

Article

Numerical Investigation of the Thermal Environment in a Data Center Under Dynamic Server Loads Driven by Artificial Neural Network Models

Chang Yue ¹, Zimu Li ^{1,2}, Jiaqiang Wang ^{1,*}, Zhang Quan ³, Zhiqiang Zhai ⁴ and Michael W. Jack ⁵

¹ School of Energy Science and Engineering, Central South University, Changsha 410083, China; yuechang928@csu.edu.cn (C.Y.); 233911008@csu.edu.cn (Z.L.)

² China Telecom Co., Ltd., Guangdong Branch, Guangzhou 510030, China

³ College of Civil Engineering, Hunan University, Changsha 410082, China; quanzhang@hnu.edu.cn

⁴ Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, CO 80309, USA; john.zhai@colorado.edu

⁵ Department of Physics, University of Otago, Dunedin 9016, New Zealand

* Correspondence: jiaqiangwang@csu.edu.cn

Abstract

This study numerically investigates the transient thermal response of a raised-floor data center under dynamic AI server loads. Power profiles of DGX A100 servers running GEMM, VGG-19, and ResNet-152 workloads were used as time-varying boundary conditions in CFD simulations. Six load scenarios were analyzed to evaluate the dynamic responses of cooling capacity, rack inlet and outlet temperatures, and net airflow rate. The results show that the response lag of air-conditioning units causes a temporary mismatch between cooling supply and server heat dissipation, leading to transient inlet-temperature fluctuations. Fluctuating workloads produce larger outlet-temperature amplitudes and higher temperature change rates than stable workloads, while mixed workloads change the overall fluctuation intensity and affect neighboring-rack airflow through cold-aisle pressure interactions. These findings can support workload-aware rack arrangement and adaptive cooling control for future AI-oriented data centers.

Keywords: airflow organization; data center; dynamic load; CFD

1. Introduction

In recent years, the proliferation of generative artificial intelligence and large language model applications has driven growth in data center demand. The global data center market reached approximately USD 269.8 billion in 2025 and is projected to expand at a CAGR of 11.10%, reaching USD 699.1 billion by 2034 [1]. This growth carries a significant energy cost. Global data center electricity consumption reached approximately 415 TWh in 2024 and is projected to exceed 945 TWh by 2030 [2]. Escalating energy demands intensify carbon emission pressures and make improving data center energy efficiency a critical industry challenge.

Cooling systems account for approximately 40% of total data center energy consumption [3,4], making them the primary target for efficiency optimization. Air cooling remains the dominant approach due to its low operational cost and broad adaptability

Academic Editor: Constantinos A. Balaras

Received: 26 May 2026

Revised: 16 June 2026

Accepted: 03 July 2026

Published: 13 July 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

[5]. Airflow optimization is widely recognized as a key strategy for reducing cooling energy consumption. Poor airflow organization can lead to hot-air recirculation and cold-air bypass, causing localized hotspots that risk damaging IT equipment. Understanding airflow distribution is therefore essential for stable operation and effective energy management [6].

Research on airflow optimization under steady-state conditions has been extensive, covering ACU placement, plenum structure, perforated tile design, and aisle containment strategies. A symmetrical ACU layout improves cooling performance [7]. Nada and Said [8] found that a perpendicular supply arrangement improves rack inlet uniformity, reduces end-rack recirculation, and suppresses mid-row cold-air bypass. Increasing plenum depth improves airflow uniformity, though structural limitations arise beyond a certain threshold [9,10]. Beitelmal [11] proposed an optimal range of 0.762–0.914 m. Zhang [12] recommended 1.0–1.2 m for open aisles and 0.6–0.8 m for containment configurations. For perforated tiles, an open area ratio of 20–25% has been commonly recommended [13,14]. Variable-ratio and active perforated floors have also been proposed to further improve supply uniformity [15–18]. Aisle containment effectively suppresses bypass and recirculation. Nada and Abbas [19] found that full cold-aisle containment outperformed partial top containment. Tsuda et al. [20] reported that hot-aisle containment required twice the fan power of cold-aisle containment and that cold-aisle containment better maintained optimal server intake temperatures. Flow control devices such as baffles and jet fans installed within plenum chambers or at rack inlets provide further improvements [21–24]. However, all aforementioned studies assume constant server loads. The thermal and airflow responses to continuously fluctuating workloads remain largely unexplored.

Data center loads are inherently dynamic and non-uniform, driven by the variability of user behavior, workload scheduling, and task allocation. Existing studies have addressed three aspects. The first is server power modeling. Researchers have categorized server power models into the additive model and the Base-Amplitude (BA) model [25]. The additive model defines total server power as the sum of component-level consumption across the CPU, memory, and cooling fans [26,27]. The BA model decomposes power into static and active components [28,29], with variants subsequently derived from both approaches [30,31]. The second is energy consumption tracking under realistic scenarios. Wang et al. [32] executed the Quantum ESPRESSO benchmark on an NVIDIA V100S GPU and recorded utilization, power, and temperature under frequency-limiting conditions using the NVML library. Krzywaniak et al. [33] ran general matrix multiplication and neural network models on a DGX A100 GPU and measured transient energy curves under power-capping conditions. The third is transient thermal response under dynamic boundary conditions. Fulpagare et al. [34] found that a sudden load increase may cause rack inlet temperatures to exceed the recommended range. Ibrahim et al. [35] showed that airflow rate changes exert far greater influence on rack inlet temperatures than rack power adjustments alone. Despite these advances, existing research remains fragmented. Few studies have coupled real-world power curves with room-scale CFD simulations to capture continuous dynamic thermal evolution under realistic AI workloads.

To address this gap, the present study makes the following novel contributions. First, transient energy consumption curves of the NVIDIA DGX A100 GPU server under three representative AI workloads are extracted from published literature and applied as time-varying power boundary conditions to a room-scale raised-floor data center model. Second, numerical simulations covering different load fluctuation levels are conducted to analyze dynamic variations in rack inlet and outlet temperatures, cooling capacity, and airflow velocity, investigating the effects of load fluctuation level on the dynamic thermal environment. The results provide a theoretical basis for precise cooling control and energy efficiency optimization under realistic AI computing scenarios.

2. Materials and Methods

2.1. Data Center Geometric Model

This section presents a baseline raised-floor room-level data center model. The facility is equipped with two precision air conditioning units and two rows of server racks totaling 12 racks, arranged in a cold-aisle containment configuration. The detailed geometric model is shown in Figure 1a,b. The ACUs are configured with supply airflow directed parallel to the cold aisle to better capture thermal differences between the left and right rack rows. Rack arrangement follows the NVIDIA DGX white paper [36]. Each row consists of five server racks and one switch rack. Each server rack houses four DGX A100 high-performance computing servers rated at 6.5 kW each and two standard servers rated at 225 W each. Each switch rack contains ten QM8790 switches, one AS4610 switch, and one SN4600 switch. Blank panels are installed in all vacant slot positions. The detailed server arrangement is shown in Figure 1c, where LE and RE denote the left and right switch racks; labels 1, 2, and 3 represent the QM8790, AS4610, and SN4600 switches; labels 4 and 5 denote the DGX A100 servers and standard servers, respectively.

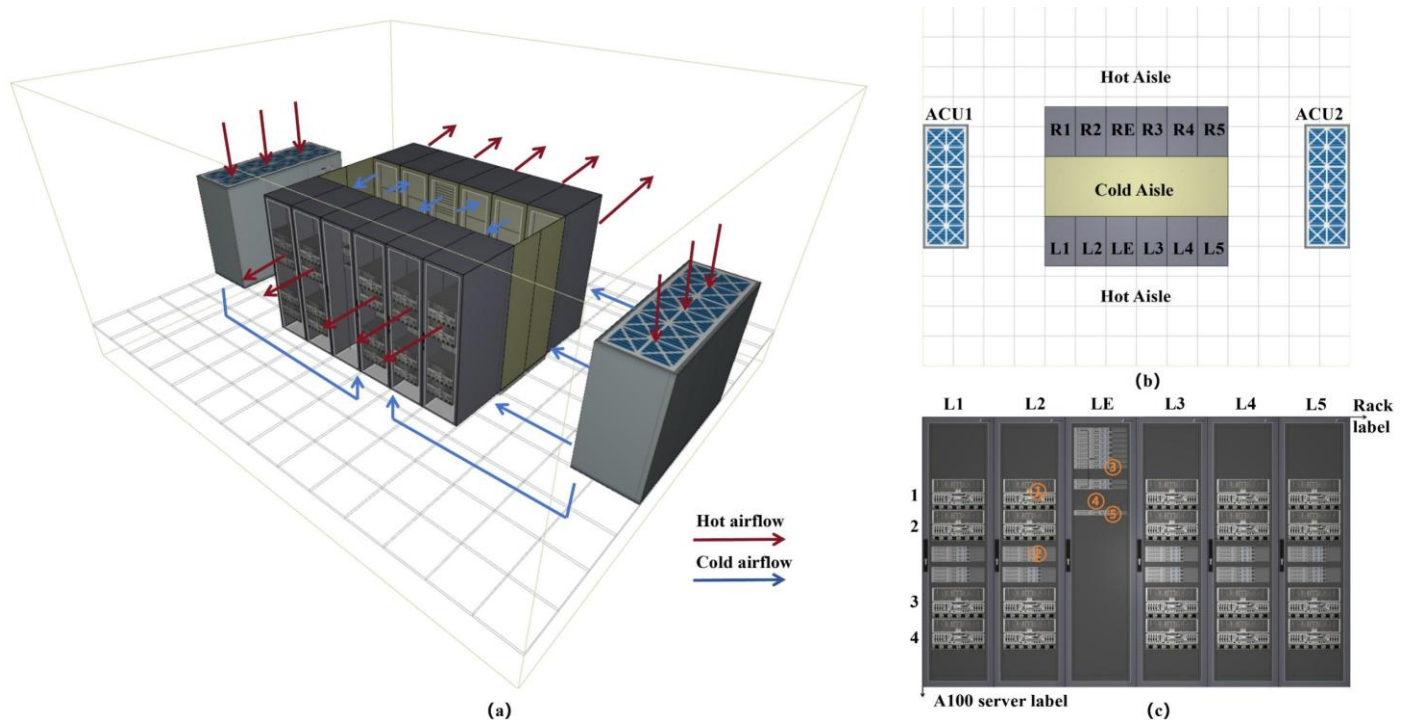


Figure 1. Layout of the studied data center: (a) 3D room view; (b) Top view and rack layout; (c) Rear view of rack and IT equipment arrangement.

Room parameters and IT equipment specifications are listed in Tables 1 and 2. Equipment parameters are sourced from official documentation and partially adapted for the rack configuration [37–39]. The AS4610 switch operates with a constant fan airflow rate of 58.63 L/s/kW. All other IT equipment adopts a PUE-based control strategy, with control curves derived from EnergyStar data. When the inlet temperature is below 27 °C, the fan airflow rate per unit load is 58.52 L/s. When the inlet temperature exceeds 27 °C, the fan airflow rate increases linearly with temperature, reaching 100.5 L/s per unit load at 35 °C.

Table 1. Main parameters of the studied DC.

Items	Description	Parameters
DC room	Area	60.48 m ²
	Room dimension	8.4 m (L) × 7.2 m (W) × 4.5 m (H)
	Air supply mode	Raised-floor cooling
	Aisle form	CAC&OHA
	Aisle dimension	3.6 m (L) × 1.2 m (W) × 2.322 m (H)
	Number of racks	12
	Plenum height	0.64 m
	Perforated tiles porosity	50%
	Perforated tiles dimension	0.6 m (L) × 0.6 m (W)
	Rack	0.997 m (L) × 0.6 m (W) × 2.322 m (H)
Rack	Dimension	0.997 m (L) × 0.6 m (W) × 2.322 m (H)
	Gap condition	Typical small gap
	empty slot	Closed
	Peak energy consumption of server racks	22.5 kW/Rack
ACU	Number	2
	Max cooling capacity	150 kW/unit
	Max airflow	12 m ³ /s
	Temperature control mode	supply air control
	Set supply air temperature	20 °C
	Airflow control mode	Fixed speed
	Fan speed	100%
	Power consumption	3.9 kW/unit

Table 2. Main parameters of IT equipment.

Equipment	Number	Energy Consumption (kW)	Size (L × W × H) (m)
DGX A100	40	6.5	0.8972×0.4445×0.2667(6U)
3U server	20	0.225	0.6×0.45×0.1334(3U)
QM8790 switch	20	0.2	0.6×0.45×0.0445(1U)
AS4610 switch	2	0.099	0.35×0.44×0.0445(1U)
SN4600 switch	2	0.2	0.6×0.45×0.0445(1U)

2.2. Server Energy-Consumption Models

Energy consumption ratio curves of the NVIDIA DGX A100 GPU under realistic operational scenarios are extracted from an existing GPU tracking study [33], as shown in Figure 2. The instantaneous power of a single DGX A100 server at time t is defined as Equation (1).

$$P_{\text{DGX}}(t) = P_{\text{rated}} \cdot r(t) = 6.5 \cdot r(t) \quad (1)$$

where $r(t) \in [0, 1]$ is the energy consumption ratio at time t , extracted from the dataset of Krzywaniak et al. [33], and $P_{\text{rated}} = 6.5$ kW is the rated maximum power of the DGX A100 server as specified in the NVIDIA datasheet [37].

The GEMM model represents the most widely used convolution algorithm in deep learning. It operates at a low load of approximately 5% energy ratio during the first 5 s before transitioning to a high-load phase. In the high-load phase, the energy ratio fluctuates between 75.8% and 85.3%, with a mean of 81.51% and a variance of 3.83%. The VGG-19 and ResNet-152 models (hereafter VGG and ResNet), both representative convolutional neural network architectures, enter their high-load phases at 124 s and 121 s, respectively. The VGG model exhibits an energy ratio ranging from 33.6% to 94.2%, with a mean of 70.25% and a variance of 287.4%. The ResNet model fluctuates between 20.4%

and 84.5%, with a mean of 52.14% and a variance of 185.9%². The GEMM model exhibits significantly smaller fluctuations and lower variance than the other two models. It is therefore classified as a stable energy consumption model, whereas VGG and ResNet are classified as fluctuating energy consumption models.

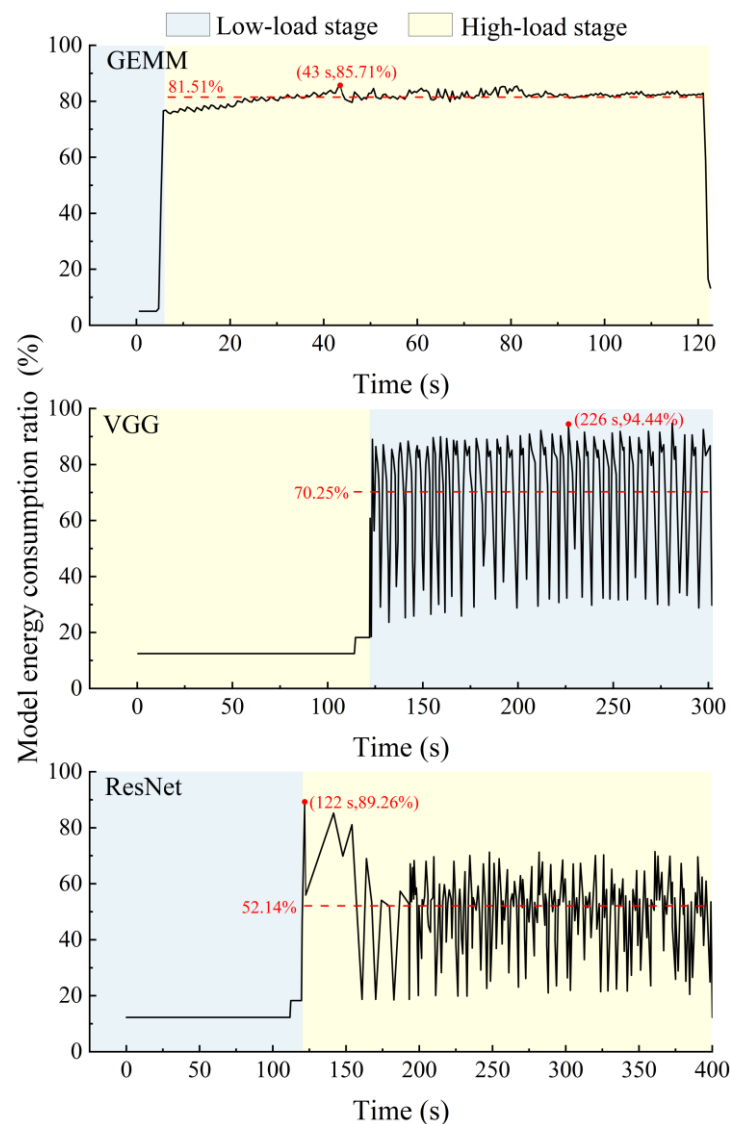


Figure 2. Energy consumption ratio curves of three artificial neural network models.

2.3. Case Configuration

Six cases were established by combining the three models introduced in Section 2.2, comprising two categories: single non-uniform load conditions and mixed non-uniform load conditions. The non-uniform load models were assigned to the DGX A100 servers, while the switch racks and standard servers within the server racks operated under constant load. The model configuration for each case is summarized in Table 3. Since model runtimes differ, the ResNet runtime of 400 s was adopted as the unified simulation duration, with the GEMM and VGG models repeatedly within this period. As shown in Table 3, for Cases 1, 3, and 4, which include the GEMM model, the proportion of the stable energy consumption model is 100%, 50%, and 60%, respectively (excluding switch racks). For the fully fluctuating Cases 2, 5, and 6, the VGG model proportion is 100%, 50%, and 60%, respectively.

Table 3. Server Rack Load Configuration.

Case	Load Type	End Racks, Left Row (L1, 5)	Middle Racks, Left Row (L2, 3, 4)	End Racks, Right Row (R1, 5)	Middle Racks, Right Row (R2, 3, 4)
1	Single load, stable	GEMM	GEMM	GEMM	GEMM
2	Single load, fluctuated	VGG	VGG	VGG	VGG
3	Mixed load, partially fluctuated	GEMM	GEMM	VGG	VGG
4	Mixed load, partially fluctuated	VGG	GEMM	VGG	GEMM
5	Mixed load, fully fluctu- ated	VGG	VGG	ResNet	ResNet
6	Mixed load, fully fluctu- ated	ResNet	VGG	ResNet	VGG

The energy consumption ratio curves for each case are shown in Figure 3. Based on the model energy consumption curves, the load operation process for each case is divided into three stages: a low-load stage, in which all models operate at low load; a high-load stage, in which all models operate at high load; and a mid-load stage covering the remaining period. This study focuses primarily on the high-load stage, as it captures the energy consumption fluctuation characteristics of all three models.

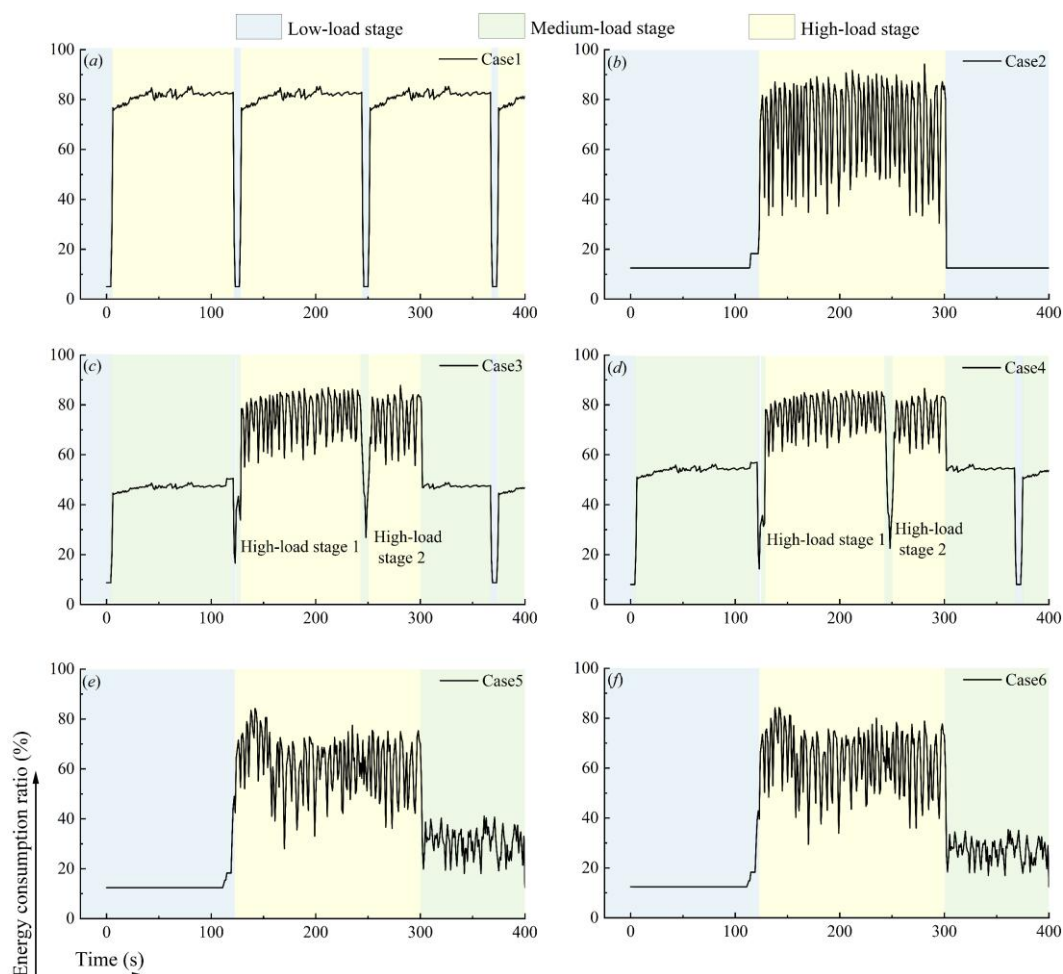


Figure 3. Energy consumption ratio curves for different cases. Subfigures (a) – (f) correspond to Case1 – Case6, respectively.

Table 4 presents the high-load phase time intervals and energy consumption ratio characteristics. During 129–244 s and 252–301 s, all cases are in the high-load stage. For Cases 1, 3, and 4, replacing GEMM with VGG decreases the mean energy consumption ratio and substantially increases its variance during 129–244 s. For Cases 2, 5, and 6, replacing VGG with ResNet decreases both the mean and variance during 124–301 s. The greater the number of substituted models, the larger the magnitude of change.

Table 4. Characteristics of the high-load stages of energy consumption ratio curves for different cases.

Case	Load Stage	Duration	Mean Energy Consumption Ratio (%)	Energy Consumption Ratio Variance (%) ²
1	High-load stage	6–121 ,129–244 ,252–367 s	81.85	3.83
2	High-load stage	124–301 s	70.25	287.4
3	High-load stage 1	129–244 s	76.02	72.02
	High-load stage 2	252–301 s	74.56	79.56
4	High-load stage 1	129–244 s	77.12	46.93
	High-load stage 2	252–301 s	75.70	51.37
5	High-load stage	124–301 s	62.05	130.49
6	High-load stage	124–301 s	63.69	143.09

2.4. Simulation Settings, Model Validation and Independence Analysis

Modeling, simulation, and post-processing were conducted using Data Center Design Pro 2023.2 (6SigmaRoom®, Future Facilities Ltd.), a dedicated simulation software for data center airflow organization. The solver resolves the three-dimensional coupled continuity, momentum, and energy equations using the finite-volume method (FVM) [40,41]. A finite-volume, pressure-based SIMPLE algorithm is employed for steady-state initialisation, followed by a segregated transient solver with first-order implicit time integration for the transient phase. The standard k - ε turbulence model was employed, as it has been extensively validated for raised-floor data center airflow simulations and is the recommended turbulence model in dedicated data center CFD tools [42,43], with transport equations for turbulent kinetic energy k and turbulent dissipation rate ε given by Equations (2) and (3).

$$\frac{\partial(\rho k)}{\partial t} + \frac{\partial(\rho k u_i)}{\partial x_i} = \frac{\partial}{\partial x_j} \left[\left(\mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] + G_k - \rho \varepsilon \quad (2)$$

$$\frac{\partial(\rho \varepsilon)}{\partial t} + \frac{\partial(\rho \varepsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_j} \left[\left(\mu + \frac{\mu_t}{\sigma_\varepsilon} \right) \frac{\partial \varepsilon}{\partial x_j} \right] + C_{1\varepsilon} \frac{\varepsilon}{k} G_k - C_{2\varepsilon} \rho \frac{\varepsilon^2}{k} \quad (3)$$

where the turbulent viscosity is $\mu_t = \rho C_\mu k^2 / \varepsilon$, with model constants $C_\mu = 0.09$, $C_{1\varepsilon} = 1.44$, $C_{2\varepsilon} = 1.92$, $\sigma_k = 1.0$, $\sigma_\varepsilon = 1.3$. Standard wall functions were applied at all solid boundaries, with the non-dimensional wall distance y^+ maintained in the range of 30–300.

A grey-box modeling strategy was adopted for all equipment, with boundary conditions defined as follows. For the ACUs, velocity inlet and pressure outlet boundary conditions are applied. The ACU operates with a fixed fan speed of 100%, maintaining a constant supply airflow rate $\dot{m}_{\text{air}}$ throughout the simulation. The cooling capacity is modulated exclusively through compressor-side adjustment to maintain the supply air temperature at the setpoint $T_{\text{supply, set}}$. The instantaneous cooling capacity is computed as Equation (4) [44].

$$Q_{\text{cool}}(t) = \dot{m}_{\text{air}} c_p \left[T_{\text{return}}(t - \Delta t_{\text{lag}}) - T_{\text{supply,set}} \right], \quad Q_{\text{cool}}(t) \leq Q_{\text{cool,max}} \quad (4)$$

where c_p is the specific heat of air, $T_{\text{return}}(t - \Delta t_{\text{lag}})$ is the return air temperature after the response time Δt_{lag} , and $Q_{\text{cool,max}}$ is the rated maximum cooling capacity per unit. The response time Δt_{lag} represents the total delay between a change in server heat load and the corresponding adjustment in cooling output, arising from the combined effect of air transit time from the server racks to the return-air sensor, sensor thermal inertia, and cooling coil thermal mass. In this study, Δt_{lag} is set to 10 s, determined with reference to [34,35] and taking into account the compact room dimensions, the higher-than-average supply airflow rate resulting from the elevated IT heat flux density, and the large-amplitude rapid fluctuations of the AI workload power profiles. When the required cooling capacity reaches $Q_{\text{cool,max}}$, the compressor output is locked at this limit and the supply air temperature rises above the setpoint until the return air temperature signal decreases, after which the cooling capacity adjusts downward accordingly.

The perforated floor tiles were modeled using a porous jump model. The pressure drop across the tiles is computed using the Darcy–Forchheimer equation [45], as given by Equation (5).

$$\Delta p = \frac{\mu}{\alpha} v \Delta m + C_2 \frac{\rho}{2} v^2 \Delta m \quad (5)$$

where v is the face-normal velocity, Δm is the tile thickness, α is the permeability, and C_2 is the inertial resistance coefficient. Resistance coefficients were taken from the software's built-in library for a tile opening ratio of 50%, consistent with Table 1.

The ends and top of the cold aisle, as well as the facility walls, were set as adiabatic surfaces. Each server was modeled as a volumetric heat source; the DGX A100 transient heat dissipation follows Equation (1), with the energy consumption ratio curves shown in Figure 3.

To validate the accuracy of the proposed numerical model, the simulated temperature and velocity distributions were compared with experimental measurements at six monitoring locations. The corresponding error analysis is summarized in Table 5, and the detailed comparison is shown in Figure 4. As indicated in Table 5, the simulation results agree well with the experimental data and accurately reproduce the spatial variation trends of both temperature and airflow velocity. For temperature, the mean absolute error (MAE), root-mean-square error (RMSE), and maximum absolute deviation are 0.283 °C, 0.303 °C, and 0.400 °C, respectively. For velocity, the corresponding values are 0.0133 m/s, 0.0163 m/s, and 0.0300 m/s, respectively. The relative errors at all measurement locations are within 10%, and most are below 8%. These results demonstrate that the developed numerical model has sufficient predictive accuracy and can be reliably used for subsequent transient simulations and analysis.

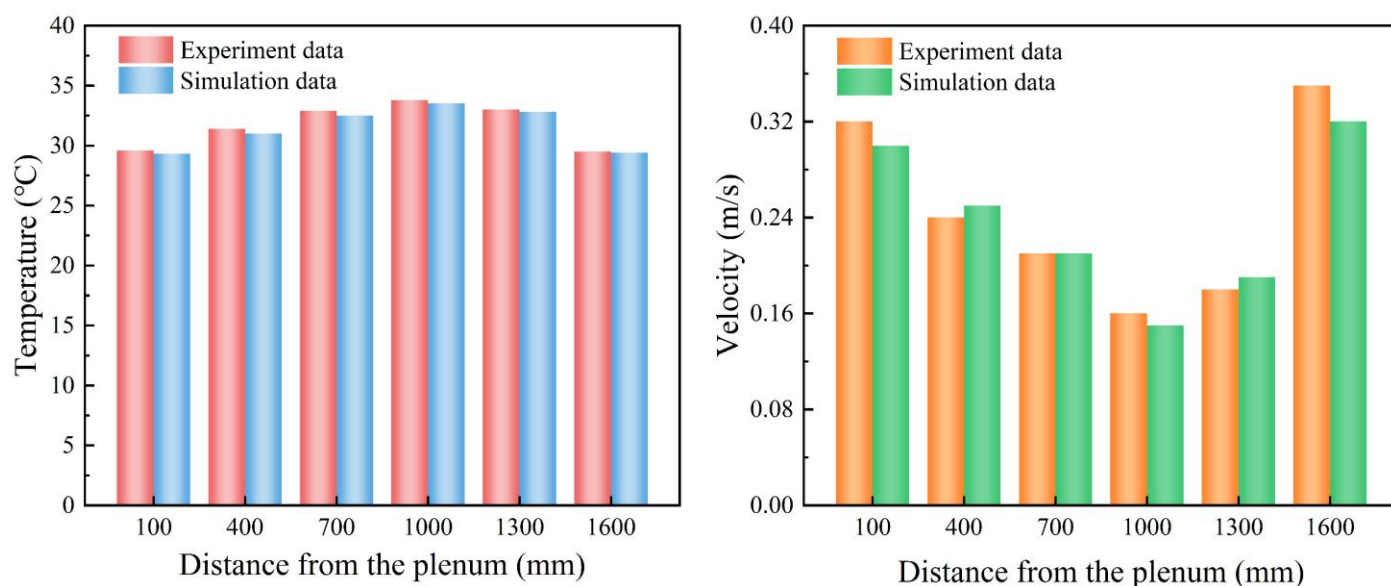


Figure 4. Experimental [46] and simulated values for outlet-air temperatures and velocities of one tested rack.

Table 5. Comparison of experimental and simulated temperatures at different measurement locations.

Location (mm)	Temperature (Exp.) (°C)	Temperature (Sim.) (°C)	Relative Deviation (%)	Velocity (Exp.) (m/s)	Velocity (Sim.) (m/s)	Relative Deviation (%)
100	29.6	29.3	1.01	0.32	0.30	6.25
400	31.4	31.0	1.27	0.24	0.25	4.17
700	32.9	32.5	1.22	0.21	0.21	0.00
1000	33.8	33.5	0.89	0.16	0.15	6.25
1300	33.0	32.8	0.61	0.18	0.19	5.56
1600	29.5	29.4	0.34	0.35	0.32	8.57

Temperature: MAE = 0.283 °C, RMSE = 0.303 °C, Maximum absolute deviation = 0.400 °C.

Velocity: MAE = 0.0133 m/s, RMSE = 0.0163 m/s, Maximum absolute deviation = 0.0300 m/s.

To ensure that the numerical results were insensitive to mesh resolution, a grid-independence study was first conducted using the Grid Convergence Index (GCI) method proposed by Roache [47]. Three mesh configurations were generated by adjusting the maximum cell size in the x -, y -, and z -directions, with maximum mesh sizes of 90, 80, and 70 mm for Cases 01, 02, and 03, respectively. The maximum indoor temperature was selected as the reference parameter for the mesh convergence assessment. As listed in Table 6, the GCI values for all adjacent mesh pairs are below 3% [48], indicating that the computational results are independent of mesh density. Therefore, the mesh with 651,640 elements was selected for the subsequent simulations.

Table 6. Grid independence verification.

Case	Grid Number	Reference Parameter (°C)	Adjacent Grid Refinement Ratio	Convergence Order	Adjacent Grid GCI Index (%)
01	651640	32.6	1.078	13.09	0.453
02	817044	32.8	1.151	13.09	0.072
03	1246476	32.9	-	-	-

After confirming mesh independence, a time-step independence study was further performed to verify the temporal accuracy of the transient simulations. Three time steps of 1 s, 0.5 s, and 0.25 s were evaluated following the same GCI-based procedure, with adjacent refinement ratios of 2.0. The maximum rack outlet temperature during the high-load stage of Case 2 was selected as the reference parameter. As shown in Table 7, the GCI values for all adjacent time-step pairs are far below 3%, confirming that the numerical results are independent of temporal discretization. Therefore, a time step of 1 s was adopted for all subsequent simulations. The key simulation parameter settings adopted in this study are summarised in Table 8.

Table 7. Time-step independence verification.

Case	Time Step (s)	Reference Parameter (°C)	Adjacent Time-Step Refinement Ratio	Convergence Order	Adjacent-Time-Step GCI Index (%)
001	1	30.138	2.0	2.59	0.0050
002	0.5	30.144	2.0	2.59	0.0008
003	0.25	30.145	-	-	-

Table 8. Simulation parameter settings.

Parameter	Detail
Flow Regimes	Turbulent Flow
Transient Simulation	On
Total Simulation Time	400 s
Time Step Size	1 s
Maximum iterations per time step	2000
ACU response time	10 s
Turbulence Model	Standard $k-\varepsilon$ model
Wall Functions	Standard Wall Function

3. Results

3.1. Cooling Capacity

Figure 5 presents the dynamic cooling-capacity curves for the six cases. In all cases, the ACU cooling capacity closely tracks the variation in server energy consumption. During the low-load stage, both ACUs maintain a low and stable output. This is consistent with the low and steady server power draw. As servers enter the high-load stage, energy consumption rises sharply and cooling capacity increases accordingly. The two ACU outputs then begin to diverge. This reflects the asymmetric return air temperatures arising from the non-uniform rack load distribution. Once divergence occurs, the rate of increase gradually slows. During the intermediate phase of the high-load stage, the cooling capacity reaches a quasi-steady elevated level. This applies to both stable and fluctuating load cases. For Case 1, this stable elevated phase lasts notably longer in the second and subsequent high-load cycles than in the first. The difference arises because the initial ACU response overshoots before settling. For Cases 2–6, continuous low-amplitude oscillations persist throughout the high-load stage. These oscillations reflect the periodic fluctuations in rack energy consumption. The onset of the cooling capacity rise lags approximately 10 s behind the sharp rise in server energy consumption. This is consistent with the ACU response time specified in Table 8.

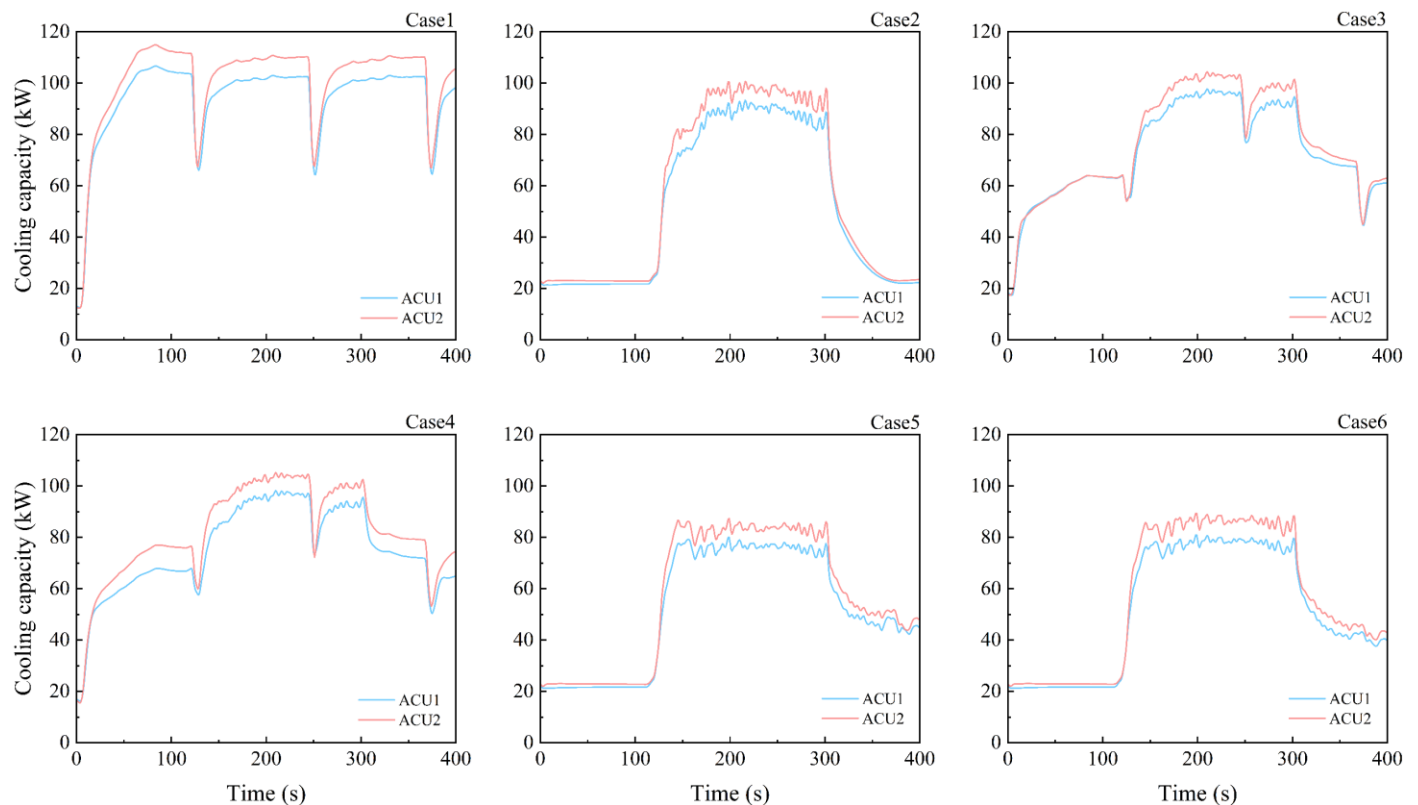


Figure 5. Dynamic cooling capacity curves for different cases.

3.2. Inlet Temperature

The mean rack inlet temperature curves for the six cases are shown in Figure 6. Three distinct variation patterns are identified. Case 1 represents the stable energy consumption case. Cases 3 and 4 are the partially fluctuating cases. Cases 2, 5, and 6 are the fully fluctuating cases.

For Case 1, the GEMM model enters its high-load stage at 6 s. Rack heat dissipation rises sharply. Due to the ACU response lag, the inlet temperature rises rapidly and reaches a maximum of 22.35 °C at 46 s. As the supply air temperature exceeds the setpoint, the ACUs progressively increase their cooling output, gradually pulling the inlet temperature back toward 20 °C. At 122 s, the GEMM model transitions to its low-load stage, causing a sudden drop in total energy consumption. The inlet temperature falls sharply, reaching a minimum of 19.23 °C at 136 s. As the inlet temperature drops below the setpoint, the ACU cooling output decreases, while the GEMM model simultaneously re-enters its high-load stage, causing the temperature to recover. During the second and subsequent operating cycles, the inlet temperature variation closely follows the same pattern, with a maximum of approximately 20.14 °C. This suggests that for cases running a single load model periodically over extended periods, the inlet temperature trend stabilizes from the second cycle onward.

For Case 2, the VGG model enters its high-load stage at 124 s, driving the inlet temperature to a maximum of 21.86 °C at 164 s. The ACUs respond by increasing cooling capacity, gradually reducing the inlet temperature to 20.25 °C by 303 s. At 303 s, the VGG model transitions to its low-load stage, causing the inlet temperature to drop sharply to a minimum of 18.19 °C at 339 s, before recovering under ACU regulation.

Cases 1, 3, and 4 are compared in Figure 6a. Cases with a higher GEMM server proportion reach higher maximum inlet temperatures. They also reach these maxima more rapidly. This is because greater total energy consumption produces faster heat accumulation. During the temperature decline phase, cases with higher energy consumption show

larger deviations from the setpoint. This leads to stronger ACU regulation and lower minimum inlet temperatures.

Cases 2, 5, and 6 show minimal differences during 0–302 s. This is attributable to the similar dynamic energy profiles of the VGG and ResNet models. After 302 s, the VGG model enters its low-load stage and the inlet temperature drops sharply. The magnitude of this drop increases with the number of VGG-loaded servers. From Cases 2 to 6, the temperature drops are 2.06, 1.43, 1.15, 1.51, and 1.66 °C, respectively.

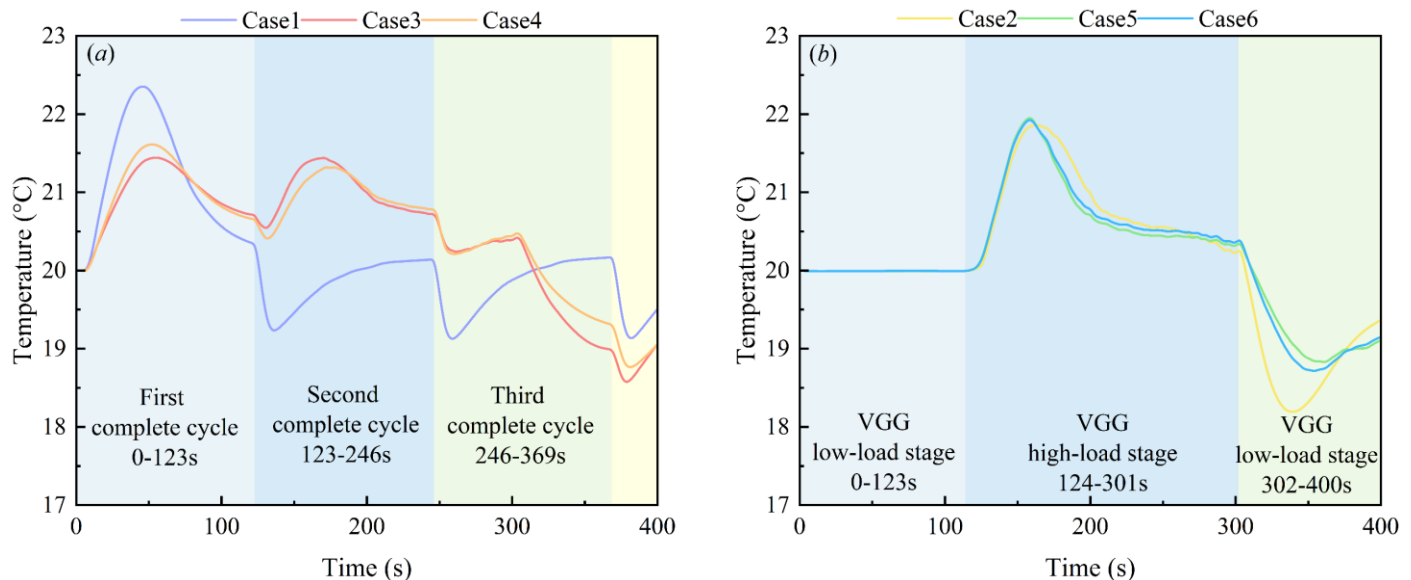


Figure 6. Mean rack inlet temperatures for different cases; (a) cases containing the stable energy consumption model; (b) cases without the stable energy consumption model.

3.3. Outlet Temperature

Figure 7 shows the dynamic rack outlet temperature curves under the two single-load cases. For Case 1, the GEMM model enters its high-load stage at 5 s, leading to a rapid increase in outlet temperature and a peak value of 33.17 °C at 44 s (taking rack R5 as a representative example, minor inter-rack differences exist). The temperature subsequently declines with small fluctuations for the remainder of the high-load stage. In the second and third high-load cycles, the outlet temperature stabilizes within 29.23–30.85 °C, without the pronounced peak observed in the first cycle. This behavior mirrors the inlet temperature pattern shown in Figure 7a, confirming that inlet temperature is the primary driver of the overall outlet temperature trend. In terms of dynamic response, the near-step-change energy profile of the GEMM model produces a sharp transient rise followed by a gradual return toward equilibrium. This is consistent with the rack outlet temperature response to abrupt load changes reported by Ibrahim et al. [35].

For Case 2, the VGG model transitions to high load at 124 s, triggering a rapid rise in outlet temperature. Throughout the high-load stage, the periodic fluctuations in VGG energy consumption drive sustained oscillations in the mean rack outlet temperature. The mean peak temperature increases from 30.69 °C at 129 s to 32.78 °C at 168 s, tracking the overall inlet temperature trend. After 300 s, the outlet temperature variation during the subsequent low-load stage deviates from the initial low-load behavior, owing to ACU-regulated changes in inlet temperature. The outlet temperature oscillations closely track the energy consumption ratio of the VGG model. This dynamic correspondence is consistent with the behavior reported in [49].

Temperature contours were extracted at representative time points—the peak and trough of Case 1 (43 s and 373 s) and a local peak and trough within the high-load stage

of Case 2 (168 s and 188 s)—to reveal the spatial evolution of the thermal field. The spatial extent of the high-temperature zone at the rack outlets is strongly correlated with instantaneous energy consumption. Under high-load conditions, prominent high-temperature zones develop at both left and right rack outlets. In Case 2, as the energy consumption ratio transitions from a local peak to a local trough, the high-temperature zone contracts markedly. Under low-load conditions, no significant high-temperature zones are observed.

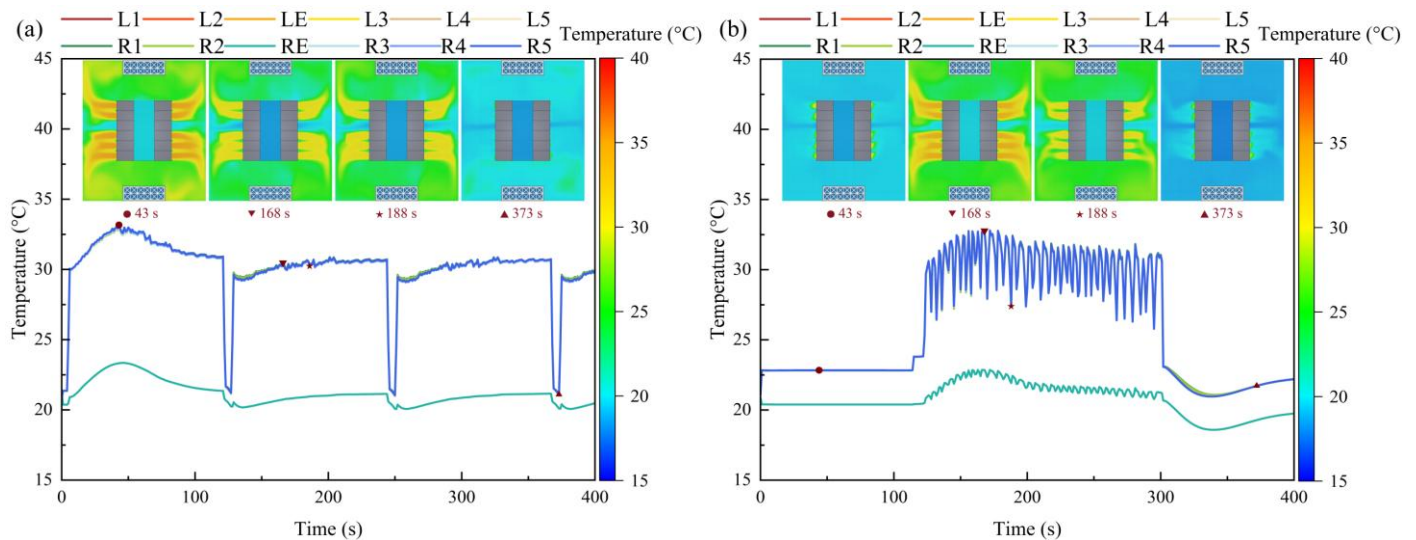


Figure 7. Mean rack outlet temperatures and temperature contours ($y = 1.66$ m) under single non-uniform load conditions; (a) Case 1; (b) Case 2.

To quantitatively characterize the dynamic thermal behavior of rack outlet temperatures, two metrics are defined. The mean fluctuation amplitude is defined as the difference between the mean peak temperature and mean trough temperature of the outlet temperature curve. The temperature change rate is the temperature difference between adjacent peaks and troughs divided by the corresponding time interval. To enable direct comparison of VGG-loaded rack dynamics across the partially and fully fluctuating cases, metrics are evaluated over 129–244 s (high-load stage 1) and 124–301 s (high-load stage), respectively. The quantitative results are summarised in Table 9. Case 1 exhibits near-zero fluctuation, with a maximum temperature change rate of only 0.37 °C/s. In contrast, Case 2 shows a mean fluctuation amplitude of 3.72 °C (129–244 s) and a maximum temperature change rate of 4.99 °C/s, both substantially exceeding the values recorded for Case 1.

Table 9. Dynamic outlet temperature characteristics of server racks during the high-load stage under single load cases (taking rack R5 as an example).

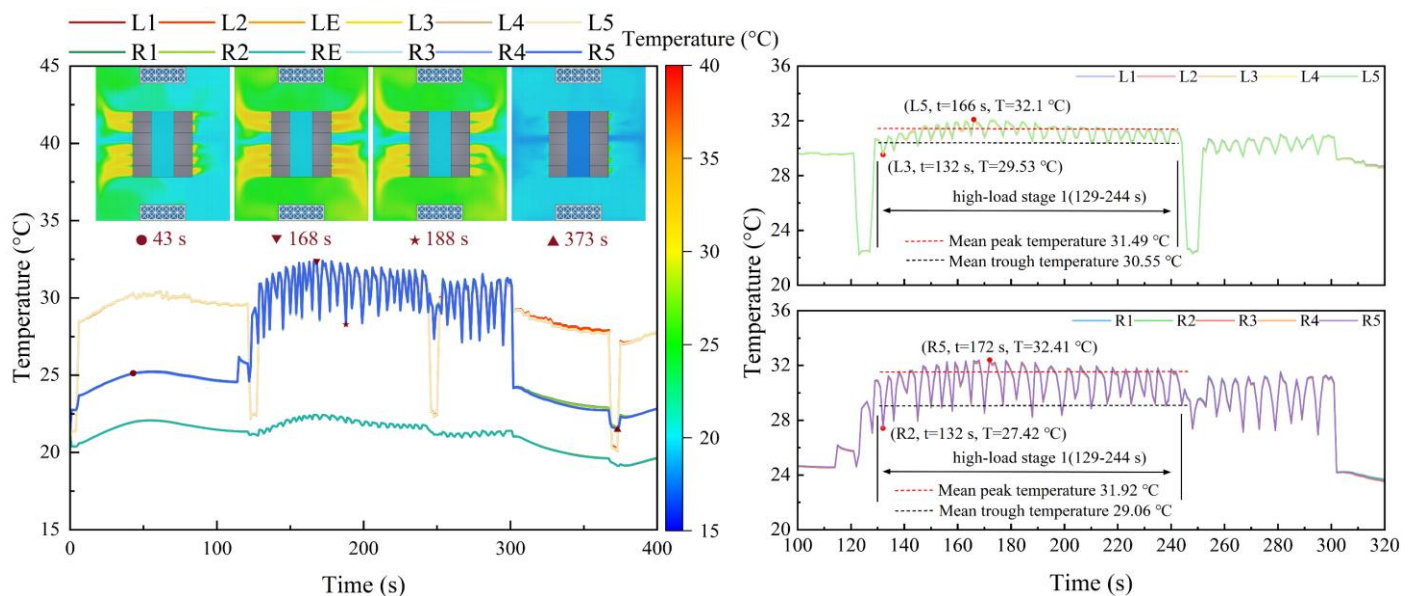
Case	Dynamic Characteristic	Value	Notes
1(129–244 s)	Maximum temperature change rate	0.37 °C/s	From 189 s– 30.67 °C to 190 s– 30.20 °C
	Minimum temperature change rate	0 °C/s	-
2(129–244 s)	Mean fluctuation amplitude	3.72 °C	Mean peak temperature: 32.07 °C Mean trough temperature: 28.35 °C
	Maximum temperature change rate	4.99 °C/s	From 188 s– 27.39 °C to 189 s– 32.38 °C

2(124–301 s)	Minimum temperature change rate	0.64 °C/s	From 138 s–31.87 °C to 141 s–29.94 °C
	Mean fluctuation amplitude	3.84 °C	Mean peak temperature: 31.81 °C
	Maximum temperature change rate	4.99 °C/s	Mean trough temperature: 27.97 °C
	Minimum temperature change rate	0.64 °C/s	From 188 s–27.39 °C to 189 s–32.38 °C
			From 138 s–31.87 °C to 141 s–29.94 °C

Figures 8 and 9 show the rack outlet temperature curves for the mixed non-uniform load cases. Because the three models transition to high-load operation at different time points, the outlet temperature curves exhibit multiple distinct stages across all four cases.

For the partially fluctuating cases, three distinct variation patterns emerge based on rack load assignment. GEMM-loaded racks exhibit five high-temperature stages, driven by the superimposed GEMM and VGG energy profiles. During 6–121 s and 129–244 s, outlet temperatures follow an increasing-then-decreasing trend under inlet temperature influence, with minor VGG-induced oscillations in the latter interval. At 301 s, the VGG model transitions to low load and the outlet temperature drops sharply. A brief GEMM low-load transition then divides the 302–400 s interval into two high-temperature sub-stages. VGG-loaded racks follow a similar five-stage pattern, with lower absolute temperatures during GEMM-dominated intervals and large-amplitude oscillations during 124–301 s. The GEMM high–low–high load transition at 244–252 s introduces two distinct high-fluctuation periods within this interval.

At 43 s, the high-temperature zone at the left rack outlets in Case 3 is marginally smaller than in Case 1. The zone at the right rack outlets is substantially reduced but remains larger than under low-load conditions. Two factors account for these differences: lower total instantaneous energy consumption in Case 3, and the higher low-load energy ratio of the VGG model (12.48%) relative to GEMM (5%). Within the same case, the extent of the high-temperature zone is strongly correlated with instantaneous energy consumption, with large fluctuations driving repeated cycles of expansion and contraction. The maximum and minimum outlet temperatures during the high-load stage under mixed operating conditions, together with their corresponding time instants and rack identifiers, are highlighted in red in Figures 8 and 9.



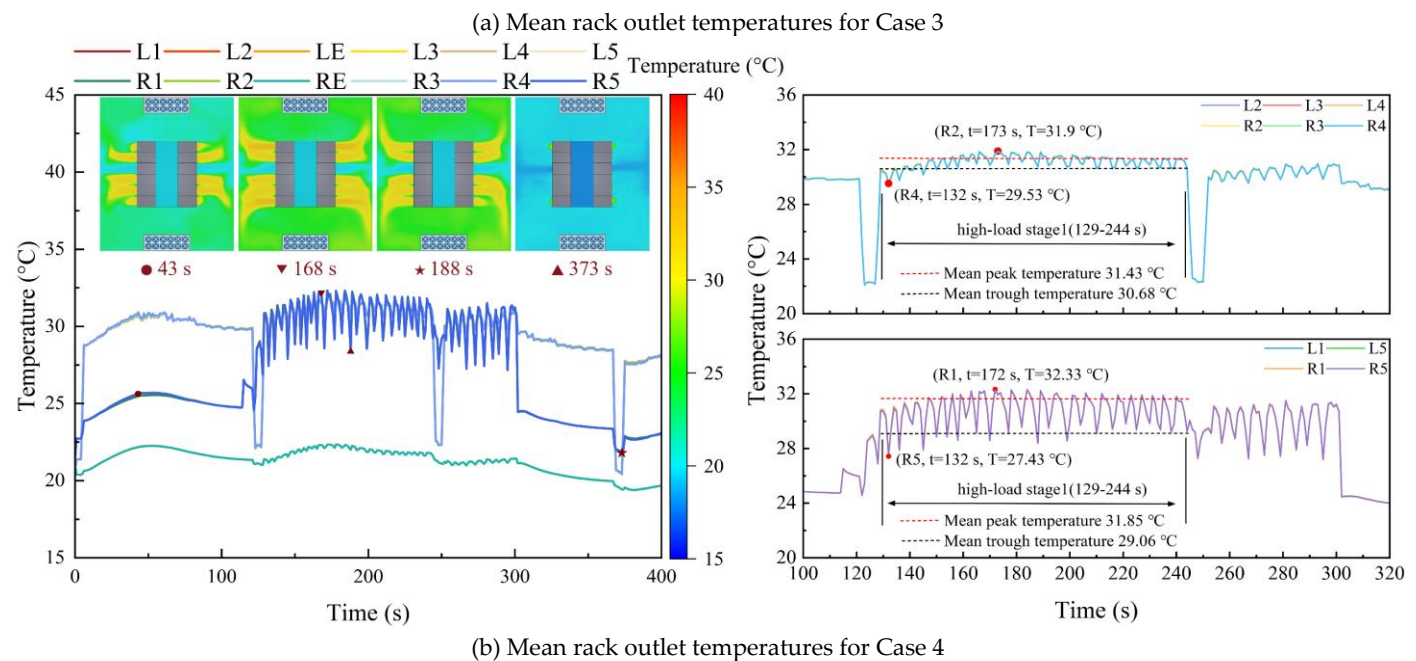


Figure 8. Mean rack outlet temperatures and temperature contours ($y = 1.66$ m) under mixed non-uniform load conditions with partial fluctuation; left: full simulation period; right: high-load stage.

Tables 10 and 11 quantify the dynamic outlet temperature characteristics of GEMM-loaded and VGG-loaded racks, respectively, during the high-load stage under partially fluctuating conditions. For the GEMM-loaded racks (Table 10), Cases 3 and 4 exhibit mean fluctuation amplitudes of 0.94 °C and 0.75 °C, respectively. Both values substantially exceeding the near-zero fluctuation of Case 1 over the same interval. These elevated fluctuations arise from the indirect thermal influence of the neighboring VGG-loaded racks through the shared cold aisle. Case 3 shows a larger increase than Case 4, attributable to its higher VGG model proportion (50% vs. 40%). The maximum temperature change rates of 1.33 °C/s and 1.09 °C/s in Cases 3 and 4 are also higher than the 0.37 °C/s recorded for Case 1. For the VGG-loaded racks (Table 11), the mean fluctuation amplitudes are 2.86 °C and 2.79 °C for Cases 3 and 4, compared with 3.72 °C for Case 2. All three dynamic metrics decrease progressively as the VGG model proportion drops from 100% in Case 2 to 50% and 40% in Cases 3 and 4. The mean fluctuation amplitudes in Cases 3 and 4 are reduced by 0.86 °C and 0.93 °C relative to Case 2, respectively. Notably, the maximum temperature change rate occurs at the same time point (188–189 s) across all three cases. This confirms that the instantaneous rate of temperature change is primarily governed by the energy consumption ratio rather than by the load composition.

Table 10. Dynamic characteristics of rack outlet temperature for the GEMM model during the high-load stage under partially fluctuating load conditions (taking rack L2 as an example).

Case	Dynamic Characteristic	Value	Notes
3(129–244 s)			Mean peak temperature: 31.49 °C
	Mean fluctuation amplitude	0.94 °C	Mean trough temperature: 30.55 °C
	Maximum temperature change rate	1.33 °C/s	From 188 s– 30.52 °C to 189 s– 31.85 °C
	Minimum temperature change rate	0.18 °C/s	From 138 s– 31.77 °C to 141 s– 31.24 °C

4(129–244 s)	Mean fluctuation amplitude	0.75 °C	Mean peak temperature: 31.43 °C
	Maximum temperature change rate	1.09 °C/s	Mean trough temperature: 30.68 °C
	Minimum temperature change rate	0.15 °C/s	From 188 s–30.78 °C to 189 s–31.87 °C
			From 138 s–30.84 °C to 141 s–30.54 °C

Table 11. Dynamic characteristics of rack outlet temperature for the VGG model during the high-load stage under partially fluctuating load conditions (taking rack R5 as an example).

Case	Dynamic Characteristic	Value	Notes
3(124–301 s)	Mean fluctuation amplitude	2.86 °C	Mean peak temperature: 31.92 °C
	Maximum temperature change rate	3.95 °C/s	Mean trough temperature: 29.06 °C
	Minimum temperature change rate	0.51 °C/s	From 188 s–28.27 °C to 189 s–32.22 °C
			From 138 s–31.57 °C to 141 s–30.04 °C
4(124–301 s)	Mean fluctuation amplitude	2.79 °C	Mean peak temperature: 31.85 °C
	Maximum temperature change rate	3.84 °C/s	Mean trough temperature: 29.06 °C
	Minimum temperature change rate	0.51 °C/s	From 188 s–28.38 °C to 189 s–32.22 °C
			From 138 s–31.29 °C to 141 s–29.76 °C

Figure 9 shows the rack outlet temperature curves for the fully fluctuating cases (Cases 5 and 6). Three distinct stages are observed. The VGG model transitions from high to low load at 301 s, defining the stage boundaries. During 0–121 s, both models operate at comparable low-load energy ratios. Outlet temperatures across all racks are nearly identical and stable. During 124–301 s, large-amplitude oscillations appear in VGG-loaded rack outlet temperatures. These oscillations are superimposed on a broader increasing-then-decreasing trend, consistent with Case 2 and with the behavior reported in [49]. At 301 s, the load transition causes an abrupt drop in outlet temperature across all racks. Subsequently, ResNet-loaded rack outlet temperatures fluctuate at a reduced level. They follow an overall decreasing-then-increasing trend, reflecting the combined influence of inlet temperature variation and the ResNet model energy profile. After 301 s, the mean outlet temperature decreases relative to the 124–301 s interval despite continued fluctuations.

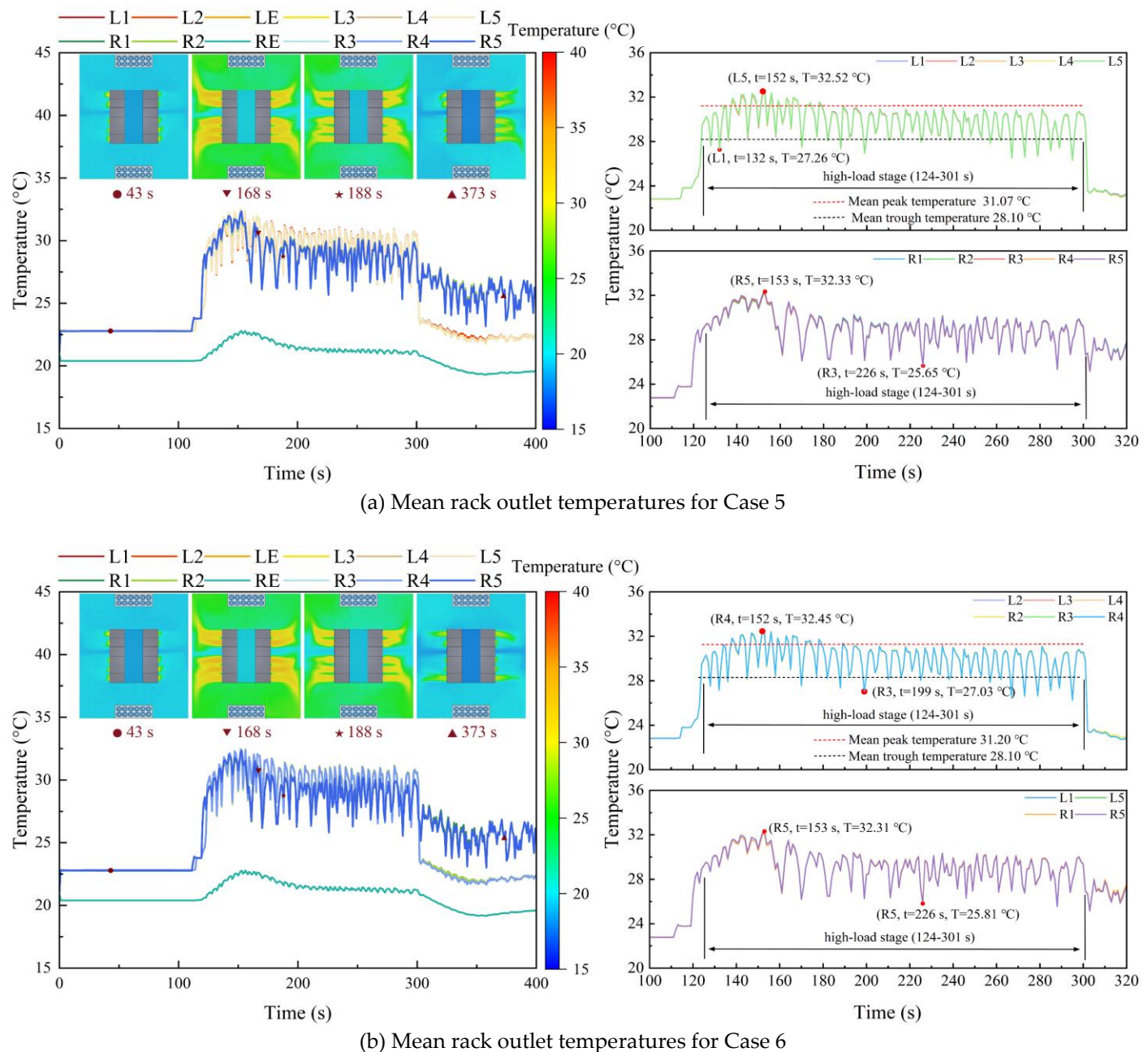


Figure 9. Mean rack outlet temperatures and temperature contours ($y = 1.66$ m) under mixed non-uniform load conditions with full fluctuation; left: full simulation period; right: high-load stage.

The dynamic outlet temperature characteristics of the VGG-loaded racks during the high-load stage under fully fluctuating conditions are summarised in Table 12, with rack L2 taken as the representative example. Comparing Cases 5 and 6 with Case 2, which operates under a single VGG load, the mean fluctuation amplitudes are 2.97°C and 3.10°C , respectively, both lower than the 3.84°C recorded for Case 2 over the same interval (124–301 s). As the proportion of the ResNet model increases from 0% in Case 2 to 40% in Case 6 and 50% in Case 5, the VGG model proportion decreases accordingly, and all three dynamic characteristics of the VGG-loaded rack outlet temperature decrease in sequence. This pattern is consistent with that observed under the partially fluctuating cases. These results indicate that, regardless of whether energy-stable or lower-amplitude fluctuation models are introduced into a fluctuating load configuration, the presence of such models reduces the dynamic thermal response of the VGG-loaded racks through the combined effect of a reduced total energy fluctuation level and a lower mean energy consumption ratio.

From a design and operational perspective, these results suggest that servers with highly fluctuating AI workloads should not be concentrated within the same rack row or thermal zone. Mixing highly fluctuating servers with stable-load, constant-load, or lower-fluctuation servers can reduce the room-level load fluctuation intensity, thereby weakening the dynamic outlet-temperature response and reducing the variation in cooling demand. This arrangement improves the temporal matching between server heat dissipation and air-conditioning cooling capacity, which is particularly important when the cooling system has a finite response delay.

In terms of reliability, large outlet-temperature fluctuation amplitudes and high temperature change rates indicate stronger transient thermal stress and a greater risk of temporary hotspot formation near rack outlets and hot aisles. Therefore, workload-aware rack arrangement and adaptive cooling control are necessary for AI-oriented data centers with rapidly varying server loads.

Table 12. Dynamic characteristics of rack outlet temperature for the VGG model during the high-load stage under fully fluctuating load conditions (taking rack L2 as an example).

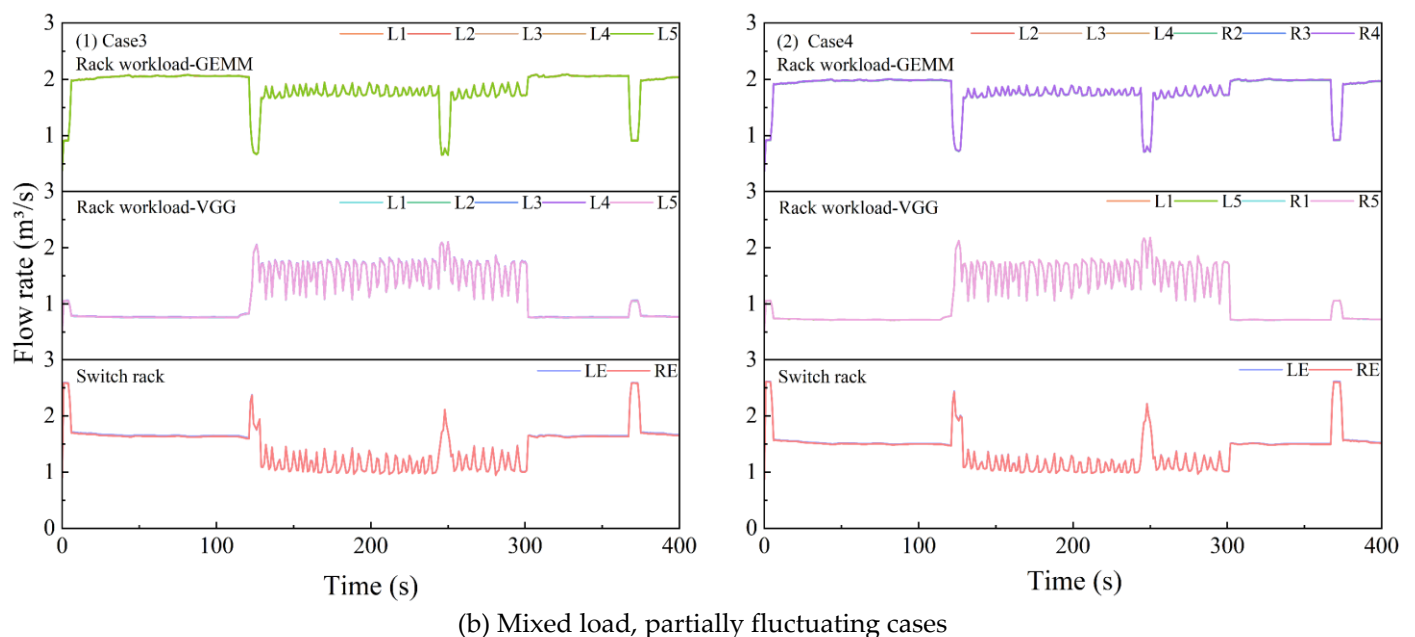
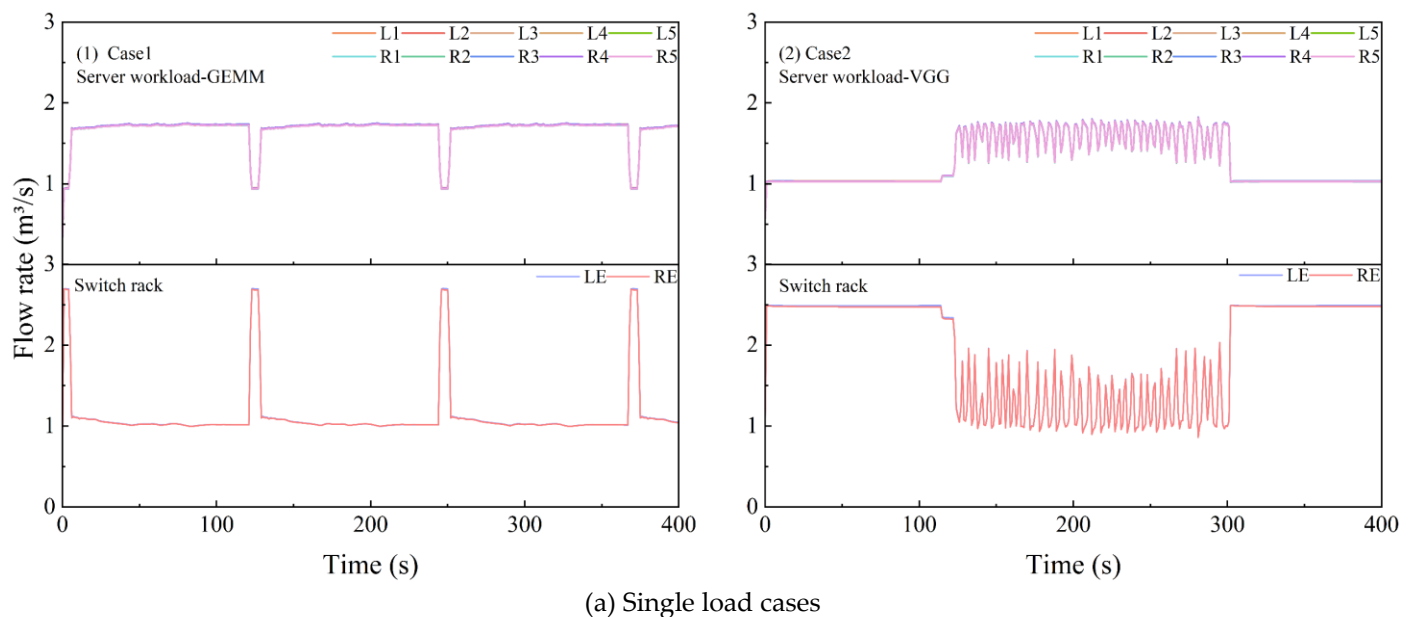
Case	Dynamic Characteristic	Value	Notes
5(124–301 s)			Mean peak temperature: 31.07 °C
	Mean fluctuation amplitude	2.97 °C	Mean trough temperature: 28.10 °C
	Maximum temperature change rate	3.77 °C/s	From 188 s–27.42 °C to 189 s–31.19 °C
	Minimum temperature change rate	0.46 °C/s	From 138 s–31.76 °C to 141 s–30.37 °C
6(124–301 s)			Mean peak temperature: 31.20 °C
	Mean fluctuation amplitude	3.10 °C	Mean trough temperature: 28.10 °C
	Maximum temperature change rate	3.99 °C/s	From 188 s–27.41 °C to 189 s–31.40 °C
	Minimum temperature change rate	0.49 °C/s	From 138 s–31.75 °C to 141 s–30.29 °C

3.4. Flow Rate

The dynamic net airflow rate histories for all six cases are shown in Figure 10. Here, the net airflow rate is defined as the surface integral of the velocity component normal to the rack outlet face over the entire outlet area. A positive value indicates that air is being exhausted from the rack into the hot aisle. For both server racks and switch racks, the net airflow rate represents the sum of the net airflow contributions of all IT equipment deployed within the rack, computed using the same surface integral definition.

In the single load cases (Figure 10a), the net airflow rate curves of both GEMM- and VGG-loaded racks closely follow their respective energy consumption curves. This is because rack inlet temperatures remain below 27 °C throughout the simulation. According to the EnergyStar fan control curve for the A100 server, the fan speed per unit load remains constant when the inlet temperature is below this threshold. Consequently, under single-load conditions, the airflow response is directly proportional to the load-induced power demand. Therefore, abrupt load transitions trigger sharp changes in server exhaust flow. These changes perturb cold-aisle static pressure and induce transient airflow variations in adjacent switch racks.

In the mixed load cases (Figure 10b,c), the GEMM- and VGG-loaded racks exhibit dynamics that differ markedly from those in Cases 1 and 2. For GEMM-loaded racks, minor net airflow oscillations are observed during the high-load stages of 129–244 s and 252–301 s. After 301 s, the net airflow increases abruptly relative to the fluctuation period. For VGG-loaded racks in Cases 3 and 4, the net airflow exhibits an increasing-then-decreasing pattern at the onset of the high-load stage and during 244–252 s. In Cases 5 and 6, minor airflow oscillations are observed in VGG-loaded racks during the constant-load stage after 301 s. These phenomena share the same underlying mechanism as in the single load cases. Abrupt large-amplitude load changes in high-load racks cause sharp static pressure fluctuations in the cold aisle. Because server fan exhaust flow depends not only on fan speed but also on back pressure, these pressure variations induce transient airflow changes in adjacent racks operating under different load conditions.



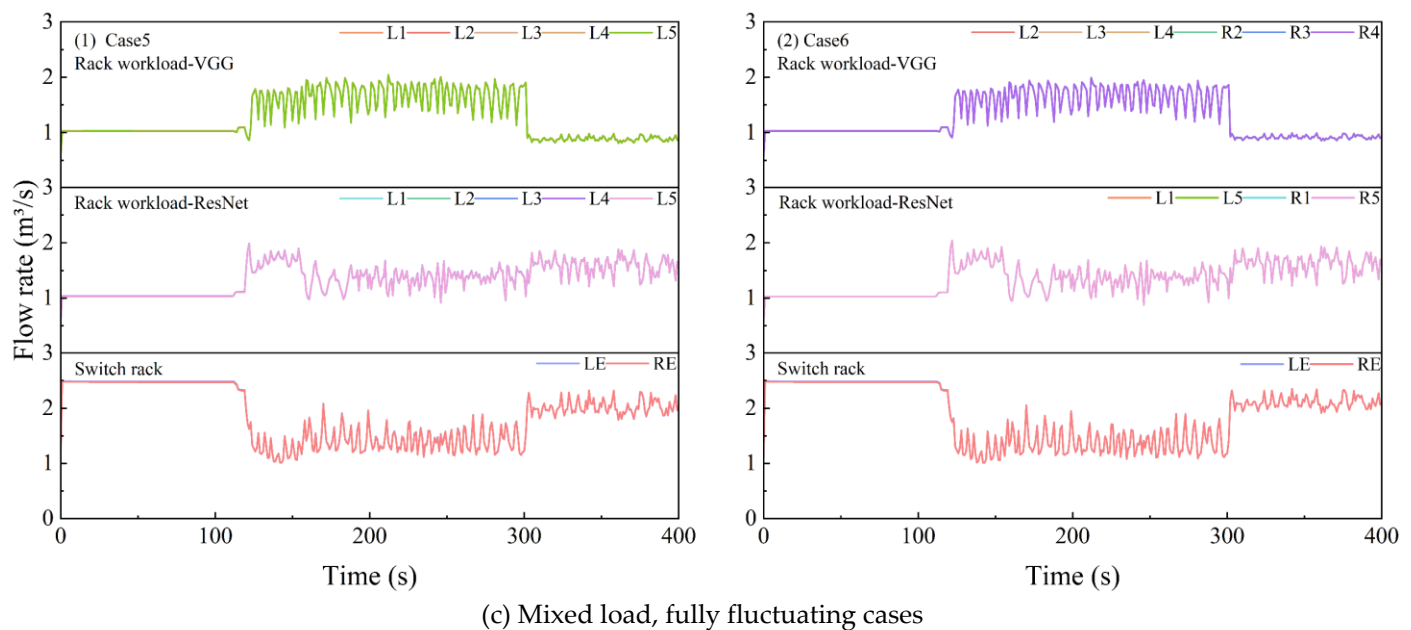


Figure 10. Dynamic net airflow rate curves of server racks for different cases.

4. Conclusions

A CFD model of a raised-floor data center was developed to investigate the dynamic thermal responses under six non-uniform load scenarios derived from representative AI workloads. Based on the simulation results, the following conclusions can be drawn.

(1) The response lag of the air-conditioning units causes a temporary mismatch between cooling supply and server heat dissipation, leading to dynamic fluctuations in rack inlet temperature.

(2) Rack outlet temperature is influenced by both inlet-temperature variation and workload fluctuations. Compared with stable workloads, fluctuating workloads produce larger temperature oscillations and higher temperature change rates. In addition, weaker load fluctuations lead to a weaker dynamic response in rack outlet temperature, with both mean fluctuation amplitude and maximum temperature change rate decreasing.

(3) The mixed workload arrangement can reduce the overall fluctuation intensity of room-level heat dissipation. When servers with highly fluctuating AI workloads are mixed with stable-load or lower-fluctuation servers, the dynamic outlet-temperature response is weakened, which helps improve the matching between cooling capacity and server heat dissipation and reduces the risk of transient hotspots.

(4) Under single-load conditions, net airflow variation generally follows the server power profile. Under mixed-load conditions, abrupt load transitions can induce cold-aisle pressure fluctuations, which affect the airflow responses of neighboring racks.

Future work should focus on integrating dynamic workload prediction with cooling-control strategies to improve thermal management and energy efficiency in AI-oriented data centers.

Author Contributions: C.Y.: Writing—Review and Editing, Supervision, Methodology, Funding acquisition. Z.L.: Writing—original draft, Software, Investigation. J.W.: Writing—Review and Editing, Methodology, Funding acquisition, Formal analysis. Z.Q.: Validation, Software, Conceptualization. Z.Z.: Validation, Methodology, Formal analysis. M.W.J.: Validation, Conceptualization, Formal analysis. All authors reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Research work of this study was supported by the National Natural Science Foundation of China (No. 52208136), the Natural Science Foundation of Hunan Province of China (2023JJ40758, 2024JJ5436), the International Science and Technology Cooperation Programme of China (2023YFE0120400).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: Author Zimu Li was employed by the company China Telecom Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. PAN AMERICAN FINANCE. *Global Data Center Market Report 2025*; Pan American Finance: Miami, FL, USA, 2025.
2. IEA. *Energy and AI*; IEA: Paris, France, 2025.
3. Rong, H.G.; Zhang, H.M.; Xiao, S.; Li, C.B.; Hu, C.H. Optimizing energy consumption for data centers. *Renew. Sustain. Energy Rev.* **2016**, *58*, 674–691. <https://doi.org/10.1016/j.rser.2015.12.283>.
4. Kurt, K.; Öztö, H.F.; Abu-Hamdeh, N.; Gür, M. An experimental study on cooling of model data center building by using eutectic phase change material in their walls. *J. Therm. Anal. Calorim.* **2023**, *148*, 7233–7258. <https://doi.org/10.1007/s10973-023-12203-3>.
5. Khalaj, A.H.; Halgamuge, S.K. A Review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system. *Appl. Energy* **2017**, *205*, 1165–1188. <https://doi.org/10.1016/j.apenergy.2017.08.037>.
6. Pan, Y.Y.; Niu, B.L.; Yang, Z.C. Research on the distribution characteristics of static pressure in data centers from the perspective of airflow transmission path. *J. Build. Eng.* **2024**, *87*, 108557. <https://doi.org/10.1016/j.job.2024.108557>.
7. Phan, L.; Bhusal, S.; Lin, C.X. Improving cooling efficiency by using mixed tiles to control airflow uniformity of perforated tiles in a data center model. In *Proceedings of the ASME 2017 Heat Transfer Summer Conference, Bellevue, WA, USA, 9–12 July 2017*; ASME: Bellevue, WA, USA, 2017.
8. Nada, S.A.; Said, M.A. Effect of CRAC units layout on thermal management of data center. *Appl. Therm. Eng.* **2017**, *118*, 339–344. <https://doi.org/10.1016/j.applthermaleng.2017.03.003>.
9. Nada, S.A.; Said, M.A. Comprehensive study on the effects of plenum depths on air flow and thermal managements in data centers. *Int. J. Therm. Sci.* **2017**, *122*, 302–312. <https://doi.org/10.1016/j.ijthermalsci.2017.09.001>.
10. Lu, H.J.; Zhang, Z.B. Numerical and experimental investigations on the thermal performance of a data center. *Appl. Therm. Eng.* **2020**, *180*, 115759. <https://doi.org/10.1016/j.applthermaleng.2020.115759>.
11. Beitelmal, A.H. Numerical investigation of data center raised-floor plenum. In *Proceedings of the ASME 2015 International Mechanical Engineering Congress and Exposition, Houston, TX, USA, 13–19 November 2015*; ASME: New York, NY, USA, 2015.
12. Zhang, M.; Zhang, Z.B.; Yu, H.; Geng, Y.; Huang, H.; Huang, Y. Effect of raised floor height on different arrangement of under-floor air distribution performance in data center. *Procedia Eng.* **2017**, *205*, 556–564. <https://doi.org/10.1016/j.proeng.2017.10.425>.
13. Abdelmaksoud, W.A.; Khalifa, H. E.; Dang, T.Q.; Elhadidi, B.; Schmidt, R.R.; Iyengar, M. Experimental and computational study of perforated floor tile in data centers. In *Proceedings of the 12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, Las Vegas, NV, USA, 2–5 June 2010*; IEEE: New York, NY, USA, 2010; pp. 1–10.
14. Patankar, S.V. Airflow and Cooling in a Data Center. *J. Heat Transfer.* **2010**, *132*, 073001. <https://doi.org/10.1115/1.4000703>.
15. Phan, L.; Hu, B.C.; LIN, C.X.; Dulikravich, G.S. Optimization and inverse design of floor tile airflow distributions in data centers using response surface method. *J. Fluids Eng.* **2022**, *144*, 011503. <https://doi.org/10.1115/1.4051971>.
16. Chu, W.X.; Wang, C.C. CFD investigation of airflow management in a small container data center. *IEEE Trans. Compon. Packag. Manuf. Technol.* **2019**, *9*, 2177–2188.
17. Athavale, J.; Joshi, Y.; Yoda, M. Experimentally validated computational fluid dynamics model for data center with active tiles. *J. Electron. Packag.* **2018**, *140*, 010902. <https://doi.org/10.1115/1.4039025>.
18. Wan, J.X.; Zhou, J.; Gui, X. Intelligent rack-level cooling management in data centers with active ventilation tiles: A deep reinforcement learning approach. *IEEE Intell. Syst.* **2021**, *36*, 42–52.
19. Nada, S.A.; Abbas, A.M. Solutions of thermal management problems for terminal racks of in-row cooling architectures in data centers. *Build. Environ.* **2021**, *201*, 107991. <https://doi.org/10.1016/j.buildenv.2021.107991>.

20. Tsuda, A.; Mino, Y.; Nishimura, S. Comparison of ICT equipment air-intake temperatures between cold aisle containment and hot aisle containment in datacenters. In *Proceedings of the 2017 IEEE International Telecommunications Energy Conference, Broadbeach Australia*, 22–26 October 2017; IEEE: Broadbeach, QLD, Australia, 2017, pp. 59–65.
21. Yuan, X.L.; Liu, J.X.; Yang, Y.J.; Wang, Y.; Yuan, X.H. Investigation and improvement of air distribution system's airflow performance in data centers. *Procedia Eng.* **2017**, *205*, 2895–2902. <https://doi.org/10.1016/j.proeng.2017.09.920>.
22. Song, Z.H. Studying the fan-assisted cooling using the Taguchi approach in open and closed data centers. *Int. J. Heat Mass Transf.* **2017**, *111*, 593–601. <https://doi.org/10.1016/j.ijheatmasstransfer.2017.04.022>.
23. Wu, L.; Mao, R.Y.; Wu, H.W.; Zhang, J.; Li, C.; Yi, X.Y.; Huang, G.Q.; Zhang, Z.J. Optimization of airflow organization in fan-wall data center via baffles. *Appl. Therm. Eng.* **2024**, *244*, 122745. <https://doi.org/10.1016/j.applthermaleng.2024.122745>.
24. Zhou, X.T.; Yuan, X.L.; Xu, X.J.; Liu, J.X.; Kosonen, R.; Liu, C. Research on the thermal performance of rack-level composite baffle diversion system for data centre. *Energy Effic.* **2020**, *13*, 1245–1262.
25. Jin, C.Q.; Bai, X.L.; Yang, C.; Mao, W.X.; Xu, X. A review of power consumption models of servers in data centers. *Appl. Energy* **2020**, *265*, 114806. <https://doi.org/10.1016/j.apenergy.2020.114806>.
26. Basmadjian, R.; Ali, N.; Niedermeier, F.; Meer, H.; Giuliani, G. A methodology to predict the power consumption of servers in data centres. In *Proceedings of the 2nd International Conference on Energy-Efficient Computing and Networking*, New York, NY, USA, 31 May–1 June 2011; ACM: New York, NY, USA, 2011; pp. 1–10.
27. Arroba, P.; Risco-Martín, J.L.; Zapater, M.; Moya, J.M.; Ayala, J.L.; Olcoz, K. Server power modeling for run-time energy optimization of cloud computing facilities. *Energy Procedia* **2014**, *62*, 401–410. <https://doi.org/10.1016/j.egypro.2014.12.402>.
28. Chen, F.; Grundy, J.; Yang, Y.; Schneider, G.J.; He, Q. Experimental analysis of task-based energy consumption in cloud computing systems. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering*, Prague, Czech Republic, 21–24 April 2013; ACM: New York, NY, USA, 2013; pp. 295–306.
29. Xiao, P.; Hu, Z.G.; Liu, D.B.; Yan, G.F.; Qu, X.L. Virtual machine power measuring technique with bounded error in cloud environments. *J. Netw. Comput. Appl.* **2013**, *36*, 818–828. <https://doi.org/10.1016/j.jnca.2012.12.002>.
30. Li, H.; Casale, G.; Ellahi, T. SLA-driven planning and optimization of enterprise applications. In *Proceedings of the first joint WOSP/SIPEW international conference on Performance engineering*, San Jose, CA, USA, 28–30 January 2010; ACM: New York, NY, USA, 2010; pp. 117–128.
31. Ham, S.-W.; Kim, M.-H.; Choi, B.-N.; Jeong, J.-W. Simplified server model to simulate data center cooling energy consumption. *Energy Build.* **2015**, *86*, 328–339. <https://doi.org/10.1016/j.enbuild.2014.10.058>.
32. Wang, Y.M.; Hao, M.; He, H.; Zhang, W.Z.; Tang, Q.Y.; Sun, X.Y.; Wang, Z. DRLCAP: Runtime GPU frequency capping with deep reinforcement learning. *IEEE Trans. Sustain. Comput.* **2024**, *9*, 712–726. <https://doi.org/10.1109/TSUSC.2024.3362697>.
33. Krzywaniak, A.; Czarnul, P.; Proficz, J. Dynamic GPU power capping with online performance tracing for energy efficient GPU computing using DEPO tool. *Future Gener. Comput. Syst.* **2023**, *145*, 396–414. <https://doi.org/10.1016/j.future.2023.03.041>.
34. Fulpagare, Y.; Bhargav, A.; Joshi, Y. Dynamic thermal characterization of raised floor plenum data centers: Experiments and CFD. *J. Build. Eng.* **2019**, *25*, 100783. <https://doi.org/10.1016/j.job.2019.100783>.
35. Ibrahim, M.; Bhopte, S.; Sammakia, B.; Murray, B.; Iyengar, M.; Schmidt, R. Effect of transient boundary conditions and detailed thermal modeling of data center rooms. *IEEE Trans. Compon. Packag. Manuf. Technol.* **2012**, *2*, 300–310.
36. NVIDIA Corporation. *NVIDIA DGX SuperPOD*; Scalable Infrastructure for AI Leadership Reference Architecture; NVIDIA Corporation: Santa Clara, CA, USA, 2021.
37. NVIDIA Corporation. *NVIDIA DGX A100 Datasheet*; NVIDIA Corporation: Santa Clara, CA, USA, 2020.
38. NVIDIA Corporation. *NVIDIA Quantum QM8700 and QM8790 Switch Systems Hardware User Manual*; NVIDIA Corporation: Santa Clara, CA, USA, 2021.
39. NVIDIA Corporation. *NVIDIA Spectrum SN4000 Series Switches Datasheet*; NVIDIA Corporation: Santa Clara, CA, USA, 2021.
40. Coşanay, H.; Öztö, H.F.; Gür, M.; Bakır, E. Analysis of turbulent wall jet impingement onto a moving heated body. *Int. J. Numer. Methods Heat Fluid Flow* **2022**, *32*, 2938–2963. <https://doi.org/10.1108/HFF-08-2021-0521>.
41. Gür, E.B.; Işık, E.; Uçar, A. Numerical analysis of PCM integrated in solar dryers for thermal efficiency. *Ain Shams Eng. J.* **2026**, *17*, 103847.
42. Wang, Y.T.; Bai, X.L.; Fu, Y.S.; Tang, Y.; Jin, C.Q.; Li, Z.Y. Field experiment and numerical simulation for airflow evaluation in a data center with row-based cooling. *Energy Build.* **2023**, *294*, 113231. <https://doi.org/10.1016/j.enbuild.2023.113231>.
43. Yuan, X.L.; Zhou, X.T.; Liu, J.X.; Wang, Y.; Kosonen, R.; Xu, X.J. Experimental and numerical investigation of an airflow management system in data center with lower-side terminal baffles for servers. *Build. Environ.* **2019**, *155*, 308–319. <https://doi.org/10.1016/j.buildenv.2019.03.039>.

44. Yu, F.J.; Chen, H.B.; Wang, W.Q.; An, J.J. Optimization strategies for underfloor air distribution in a small-scale data center. *Buildings* **2025**, *15*, 428. <https://doi.org/10.3390/buildings15030428>.
45. Li, S.; Davidson, L.; Peng, S.-H. A pressure-loss model for flow-through round-hole perforated plates of moderate porosity and thickness in laminar and turbulent flow regimes. *Int. J. Heat Mass Transf.* **2024**, *226*, 125490. <https://doi.org/10.1016/j.ijheatmasstransfer.2024.125490>.
46. Zhu, Y.; Song, P.; Zhang, Z.B. Numerical and experimental investigation of thermal performance in data center with different deflectors for cold aisle containment. *Build. Environ.* **2021**, *200*, 107961. <https://doi.org/10.1016/j.buildenv.2021.107961>.
47. Roache, P.J. Perspective: A method for uniform reporting of grid refinement studies. *J. Fluids Eng.* **1994**, *116*, 405–413.
48. Manna, P.; Dharavath, M.; Sinha, P.K.; Chakraborty, D. Optimization of a flight-worthy scramjet combustor through CFD. *Aerosp. Sci. Technol.* **2013**, *27*, 138–146. <https://doi.org/10.1016/j.ast.2012.07.005>.
49. Tang, Y. Research on rack thermal environment in data center based on dynamic power consumption; Chongqing University: Chongqing, China, 2022.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.