Numerical Study of Workload-Dependent Air-Side Convective Heat Transfer over a High-Density AI Server Board under Large Language Model Inference

Nafisa Anjum Tanisha[1, a)] and Anindya Guha[1, b)]

[1]*Department of Mechanical Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh.*

[a)] Corresponding author: nafisaanjum1004@gmail.com

[b)] ag.anindya21@gmail.com

Abstract. Large language model (LLM) inference has pushed data center servers toward higher and more uneven heat loads. System-level thermal models usually give each component a fixed convective heat transfer coefficient, although the heat load changes several-fold with the model being served. This work tests that assumption on a high-density AI server board based on the NVIDIA DGX A100 layout, modelled with separate regions for two compute boards (CB1 and CB2), four DIMM banks and two PCIe cards, and solved in ANSYS Fluent as a steady three-dimensional conjugate case with the k-ε turbulence model at inlet velocities of 1, 3 and 5 m/s. Block heat loads were derived from token-level power estimates for LLaMA 7B, LLaMA 13B, GPT-3 Small and GPT-3 175B at 1 to 70 generated tokens. At 70 tokens the load on one compute board ranged from 64 to 359 W, and at 1 m/s the coefficient changed with the workload in every block: from 63.6 to 64.4 W/m^{2}K on CB1, 60.2 to 61.8 on CB2, 222 to 288 on the finned CB1, 11.0 to 12.7 on the DIMM banks and 11.0 to 17.9 on the PCIe cards, while the CB1 temperature rise went from 72 to 400 K. The coefficient rose with velocity as $Nu = C\,Re^m\,Pr^{1/3}$, with m = 0.37 to 0.39 for the compute boards and 0.6 to 0.7 for the DIMM and PCIe blocks and it depended strongly on position: the DIMM bank lying in the wake of a PCIe card and a compute board had less than half the coefficient of the others. The downstream compute board was 3 to 6 percent below the upstream one without fins and 30 to 39 percent below it with fins. Rectangular fins with a thermal paste interface raised the effective coefficient 3.5 to 4.5 times on CB1 and 2.6 to 2.8 times on CB2, cutting the temperature rise by 71 to 78 and 61 to 65 percent. The study characterizes the air-side coefficient of each block and the reported ranges show how far a coefficient found for one workload can be carried to another.

INTRODUCTION

Servers built for artificial intelligence now dissipate several kilowatts each. An NVIDIA DGX A100 node is rated at about 6.5 kW [1], and racks filled with such nodes push conventional air cooling close to its limit [2]. During LLM inference the power drawn also depends on the model being served and on the response length, and measurements on production GPU clusters show that it rarely sits at the rated value [3, 4]. The heat that the cooling system has to remove is therefore set by the workload, not only by the hardware.

Most thermal models of data center equipment represent a server as a lumped mass or a black-box region with a uniform heat load [5, 6]. Each component is linked to the cooling air by a convective heat transfer coefficient, h, which is taken as fixed once the geometry and airflow are fixed. Whether this still holds when the load on a board changes several-fold between workloads is rarely checked. Components also interact: downstream parts receive disturbed, warmed air [5], and memory temperatures are raised by processor heat conducted through the printed circuit board (PCB) [7].

This paper examines the air-side coefficient of a high-density AI board under realistic LLM workloads. Block heat loads are derived from token-level power estimates for four models (LLaMA 7B, LLaMA 13B, GPT-3 Small and GPT-3 175B [9, 10]) over 1 to 70 generated tokens. A steady CFD model of a DGX A100-style board is then used to extract the surface temperature, h and the Nusselt number of each block at inlet velocities of 1, 3 and 5 m/s, with and without fins on the compute boards. The aims are to check whether h changes with workload on identical hardware,

to give a Nusselt–Reynolds correlation for each block, and to quantify the upstream–downstream difference and the benefit of fins.

METHODOLOGY

Workload-Resolved Heat Load

The heat load of each run was estimated from the model architecture and the number of generated tokens, N_t. The compute demand was approximated from the number of layers, n, and the hidden width, d, and converted to a compute utilization, U_c, by dividing by the peak rate of six accelerators at 40 TFLOP/s each:

$$F = 2\,Nt\,n\,d2, \quad Uc = F / (6 \times 4 \times 1013) \tag{1}$$

The memory utilization, U_m, was taken to rise in proportion to N_t at a model-specific rate. The two were blended into one load fraction and mapped linearly between the idle and peak node power, following utilization-based GPU power modelling [8]:

$$U = 0.7\,Uc + 0.3\,Um, \quad P = Pidle + (Ppeak - Pidle)\,U, \quad TDP = 0.95\,P \tag{2}$$

with P_{idle} = 65 W and P_{peak} = 6500 W. The thermal design power (TDP) was split into fixed shares of 60 % for the compute boards, 15 % for the DIMM banks and 5 % for the PCIe cards; the remaining 20 % (fans, regulators and power supply) was not modelled. The compute share was divided equally between CB1 and CB2, the DIMM share over four banks and the PCIe share over two cards, and each block load was applied as a volumetric heat source. Table 1 lists the inputs and the loads at 70 tokens.

TABLE 1. Model inputs and resulting loads at 70 generated tokens.

Model	n	d	U_m	t_{inf} (s)	TDP (W)	Load per CB (W)
GPT-3 Small	12	768	0.08	2.0	208.5	64.1
LLaMA 7B	32	4096	0.10	2.6	246.5	76.2
LLaMA 13B	40	5120	0.19	4.8	414.2	128.0
GPT-3 175B	96	12288	0.58	35.0	1161.6	358.8

Inference is intermittent: a burst lasting the inference time, t_{inf}, is followed by an idle period. A 600 s profile was therefore built for every block by alternating an idle floor (the 65 W idle power, split with the same shares) with the active load, with a ±3 % ripple. The steady CFD cases use the peak of this profile, so they give an upper bound on the temperature reached during a burst.

Geometry and Numerical Setup

The board follows the NVIDIA DGX A100 layout [1] and keeps only the heat-generating regions that matter for the air side (Fig. 1): two compute boards (CB1 upstream and CB2 downstream), four DIMM banks and two PCIe cards on a PCB. Each compute board is a 100 × 100 × 10 mm block with an exposed area of 0.014 m^2, and each DIMM bank and PCIe card is a 300 × 135 × 30 mm block with an exposed area of 0.0666 m^2. The thermal conductivity was 140 W m^{-1} K^{-1} for the compute boards, 0.8 W m^{-1} K^{-1} for the DIMM and PCIe blocks and 3 W m^{-1} K^{-1} for the PCB. The board was placed in a rectangular air enclosure with an inlet area of 0.127 m^2. In the finned cases, rectangular aluminium plate fins were bonded to CB1 and CB2 through a thin thermal paste layer.

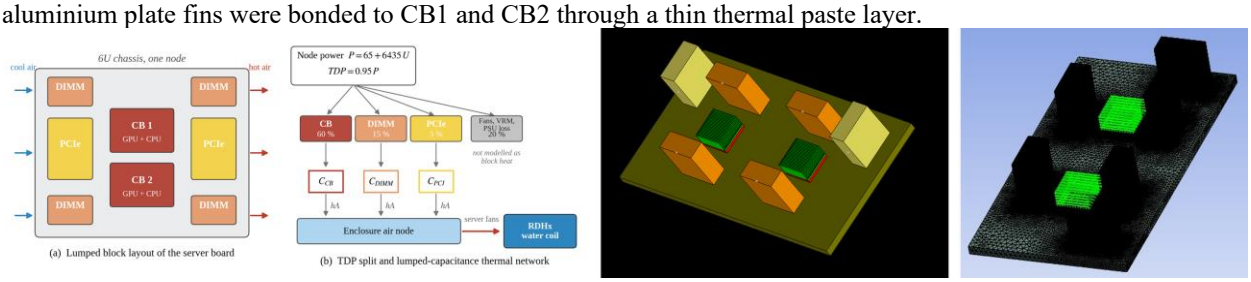


FIGURE 1. (a) Block layout of the server board with the TDP split (flow from left to right); (b) CAD model in the air enclosure; (c) tetrahedral mesh with refined body sizing on the compute boards and fins.

The steady, incompressible Reynolds-averaged flow and energy equations were solved in ANSYS Fluent with the standard k-ε turbulence model [11] and enhanced wall treatment, and conduction in all solids was solved together with the flow. Air was given constant properties at 298 K; gravity and radiation were not included. A uniform velocity at

298 K was set at the inlet, zero gauge pressure at the outlet, symmetry on the side and top faces of the enclosure and no-slip walls on all solid surfaces. Six 60 cfm fans would deliver about 1.3 m/s through this inlet, so 1 m/s was used as the baseline and 3 and 5 m/s were added to cover higher fan speeds. These conditions serve to characterize the coefficient of each block, not to design the cooling for a given model. All cases converged within about 100 iterations (scaled residuals below 10^{-3}, 10^{-6} for energy) and were run to 500–600 iterations to confirm that the temperature monitors stayed flat.

The domain was meshed in ANSYS Meshing with unstructured tetrahedral elements (Fig. 1(c)). A global element size was applied to the air and the PCB, and two body sizings refined the compute boards and the fin region. Mesh independence was tested on the LLaMA 7B, 70-token case at 1 m/s by refining both sizes in four steps while monitoring the CB1 temperature and coefficient (Table 2). The change in h fell from 2.7 % to 0.2 %, so the fine mesh (15 mm global, 3 mm body sizing) was used for all cases.

TABLE 2. Mesh independence study (LLaMA 7B, 70 tokens, 1 m/s, no fin

Mesh	Global / body (mm)	Elements (×10^6)	T_{CB1} (K)	h_{CB1} (W m^{-2} K^{-1})	Change (%)
Coarse	25 / 5	0.64	386.4	61.5	–
Medium	20 / 4	1.31	384.1	63.2	2.7
Fine (used)	15 / 3	2.68	383.0	64.0	1.3
Very fine	10 / 2	5.12	382.8	64.2	0.2

The solver settings were then checked on a reference case with a known correlation: an isothermal square cylinder of 0.10 m side in cross-flow, meshed with the same sizing and solved with the same turbulence model and wall treatment. The predicted Nusselt number was compared with $\mathrm{Nu} = 0.102\,\mathrm{Re}^{0.675}\,\mathrm{Pr}^{1/3}$, valid for $5 \times 10^3 < \mathrm{Re} < 10^5$ [12], at the three Reynolds numbers of the compute boards (Table 3). The deviation was 2 to 7 %, within the usual accuracy of such correlations.

TABLE 3. Validation against the cross-flow correlation for a square cylinder [12]

Inlet velocity (m/s)	Re	Nu, correlation	Nu, present model	Deviation (%)
1	6293	33.3	35.6	+6.9
3	18 880	69.9	73.1	+4.6
5	31 466	98.7	100.8	+2.1

For each block, the area-averaged surface temperature, T_s, was extracted and compared with the ambient temperature, $T_{amb} = 298$ K. The coefficient, Nusselt number and Reynolds number were then found as

$$h = Q / [A\,(T_s - T_{amb})], \quad \mathrm{Nu} = h\,L / k, \quad \mathrm{Re} = V\,L / \nu \tag{3}$$

where Q is the block heat load, A its exposed area, V the inlet velocity and L the streamwise length of the block (0.10 m for the compute boards, 0.135 m for the DIMM and PCIe blocks). For the finned boards A was kept at the unfinned exposed area, so the finned value is an effective, area-augmented coefficient.

RESULTS AND DISCUSSION

Heat Load and Its Transient Character

Figure 2(a) shows that the peak load on one compute board ranges from about 20 W at a single token, where every model sits close to idle, to 359 W for GPT-3 175B at 70 tokens, a factor of 18 on the same hardware. The load also differs in duration: at 70 tokens a LLaMA 7B burst lasts 2.6 s, while GPT-3 175B holds its peak for 35 s (Fig. 2(b, c)). With a silicon-like heat capacity ($\rho c \approx 1.63$ MJ m^{-3} K^{-1}) a compute board has a lumped time constant $\tau = \rho c V/(hA)$ of about 180 s without fins (Biot number about 0.02). A 2.6 s burst then reaches only about 1.4 % of the steady temperature rise, while a 35 s GPT-3 175B burst reaches about 17 %. The steady peak-load results are therefore an upper bound, tightest for the heaviest model.

Effect of Workload at Fixed Airflow

Table 4 and Fig. 4(a) give the results at 70 tokens and 1 m/s without fins, and Fig. 3 shows a typical temperature field. The load on a compute board rises 5.6 times from GPT-3 Small to GPT-3 175B, and the coefficient changes with it in every block. On CB1 it ranges from 63.6 to 64.4 W m^{-2} K^{-1} and on CB2 from 60.2 to 61.8 W m^{-2} K^{-1}; with fins the CB1 value ranges from 222 to 288 W m^{-2} K^{-1} and the CB2 value from 156 to 175 W m^{-2} K^{-1}. The mean coefficient of the DIMM banks ranges from 11.0 to 12.7 W m^{-2} K^{-1} and that of the PCIe cards from 11.0 to 17.9 W m^{-}

^{2} K^{-1}. The change is 1.3 % and 2.6 % on the unfinned compute boards, 30 % and 12 % on the finned ones, 15 % on the DIMM banks and 63 % on the PCIe cards. On the unfinned compute boards part of the change is comparable with the 1 K reading resolution.

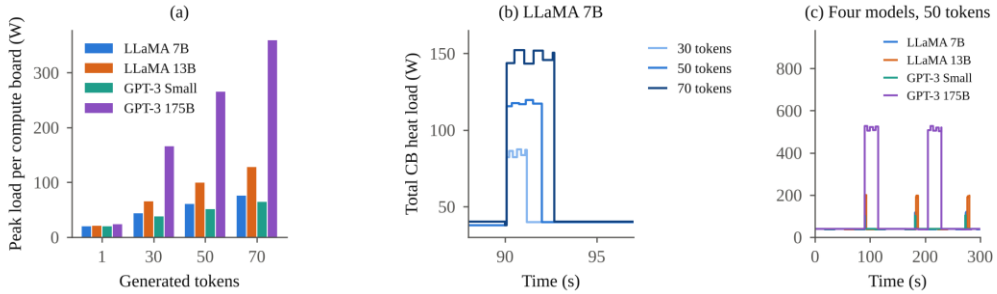


FIGURE 2. (a) Peak load per compute board at 1–70 tokens; (b) transient compute-block load for LLaMA 7B; (c) transient load of the four models at 50 tokens.

TABLE 4. Results at 70 tokens and 1 m/s without fins (h in W m^{-2} K^{-1}; DIMM as mean and range of the four banks).

Model	Q_{CB} (W)	ΔT_{CB1} (K)	h_{CB1}	ΔT_{CB2} (K)	h_{CB2}	h_{DIMM}	h_{PCIe}
GPT-3 Small	64.1	72	63.6	76	60.2	11.0 (6.0–15.0)	13.6
LLaMA 7B	76.2	85	64.0	88	61.8	12.0 (7.0–15.8)	11.0
LLaMA 13B	128.0	142	64.4	151	60.5	11.9 (7.0–16.0)	14.5
GPT-3 175B	358.8	400	64.0	423	60.6	12.7 (8.0–16.6)	17.9

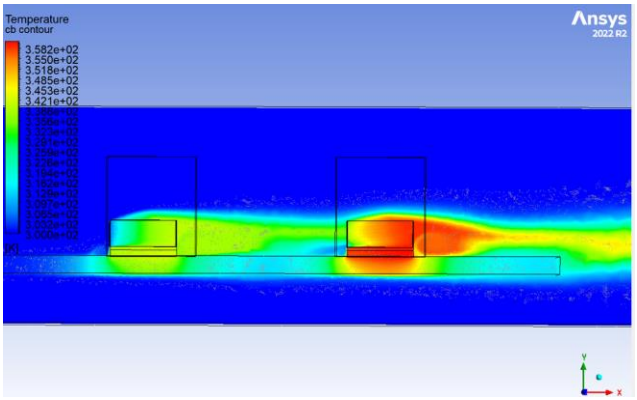


FIGURE 3. Representative temperature field on a vertical mid-plane through the compute boards; flow from left to right.

Because the change on the compute boards is small compared with the change in load, their temperature rise follows the load almost linearly (Fig. 4(b)): 72 K for GPT-3 Small, 85 K for LLaMA 7B, 142 K for LLaMA 13B and 400 K for GPT-3 175B on CB1. The workload therefore acts mainly through the surface temperature, and a coefficient found for one workload can be carried to another within the ranges given above. The 1 m/s condition is used here to characterize the coefficient and is not a cooling design for the largest model; the 698 K reached by CB1 under GPT-3 175B is outside the range where constant air properties are accurate and is indicative only. The compute-board coefficients are about five times the laminar flat-plate value of 12.3 W m^{-2} K^{-1} [12] because much of their heat leaves through the PCB; they are effective design values.

Within the DIMM group the change differs from bank to bank: 15.0–16.6 W m^{-2} K^{-1} for the first bank, 12.5–13.0 for the second, 10.5–13.0 for the third and 6–8 for the fourth. Position matters far more than workload: in every case the first three banks lie between 10.5 and 16.6 W m^{-2} K^{-1}, while the fourth reaches only 6–8 W m^{-2} K^{-1}, less than half the value of the first. The fourth bank is the last block in a densely packed row, behind a PCIe card and a compute board, so the air has to pass both obstacles before reaching it. It therefore sits in their combined wake, where the velocity is low and the air has already been heated by the upstream blocks. Because h is referenced to the ambient temperature, both effects lower it. The weaker response of this bank to velocity ($m \approx 0.4$–0.5 against 0.69 for the first bank) supports this explanation, since separated wake flow gains less from a faster stream than attached flow. The PCIe cards change the most. Their mean coefficient ranges from 11.0 W m^{-2} K^{-1} for LLaMA 7B to 17.9 W m^{-2} K^{-1} for

GPT-3 175B. Three causes are likely: the cards rise only 5–12 K for the small models, so a reading error of 0.5 K changes h by 4–10 %; their low conductivity makes the surface temperature converge slowly in a steady solver; and they share the PCB with the neighbouring compute boards, so their temperature is not set by their own load alone.

Effect of Inlet Velocity and Nusselt Number Correlation

The GPT-3 175B case was repeated at 3 and 5 m/s (Table 5). The CB1 coefficient rose from 64.0 to 96.5 and 116.2 W m^{-2} K^{-1}, and its temperature rise fell from 400 K to 265 K and 221 K. A five-fold increase in velocity therefore raises the coefficient by about 80 %. The results of each block were then fitted by least squares to

$$\mathrm{Nu} = C\,\mathrm{Re}^{m}\,\mathrm{Pr}^{1/3} \tag{4}$$

For the compute boards the fit reproduces every case within 1.7 %, with m = 0.37 for CB1 and 0.39 for CB2 (Fig. 4(c)). This is below the laminar flat-plate value of 0.5 [12], because the heat leaving through the PCB passes a solid resistance that does not change with velocity. Adding a factor $(T_s/T_{amb})^n$ gave n = 0.01 for CB1 and −0.01 for CB2 at 1 m/s, so the workload change on the compute boards is too small to need a separate term.

TABLE 5. GPT-3 175B coefficient at three velocities and constants of Eq. (4) (12 cases per block)

Block	h at 1 / 3 / 5 m/s (W m^{-2} K^{-1})	C	m	Deviation mean / max (%)
CB1	64.0 / 96.5 / 116.2	10.59	0.37	0.5 / 1.7
CB2	60.6 / 93.0 / 114.4	8.21	0.39	0.7 / 1.6
DIMM (mean)	12.7 / 27.1 / 37.4	0.136	0.69	3.7 / 8.2
PCIe (mean)	17.9 / 33.3 / 44.9	0.439	0.58	13.5 / 27.4

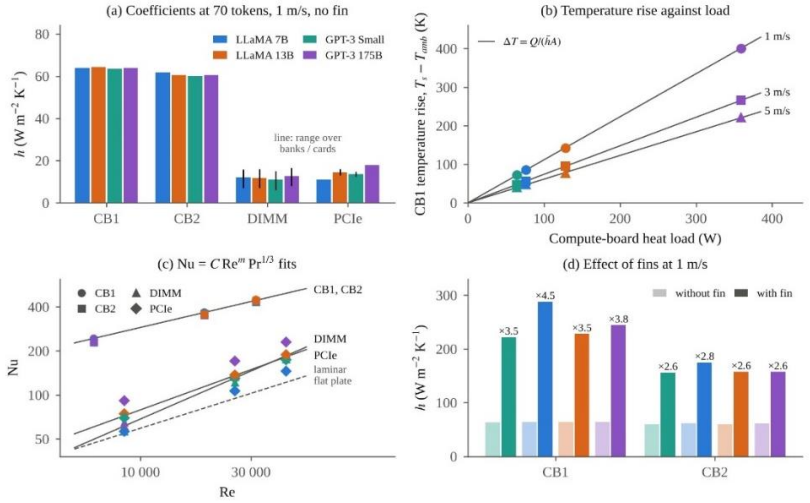


FIGURE 4. (a) Block coefficients without fins; (b) CB1 temperature rise against load, with lines from the mean coefficient; (c) Nusselt number against Reynolds number with the fits of Eq. (4); (d) compute-board coefficient without and with fins, models in order of load.

The DIMM and PCIe blocks give m = 0.69 and 0.58, between the laminar (0.5) and turbulent (0.8) flat-plate values, as expected for blocks in a stream already disturbed by upstream components. At 1 m/s the mean DIMM coefficient lies 4 to 19 % above the laminar flat-plate value of 10.6 W m^{-2} K^{-1}, which supports the model. The DIMM fit holds within 8 %; the PCIe fit is looser (up to 27 %) because of the model-to-model scatter noted above, not the velocity trend. Because the exponent differs between the exposed and the sheltered banks, one correlation for all banks is a compromise.

Effect of Rectangular Fins and Board Position

Table 6 and Fig. 4(d) compare the unfinned and finned boards at 1 m/s for the four models. With fins the CB1 temperature rise fell from 72 K to 20.6 K for GPT-3 Small, from 85 K to 18.9 K for LLaMA 7B, from 142 K to 40 K for LLaMA 13B and from 400 K to 105 K for GPT-3 175B, a reduction of 71–78 %. The effective coefficient, still referenced to the unfinned area, rose from about 64 to 222–288 W m^{-2} K^{-1}, a gain of 3.5 to 4.5 times. On CB2 the

temperature rise fell by 61–65 % and the coefficient rose 2.6 to 2.8 times, to 156–175 W m^{-2} K^{-1}. The gain comes mainly from the added surface area.

TABLE 6. Effect of fins at 70 tokens and 1 m/s (ΔT: without fin → with fin; h: finned value in W m^{-2} K^{-1}).

Model	ΔT_{CB1} (K)	Drop (%)	h_{CB1}	Gain	ΔT_{CB2} (K)	Drop (%)	h_{CB2}	Gain
GPT-3 Small	72 → 20.6	71	222.0	×3.49	76 → 29.3	61	156.0	×2.59
LLaMA 7B	85 → 18.9	78	288.0	×4.50	88 → 31.1	65	175.0	×2.83
LLaMA 13B	142 → 40	72	228.0	×3.54	151 → 58	62	157.6	×2.60
GPT-3 175B	400 → 105	74	244.1	×3.81	423 → 163	61	157.2	×2.59

Fins also change the balance between the two boards. CB2 sits in the wake of CB1 (Fig. 3) and takes in air that is obstructed and slightly warmed. Without fins the penalty is small: its coefficient is 3.4 to 6.0 % lower at 1 m/s, and for GPT-3 175B the gap narrows to 3.6 % at 3 m/s and 1.5 % at 5 m/s as the faster stream flushes the wake. With fins CB2 is 30–39 % below CB1, because the finned CB1 blocks more of the stream and releases its heat directly ahead of CB2. The finned coefficient also changes more with workload than the unfinned one, by 30 % on CB1 and 12 % on CB2, and the change does not follow the load: LLaMA 7B gives the highest values (288 and 175 W m^{-2} K^{-1}), while the lighter GPT-3 Small and the heavier LLaMA 13B and GPT-3 175B lie between 222 and 244 W m^{-2} K^{-1} on CB1. Possible causes are the small temperature rise of the finned boards (19–40 K for the three lighter models), which makes h sensitive to the reading, and slower convergence in the thin fin and paste regions; the LLaMA 7B case should be rechecked.

CONCLUSION

The main findings are:

1. The load on one compute board spans a factor of 18 across models and token counts.
2. At a fixed airflow the coefficient changed with workload in every block while the load changed 5.6 times: from 63.6 to 64.4 W m^{-2} K^{-1} on CB1, 60.2 to 61.8 on CB2, 222 to 288 on the finned CB1, 11.0 to 12.7 on the DIMM banks and 11.0 to 17.9 on the PCIe cards. The CB1 temperature rise went from 72 to 400 K.
3. The coefficient follows $Nu = C\,Re^{m}\,Pr^{1/3}$ with m = 0.37–0.39 for the compute boards (within 1.7 %) and 0.58–0.69 for the DIMM and PCIe blocks (within 8 % and 27 %); the DIMM bank in the wake of a PCIe card and a compute board has less than half the coefficient of the first.
4. The downstream compute board was 3–6 % below the upstream one without fins and 30–39 % below it with fins.
5. Fins raised the effective coefficient 3.5–4.5 times on CB1 and 2.6–2.8 times on CB2 and cut the temperature rise by 71–78 % and 61–65 %.

Future work will add temperature-dependent air properties and resolve the transient response.

REFERENCES

1. NVIDIA Corporation, *NVIDIA DGX A100 System Architecture*, Technical White Paper (NVIDIA, Santa Clara, 2020).
2. A. A. Alkrush, M. S. Salem, O. Abdelrehim and A. A. Hegazi, Int. J. Refrig. **160**, 246–262 (2024).
3. P. Patel *et al.*, "Characterizing power management opportunities for LLMs in the cloud," in *Proceedings of the 29th ACM ASPLOS Conference* (ACM, New York, 2024), pp. 207–222.
4. S. Samsi *et al.*, "From words to watts: Benchmarking the energy costs of large language model inference," in *2023 IEEE High Performance Extreme Computing Conference* (IEEE, 2023), pp. 1–9.
5. W.-X. Chu and C.-C. Wang, Appl. Energy **240**, 84–119 (2019).
6. Z. Li *et al.*, Appl. Therm. Eng. **251**, 123612 (2024).
7. Z.-X. Wang and W.-Q. Tao, Appl. Therm. Eng. **280**, 128111 (2025).
8. V. Kandiah *et al.*, "AccelWattch: A power modeling framework for modern GPUs," in *Proceedings of the 54th IEEE/ACM MICRO Symposium* (ACM, New York, 2021), pp. 738–753.
9. H. Touvron *et al.*, arXiv:2302.13971 (2023).
10. T. B. Brown *et al.*, Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020).
11. ANSYS Inc., *ANSYS Fluent Theory Guide, Release 2022 R2* (ANSYS Inc., Canonsburg, 2022).
12. M. Jakob, Heat Transfer, Vol. 1, John Wiley & Sons, New York, 1949.