Supplementary Information

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

Byeongheon Lee and Young-Chai Ko

Korea University

Supplementary Results

Workload statistics and session-threshold sensitivity

The main Methods define the BurstGPT request-pair construction, the exhaustive distance-based occupancy mapping, the independent range-window estimator for the type-specific active-user fractions, and the day-level bootstrap procedure. This section reports the corresponding baseline workload statistics, bootstrap intervals and session-threshold sensitivity results; it does not introduce an alternative preprocessing, mapping or uncertainty-estimation method.

Supplementary Table 1: Baseline workload statistics and day-bootstrap activity intervals.

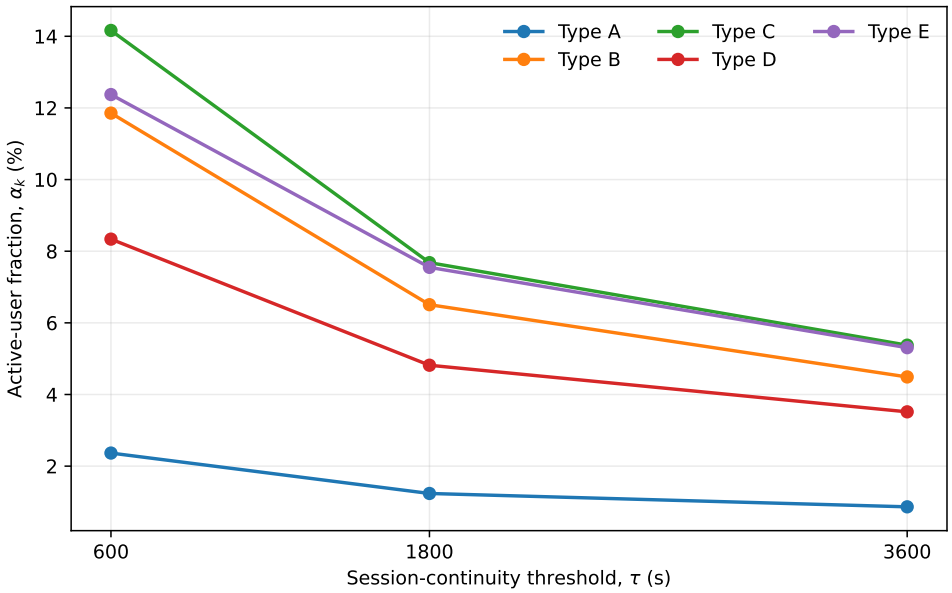
Type	Distance count	p_k (%)	α_k (%)	95% interval (%)	Range pairs
A	75,786	31.970	1.237	1.172–1.322	13,543
B	17,858	29.151	6.507	5.939–7.255	3,226
C	674	2.239	7.681	7.029–8.407	1,716
D	32,550	35.295	4.817	4.505–5.199	6,241
E	337	1.345	7.549	6.603–8.745	505

Distance counts sum to 127,205 valid request pairs. The range counts are local-window counts and need not sum to the distance population or be mutually exclusive. Intervals describe bootstrap sampling variability, not a 95% service reliability criterion. Values are transcribed from the final Fig. 2b–c result CSVs.

Supplementary Table 2: Activity estimates at the evaluated session-continuity thresholds.

τ (s)	α_A (%)	α_B (%)	α_C (%)	α_D (%)	α_E (%)
600	2.365	11.855	14.163	8.337	12.373
1800	1.237	6.507	7.681	4.817	7.549
3600	0.864	4.490	5.377	3.516	5.307

These are range/no-priority recheck estimates from the supplied threshold-sensitivity table. They characterize sensitivity of the activity estimator only; downstream N_{95} values at 600 and 3600 s are not inferred from this table.



Supplementary Fig. 1: Sensitivity of the inferred active-user fractions to the session-continuity threshold. The range-based activity estimates are shown for $\tau = 600$, 1800 and 3600 s. The threshold changes the absolute activity fractions because longer inter-request intervals are retained as candidate inactive periods, but the separation between the low-activity Type A population and the higher-activity workload classes remains visible. This figure characterizes the activity estimator only; it is not a downstream N_{95} sensitivity calculation.

Serving-capacity and interpolation sensitivity

The main Methods define the SLO-constrained capacity extraction, baseline piecewise-linear crossing and the alternative interpolation rules. This section reports the corresponding configuration-specific capacity equivalents and interpolation sensitivity; it introduces no additional serving-capacity extraction procedure. Configuration labels, numerical precision and tensor-parallel degree are retained as reported in the archived benchmark registry. In particular, the evaluated 70B/8×A100 configuration uses BF16 TP8, whereas the other configurations listed below use FP8; these configurations therefore should not be interpreted as a controlled comparison of accelerator generation or precision alone.

Supplementary Table 3: General-SLO capacity equivalents for the seven configuration labels in the benchmark registry.

Configuration	Precision / TP	$C_{\max,A}$	$C_{\max,B}$	$C_{\max,C}$	$C_{\max,D}$	$C_{\max,E}$
8B / 1×H100	FP8 / 1	250.000	250.000	203.435	164.019	18.961
8B / 1×H200	FP8 / 1	250.000	233.530	172.545	250.000	27.686
8B / 1×L40S	FP8 / 1	250.000	213.313	25.000	57.168	3.013
70B / 2×H100	FP8 / 2	250.000	224.507	168.631	58.137	1.620
70B / 2×H200	FP8 / 2	250.000	250.000	250.000	61.920	1.393
70B / 4×H100	FP8 / 4	250.000	250.000	250.000	175.575	4.386
70B / 8×A100	BF16 / 8	250.000	250.000	150.000	63.810	0.000

Non-integer entries are interpolated capacity equivalents, not literal fractional concurrent requests. Values capped at the largest tested concurrency do not prove that this is the hardware's absolute limit. The 70B/8×A100 BF16 TP8 case is infeasible for Type E under the general SLO. Configuration labels are retained from the archived source-data table; a versioned source-to-row NIM provenance record is provided in the archived source data.

To keep interpolation sensitivity separate from the downstream reliability calculation, we propagate

each interpolation rule only through the deterministic nominal-capacity stage. Accordingly, Supplementary Table 4 reports the resulting $C_{\max,E}$, C_{mix} and N_{harm} values; interpolation-dependent N_{95} values are not used as an additional reported sensitivity result.

Supplementary Table 4: Interpolation sensitivity of the Type-E capacity, nominal mixed active capacity and nominal harmonic user-equivalent capacity under the general SLO.

Configuration	Interpolation	$C_{\max,E}$	C_{mix}	N_{harm}
8B / 1×H100	linear	18.961	184.628	7,038
8B / 1×H100	lower_step	5.000	131.571	5,015
8B / 1×H100	pchip	20.170	191.670	7,306
8B / 1×H100	log_linear	15.378	178.552	6,806
70B / 2×H100	linear	1.620	58.529	2,231
70B / 2×H100	lower_step	1.000	42.739	1,629
70B / 2×H100	pchip	1.672	63.281	2,412
70B / 2×H100	log_linear	1.283	51.260	1,954

$C_{\max,E}$ is taken from the supplied interpolation table; C_{mix} is recalculated algebraically from all five stored capacities and the unchanged baseline p_k. N_{harm} is then obtained deterministically from the baseline p_k and α_k using the main-text harmonic population model and is rounded to the nearest user equivalent. The interpolation sensitivity is intentionally reported only through N_{harm} so that it remains distinct from the reliability-threshold calculation and its separate numerical validation.

Thermal robustness, higher-load scaling and validation checks

The main Methods define the thermal-network topology, reference parameterisation, fan-speed search and terminal convergence criteria. This section reports supporting thermal results only and introduces no additional thermal-model or search procedure.

For the common duct-area sweep, the selected factor of 1.5 relative to the baseline geometry gives 2675 rpm and $P_{\mathrm{fan}} = 8.634443$ W, corresponding to 0.784949% of the 1100-W IT reference. The converged server and chamber means are approximately 322.340 and 299.496 K, respectively. Thermal-margin tests that tighten both temperature limits by 0, 2.5 and 5 K give archived electrical fan powers of 8.634, 9.698 and 11.141 W, respectively. Environmental-corner results at the selected duct-area design are listed in Supplementary Table 7.

To test whether the selected fan-to-IT-power ratio can be extended beyond the 1.1-kW reference point, we additionally evaluated higher heat loads using the same steady-state temperature limits and fan-power definition. The principal results are summarized in Supplementary Table 5. Retaining the selected 1.5-times duct geometry while increasing the IT heat load to 1.275 kW increased the minimum fan power to 14.414 W, or 1.130% of the IT heat load. Increasing the external heat-rejection conductance from 100 to 175 W K^{-1} then reduced the required fan power to 9.976 W (0.782%), recovering a fan-to-IT-power ratio essentially equal to the 1.1-kW reference. Thus, within this single-module load range, modest resizing of the external heat-rejection path was sufficient to preserve the low-overhead regime.

At 2.55 kW, the unchanged thermal geometry did not recover a feasible operating point within the 12,000-rpm hard search cap. The internal server-to-air path explains the first part of this scaling penalty. With $h_1 = 120$ W m^{-2} K^{-1} and $A_1 = 0.40$ m^2, the reference internal conductance is $UA_{\mathrm{int}} = h_1 A_1 = 48$ W K^{-1}. At a 2.55-kW heat load, the corresponding lumped temperature rise is approximately $2550/48 = 53.1$ K. If the chamber is allowed to approach its 300-K limit while the server is constrained to 330 K, the same lumped relation requires at least

$$UA_{\mathrm{int}} \gtrsim \frac{2550}{330 - 300} = 85 \text{ W K}^{-1}.$$

Supplementary Table 5: Higher-load thermal scaling relative to the selected 1.1-kW reference design.

Case	Q_{IT} (kW)	UA_{ext} (W K^{-1})	A_1 scale	Duct scale	Minimum rpm	P_{fan} (W)	Fan / IT (%)
Selected reference	1.100	100	1.00	1.50	2,675	8.634	0.785
Fixed geometry	1.275	100	1.00	1.50	3,150	14.414	1.130
Resized rejection path	1.275	175	1.00	1.50	2,800	9.976	0.782
Fixed geometry	2.550	100	1.00	1.50	–	–	–
Best evaluated co-design	2.550	1,000	2.25	1.25	5,125	62.64	2.456

A_1 denotes the effective internal heat-transfer area and its scale is relative to the 0.40-m^2 reference. The fixed 2.55-kW geometry did not satisfy both $T_{server} \leq 330$ K and $T_{chamber} \leq 300$ K within the 12,000-rpm hard search cap. The 5,125-rpm co-design point lies just above the preferred 5,000-rpm range and is therefore interpreted as a thermal-envelope result rather than a flight-qualified fan operating point.

Doubling the internal heat-transfer area raises this conductance to 96 W K^{-1}, and the co-design sweep showed little additional reduction in minimum fan overhead once the area scale reached approximately two. The dominant limitation therefore shifted from the server-to-air path to the chamber and external heat-rejection path.

We then extended the 2.55-kW co-design over internal-area scales of 1.75, 2.0 and 2.25, duct-area scales of 1.0, 1.25, 1.5 and 1.75, and UA_{ext} = 300–1000 W K^{-1}, giving 84 evaluated design combinations. Increasing external heat-rejection capability reduced both the minimum required fan speed and the fan-power fraction, but with diminishing returns (Supplementary Table 6). The best point in this sweep used $A_1 = 0.90$ m^2 (2.25 times the reference), a 1.25-times duct scale and $UA_{ext} = 1000$ W K^{-1}. It required 5,125 rpm and 62.64 W of electrical fan power, corresponding to 2.456% of the 2.55-kW IT load, with converged server and chamber mean temperatures of approximately 323.50 and 299.89 K. No evaluated 2.55-kW design reached a fan-power fraction of 2% or less, and none satisfied the thermal constraints within the preferred 5,000-rpm range.

Supplementary Table 6: Best 2.55-kW co-design envelope as the external heat-rejection conductance is increased.

UA_{ext} (W K^{-1})	Equivalent A_2 at $h_2 = 50$ (m^2)	Best rpm	Best fan / IT (%)
300	6	5,800	3.56
350	7	5,650	3.29
400	8	5,550	3.12
500	10	5,400	2.87
600	12	5,300	2.72
750	15	5,225	2.60
1,000	20	5,125	2.456

Each row reports the lowest-overhead point among the evaluated internal-area and duct-area combinations at the specified UA_{ext}. The equivalent A_2 values follow from the fixed $h_2 = 50$ W m^{-2} K^{-1} used in this sweep. Large-area points are design-envelope sensitivities and should not be interpreted as flight-integrated radiator layouts; radiator mass, volume and structural integration are not included.

These higher-load results motivate the module-level interpretation used for platform power accounting. They do not show that the 0.785% ratio is preserved when a full 2.55-kW heat load is imposed on one unchanged thermal loop. Rather, the 1.275-kW result shows that a low-overhead module can be recovered by resizing its rejection path. If two independent 1.275-kW thermal modules were replicated, each operating at the evaluated 9.976-W point, their aggregate fan power would be approximately 19.95 W for 2.55 kW of IT load, preserving the 0.782% ratio arithmetically. This is a system-integration assumption, not an integrated two-module thermal validation: shared-radiator

interactions, additional heat-exchanger and radiator mass, packaging volume, structural coupling and common-flow contention are not included. Accordingly, proportional fan scaling in the main platform model should be interpreted as thermal-system resizing or module replication rather than fixed-geometry scaling of a single thermal loop.

Supplementary Table 7: Environmental-corner results at the selected 1.5-times duct-area design.

T_{amb} (K)	Q_{solar} (W)	ϵ_{rad}	Speed (rpm)	P_{fan} (W)	Fraction (%)
223.15	500	off	2,675	8.634	0.785
223.15	500	0.3	2,650	8.382	0.762
223.15	500	0.9	2,650	8.383	0.762
223.15	1,500	off	3,175	14.765	1.342
223.15	1,500	0.3	3,125	14.052	1.277
223.15	1,500	0.9	3,075	13.363	1.215
243.15	500	off	4,000	30.165	2.742
243.15	500	0.3	3,975	29.589	2.690
243.15	500	0.9	3,900	27.905	2.537
243.15	1,500	off	5,600	84.075	7.643
243.15	1,500	0.3	5,450	77.468	7.043
243.15	1,500	0.9	5,225	68.198	6.200

"off" denotes $\epsilon_{\mathrm{rad}} = 10^{-15}$. Points above 5000 rpm are model extrapolations, not validated operating points of a flight-qualified fan. The ratio is normalised by 1100 W.

The warm/high-solar cases require substantially higher fan power than the selected reference and the points above 5000 rpm should be interpreted as model extrapolations. Motor losses were included in the electrical-power accounting but were not injected as an additional thermal source. At the selected reference, the electrical-to-shaft power difference is approximately 1.295 W (0.118% of the IT heat load), whereas this omission becomes less conservative as fan power increases.

The archived nominal-duration and doubled-duration checks are listed in Supplementary Table 8. The adverse case moves by one 25-rpm grid increment and approximately 1.39% in electrical fan power; the corresponding simulations therefore do not support a claim of exactly identical thresholds for every duration check.

Supplementary Table 8: Nominal-duration and doubled-duration thermal checks.

Case	rpm (1x)	rpm (2x)	Power (1x) (W)	Power (2x) (W)	Change (%)
baseline	2,675	2,675	8.63444	8.63404	-0.005
adverse environment	5,450	5,475	77.46751	78.54500	1.391
near boundary	3,625	3,625	22.27124	22.27078	-0.002

Downlink activity and link-budget outputs

The main Methods define the Binomial activity-quantile calculation, SLO-implied downlink-rate conversion and reference free-space link budget. This section reports supporting numerical outputs only and introduces no additional traffic or link-budget method.

For comparison, replacing the P95 active-user count with the all-active population gives 8.46144 and 1.42704 Mbit s^{-1} for the 8B/1×H100 and 70B/2×H100 cases, respectively, at six bytes/token. This is an all-active comparison under the assumed per-source delivery rate rather than a universal bound on application traffic. Under the 0-dB receiver-SNR-floor stress case, the required payload-bearing

Supplementary Table 9: Activity quantiles and SLO-implied downlink design rates.

Configuration	N_{95}	Mean	P50	P95	P99	P95 rate (Mbit/s)
8B / 1×H100	5,876	154.141	154	174	183	0.16704–0.25056
70B / 2×H100	991	26.006	26	34	38	0.03264–0.04896

The two rate endpoints correspond to 4 and 6 bytes/token with a factor of three for protocol/implementation overhead. Quantiles are obtained from the exact convolution of the five Binomial active-user distributions; they are not empirical measurements of network traffic.

radiated power is approximately 0.070783 W at 20 km and 0.637046 W at 60 km with other reference parameters fixed. The complete scenario grid is retained in the source-data link tables.

These outputs correspond to an aggregate-traffic, free-space single-link design surrogate. They do not represent explicit multi-user scheduling, a complete cellular control/reference waveform, quantified uplink traffic or a joint end-to-end 95% reliability guarantee.

Hardware-allocation summary for platform integration

The main Methods define the system-equivalent hardware allocation and payload-feasibility rule. Supplementary Table 10 reports the configuration-level IT-power and mass values used in Fig. 6 together with the exact single-instance reliability-aware operating points; it introduces no additional payload-integration model. The DGX-derived values are based on manufacturer system specifications [6, 7], whereas the L40S shared-host allocation uses the adopted host and board references [8, 9].

Supplementary Table 10: Supporting engineering IT-power and mass allocations and exact single-instance reliability-aware operating points.

Configuration	Precision / TP	IT power (kW)	Allocated mass (kg)	N_{95}
8B / 1×H100	FP8 / TP1	1.275	16.306	5,876
8B / 1×H200	FP8 / TP1	1.275	16.306	7,245
8B / 1×L40S	FP8 / TP1	1.100	31.400	1,337
70B / 2×H100	FP8 / TP2	2.550	32.612	991
70B / 2×H200	FP8 / TP2	2.550	32.612	936
70B / 4×H100	FP8 / TP4	5.100	65.225	2,784
70B / 8×A100	BF16 / TP8	6.500	123.160	–

The DGX allocations use maximum-system specification values distributed across GPUs; they are not measured inference operating powers or bare-card masses. L40S uses a separately defined shared-host engineering allocation. The SLO-infeasible A100 case has no finite mixed-workload N_{95}.

Platform-level pooled capacity and request-assignment sensitivity

The main Methods define the pooled platform-capacity calculation, the linear mN_{95} reference and the static balanced-partition sensitivity. This section reports the corresponding platform-level numerical results only and introduces no additional reliability or fleet-dimensioning method.

For each HAPS class, the feasible configuration with the largest $N_{\mathrm{HAPS},95}$ gives 7,245 user equivalents for Sunglider and 23,144 for Stratobus. The corresponding compute-capacity dimensioning at illustrative target populations of 10,000, 25,000, 50,000 and 100,000 user equivalents gives 2, 4, 7 and 14 active Sunglider-class platforms and 1, 2, 3 and 5 active Stratobus-class platforms, respectively.

These counts are compute-capacity requirements, not geographical coverage requirements or fleet-wide 95% reliability guarantees.

Supplementary Table 11: Exact pooled platform capacities and linear single-instance references.

Platform	Configuration	m	mN_{95}	$N_{\mathrm{HAPS},95}$	Gain (%)
Sunglider	8B / 1×H100	1	5,876	5,876	0.00
Sunglider	8B / 1×H200	1	7,245	7,245	0.00
Sunglider	8B / 1×L40S	2	2,674	3,507	31.15
Stratobus	8B / 1×H100	3	17,628	19,104	8.37
Stratobus	8B / 1×H200	3	21,735	23,144	6.48
Stratobus	8B / 1×L40S	10	13,370	22,517	68.41
Stratobus	70B / 2×H100	1	991	991	0.00
Stratobus	70B / 2×H200	1	936	936	0.00

For $m = 1$, the exact single-instance operating point is inherited directly. For $m \geq 2$, pooled capacity is recomputed by direct evaluation of the platform-level overload probability defined in the main Methods. The linear quantity mN_{95} is a neutral reference rather than a platform-level reliability guarantee under static partitioning.

Supplementary Table 12: Sensitivity of multi-instance platform capacity to pooled versus static balanced request assignment.

Platform	Configuration	m	Static partition	mN_{95}	Pooled	Static overload at mN_{95} (%)
Sunglider	8B / 1×L40S	2	2,451	2,674	3,507	8.935
Stratobus	8B / 1×H100	3	16,684	17,628	19,104	13.360
Stratobus	8B / 1×H200	3	20,781	21,735	23,144	14.212
Stratobus	8B / 1×L40S	10	9,506	13,370	22,517	35.702

Static balanced partitioning fixes each workload type across the serving instances as defined in the main Methods, whereas the pooled model treats aggregate type-specific capacity as shareable. The ordering static $< mN_{95} <$ pooled is observed for these evaluated cases only and is not a general mathematical identity.

Supplementary Discussion

1. Sensitivity to activity-window overlap and global mapping

The B–C overlap contains 1,074 retained pairs in the supplied $\tau = 1800$-s diagnostics. All of these are assigned to B by nearest-centre resolution among eligible windows. Consequently, overlap-only distance and B-over-C priority give the same activity vector for this trace. They retain the baseline Type-B window but remove shared pairs from Type C. Type-C activity increases from approximately 7.681% to 8.198%. Conversely, C-over-B retains the Type-C window and changes Type-B activity from approximately 6.507% to 6.025%. It is therefore inaccurate to say that C-over-B primarily changes α_C.

Under exact probability evaluation, overlap-only distance and B-over-C reduce the 8B operating point from 5,876 to 5,873 user equivalents and the 70B operating point from 991 to 990, corresponding to changes of only -0.051% and -0.101%, respectively. C-over-B changes the exact operating points to 5,930 and 999, or +0.919% and +0.807% relative to baseline. These results indicate that the local rule used to resolve the observed B–C overlap has little effect on the reliability-aware capacity for the evaluated configurations.

Global distance changes the complete activity-estimation population rather than only resolving ambiguity; it increases exact N_{95} by 3.387% for 8B and 8.880% for 70B. The broader estimator

Supplementary Table 13: Activity fractions under alternative mapping rules at a 1800-s session threshold.

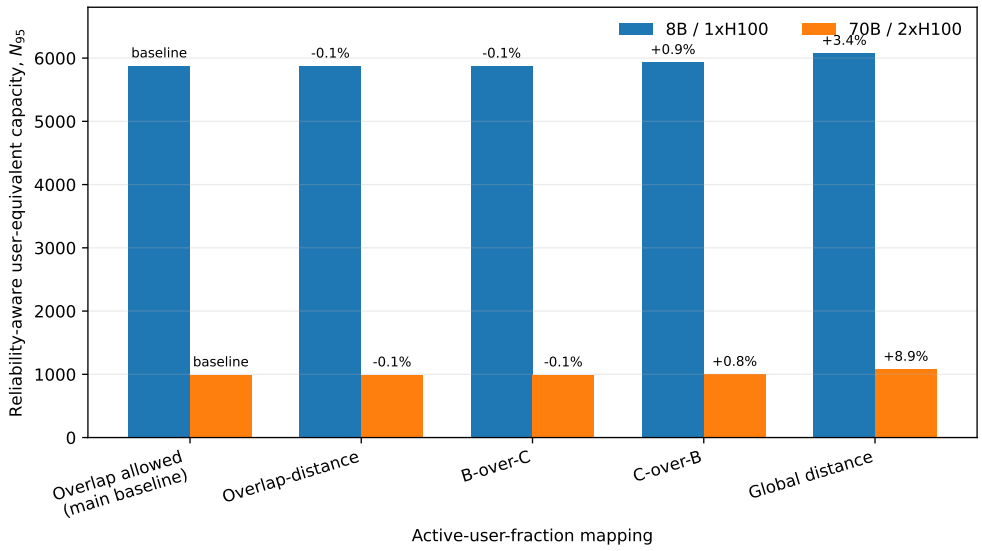
Type	Baseline	Overlap-only distance	B-over-C	C-over-B	Global distance
A	1.23708	1.23708	1.23708	1.23708	1.43915
B	6.50737	6.50737	6.50737	6.02540	4.79578
C	7.68103	8.19799	8.19799	7.68103	9.98721
D	4.81712	4.81712	4.81712	4.81712	3.33688
E	7.54857	7.54857	7.54857	7.54857	13.13352

All entries are percentages. The baseline is read from the baseline values used in Fig. 2c rather than overwritten with a recheck estimate. p_k and $C_{\max,k}$ are held fixed; the inferred population mixture q_k is recomputed when α_k changes.

Supplementary Table 14: Exact-under-model sensitivity of the reliability-aware N_{95} operating point to activity mapping.

Mapping	8B N_{95}	Change (%)	70B N_{95}	Change (%)
Overlapping windows	5,876	+0.000	991	+0.000
Overlap-only distance	5,873	-0.051	990	-0.101
B-over-C priority	5,873	-0.051	990	-0.101
C-over-B priority	5,930	+0.919	999	+0.807
Global distance	6,075	+3.387	1,079	+8.880

Percent changes are relative to the exact baseline values of 5,876 and 991 user equivalents. For each mapping rule, the full-precision activity fractions are propagated through the same deterministic exact-under-model reliability calculation while p_k and $C_{\max,k}$ are held fixed.



Supplementary Fig. 2: Sensitivity of the reported reliability-aware N_{95} operating point to active-user-fraction mapping. The overlap-allowed range-based estimator used in the main analysis is compared with nearest-centre resolution of overlapping Type-B/Type-C samples, B-over-C and C-over-B priority assignments, and global nearest-centre mapping. Values above the bars indicate the change in exact N_{95} relative to the overlap-allowed baseline. All cases use the same deterministic exact-under-model reliability calculation with only the activity-fraction mapping changed. Nearest-centre overlap resolution and B-over-C remain essentially unchanged from baseline, whereas the broader global-distance estimator produces the largest shift.

therefore produces the largest mapping-dependent shift, while the qualitative ordering of the two tested serving configurations is retained. No conclusion about all SLOs, all seven configurations or the full HAPS fleet follows without additional calculations.

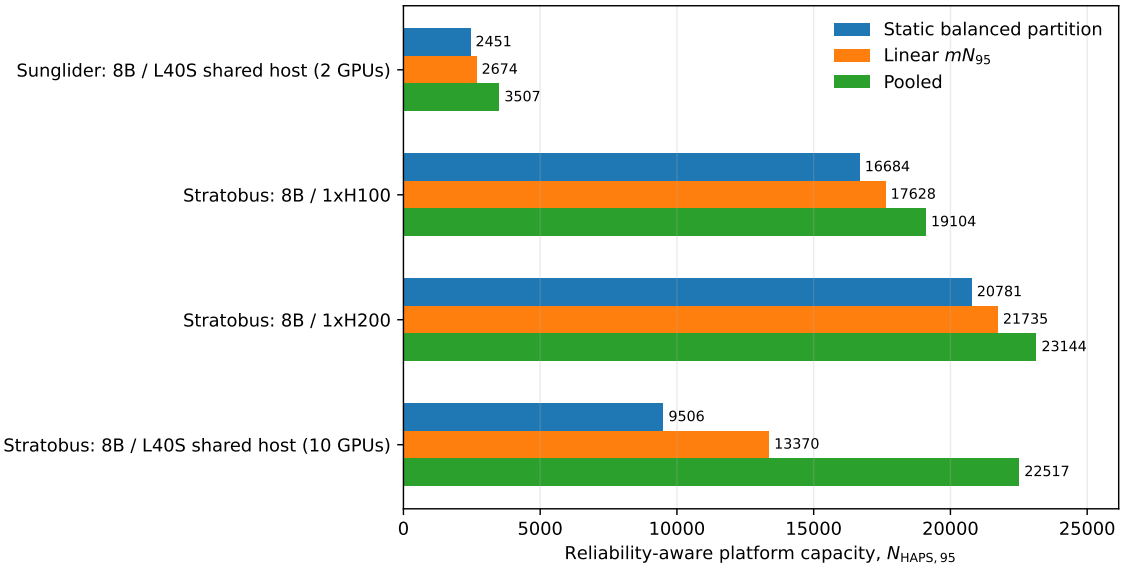
2. Sensitivity to platform-level request assignment

The main platform-capacity analysis assumes pooled type-specific serving resources and therefore represents an aggregate dynamic-routing reference. To test the dependence of the platform result on this assumption, we reran the platform reliability calculation using a static balanced-partition model. For each workload type, the integer user-equivalent population was divided as evenly as possible among the deployed serving instances, and the platform was classified as overloaded if any serving instance exceeded its own normalized serving budget.

The completed sensitivity results are summarized in Supplementary Table 12 and Supplementary Fig. 3. Static partitioning consistently reduced the platform-level capacity relative to pooled operation. The effect is modest for the three-instance H100/H200 cases but becomes pronounced for the ten-instance shared-host L40S case, where the pooled capacity is 22,517 user equivalents compared with 9,506 under static balanced partitioning. The linear reference mN_{95} lies between the exact static and pooled thresholds for every evaluated multi-instance case.

This ordering can be understood from the reliability definitions. Under pooled operation, short-term demand fluctuations can be absorbed by idle capacity elsewhere in the aggregate resource pool. Under static partitioning, spare capacity on one instance cannot rescue another partition, and the platform-level failure event is triggered when any instance overloads. Consequently, simply multiplying the single-instance N_{95} by m does not preserve a 5% platform-level overload criterion. As an intuition only, if m identical partitions each had an independent 5% overload probability, the probability that at least one partition overloaded would be $1 - 0.95^m$, which increases with m. At the platform threshold, the individual partitions need not each operate at exactly 5% overload probability, so $1 - 0.95^m$ is only an intuition rather than the calculation used here. The exact static-balanced result is obtained from the per-instance Binomial overload probabilities under the adopted independent-user activity model.

These results support interpreting the main pooled capacities as architectural references. Achieving them requires sufficiently flexible routing. They do not demonstrate that the pooled capacity will be achieved by a specific scheduler or interconnect, because indivisible-request placement, shared-host contention and scheduling overhead remain outside the aggregate model.



Supplementary Fig. 3: Sensitivity of platform-level reliability-aware capacity to request assignment across multiple serving instances. Static balanced partitioning, the neutral linear reference mN_{95}, and the pooled aggregate-capacity model are compared for the four multi-instance deployments. The pooled model provides the largest capacity in the evaluated cases because instantaneous idle capacity can be shared across instances in the aggregate model. Static balanced partitioning can fall below mN_{95} because the platform-level failure event is defined by any partition becoming overloaded.

3. Monte-Carlo validation of the exact reliability calculation

The main reliability results are obtained by direct evaluation of the overload probability implied by the independent Binomial activity model, so the reported N_{95} values do not depend on a random seed. We nevertheless retained Monte-Carlo calculations as a numerical validation of the deterministic probability evaluation. For each primary serving configuration, 20 independent random seeds with 10^5 activity snapshots per seed were pooled at each tested population, giving 2×10^6 snapshots per population near the exact threshold.

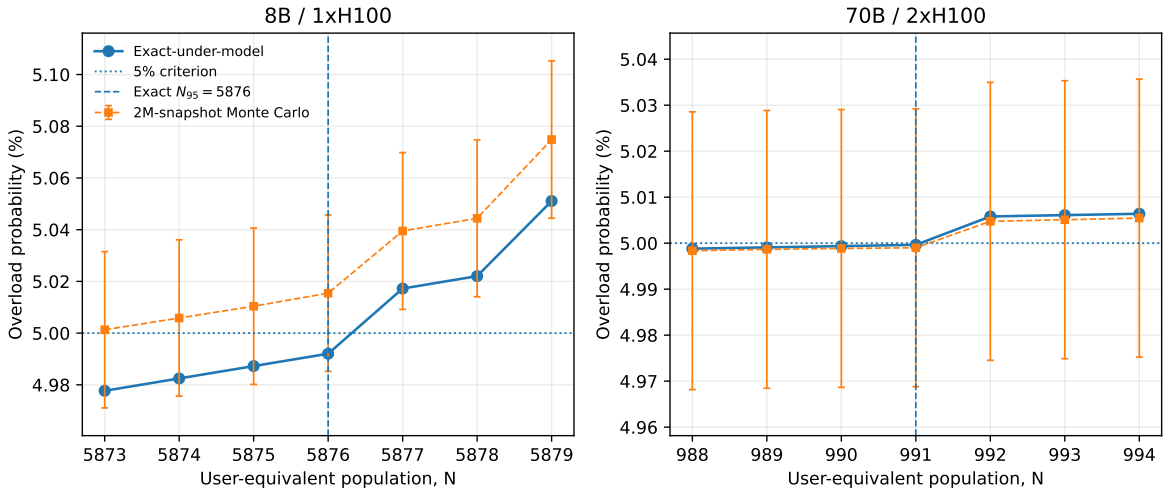
For 8B/1×H100, the exact threshold is $N_{95} = 5{,}876$. The pooled Monte-Carlo threshold is 5,872, a difference of four user equivalents (0.068%). At the exact threshold, the deterministic overload probability is 4.9920%, while the pooled Monte-Carlo estimate is 5.01545%; its normal-approximation 95% interval, 4.9852–5.0457%, contains the exact value. For 70B/2×H100, both methods give $N_{95} = 991$. At $N = 991$, the exact overload probability is 4.99964% and the pooled Monte-Carlo estimate is 4.9990%, with a 95% interval of 4.9688–5.0292%.

The earlier 20-seed, 10^5-snapshot threshold searches are retained as a diagnostic of finite-sample threshold variability. Their thresholds span 5,866–5,888 for 8B/1×H100 and 936–1,048 for 70B/2×H100. These ranges do not imply seed dependence of the underlying reliability model; they quantify the variability of a finite Monte-Carlo estimator when the true overload probability lies close to the 5% decision boundary. The effect is particularly visible for the 70B configuration because the exact probability changes only slightly over several adjacent integer populations near the threshold.

Supplementary Table 15: Monte-Carlo validation and finite-sample threshold variability relative to the exact reliability calculation.

Configuration	Exact N_{95}	2M-snapshot MC N_{95}	ΔN	20-seed mean	Median	Std.	Range
8B / 1×H100	5,876	5,872	-4	5,873.75	5,873.5	5.12	5,866–5,888
70B / 2×H100	991	991	0	993.95	1,001.5	39.45	936–1,048

The pooled Monte-Carlo calculation combines 20 independent seeds with 10^5 activity snapshots per seed, for 2×10^6 snapshots per tested population. The 20-seed mean, median, standard deviation and range summarize separate fully bracketed 10^5-snapshot threshold searches retained as a finite-sampling diagnostic. These statistics are not confidence intervals on the workload model.



Supplementary Fig. 4: Monte-Carlo validation of the exact overload probability near the 5% reliability boundary. Exact-under-model overload probabilities are compared with estimates obtained by pooling 2×10^6 Monte-Carlo activity snapshots per population. Error bars show normal-approximation 95% Monte-Carlo intervals. Vertical dashed lines mark the exact N_{95} values of 5,876 and 991. The 8B Monte-Carlo threshold differs by four user equivalents, while the 70B threshold coincides with the exact result; in both cases the probability estimates are statistically compatible with the exact values at the threshold.

4. Scope of the available checks

The available tables support the baseline day-bootstrap intervals, exact mapping-dependent N_{95} comparison, deterministic interpolation sensitivity through N_{harm}, thermal environmental corners, duct reruns, higher-load thermal scaling and co-design checks, exact pooled and static-balanced platform thresholds for the evaluated multi-instance cases, and Monte-Carlo validation for the two primary Fig. 3d configurations. The exact reliability values are conditional on the adopted independent Binomial activity model, integer population-allocation rule and type-specific serving capacities; they remove Monte-Carlo sampling error but do not remove uncertainty in those modelling assumptions. The Monte-Carlo checks are therefore used as numerical validation rather than as the definition of the reported reliability threshold. Reproducibility, parameter sensitivity and validation of a modelling approximation remain distinct questions.

Supplementary References

[1] Wang, Y. et al. BurstGPT: A real-world workload dataset to optimize LLM serving systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 2, pages 5831–5841. Association for Computing Machinery, 2025.

[2] 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.

[3] MathWorks. Conductive heat transfer. https://www.mathworks.com/help/simscape/ref/conductiveheattransfer.html. Simscape reference documentation. Accessed 11 September 2026.

[4] MathWorks. Convective heat transfer. https://www.mathworks.com/help/simscape/ref/convectiveheattransfer.html. Simscape reference documentation. Accessed 11 September 2026.

[5] MathWorks. Radiative heat transfer. https://www.mathworks.com/help/simscape/ref/radiativeheattransfer.html. Simscape reference documentation. Accessed 11 September 2026.

[6] 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.

[7] 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.

[8] 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.

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

[10] Shiro Fukumoto. 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, October 2020. SoftBank Corp. and HAPSMobile Inc., accessed August 2026.

[11] 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, March 2023. Press release, accessed August 2026.