RELIABILITY IN MODERN CLOUD SYSTEMS

Summer 2025

LOGISTICS

ASSIGNMENT 4

- Check-In #4 on Monday
- Final Presentations next Wednesday
- Assigned Time Slots for Efficiency:
  - 2:00-2:10: Lukas + Bastien
  - 2:10-2:20: Paritosh
  - 2:20-2:30: Jinhao + Bekhrouz
  - 2:30-2:40: Ali Fahad + Zawayar
  - 2:40-2:50: Aiman + Asim
  - 2:50-3:00: Felix + Marius
  - 3:00-3:10: Talal + Umair

ASSIGNMENT 4 PRESENTATIONS

Suggested Structure (Content: 25% + Style: 5%)

- ❖ Introduce the project theme
  - ❖ Eg: “How does vertical scaling help systems during system overload?”
- ❖ Explain the specific experiments & blueprint implementation
  - ❖ Eg: “We overload system A and then configure vertical scaling”
- ❖ Present results from the experiments
- ❖ Present final insights

Q/A (10%)

ASSIGNMENT 4 TECHNICAL SUBMISSION

Technical submission due on Monday 21st July, 2024

How to submit?

- Send the course staff an email with a link to the codebase (ensure we have the right access)
- Alternatively you could send us a zip file
- There should be a readme of how to run your code and your experiments

COURSE SURVEY

Survey Link: https://qualis.uni-saarland.de/eva/?l=157990&p=1uwioz

OPPORTUNITY - CERULEAN

We are developing a new tool that uses LLMs to generate microservice systems

- We would like to use the time taken to complete the assignments in the course
  - We need your consent for using this information
- We are looking for volunteers to test out the system as part of the user study starting next week. 10 euros / hour for 1 hour.

HARDWARE RELIABILITY DISCUSSION

DISCUSSION THEMES

- How to do resource allocation and utilization in the presence of hardware failures?
- When using RAID, when can you consider written data to be persistent?
- What are some ways you could potentially detect Silent Data Corruption failures during normal workload execution?

DISCUSSION THEMES

- What are some ways you could potentially detect Silent Data Corruption failures during normal workload execution?

DISCUSSION THEMES

- What are some ways you could potentially detect Silent Data Corruption failures during normal workload execution?

You would need to at least triple the amount of work and computation to be able to detect if something went wrong

DISCUSSION THEMES

- When using RAID, when can you consider written data to be persistent?

DISCUSSION THEMES

- When using RAID, when can you consider written data to be persistent?

Must be written to all replicas with all the parity blocks be written correctly. Even then the Power Management Unit (PMU) can destroy the drive.

DISCUSSION THEMES

- How to do resource allocation and utilization in the presence of hardware failures?

DISCUSSION THEMES

- How to do resource allocation and utilization in the presence of hardware failures?

Need some buffer capacity that can be used to deal with issues or unavailability

PAPER SUMMARY

PAPER SUMMARY

- Organizer servers in a logical cluster
- RAS uses reservation as an abstraction for capacity
  - Servers assigned to a reservation
- Runs a Multi-Integer Programming Solver to continuously optimize assignments of resources
- Shared buffer to deal with failures
  - Buffer is just extra servers that are available for reservations
  - Pre-allocate buffers into reservation to minimize impact

BUILDING RELIABLE DATA CENTERS

REGIONS AND DATA CENTERS

REGIONS AND DATA CENTERS

A typical full-scale Meta region consists of 5-6 Data Centers

DATA CENTER POWER

DATA CENTER POWER

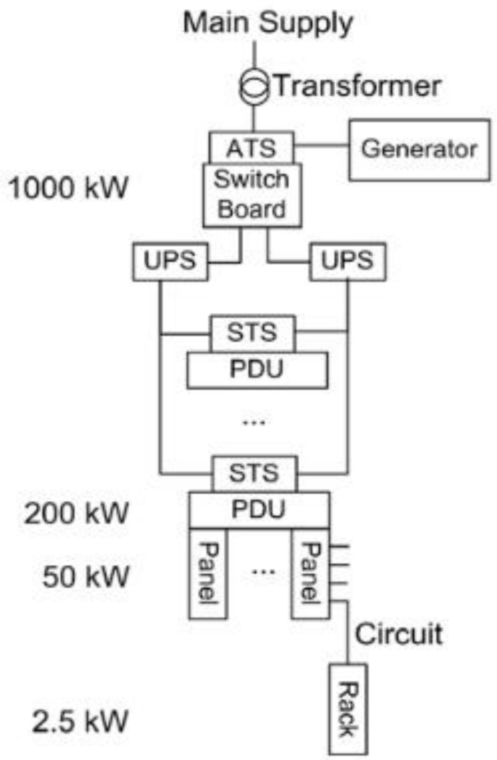
Automatic Transformer Switch (ATS) switches between main supply and generator

Uninterruptible Power Supply (UPS) provides backup until the generators kick in

Power Distribution Units (PDUs) provide monitoring equipment and supply power to panels

Panels power circuits which power racks

Rack (7ft tall, 19 inch wide): multiple servers



: Simplified datacenter power distribution hierarchy.

Power Provisioning for a Warehouse-sized Computer, ISCA'07, Fan et al

DATA CENTER POWER AT META

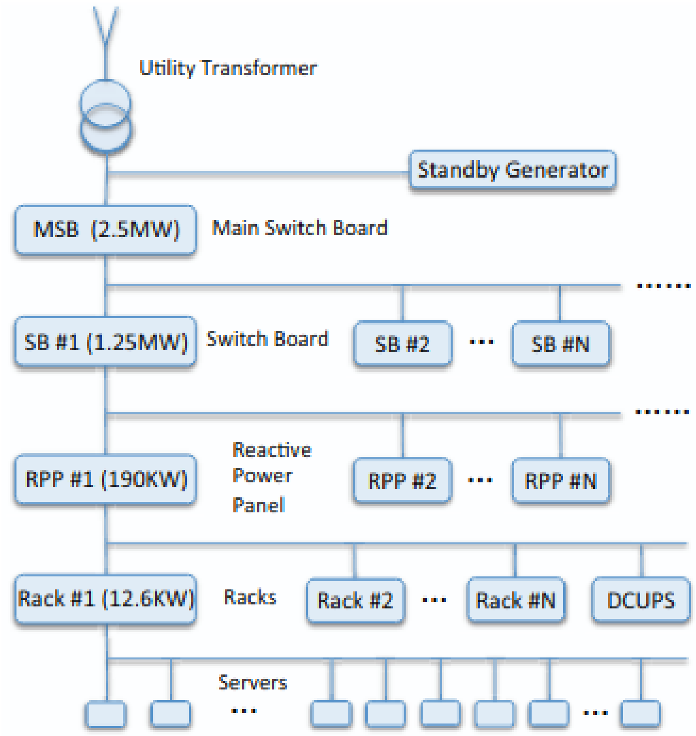
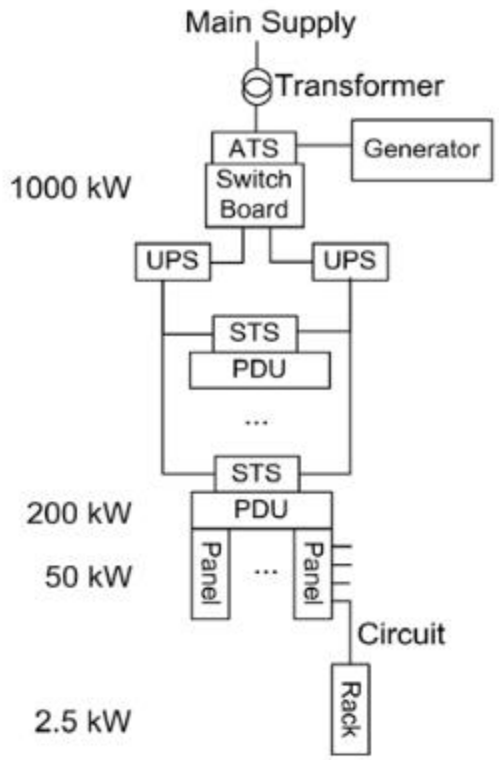


Figure 2. Typical Facebook data center power delivery infrastructure [14].



: **Simplified datacenter power distribution hierarchy.**

Dynamo: Facebook's Data Center-Wide Power Management System, ISCA'16, Wu et al

POWER OVERSUBSCRIPTION

Add more servers and devices than what the max power supply can handle

Not all devices will peak at the same time

Use the same amount of power to host more devices and capacity

Similar to oversubscription for resources like CPU, Memory

POWER OVERSUBSCRIPTION

Add more servers and devices than what the max power supply can handle

Not all devices will peak at the same time

Use the same amount of power to host more devices and capacity

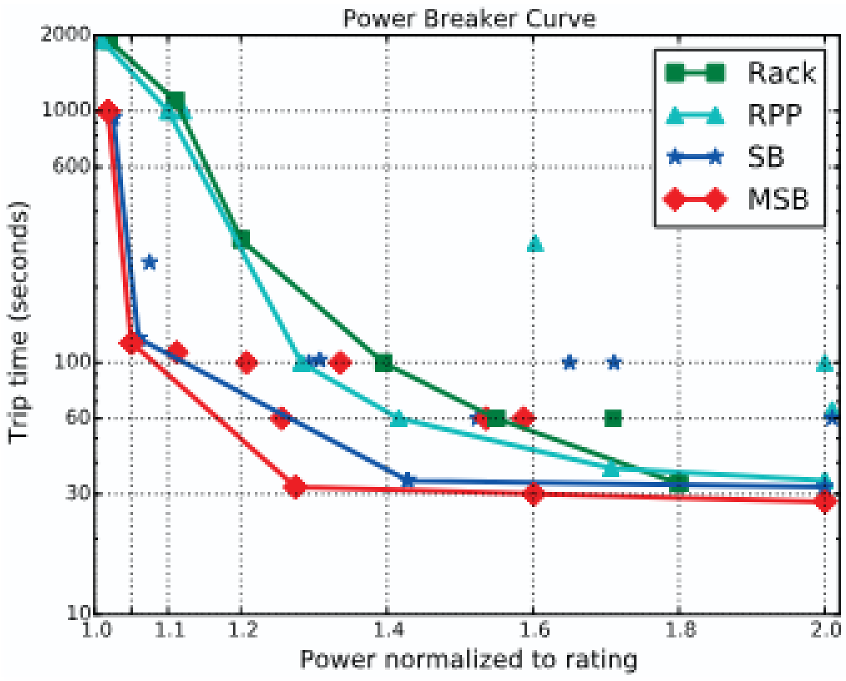
Similar to oversubscription for resources like CPU, Memory

What if the drawn power goes past the max?

CIRCUIT BREAKERS

Circuit Breakers prevent the power draw from exceeding the max power draw that the datacenter was designed for

- Trip under large spikes very quickly
- For small overdraws, it will trip more slowly



Circuit breakers tripping is very bad

DEMAND CURTAILMENT

A **demand curtailment** request is one where the grid provider asks the consumer (a data center in this case) to shed load by a considerable amount—50 to 100 percent—within a defined window of time.

Could be in response to consistently overdrawing more than what was negotiated

Could be due to decrease in power supply for more critical infrastructure

POWER MONITORING - REQUIREMENTS

Sampling Interval: 1-minute

- ❖ High Variation within a 60s window (3% for MSB, 30% for rack)

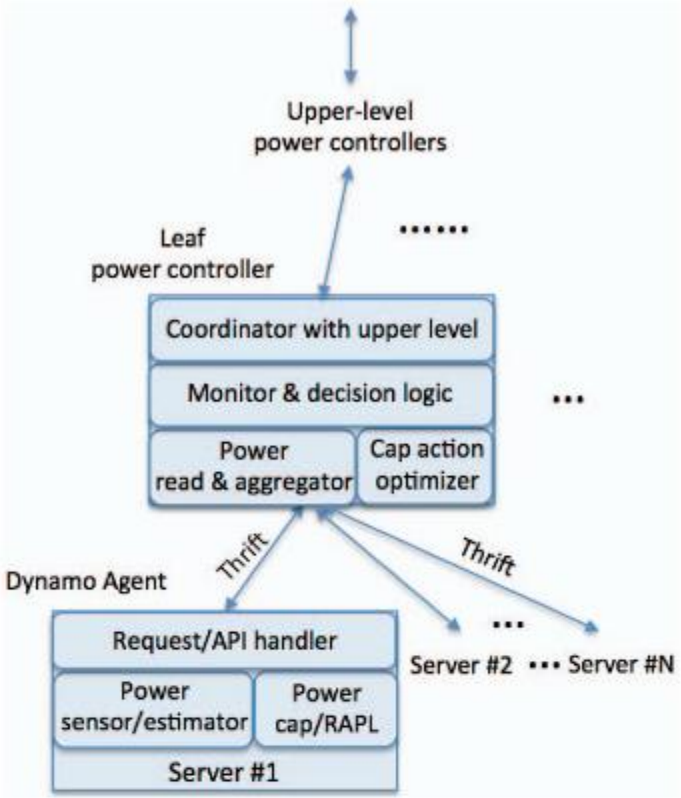
POWER MONITORING

Add an agent on each device

- Provides power measurements
- Performs actions to cap power

Add hierarchical controllers

- Controller collects info and decides specific actions to be taken



POWER CAPPING

Capping Threshold: Max Power that you want a server to use

Capping Target: The safe power usage you want to bring server back to. Acts as the limit until removed.

- Uses RAPL for placing caps

Uncapping Threshold: Power usage at which you remove limits

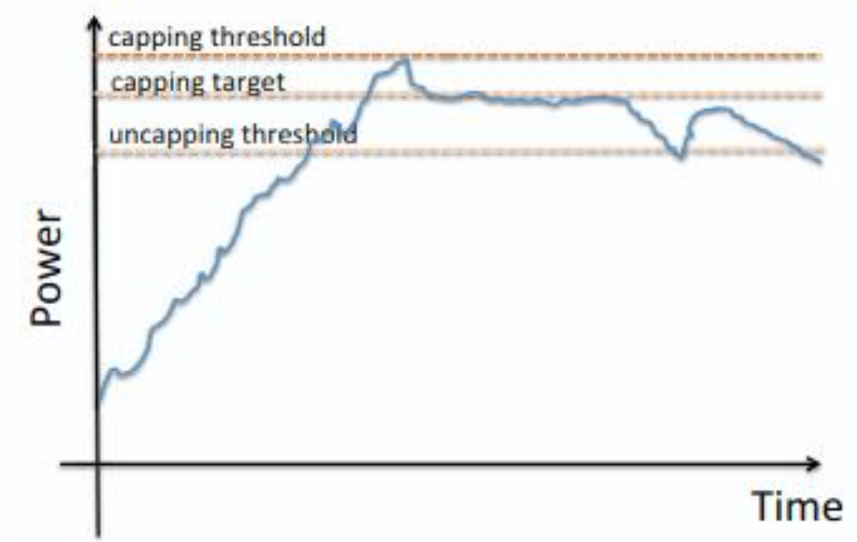


Figure 10. An illustration of the three-band power capping and uncapping algorithm.

UNDERCLOCKING

Decrease the clock frequency

- ❖ Reduces power draw
- ❖ Reduces app performance

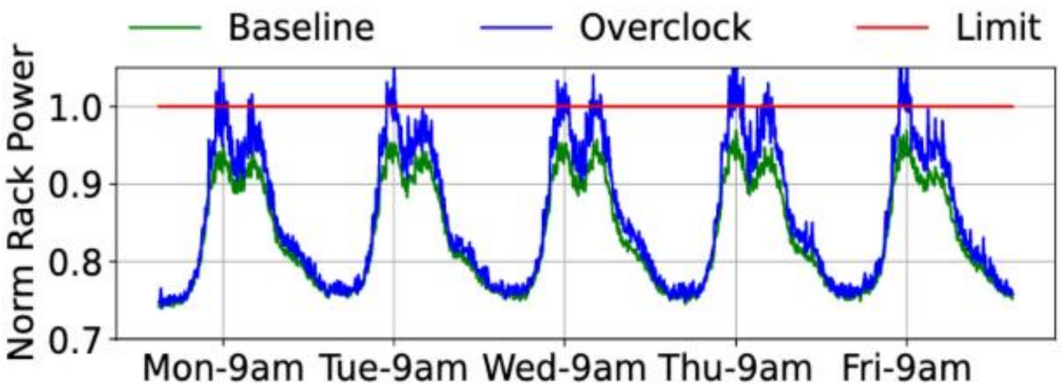
Uses DVFS (Dynamic Voltage and Frequency Scaling)

OVERCLOCKING

Increase the clock frequency to improve the performance

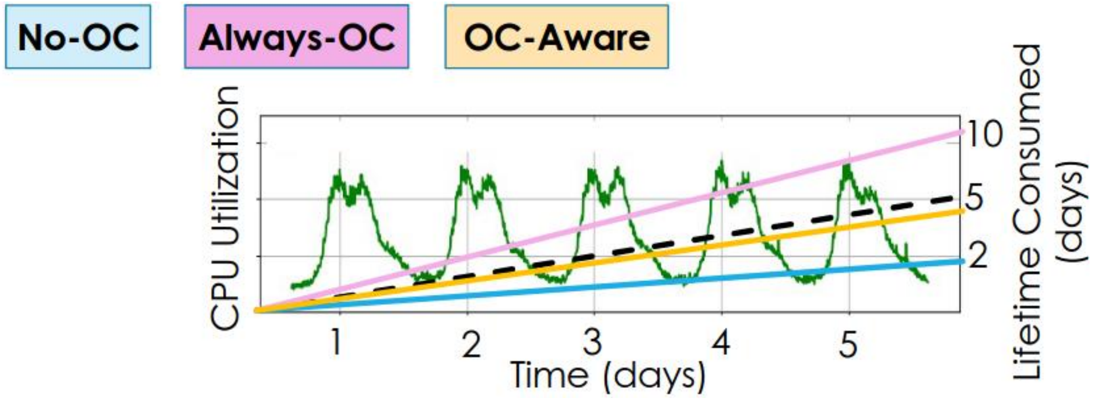
OVERCLOCKING

Increase the clock frequency to improve the performance

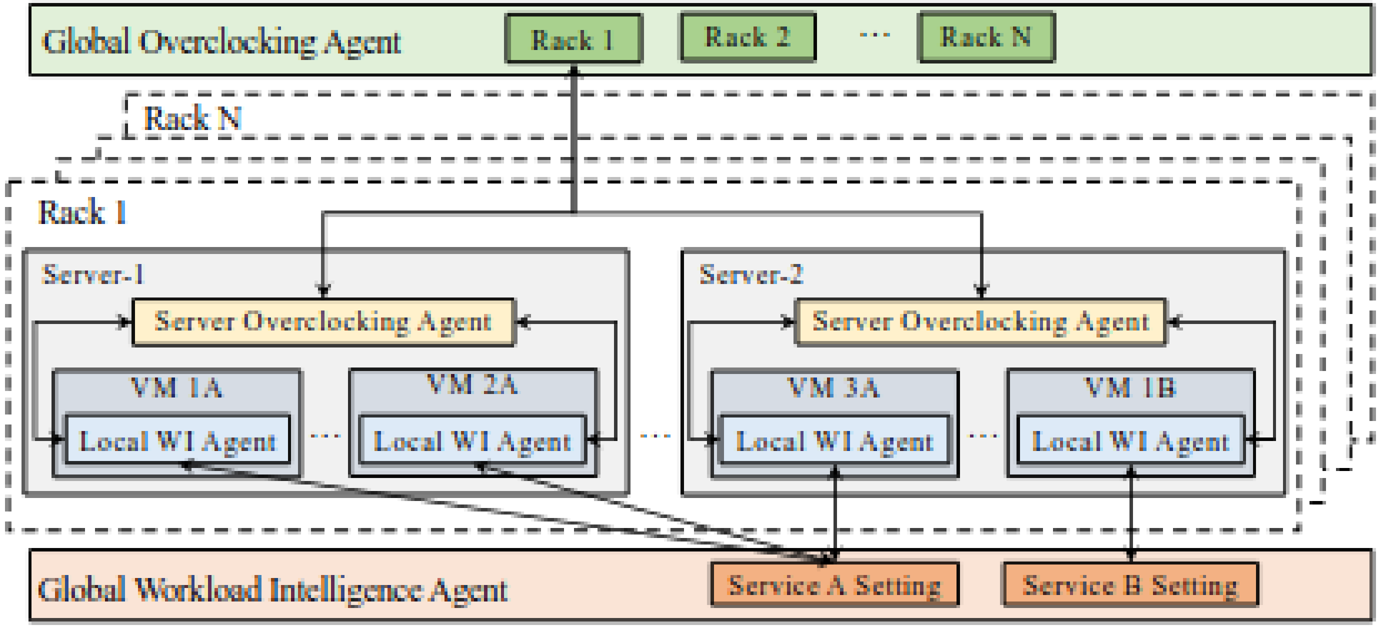


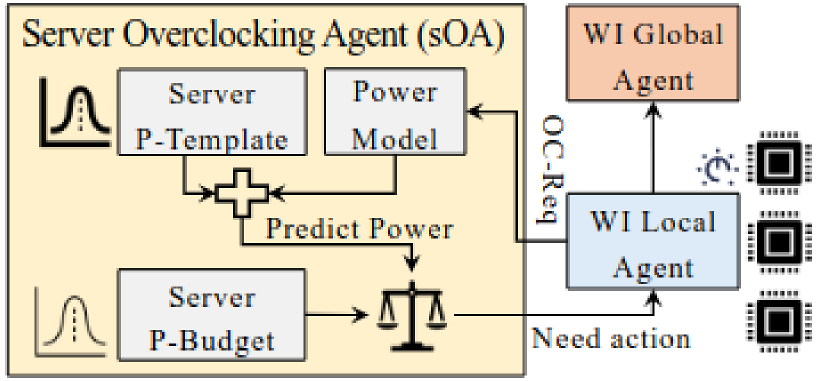
Always overclocking is problematic

- ❖ Can cause power-capping
- ❖ Degrade the lifetime



WORKLOAD-AWARE OVERCLOCKING





SmartOClock: Workload- and Risk-Aware Overclocking in the Cloud, ISCA'24, Stojkovic et al

DATA CENTER COOLING

Datacenter temperature must be maintained properly to avoid overheating

Hot DC can damage equipment and cause data and monetary loss

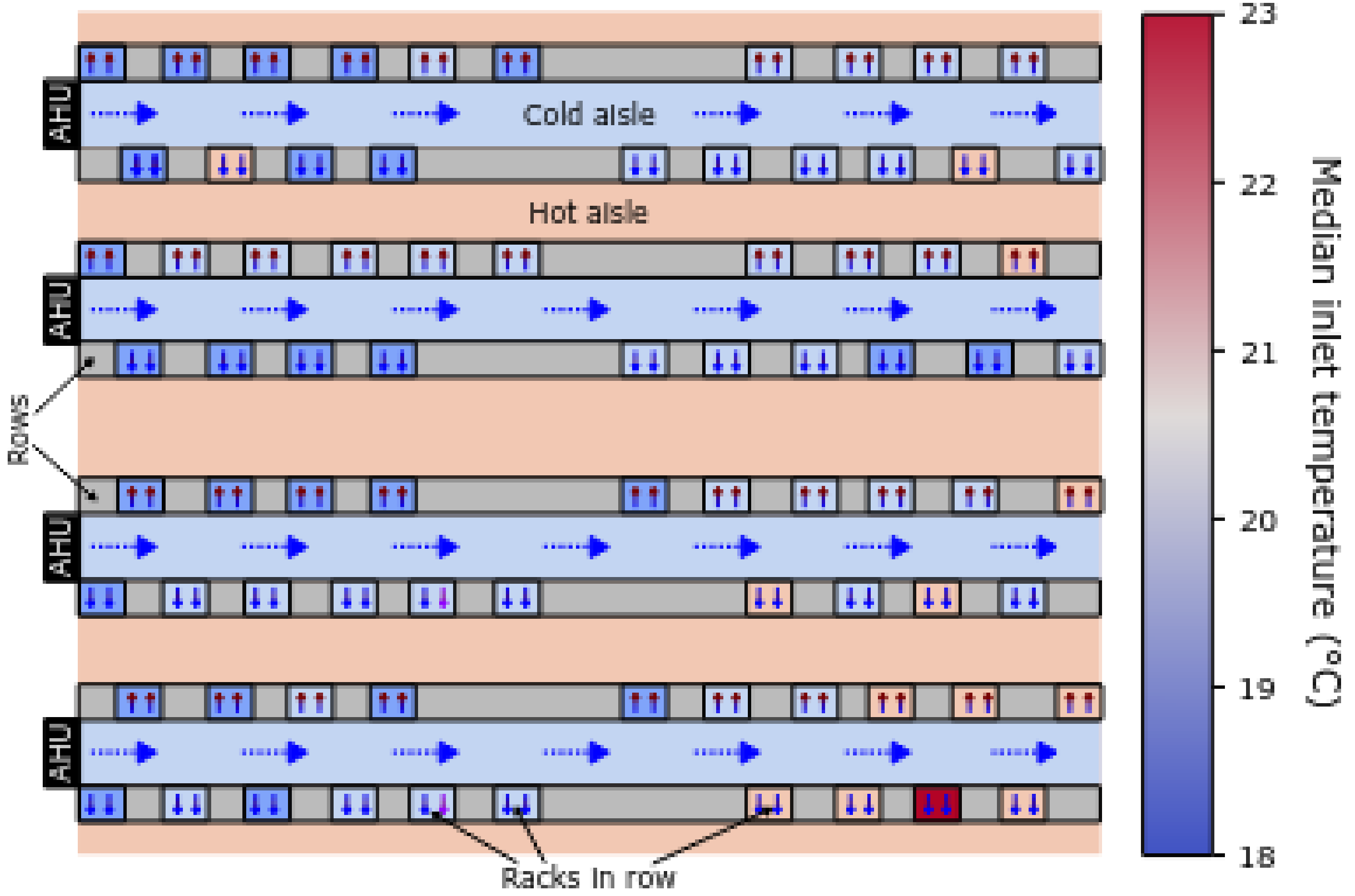
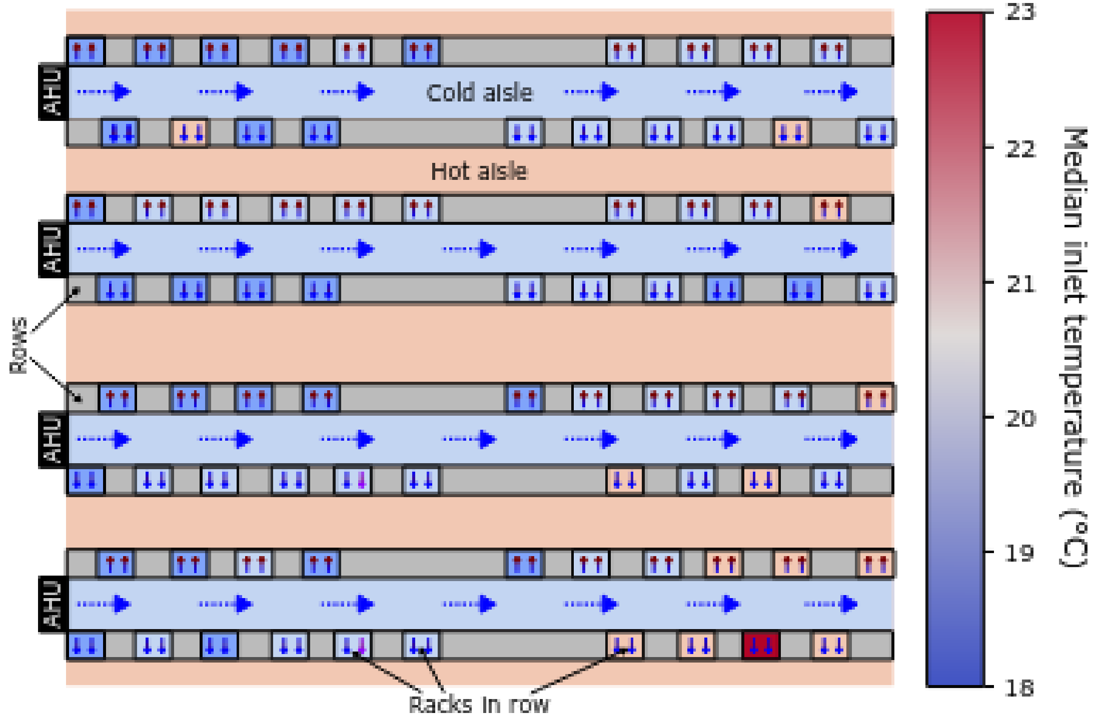


Fig from TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms, ASPLOS'25, Stojkovic et al

AIR-COOLING

Use large fans and cooler outside air

Have to do humidity control

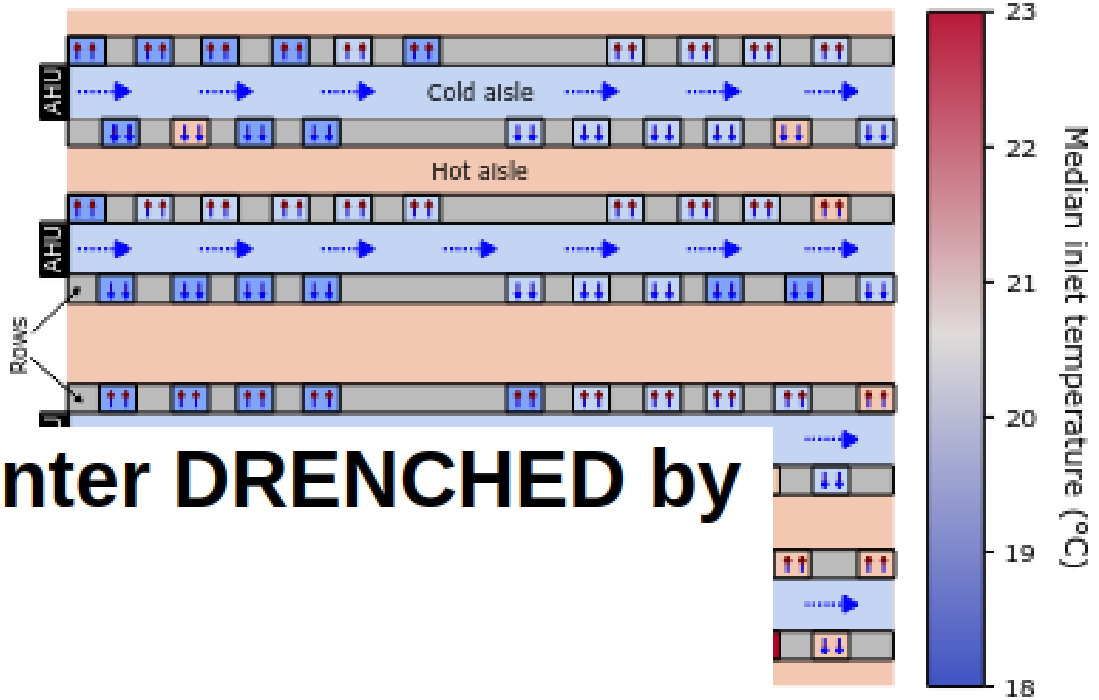


AIR-COOLING

Use large fans and cooler outside air

Have to do humidity control

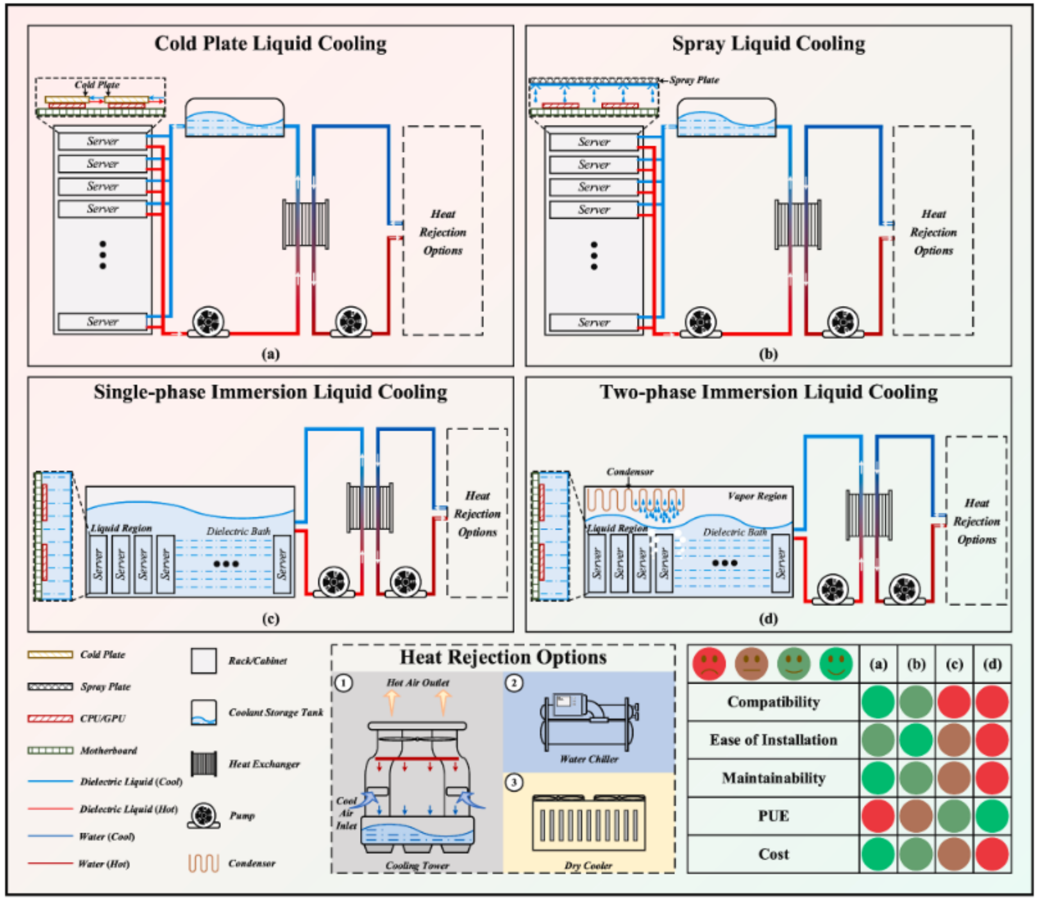
- Otherwise you will have a cloud inside your cloud datacenter



Facebook's first data center DRENCHED by ACTUAL CLOUD

Revealed: Cloud downed by ... cloud!

LIQUID COOLING



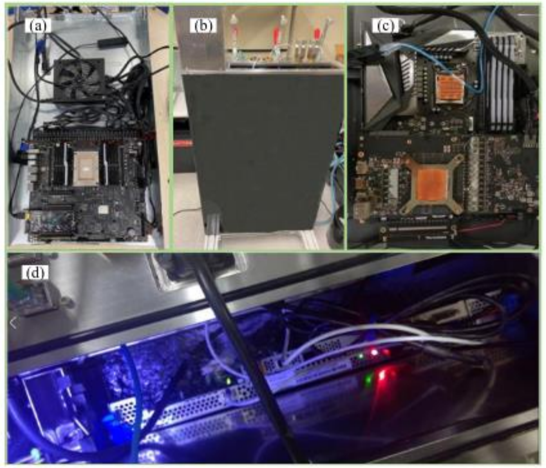
A comprehensive review of cold plate liquid cooling technology for data centers, Chemical Engineering Science Vol. 310, May 2025, Wu et al

2-PHASE LIQUID IMMERSION COOLING

Servers are placed in a liquid with a low boiling point

Heat boils the liquid, liquid evaporates

Condensers on top condense the liquid, causing it to precipitate



Cost-Efficient Overclocking in Immersion-Cooled Datacenters, ISCA'21, Jalili et al

DATA CENTER LOCATION

Power: Power Supply Reliability, Renewable Energy Options, Energy Costs

Climate Considerations: Temperature and Cooling, Natural Disasters, Environmental Regulations

Utility Infrastructure: Water Supply, Waste Management

Security: Physical Security, Political Stability

Source: https://www.landgate.com/news/factors-to-consider-for-data-center-site-selection

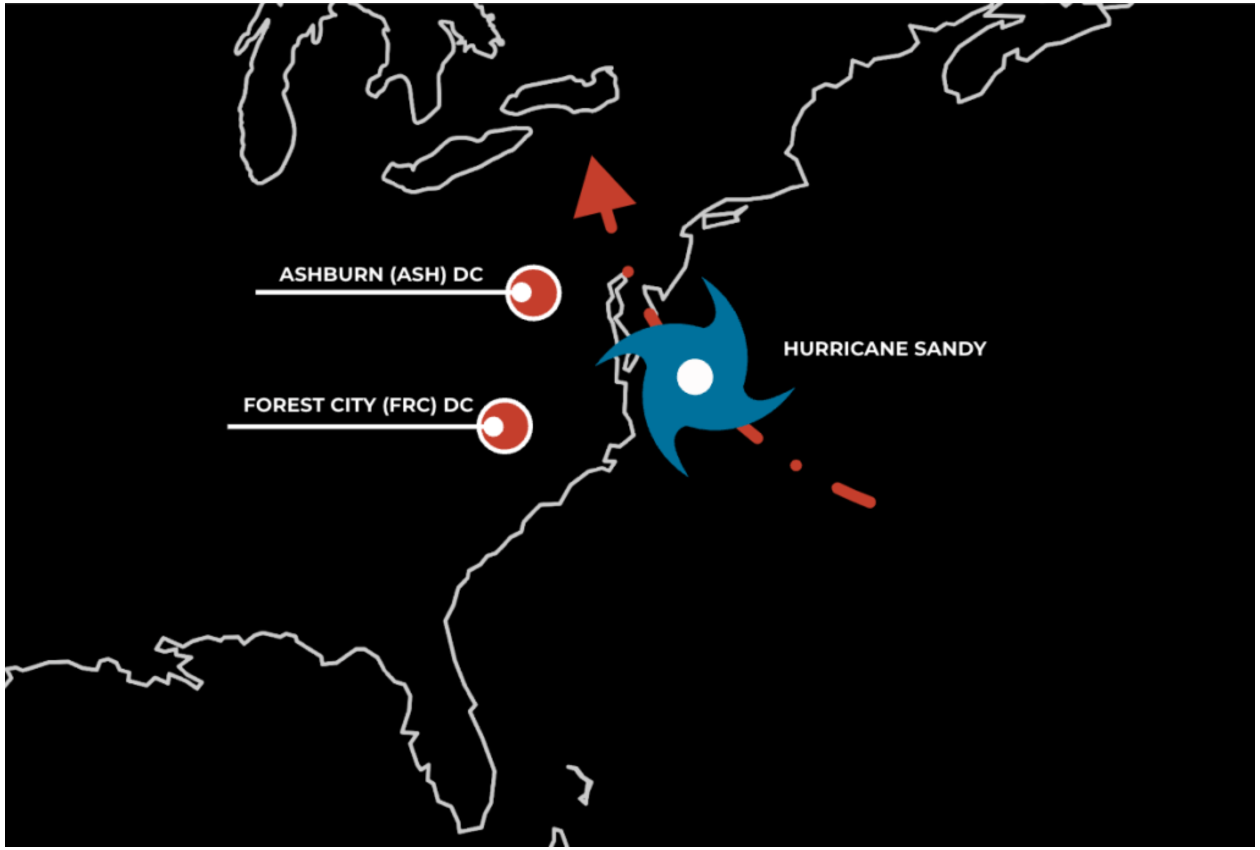
DISASTER AT THE DOOR

In Oct 2012, Meta had 3 DCs

ASH connected Meta to the World

FC had all MySQL primaries

Meta nearly lost all of its data



DISASTER RECOVERY

Disaster Recover (DR) buffer

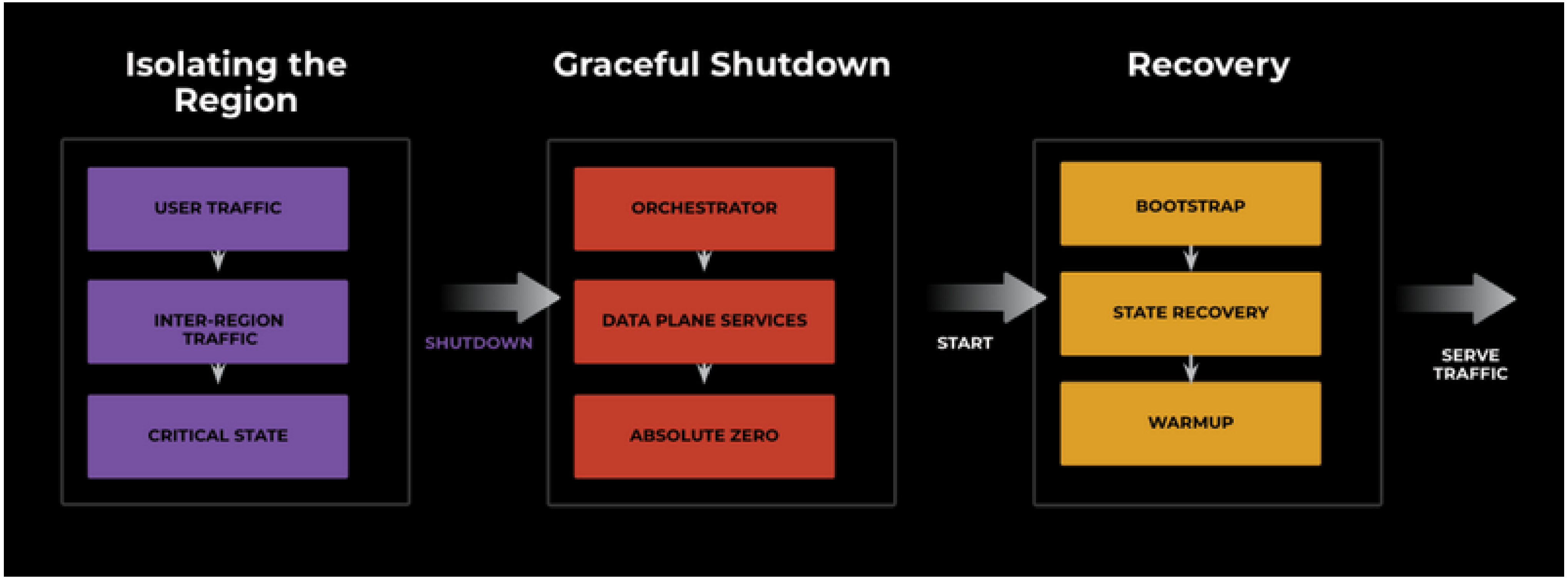
Extra capacity to enable the healthy, running regions to absorb the traffic from the faulty region without overload risks or user impact

DR Storms: Disaster Readiness Exercises

Isolate a production region to validate the end-to-end (E2E) readiness of the DR buffer and service placement.

POWER STORM

DR exercise where a typical production region is brought to a complete stop, transitioned to a powered down state, and then restored fully



FUTURE CONSIDERATIONS

SUSTAINABILITY

LLMs are power hungry applications/workloads

More power means more emissions

Datacenters want to be more sustainable while still providing the same performance guarantees as well as reliability guarantees

Sustainability will influence every efficiency + reliability decision in the future

CARBON EMISSIONS

Measured in CO2-kg (how many CO2 equivalent kgs of greenhouse gases are emitted in a given time period

Operational Emissions

- Scope 1 (direct emissions) + Scope 2 (purchased energy supply)

Embodied Emissions

- Scope 3 (indirect emissions from supply chain, transport, etc)

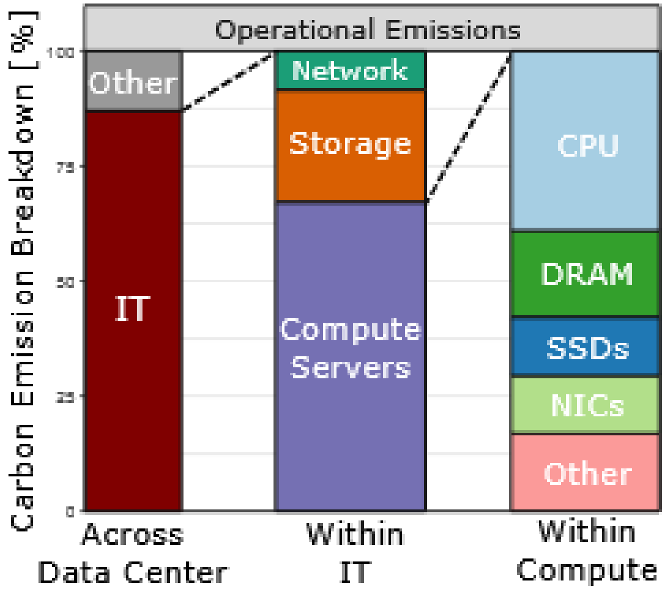
CARBON EMISSIONS OF A DC

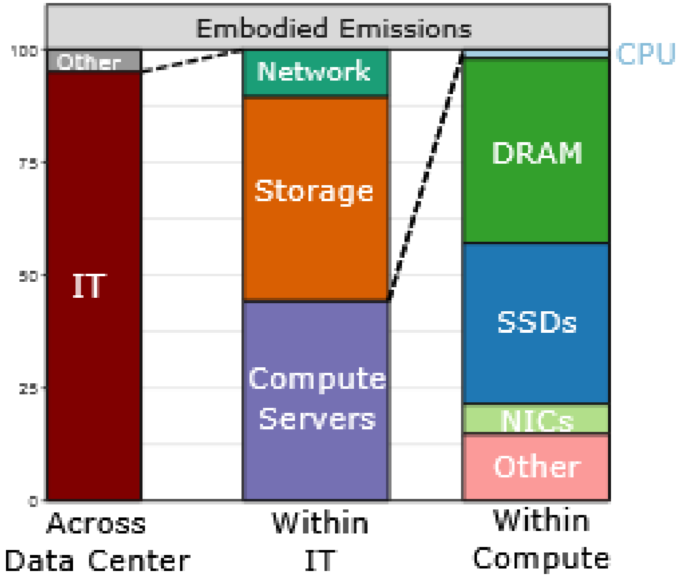
Operational Emissions	CPU	DRAM	SSD	HDD	Other
Compute Rack	42%	18%	19%	0%	21%
SSD Rack	32%	8%	38%	1%	21%
HDD Rack	26%	5%	7%	41%	21%

Table 2: Operational emission breakdown for Azure rack types.

Embodied Emissions	CPU	DRAM	SSD	HDD	Other
Compute Rack	4%	40%	30%	0%	26%
SSD Rack	1%	9%	80%	1%	9%
HDD Rack	2%	11%	14%	41%	33%

Table 3: Embodied emission breakdown for Azure racks.





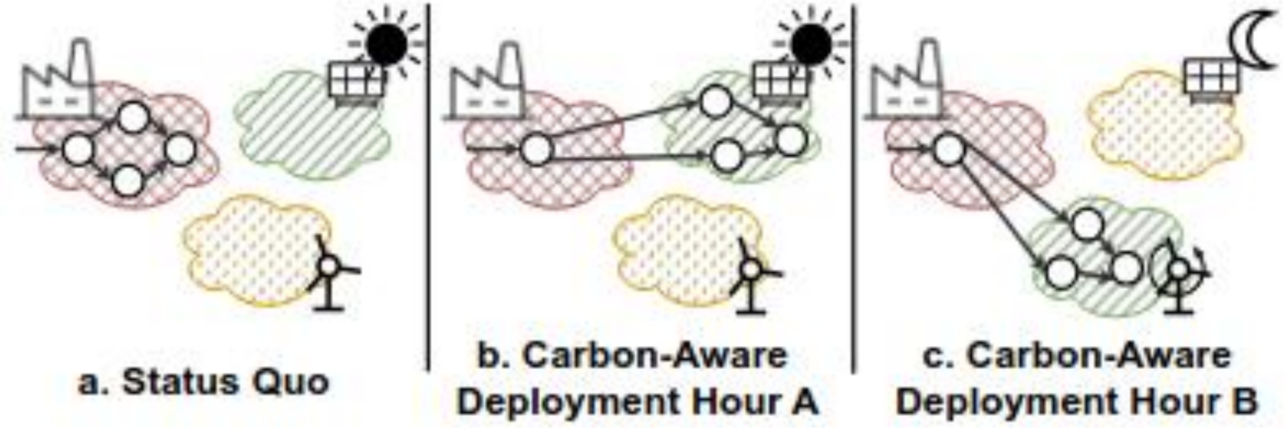
Tables from “A Call for Research on Storage Emissions”, HotCarbon’24, MacAllister et al

Figures from “Designing Cloud Servers for Lower Carbon”, ISCA’24, Wang et al

REGION-AWARE WORKLOAD PLACEMENT

Dynamically change the region in which a workload is hosted

Pick the best region from a clean energy perspective for any given time period



Caribou: Fine-Grained Geospatial Shifting of Serverless Applications for Sustainability, SOSP'24, Gsteiger et al

FIELD TRIP

To the OS group server room