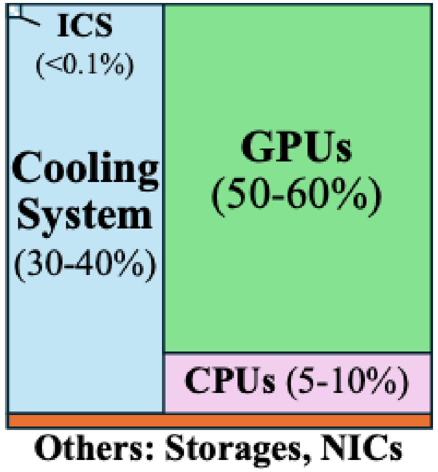
A Thermal-aware Workload Scheduler for High-performance LLM Inference in Cooling-regulated Datacenters

Rui Lu, Dan Wang
Department of Computing
The Hong Kong Polytechnic University

AI Datacenters have Immense Energy Consumptions

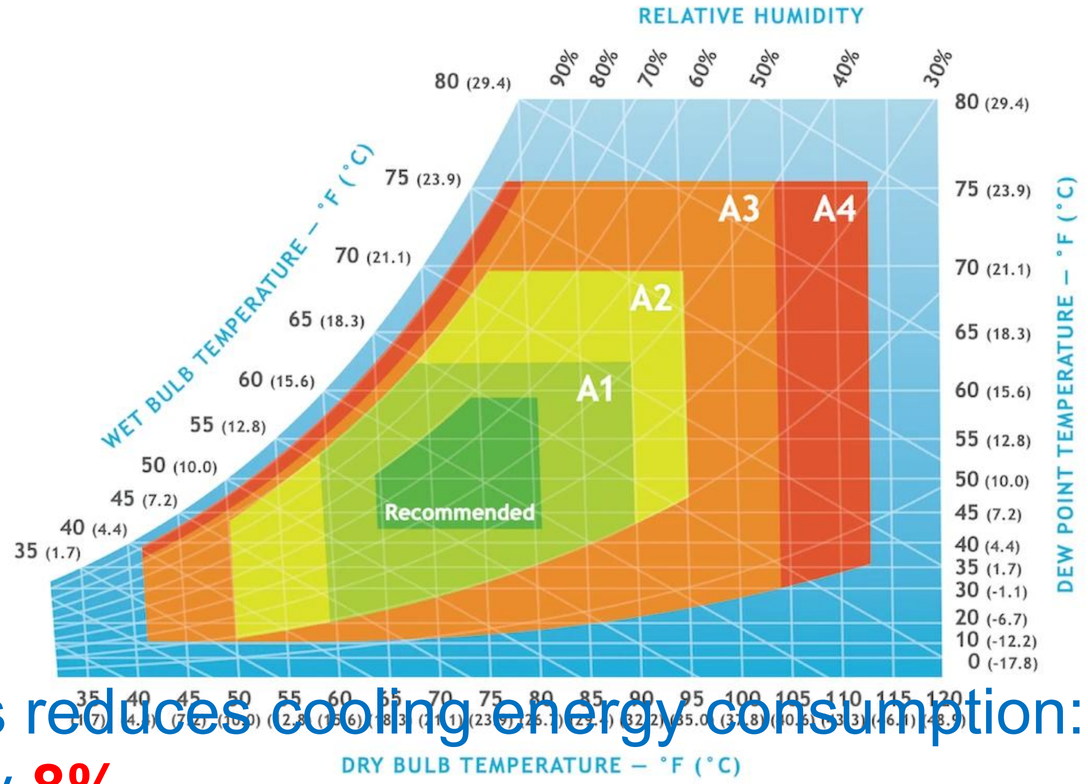
- Modern AI datacenters are experiencing rapid growth of AI applications, e.g., Large Language Models (LLMs).

- The cooling infrastructure still consumes a significant amount of energy, 30%+.



Cooling Regulations in Datacenters

- Guideline to set the room ambient temperature.
  - US: ASHRAE 2004/2008



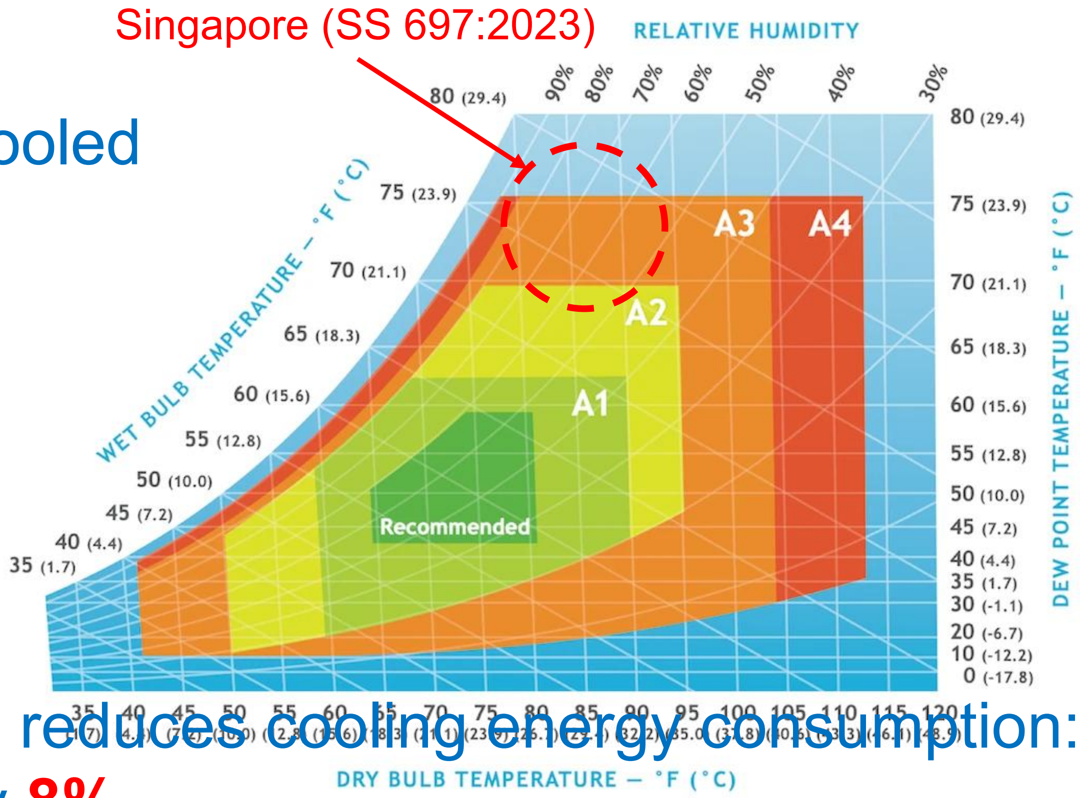
Increasing the room ambient temperatures reduces cooling energy consumption: 1°C increase can reduce cooling energy by 8%.

Cooling Regulations in Datacenters

- Guideline to set the room ambient temperature.
  - US: ASHRAE 2004/2008

- Cooling regulations: Cool but not over-cooled

→ conserve energy
  - Singapore SS 697:2023: 28–32°C



Increasing the room ambient temperatures reduces cooling energy consumption: **1°C** increase can reduce cooling energy by **8%**.

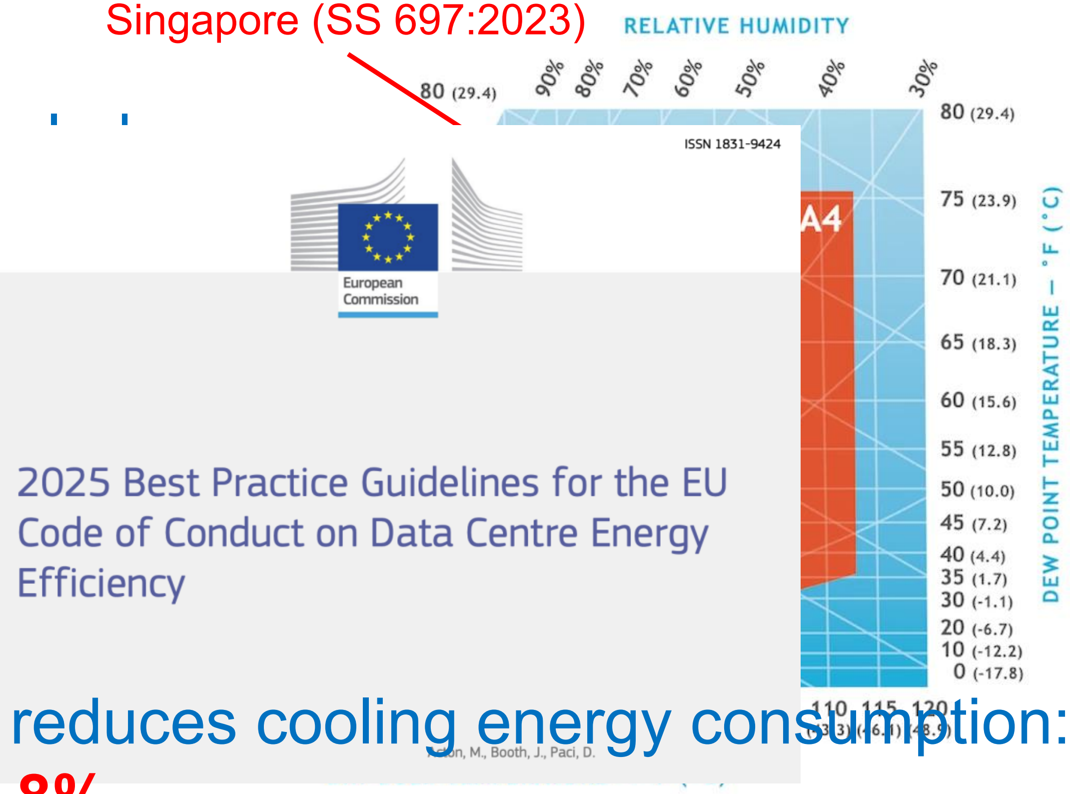
Cooling Regulations in Datacenters

- Guideline to set the room ambient temperature.
  - US: ASHRAE 2004/2008

- Cooling regulations: Cool but not over-cc

→ conserve energy
  - Singapore SS 697:2023: 28–32˚C
  - EU-ISSN 1831-9424: 35˚C

Singapore (SS 697:2023)



Increasing the room ambient temperatures reduces cooling energy consumption: **1˚C** increase can reduce cooling energy by **8%**.

Studies from E & M scholars

Cell Reports
Physical Science

Article

The global energy impact
of raising the space temperature
for high-temperature data centers

- To study societal benefit if datacenters can tolerate higher temperature

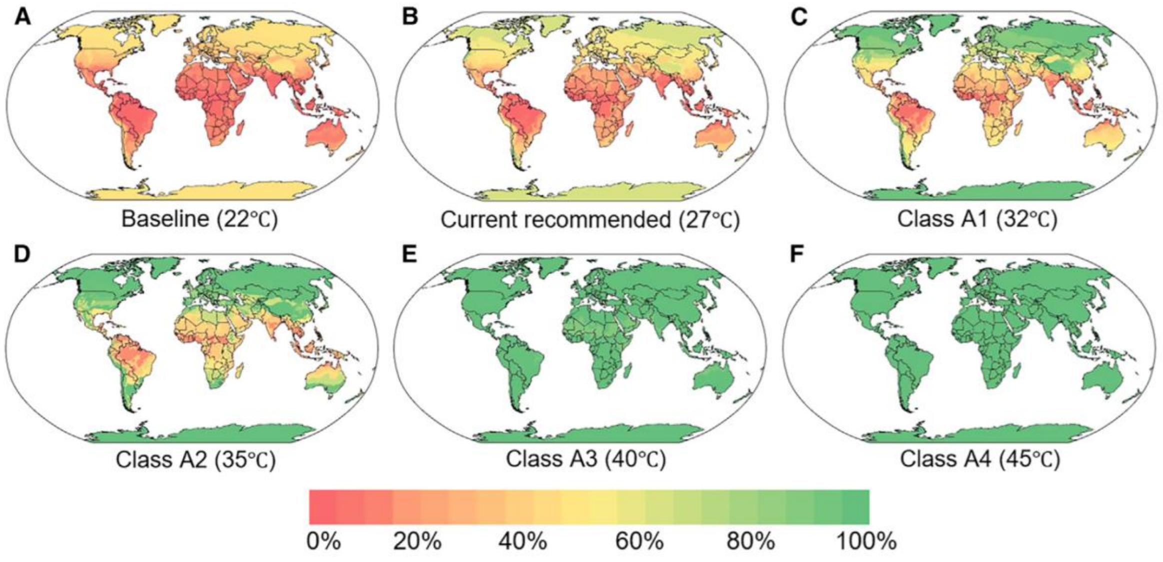


Figure 4. Global maps of annual free-cooling ratio at different space temperatures

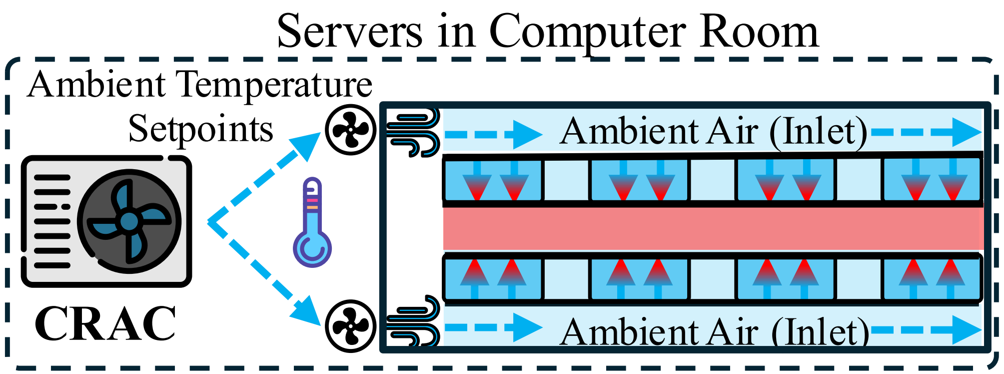
(A) At a baseline space temperature of 22°C.
(B) At 27°C (upper limit of current recommendation).
(C) At 32°C (upper limit of class A1).
(D) At 35°C (upper limit of class A2).
(E) At 40°C (upper limit of class A3).
(F) At 45°C (upper limit of class A4).

'Global Free Cooling Temperature': **41°C**

The cooling system in a datacenter

- **CRAC: Computer Room Air Conditioner**
- **ICS: Inner Cooling System**

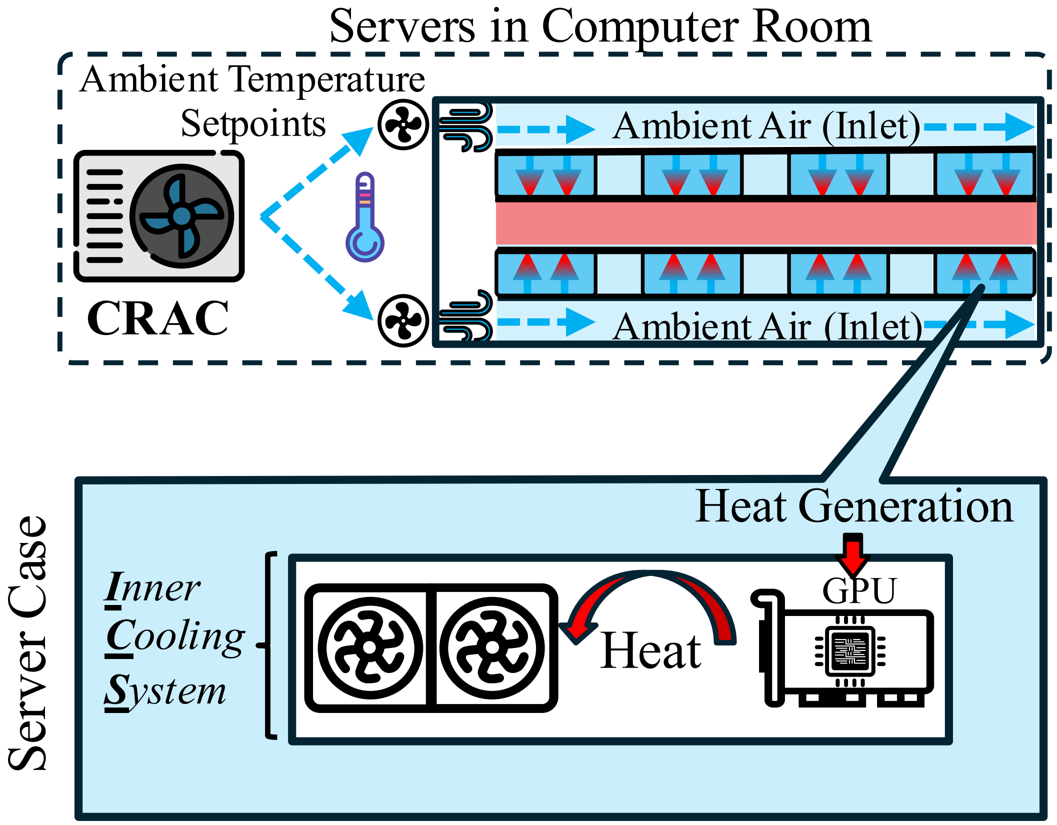
- Cooling Process
  1. External CRAC units chill the air, push the air to the server room, and establishes the room's ambient temperature.
  2. GPU computing generates heat
  3. Inner coolers (air or liquid) remove the heat to the room's air.
  4. Room air exits to the CRAC, re-chilled and re-circulated.



The cooling system in a datacenter

- **CRAC: Computer Room Air Conditioner**
- **ICS: Inner Cooling System**

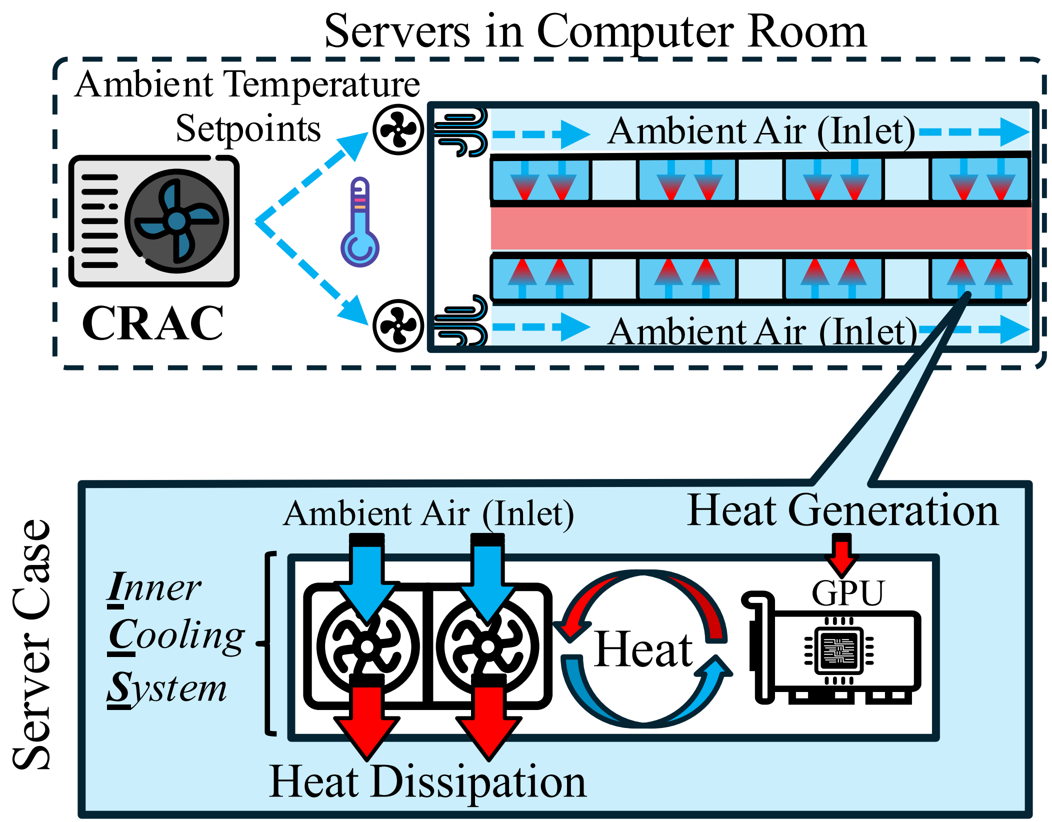
- Cooling Process
  1. External CRAC units chill the air, push the air to the server room, and establishes the room's ambient temperature.
  2. GPU computing generates heat
  3. Inner coolers (air or liquid) remove the heat to the room's air.
  4. Room air exits to the CRAC, re-chilled and re-circulated.



The cooling system in a datacenter

- **CRAC: Computer Room Air Conditioner**
- **ICS: Inner Cooling System**

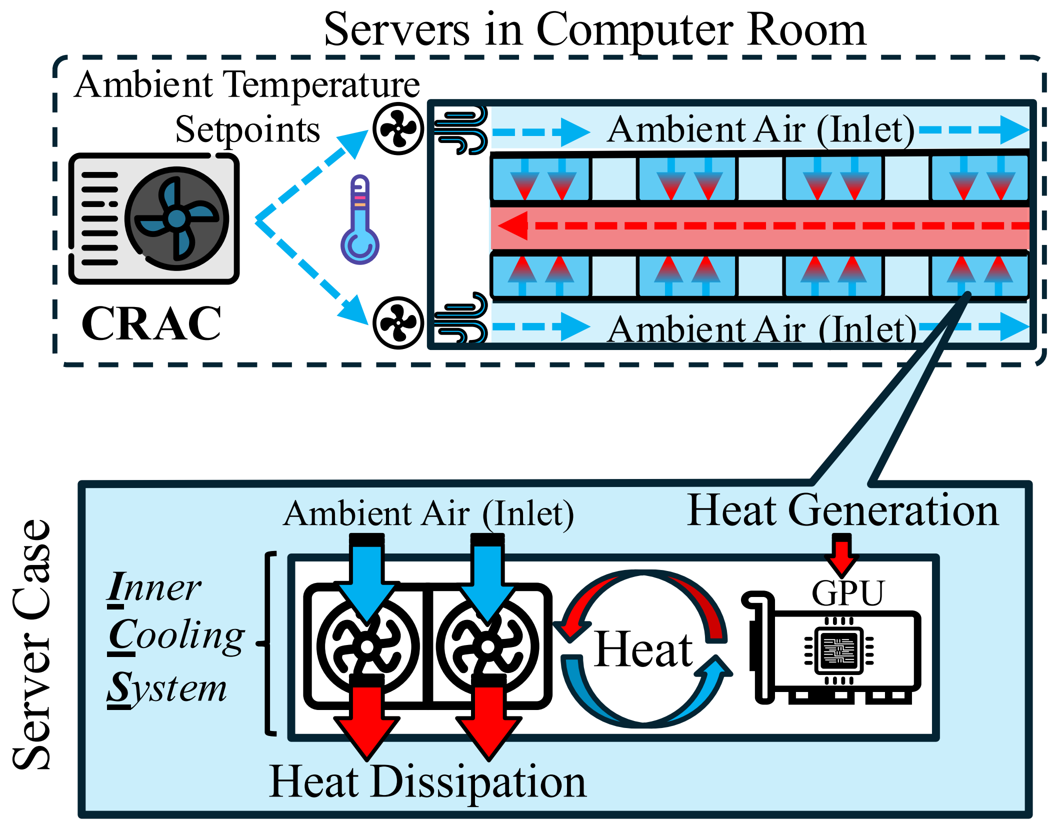
- Cooling Process
  1. External CRAC units chill the air, push the air to the server room, and establishes the room's ambient temperature.
  2. GPU computing generates heat
  3. Inner coolers (air or liquid) remove the heat to the room's air.
  4. Room air exits to the CRAC, re-chilled and re-circulated.



The cooling system in a datacenter

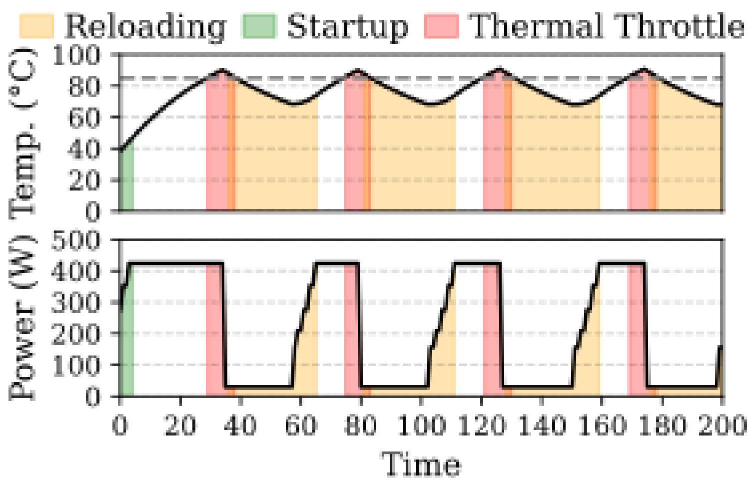
- **CRAC: Computer Room Air Conditioner**
- **ICS: Inner Cooling System**

- Cooling Process
  1. External CRAC units chill the air, push the air to the server room, and establishes the room's ambient temperature.
  2. GPU computing generates heat
  3. Inner coolers (air or liquid) remove the heat to the room's air.
  4. Room air exits to the CRAC, re-chilled and re-circulated.



Motivation: Thermal Throttles in GPU

- A Motivation Experiment
  - Ambient Temperature: 41°C
  - GPU: RTX 4090-Air Cooling/Power: 600 W
  - CPU/RAM: Intel i9-13900K/128GB RAM
  - Outlet Fan Speed: 12.5% (To roughly simulate in 8-GPU cases)
  - LLM Inference: LLAMA3-8B with 128-length prompts
- Observations
  - GPU Thermal Throttle Triggered (>83°C)
  - Data (LLM parameters/weights/intermediate results) in Memory Chips Reload



- Thermal throttle
  - A self-protection mechanism to decrease the voltage and frequency.
  - If serious, throttle triggers GPU reset and reload.

Cooling regulations may affect system performance

Can we do better if there are cooling regulations

LLM inference in AI datacenters

- An AI datacenter serves LLM inference requests on a cluster of GPUs. The objective is to maximize throughput (tokens per second).
- The Ray Serve scheduler assigns LLM inference jobs to a cluster of GPUs and determines the batch size on each GPU to maximize the throughput (tokens per second).
- Existing schedulers implicitly assume sufficient heat dissipation capacity regardless what the workload assignment is.
- In cooling-regulated datacenters, the overall heat dissipation capacity is sufficient, but no longer unlimited. There are variances; and if agnostic to the thermal environment, the performance may be affected

Problem Formulation

- Given: LLM inference tasks and available GPUs $\{g_1, g_2, ...\}$, the Thermal Throttle temperature limit TT_i of each GPU g_i, and the cooling regulation ambient temperature ΔT_{ENV}

- Determine: the batch size b_i assigned on GPU g_i, and GPU g_i's frequency and voltage

- Maximize: throughput TP (tokens per second).

- Challenge 1: A model on the GPU temperature given workloads.
- Challenge 2: A thermal aware workload scheduler

System models

- GPU Temperature Model $TG_i(t)$ at time t

$$TG_i(t) = TG_i(t-\Delta t) + \frac{[\dot{Q}G_i(t) - \dot{Q}R_i(t)] \cdot \Delta t}{m_i \cdot c_i}$$

Previous temperature — Mass — GPU Heat capacity

Heat generation – Heat dissipation

- Heat Generation Model

$$\dot{Q}G_i(t) \sim P_i(t) = \alpha_i v_t + \beta_i C_i v_t^2 f_t + Pc_i$$

GPU Core Voltage — GPU Core Frequency

- Heat Dissipation Model
  - Air-cooling

$$\dot{Q}R_A = \rho v_a A_{hs} \cdot c_p \cdot \Delta T,$$

Contact Surface Area — Density of Air — Air Flow Rate — Air Heat capacity — Temp. Difference

  - Water cooling → details in the paper

System models

- The latency model of the LLM inference jobs: latency *tw* of each layer w in batch size b_i under frequency *fi*

$$t_w(\mathbf{g}_i) = \frac{FLOP_w}{FLOPS(f_i) \cdot eff_{i,w}}$$

Computation Complexity

GPU Computation Capability

Computation Efficiency

$$FLOPS(f_i) \propto N_{core}^{i} \cdot f_i \cdot 2$$

- GPU Computation Capability is proportion to GPU frequency and Number of Cores

A Thermal-aware Workload Scheduler (TAWS)

- TAWS makes decisions on workload assignment to a GPU and frequencies, given its thermal conditions (time-to-throttle) through RL to adapt to the dynamic workloads of the inference jobs

Reinforcement learning

- States

Estimation of time-to-throttle

$$s_k = \left[\Delta T_{\mathrm{ENV}},\ \{TG_{i,k}, f_{i,k}, v_{i,k}, \tau_{i,k}, b_{i,k}\}_{i=1}^{N},\ E_k^{\mathrm{extra}},\ E_k^{\mathrm{save}}\right],$$

- Actions
  1. Choose GPUs to execute the next batch
  2. A frequency–voltage pair (f_i, v_i) for each selected GPU
  3. Batch size assigned to every chosen GPU
- Reward

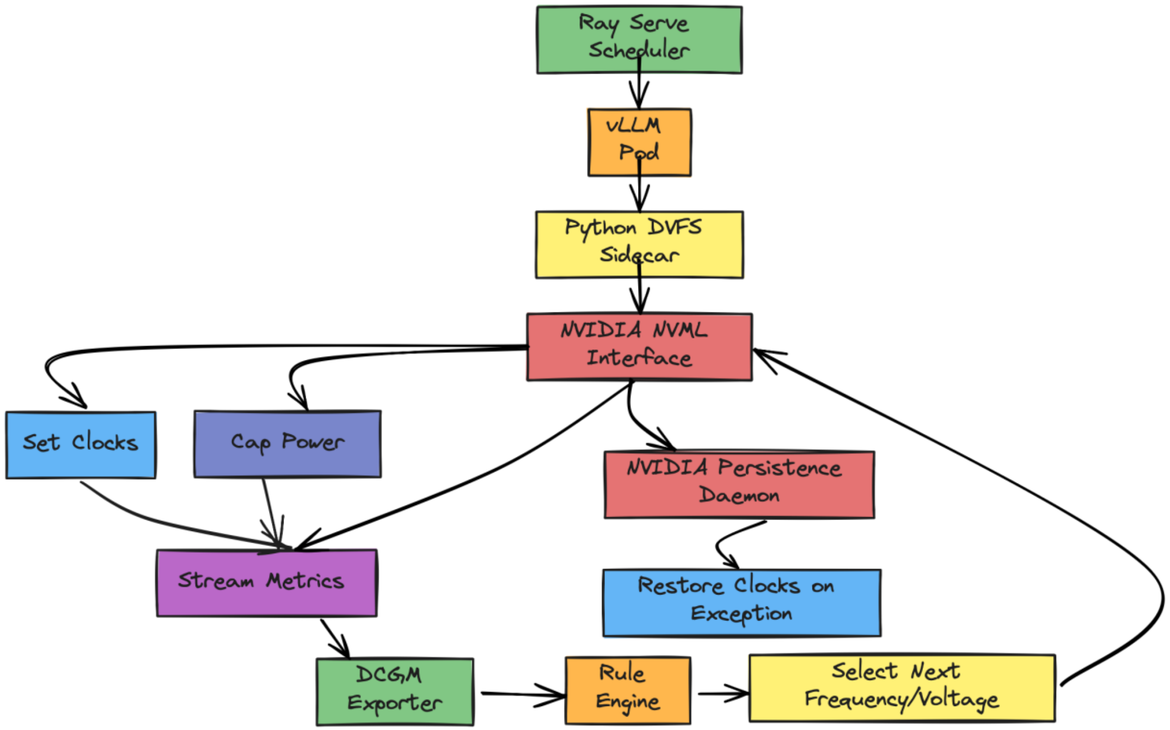
$$r_k = a\,\mathrm{TP}_k - b\ \Delta E_{inf} - d\ \max\left(0,\ TG_{i,k} - T^{\max}\right)$$

Tokens Per Second

Constraint:
Overheat Too Long

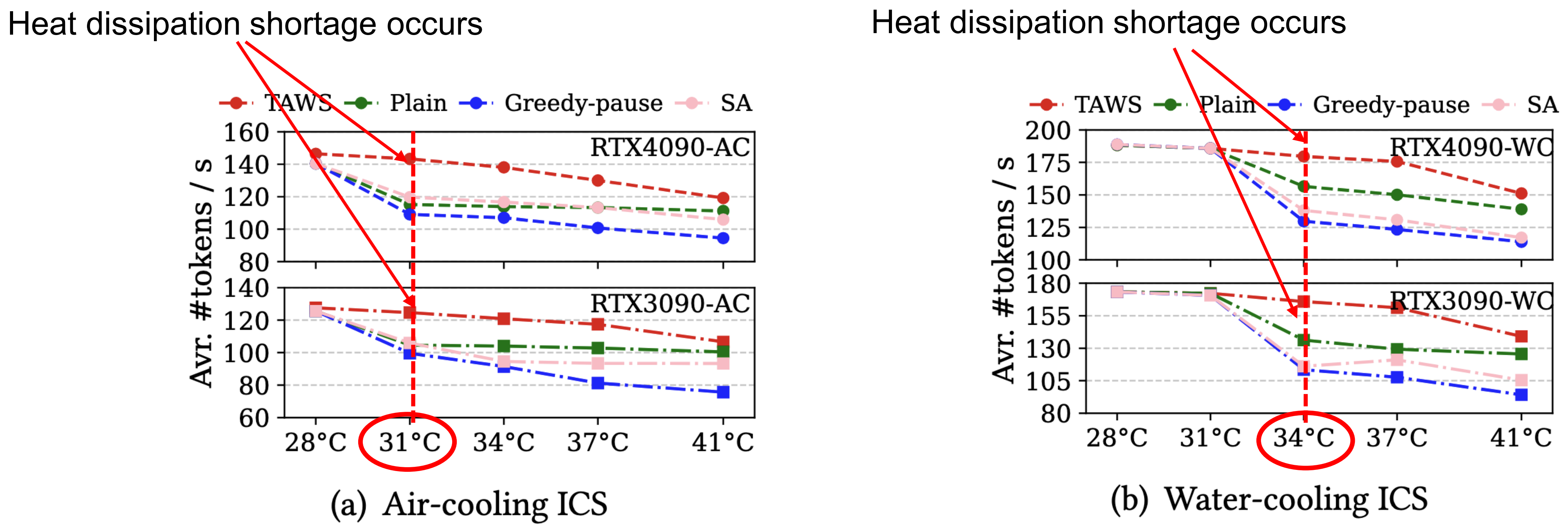
Evaluation: Setup

- Servers
  - RTX 3090 & 4090
  - Air-cooling & Water-cooling
- LLM Inference
  - Models: Llama3-8B
  - Prompts: IBM-TGIS
- Ambient Configs
  - 27°C, 31°C, 34°C, 37°C, 41°C
- Baseline
  - Plain: Agnostic to handling thermal throttle
  - Greedy-Pause: If thermal throttle, pause until a preset temperature
  - SA: Allocate the tasks using Simulated Annealing



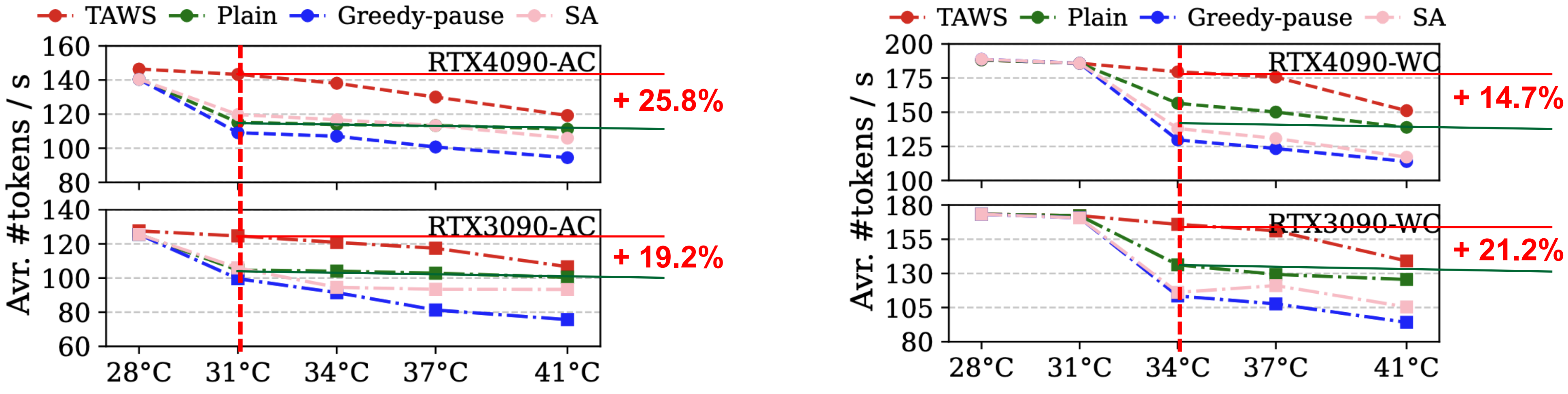
Implementation Details

Evaluation



- Observation 1: Heat dissipation shortage occurs Air cooling: 31°C; Water cooling: 34°C. Higher heat dissipation capacity of water

Evaluation

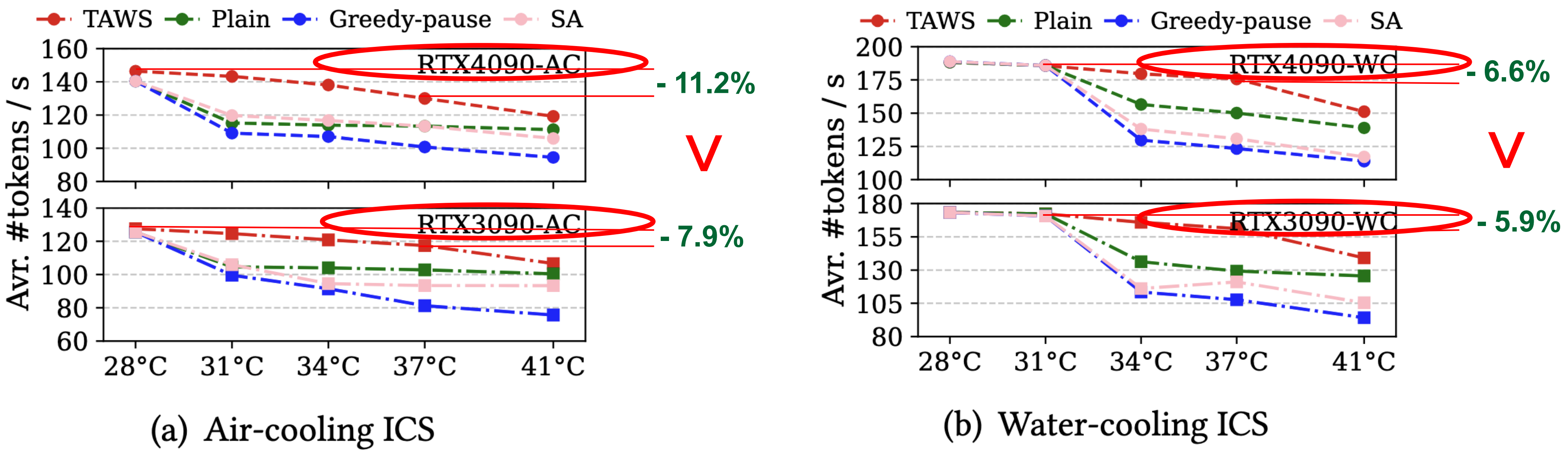


(a) Air-cooling ICS

(b) Water-cooling ICS

- Observation 2: TAWS improves the performance by 25.8% & 19.2% for air-cooling ICS; and 14.7% & 21.2% for water-cooling ICS.

Evaluation



- Observation 3: the performance loss of 4090 is greater than that of 3090

  Advanced Nano Manufactory (4090 4nm vs. 3090 8nm) get more influenced (likely with greater heat concentration, thus requires greater heat removal capacity)

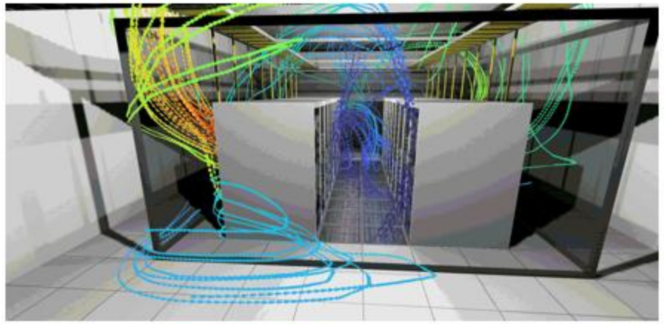
Conclusion

- There are argues on cooling-regulations for AI datacenters.

- We observed that under cooling regulations, the heat dissipation capacity is no longer unlimited. The thermal environment should be taken into scheduling consideration.
- Future work: more systematic measurement to have comprehensive understanding

- We model the heat generation and dissipation process and develop an RL-based thermal-aware scheduler for high-performance LLM inference in cooling-regulated datacenters.
- Future work: Computational Fluit Dynamics (CFD) simulators

to better capture the thermal environment



- We carried out experiments and the results show that TAWS improves the throughput by 32.62%.

Thank you!
Q&A

Related work

- Power capping
  - Proactive reducing frequency vs. reactive in reducing frequency and workload assignment
  - Saving of the GPU energy vs. High-performance computing under cooling regulations