HotGPU: A Thermal Profile Dataset for Immersion-Cooling AI Datacenters

HENGJIA ZHANG, The Hong Kong Polytechnic University
SHUNTAO ZHU, The Hong Kong Polytechnic University
RUI LU, The Hong Kong Polytechnic University
DAN WANG, The Hong Kong University of Science and Technology

With the increasing power density of modern GPUs, thermal management in AI datacenters has attracted great attention. Unfortunately, there are very few datasets on the GPU thermal profiles. In this paper, we develop HotGPU, a thermal profile dataset for AI datacenters in the new immersion cooling systems, where servers are immersed in dielectric coolant. HotGPU reports the end-to-end thermal process from the heat generation of AI workloads to the heat dissipation of immersion cooling systems. HotGPU is generated by high-fidelity Computational Fluid Dynamics (CFD) simulations validated by real-world experiments. HotGPU covers major AI workloads, diverse GPU types, and coolant types; and has rich features on the GPU temperature distribution, evolution, the dynamics in the coolant heat absorption process, and multi-GPU interactions. We study the use cases of the HotGPU datasets in the thermal-aware scheduling and capacity planning of AI datacenters.

CCS Concepts: • **Social and professional topics** → **Sustainability**; • **Information systems** → **Data centers**; • **Hardware** → **Temperature simulation and estimation**; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: GPU thermal profile dataset, immersion cooling, AI datacenters, computational fluid dynamics

1 Introduction

AI datacenters consume huge amounts of energy [5, 12, 33, 38], and cooling has become a key component of their operational cost [19]. The power density of modern GPUs is increasing rapidly and hotspots frequently occur even when AI datacenters are now supported by advanced cooling systems [7, 15, 45]. Thus, thermal management has attracted increasing attention [34, 36, 44]. Unfortunately, there are very few datasets on the GPU thermal profiles, in particular with the new immersion cooling systems [26, 35] for high-end AI datacenters, where servers are immersed in dielectric coolant (Fig. 1). In this paper, we develop HotGPU, a thermal profile dataset for AI datacenters in immersion cooling systems[1]. As compared to existing datasets, HotGPU is new with: (1) data related to immersion cooling systems, e.g., coolant type, bubble fractions, etc., (2) data reflecting the end-to-end thermal process, e.g., from heat generation of GPUs driven by AI workloads to heat dissipation of immersion cooling systems, and (3) data in high spatial and temporal resolution with a granularity of 0.43mm-15mm and 0.01 seconds.

HotGPU is generated by high-fidelity Computational Fluid Dynamics (CFD) simulations at a computational cost of over 478,000 CPU core-hours, validated by real-world experiment traces. In CFD simulations, the space is constructed into a grid. In each cell of the grid, the simulation follows the physics of the thermal process, e.g., the mass conservation equation, the momentum conservation equation, the energy conservation equation, etc. The thermal process consists of multiple stages: from heat source simulation to heat dissipation simulation. The heat is generated by GPU executing AI workloads, and the heat is dissipated by the coolant, which can be further divided into the *natural convection stage*, where the heat is dissipated primarily by convection, and the *nucleate boiling stage*, where the heat is dissipated primarily by a phase change of the coolant from liquid to vapor to absorb heat (this stage is also called the *bubble stage*). Different stages are triggered by the power density of GPUs. Our CFD simulates the corresponding physics of different stages. In our dataset, we simulate the main heat sources of a GPU, e.g., the die board and printed circuit board (PCB); and for the minor sources that may vary across GPUs, e.g., memory, we develop a *thermal bias correction scheme*.

To validate HotGPU, we develop a set of real-world measured data. Our measurement platform consists of a single GPU immersed in an immersion cooling tank filled with selected coolants. We measured the main AI workloads, diverse GPU types, and coolant types, covering the heat generation and heat dissipation processes with the natural convection stage and nucleate boiling stage.

We evaluate (1) the accuracy of the temperature and (2) the consistency of the temperature turning-points, i.e., the increase-decrease turning points of the temperature. The turning point is triggered by the dynamics of heat generation (driven by the dynamics of AI workloads) or heat dissipation (driven by the vapor concentration of the coolant in the boiling stage). Our results show an RMSE of 1.47°C and 1.64°C for the average and peak temperature. For turning point consistency, HotGPU achieves 90.06%.

HotGPU provides rich features of the GPU thermal profiles to describe average and hotspot temperature evolutions, bubble dynamics, and multi-GPU thermal interactions. These features can be further transformed into task-oriented metrics for datacenter thermal management applications, e.g., *thermal-aware scheduling, cooling capacity planning, etc.* For thermal-aware scheduling, thermal cost matrices can be derived from average- and hotspot-temperature evolutions, and such metrics help schedulers avoid workload placements that increase temperature or create hotspots. For capacity planning, metrics such as safe deployment-density estimation, and thermal-headroom indicators can be derived from bubble-fraction dynamics and multi-GPU thermal interaction factors, since these features reveal boiling onset, vapor accumulation, and spatial thermal coupling among neighboring GPUs.

[1]HotGPU is publicly available at https://github.com/Jacob-ZHANG-2025/HotGPU

Authors' Contact Information: Hengjia Zhang, heng-jia.zhang@connect.polyu.hk, The Hong Kong Polytechnic University, ; Shuntao Zhu, shun-tao.zhu@connect.polyu.hk, The Hong Kong Polytechnic University, ; Rui Lu, rui2020.lu@connect.polyu.hk, The Hong Kong Polytechnic University, ; Dan Wang, wangdan@ust.hk, The Hong Kong University of Science and Technology,

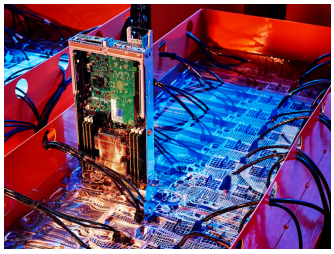


Fig. 1. AI Datacenters with Immersion Cooling

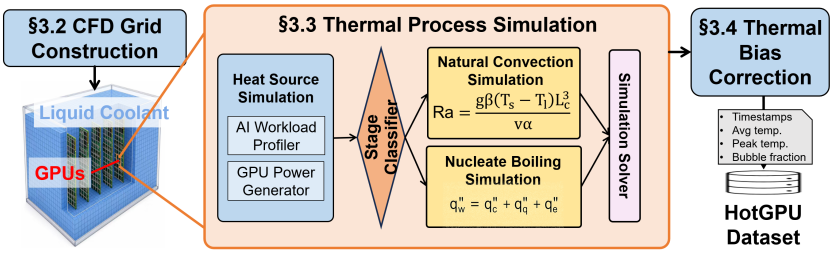


Fig. 2. Overview of the CFD simulation for the HotGPU dataset

Existing datasets: The computer science community has developed datasets on AI workloads and GPU power traces. For example, Philly [24] and Helios [22] report on the workload characteristics across various AI models and hardware platforms, allowing analysis of AI workloads. AcmTrace [23] and GenAI Power [46] report GPU power and thermal measurements under different workloads and operating conditions. The mechanical and thermal engineering community has developed datasets on immersion cooling technologies. There are real-world datasets and CFD simulation datasets. For real-world datasets, MOSFETs reports on the thermal temperatures of immersion-cooling systems [39] and Krishnadasan et. al. [28] reports on the bubble stage behaviors of immersion cooling systems. For CFD simulation datasets, BubbleML [20] reports CFD-based simulation of the bubble stage behaviors.

Limitation of existing datasets: Existing datasets are divided into separated stages of the thermal process. AI workloads and GPU power datasets lack cooling systems in general and immersion cooling in particular. Thus, there will be errors when constructing the thermal profiles of GPUs directly using these data. Immersion cooling datasets do not have GPU specifics. For example, the heat sources of MOSFET are two stable power settings (50W and 100W), the chips of [28] are 33-ohm power resistor chips, and BubbleML [20] only employs a 2D smooth planar surface. These limitations are partially because developing a GPU thermal profile dataset requires knowledge from interdisciplinary fields. HotGPU fills this gap.

We comment that a direct concatenation of the datasets of GPU powers and the datasets of cooling systems is not viable. There are fundamental inconsistencies: (1) temporal granularity mismatch: the datasets are often collected at different sampling granularity due to different emphasis; (2) an absence of thermal throttling effects: the GPU power datasets assume an ideal cooling environment and there is a lack of the thermal throttle events in realistic cooling systems; and (3) mismatch across GPU types: the thermal characteristics of GPUs have differences, and it is not easy to synchronize GPU types to immersion cooling datasets.

2 The Thermal Process of AI Datacenters

Heat Generation: The heat generation process in AI datacenters fundamentally originates from the execution of AI workloads (training and inference tasks) [18], where the computation is translated into GPU power consumption and eventually dissipated as heat. To characterize the diverse computational behaviors of AI workloads, high-level model operations can be decomposed into sequences of GPU kernels [11, 31], e.g., matrix multiplication (GEMM), attention functions, activation functions (e.g., ReLU or GELU), Softmax, etc. For example, training or serving large language models (LLMs) such as Llama involves repeated execution of GEMM and attention kernels [14, 21]. As such, the heat generation of AI models can be captured by their kernels together with GPU types.

Heat Dissipation: The heat dissipation of GPUs in immersion cooling systems is governed by heat transfer mechanisms, including *natural convection* [10] and *nucleate boiling* [17, 41]. When GPUs generate heat during workload execution, thermal energy is transferred from the GPU surface to the surrounding dielectric coolant. When heat generation is low (e.g., 100W), heat is mainly removed through natural convection. When heat generation increases, the coolant near the heated surface reaches its saturation temperature, triggering nucleate boiling. Bubbles begin to form at the hotspots on the GPU surface. This liquid-vapor phase change absorbs a substantial amount of heat. This is the reason why a (two-phase) immersion cooling system has a greater heat dissipation capacity. Finally, bubbles will impact a condense water pipe, which brings the heat to ambient environment and the bubbles vanish to liquid forms. During this process, the fraction of bubbles is dynamic and it is an important feature on the heat dissipation capacity.

3 HotGPU Data Generation

HotGPU links AI workload execution with immersion-cooling behaviors by converting measured GPU power traces [18] into heat sources, simulating boiling-aware heat dissipation, and correcting GPU-specific thermal bias. We first introduce the overall generation pipeline and then describe the core modules. Finally, we summarize dataset features, including temperature evolution, hotspot behavior, bubble dynamics, and multi-GPU thermal interactions.

3.1 Simulation Overview

The overview of HotGPU's dataset generation pipeline is in Fig. 2. The pipeline contains three main modules: *CFD Grid Construction*, *Thermal Process Simulation*, and *Thermal Bias Correction*.

First, *CFD Grid Construction* constructs a grid structure given a 3D geometric environment of an immersion cooling tank with GPUs. It generates the cells with the resolutions to capture bubbles, with a size of 0.5mm in their generation to 50mm in their vanishing. We thus generate a grid structure with multiple resolutions.

Second, *Thermal Process Simulation* generates the thermal profiles in each cell of the grid. This is an iterative procedure. For each cell at each time step, the simulation consists of five modules: (1) the heat

source simulation converts the traces containing the characteristics of AI workloads and GPU types to heat source representation; (2) the stage classifier identifies one of the two heat-transfer physical mechanisms of immersion cooling systems; (3) the natural convection stage simulation through physics equations, e.g., mass, momentum, and energy, (4) the nucleate boiling stage simulation, also through physics equations; and (5) the simulation solver updates the necessary physics fields, which will become the inputs for the next iteration.

Finally, *Thermal Bias Correction* compensates for a systematic deviation caused by simplified GPU modeling.

3.2 CFD Grid Construction

The goal is to construct a grid structure given a 3D geometric environment of an immersion cooling tank. We configure three types of objects in this environment, the tank, a number of GPUs, and the condensed pipe. We configure their shape, e.g., length, width, height. One decision is to find the appropriate resolutions of the cells of the grid. We need to capture the bubbles. Thus, the cells should be smaller than the bubbles. The bubbles are generated on the GPUs and they vanish on the condensed pipes at the periphery of the 3D environment. Their sizes evolve from 0.5mm^2 when generated to 50mm^2 when vanished. In our setting we choose multiple resolutions, 0.5mm as the length of the cells around GPUs (a cell will be 0.25mm^2, allowing two to three cells to capture the bubble activities) and 15mm as the length of the cells at the periphery of the 3D environment. We then leave it to state-of-the-art software (e.g., Ansys) to generate the grid structure [3]. The software will make slight adjustment (the cells at the GPUs are adjusted to 0.43mm) so as to construct the grid with minimum orthogonal quality [6], i.e., to maximize the number of 90° angles for the cells.

3.3 Thermal Process Simulation

The thermal process simulation is to update the transient thermal profiles (e.g., temperature, bubble fractions, etc.) in each cell. The thermal evolution is computed in an iterative manner for the cells, where the temperature, heat flux, and phase information are updated sequentially over time steps. In what follows, we present, in each time step, how the thermal profile of a cell is updated.

Heat Source Simulation. The objective of this module is to construct the (volumetric) heat flux of a GPU in a transient and kernel-aware heat source of a specific GPU type. The heat flux is defined as the rate of heat transfer per unit volume. Specifically, we observe that the most easily available traces are the power traces. Thus, given the power traces $P(t)$ and the GPU type, our simulation generates a time-series of the heat flux of this GPU q''':

$$q''' = P(t)/V_{\text{die}}, \tag{1}$$

where V_{die} is the volume of the GPU die.

We also develop a profiler to approximate a time-series of kernel functions along-side the heat flux. This can approximate the heat flux generated at the kernel function level instead of AI model level, which provides finer granularity. We have two steps (1) to map the execution of an AI model to a time-series of execution of kernel functions and (2) to map the kernel functions into power. The first step can be done by existing AI workload profilers. For example, Pytorch Profiler [40], TensorFlow Profiler [1], etc., can follow an AI model structure to generate the kernel functions, e.g., GEMM, or ReLU, in execution. For the second step, we develop a hand-crafted approximation between the kernel functions and their power trace, and we use the first step for calibration.

Stage Classifier. The stage classifier enables dynamic regime-aware thermal simulation by identifying the current heat-transfer mechanism during runtime. This is a threshold classification according to different types of coolant. For example, for coolant Noah-2100A, if the temperature is greater than 47°C, it is dominated by *nucleate boiling*; otherwise, it is dominated by *natural convection.*

Natural Convection Simulation. This module has the physical equations of natural convection stage. We avoid explaining these equations as they are complex and well-known equations on mass, momentum, and energy [4]. We only present one parameter of natural convection, the Rayleigh number Ra, which describes the strength of natural convection:

$$Ra = g\beta(T_s - T_l)L_c^3/(\nu\alpha), \tag{2}$$

where g, β, T_s, T_l, L_c, ν, and α denote the gravitational acceleration, thermal expansion coefficient, GPU surface temperature, liquid temperature, characteristic length, kinematic viscosity, and thermal diffusivity, respectively [8, 9].

The input of this module is an initial/previous field (the temperature field, velocity field, thermal field, etc), and this module will generate the heat flux by physics equations (through state-of-the-art software (e.g., Ansys), calibrating parameters such as those for Ra). As said, this process will continue iteratively.

Nucleate Boiling Simulation. Similarly, this module has the equations for the nucleate boiling stage. This is also a complex physics process. Once boiling occurs, vapor bubbles form on the heated GPU die surface, and wall heat transfer shifts from single-phase convection to boiling heat transfer. The total wall heat flux is partitioned into convective, quenching, and evaporative components [27, 29]. A major equation is the total amount of heat transferred from the solid heated wall to the surrounding fluid, q''_w:

$$q''_w = q''_c + q''_q + q''_e, \tag{3}$$

where q''_c, q''_q, and q''_e are the convective, quenching, and evaporative heat fluxes, respectively.

Simulation Solver. The simulation solver takes the transient heat source and the updated stage-dependent parameters from the aforementioned modules as inputs. It then updates the cell-centered solution fields, including the temperature, velocity, pressure, and phase-fraction fields, by solving the discretized governing equations over the computational cells [30, 32].

3.4 Thermal Bias Correction

Due to model simplifications and the incomplete availability of material thermophysical properties, the CFD simulations tend to underpredict the actual GPU die temperature. To compensate for this discrepancy, an empirical GPU-specific temperature bias correction is introduced. The corrected temperature is defined as

$$T^{(g)}_{\text{corr}}(t) = T^{(g)}_{\text{CFD}}(t) + B_g, \tag{4}$$

where $T^{(g)}_{\text{corr}}(t)$ and $T^{(g)}_{\text{CFD}}(t)$ denote the corrected and raw CFD-predicted temperatures of GPU type g, respectively, and B_g is a constant bias term. To account for GPU-dependent differences , B_g is modeled as a function of the normalized peak power, normalized peak heat flux, and CUDA compute capability:

$$B_g = \alpha_0 + \alpha_1 \frac{P_{\text{peak},g}}{P_{\text{TDP},g}} + \alpha_2 \frac{P_{\text{peak},g}/A_{\text{die},g}}{\overline{P_{\text{peak}}/A_{\text{die}}}} + \alpha_3 C^*_g, \tag{5}$$

Here, $P_{\text{peak},g}$, $P_{\text{TDP},g}$, $A_{\text{die},g}$, and C^*_g represent the peak workload power, thermal design power, die area, and CUDA compute capability of GPU g, respectively. The coefficients α_0–α_3 are empirical calibration parameters. After correction, the temperature trace is aligned with kernel labels, GPU power, and wall-averaged vapor volume fraction along a unified time axis to construct HotGPU.

3.5 HotGPU Dataset Features

HotGPU provides two-phase immersion-cooled GPU thermal profiles. It has GPU types, coolant types, and power dynamics, which are the input data; it approximates the types of the kernel functions, and it exposes four key features:

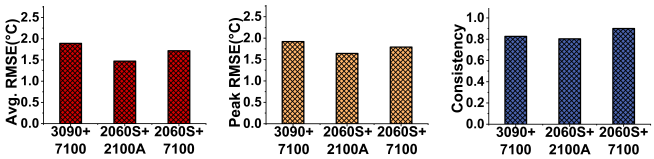
- **GPU Temperature** is provided by HotGPU across diverse GPUs and coolants, enabling fine-grained characterization of thermal behaviors during heating, steady, and cooling periods. They support the study of GPU thermal dynamics.
- **GPU Peak Temperature** is to characterize the localized thermal concentration regions on GPUs. Unlike average GPU temperature, these observations capture spatial thermal heterogeneity and localized overheating, which are critical to thermal throttling and hardware reliability.
- **Bubble Dynamic Fraction** is to characterize the unique cooling behaviors in two-phase immersion, reflecting local vapor-liquid phase distributions around GPUs and enabling investigation of the dynamics of heat dissipation capacity.
- **Multi-GPU Thermal Interaction Temperature** is to characterize the GPU temperature under a neighboring GPU with different spatial deployment distances. This feature allows the examination of spatial thermal dependencies and collective thermal behaviors in multi-GPU environments.

4 HotGPU Data Validation

4.1 Validation Setup

Experimental Platform: To validate the fidelity of HotGPU in capturing GPU thermal evolution from workloads to temperature dynamics, we collect real-world thermal traces from a two-phase immersion cooling testbed. The measurement platform consists of a single GPU immersed in a two-phase cooling tank filled with the selected coolant. Our evaluation uses three GPU-coolant setups: RTX 3090 + Novec-7100 (S1) [13], RTX 2060S with Novec-7100 (S2), and RTX 2060S with Noah-2100A (S3) [25].

Metrics: We validate HotGPU by comparing its simulated thermal profile against measured thermal traces. We adopt two complementary metrics: 1.) *Root Mean Square Error (RMSE)* for quantify the deviation between simulated and measured temperature evolutions;



(a) Avg. RMSE. (b) Peak RMSE. (c) Consistency.

Fig. 3. Validation results across GPU–coolant cases.

2.) *thermal transition consistency* for whether HotGPU correctly captures turning points where the temperature shifts from increasing to decreasing, or vice versa.

Workload Traces: During validation, we select the most representative kernels, including GEMM, Attention, and LayerNorm, to cover diverse computation and thermal patterns in AI workloads.

4.2 Validation Results

Validation on Average Temperature Accuracy. Fig. 3 summarizes the validation results across three GPU–coolant cases. As shown in Fig. 3a, HotGPU achieves average RMSE values of 1.89 °C, 1.72 °C, and 1.47 °C for S1, S2, and S3, respectively. These results indicate that HotGPU can reproduce the overall GPU temperature evolution with low average error.

Validation on Peak Temperature Accuracy. HotGPU maintains low peak-temperature errors across all evaluated cases. In Fig. 3b, the peak RMSE values are 1.92 °C, 1.79 °C, and 1.64 °C for the three cases, respectively. This demonstrates that HotGPU can preserve the high-temperature response of the GPU, which is important for thermal-risk analysis and hotspot-aware scheduling [47].

Validation on Workload Dynamics and Cooling Behavior. The consistency results further evaluate if HotGPU preserves transient thermal dynamics. In Fig. 3c, HotGPU achieves consistency values of 82.67%, 80.36%, and 90.06% for S1, S2, and S3, respectively. These results indicate that HotGPU can capture the thermal patterns induced by workload dynamics and coolant boiling behavior, rather than only matching the average temperature level.

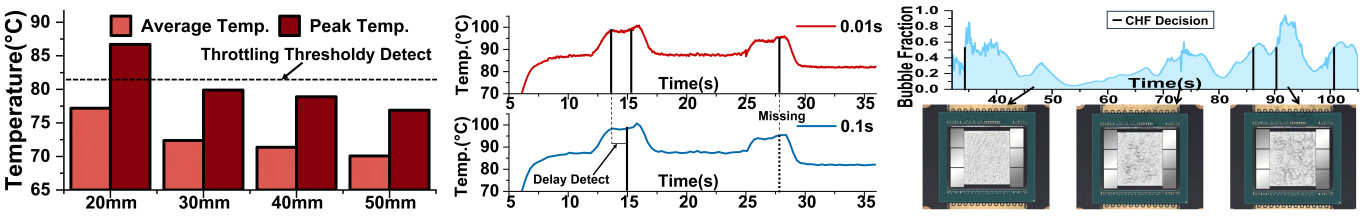
Ablation Study. To evaluate the effect of bias correction, we compare HotGPU with HotGPU-raw by removing the thermal bias correction step. Table 1 shows that the average RMSE of HotGPU-raw increases to 4.98 °C, 5.63 °C, and 3.38 °C, while the peak RMSE increases to 7.47 °C, 8.80 °C, and 6.23 °C for the three cases, respectively. These results show that raw CFD outputs contain larger systematic errors, and bias correction is necessary for reliable HotGPU dataset construction.

Table 1. Thermal bias correction across validation setups.

Setup	GPU	Coolant	Avg. RMSE		Peak RMSE	
			Bias	No Bias	Bias	No Bias
S1	RTX 3090	Novec-7100	1.89	4.98	1.92	7.47
S2	RTX 2060S	Novec-7100	1.72	3.38	1.79	6.23
S3	RTX 2060S	Noah-2100A	1.47	5.63	1.64	8.80

4.3 Validation Scope and Limitation

GPU heat-source modeling scope. HotGPU models the dominant GPU heat-generation paths, including the die board and PCB, and represents secondary sources such as memory through an aggregate



(a) Effect of Spacing on Temperature (b) Thermal throttling observation (c) Bubble-fraction-guided CHF decision making

Fig. 4. HotGPU case studies for capacity planning (a) and online thermal-aware scheduling (b)(c).

thermal abstraction. This scope supports scalable device-level thermal profiling for immersion-cooled AI datacenters. Thermal bias correction aligns simulated traces with measured GPU behavior. Since GPU packages vary across vendors and generations, e.g., chiplet-based designs may exhibit distributed hotspots while monolithic dies concentrate heat differently, we plan future work to incorporate package-aware heat-source layouts.

Device coverage in empirical validation. HotGPU provides profiles for multiple GPU specifications, including Nvidia H100, B200, RTX 3090, and RTX 2060S. Empirical validation uses RTX 3090 and 2060S platforms with representative coolant configurations. For other GPUs, HotGPU applies the same specification-driven CFD methodology using power, die area, and thermal design parameters. We believe the resulting data accuracy remains consistent for other GPU types covered by this pipeline.

Single-GPU validation and multi-GPU profile generation. Measured validation focuses on controlled single-GPU thermal dynamics. Multi-GPU profiles extend this validated pipeline with distance-dependent thermal coupling from neighboring GPUs. Future work will validate the simulated impact of neighboring GPUs using real-world multi-GPU immersion-cooling experiments.

5 Case Studies

5.1 Capacity Planning

We implement our dataset and benchmark on several cases. The firsts case is capacity planning for two-phase immersion-cooled LLM serving systems. Operators must choose deployment density, GPU spacing, and thermal headroom under expected workloads and coolant conditions. HotGPU supports this decision by exposing both *bubble-fraction dynamics* and *multi-GPU thermal interaction factors*, which reveal boiling onset, vapor accumulation, and spatial thermal coupling among neighboring GPUs.

Fig. 4a gives an example of how HotGPU can be used for spacing decisions during capacity planning. As GPU spacing increases from 20 mm to 50 mm, both average and peak temperatures decrease, indicating weaker thermal coupling and improved vapor removal between adjacent GPUs. The 355 K (81.85°C) line marks an operator-defined thermal safety threshold. Under this constraint, spacing settings tighter than 30 mm touch the unsafe region, while larger spacing provides more thermal headroom. HotGPU reveals that dense placement can improve rack capacity, but spacing below the threshold increases throttling risk and needs to be avoided.

This example illustrates the broader role of HotGPU in capacity planning. HotGPU is not a static lookup table for every deployment. Instead, it provides a structured set of CFD-generated scenarios across deployment density, workload mix, coolant condition, spacing, and initial thermal state. These scenarios can be used to derive capacity-planning labels, including maximum safe concurrency, safe deployment density, throttling probability, and remaining thermal headroom under temperature or bubble-fraction constraints.

Such labels can also train a low-cost surrogate model, such as an MLP, that maps system descriptors, workload descriptors, bubble-fraction trajectories, and interaction factors to planning-oriented outputs. This model can compare candidate density and concurrency settings without rerunning CFD simulations for nearby operating points [42, 43]. For example, a denser deployment may offer higher nominal compute capacity, but stronger thermal coupling and earlier vapor accumulation can reduce sustainable capacity. A moderately dense deployment may therefore achieve better utilization and thermal robustness. This case study shows how HotGPU helps operators set concurrency limits, reserve thermal headroom, and decide when scaling out is safer than further densification.

5.2 Online Thermal-aware Scheduling

In online thermal-aware scheduling for two-phase immersion-cooled AI clusters, operators must assign jobs across GPUs while maintaining throughput and avoiding thermal throttling [2, 44]. HotGPU supports this decision by exposing average-temperature evolution, hotspot-temperature evolution, and bubble-fraction dynamics. These signals allow thermal cost matrices to be derived for candidate workload placements, so that schedulers can avoid increasing overall temperature or creating localized hotspots.

Fig. 4b gives an example of how HotGPU can be used for throttling-aware runtime control. Compared with 0.1 s sampling, our 0.01 s temperature trace captures short temperature spikes and rapid rising segments that are delayed or missed by coarse monitoring. These missed events are important for scheduling because delayed detection can leave little time for mitigation before the GPU approaches a throttling state. With fine-grained traces, the scheduler can identify risky temperature trends earlier and trigger actions such as reducing concurrency, migrating a task, or reallocating workloads to avoid throttling.

Fig. 4c further shows how bubble fraction provides an immersion-specific scheduling signal. In two-phase immersion cooling, local vapor accumulation near the GPU die can weaken heat transfer before a clear temperature violation is observed [16, 37]. A rising bubble fraction therefore indicates reduced cooling headroom and possible approach toward critical heat flux, providing an early cue for upcoming thermal changes. This signal complements temperature traces by revealing not only the current thermal state, but also the near-future cooling capacity.

This example illustrates the broader role of HotGPU in online scheduling. HotGPU provides structured CFD-generated scenarios that connect workload placement, GPU spacing, coolant condition, average temperature, hotspot temperature, and bubble-fraction evolution. These scenarios can be converted into scheduling labels or thermal cost matrices, including throttling risk, hotspot severity, cooling-capacity degradation, and remaining headroom. A scheduler can then compare candidate placements without rerunning CFD simulation. This case study shows how HotGPU helps build scheduling policies that jointly consider workload demand, transient temperature, and phase-change cooling behavior.

6 Conclusion and Future Work

This paper presents HotGPU, a thermal profile dataset for immersion-cooling AI datacenters. HotGPU is generated by high-fidelity CFD simulations validated by real-world experiments. HotGPU has rich features on GPU temperature evolution, bubble dynamics, and multi-GPU thermal interactions and can be used for capacity planning, thermal-aware scheduling for AI datacenters.

With the increasing power density of GPUs, there are demands to control GPU thermal behavior. In future work, we plan to further investigate heat-generation and heat-transfer characteristics under immersion cooling for diverse multi-GPU configurations. We will also extend our validation from consumer-grade graphics cards to datacenter-grade compute accelerators to better reflect the thermal behavior of production AI systems. In addition, we plan to collect and align real-world datacenter workload traces with thermal measurements, allowing future versions of HotGPU to support more realistic workload-aware thermal modeling and scheduling studies.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. Dan Wang's work is supported in part by RGC GRF 15201322, 15230624, 15239925, CRF C5020-25GF, ITC ITS/052/23MX, and a start up fund of HKUST.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 265–283.

[2] Nardos Abera and Yize Chen. [n. d.]. Joint Cooling and Computing Optimization for Language Model Serving. In *UrbanAI: Harnessing Artificial Intelligence for Smart Cities*.

[3] Xudong An, Manish Arora, Wei Huang, William C Brantley, and Joseph L Greathouse. 2018. 3D numerical analysis of two-phase immersion cooling for electronic components. In *2018 17th IEEE intersociety conference on thermal and thermomechanical phenomena in electronic systems (ITherm)*. IEEE, 609–614.

[4] John D. Anderson, Jr. 1995. *Computational Fluid Dynamics: The Basics with Applications*. McGraw-Hill, New York.

[5] Thomas Anderson, Adam Belay, Mosharaf Chowdhury, Asaf Cidon, and Irene Zhang. 2023. Treehouse: A case for carbon-aware datacenter software. *ACM SIGENERGY Energy Informatics Review (HotCarbon'22)* 3, 3 (2023), 64–70.

[6] ANSYS, Inc. 2025. Ansys Fluent. General-purpose computational fluid dynamics software for fluid flow, heat and mass transfer, and multiphase modeling.

[7] Mohammad Azarifar, Mehmet Arik, and Je-Young Chang. 2024. Liquid cooling of data centers: A necessity facing challenges. *Applied Thermal Engineering* 247 (2024), 123112.

[8] Adrian Bejan. 2013. *Convection heat transfer*. John wiley & sons.

[9] Theodore L Bergman. 2011. *Fundamentals of heat and mass transfer*. John Wiley & Sons.

[10] Pin Chen, Souad Harmand, and Safouene Ouenzerfi. 2020. Immersion cooling effect of dielectric liquid and self-rewetting fluid on smooth and porous surface. *Applied Thermal Engineering* 180 (2020), 115862.

[11] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. 2018. {TVM}: An automated {End-to-End} optimizing compiler for deep learning. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*. 578–594.

[12] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd workshop on sustainable computer systems (HotCarbon'23)*. 1–7.

[13] 3M Company. 2023. 3M Novec 7100 Engineered Fluid. https://multimedia.3m.com/mws/mediawebserver?mwsId=SSSSSuUn_zu8lZNUl8mBm8mePv70k17zHvu9lxtD7SSSSSS--.

[14] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems* 35 (2022), 16344–16359.

[15] Virmani Darpan and Chatterjee Baibhab. 2025. Thermal Management Challenges in 2.5 D and 3D Chiplet Integration: A Review on Architecture–Cooling Co-Design. *Eng* 6, 12 (2025), 373.

[16] Faizan Ejaz and Beomjin Kwon. 2024. Two-phase active immersion cooling for vertically mounted electronics with interchip component-assisted bubble departure. *International Communications in Heat and Mass Transfer* 159 (2024), 107981.

[17] Mohamed S El-Genk and Jack L Parker. 2008. Nucleate boiling of FC-72 and HFE-7100 on porous graphite at different orientations and liquid subcooling. *Energy conversion and management* 49, 4 (2008), 733–750.

[18] Ahmed Abd Elaziz Elsayed, Abdullah Azhar Al-Obaidi, and Hany EZ Farag. 2025. Characterization of high-resolution AI data center training workloads on single and multiple GPU nodes. (2025).

[19] Wedan Emmanuel Gnibga, Andrew A Chien, Anne Blavette, and Anne Cécile Orgerie. 2024. FlexCoolDC: datacenter cooling flexibility for harmonizing water, energy, carbon, and cost trade-offs. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*. 108–122.

[20] Sheikh Md Shakeel Hassan, Arthur Feeney, Akash Dhruv, Jihoon Kim, Youngjoon Suh, Jaiyoung Ryu, Yoonjin Won, and Aparna Chandramowlishwaran. 2023. Bubbleml: A multiphase multiphysics dataset and benchmarks for machine learning. *Advances in Neural Information Processing Systems* 36 (2023), 418–449.

[21] Ke Hong, Guohao Dai, Jiaming Xu, Qiuli Mao, Xiuhong Li, Jun Liu, Kangdi Chen, Yuhan Dong, and Yu Wang. 2023. Flashdecoding++: Faster large language model inference on gpus. *arXiv preprint arXiv:2311.01282* (2023).

[22] Qinghao Hu et al. 2021. Characterization and Prediction of Deep Learning Workloads in Large-Scale GPU Datacenters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '21)*. doi:10.1145/3458817.3476223

[23] Qinghao Hu, Zhisheng Ye, Zerui Wang, Guoteng Wang, Meng Zhang, Qiaoling Chen, Peng Sun, Dahua Lin, Xiaolin Wang, Yingwei Luo, Yonggang Wen, and Tianwei Zhang. 2024. Characterization of Large Language Model Development in the Datacenter. arXiv:2403.07648 [cs.DC] https://arxiv.org/abs/2403.07648

[24] Myeongjae Jeon, Shivaram Venkataraman, Amar Phanishayee, Junjie Qian, Wencong Xiao, and Fan Yang. 2019. Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*. USENIX Association, Renton, WA, 947–960. https://www.usenix.org/conference/atc19/presentation/jeon

[25] Hanying Jiang, Xiucong Zhao, and Meng Zhang. 2024. Boiling Heat Transfer Characteristics of Noah-2100A and HFE-649 in Pin-Fin Microchannel Heat Sink. *Energies* 17, 24 (2024), 6216.

[26] Baris Burak Kanbur, Chenlong Wu, Simiao Fan, Wei Tong, and Fei Duan. 2020. Two-phase liquid-immersion data center cooling system: Experimental performance and thermoeconomic analysis. *International Journal of Refrigeration* 118 (2020), 290–301.

[27] Eckhard Krepper and Roland Rzehak. 2011. CFD for subcooled flow boiling: Simulation of DEBORA experiments. *Nuclear Engineering and Design* 241, 9 (2011), 3851–3866.

[28] VB Krishnadasan, Pratheek Suresh, and C Balaji. 2025. Experimental investigations on the performance and bubble dynamics of two-phase immersion cooling system with multiple chips. *International Journal of Heat and Mass Transfer* 245 (2025), 126977.

[29] N Kurul and Michael Z Podowski. 1990. Multidimensional effects in forced convection subcooled boiling. In *International Heat Transfer Conference Digital Library*. Begel House Inc.

[30] Brian Edward Launder and Dudley Brian Spalding. 1983. The numerical computation of turbulent flows. In *Numerical prediction of flow, heat transfer, turbulence and combustion*. Elsevier, 96–116.

[31] Seonho Lee, Amar Phanishayee, and Divya Mahajan. 2025. Forecasting GPU Performance for Deep Learning Training and Inference. In *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'25)*. Rotterdam, Netherlands.

[32] Wen Ho Lee. 1980. A pressure iteration scheme for two-phase flow modeling. *Multiphase transport fundamentals, reactor safety, applications* 1 (1980), 407–431.

[33] Amy Li, Sihang Liu, and Yi Ding. 2024. Uncertainty-aware decarbonization for datacenters. *ACM SIGENERGY Energy Informatics Review (HotCarbon'24)* 4, 5 (2024), 141–147.

[34] Jianpeng Lin, Weiwei Lin, Huikang Huang, Wenjun Lin, and Keqin Li. 2023. Thermal modeling and thermal-aware energy saving methods for cloud data centers: A review. *IEEE Transactions on Sustainable Computing* 9, 3 (2023), 571–590.

[35] Cheng Liu and Hang Yu. 2021. Evaluation and optimization of a two-phase liquid-immersion cooling system for data centers. *Energies* 14, 5 (2021), 1395.

[36] Rui Lu and Dan Wang. 2025. A Thermal-aware Workload Scheduler for High-performance LLM Inference in Cooling-regulated Datacenters. *ACM SIGENERGY Energy Informatics Review (HotCarbon'25)* 5, 2 (2025), 98–104.

[37] Yu Ma, Yuchen Bao, and Ji Li. 2025. Heat transfer dependence of power usage effectiveness of an augmented two-phase immersion cooling system for high-power servers. *Energy* 323 (2025), 135853.

[38] Sophia Nguyen, Beihao Zhou, Yi Ding, and Sihang Liu. 2024. Towards sustainable large language model serving. *ACM SIGENERGY Energy Informatics Review (HotCarbon'24)* 4, 5 (2024), 134–140.

[39] Hari Pandey, Xinyuan Du, Ethan Weems, Stephen Pierson, Ahmad Al-Hmoud, Yue Zhao, and Han Hu. 2024. Two-Phase Immersion Cooler for Medium-Voltage Silicon Carbide MOSFETs. In *2024 23rd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. IEEE, 1–6.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[41] Warren M Rohsenow. 1952. A method of correlating heat-transfer data for surface boiling of liquids. *Transactions of the American Society of Mechanical Engineers* 74, 6 (1952), 969–975.

[42] Soumyendu Sarkar, Antonio Guillen-Perez, Zachariah Carmichael, Vineet Gundecha, Avisek Naug, Ricardo Luna Gutierrez, Ashwin Ramesh Babu, and Cullen Bash. 2024. Cfd surrogates for data center sustainability using 3d u-net convolutional neural network. In *2024 23rd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. IEEE, 1–9.

[43] Soumyendu Sarkar, Antonio Guillen-Perez, Zachariah J Carmichael, Avisek Naug, Refik Mert Cam, Vineet Gundecha, Ashwin Ramesh Babu, Sahand Ghorbanpour, and Ricardo Luna Gutierrez. 2026. Fast 3D Surrogate Modeling for Data Center Thermal Management. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 39201–39210.

[44] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Esha Choukse, Haoran Qiu, Rodrigo Fonseca, Josep Torrellas, and Ricardo Bianchini. 2025. Tapas: Thermal-and power-aware scheduling for LLM inference in cloud platforms. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 1266–1281.

[45] Xiaoqing Sun, Zongwei Han, and Xiuming Li. 2022. Simulation study on cooling effect of two-phase liquid-immersion cabinet in data center. *Applied Thermal Engineering* 207 (2022), 118142.

[46] Roberto Vercellino, Jared Willard, Gustavo Campos, Weslley da Silva Pereira, Olivia Hull, Matt Selensky, and Juliane Mueller. 2026. Dataset of Generative AI Workload Power Profiles. doi:10.7799/3025227 Dataset.

[47] Jiaxi Wu, Pavel Popov, Wenquan Yang, Andrei Gudkov, Elizaveta Ponomareva, Xinming Han, Yunzhe Qiu, Jie Song, and Stepan Romanov. 2024. Hotspot-Aware Scheduling of Virtual Machines With Overcommitment for Ultimate Utilization in Cloud Datacenters. *IEEE Transactions on Automation Science and Engineering* 22 (2024), 6809–6821.