Detecting Data Center Cooling Problems Using a Data-driven Approach

Charley Chen, Guosai Wang, Jiao Sun and Wei Xu

Tsinghua University

Data Center Cooling Problems Are Important

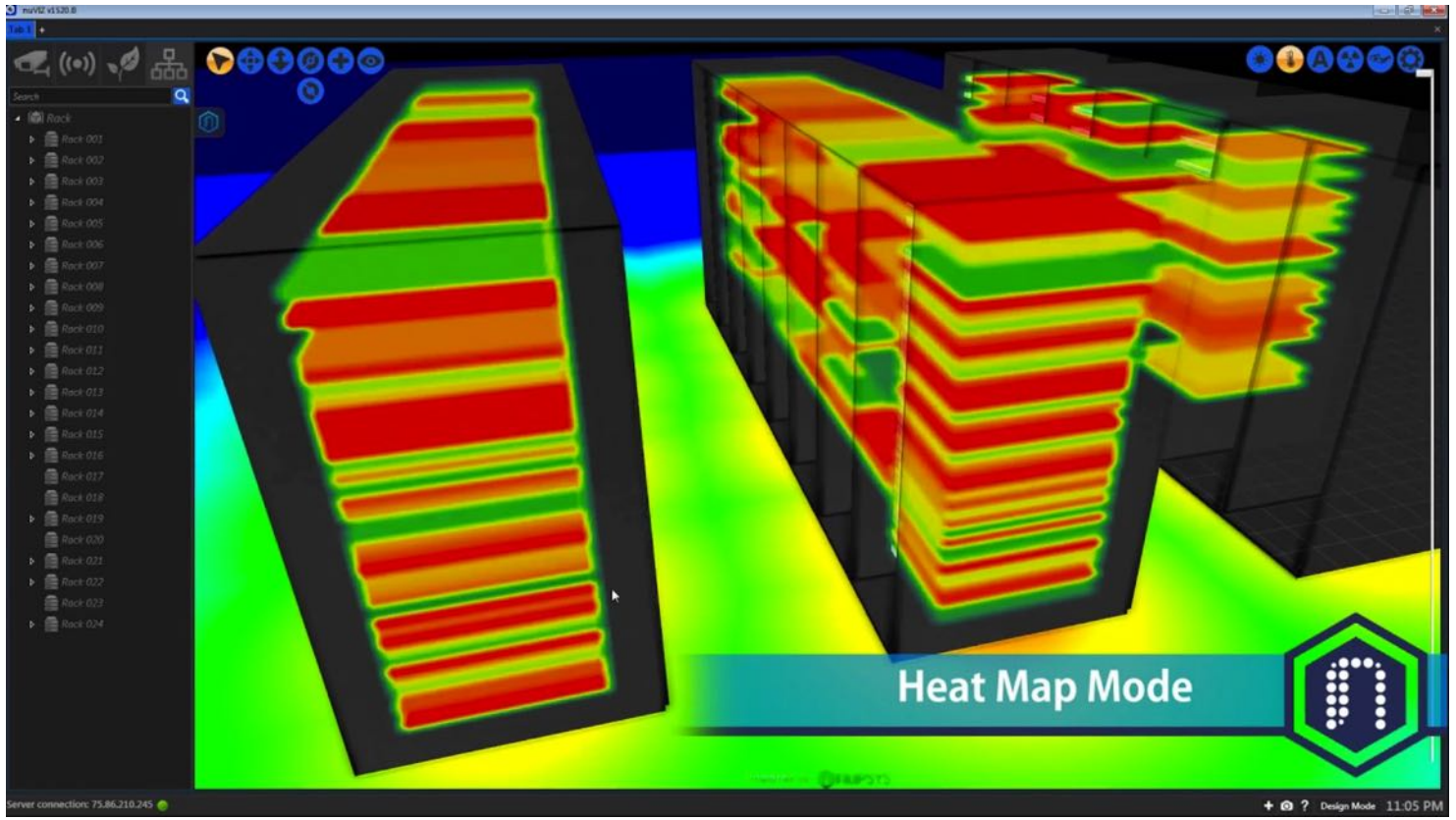
- 32% of the system errors are caused by hardware and cooling problems
- Avoid cooling problem is to reduce the room temperature to ensure a safe margin.
- With the safe margin, servers cooling problem hide anywhere
- High power consumption

"It's hot here, I just need to lower the temperature."

Data Center Cooling Problems Are Important

Servers gets hot anyway when the CPU utilization raise and we cannot say it has cooling problem.

All servers temperature mainly depends on workload, but only with the **overall** workload situation we can detect the hidden cooling problems

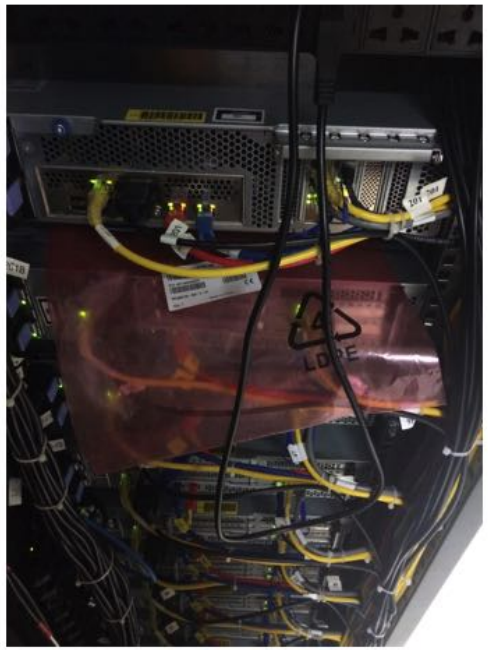


Reference https://www.youtube.com/watch?v=5xLiDYfEQD0

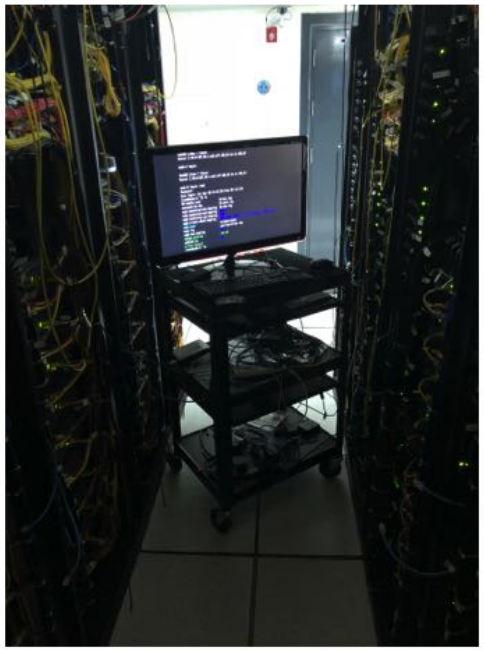
Data Center Cooling Problems

- Transient & Lasting cooling failures

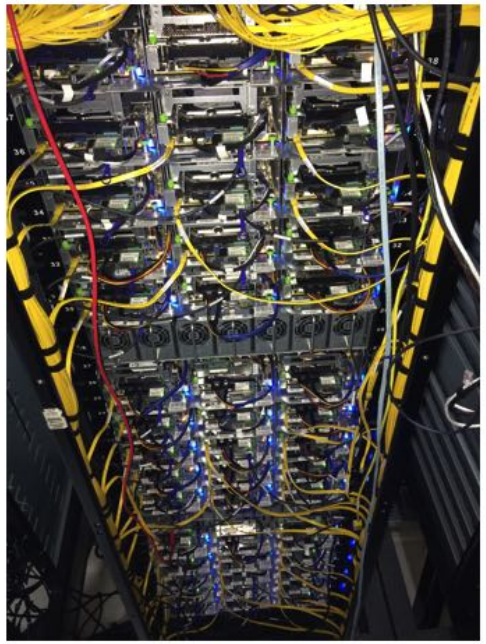
Gap between the tiles



Plastic bag block inlet



Monitor cart forget to remove



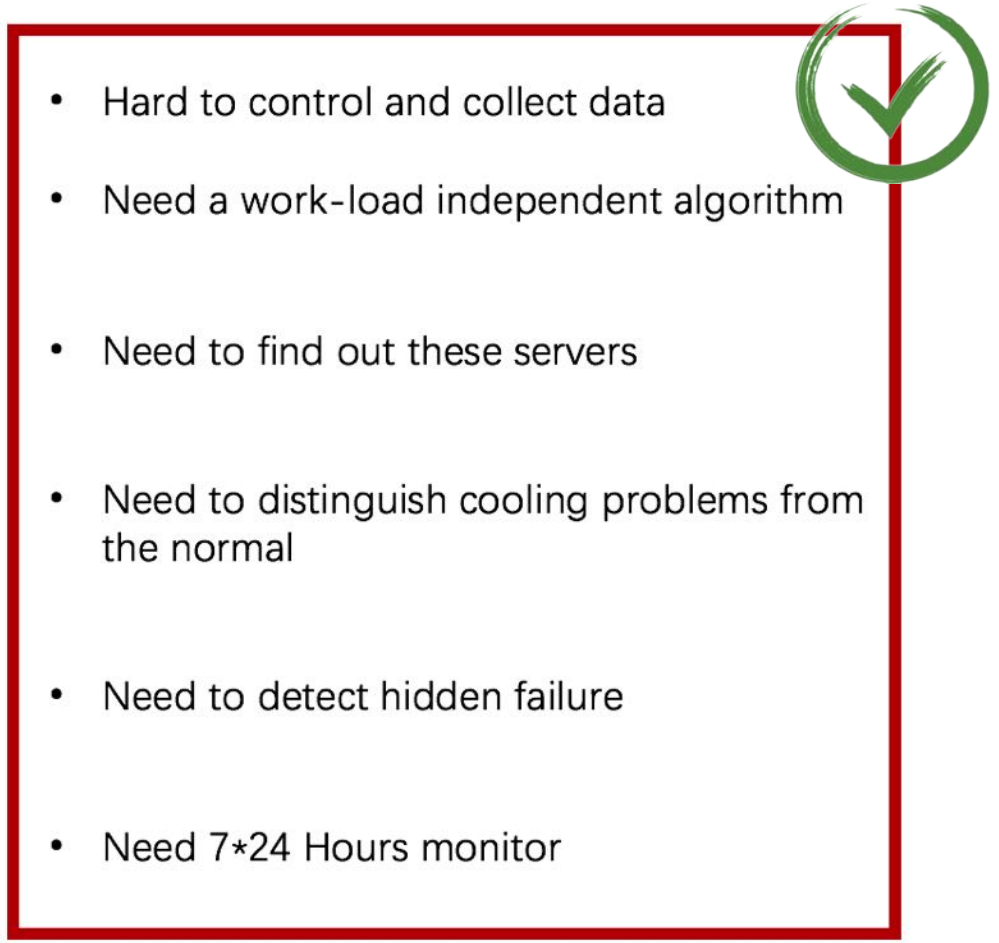
Rack design failure

Data Center Cooling Problems Are Hard to Detect

1. Servers get hot anyways when the CPU utilization increases	• Need to distinguish cooling problems from the normal
2. Servers have a poor cooling behavior to begin with	• Need to find out these servers
3. Operators design layers of hardware, software and operation procedures to tolerate cooling problems.	• Need to detect hidden failure
4. Unexpected situation happens at any moment	• Need 7*24 Hours monitoring
5. Heterogeneous equipment and data centers	• Hard to control and collect data
6. Servers are running tasks and can not stop all job for thermal modeling.	• Need a workload independent algorithm

Contribution

- We propose a novel model called cooling profile to capture the intrinsic cooling behavior of a server that is independent of current workload.
- We design a machine-learning based approach to detect both transient and lasting cooling problems.
- We applied our approach in three distinct data centers and found many real world cooling problems.



Previous Work with Thermal Modeling

- Researchers have used *Computational Fluid Dynamics (CFD)* to model airflow and heat transfer

Need special knowledge of physics and implement sensor

- Researchers have implemented neural networks optimizing the power utilization efficiency

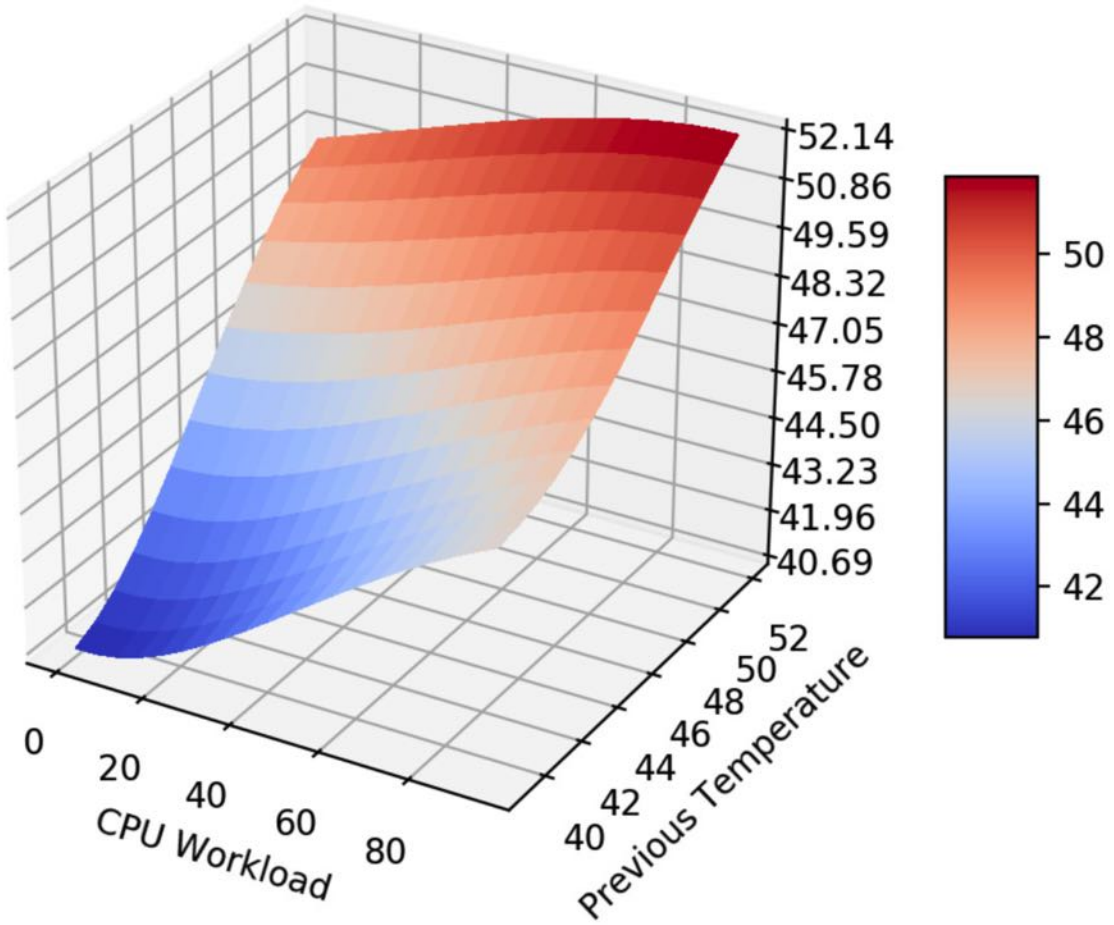
Tools to avoid the hidden cooling problem not to fix it

- Job placement and scheduling with in the data center to help both thermal and power control.

Build Up Cooling Profile

$$\Phi : (\mathrm{T_0}, \mathrm{W},) \rightarrow \mathrm{T} \quad \longrightarrow$$

$\boldsymbol{T_0}$ represents the current temperature (Inlet/Outlet temp, CPU temp)
$\boldsymbol{W}$ represents the workload (Power Sum, CPU usage, Memory)
$\boldsymbol{T}$ is the prediction CPU temperature

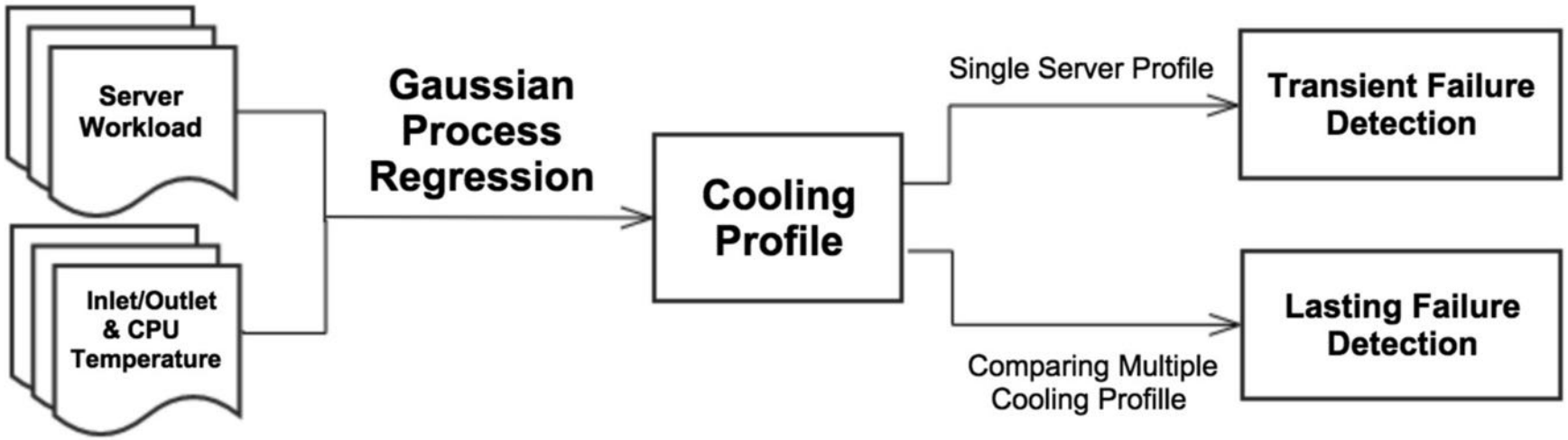


Build Up Cooling Profile

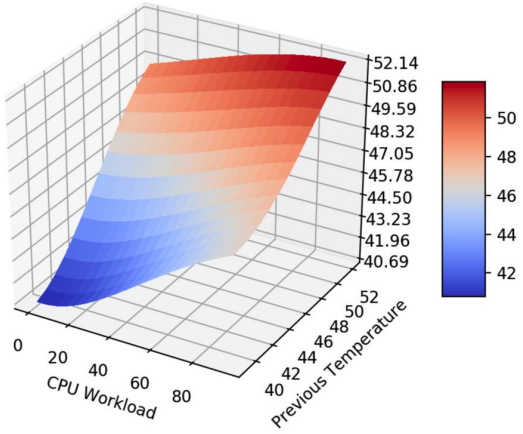
Gaussian Process Regression (GPR)

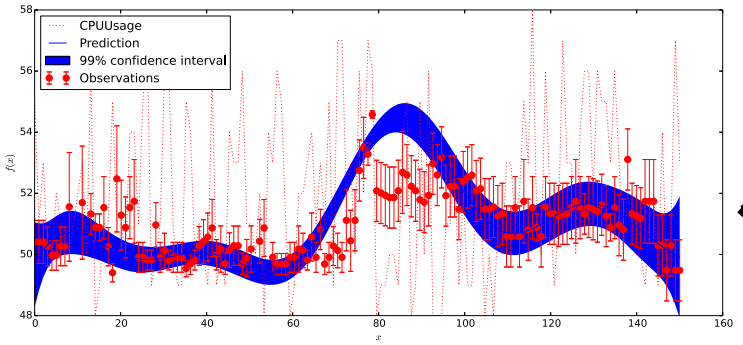
Algorithms	Accuracy	Time
Linear Regression	90.12%	5 Sec
Support Vector Machines	79.84%	2 Min
Gaussian Process Regression	95.24%	35 Min
Conditional Random Field	94.64%	13 Hours

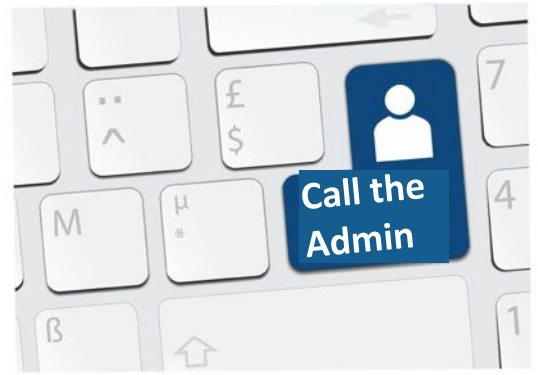
Cooling Profile Model

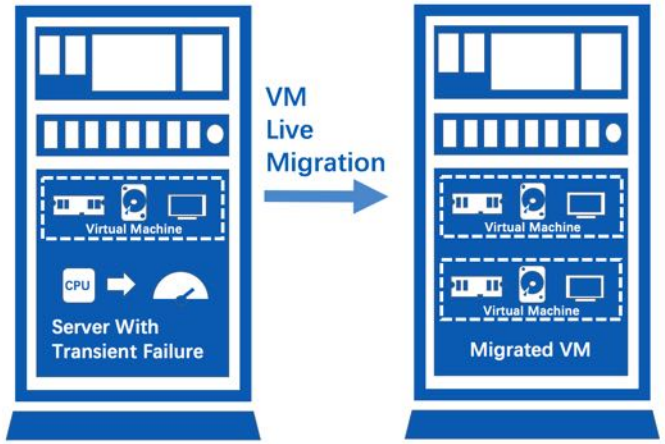


Cooling Profile Detects Transient Failure







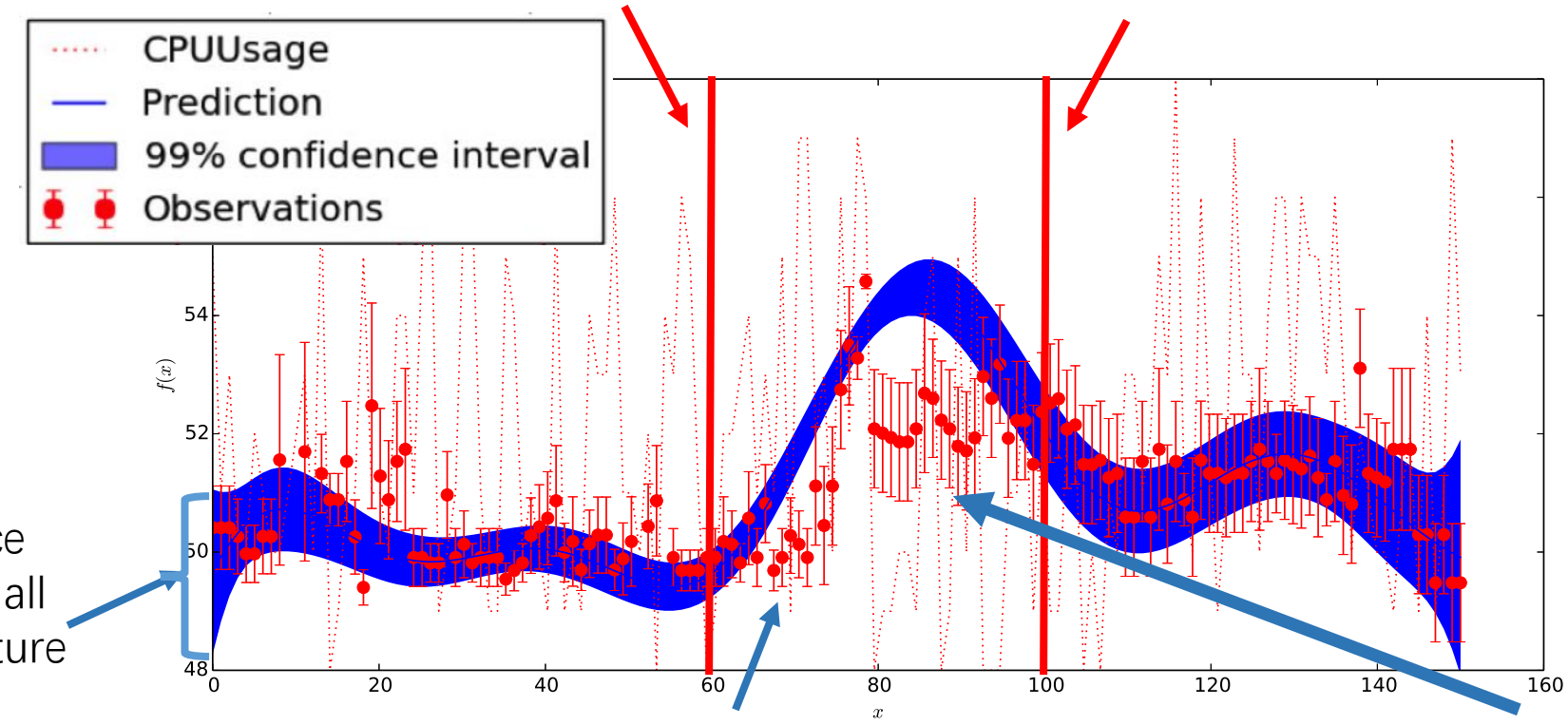


Live Migration to the available server with good cooling profile

Detecting Transient Failures

60-th we seal the inlet/outlet

100-th release the block



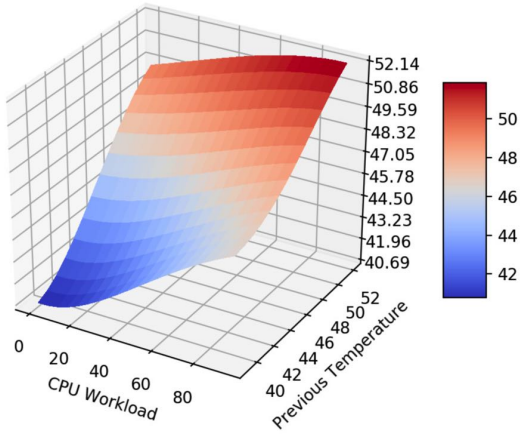
99% confidence interval cover all CPU temperature under normal case

Time series

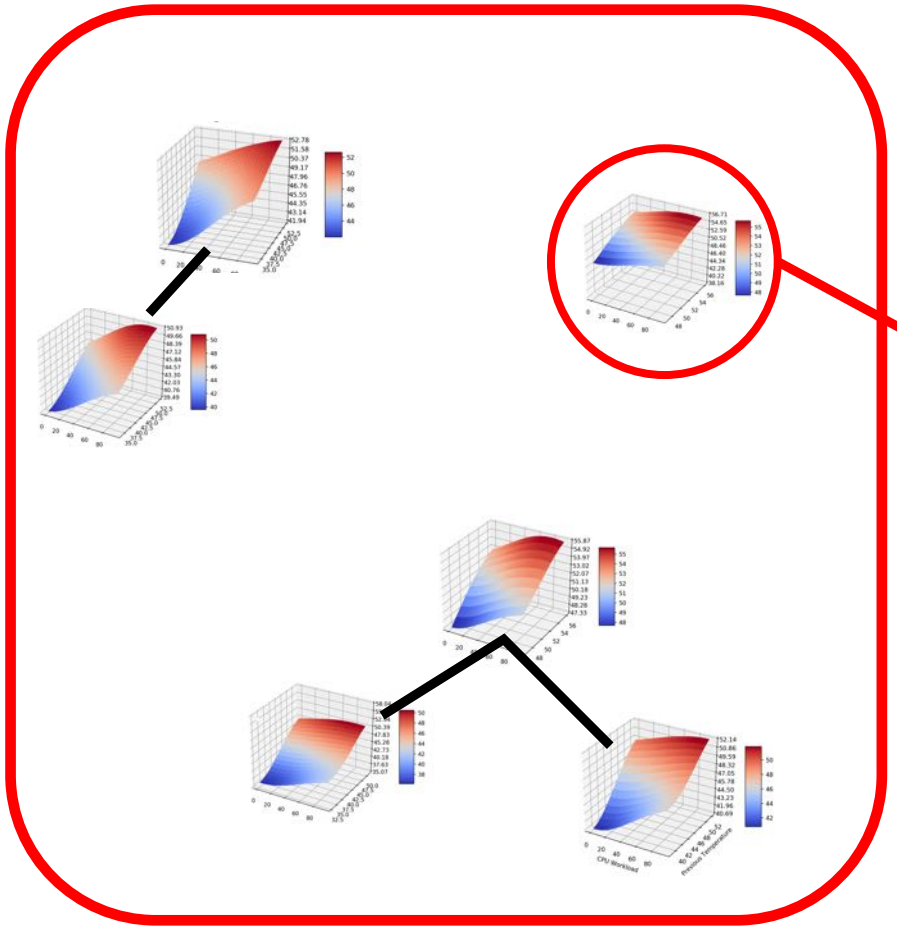
70-th cooling profile detect transient failure

Anomaly CPU temperature raise the fan speed so the actual temperature lower than the prediction.

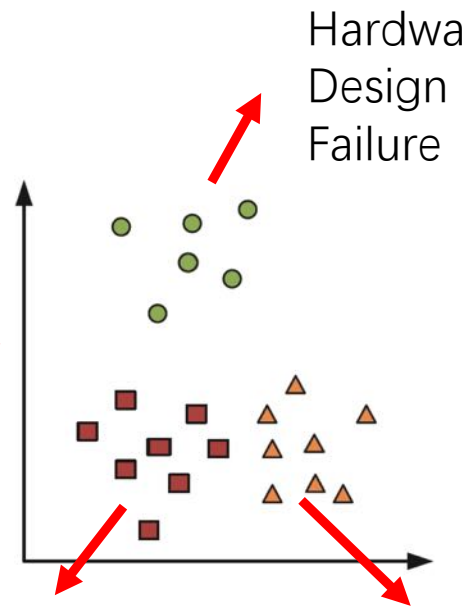
Cooling Profile Detects Lasting Failure



Unsupervised Anomaly Detection



K-means

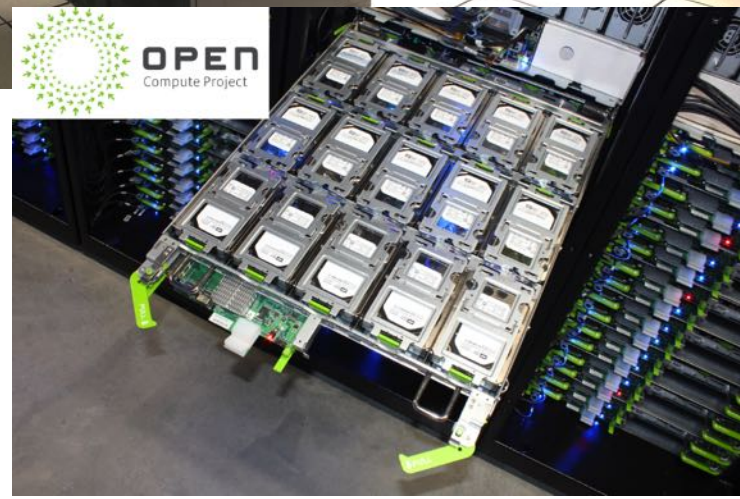


Hardware Design Failure

Non-fatal Server

Poor Cooling Position

Evaluation Setup



DC-A

- Host 200+ 2U rack servers.
- Four rows of racks, six per row.
- Two air conditioner units uses under floor cooling.

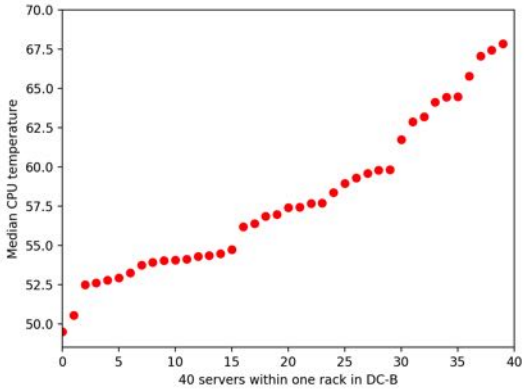
DC-B

- Host 150+ Open Compute Project (OCP) servers.
- Four Open Compute Project (OCP) standard racks.
- A single air conditioner uses overhead cooling.

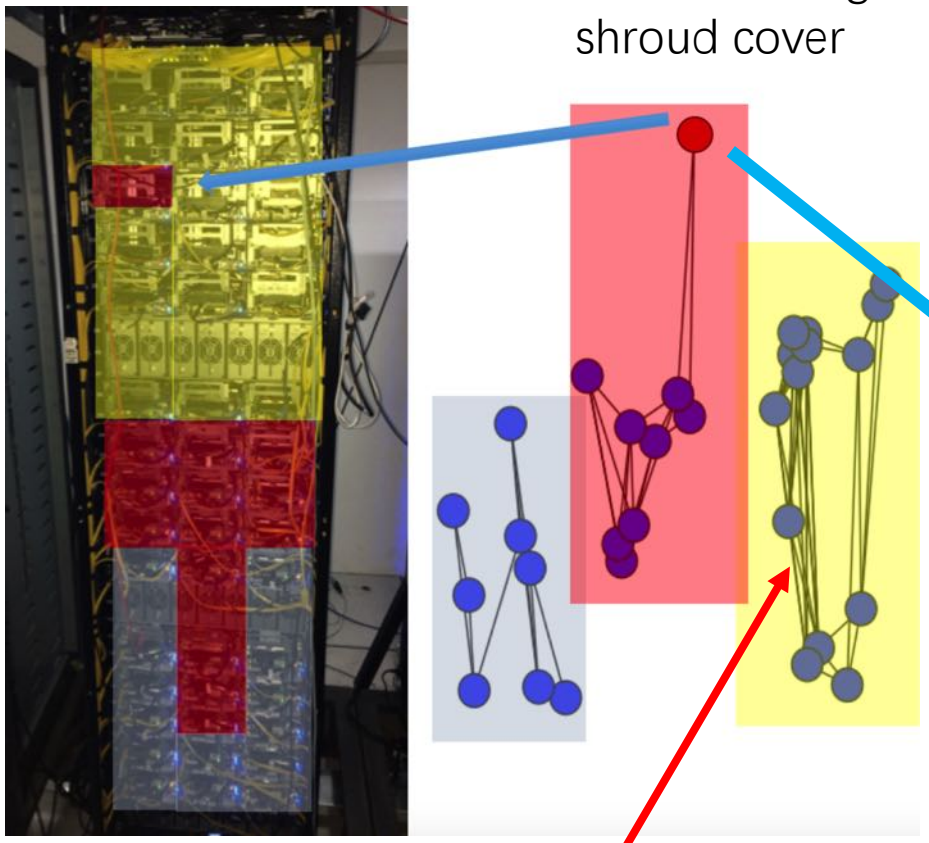
DC-C

- Host over a hundred thousand servers serving real production jobs for a large-scale Internet service company.
- We do not have information of servers and air conditioner.

Detecting lasting problems



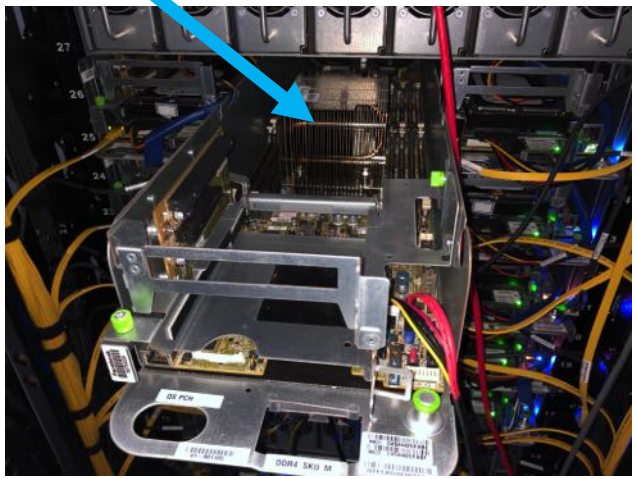
With two obvious inflexions we determine K=3 when using k-means clustering algorithm.



Server missing shroud cover

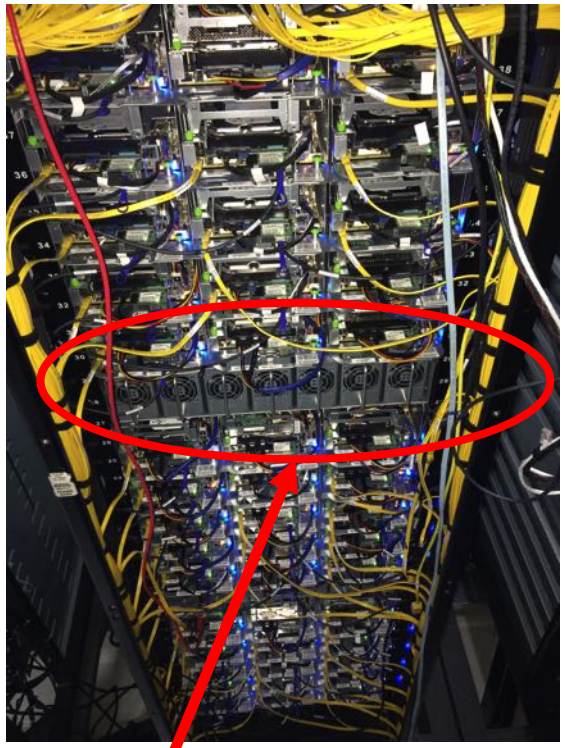
Euclidean distance between server to server

Normal Server

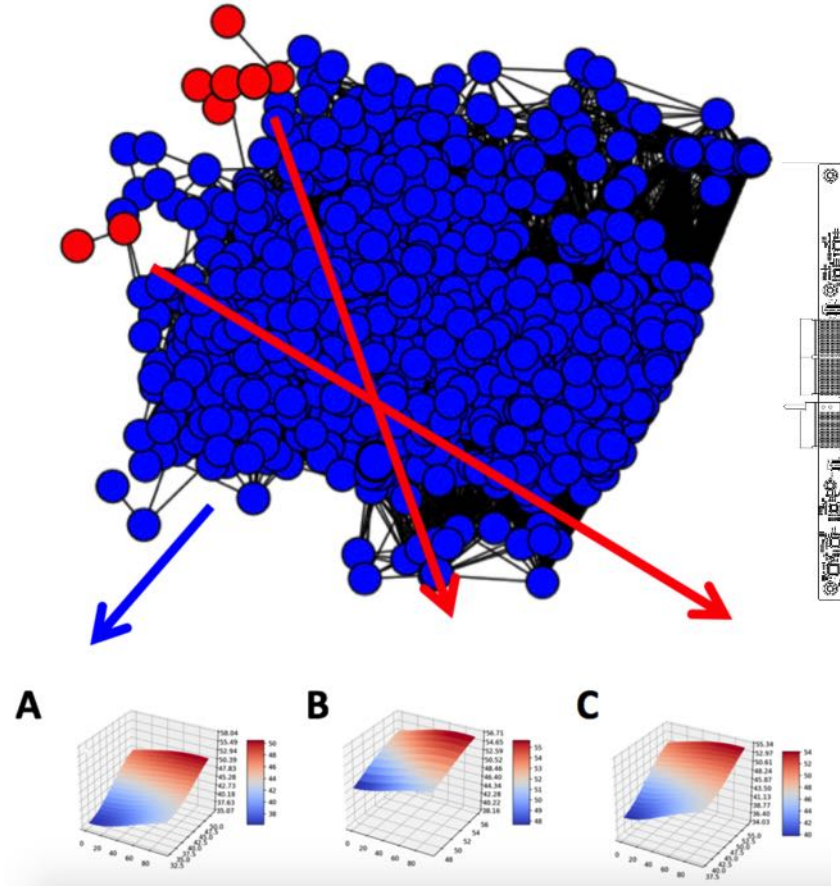


Detecting lasting problems

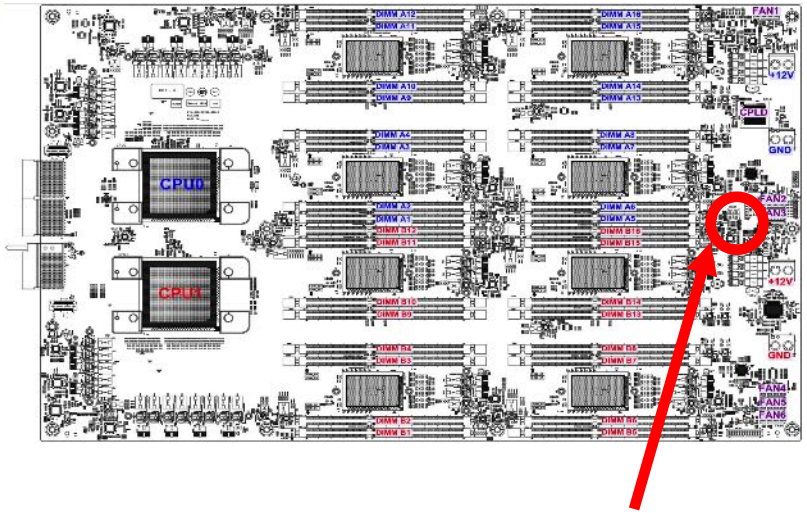
Design Failure



Power supply gets over heat and affects nearby servers



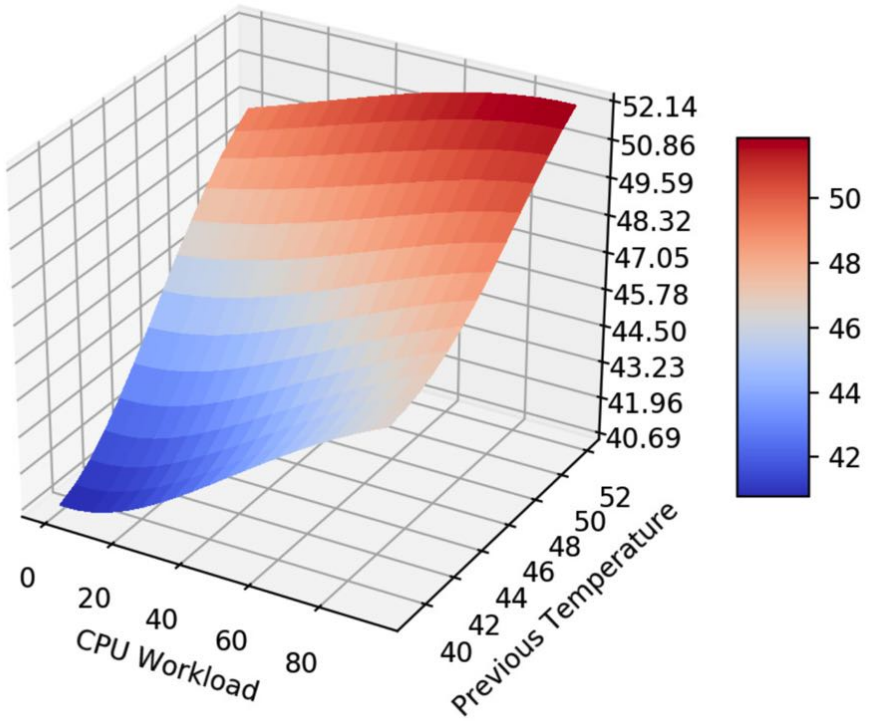
Non-fatal devices



Over Heat

Conclusion

- Cooling profile definition: We capture the ***overall*** cooling capability of each individual server with Gaussian Process Regression model.
- We can use cooling profile to detect transient & lasting cooling problems
- Data we use readily available metrics while the data center is running production workload.



Thank you!