TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms

ASPLOS 2025

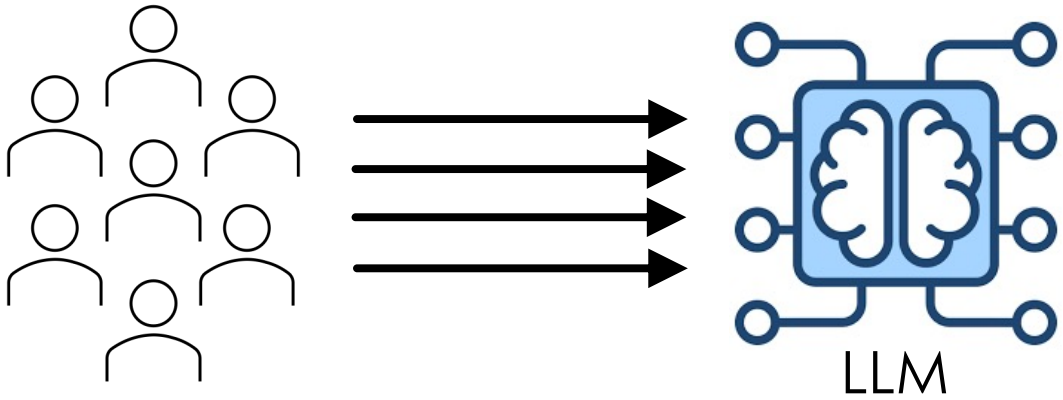
Jovan Stojkovic*, Chaojie Zhang, Íñigo Goiri, Esha Choukse, Haoran Qiu, Rodrigo Fonseca, Josep Torrellas*, Ricardo Bianchini

*University of Illinois at Urbana-Champaign, Azure Research – Systems

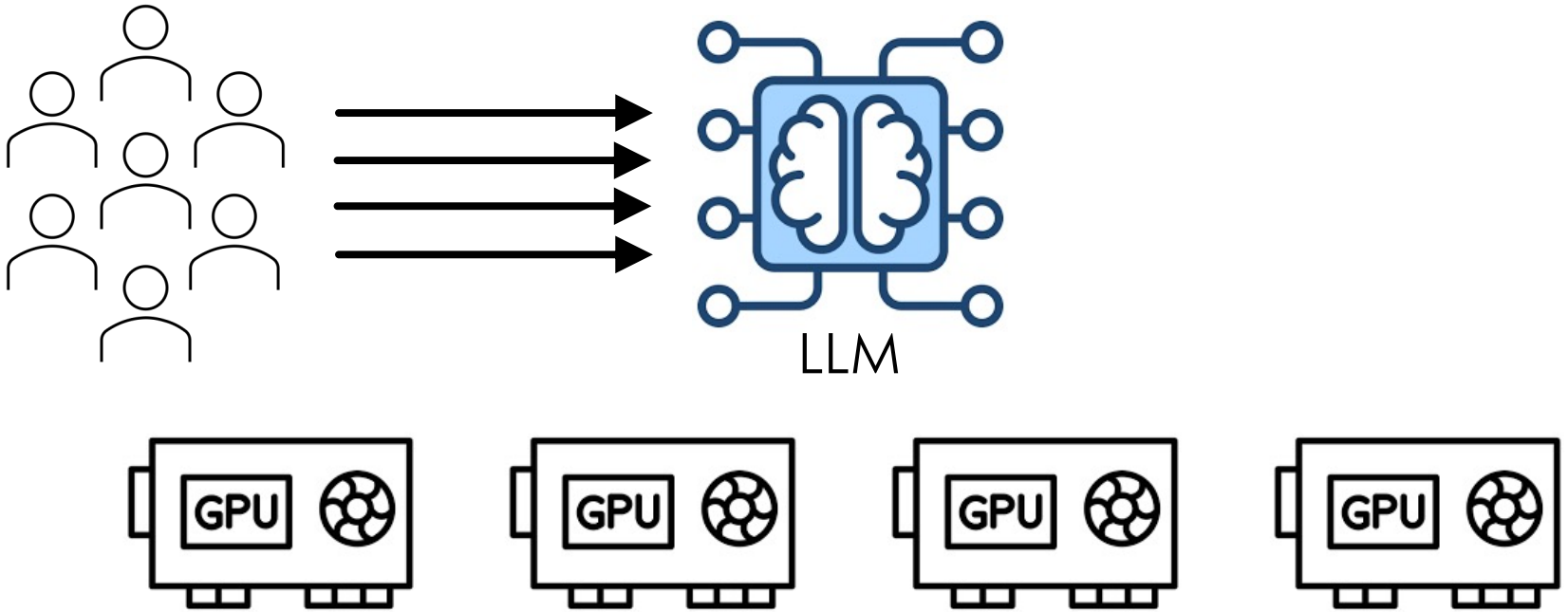
LLM Inference is Emerging in the Cloud

- Modern generative LLMs are turning ubiquitous
  - Use cases: programming, chat-bots, education, healthcare

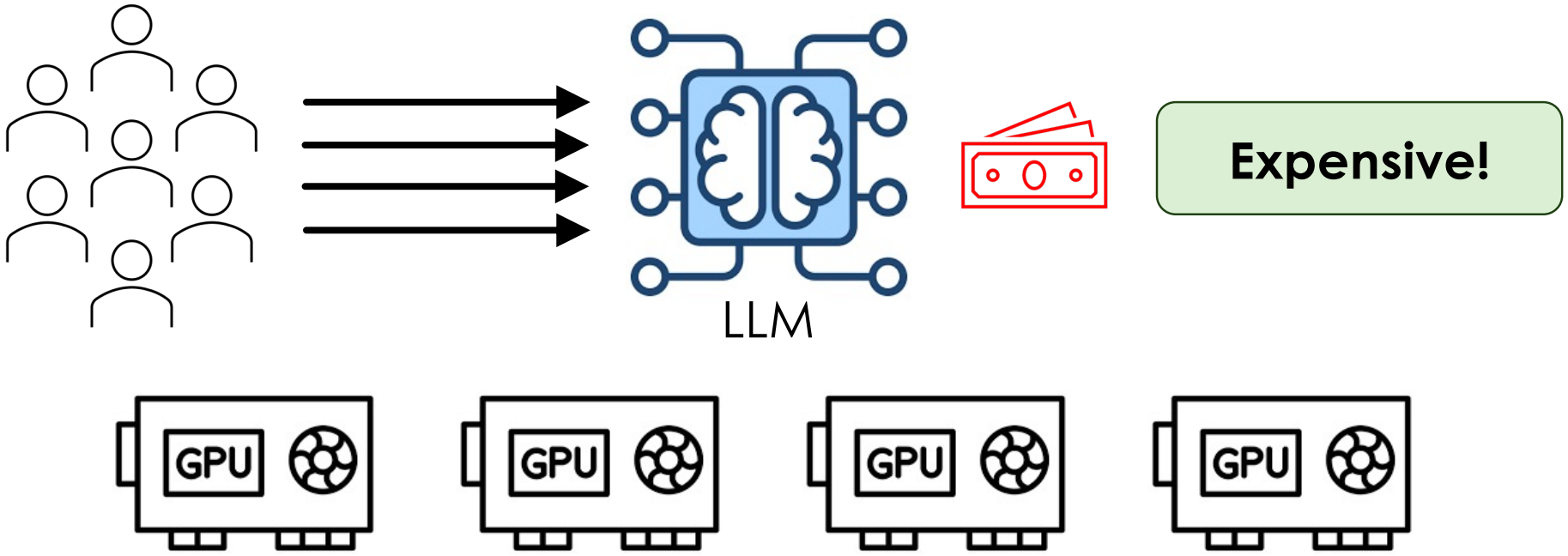
LLM Inference Stresses Cloud Infrastructure



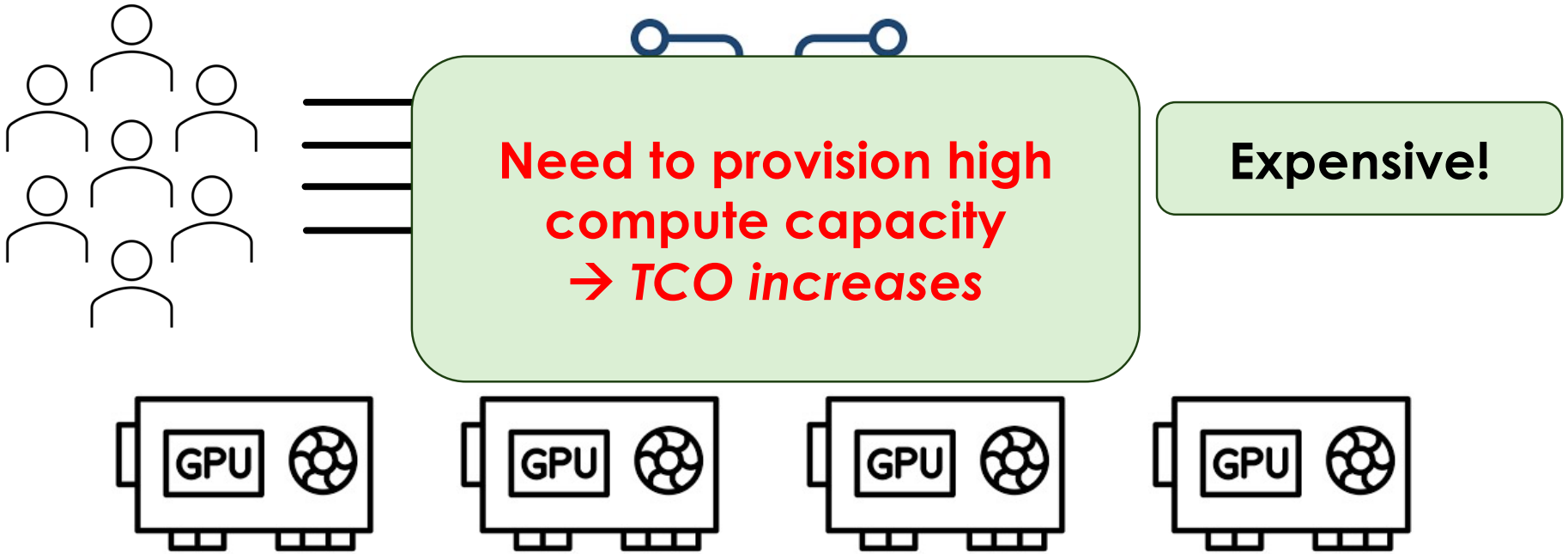
LLM Inference Stresses Cloud Infrastructure



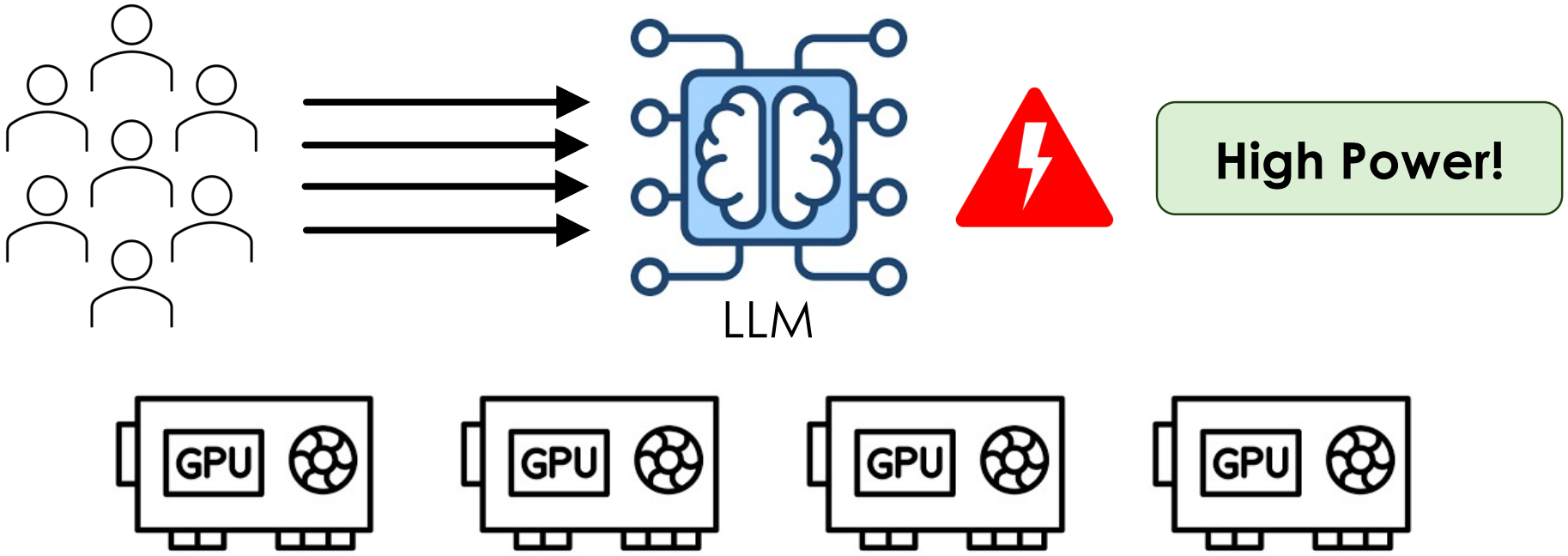
LLM Inference Stresses Cloud Infrastructure



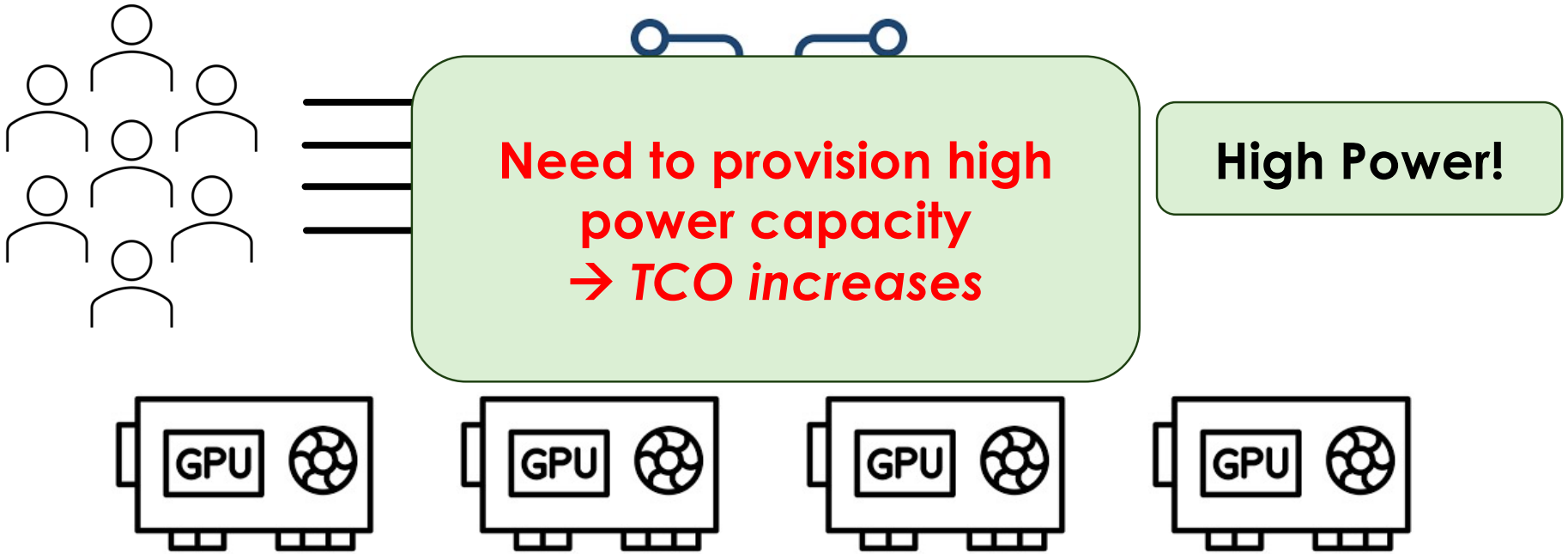
LLM Inference Stresses Cloud Infrastructure



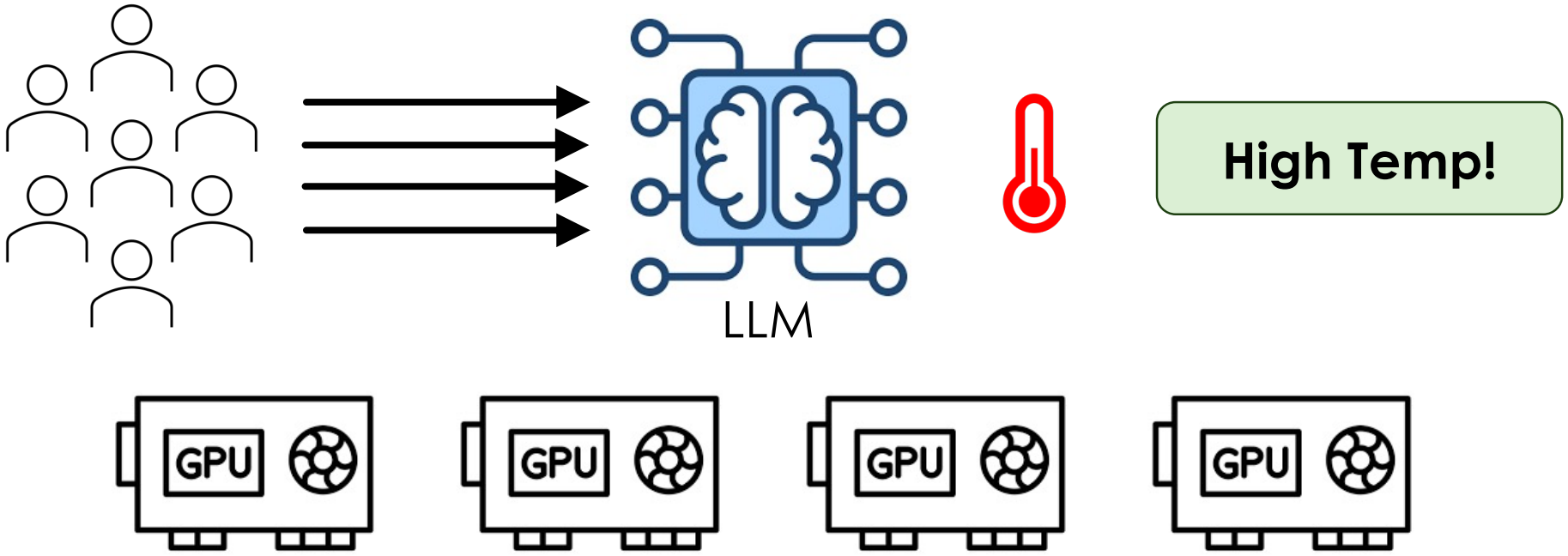
LLM Inference Stresses Cloud Infrastructure



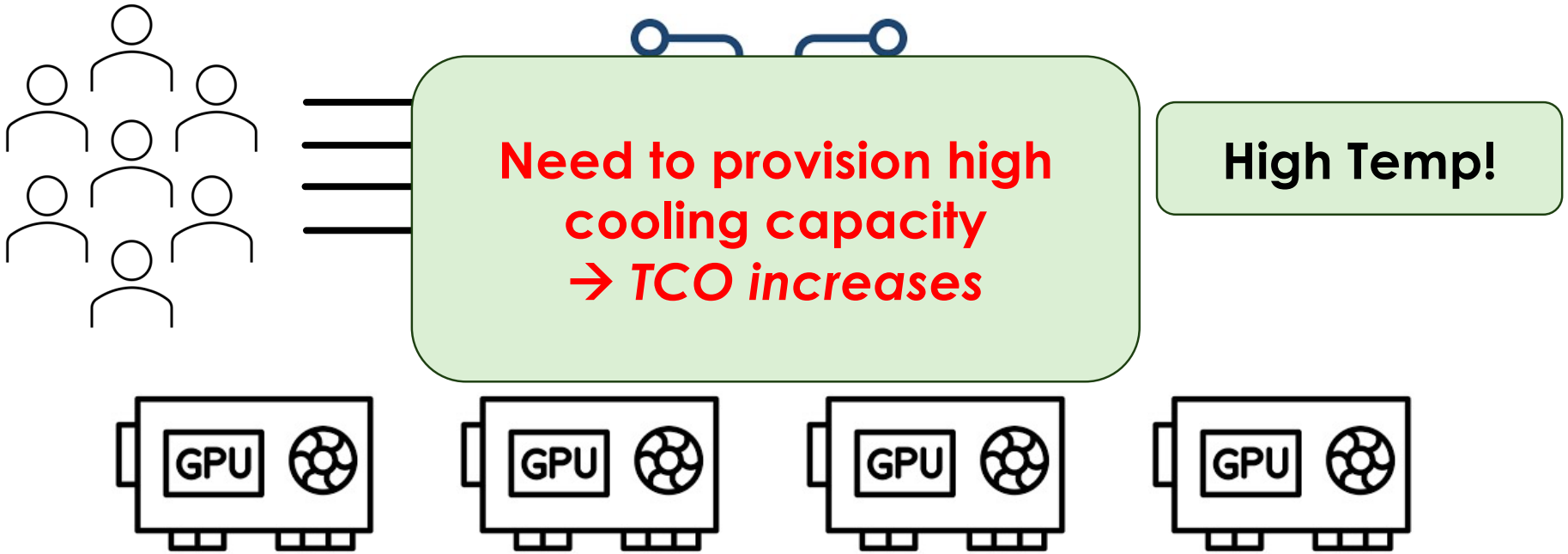
LLM Inference Stresses Cloud Infrastructure



LLM Inference Stresses Cloud Infrastructure

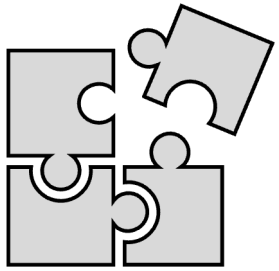


LLM Inference Stresses Cloud Infrastructure



How to Tame the LLMs?

- Lots of work on performance, accuracy, scalability
- LLMs cause the datacenter cost, power, and cooling requirements to skyrocket



Thermal and power efficiency of LLMs is a missing piece of a puzzle!

Contributions

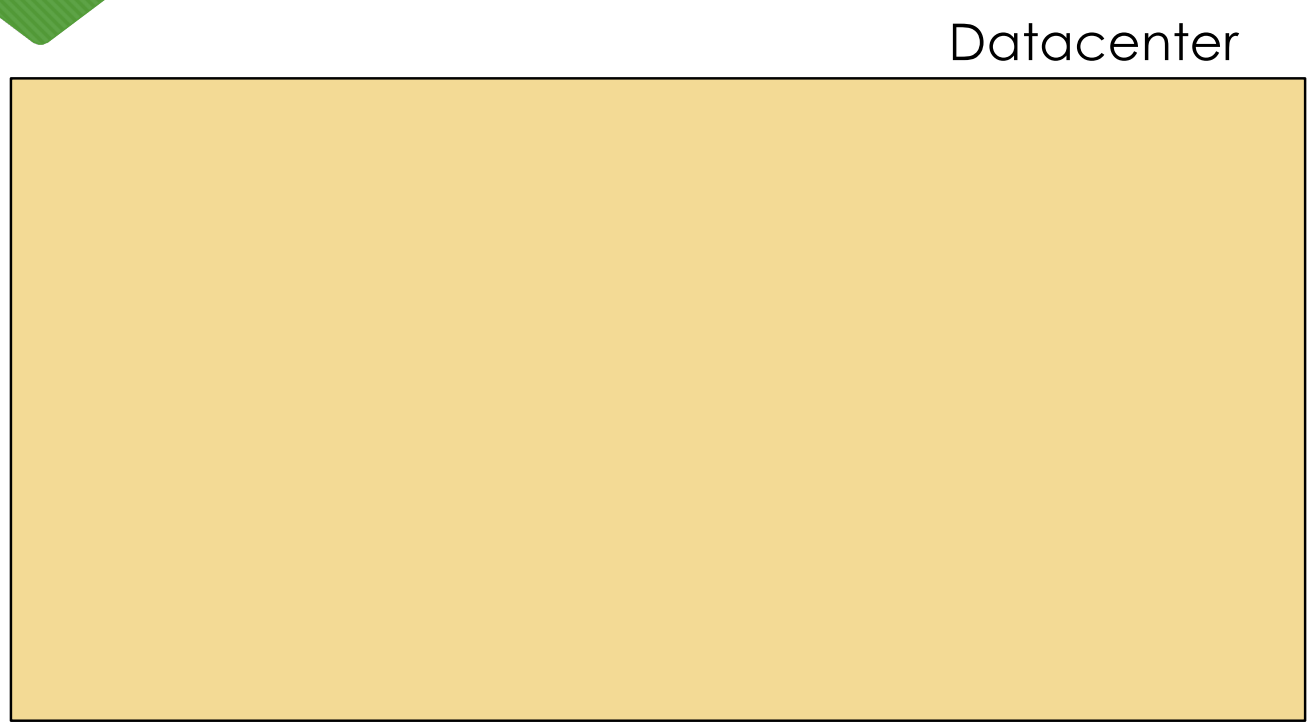
- Characterize thermal and power properties of LLMs and their behavior at production scale
- **TAPAS:** thermal- and power-aware scheduling for LLM inference clusters in the cloud
- Thorough evaluation on a GPU cluster + production traces
  - 30%-40% reduction in peak temperature and power

Goal: Make LLMs Thermal- and Power-Efficient

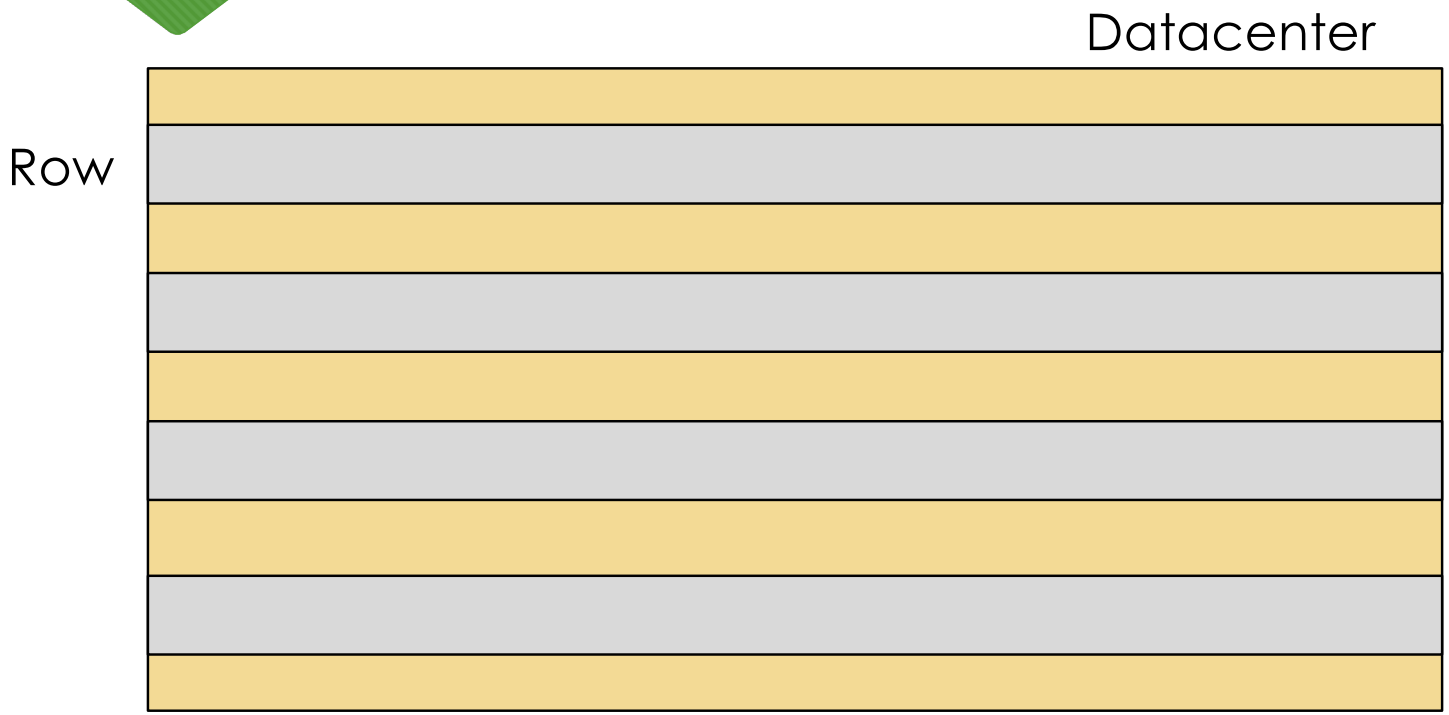
Goal: Make LLMs Thermal- and Power-Efficient

- Challenges
  - Cooling management
  - Power management
  - Workload heterogeneity

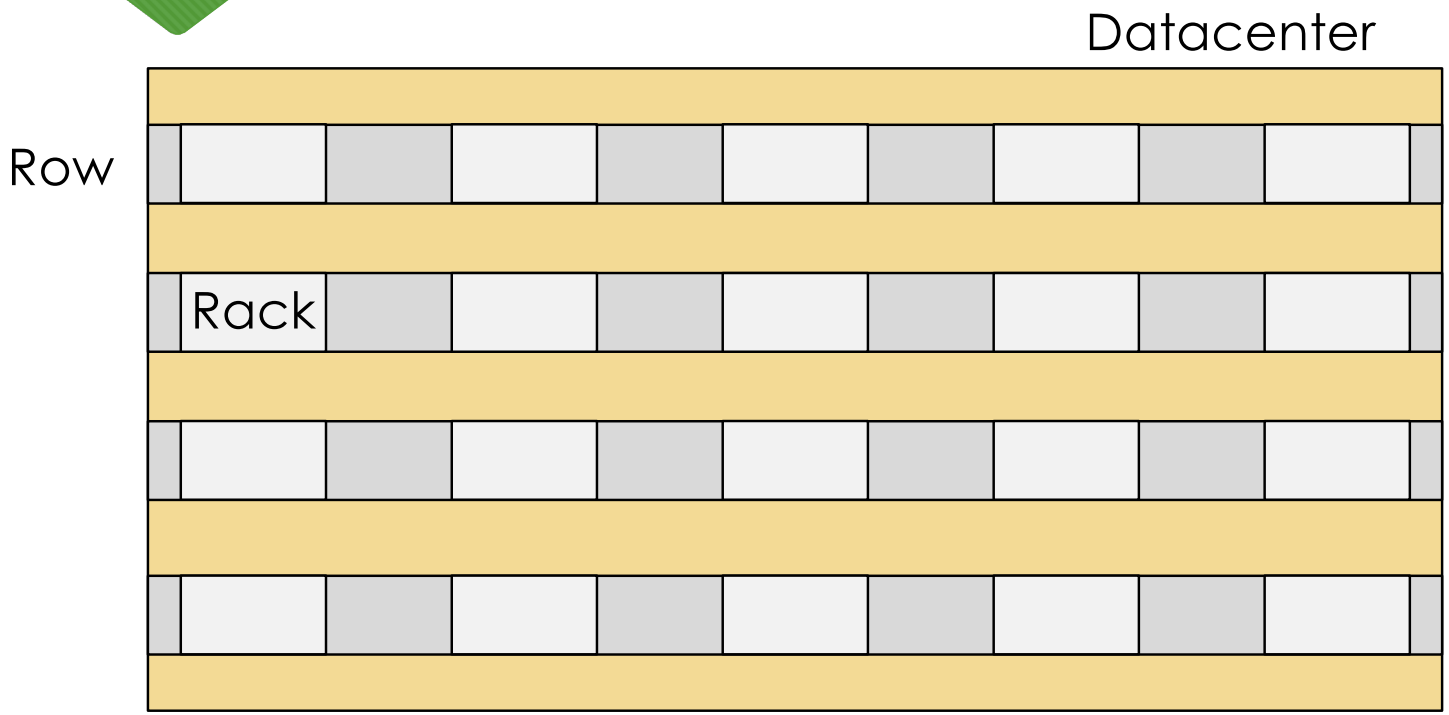
Challenge #1: Cooling Management



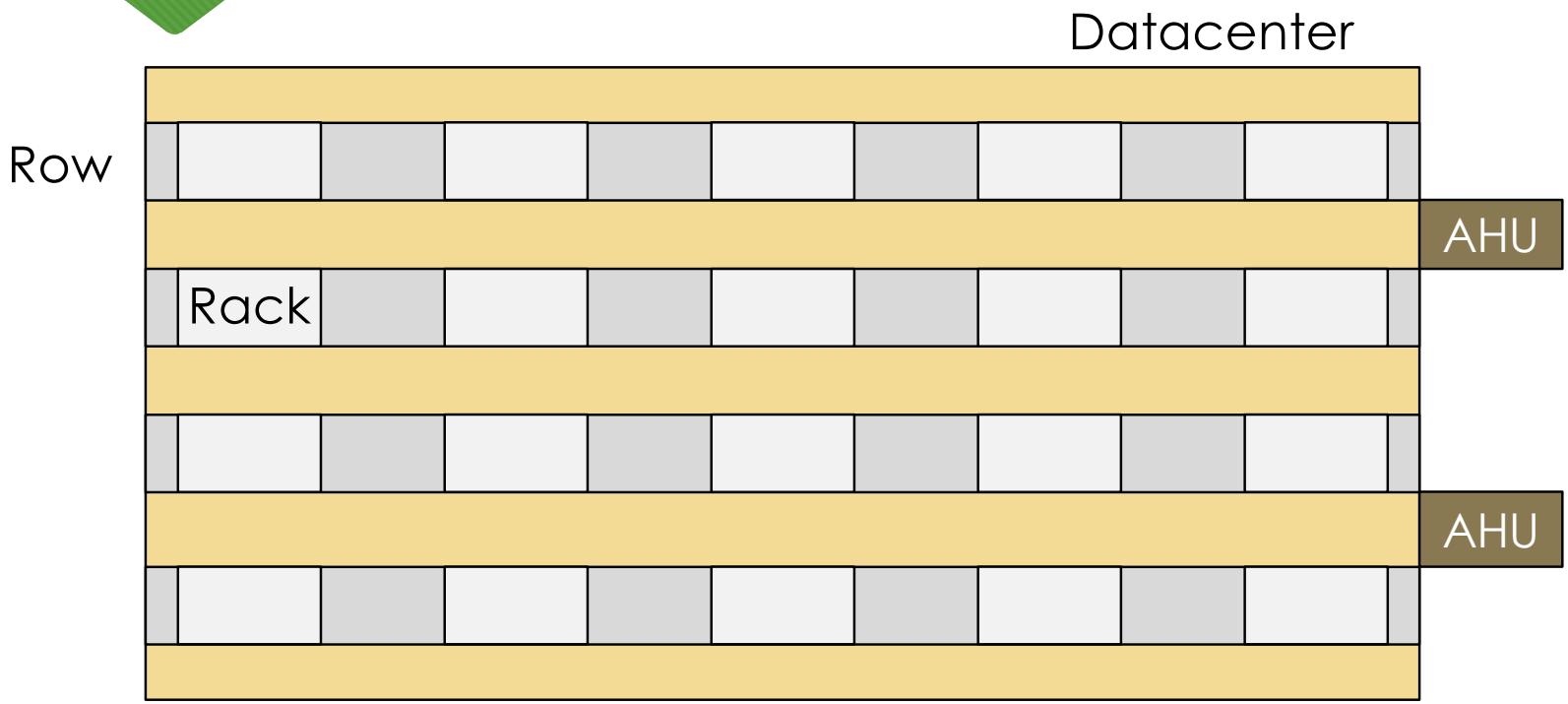
Challenge #1: Cooling Management



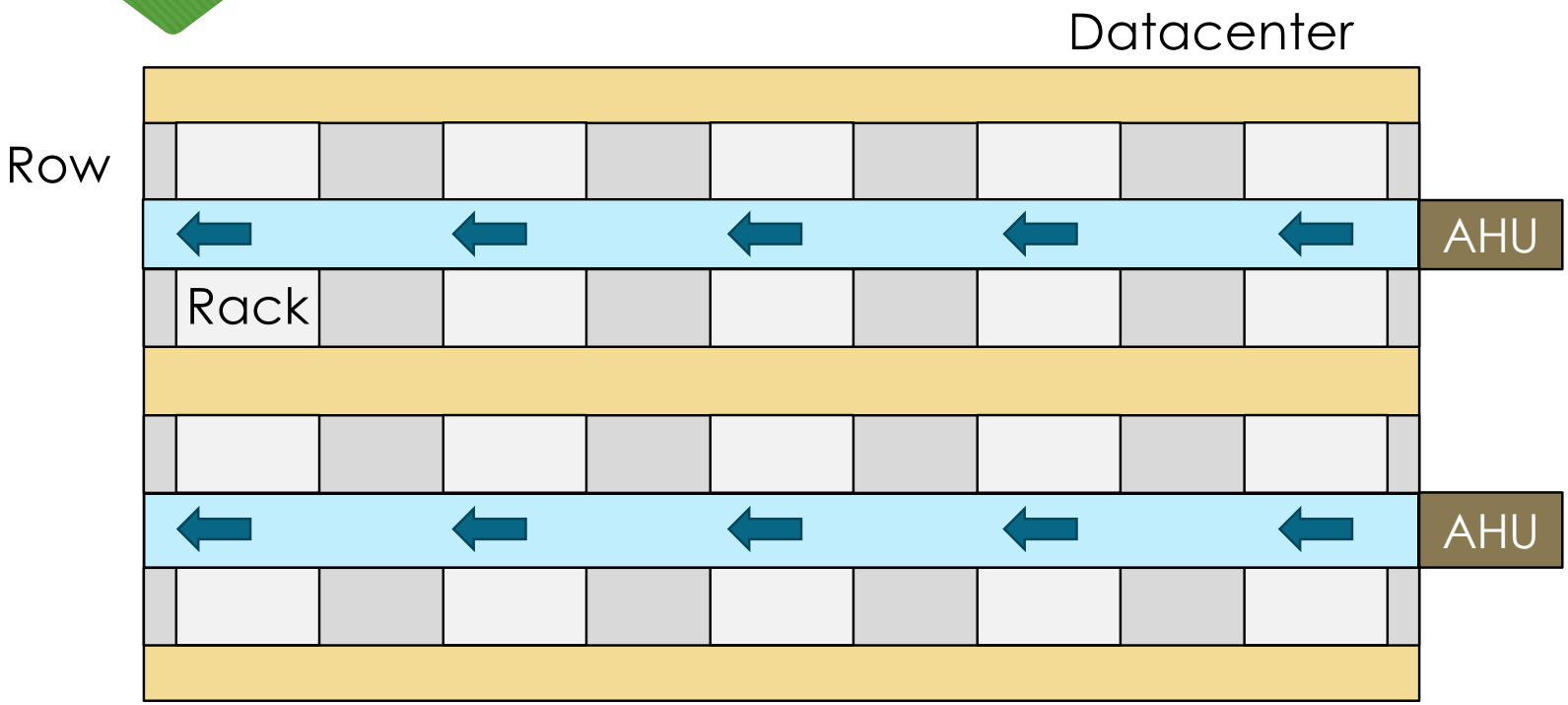
Challenge #1: Cooling Management



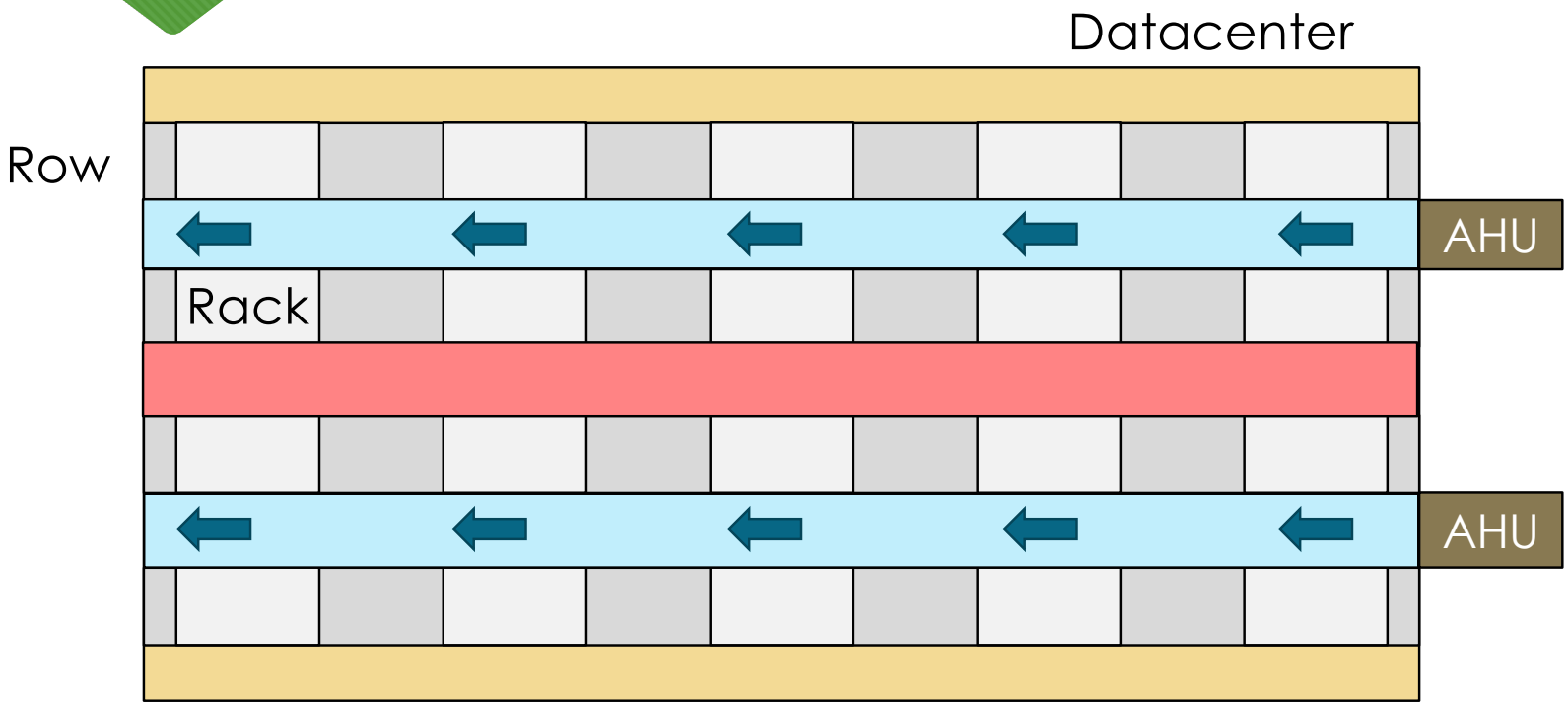
Challenge #1: Cooling Management



Challenge #1: Cooling Management



Challenge #1: Cooling Management

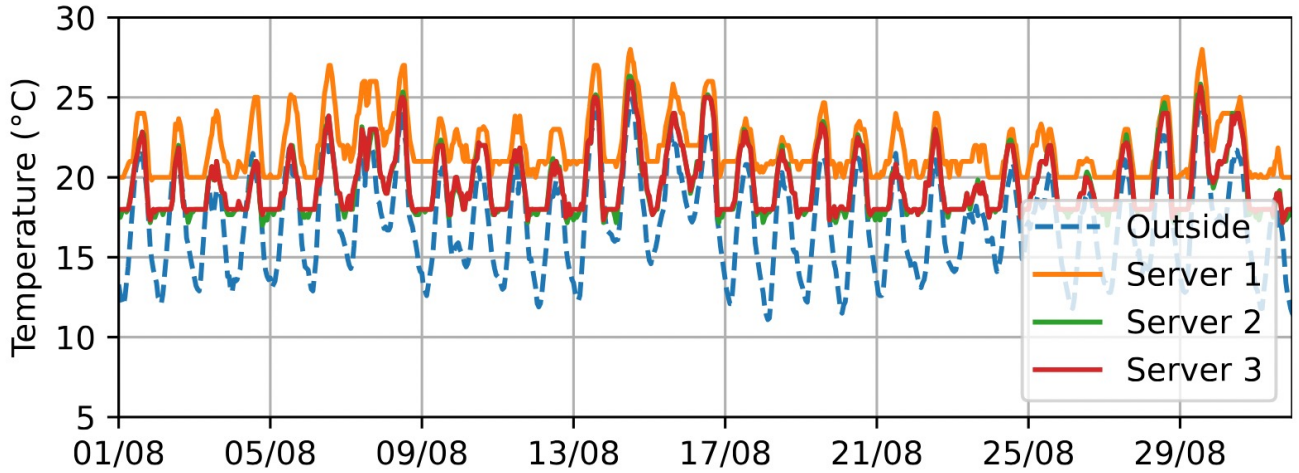


Challenge #1: Cooling Management

- Datacenter operators provision cooling infrastructure to sustain ***peak load***
  - Enough airflow in each aisle to prevent heat recirculation
  - Enough cooling to lower the temperature within operating conditions

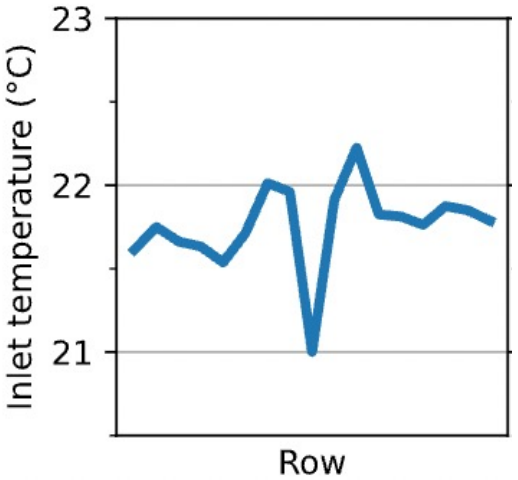
Challenge #1: Thermal Hotspots!

- Temperature of a GPU server depends on:
  1. Outside temperature



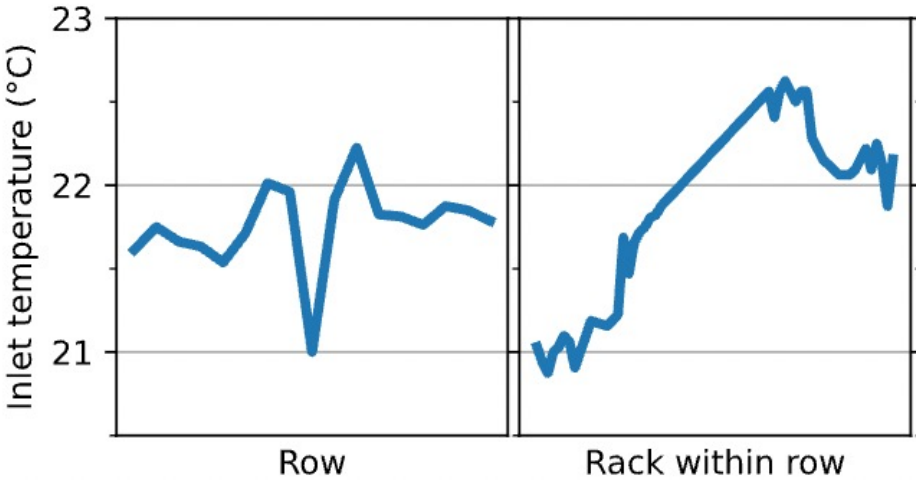
Challenge #1: Thermal Hotspots!

- Temperature of a GPU server depends on:
  2. Datacenter layout



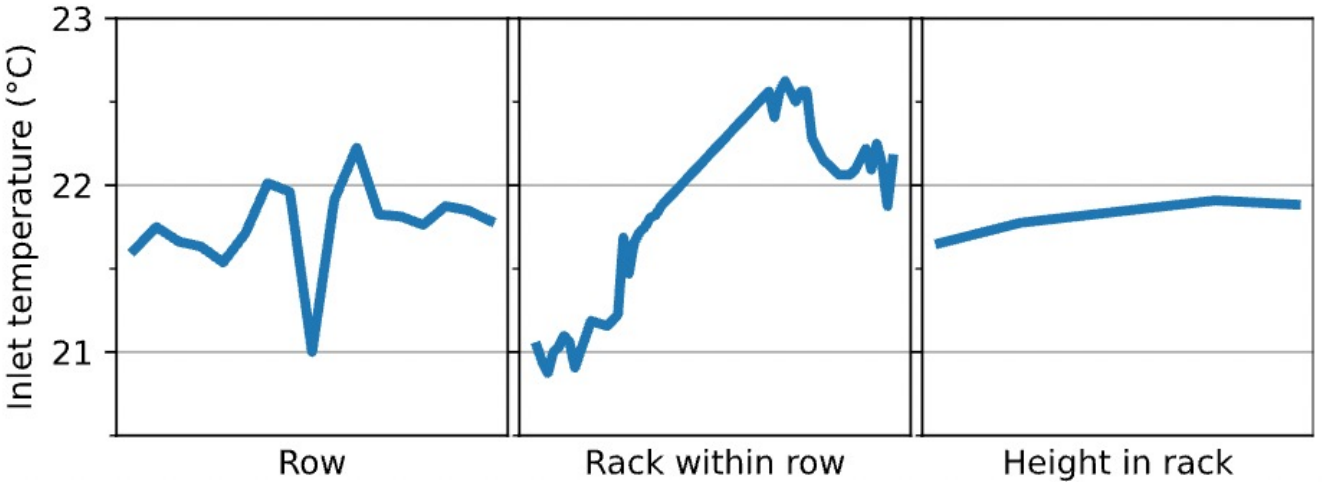
Challenge #1: Thermal Hotspots!

- Temperature of a GPU server depends on:
  2. Datacenter layout



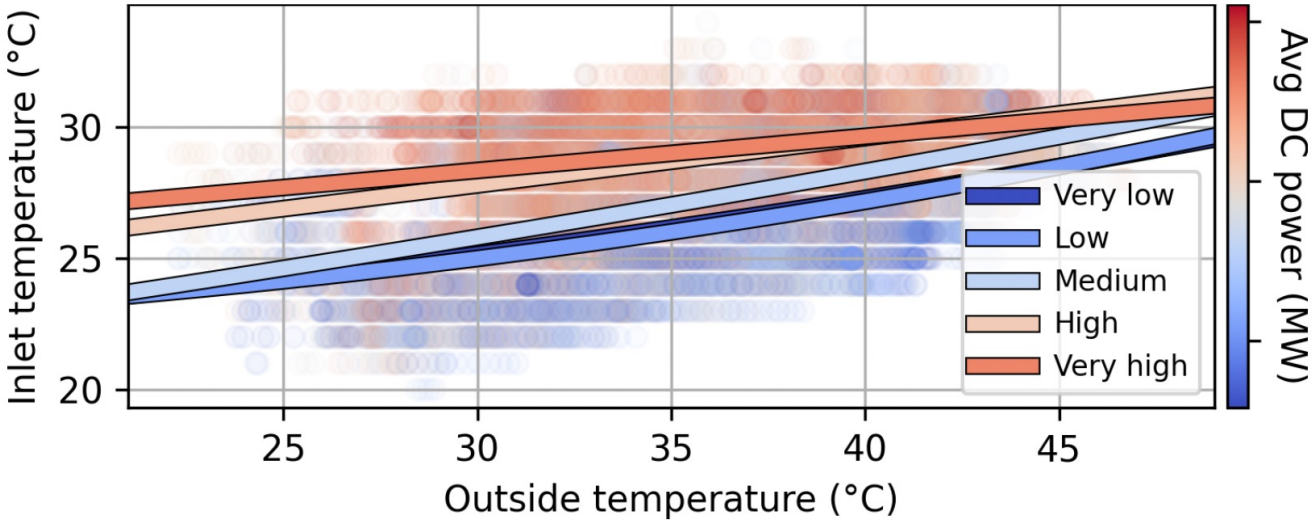
Challenge #1: Thermal Hotspots!

- Temperature of a GPU server depends on:
  2. Datacenter layout



Challenge #1: Thermal Hotspots!

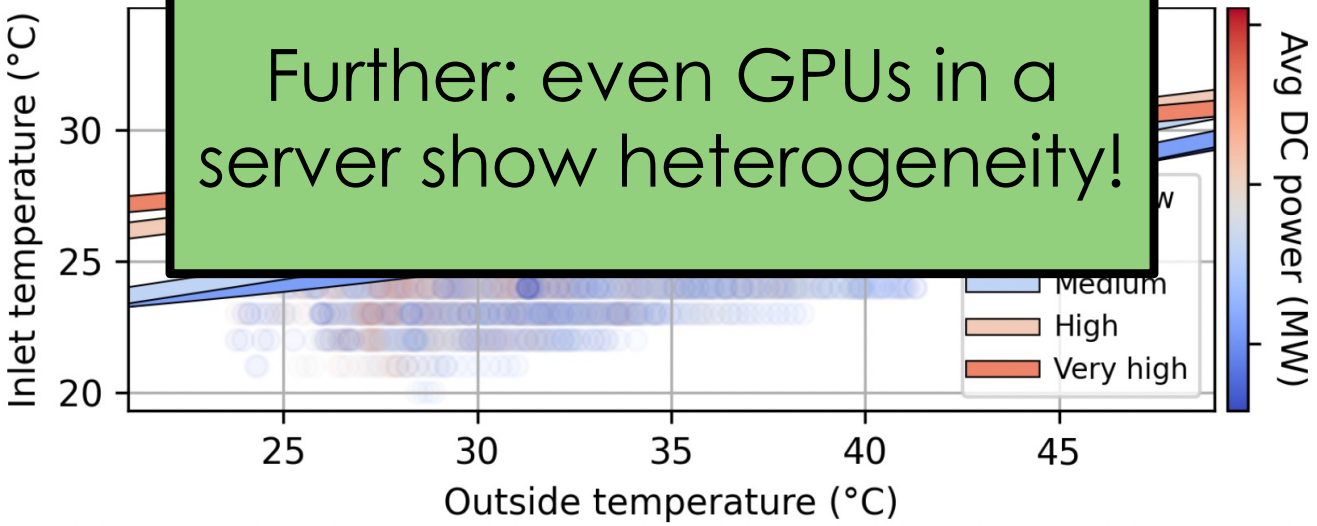
- Temperature of a GPU server depends on:
  3. Server's load



Challenge #1: Thermal Hotspots!

- Temperature of a GPU server depends on:
  3. Server's load

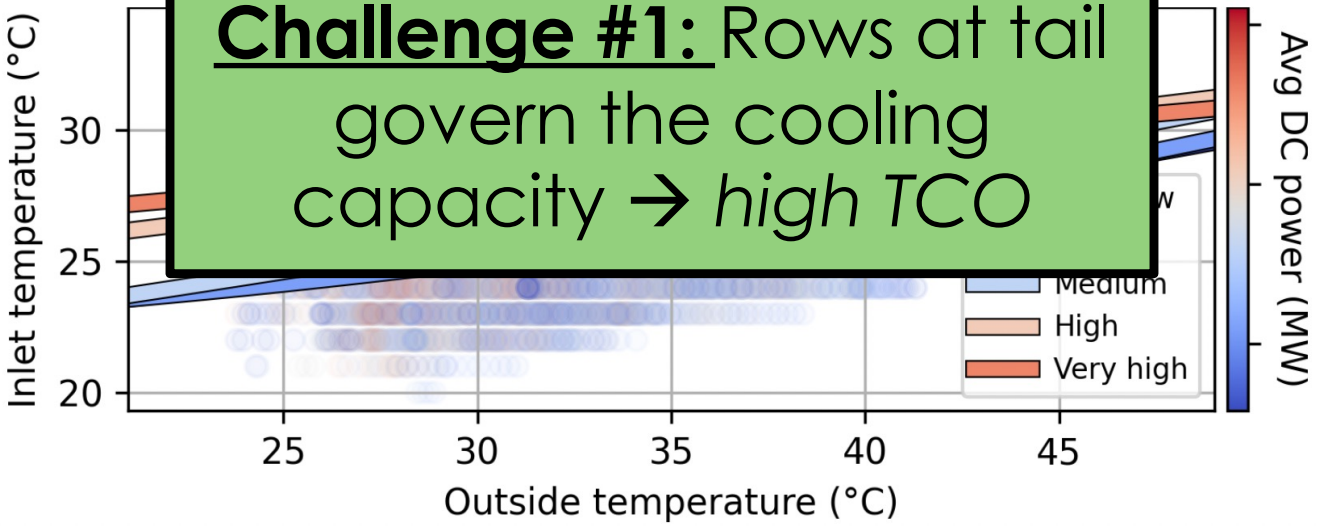
Further: even GPUs in a server show heterogeneity!



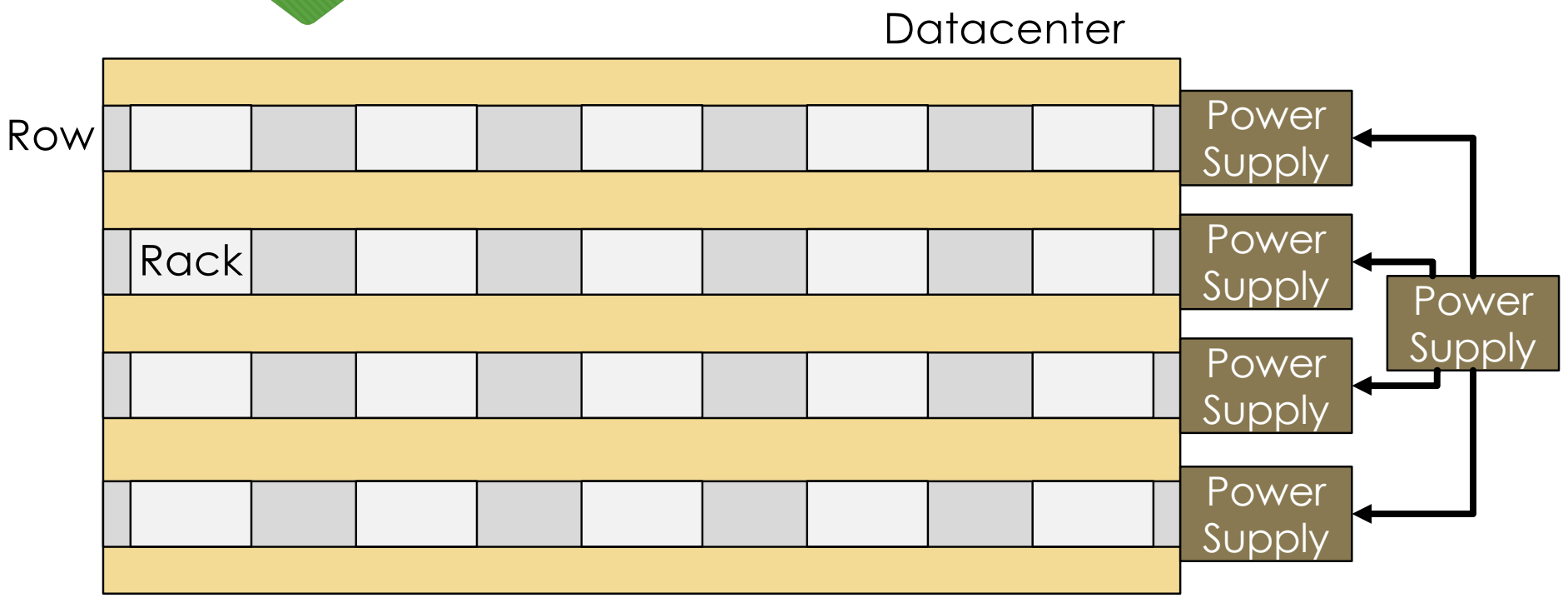
Challenge #1: Thermal Hotspots!

- Temperature of a GPU server depends on:
  3. Server's load

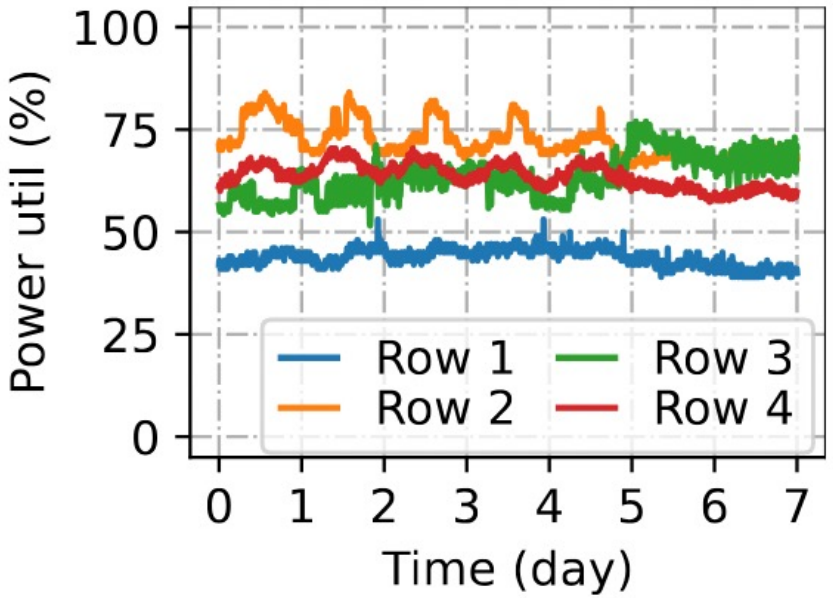
Challenge #1: Rows at tail govern the cooling capacity → *high TCO*



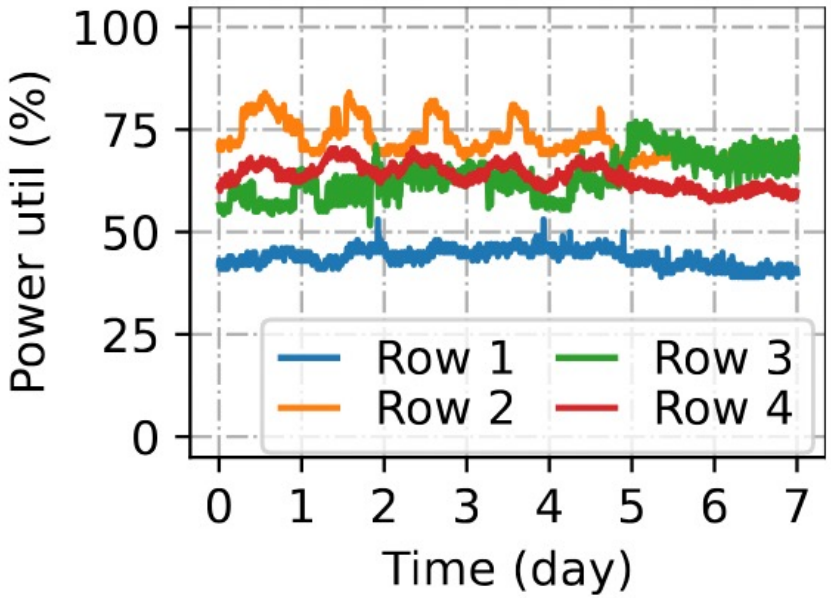
Challenge #2: Power Management

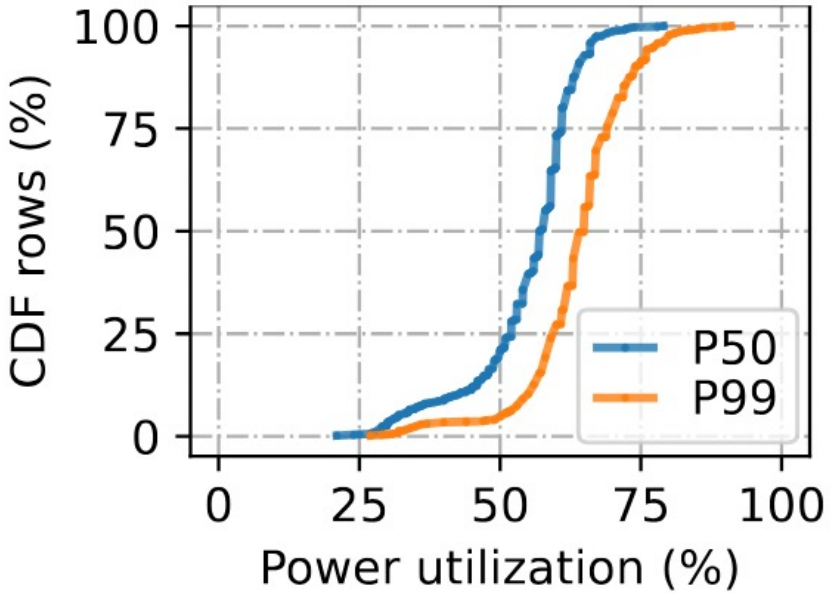


Challenge #2: Power Imbalance!

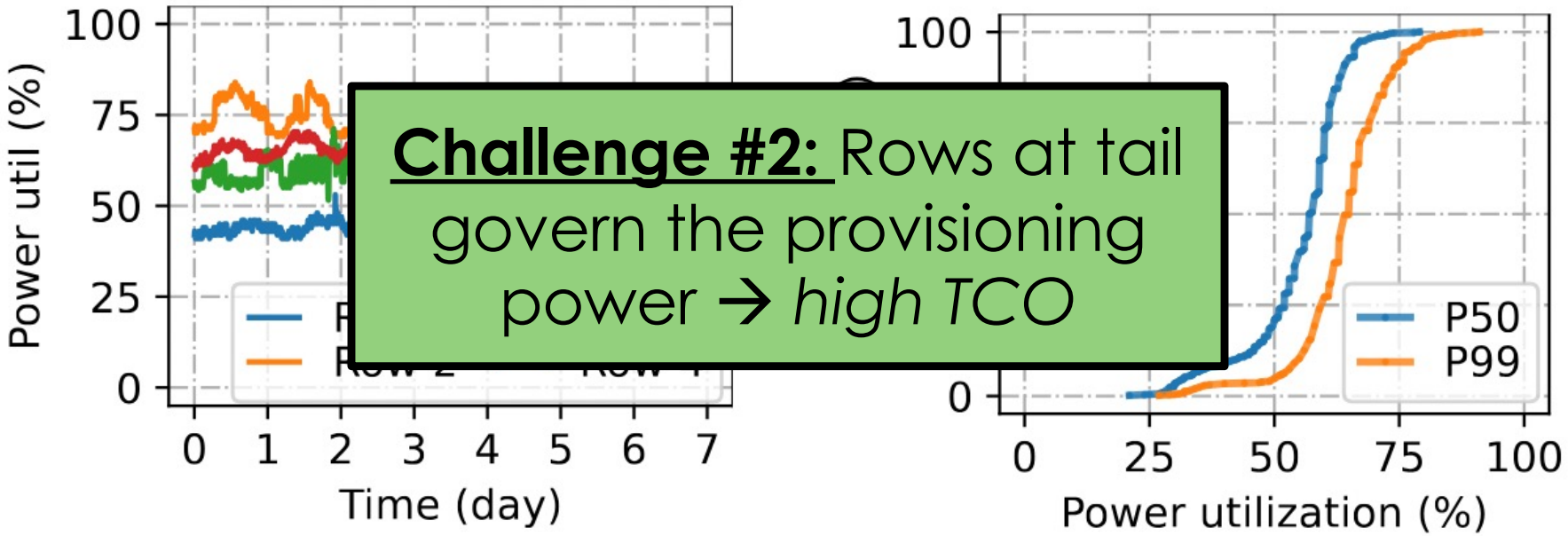


Challenge #2: Power Imbalance!



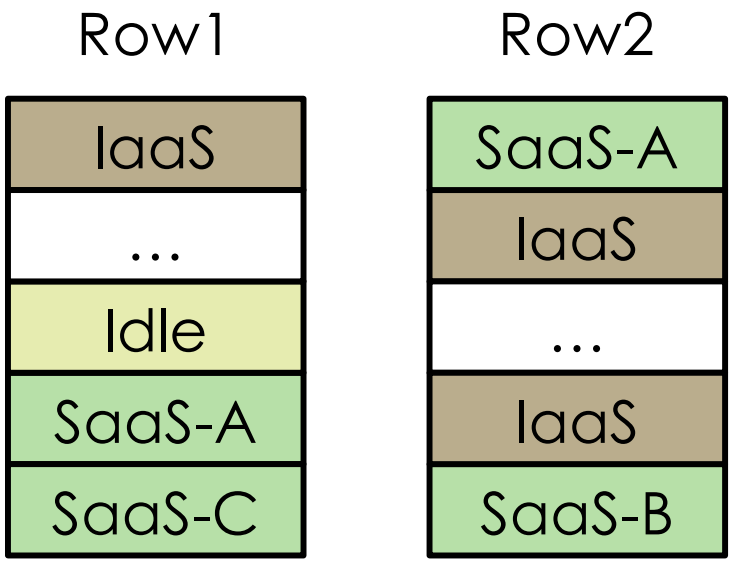


Challenge #2: Power Imbalance!

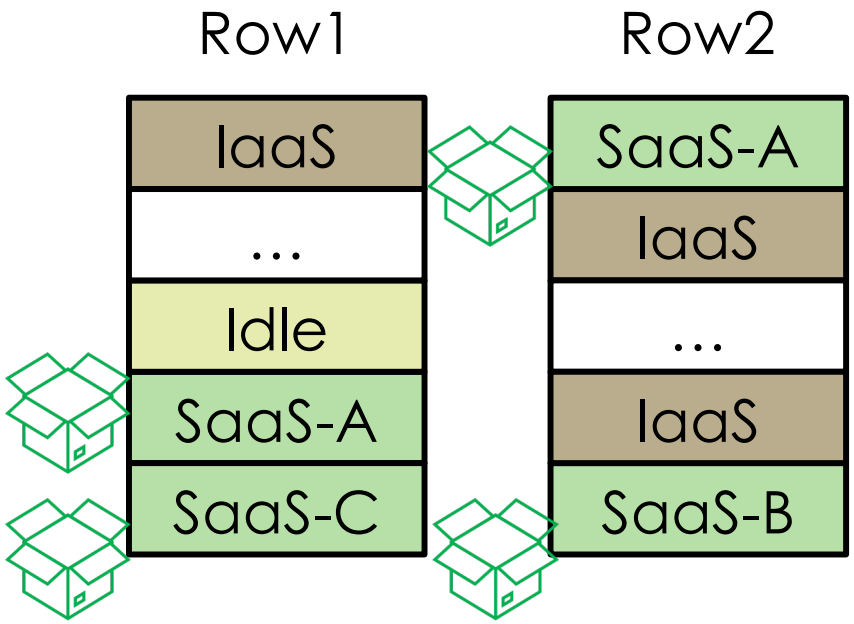


Challenge #2: Rows at tail govern the provisioning power → *high TCO*

Challenge #3: Heterogeneous Workloads



Challenge #3: Heterogeneous Workloads



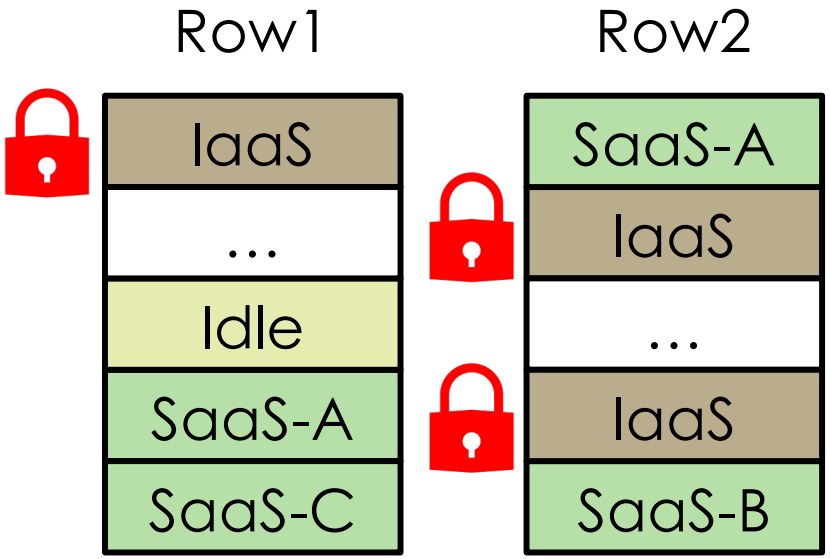
SaaS services

First-party

Good visibility

Good flexibility

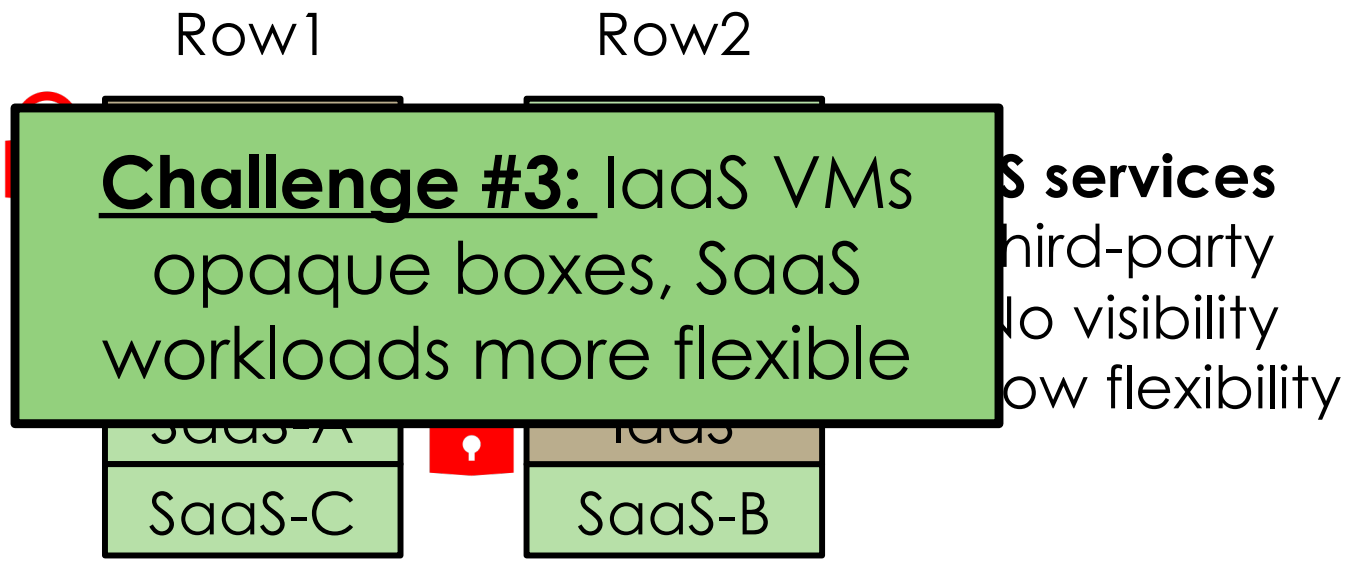
Challenge #3: Heterogeneous Workloads



IaaS services

- Third-party
- No visibility
- Low flexibility

Challenge #3: Heterogeneous Workloads



Challenge #3: IaaS VMs opaque boxes, SaaS workloads more flexible

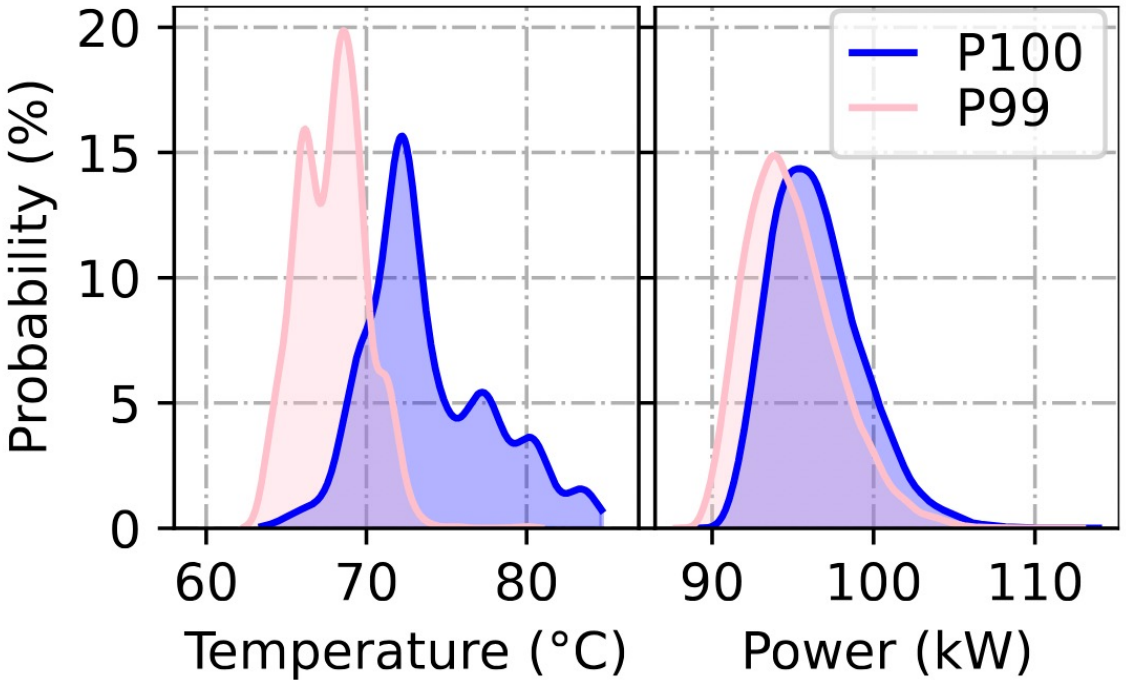
S services
hird-party
lo visibility
ow flexibility

How to Address the Challenges?

- Thermal- and power-aware:
  - Workload placement
  - Instance configuration
  - Request scheduling

Opportunity #1: Smart Workload Placement

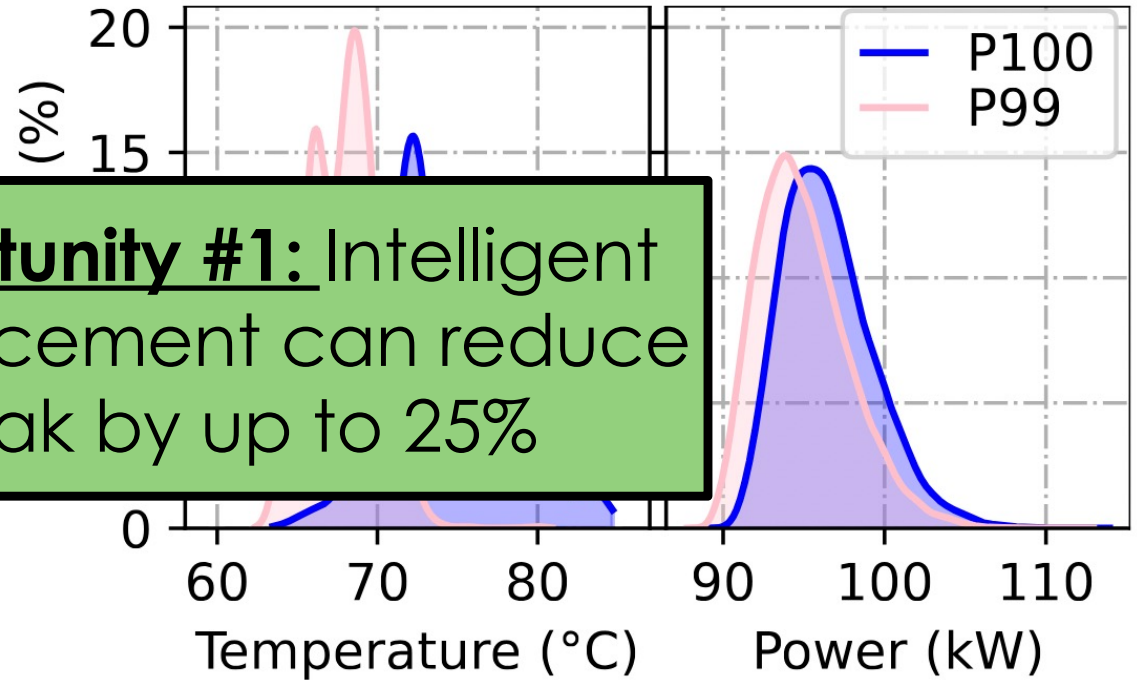
- 100K random VM placements
- Generate diverse thermal and power distributions



Opportunity #1: Smart Workload Placement

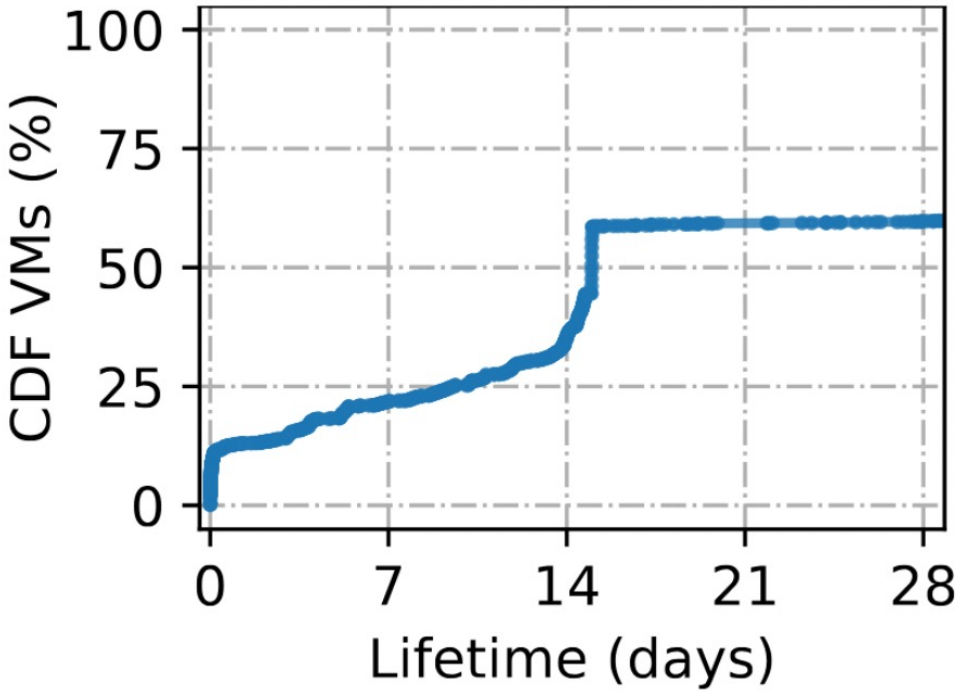
- 100K random VM placements
- Generate c
  thermal anc
  distributions

Opportunity #1: Intelligent VM placement can reduce peak by up to 25%



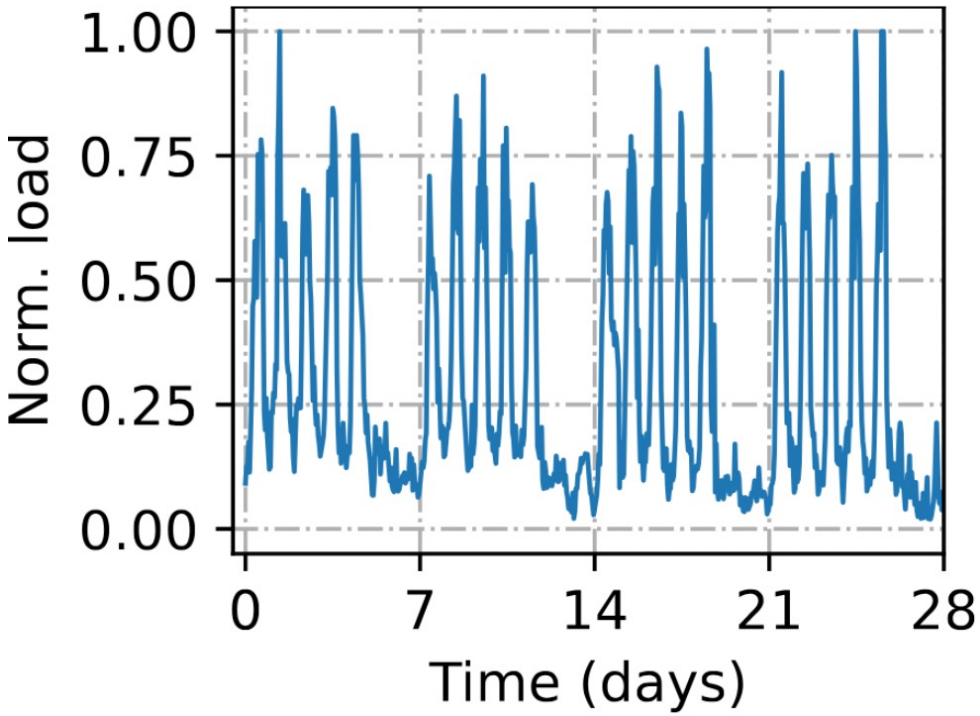
Opportunity #1: Smart Workload Placement

- LLM inference VMs are long lived



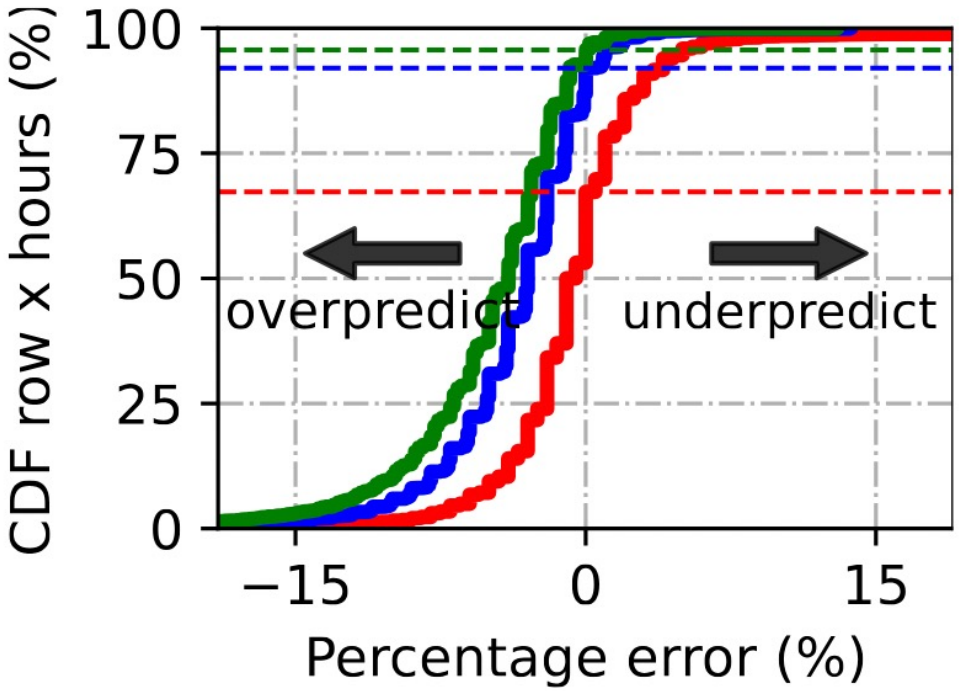
Opportunity #1: Smart Workload Placement

- LLM inference loads are periodic



Opportunity #1: Smart Workload Placement

- Long-lived instances with periodic loads → high predictability
- Accuracy over 90%



Opportunity #2: Tune Flexible SaaS LLM VMs

Knob	Power	Temp	Perf	Quality

Opportunity #2: Tune Flexible SaaS LLM VMs

Knob	Power	Temp	Perf	Quality
Model Size ↓	↓	↓	↑	↓↓

Opportunity #2: Tune Flexible SaaS LLM VMs

Knob	Power	Temp	Perf	Quality
Model Size ↓	↓	↓	↑	↓↓
Quantize ↓	↓	↓	↑	↓

Opportunity #2: Tune Flexible SaaS LLM VMs

Knob ↓	Power ↓	Temp ↓	Perf ↑	Quality ↓↓
Model Size ↓	↓	↓	↑	↓↓
Quantize ↓	↓	↓	↑	↓
Parallelism ↓	↓	↑	↑	

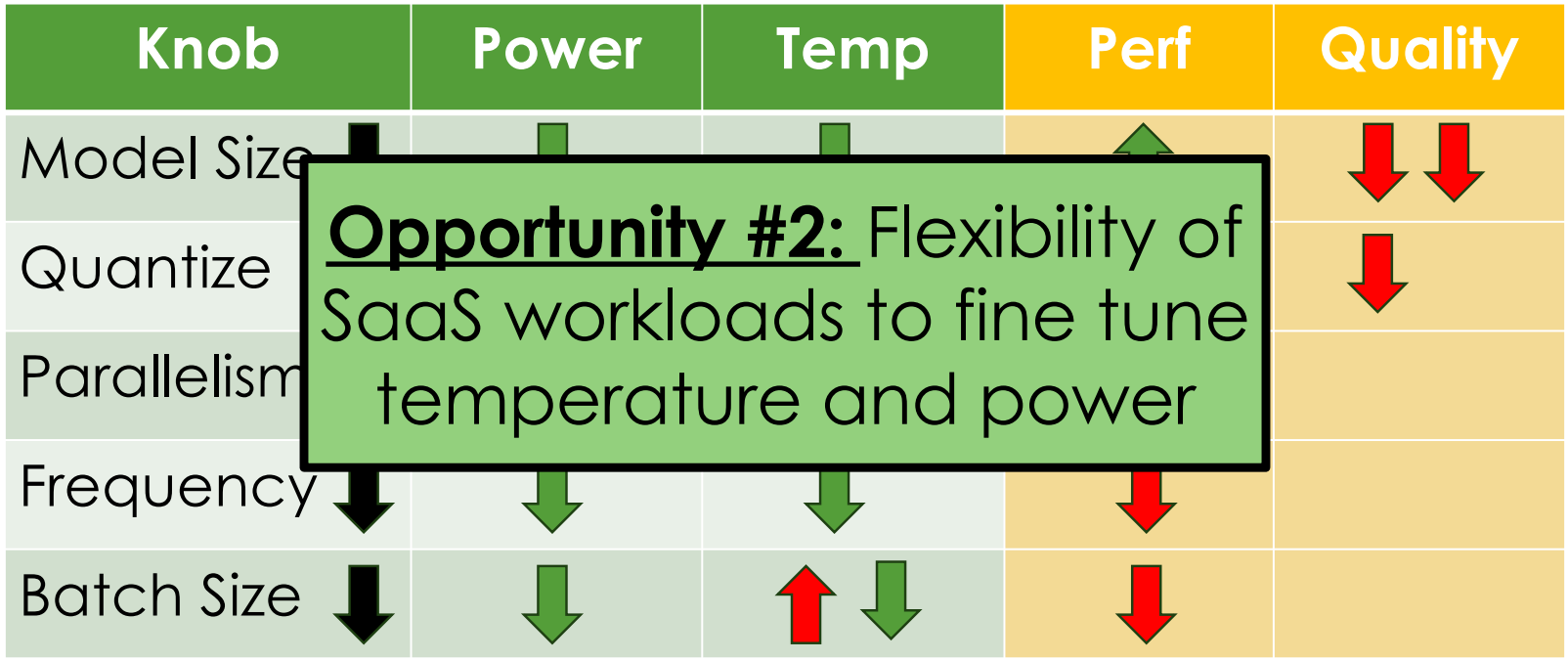
Opportunity #2: Tune Flexible SaaS LLM VMs

Knob	Power	Temp	Perf	Quality
Model Size ↓	↓	↓	↑	↓↓
Quantize ↓	↓	↓	↑	↓
Parallelism ↓	↓	↑	↑	
Frequency ↓	↓	↓	↓	

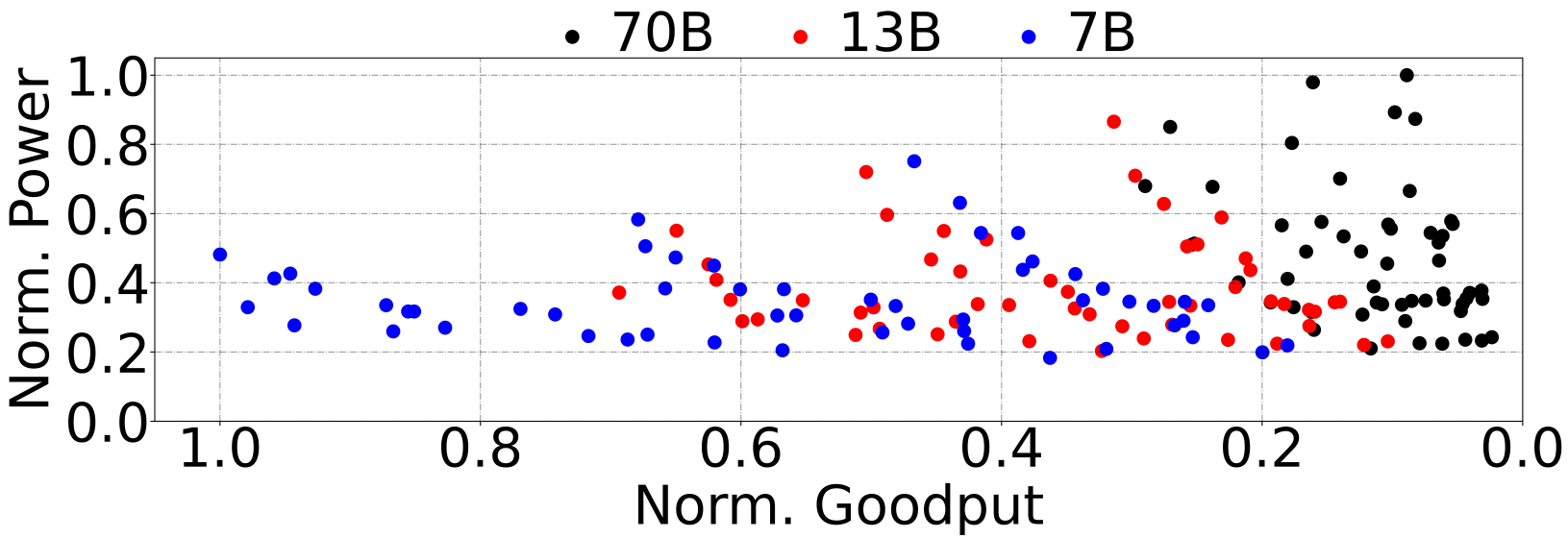
Opportunity #2: Tune Flexible SaaS LLM VMs

Knob ↓	Power ↓	Temp ↓	Perf ↑	Quality ↓↓
Model Size ↓	↓	↓	↑	↓↓
Quantize ↓	↓	↓	↑	↓
Parallelism ↓	↓	↑	↑	
Frequency ↓	↓	↓	↓	
Batch Size ↓	↓	↑↓	↓	

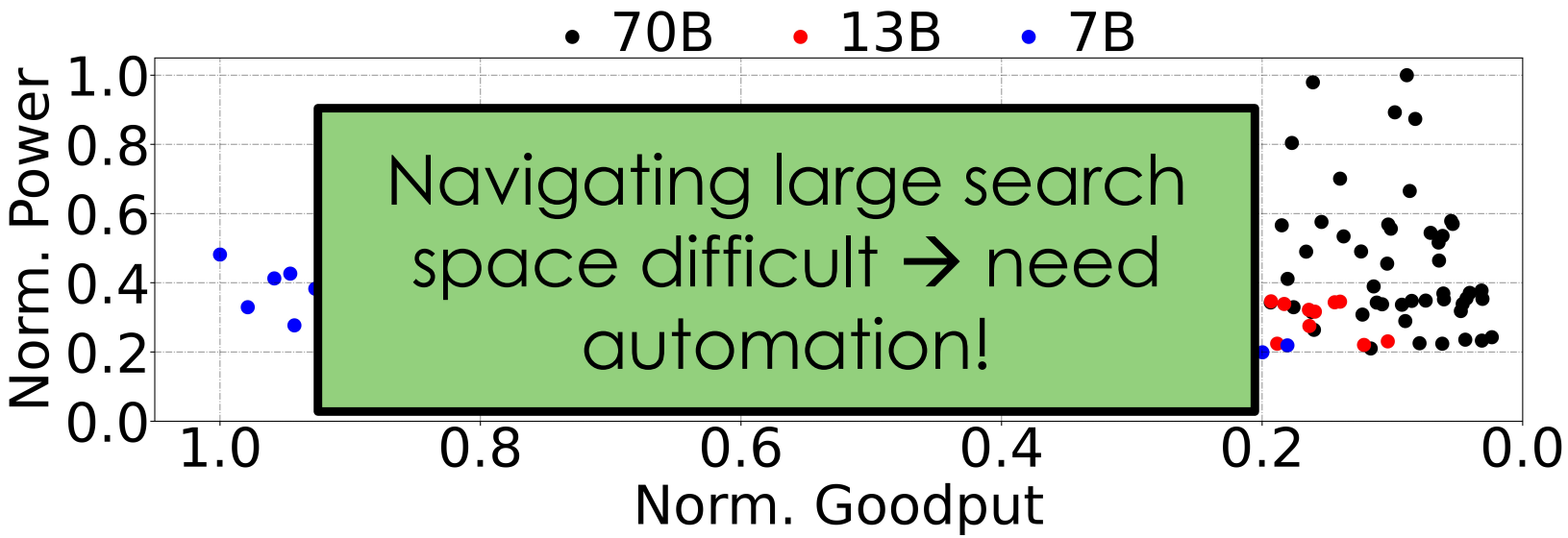
Opportunity #2: Tune Flexible SaaS LLM VMs



Opportunity #2: Tune Flexible SaaS LLM VMs

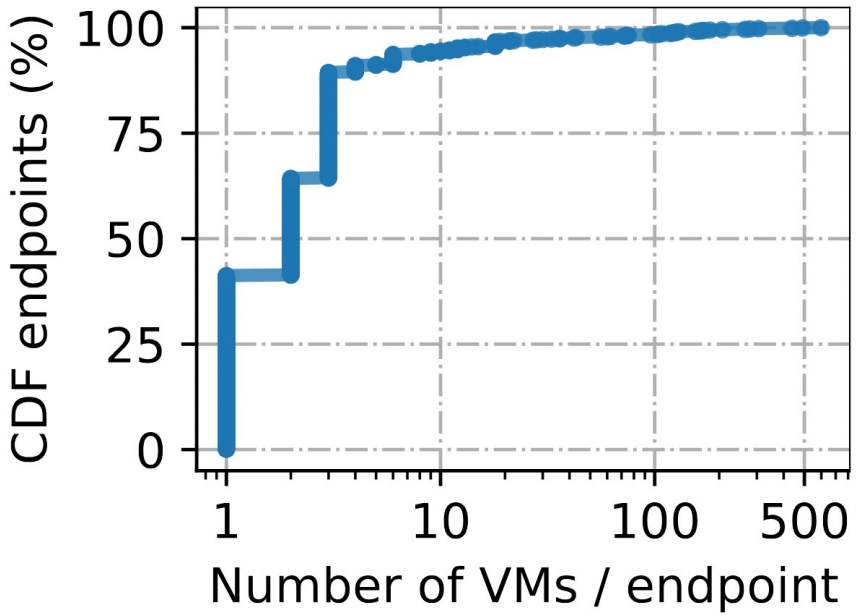


Opportunity #2: Tune Flexible SaaS LLM VMs



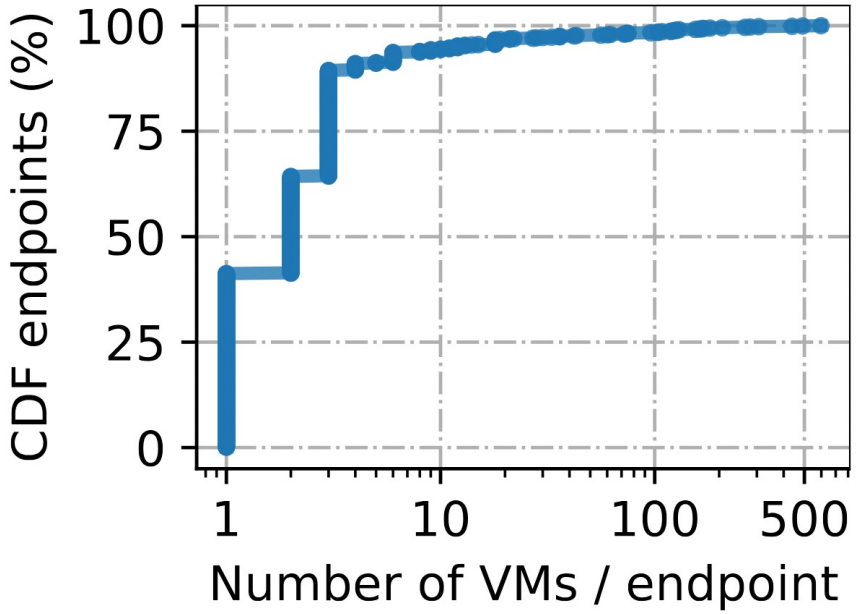
Opportunity #3: Smart Request Routing

- LLM inference service spans many VMs



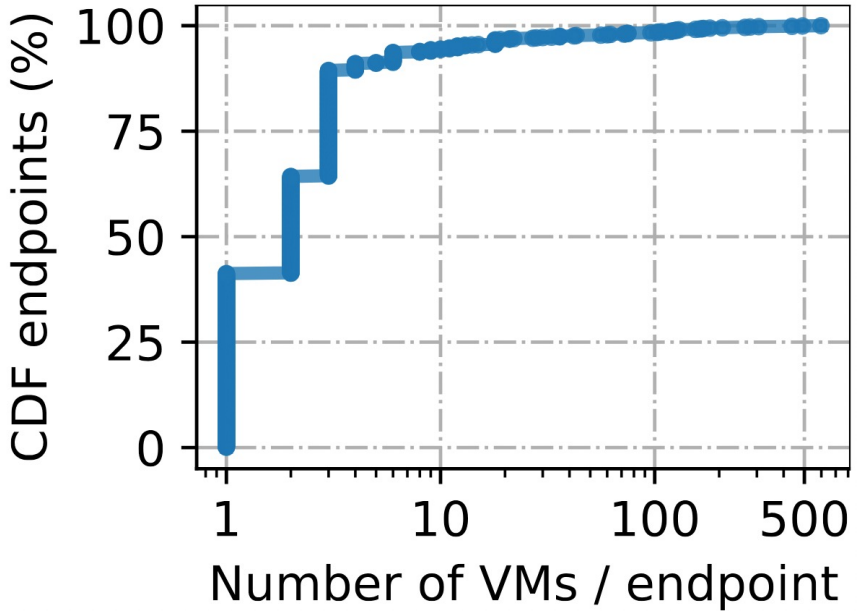
Opportunity #3: Smart Request Routing

- LLM inference service spans many VMs
- VMs of a given service can be placed in different rows



Opportunity #3: Smart Request Routing

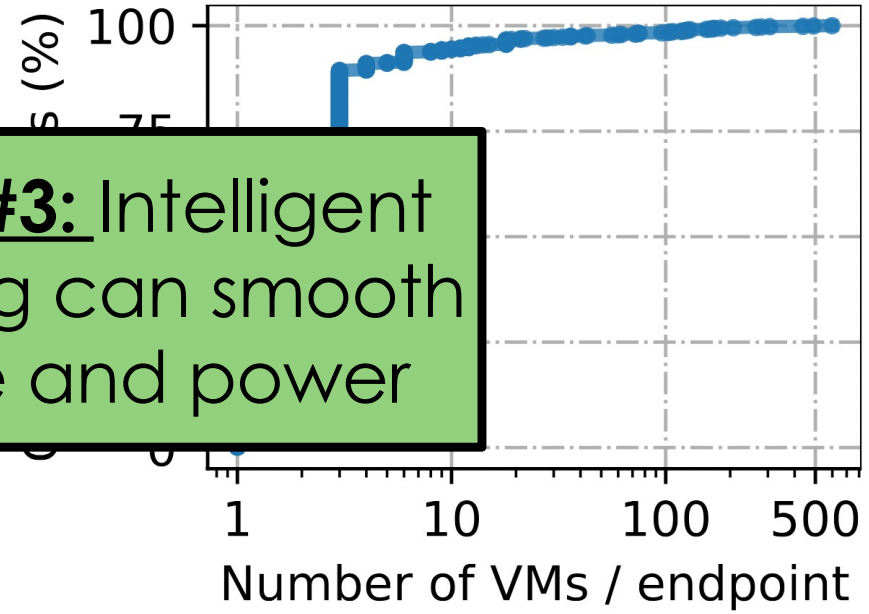
- LLM inference service spans many VMs
- VMs of a given service can be placed in different rows
- Load balance the requests in a thermal- and power-aware manner



Opportunity #3: Smart Request Routing

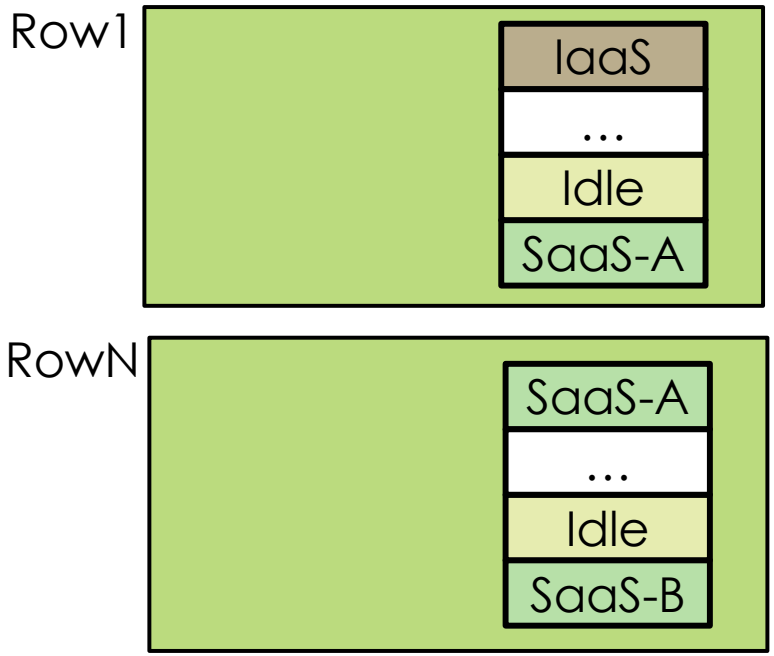
- LLM inference service spans many VMs
- VMs of a gi… be placed i…
- Load balan… in a thermal- and power-aware manner

Opportunity #3: Intelligent request routing can smooth temperature and power

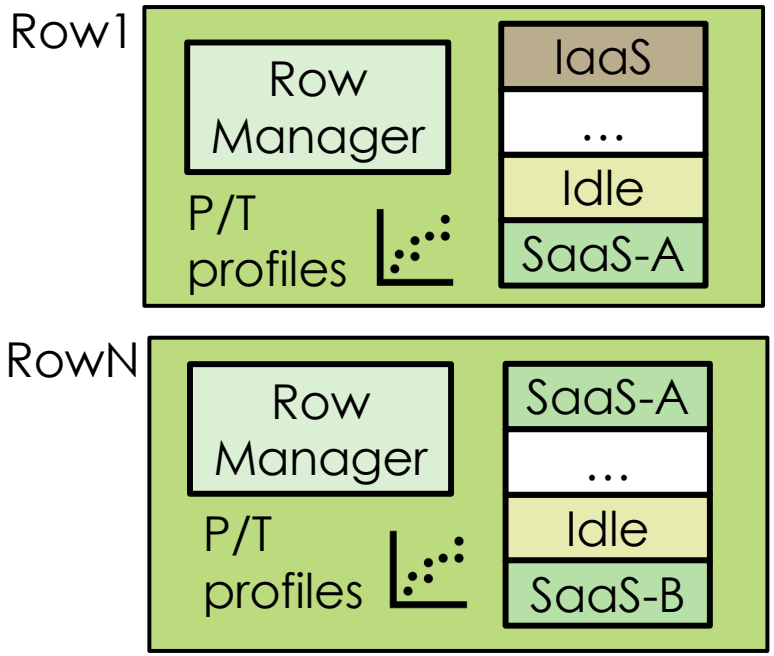


TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in the Cloud

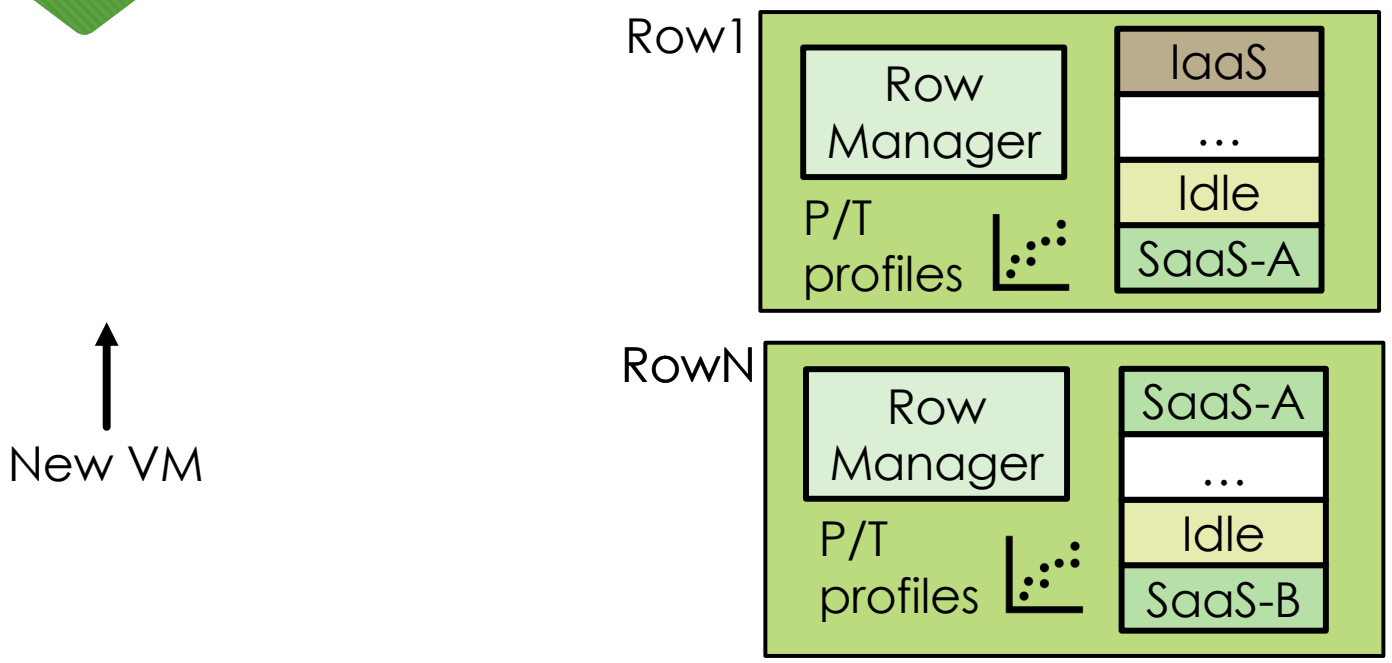
TAPAS: VM Placement



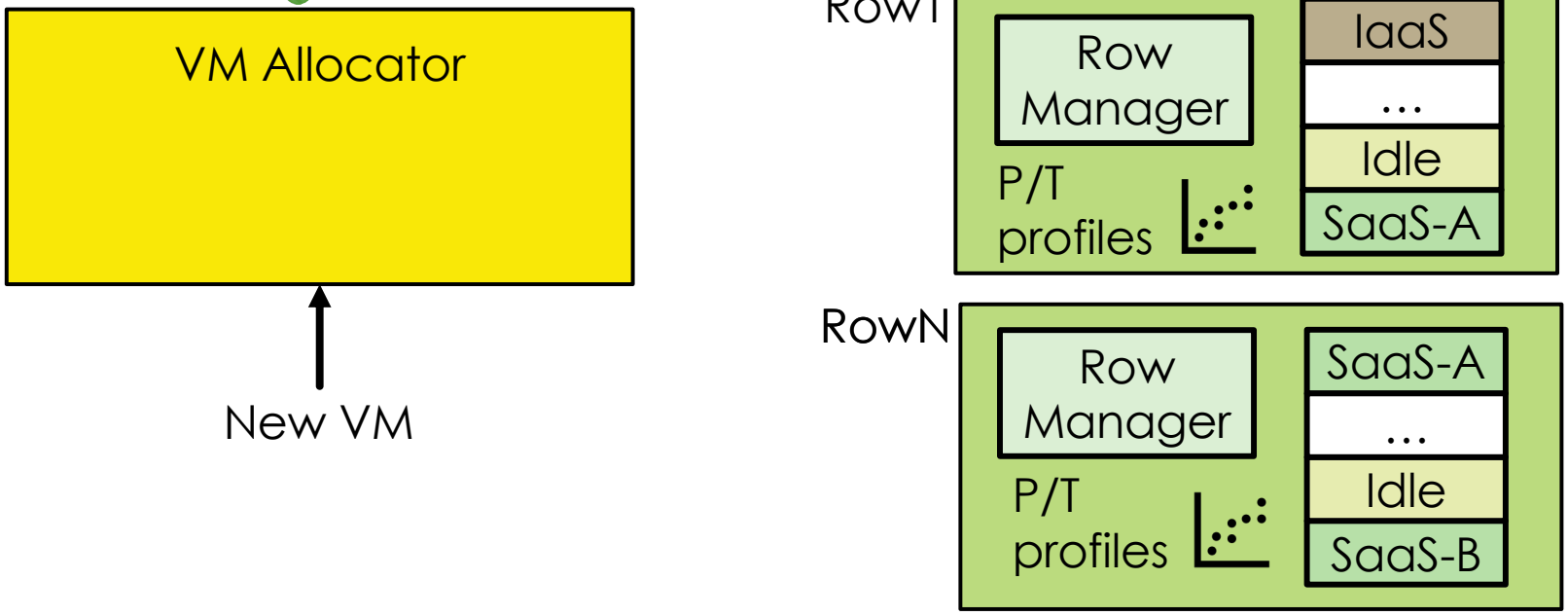
TAPAS: VM Placement



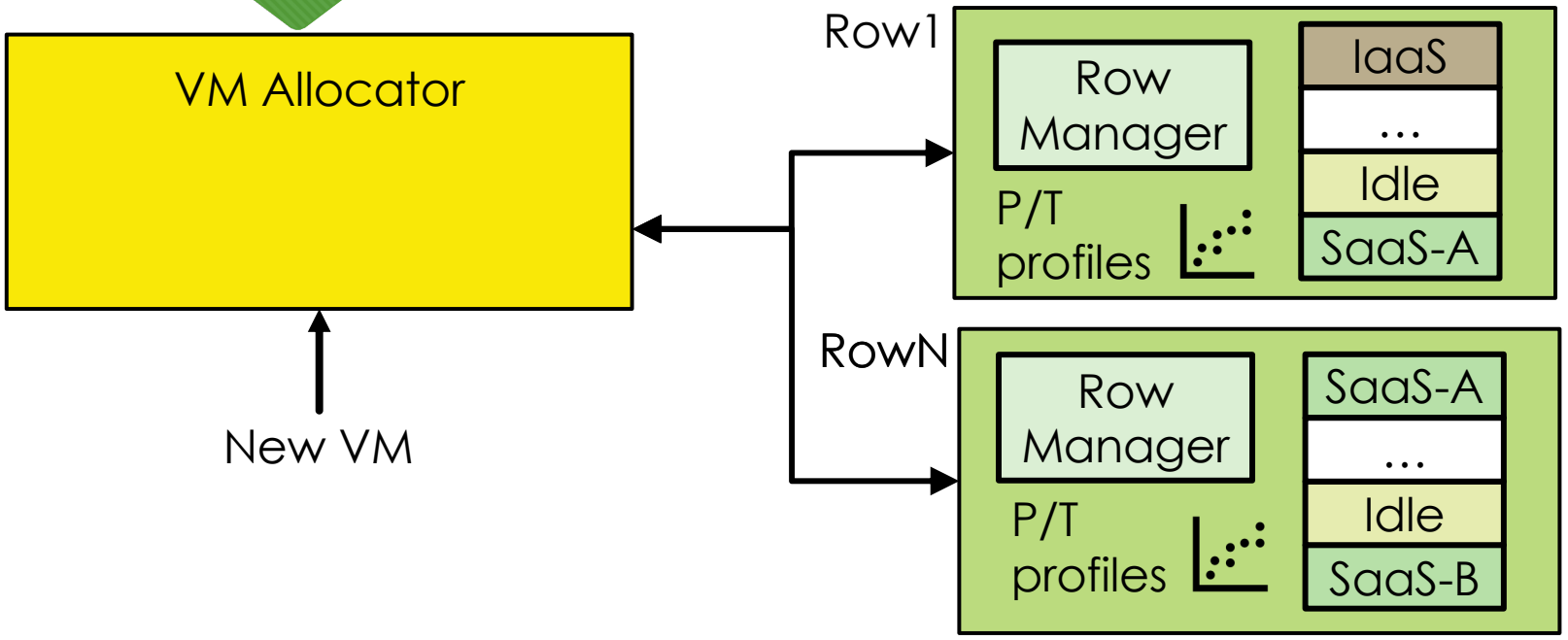
TAPAS: VM Placement



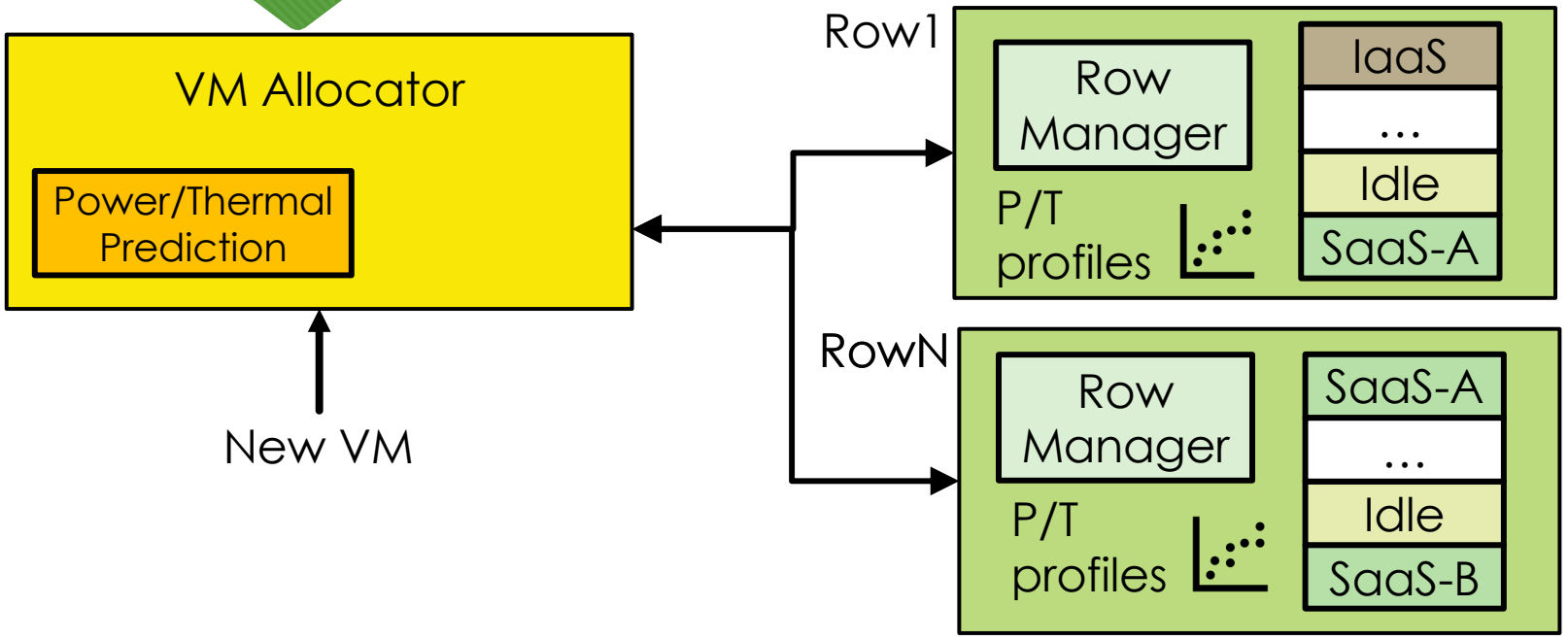
TAPAS: VM Placement



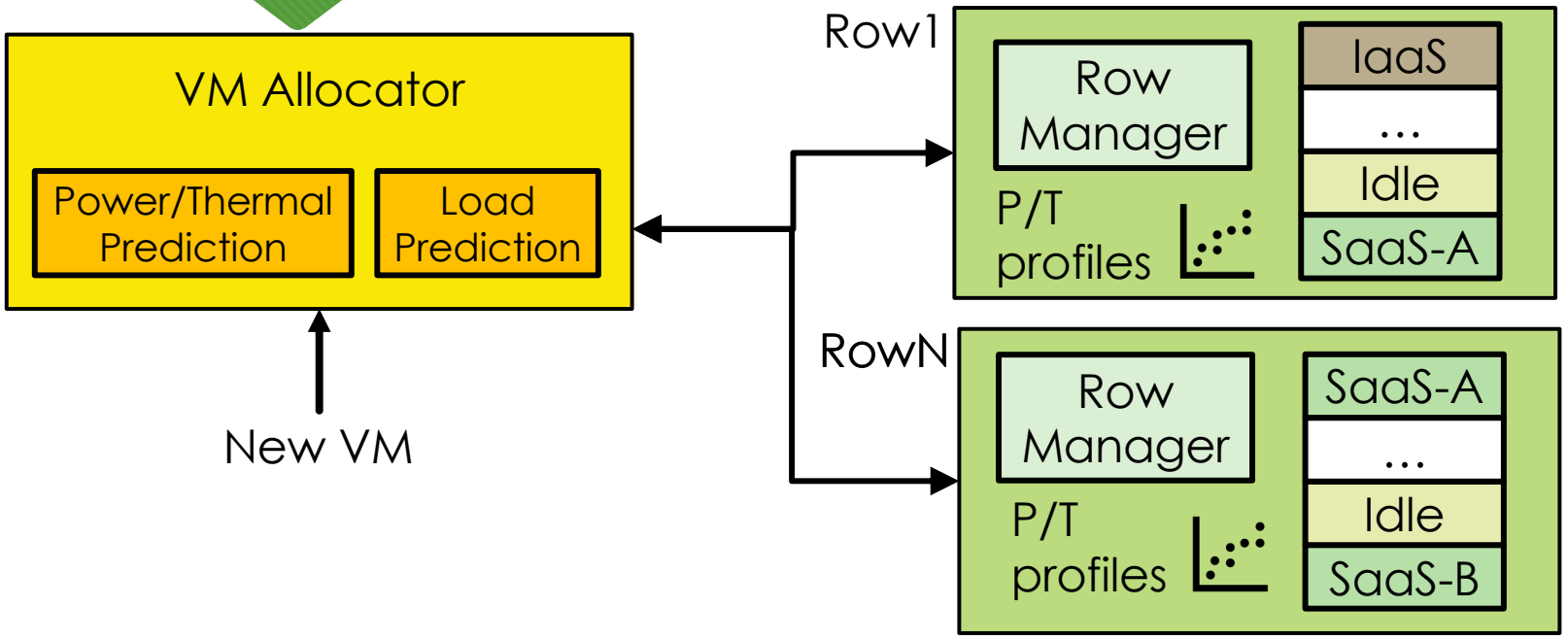
TAPAS: VM Placement



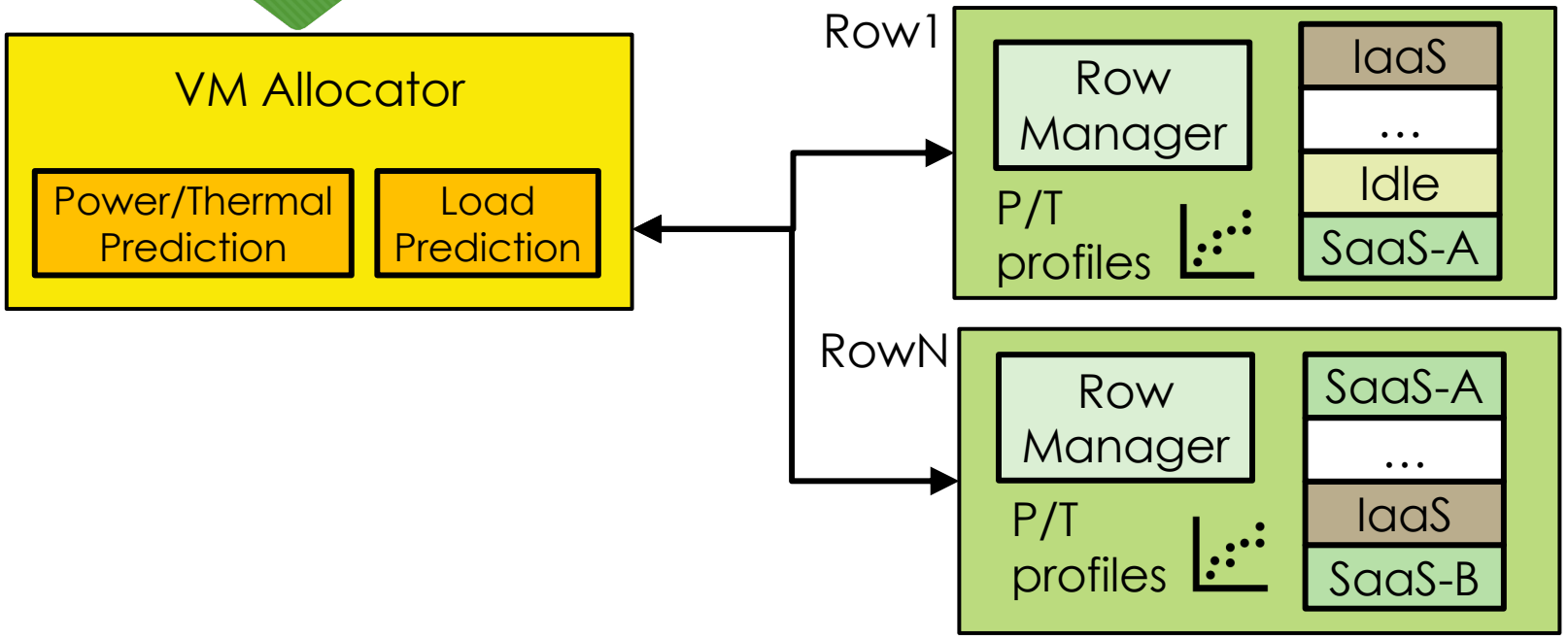
TAPAS: VM Placement



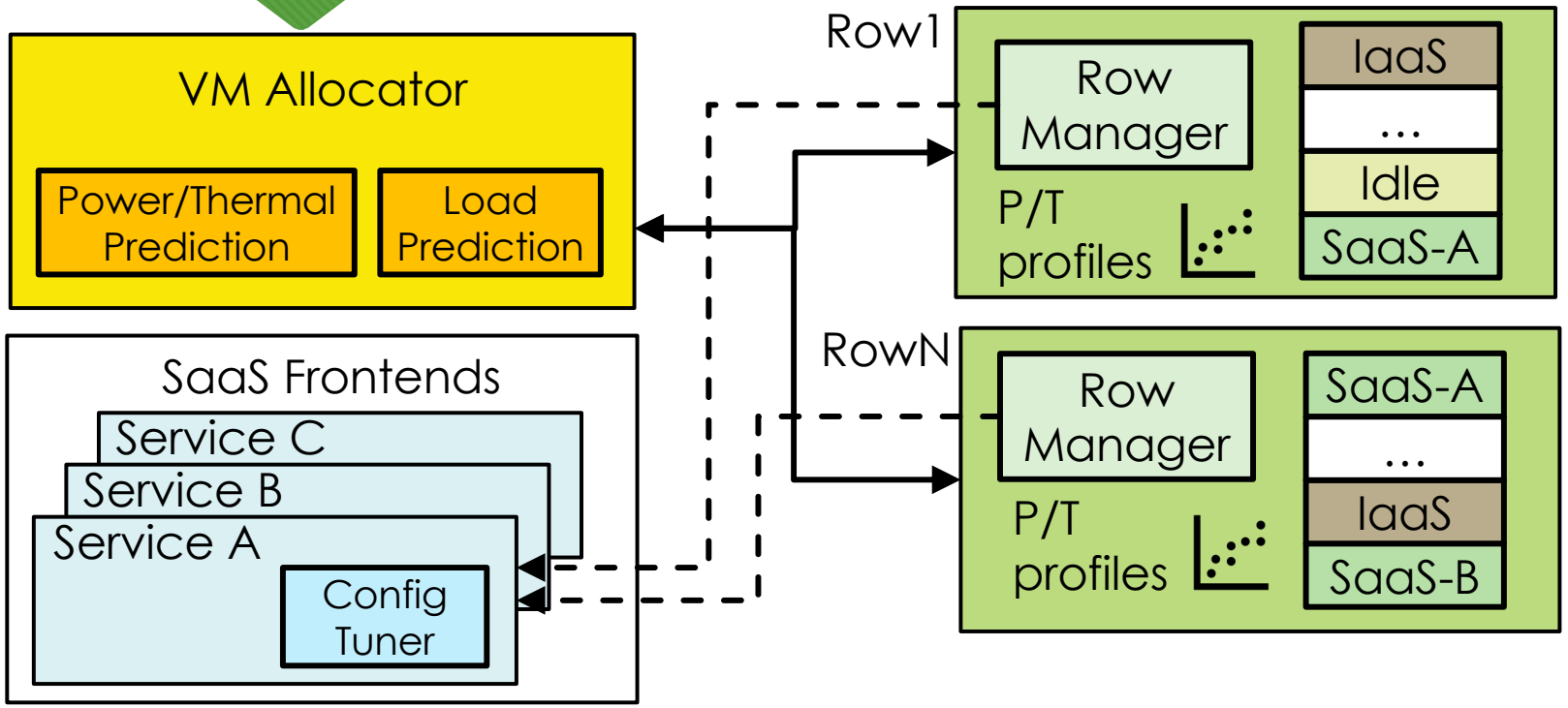
TAPAS: VM Placement



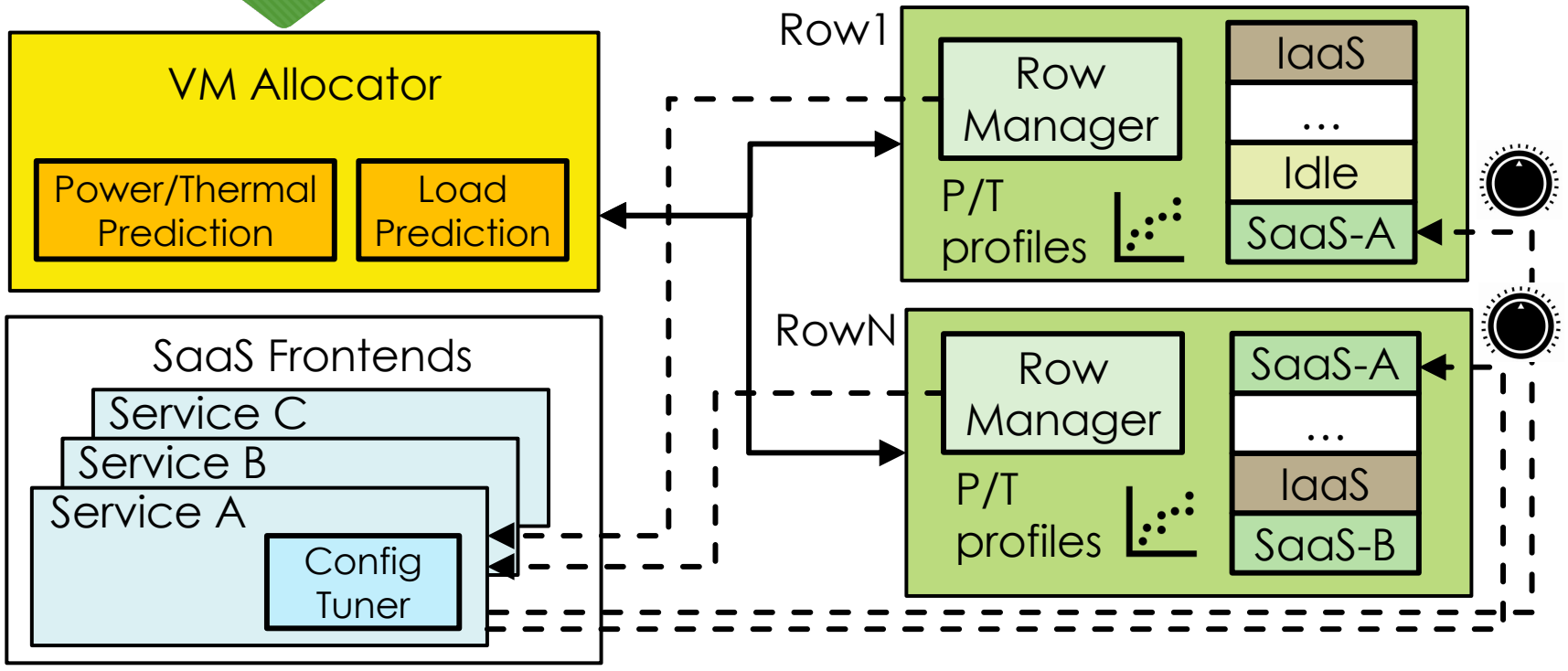
TAPAS: VM Placement



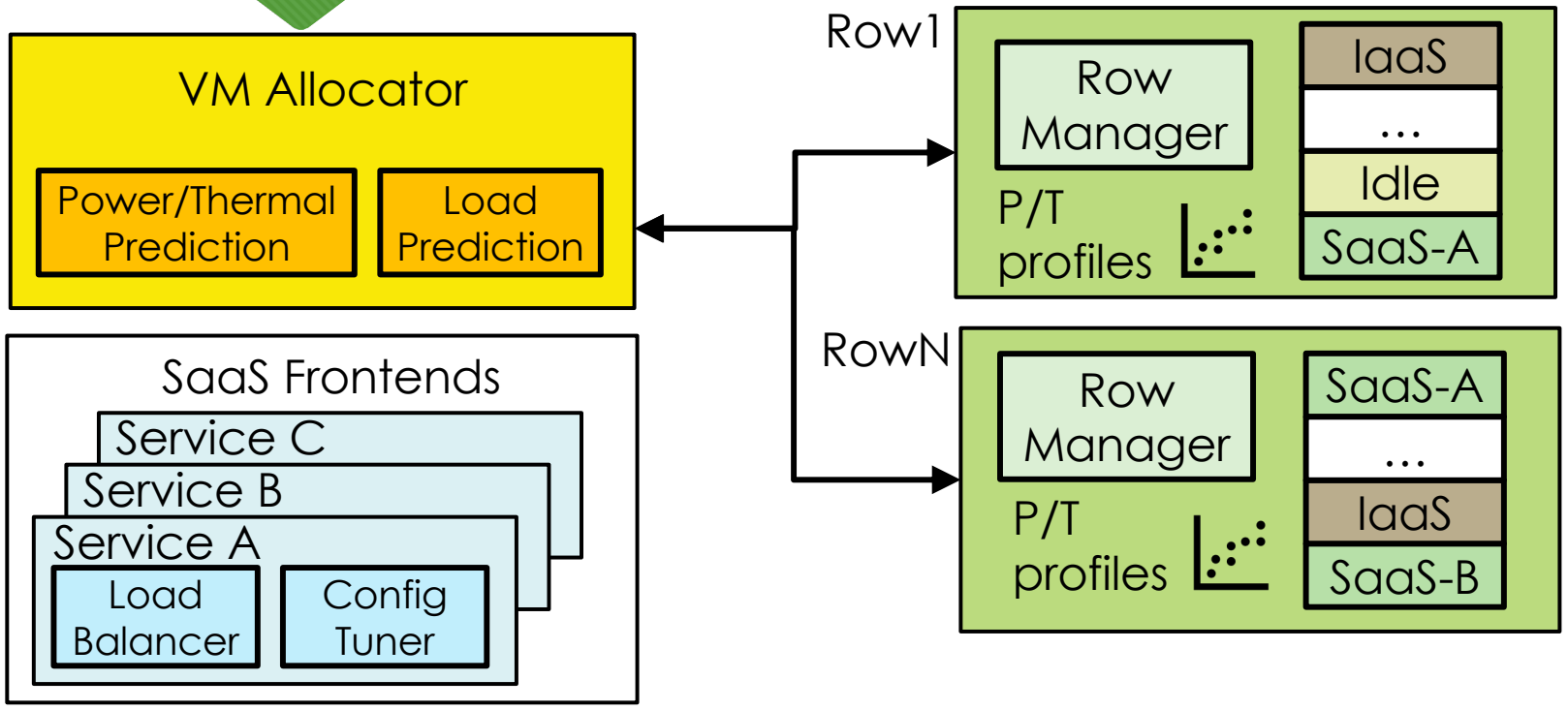
TAPAS: Configuration Tuning



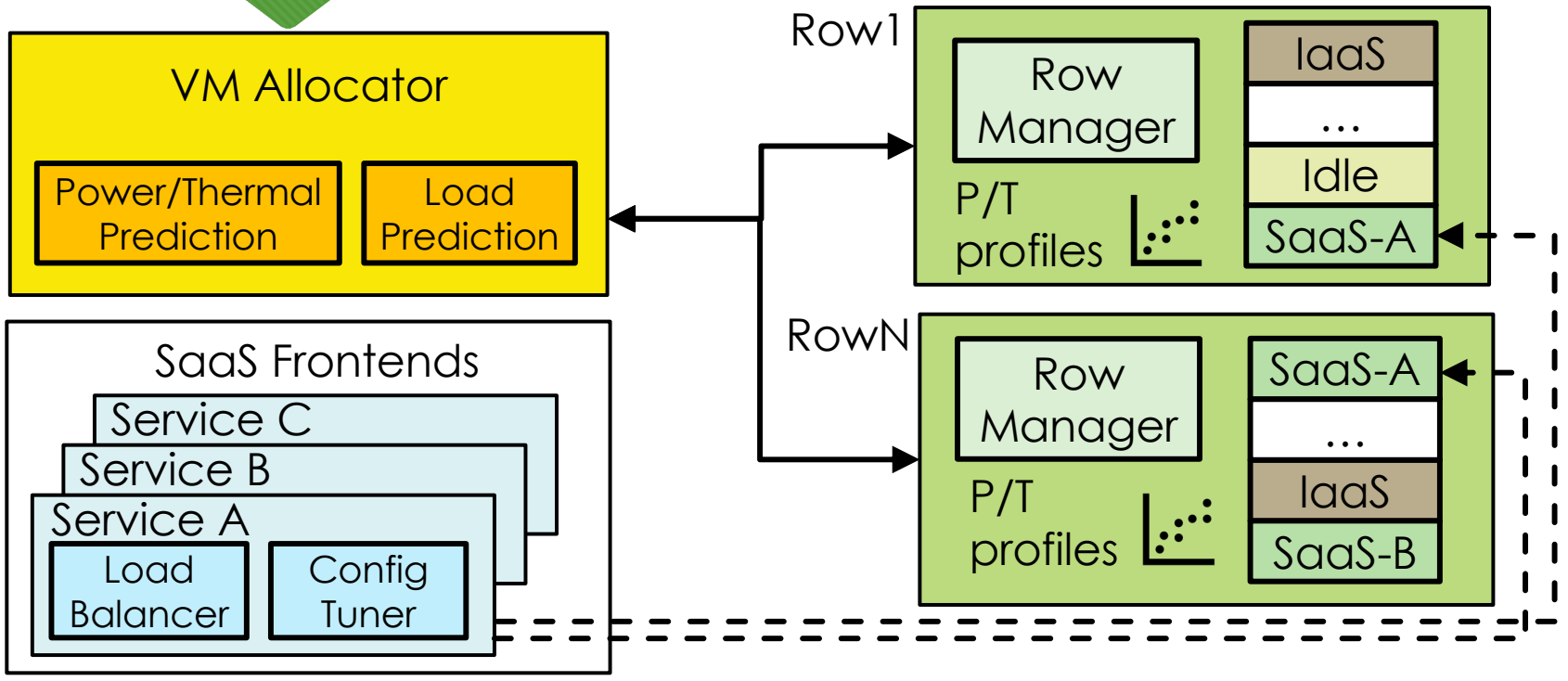
TAPAS: Configuration Tuning



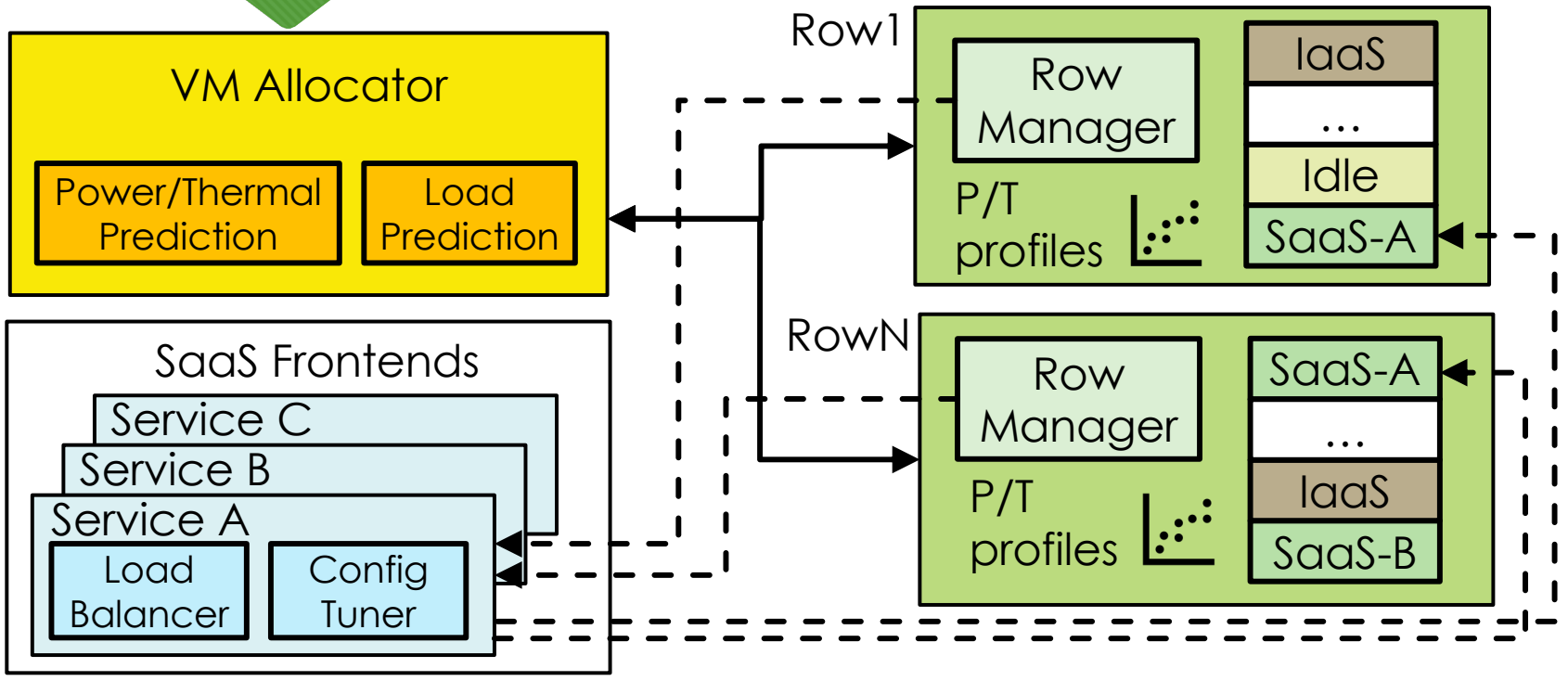
TAPAS: Request Routing



TAPAS: Request Routing



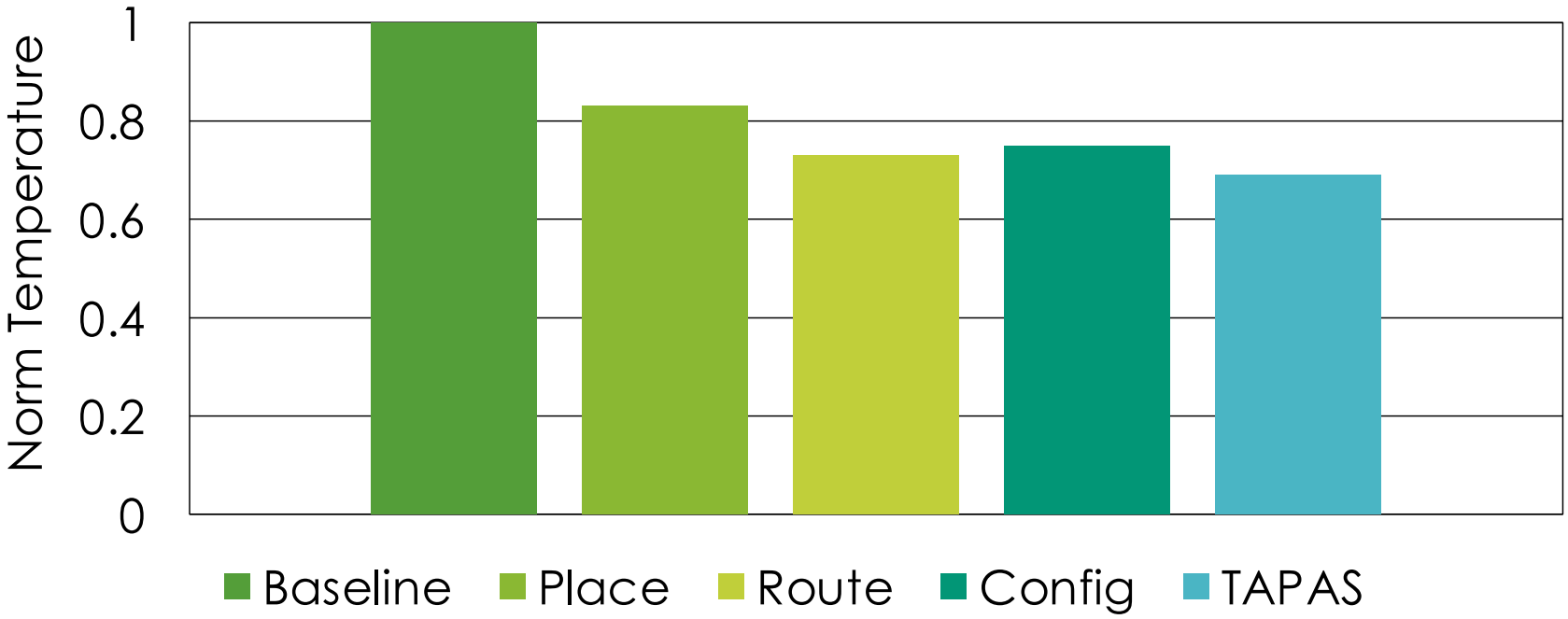
TAPAS: Request Routing



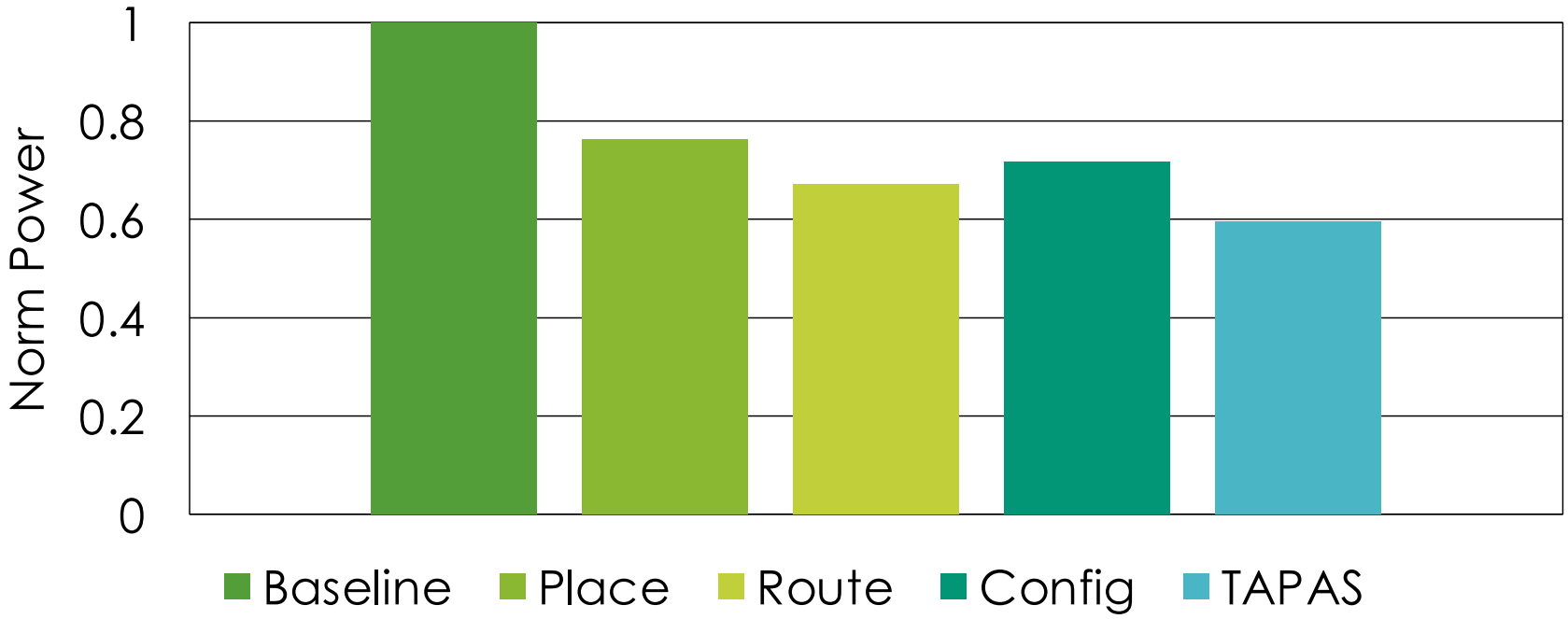
TAPAS Evaluation

- Real cluster experiments
  - Two rows of 80 A100 servers for 1 hour
  - IaaS: direct power and thermal readings
  - SaaS: reply production traces with Llama2 models
- Simulations
  - Week production trace from one of the A100 datacenters
  - Thermal and power modeling using our derived equations

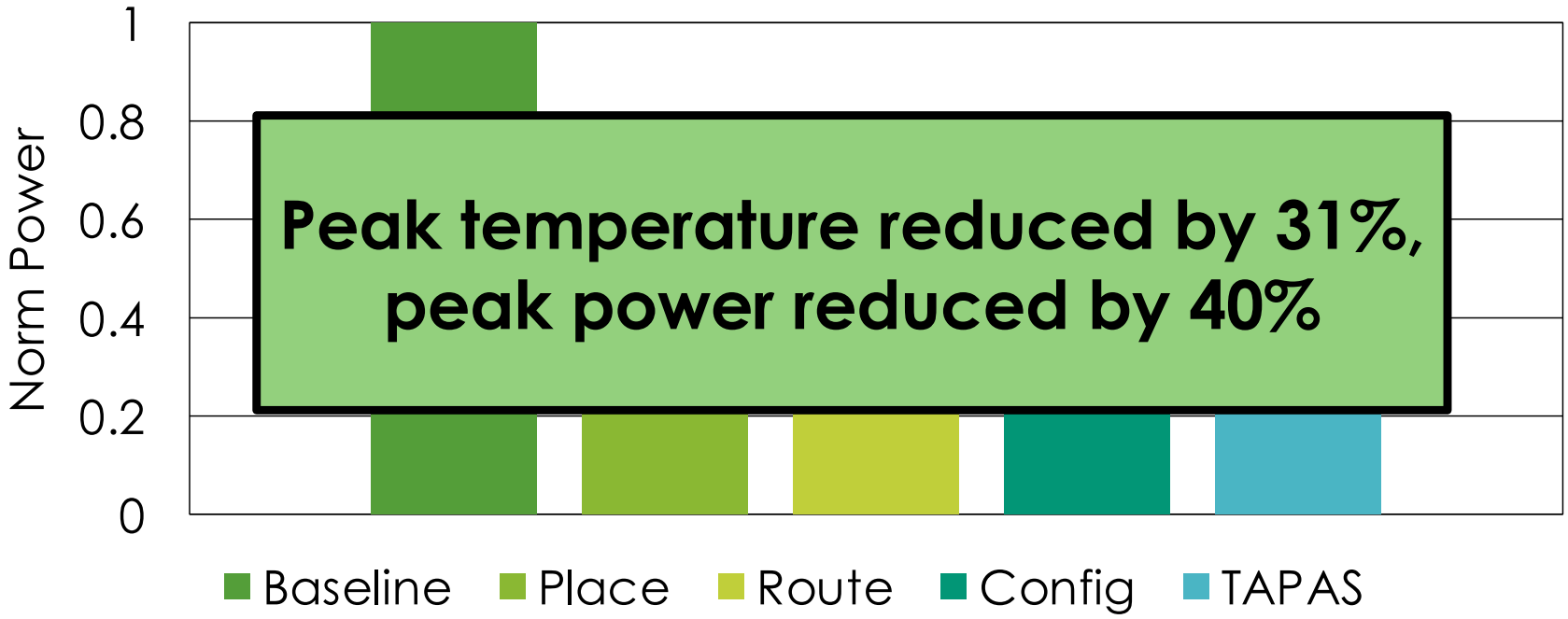
TAPAS Reduces Peak Temperature and Power!



TAPAS Reduces Peak Temperature and Power!



TAPAS Reduces Peak Temperature and Power!



Conclusion

- LLM inference emerging workload in the cloud
  - Its execution thermal- and power-inefficient
- Address these challenges for cost-effective datacenters
- ***TAPAS*** as a step towards our goals!

TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms

ASPLOS 2025

Jovan Stojkovic*, Chaojie Zhang, Íñigo Goiri, Esha Choukse, Haoran Qiu, Rodrigo Fonseca, Josep Torrellas*, Ricardo Bianchini

*University of Illinois at Urbana-Champaign, Azure Research – Systems