Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Direct wireless large language model inference from high-altitude platform stations under payload constraints

Young-Chai Ko

koyc@korea.ac.kr

Korea University

Byeongheon Lee

Korea University

Article

Keywords: High-altitude platform stations (HAPS), large language model inference, non-terrestrial networks, aerial computing, edge computing

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11385664/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Direct wireless large language model inference from high-altitude platform stations under payload constraints

Byeongheon Lee[1] and Young-Chai Ko[1*]

[1]School of Electrical and Computer Engineering, Korea University, Seoul, 02841, Korea.

*Corresponding author(s). E-mail(s): koyc@korea.ac.kr;
Contributing authors: tbio@korea.ac.kr;

Abstract

High-altitude platform stations (HAPS) could extend 6G non-terrestrial networks from wireless access to onboard AI inference, but serving capacity is jointly constrained by computing, thermal management, radio circuitry and finite payload resources. We develop a cross-layer framework combining a 110-day LLM workload trace, concurrency-dependent latency, stochastic user activity, thermal simulation, radio-power modelling and reported HAPS payload limits. Long-context requests account for only 1.35% of elapsed-time occupancy yet consume 48.6% of the serving budget for the evaluated 70B configuration. Under a 5% model-defined compute-overload criterion, the 8B/1×H100 and 70B/2×H100 configurations support 5,876 and 991 user equivalents. Integrating computing, thermal, communication, power and mass constraints yields platform-level capacities up to 7,245 and 23,144 user equivalents for Sunglider-class and Stratobus-class platforms. These results show that long-context demand, heat-rejection capability and payload power—not accelerator throughput alone—govern HAPS-based LLM inference feasibility in 6G NTN.

Keywords: High-altitude platform stations (HAPS), large language model inference, non-terrestrial networks, aerial computing, edge computing

1 Introduction

The rapid adoption of generative artificial intelligence (AI) is driving a corresponding expansion of inference infrastructure. Unlike conventional cloud workloads, large-language-model (LLM) inference combines substantial accelerator demand with highly variable request sizes, interactive latency requirements and temporally bursty user activity. At scale, the resulting growth in computing infrastructure increases pressure on electrical grids, cooling systems, water resources, physical space and local infrastructure[1,2]. These constraints have motivated increasing interest in computing facilities that exploit unconventional physical environments or infrastructure. Examples include underwater and underground data centres, which have explored alternative approaches to site availability, thermal management and energy integration[3,4].

An emerging extension of this concept is to move computing infrastructure above the terrestrial environment. Orbital or space-based computing could offer broad geographical reach and access to solar energy, but introduces fundamentally different engineering constraints, including launch mass and volume, radiation exposure, difficult maintenance and the need to reject heat radiatively in vacuum[5,6]. On the other hand, high-altitude platform stations (HAPS) occupy

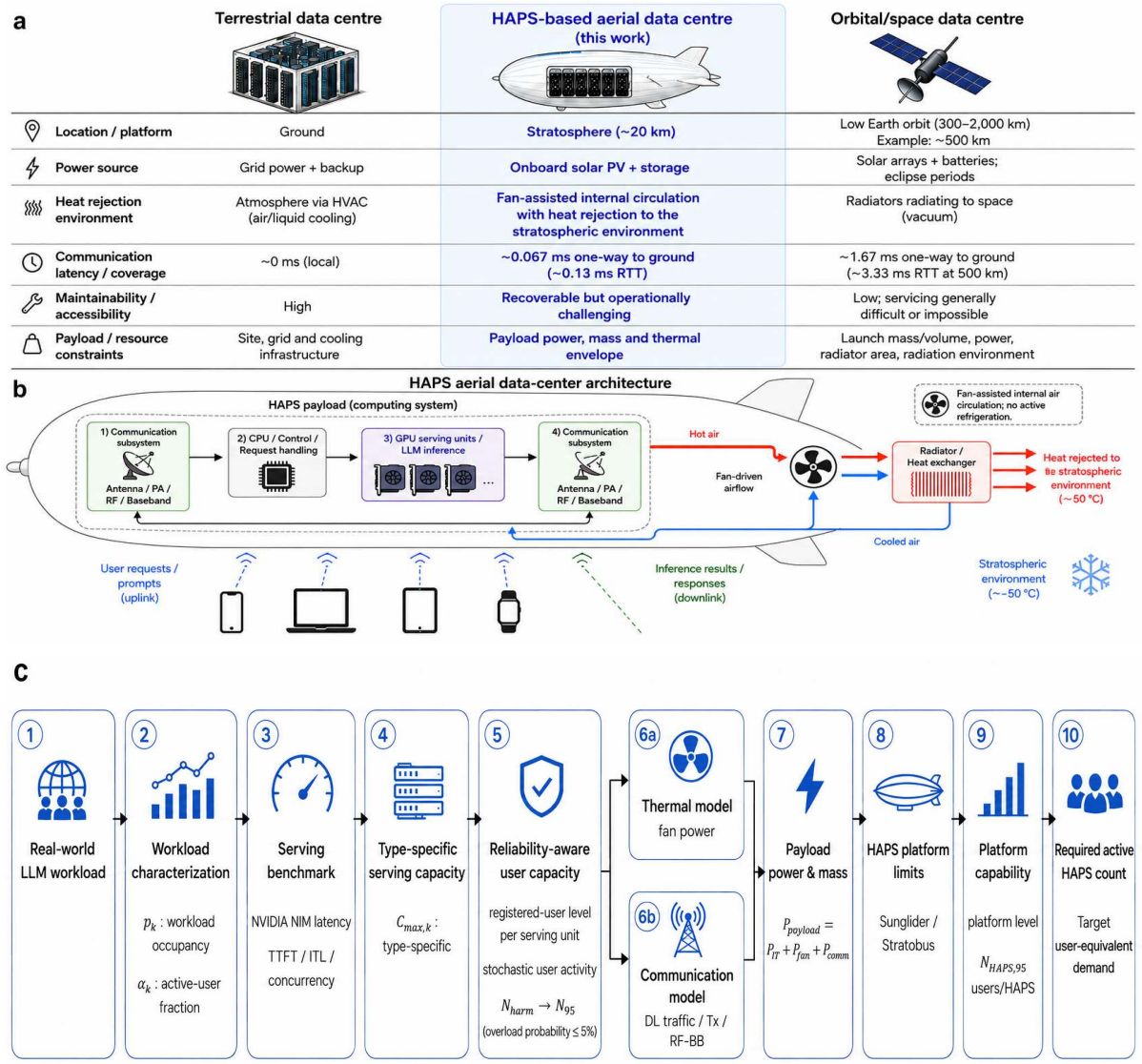


Fig. 1 Concept and cross-layer framework for HAPS-based aerial LLM inference. a, Conceptual comparison of terrestrial, HAPS-based and orbital/space data-centre regimes. Propagation-delay values indicate propagation only. **b,** Proposed HAPS architecture integrating onboard LLM inference, direct ground-user communication and fan-assisted heat rejection to the stratospheric environment. **c,** Cross-layer analysis from trace-derived workload characterization and SLO-constrained serving performance to thermal and radio power, platform constraints and reliability-aware HAPS capacity.

an intermediate physical regime between terrestrial and orbital infrastructure (Fig. 1a). Operating in the stratosphere at approximately 20 km altitude, HAPS are more geographically flexible than fixed terrestrial facilities while remaining far more accessible and potentially recoverable than orbital platforms. At the same time, they are subject to stringent payload-power and payload-mass constraints. The cold stratospheric environment also provides a potentially favourable heat sink for computing payloads, although exploiting this environment requires sufficient thermal coupling between the onboard electronics and the external air.

From a 6G perspective, HAPS form an airborne layer of the broader non-terrestrial network (NTN) architecture. While NTN research has traditionally emphasized radio access and coverage,

recent studies have also considered HAPS as components of 6G NTN and edge-computing architectures[7–9]. Equipping HAPS with onboard AI computing could further extend this role from communication relays toward joint communication-and-computing nodes, allowing selected user traffic to be processed directly within the non-terrestrial segment.

Previous studies have examined HAPS as communication relays, upper-tier mobile-edge-computing nodes, workload-offloading platforms and, more recently, aerial computing or data-centre platforms[10,11]. However, these studies do not establish how many interactive generative-AI users can be supported when workload heterogeneity, concurrency-dependent latency, stochastic user activity, thermal rejection, radio circuitry and finite HAPS payload resources are considered jointly.

This distinction is particularly important for text-generation workloads. Real LLM demand spans orders of magnitude in prompt and response length, and different request classes need not contribute equally to either temporal occupancy or instantaneous serving demand. A workload that appears rarely in a trace can nevertheless become a dominant resource bottleneck if its supported concurrency is small. Similarly, an average-load capacity estimate does not quantify the probability that instantaneous user activity exceeds the available serving resources. Reliable user capacity must therefore be derived jointly from workload composition, latency-constrained serving performance and stochastic activity, rather than inferred directly from peak accelerator throughput.

We consider a HAPS-based aerial data-centre architecture in which ground users communicate directly with the airborne computing payload (Fig. 1b). User prompts are transmitted to the HAPS through the onboard communication subsystem, processed by the host and GPU serving system, and the generated responses are returned directly to the users over the downlink. The corresponding information technology (IT) heat is removed by a closed internal airflow loop: fan-driven circulation transports heat from the computing payload to a radiator or heat exchanger, from which it is rejected to the stratospheric environment without a dedicated vapour-compression chiller. The computing, thermal-management and communication subsystems therefore compete for the same finite payload-resource envelope.

Importantly, we treat the publicly reported payload-power and payload-mass allowances of representative HAPS platform concepts as resources made available to the mission payload[12,13]. Aircraft-level functions required to sustain stratospheric operation—such as propulsion or station keeping, avionics, flight control, solar-energy conversion and energy storage—are consequently treated as upstream platform functions associated with providing this payload allowance, rather than being charged again to the computing payload. Our objective is therefore not to reconstruct the complete aircraft energy balance, but to determine whether an LLM inference payload can operate within the power and mass envelope made available by the HAPS platform.

The central question addressed in this work is therefore: under what workload, serving-performance, thermal, radio and payload-resource conditions can a HAPS operate as a practical AI inference node within a non-terrestrial network, and what reliability-aware user-equivalent capacity follows from those conditions? To answer this question, we develop the cross-layer analysis summarized in Fig. 1c. We first derive workload occupancy and type-specific active-user fractions from the BurstGPT trace[14]. Concurrency-dependent NVIDIA Inference Microservices (NIM) latency measurements[15] are then used to obtain service level objective (SLO) constrained serving capacities for representative workload classes, which are combined with stochastic user activity to directly evaluate the model-defined overload probability and derive a reliability-aware user-equivalent operating point under a 5% compute-overload criterion. We next quantify the minimum fan power required for thermal feasibility and the downlink traffic, radiated power and radio-circuit power required for direct user access. Finally, these serving, thermal and communication requirements are integrated with platform payload-power and payload-mass constraints, and platform-level reliability is recomputed when multiple serving instances are pooled.

The resulting framework identifies the conditions under which reliability-constrained LLM inference can fit within the thermal, radio and payload-resource envelope of a HAPS. It further provides a direct mapping from trace-derived

demand to reliable user equivalents per platform and the active-HAPS count required for a target service population, thereby quantifying the feasibility of HAPS as an integrated communication-and-computing node for future 6G NTN deployments.

Results

Trace-derived LLM demand is heterogeneous in workload occupancy and user activity

Real-world LLM demand cannot be represented by a single request profile because workload classes differ not only in token demand, but also in their contribution to service occupancy and their probability of being active. We therefore characterize the BurstGPT trace using both request–response token pairs and temporal usage information. Figure 2a shows the joint distribution of input and output tokens, together with the five representative workload types considered in this study. The trace spans several orders of magnitude in both input and output length, demonstrating that user demand is distributed over a broad token space rather than concentrated around a single representative request. The five workload types consequently capture distinct regions of this space, ranging from short conversational requests to long-context and long-generation workloads. Types C and E occupy comparatively sparse regions of the trace but represent distinct long-generation and long-context demand regimes, respectively.

Request frequency alone, however, does not reflect how long each workload occupies the serving system. We therefore define the workload occupancy probability p_k from the elapsed-time share of workload type k, rather than from its request-count share. As shown in Fig. 2b, Type A accounts for approximately 60% of all requests and is therefore dominant when the workload is described solely by request frequency. Its contribution decreases markedly when requests are weighted by elapsed time, consistent with its relatively short input–output profile and processing duration. In contrast, Types B and D contribute substantially larger fractions of elapsed-time occupancy than suggested by their request counts. Consequently, the elapsed-time representation is no longer dominated by a single workload class: Types D, A and B jointly account for most of the observed service occupancy at comparable shares. This discrepancy between request frequency and temporal occupancy indicates that a single "typical request" would provide a poor representation of the workload seen by an LLM serving system.

Heterogeneity also appears in the inferred type-specific activity fractions. Figure 2c reports the type-specific active-user fraction α_k, together with bootstrap confidence intervals obtained from the trace. Type A consistently exhibits a substantially lower active-user fraction than the other workload classes, indicating that users associated with short conversational requests spend a smaller fraction of time actively generating service demand. Type D shows an intermediate activity level, whereas Types B, C and E exhibit higher active-user fractions. The bootstrap intervals preserve the pronounced separation between Type A and the remaining workload classes, while the intervals for the higher-activity classes partially overlap; we therefore do not interpret small differences among Types B, C and E as a strict ranking. These results show that heterogeneity in token demand is accompanied by heterogeneity in user activity, and that request frequency alone is insufficient to describe the instantaneous population presented to the serving system.

Together, the elapsed-time occupancy p_k and active-user fraction α_k define the workload-dependent stochastic population model used in the following reliability-aware capacity analysis. By combining these trace-derived demand characteristics with type-specific serving limits, we next quantify how workload heterogeneity translates into nominal and reliability-aware user capacity.

Long-context demand and stochastic activity separate nominal from reliable serving capacity

We next examine how the heterogeneous workload identified above translates into serving capacity under interactive latency requirements. Figure 3a shows the measured time-to-first-token (TTFT) and inter-token latency (ITL) as functions of concurrency for two representative serving configurations: Llama 3.1 8B (approximately 8 billion parameters) on 1×H100 (FP8 TP1) and Llama 3.3 70B on 2×H100 (FP8 TP2). Latency increases

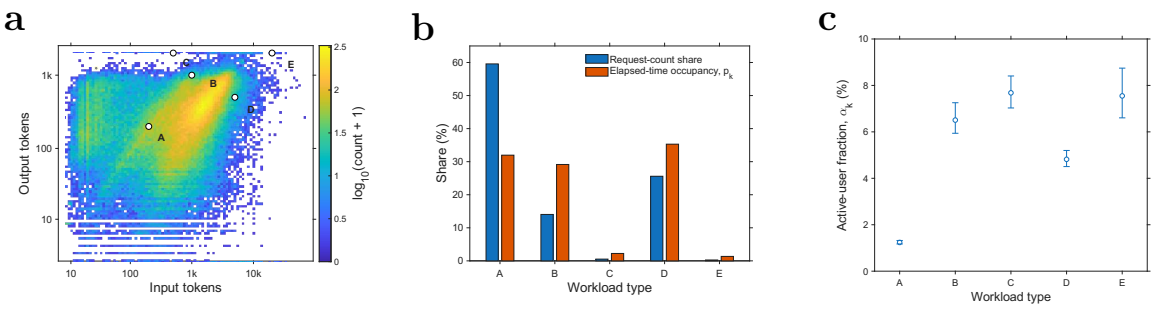


Fig. 2 Trace-derived workload composition and user activity. a, Joint distribution of input and output token counts in the BurstGPT trace; colour indicates $\log_{10}(\text{count}+1)$. Markers denote the representative workload types (input/output tokens): A, 200/200; B, 1000/1000; C, 500/2000; D, 5000/500; and E, 20000/2000. **b,** Request-count share and elapsed-time occupancy p_k for each workload type. **c,** Type-specific active-user fraction α_k for the session-continuity threshold $\tau = 1800$ s. Markers show estimates from the original trace and error bars denote the 2.5th–97.5th percentile interval from day-level bootstrap resampling.

non-uniformly across workload types, indicating that the concurrency depends strongly on the input–output token profile. The smaller-model configuration sustains substantially higher concurrency under the same latency constraints, whereas the larger model becomes particularly sensitive to long-context demand. This behavior is most pronounced for the long-context Type E workload, for which TTFT becomes the dominant bottleneck even at low concurrency in the 70B configuration.

We converted these latency–concurrency measurements into the maximum concurrency $C_{\max,k}$ satisfying the prescribed latency SLOs for each workload type (Fig. 3b). Under the general SLO of TTFT ≤ 2 s and ITL ≤ 100 ms, the 8B configuration supports high concurrency for Types A–D but falls to approximately 19 concurrent requests for Type E. For the 70B configuration, the interpolated Type-E capacity is only 1.62 concurrent requests. Tightening the TTFT requirement further reduces this capacity and renders Type E infeasible under the evaluated TTFT-stringent SLO. Thus, long-context requests impose a serving bottleneck that is not apparent from their occurrence frequency alone.

We next combined p_k and $C_{\max,k}$ to estimate the nominal mixed-workload capacity using the weighted harmonic resource-budget model defined in the Methods section. Figure 3c compares the resulting normalized serving-budget contribution with the elapsed-time occupancy measured from the trace. Type E exhibits a pronounced mismatch between workload prevalence and serving impact: although it accounts for only approximately 1.35% of elapsed-time occupancy, it consumes 13.1% of the serving budget for the 8B configuration and 48.6% for the 70B configuration. Thus, a rare workload can become a dominant capacity constraint when its type-specific serving limit is sufficiently small.

The harmonic estimate provides a nominal mixed-workload capacity, but it does not guarantee reliability under stochastic user activity. To distinguish the modelled population from counts of uniquely identified subscribers or accounts, we define a registered-user equivalent as a modelled demand source characterized by the trace-derived workload mixture and type-specific activity fractions. Hereafter, we refer to such modelled sources simply as user equivalents.

We define N_{95} as the largest user-equivalent population for which the model-defined compute-overload probability does not exceed 5%. For each candidate population, type-specific active-user counts follow the independent Binomial activity model defined in the Methods section, and the resulting overload probability is evaluated directly rather than estimated from finite Monte-Carlo samples. Figure 3d shows this overload probability as the user-equivalent population is increased relative to the nominal harmonic estimate N_{harm}. For both configurations, N_{harm} lies near a regime in which the overload probability is approximately one half. This behavior arises because the harmonic estimate places the mean offered load close to the serving limit and therefore leaves little headroom for instantaneous fluctuations in the

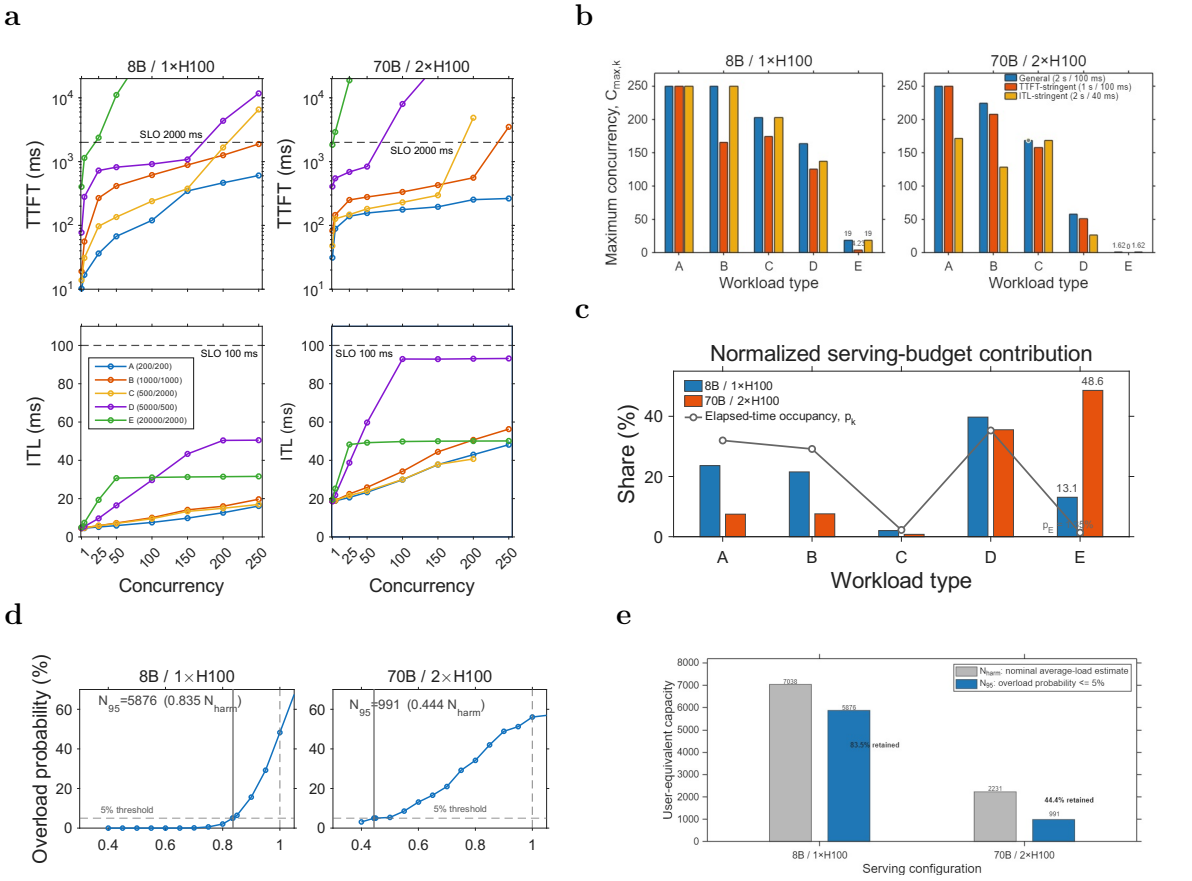


Fig. 3 Long-context demand and stochastic activity separate nominal from reliable serving capacity. a, Time to first token (TTFT) and inter-token latency (ITL) as functions of concurrency for the five representative workload types on the 8B/1×H100 and 70B/2×H100 serving configurations. Dashed lines indicate the general SLO thresholds of 2000 ms TTFT and 100 ms ITL. **b,** Type-specific SLO-constrained serving capacity $C_{\max,k}$ under the general SLO and two latency-stringent sensitivity conditions. Non-integer values denote interpolated SLO-crossing capacities. **c,** Normalized serving-capacity budget contribution of each workload type under the general SLO; the grey line shows the elapsed-time occupancy p_k. **d,** Overload probability as a function of user-equivalent population normalized by the harmonic estimate N_{harm}. Dashed lines indicate the 5% overload criterion and N_{harm}, and solid vertical lines indicate N_{95}. The displayed probabilities and N_{95} values are obtained by direct evaluation of the adopted Binomial activity model. Monte-Carlo validation of the probability calculation is reported in Supplementary Fig. 4. **e,** Nominal harmonic capacity N_{harm} and reliability-aware capacity N_{95} for the two serving configurations.

number and composition of active users. Consequently, a substantially smaller user-equivalent population is required to satisfy the 5% overload criterion.

The reliability penalty is strongly configuration dependent (Fig. 3e). For the 8B/1×H100 configuration, the nominal estimate of $N_{\text{harm}} = 7{,}038$ user equivalents decreases to $N_{95} = 5{,}876$, retaining 83.5% of the nominal capacity. For the 70B/2×H100 configuration, $N_{\text{harm}} = 2{,}231$ user equivalents decreases to $N_{95} = 991$, corresponding to 44.4% of the nominal estimate. The larger reduction in the 70B configuration results from the same stochastic user activity being weighted by much smaller type-specific serving capacities, most notably for the long-context Type-E workload. Hence, average-load capacity alone does not determine reliable serving capability: the sensitivity of each serving configuration to rare but expensive workload types controls the headroom required against stochastic demand fluctuations.

We therefore use N_{95}, rather than N_{harm}, as the serving-population input to the communication and platform analyses. The thermal analysis

is evaluated independently at a representative 1.1-kW IT heat load and subsequently supplies an auxiliary-power reference.

Low-power thermal management emerges within a bounded heat-rejection envelope

We next evaluate whether the onboard computing payload can be thermally managed in the stratospheric environment without a dedicated vapour-compression chiller. The thermal architecture considered in this work is illustrated in Fig. 4a. Heat generated by the onboard IT system is transported by fan-driven internal airflow to an internal heat-exchanger node, conducted through an equivalent wall/feedthrough thermal path to the external radiator, and ultimately rejected to the cold stratospheric ambient. The cooled internal air is subsequently recirculated to the computing payload, forming a closed internal airflow loop. At the external radiator, the model additionally accounts for absorbed solar heat and radiative exchange with the environment. We characterize the resulting fan requirement using a reference IT heat load of $Q_{\mathrm{IT,ref}} = 1.1$ kW, with the minimum fan electrical power P_{fan} required to satisfy the prescribed server- and cabin-temperature constraints as the principal thermal-management metric.

Figure 4b maps the resulting thermal feasibility as a function of the effective external convective heat-rejection conductance UA_{ext} and the effective wall/feedthrough conductance G_{wall}. A broad region of the parameter space supports thermal regulation with only a small fan-power contribution, whereas the required circulation power rises sharply as the available heat-rejection paths are weakened. Notably, the 5% fan-power contour closely follows the boundary corresponding to a fan speed of approximately 5,000 rpm. This near coincidence indicates that the onset of high-speed fan operation is accompanied by a rapid energetic penalty rather than providing an unlimited extension of the low-power operating region. At still lower heat-rejection capability, no solution satisfies the thermal constraints within the imposed fan-speed range. Increasing fan speed alone therefore cannot compensate indefinitely for insufficient external heat rejection; adequate thermal coupling to the stratospheric environment is a prerequisite for feasible operation.

Internal airflow design provides an additional degree of freedom in reducing this auxiliary-power requirement. Figure 4c shows the minimum fan electrical power as the common duct cross-sectional area is varied relative to the baseline design. The dependence is non-monotonic: increasing the duct area initially reduces the required fan power, but continued enlargement eventually reverses this trend. The baseline duct geometry requires approximately 0.91% of the reference IT power for circulation, whereas a minimum of approximately 0.78% is obtained at a common duct-area scale of 1.5. This operating point corresponds to a fan electrical power of approximately 8.63 W for the 1.1-kW reference heat load. The existence of this minimum shows that simply enlarging the internal airflow passages does not monotonically improve thermal efficiency; instead, an intermediate duct size balances the competing changes in flow resistance and volumetric circulation.

We finally test whether the selected low-power operating point persists under thermal-design margins and adverse environmental conditions (Fig. 4d). The selected reference remains below 1% fan electrical overhead, and imposing additional thermal margins increases the required circulation power only modestly, to approximately 1% of the reference IT power for the evaluated margin range. Environmental perturbations produce a substantially wider response. Several operating corners remain within the low-power regime, whereas combinations of elevated ambient temperature and strong absorbed solar loading drive the required fan power above the 2% level. Under the most adverse evaluated conditions, including an ambient temperature of 243 K and 1.5 kW of absorbed solar heat with negligible radiative assistance, the required fan overhead exceeds 5% and approaches 8%. The cooling advantage is therefore robust over a broad, but explicitly bounded, operating envelope rather than being insensitive to the external thermal environment.

Together, these results show that fan-assisted internal circulation can support low-power thermal management when sufficient passive heat rejection to the stratospheric environment is available. Under the selected reference condition, the resulting fan-power requirement is less than 1%

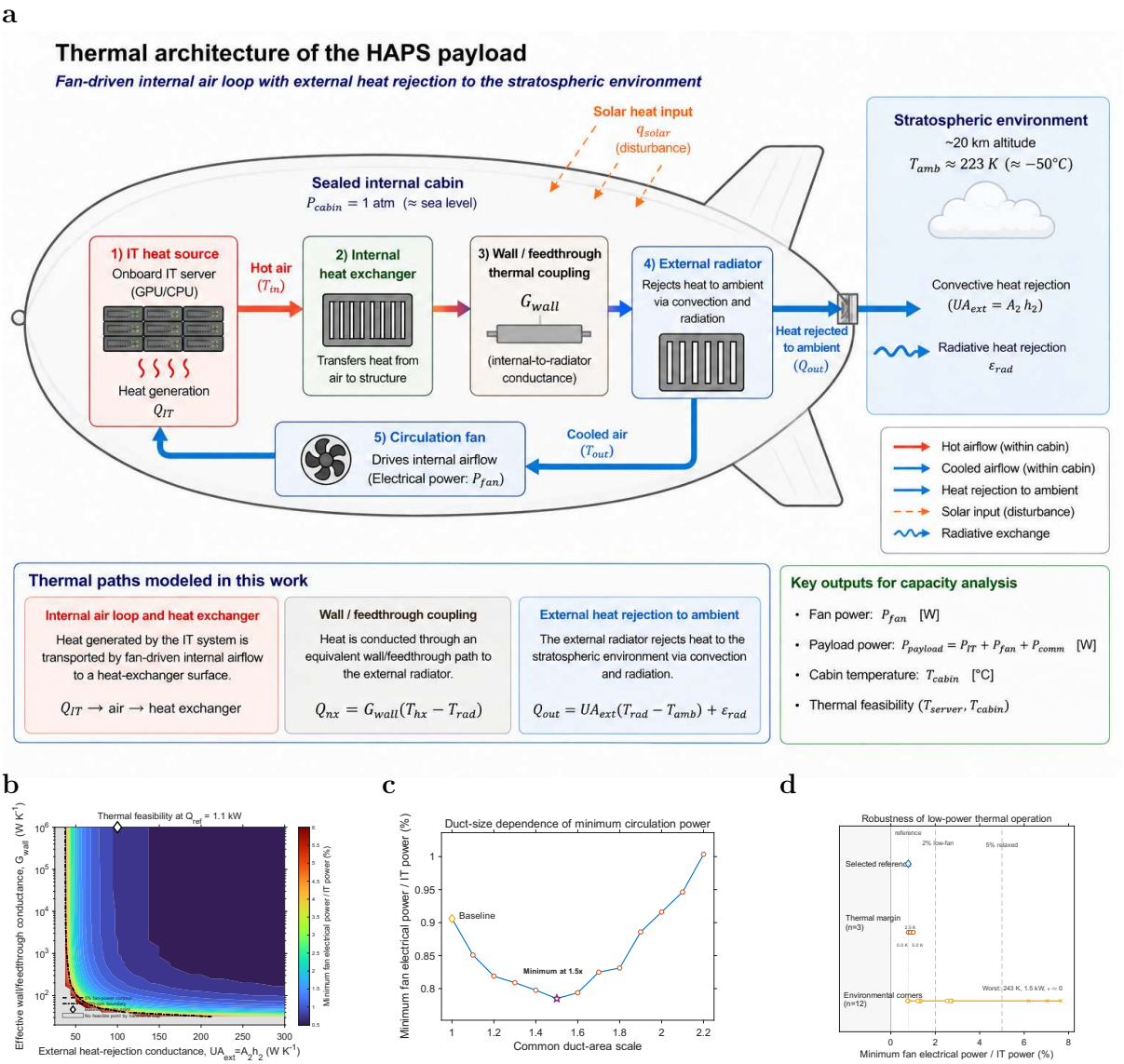


Fig. 4 Low-power thermal management emerges within a bounded heat-rejection envelope. a, Thermal architecture of the HAPS computing payload. **b,** Minimum fan-power requirement over the external heat-rejection and wall/feedthrough conductance space at the $Q_{\mathrm{IT,ref}} = 1.1$ kW heat load; lines indicate the 5% fan-power contour and 5000-rpm boundary. **c,** Duct-size dependence of minimum fan power, showing a minimum near 1.5× the baseline area. **d,** Sensitivity of minimum fan power to thermal-margin and environmental conditions.

of the IT heat-load reference, but the auxiliary-power advantage deteriorates rapidly near the heat-rejection boundary. We therefore use the selected $P_{\mathrm{fan}} \simeq 8.63$ W operating point as the representative thermal auxiliary-power contribution in the subsequent payload-level power analysis, while retaining the environmental-corner results as bounds on this reference assumption.

Radio circuitry, rather than radiated transmission, sets the communication-energy penalty.

We next quantified the communication power required for the HAPS aerial data centre to directly serve ground users. Using the reliability-aware user populations derived above, we first convert the user-equivalent capacity into simultaneous active-user demand and the corresponding

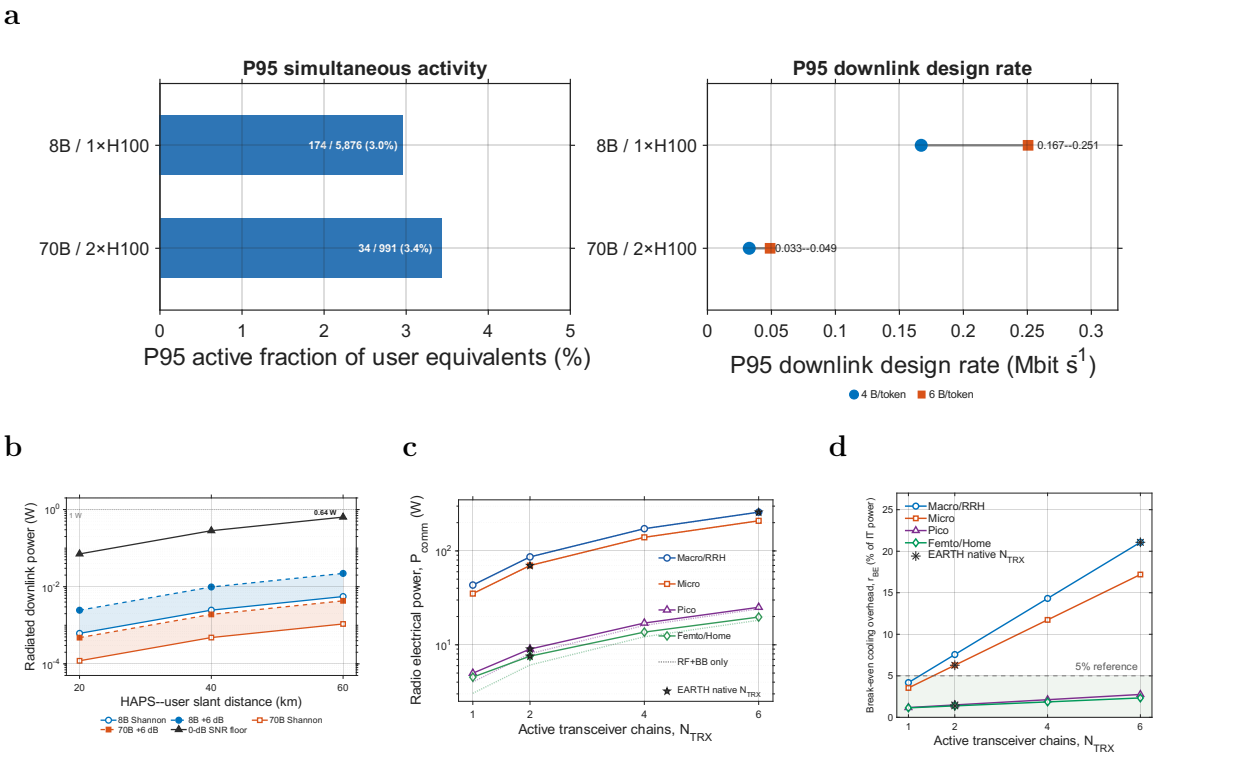


Fig. 5 Radio circuitry, rather than payload-bearing transmission, sets the communication-energy penalty. **a,** P95 simultaneous user activity and the corresponding downlink design rates for 4 and 6 bytes per token. **b,** Payload-bearing radiated downlink power versus HAPS–user slant distance under Shannon, +6-dB implementation-gap and 0-dB SNR-floor stress conditions. **c,** EARTH-derived radio electrical power versus the number of active transceiver chains. Dotted curves show the class-specific RF/baseband circuit-power contribution excluding the PA term, and stars indicate the native EARTH chain counts; EARTH cooling and main-supply losses are excluded. **d,** Break-even terrestrial cooling overhead for the evaluated radio architectures, normalized to the 1.275-kW IT-power reference of the 8B/1×H100 serving configuration. The fan contribution is scaled from the 1.1-kW thermal reference using the corresponding fan-to-IT-power ratio. The dashed line marks an illustrative 5% terrestrial cooling reference.

downlink traffic requirement. The communication analysis focuses on HAPS-to-user downlink transmission. User-equipment transmit power associated with the uplink is outside the HAPS payload-power boundary. Receive-side RF and baseband processing at the HAPS is represented through the fixed radio-circuit contribution. Uplink traffic volume and bidirectional scheduling are not explicitly modelled and are considered a limitation of the present analysis. Using the exact reliability-aware N_{95} operating points and the corresponding type-specific populations $N_{k,95}$, we convolve the five Binomial activity distributions to obtain the total active-user distribution and its 95th-percentile value. The resulting design traffic levels are summarized in Fig. 5a. For the representative compact-model and large-model serving units, only a small fraction of the user-equivalent population is simultaneously active under the 95^{th}-percentile activity condition, and the corresponding aggregate downlink rate remains in the sub-Mbit s^{-1} regime.

Using the downlink traffic requirement in Fig. 5a, we next evaluate the payload-bearing radiated power over representative HAPS–user slant ranges using the link-budget model described in the Methods section. Even under the evaluated implementation-gap and SNR-floor stress conditions, the required payload-bearing downlink power remained below 1 W across the considered distances (Fig. 5b). This outcome reflects the low aggregate bit rate [bps] of generated text responses under the evaluated activity model.

However, low radiated power does not imply negligible communication-energy overhead. Practical wireless service also requires radio-frequency

and baseband circuitry, including the power amplifier, transceiver chain, and digital processing blocks. When these circuit contributions are included, the communication power is no longer governed by the small radiated transmit power, but instead by the fixed radio hardware overhead, as shown in Fig. 5c. In other words, the communication-energy penalty of airborne deployment is set primarily by circuitry rather than by over-the-air transmission itself.

Finally, we compare the combined HAPS fan and communication power with a terrestrial cooling-power reference expressed as a fraction r_{cool} of IT power. Rather than assuming a universal terrestrial cooling fraction, Fig. 5d reports the break-even value required for auxiliary-power parity. Low-power Pico- and Femto-class radio references remain below a few per cent of IT power, whereas larger Micro- and Macro/RRH-class circuitry requires substantially higher terrestrial cooling overhead before parity is reached. The 5% line is shown only as an illustrative low-overhead reference.

Payload power converts serving efficiency into platform-scale HAPS capability

We finally combine the reliability-aware serving capacity with the power and mass requirements of the computing payload to translate serving-unit performance into platform-scale HAPS capability. Figure 6a compares the reliability-aware serving efficiency, N_{95}/P_{IT}, across the evaluated inference configurations. Within the 8B family, the 1×H200 configuration provides the highest efficiency, reaching approximately 5,682 user equivalents per kilowatt of IT power. The feasible 70B configurations achieve substantially lower efficiencies, spanning approximately 367–546 user equivalents kW^{-1}. These results show that accelerator count or nominal compute capability alone does not determine suitability for airborne deployment; the relevant metric is the reliable serving capacity delivered per unit of IT power.

The evaluated configurations occupy distinct power–mass regimes (Fig. 6b). Compact 8B configurations remain near the low-power, low-mass region, whereas multi-GPU 70B configurations require substantially greater payload resources. The shared-host L40S trajectory further illustrates that system architecture can modify the relationship between accelerator count and total payload burden. We therefore impose the publicly reported payload-resource envelopes of two representative HAPS platform concepts (Fig. 6c): 75 kg and 1.5 kW for the Sunglider-class platform and 250 kg and 5.0 kW for the Stratobus-class platform. Under the adopted IT, thermal-auxiliary and radio-power models, the Sunglider-class envelope admits only compact 8B deployments, whereas the larger Stratobus-class envelope accommodates multiple 8B serving instances and selected two-GPU 70B configurations. Payload power is the binding constraint for most of the evaluated configurations, while the Stratobus L40S case is instead capped by the adopted ten-GPU shared-host limit.

The benchmarked 70B/8×A100 BF16 TP8 configuration is excluded from the reliable-capacity comparison because Type E is already infeasible under the general latency SLO (Supplementary Table 3). Because accelerator generation, numerical precision and tensor-parallel degree differ simultaneously from the other evaluated 70B configurations, this observation is treated as configuration-specific rather than being attributed to any single hardware or precision factor.

For platforms containing multiple serving instances, platform-level reliability cannot be inferred simply by multiplying the single-unit N_{95} by the number of deployed instances. We therefore recompute the overload probability after pooling the available serving capacity and define $N_{\mathrm{HAPS},95}$ as the largest user-equivalent population for which the platform-level overload probability remains at or below 5% (Fig. 6d). Each serving instance retains its benchmark-derived type-specific capacity in this calculation; no additional compute acceleration from cross-instance GPU parallelism is assumed. Under the modelled independent-activity process, statistical multiplexing increases $N_{\mathrm{HAPS},95}$ above the linear mN_{95} reference in the evaluated multi-instance cases, with gains ranging from 6.5% to 68.4%. These gains therefore represent a reliability benefit from pooling stochastic demand rather than a superlinear GPU-compute speed-up.

Among the feasible configurations, 8B/1×H200 provides the largest platform-level capacity for both HAPS classes, yielding

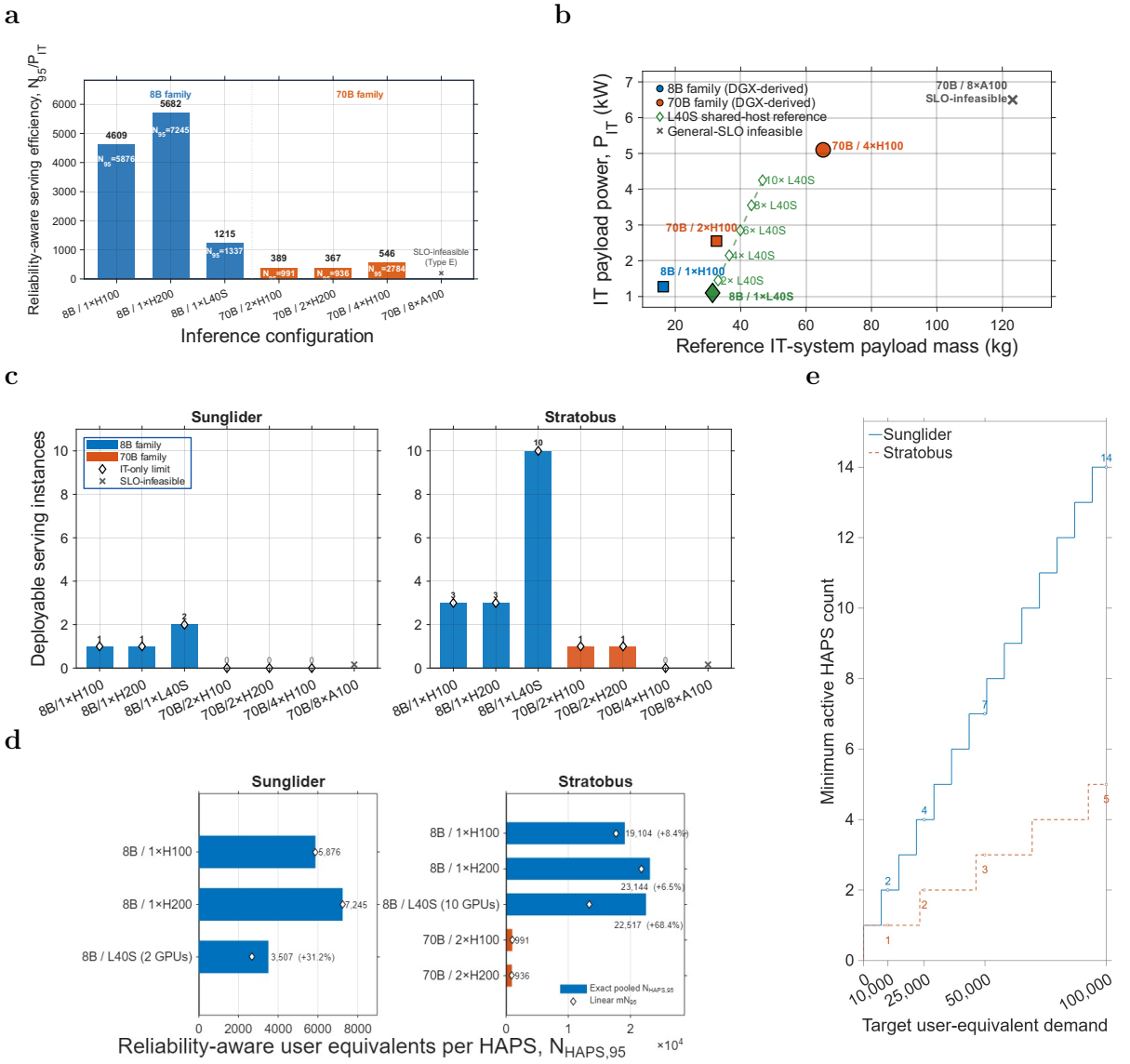


Fig. 6 Payload constraints convert serving efficiency into platform-scale HAPS capability. a, Reliability-aware serving efficiency N_{95}/P_{IT} across the evaluated inference configurations. **b,** Reference IT-payload power–mass regimes of the serving configurations, including shared-host L40S scaling. **c,** Deployable serving counts under the payload limits of Sunglider and Stratobus; diamonds indicate IT-only deployment limits. **d,** Platform-level reliability-aware capacity $N_{\mathrm{HAPS},95}$ after recomputing overload reliability for pooled serving instances; diamonds show the corresponding linear mN_{95} estimate. **e,** Minimum active-HAPS count required to serve a target user-equivalent demand using the highest-capacity feasible configuration on each platform, corresponding to 7,245 user equivalents per Sunglider-class platform and 23,144 user equivalents per Stratobus-class platform.

$N_{\mathrm{HAPS},95}$ = 7,245 and 23,144 user equivalents for the Sunglider- and Stratobus-class platforms, respectively. Figure 6e translates these capacities into the minimum number of concurrently active HAPS required for a target user-equivalent population. For example, a target of 100,000 user equivalents requires 14 Sunglider-class or 5 Stratobus-class platforms under the evaluated configuration. The staircase behaviour arises from the discrete nature of platform deployment, such that an improvement in per-platform capacity reduces the required fleet size only after the next integer deployment threshold is crossed.

Together, these results show that platform-scale HAPS capability is governed not by accelerator performance alone, but by the reliable serving efficiency achievable within a finite payload-resource envelope. Among the modelled constraints, payload power is the principal limitation on deployable serving capacity, while platform-level statistical pooling can recover part of the reliability margin when multiple serving instances share the available capacity.

Discussion

Our results support a conditional, rather than universal, feasibility claim for HAPS-based aerial data centres. By jointly accounting for trace-derived workload heterogeneity, interactive latency constraints, stochastic user activity, thermal management, radio circuitry and finite platform resources, we find reliability-aware capacities of 7,245 user equivalents for the highest-capacity feasible Sunglider-class configuration and 23,144 for the corresponding Stratobus-class configuration. These values satisfy the adopted 5% compute-overload criterion and represent modelled demand populations rather than numbers of continuously or simultaneously active users.

The workload analysis shows that serving capacity cannot be inferred from a representative request or accelerator benchmark alone. In particular, the rare long-context Type E workload has a disproportionately large serving impact because of its low SLO-constrained concurrency, especially for the evaluated 70B configuration. The resulting separation between N_{harm} and N_{95} demonstrates that mean-load dimensioning can leave insufficient headroom against stochastic variation in both the number and composition of active users. Within the evaluated configurations, the 8B family also provides substantially more reliability-aware user equivalents per unit IT power than the 70B family. This serving-efficiency comparison should not be interpreted as a model-quality comparison, because semantic accuracy, task utility, hallucination rate and other application-level quality metrics are outside the scope of this analysis.

The thermal, radio and platform results further show that useful HAPS capacity depends on subsystem co-design rather than accelerator choice alone. Thermal feasibility is governed primarily by the available external heat-rejection path; once this path becomes limiting, increasing fan speed cannot indefinitely recover feasibility. For the evaluated text-generation traffic, payload-bearing radiated power is small, whereas fixed or weakly load-dependent RF and baseband circuitry dominates the communication overhead[16]. At the platform level, payload power is the binding constraint for most evaluated configurations, so capacity can be improved not only through larger power allowances but also through higher reliability-aware serving efficiency and amortization of fixed host or radio overheads. These results characterize payload-level computing feasibility within the adopted HAPS resource envelopes, not whole-aircraft energy superiority over terrestrial infrastructure; likewise, the 5% terrestrial cooling ratio used in Fig. 5 is only an illustrative reference.

Multi-instance capacity also depends on how requests are assigned. Under idealised pooling, stochastic demand can be multiplexed across the aggregate serving budget, producing $N_{\mathrm{HAPS},95} > mN_{95}$ in the evaluated cases. In contrast, static balanced partitioning can yield capacity below the linear mN_{95} reference because overload of any instance causes platform-level failure. The pooled result should therefore be interpreted as an idealised dynamic-routing reference rather than an automatic gain from installing additional accelerators; indivisible-request placement and shared-resource contention are not explicitly represented. Although the exact numerical capacities vary across modelling choices, the main bottleneck hierarchy remains consistent across the evaluated sensitivities: long-context demand constrains large-model serving, heat-rejection capability sets the thermal boundary, fixed radio circuitry dominates at low traffic load, and payload power limits most platform configurations. Additional workload-mapping, session-construction, serving-capacity and request-assignment sensitivities are reported in the Supplementary Information. Large-sample Monte-Carlo calculations additionally reproduce the directly evaluated overload probabilities within finite-sampling uncertainty. The remaining Monte-Carlo threshold variation therefore reflects numerical estimation variability rather than seed dependence of the underlying stochastic activity model, as detailed in the Supplementary Information.

Several limitations define the scope of these conclusions. The platform analysis uses publicly reported payload-power and payload-mass allowances as boundary conditions rather than reconstructing a complete aircraft energy balance over seasonal, diurnal or meteorological conditions. The thermal system is represented by a lumped MATLAB/Simulink Simscape model described in the Methods. The module-level fan-to-IT-power ratio derived from the 1.1-kW reference operating point is recovered at 1.275 kW after resizing the heat-rejection path, but is not preserved when the full 2.55-kW load is imposed on a single thermal loop. Platform-level proportional scaling therefore represents thermal-system resizing or module replication rather than a configuration-specific validation of an unchanged cooling geometry. Payload-mass values are engineering allocation proxies and do not fully include thermal, radio, structural and qualification hardware. The radio model similarly uses terrestrial EARTH component values as a reference envelope because measurements of a HAPS-specific radio for this service class are not available[16]. Consequently, the reported platform capacities and fleet counts should be interpreted as compute-capacity estimates within the adopted payload envelopes rather than flight-qualified subsystem counts or geographical coverage requirements.

The communication and demand models introduce additional limitations. The radio analysis focuses on the HAPS-to-user response stream; uplink traffic volume, bidirectional scheduling, medium-access overhead and network-level coverage are not explicitly modelled. The demand model is derived from a single BurstGPT trace[14] and assumes independent user activity, so provider-specific usage patterns, temporal correlation and common demand bursts are not captured. In addition, the trace-derived occupancy and activity fractions partly inherit response times from the source serving system and are treated here as transferable workload descriptors when evaluating the target NIM configurations. The harmonic resource model combines homogeneous benchmark measurements through additive normalized costs and does not explicitly reproduce scheduler-, cache-, memory- or mixed-batch interference under a true heterogeneous workload.

Future work should therefore connect the present payload-level framework to higher-fidelity platform, thermal, bidirectional-radio and request-routing models, including experimental validation of the heat-rejection envelope and multi-HAPS workload coordination. Taken together, the results establish HAPS-based LLM inference as a cross-layer design problem rather than a question of accelerator performance alone. From a 6G NTN perspective, these results suggest a broader role for HAPS beyond wireless connectivity: a stratospheric platform can potentially operate as an integrated communication-and-computing node that serves selected AI workloads directly within the non-terrestrial segment. Within the bounded design space evaluated here, HAPS can therefore provide a finite and potentially useful aerial inference layer, while the remaining aircraft-level, network-level and implementation uncertainties define the path towards practical deployment.

Methods

BurstGPT trace processing and workload parameterization

We characterize text-generation demand using the BurstGPT serving trace[14]. Requests are ordered chronologically within each recorded session and paired with the subsequent request after removing invalid token-count or elapsed-time records. Requests without a usable Session ID are excluded from session-pair construction, and no session identity is imputed. The main analysis retains within-session pairs satisfying $0 \leq \Delta t_i \leq \tau$ with $\tau = 1800$ s, yielding 127,205 valid request pairs. The five representative workload types A–E are defined by the input/output-token profiles shown in Fig. 2a.

Two complementary mappings are used to derive the elapsed-time occupancy p_k and active-user fraction α_k. For occupancy assignment, input and output token counts are transformed as $x_i = \log(1 + n_i^{\mathrm{in}})$ and $y_i = \log(1 + n_i^{\mathrm{out}})$, and each retained pair is assigned to the nearest representative workload centre according to

$$k_i = \arg\min_k \left[\left(\frac{x_i - x_k^\star}{s_x} \right)^2 + \left(\frac{y_i - y_k^\star}{s_y} \right)^2 \right], \quad (1)$$

where $(x_k^\star, y_k^\star)$ is the corresponding log-transformed centre and s_x, s_y are the population standard deviations. The elapsed-time occupancy is then

$$p_k = \frac{\sum_{i:k_i=k} R_i}{\sum_i R_i}, \tag{2}$$

where R_i is the elapsed processing time of the initiating request.

The type-specific active-user fraction was estimated independently from local input/output-token windows around the five representative workloads: A, 100–400/80–400; B, 600–1400/600–1400; C, 100–1500/800–4000; D, 3000–8000/200–900; and E, $\geq 8000/\geq 280$ tokens. For consecutive requests, $Z_i = \max(\Delta t_i - R_i, 0)$ denotes the inferred inactive interval. With $\mathcal{W}_k$ denoting the retained pairs in activity window k,

$$\alpha_k = \frac{\sum_{i\in\mathcal{W}_k} R_i}{\sum_{i\in\mathcal{W}_k}(R_i + Z_i)}. \tag{3}$$

Thus, distance-based assignment provides an exhaustive partition for estimating p_k, whereas the local windows provide independent type-specific populations for estimating α_k.

Bootstrap uncertainty analysis

Sampling uncertainty in α_k was quantified using a non-parametric day-level block bootstrap. Valid request pairs are grouped by initiating day, and 1,000 bootstrap replicates are generated by resampling the observed day blocks with replacement and recomputing all five α_k values using the same $\tau = 1800$ s activity definition. The reported intervals in Fig. 2c are the 2.5th–97.5th percentiles of the bootstrap estimates. This procedure quantifies sampling variability in the inferred activity fractions and does not validate the independent-activity assumption used below.

SLO-aware serving-capacity extraction from NIM benchmarks

Workload-specific serving capacities were derived from concurrency-dependent time-to-first-token (TTFT) and inter-token latency (ITL) measurements in the NVIDIA NIM benchmark dataset[15]. For each configuration and workload type, TTFT and ITL are independently replaced by their non-decreasing envelopes to suppress isolated downward fluctuations, and SLO crossings between measured concurrency points are estimated by piecewise-linear interpolation. No extrapolation beyond the measured concurrency range is used. The SLO-constrained serving capacity is

$$C_{\max,k} = \min(C_{\mathrm{TTFT},k}, C_{\mathrm{ITL},k})\,. \tag{4}$$

The primary latency SLO and the additional sensitivity conditions are summarized in Table 1. If any workload type is infeasible under a selected SLO, the corresponding mixed-workload configuration is classified as SLO-infeasible.

We interpret $1/C_{\max,k}$ as the normalized serving-resource demand of an active request of type k. Because p_k represents the active workload mixture, $\sum_k p_k/C_{\max,k}$ gives the mean normalized resource demand per active request, and its reciprocal gives the nominal mixed-workload capacity. This aggregation is a weighted harmonic mean of the type-specific capacities, consistent with the use of harmonic averaging for rate-like quantities when their reciprocals have additive physical meaning[17]:

$$C_{\mathrm{mix}} = \left(\sum_k \frac{p_k}{C_{\max,k}}\right)^{-1}. \tag{5}$$

For Fig. 3c, the normalized serving-budget contribution is calculated as $b_k = (p_k/C_{\max,k})/\sum_j (p_j/C_{\max,j})$.

Reliability-aware user-equivalent capacity under stochastic activity

The nominal user-equivalent capacity associated with the harmonic operating point and its type composition are

$$N_{\mathrm{harm}} = C_{\mathrm{mix}} \sum_k \frac{p_k}{\alpha_k}, \tag{6}$$

$$q_k = \frac{p_k/\alpha_k}{\sum_j p_j/\alpha_j}. \tag{7}$$

This construction ensures that the expected composition of active user equivalents satisfies $q_k\alpha_k/\sum_j q_j\alpha_j = p_k$, thereby preserving the active-workload mixture used in the harmonic resource-budget model. This harmonic resource-budget model treats homogeneous-benchmark costs as additive and does not directly reproduce

scheduler, cache, memory or mixed-batch interference. A user equivalent denotes a modelled demand source rather than a uniquely identified subscriber or account.

For each candidate population N, integer type populations N_k are assigned to track q_k while satisfying $\sum_k N_k = N$. For an activity snapshot,

$$n_k(t) \sim \text{Binomial}(N_k, \alpha_k), \tag{8}$$

with independent activity across user-equivalent sources and workload types. The instantaneous normalized serving demand is

$$B(t) = \sum_k \frac{n_k(t)}{C_{\max,k}}, \tag{9}$$

and a snapshot is overloaded when $B(t) > 1$. The model-defined overload probability is evaluated directly from the independent Binomial activity distributions as

$$\begin{aligned} P_{\text{OL}}(N) &= \Pr\left[\sum_k \frac{n_k}{C_{\max,k}} > 1\right] \\ &= \sum_{\mathbf{n}:\sum_k n_k/C_{\max,k}>1} \prod_k \binom{N_k}{n_k} \alpha_k^{n_k}(1-\alpha_k)^{N_k-n_k}, \\ N_{95} &= \max\{N : P_{\text{OL}}(N) \le 0.05\}. \end{aligned} \tag{10}$$

The probability sum is evaluated deterministically; the numerical implementation conditions on four workload types and evaluates the remaining Binomial contribution through its tail probability to reduce the enumeration cost. Consequently, the reported N_{95} contains no Monte-Carlo sampling uncertainty for fixed p_k, α_k, $C_{\max,k}$ and integer population allocation. Monte-Carlo calculations are used only as an independent numerical validation in the Supplementary Information. The subscript 95 denotes the adopted 95% non-overload criterion rather than a statistical confidence interval.

Thermal model and minimum fan-power estimation

The airborne computing payload is evaluated using a MATLAB/Simulink Simscape model of a sealed internal air loop coupled to an external radiator (Fig. 4a). Heat is transferred from the IT payload through the internal heat exchanger and equivalent wall/feedthrough coupling G_{wall} to the external radiator, and rejected to the stratospheric environment by convection and radiation. The circulation fan is the only active thermal-management component included in the payload-power budget.

The principal heat-transfer relations are

$$\dot{Q}_{\text{conv}} = UA_{\text{ext}}(T_{\text{rad}} - T_{\text{amb}}), \tag{11}$$

$$\dot{Q}_{\text{cond}} = G_{\text{wall}}(T_A - T_B), G_{\text{wall}} = \frac{k_{\text{wall}}A_{\text{wall}}}{t_{\text{wall}}}, \tag{12}$$

$$\dot{Q}_{\text{rad}} = \epsilon_{\text{rad}}\sigma A_{\text{rad}}\left(T_{\text{rad}}^4 - T_{\text{env}}^4\right). \tag{13}$$

Here, $UA_{\text{ext}} = h_2 A_2$ denotes the effective external convective conductance, where A_2 is the effective heat-rejection area and h_2 is the corresponding external convective heat-transfer coefficient. The reference conditions are summarized in Table 1. The selected $G_{\text{wall}} = 10^6$ W K^{-1} represents a near-ideal equivalent coupling rather than a measured pressure-vessel-wall conductance[18,19].

For each thermal-design condition, fan speed is varied to identify the minimum tested electrical fan power satisfying the server and chamber temperature limits in Table 1. Steady operation is assessed from the terminal portion of each transient simulation, and

$$P_{\text{fan}} = \frac{|\overline{P}_{\text{shaft}}|}{\eta_{\text{motor}}}, \qquad r_{\text{fan}} = \frac{P_{\text{fan}}}{Q_{\text{IT,ref}}}. \tag{14}$$

The feasibility analysis varies UA_{ext}, G_{wall} and the internal duct-area scale, with additional perturbations in ambient temperature, absorbed solar heat and radiative properties. The 1.5× duct-area case is selected as the reference operating point used in the subsequent payload-level analysis.

To assess whether the reference fan-to-IT-power ratio remains representative at higher serving loads, we additionally evaluate 1.275- and 2.55-kW heat loads. At 1.275 kW, modest resizing of the heat-rejection path recovers a fan-to-IT-power ratio comparable to the 1.1-kW reference, whereas a single 2.55-kW thermal loop requires substantially stronger thermal transport and exhibits higher fan overhead. Accordingly, the reference ratio used in the platform-level analysis is interpreted as a module-level design assumption

Table 1 Key reference parameters used in the cross-layer analysis. Detailed implementation constants and complete sweep grids are retained in the archived code and source-data files.

Category	Parameter	Main value or range
Workload	Session-continuity threshold	$\tau = 1800$ s; sensitivity at 600 and 3600 s
Bootstrap	Day-level replicates	1,000
Serving	Primary latency SLO	TTFT ≤ 2000 ms; ITL ≤ 100 ms
Serving	Sensitivity SLOs	$(1000, 100)$, $(2000, 40)$, $(1000, 40)$ and $(500, 40)$ ms for (TTFT, ITL)
Reliability	Overload criterion	$P_{\mathrm{OL}} \leq 0.05$; direct evaluation of the independent Binomial activity model
Thermal	Reference operating point	$Q_{\mathrm{IT,ref}} = 1.1$ kW; $T_{\mathrm{amb}} = 223.15$ K; $UA_{\mathrm{ext}} = 100$ W K^{-1}; $G_{\mathrm{wall}} = 10^6$ W K^{-1}; absorbed solar load 500 W
Thermal	Feasibility limits	$T_{\mathrm{server}} \leq 330$ K; $T_{\mathrm{chamber}} \leq 300$ K; $\eta_{\mathrm{motor}} = 0.85$
Traffic	Response-rate conversion	$T_{\mathrm{ITL}} = 100$ ms; $b = 4$–6 bytes/token; $\lambda_{\mathrm{OH}} = 3$
Radio link	Reference link budget	2 GHz carrier; 20 MHz bandwidth; 20/40/60 km slant ranges; 20/0 dBi HAPS/UE gains; 7 dB noise figure; 3 dB miscellaneous loss; 5 dB link margin
Radio link	Stress conditions	0/3/6 dB implementation gap and $-10/-5/0$ dB receiver-SNR floors
Platform	Sunglider-class envelope	75 kg and 1.5 kW payload mass/power
Platform	Stratobus-class envelope	250 kg and 5.0 kW payload mass/power

that requires thermal-system resizing or replication as serving instances are added; the higher-load sensitivity is reported in the Supplementary Information.

Reliability-aware downlink traffic and link-budget model

The communication analysis uses the reported reliability-aware population rather than assuming that all user equivalents generate traffic simultaneously. For each configuration, the total active-user distribution is obtained by convolving the five type-specific Binomial activity distributions, and $N_{\mathrm{act},95}$ is defined as its 95th-percentile active-user count.

For active-user quantile u, the SLO-implied downlink rate is

$$R_{\mathrm{DL},u} = 8\lambda_{\mathrm{OH}} b \frac{1000}{T_{\mathrm{ITL}}} N_{\mathrm{act},u}. \tag{15}$$

The traffic-conversion and link-budget parameters are summarized in Table 1. The P95 rate is mapped to payload-bearing radiated power using a free-space single-link budget,

$$P_{\mathrm{TX,dBm}} = P_{\mathrm{N,dBm}} + \gamma_{\mathrm{req,dB}} + L_{\mathrm{FSPL,dB}} + L_{\mathrm{misc,dB}} - G_{\mathrm{HAPS,dBi}} - G_{\mathrm{UE,dBi}} + M_{\mathrm{link,dB}}, \tag{16}$$

where the required receiver SNR was obtained from the Shannon spectral-efficiency relation subject to the stress conditions in Table 1.

The resulting P_{TX} is the payload-bearing downlink requirement rather than the complete radiated power of a fully specified cellular air interface. Uplink traffic volume, bidirectional scheduling, control signalling and network-level coverage are not explicitly modelled. The 5% compute-overload criterion and P95 traffic design point are separate marginal criteria and do not establish a joint 95% end-to-end reliability guarantee.

Radio electrical-power model

Radio electrical power is estimated using the EARTH component-level base-station model[16] as a reference envelope for the power amplifier, RF transceiver, baseband processing and DC–DC conversion. EARTH cooling power is excluded because thermal management is modelled separately, and terrestrial AC-main-supply

losses were excluded from the adopted HAPS payload boundary. The Macro, RRH, Micro, Pico and Femto/Home classes are evaluated over the active-chain counts shown in Fig. 5c; the 10-MHz EARTH baseband term is scaled linearly to the 20-MHz main bandwidth.

The onboard radio electrical power is calculated as

$$P_{\mathrm{comm}} = \frac{P_{\mathrm{TX}}/\eta_{\mathrm{PA}} + N_{\mathrm{TRX}}\left[P_{\mathrm{RF}} + P_{\mathrm{BB,ref}} W/W_{\mathrm{ref}}\right]}{1-\ell_{\mathrm{DC}}}. \tag{17}$$

The EARTH values are terrestrial reference components rather than measurements of a HAPS-specific radio. For the platform-level power budget, the selected one-chain Pico-like reference under the P95 traffic stress case gives $P_{\mathrm{comm}} = 4.983$ W.

For the auxiliary-power comparison in Fig. 5d, the break-even terrestrial cooling fraction is

$$r_{\mathrm{BE}} = \frac{P_{\mathrm{fan}} + P_{\mathrm{comm}}}{P_{\mathrm{IT}}}. \tag{18}$$

The comparison uses the 1.275-kW system-equivalent IT-power reference of the 8B/1×H100 configuration. For consistency with the module-level thermal scaling adopted in the platform analysis, the plotted fan contribution ($P_{\mathrm{fan}} = 10.008$ W) is obtained from the selected 1.1-kW fan-to-IT-power ratio. A separate 1.275-kW thermal sensitivity recovered a nearly identical low-overhead operating point after resizing the heat-rejection path, as reported in the Supplementary Information. The 5% terrestrial cooling line in Fig. 5d is an illustrative reference rather than a universal cooling fraction.

HAPS payload integration, reliability pooling and fleet sizing

For each inference configuration j, the reliability-aware capacity $N_{95,j}$ is combined with system-equivalent IT power and payload-mass allocations. The power-normalized serving efficiency is $\eta_{\mathrm{serve},j} = N_{95,j}/P_{\mathrm{IT},j}$.

For H100/H200 and A100 configurations, the maximum power and mass of the corresponding eight-GPU DGX systems are allocated uniformly across the installed GPUs[20,21], giving 1.275 kW and 16.306 kg per H100/H200-equivalent GPU and 0.8125 kW and 15.395 kg per A100-equivalent GPU. The L40S model used a shared-host reference[22], comprising 29.7 kg and 0.75 kW for the host. Each installed L40S accelerator is assigned 1.7 kg based on an HPE product specification[23] and 0.35 kW based on the NVIDIA maximum board-power specification[24], with the configuration capped at ten GPUs. These are engineering system-equivalent allocations rather than bare-card masses or measured inference power.

For a candidate number m of serving instances, payload feasibility requires

$$\begin{aligned} P_{\mathrm{IT},j}(m) + P_{\mathrm{fan},j}(m) + P_{\mathrm{comm}} &\le P_{\mathrm{lim}}, \\ M_j(m) &\le M_{\mathrm{lim}}. \end{aligned} \tag{19}$$

The platform envelopes are given in Table 1. Fan power is scaled in proportion to aggregate IT power using the selected module-level thermal fan-to-IT ratio, whereas the 4.983-W radio reserve is included once per HAPS. This scaling represents thermal-system resizing or module replication rather than unchanged-geometry scaling of a single thermal loop. Aircraft propulsion, station keeping, avionics and energy storage are treated as upstream platform functions associated with providing the reported payload envelope.

For m serving instances, pooled type-specific serving capacity is modelled as $C_{\mathrm{max,HAPS},k} = mC_{\mathrm{max},k}$. The same Binomial activity model is then applied and the platform-level overload probability is evaluated directly to obtain $N_{\mathrm{HAPS},95}$. The pooled model assumes that type-specific capacity can be shared across serving instances and does not assume additional compute acceleration from cross-instance GPU parallelism. A static balanced-partition sensitivity instead distributed each workload type as evenly as possible across instances. The overload probability of each instance was evaluated exactly under the same Binomial activity model. Because the assigned user-equivalent populations are disjoint under the adopted independent-activity assumption, the platform-level static overload probability is calculated as $1 - \prod_i (1 - p_i)$, where p_i is the overload probability of serving instance i.

Finally, the minimum concurrently active HAPS count for target demand is

$$H_{\text{active}} = \left\lceil \frac{N_{\text{target}}}{N^{\star}_{\text{HAPS},95}} \right\rceil, \tag{20}$$

where $N^{\star}_{\text{HAPS},95}$ is the largest reliability-aware platform capacity among the feasible configurations for the corresponding HAPS class. This quantity is a compute-capacity dimensioning rule and does not include geographical coverage or impose a fleet-wide overload-probability guarantee.

Additional robustness analyses vary workload mapping, session threshold, serving-capacity interpolation, thermal conditions and multi-instance request assignment; the corresponding numerical results are reported in the Supplementary Information. The exact-probability search grids, Monte-Carlo validation seeds, block-level Simscape constants, component tables and complete stress-case grids are retained in the archived code and source-data files.

Data availability

The BurstGPT workload trace used in this study is publicly available from the original BurstGPT repository and is cited in the main text. The NVIDIA NIM latency–concurrency measurements and the public HAPS and hardware specifications used as model inputs are available from the original sources cited in the manuscript.

The processed workload statistics, serving-capacity tables, exact reliability outputs, Monte-Carlo validation outputs, thermal-simulation summaries, communication-model outputs, platform-level results, and source data underlying the figures and tables generated in this study are available at Zenodo[25].

Code availability

The custom MATLAB scripts and MATLAB/Simulink Simscape model used for workload processing, SLO-constrained serving-capacity extraction, exact Binomial reliability analysis, Monte-Carlo validation, thermal simulation, communication link-budget and radio-power calculations, and HAPS payload integration are available at Zenodo[25] and GitHub (https://github.com/BHlee1013/haps-aerial-data-center). The archived Zenodo record corresponds to release v1.0.0 used to generate the results reported in this article.

The repository also provides the numerical inputs, exact-probability outputs, Monte-Carlo validation seeds, and processed source data required to reproduce the reported figures and tables. Software-version and toolbox requirements are documented in the repository.

Author contributions

B.L. contributed to conceptualization, methodology, software, validation, formal analysis, data curation, visualization, and writing of the original draft. Y.-C.K. contributed to conceptualization, supervision, interpretation of the results, and review and editing of the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Yang, L., Deng, Y., Yang, L. T. & Lin, R. Reducing the cooling power of data centers by intelligently assigning tasks. *IEEE Internet of Things Journal* **5**, 1667–1678 (2018).

[2] Mytton, D. Data centre water consumption. *npj Clean Water* **4**, 11 (2021).

[3] Periola, A., Alonge, A. & Ogudo, K. Heat wave resilient systems architecture for underwater data centers. *Scientific Reports* **12**, 17161 (2022).

[4] Ikeda, H. *et al.* Digital twin technology in data center simulations: Evaluating the feasibility of a former mine site. *Sustainability* **15**, 16176 (2023).

[5] Yin, Z. *et al.* A comprehensive survey of orbital edge computing: Systems, applications, and algorithms. *Chinese Journal of Aeronautics* (2025).

[6] Al Ahmad, M., Memon, Q. & Pecht, M. Reliability and risk in space-based data centers: A lifecycle assessment of orbital cloud infrastructure. *Applied Sciences* **16**, 5247 (2026).

[7] Karabulut Kurt, G. *et al.* A vision and framework for the high altitude platform station (haps) networks of the future. *IEEE Communications Surveys & Tutorials* **23**, 729–779 (2021).

[8] Baltaci, A. *et al.* A survey of wireless networks for future aerial communications (facom). *IEEE Communications Surveys & Tutorials* **23**, 2833–2884 (2021).

[9] Pham, Q.-V. *et al.* Aerial computing: A new computing paradigm, applications, and challenges. *IEEE Internet of Things Journal* **9**, 8339–8363 (2022).

[10] Abderrahim, W., Amin, O. & Shihada, B. How to leverage high altitude platforms in green computing? *IEEE Communications Magazine* **61**, 134–140 (2023).

[11] Abderrahim, W., Amin, O. & Shihada, B. Data center-enabled high altitude platforms: A green computing alternative. *IEEE Transactions on Mobile Computing* **23**, 6149–6162 (2024).

[12] Fukumoto, S. Mobile connectivity with HIBS. ITU Radiocommunication Seminar 2020 for Asia-Pacific Region, https://www.itu.int/en/ITU-R/seminars/rrs/2020-Asia-Pacific/Forum/Session%205_IMT%20and%20BB/SOFTBANK_HIBS_RRS20APC.pdf (2020). SoftBank Corp. and HAPSMobile Inc., accessed August 2026.

[13] Thales Alenia Space. Thales alenia space signs contract with european commission and announces kickoff of EuroHAPS project for the demonstration of stratospheric platforms. https://www.thalesaleniaspace.com/en/press-releases/thales-alenia-space-signs-contract-european-commission-and-announces-kickoff (2023). Press release, accessed August 2026.

[14] Wang, Y. *et al.* *BurstGPT: A real-world workload dataset to optimize LLM serving systems*, Vol. 2, 5831–5841 (Association for Computing Machinery, 2025).

[15] NVIDIA. NVIDIA NIM benchmarking: Large language models performance. https://docs.nvidia.com/nim/benchmarking/llm/1.0.0/performance.html. Documentation release 1.0.0; page includes NIM container 1.8.0 results. Accessed 11 September 2026.

[16] Auer, G. *et al.* How much energy is needed to run a wireless network? *IEEE Wireless Communications* **18**, 40–49 (2011).

[17] Jain, R. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling* (John Wiley & Sons, New York, 1991).

[18] Bergman, T. L., Lavine, A. S., Incropera, F. P. & DeWitt, D. P. *Fundamentals of Heat and Mass Transfer* 8 edn (John Wiley & Sons, Hoboken, NJ, 2017).

[19] Siegel, R. & Howell, J. R. *Thermal Radiation Heat Transfer* 4 edn (Taylor & Francis, New York, 2002).

[20] NVIDIA. Introduction to the NVIDIA DGX H100/H200 system. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html. System specifications. Accessed 11 September 2026.

[21] NVIDIA. Introduction to the NVIDIA DGX A100 system. https://docs.nvidia.com/dgx/dgxa100-user-guide/introduction-to-dgxa100.html. System specifications. Accessed 11 September 2026.

[22] Supermicro. SYS-421GE-TNRT product specifications. https://www.supermicro.com/zh-tw/products/system/gpu/4u/sys-421ge-tnrt. Shared-host specification. Accessed 11 September 2026.

[23] Hewlett Packard Enterprise. NVIDIA L40S 48GB PCIe Accelerator Data Sheet. https://www.hpe.com/psnow/doc/PSN1014786

966USEN. Product specifications report a weight of 1.7 kg. Accessed 22 September 2026.

[24] NVIDIA. L40S gpu specifications. https://www.nvidia.com/es-es/data-center/l40s/. Maximum board-power specification. Accessed 11 September 2026.

[25] Lee, B. & Ko, Y.-C. Direct wireless large language model inference from high-altitude platform stations under payload constraints: reproduction code and data (2026). URL https://doi.org/10.5281/zenodo.23074651. Version 1.0.0, Zenodo.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SIADCBH.pdf