative weather files for California climate zones; CZ06 uses Los Angeles International Airport, WMO 722950; EPW TMY based on 2000–2020 (2026).
- [24] Patterson, M., et al. Water Usage Effectiveness (WUE): A Green Grid Data Center Sustainability Metric. *The Green Grid White Paper* 35 (2011).
- [25] National Renewable Energy Laboratory. Life Cycle Greenhouse Gas Emissions from Electricity Generation: Update. NREL/FS-6A50-80580 (2021).
- [26] Pandey, I., et al. Life Cycle Analysis of Natural Gas Supply Chain and End Use Applications in the United States. *Environmental Science & Technology* 60 (2026) 3148–3158. <https://doi.org/10.1021/acs.est.5c14132>
- [27] U.S. Environmental Protection Agency. GHG Emission Factors Hub (2025).
- [28] Petrovich, M., Numkam Fokoua, E., Chen, Y., et al. Broadband optical fibre with an attenuation lower than 0.1 decibel per kilometre. *Nature Photonics* 19 (2025) 1203–1208. <https://doi.org/10.1038/s41566-025-01747-5>
- [29] Microsoft Azure Networking. Microsoft Azure scales Hollow Core Fiber (HCF) production through outsourced manufacturing. *Microsoft Tech Community*, 23 September 2025. <https://techcommunity.microsoft.com/blog/azurenetworkingblog/microsoft-azure-scales-hollow-core-fiber-hcf-production-through-outsourced-manuf/4455953>
- [30] Gemini Team, Google. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805* (2023). <https://doi.org/10.48550/arXiv.2312.11805>
- [31] Douillard, A., Feng, Q., Rusu, A. A., Chhaparia, R., Donchev, Y., Kuncoro, A., Ranzato, M., Szlam, A., Shen, J. DiLoCo: distributed low-communication training of language models. *arXiv:2311.08105* (2023). <https://doi.org/10.48550/arXiv.2311.08105>
- [32] Douillard, A., et al. Streaming DiLoCo with overlapping communication: towards a distributed free lunch. *arXiv:2501.18512* (2025). <https://doi.org/10.48550/arXiv.2501.18512>
- [33] Jaghouar, S., Ong, J. M., Basra, M., et al. INTELLECT-1 technical report. *arXiv:2412.01152* (2024). <https://doi.org/10.48550/arXiv.2412.01152>
- [34] Lakens, D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 8 (2017) 355–362. <https://doi.org/10.1177/1948550617697177>
- [35] Chow, S.-C., Shao, J., Wang, H. *Sample Size Calculations in Clinical Research*, 2nd ed. Chapman & Hall/CRC, Boca Raton (2008).