ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Rule-based Thermal Anomaly Detection for Tier-0 HPC systems

Mission: *Improving the datacenter' performance by predicting the anomalies (in particular thermal anomalies) in advance.*

Mohsen Seyedkazemi Ardebili *, Andrea Bartolini *, Andrea Acquaviva*, Luca Benini * †

{mohsen.seyedkazemi, a.Bartolini, andrea.acquaviva, luca.benini}@unibo.it, lbenini@iis.ee.ethz.ch

*Department of Electrical, Electronic and Information Engineering, University of Bologna, Italy

†Integrated Systems Laboratory, ETH Zurich, Switzerland

European Processor Initiative

THE EUPILOT

REGALE
Open Architecture for Exascale Supercomputers

CINECA

Outline

1. Introduction
2. Contributions
3. Experimental Results
4. Conclusions and Future Works

Outline

1. Introduction
2. Contributions
3. Experimental Results
4. Conclusions and Future Works

What Are Datacenter Thermal Hazards/Anomalies?

Heat generated by the computing nodes

Heat removed by the cooling systems

- Thermal Hazard/Anomaly:
  - Thermal hazards can lead to IT and facility equipment damage as well as an outage of the datacenter
- Why does thermal hazard occur?
  - Large electrical power consumption in the range of megawatts which gets completely transformed into (a lot of) heat.
  - Thermal bottlenecks in real-life production workload due to significant spatial and temporal thermal heterogeneity [1,2].
  - A minor thermal issue can potentially start a chain of events that leads to an imbalance of thermal conditions and originating thermal hazards.

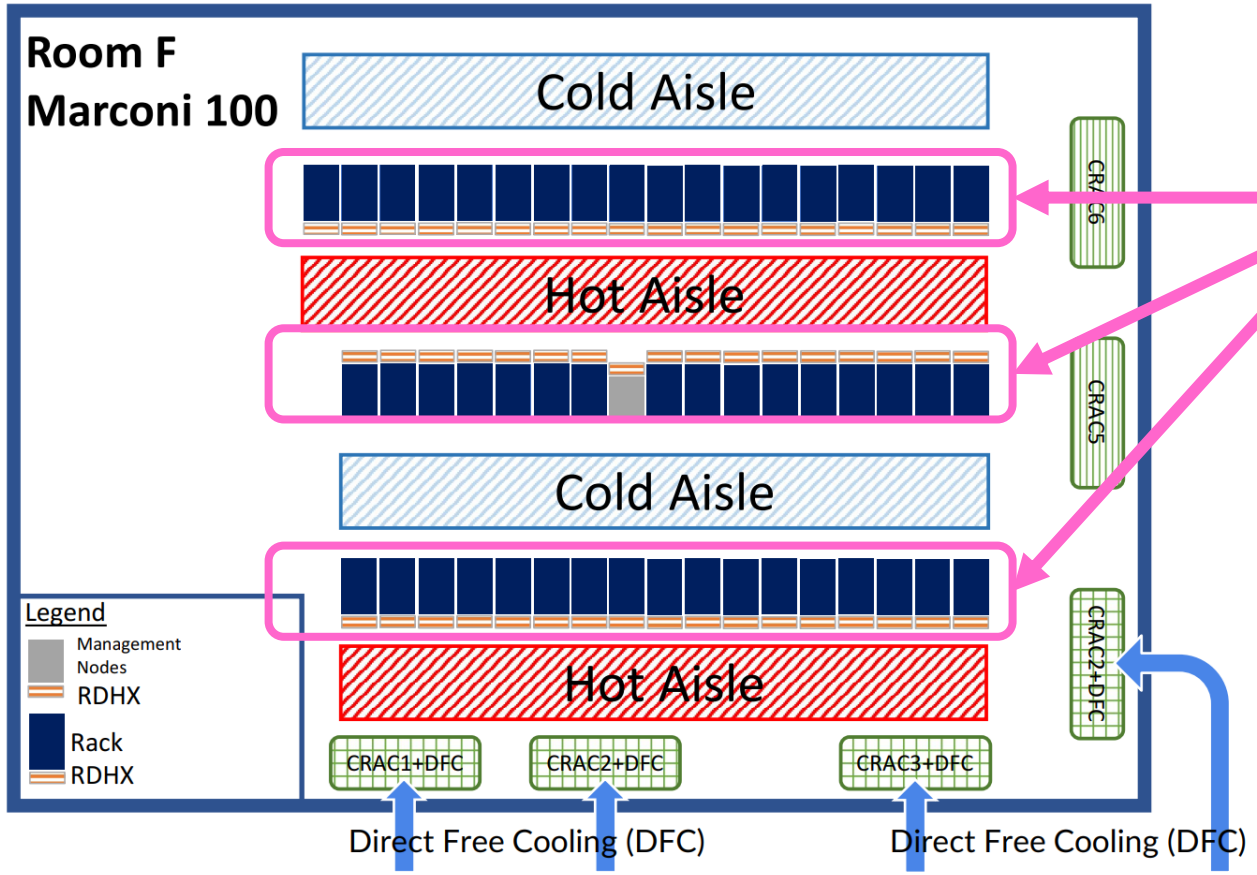
[1] Mohsen Seyedkazemi Ardebili, Carlo Cavazzoni, Luca Benini, and Andrea Bartolini. Thermal characterization of a tier0 datacenter room in normal and thermal emergency conditions. In International Conference on High Performance Computing in Science and Engineering, pages 1–16. Springer, 2019.
[2] Mohsen Seyedkazemi Ardebili, Davide Brunelli, Tommaso Polonelli, Luca Benini, and Andrea Bartolini. A full-stack and end-to-end iot framework for room temperature modelling on large-scale. Available at SSRN 4075667.

SoA Thermal Anomaly Preventive Actions in Datacenter

- Careful design and worst-case design.
  - It has some limitations:
    - The different life spans for facilities and computes nodes/ supercomputer
    - Heat waves and climate change
    - It is not efficient and has more carbon footprints.
- Human in control loop:
  - Datacenter experts constantly (24x7) monitor the datacenter thermal/power conditions to identify the possible signature of the failure/future failure and update the setpoint to prevent the failure. This is expensive and not effective.

Early identifying failures is a complex task that requires automated tools.

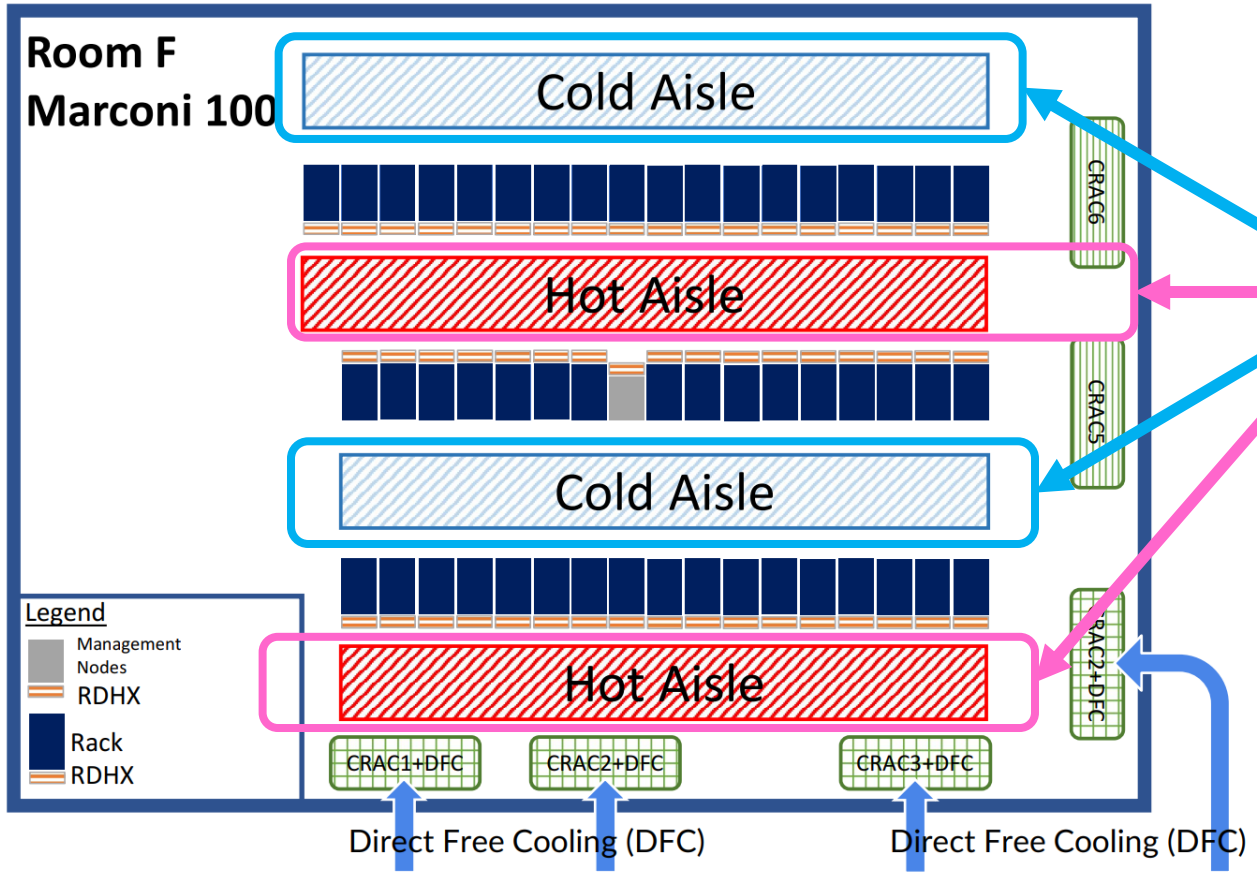
Complex Cooling Systems



- CINECA is a Large Scale HPC Facility in Europe and a PRACE Tier-0 Hosting Site [1]
- Marconi -100: Ranked 9th in June 2020 Top500 list. [2]
  - 980 Compute Node each with 2 CPUs and 4 GPUs. (~32 PFlop/s)
  - Cooling system is general to most of datacenters

[1] https://www.hpc.cineca.it/hardware/marconi100
[2] https://www.top500.org/
[3] Bartolini, A., et al.: Paving the way toward energy-aware and automated datacentre.ICPP 2019, pp. 8:1–8:8. ACM, New York (2019)

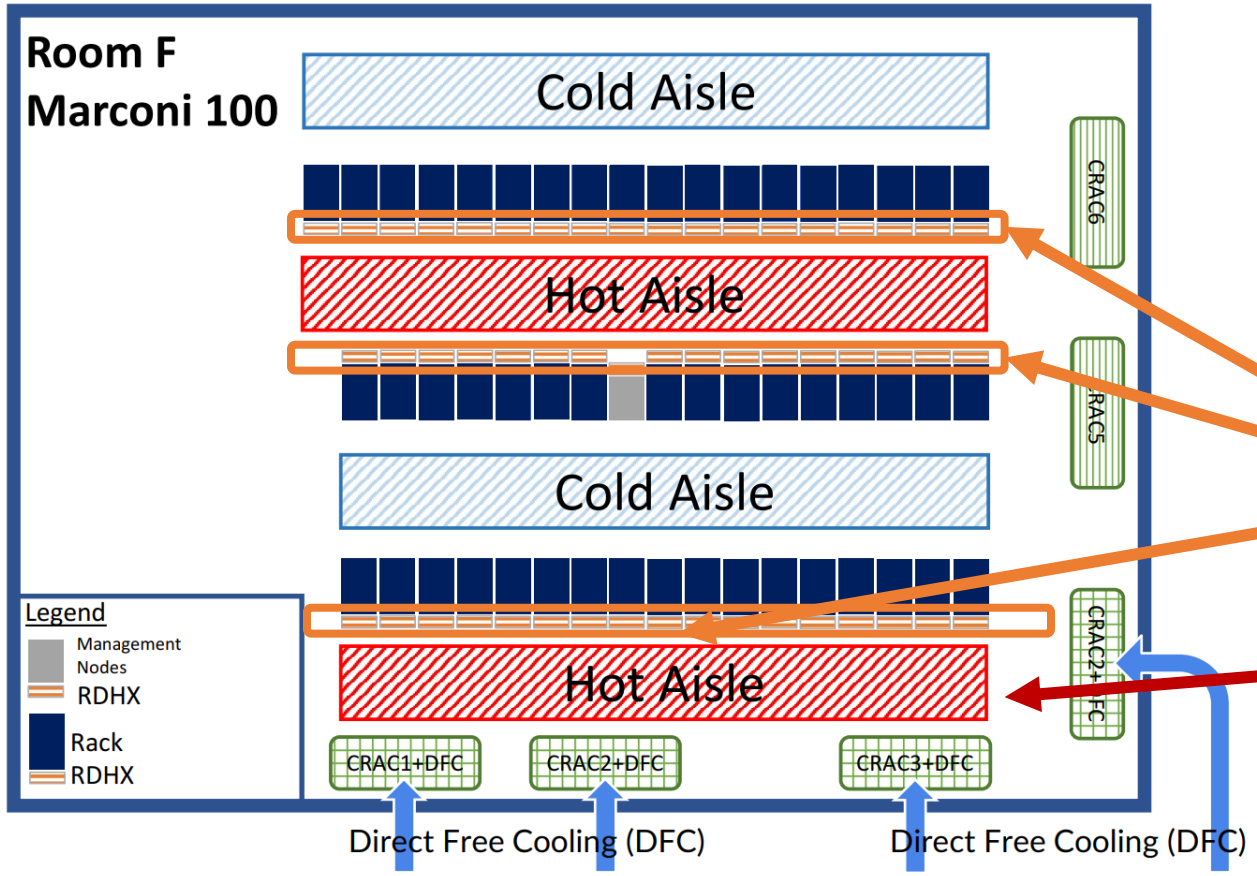
Complex Cooling Systems



- CINECA is a Large Scale HPC Facility in Europe and a PRACE Tier-0 Hosting Site [1]
- Marconi -100: Ranked 9th in June 2020 Top500 list. [2]
  - 980 Compute Node each with 2 CPUs and 4 GPUs. (~32 PFlop/s)
  - Hot/Cold Aisle

[1] https://www.hpc.cineca.it/hardware/marconi100
[2] https://www.top500.org/
[3] Bartolini, A., et al.: Paving the way toward energy-aware and automated datacentre.ICPP 2019, pp. 8:1–8:8. ACM, New York (2019)

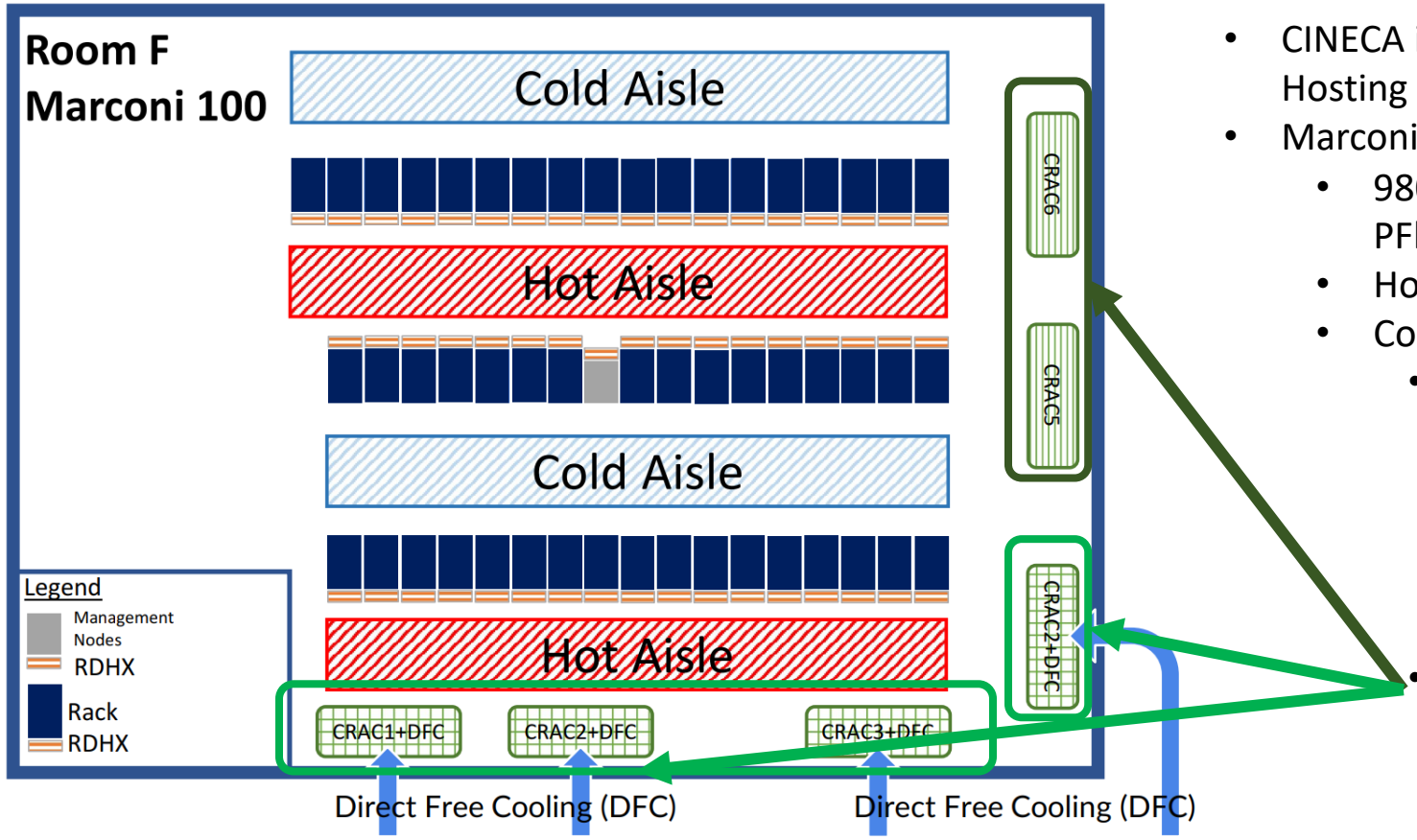
Complex Cooling Systems



- CINECA is a Large Scale HPC Facility in Europe and a PRACE Tier-0 Hosting Site [1]
- Marconi -100: Ranked 9th in June 2020 Top500 list. [2]
  - 980 Compute Node each with 2 CPUs and 4 GPUs. (~32 PFlop/s)
  - Hot/Cold Aisle
  - Cooling Systems:
    - Water Cooling System:
      - Rear Door Heat Exchangers (RDHX)
      - The RDHX device is placed in front of the hot outlet airflow of the compute node.
      - All racks are equipped with RDHX
      - RDHX of racks are in the hot aisle

[1] https://www.hpc.cineca.it/hardware/marconi100
[2] https://www.top500.org/
[3] Bartolini, A., et al.: Paving the way toward energy-aware and automated datacentre.ICPP 2019, pp. 8:1–8:8. ACM, New York (2019)

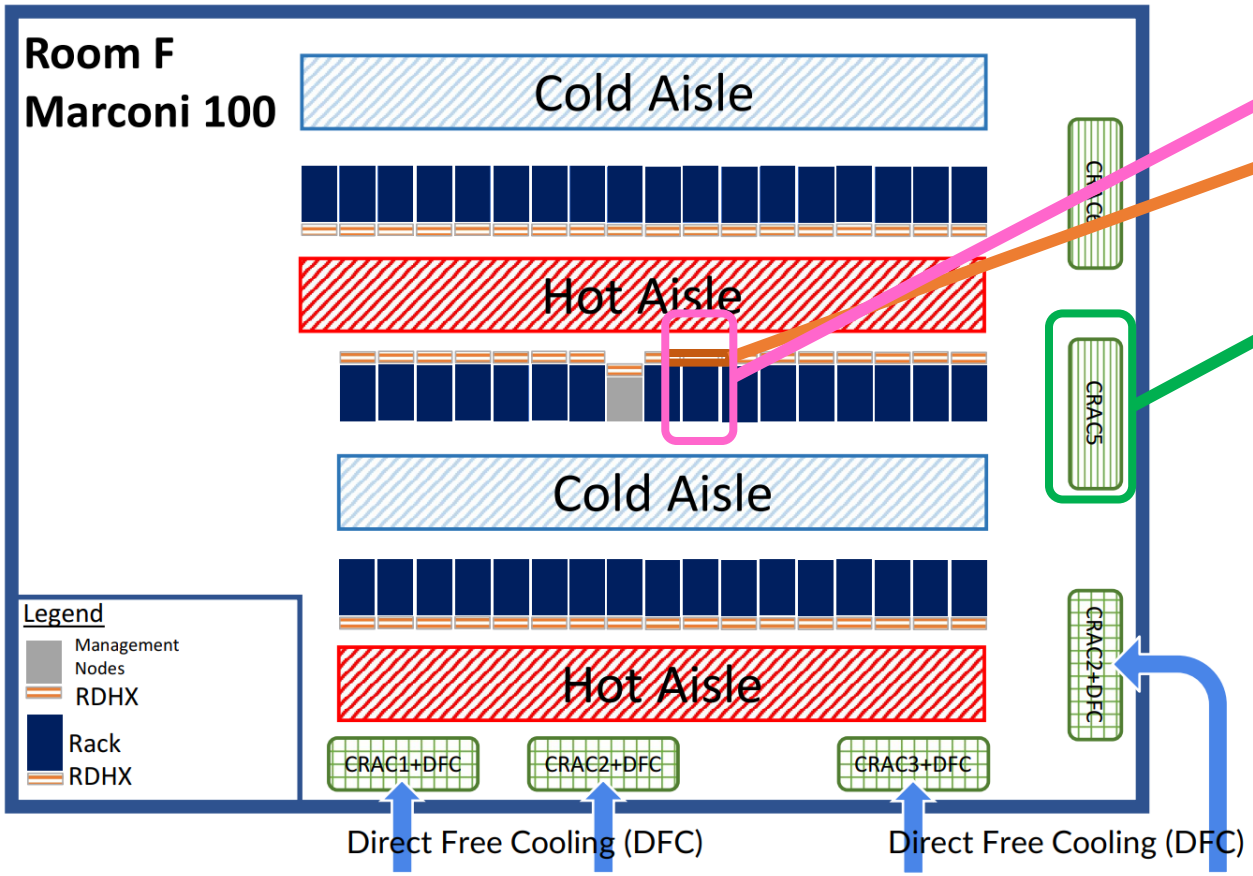
Complex Cooling Systems

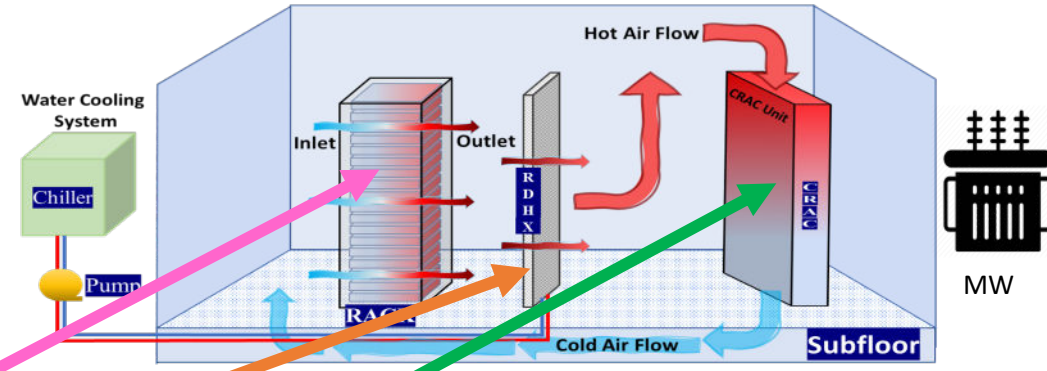


- CINECA is a Large Scale HPC Facility in Europe and a PRACE Tier-0 Hosting Site [1]
- Marconi -100: Ranked 9th in June 2020 Top500 list. [2]
  - 980 Compute Node each with 2 CPUs and 4 GPUs. (~32 PFlop/s)
  - Hot/Cold Aisle
  - Cooling Systems:
    - Water Cooling System:
      - Rear Door Heat Exchangers (RDHX)
      - The RDHX device is placed in front of the hot outlet airflow of the compute node.
      - All racks are equipped with RDHX
      - RDHX of racks are in the hot aisle.
    - Six Computer Room Air Conditioning (CRAC) units
      - Four CRAC units support the Direct Free Cooling (DFC)

[1] https://www.hpc.cineca.it/hardware/marconi100
[2] https://www.top500.org/
[3] Bartolini, A., et al.: Paving the way toward energy-aware and automated datacentre.ICPP 2019, pp. 8:1–8:8. ACM, New York (2019)

Complex Cooling Systems

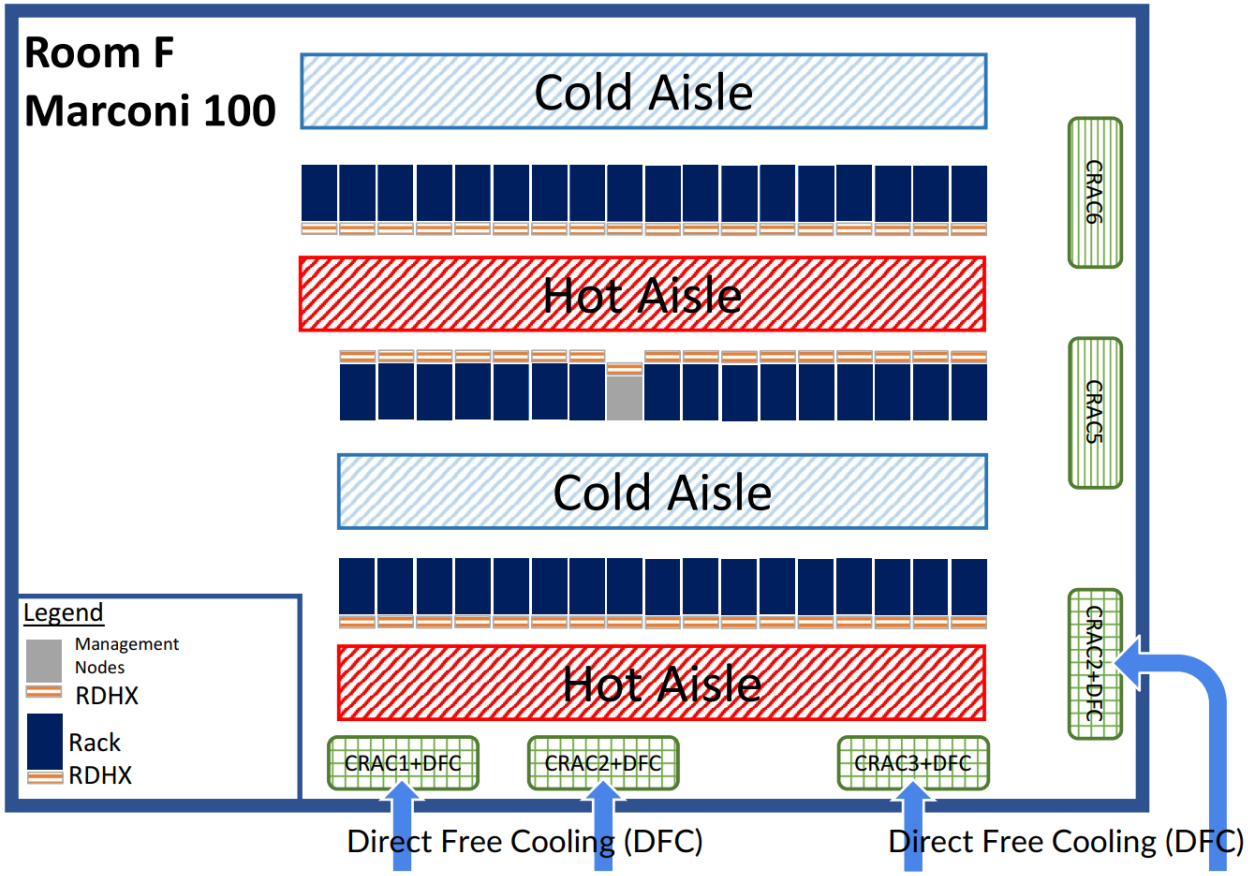


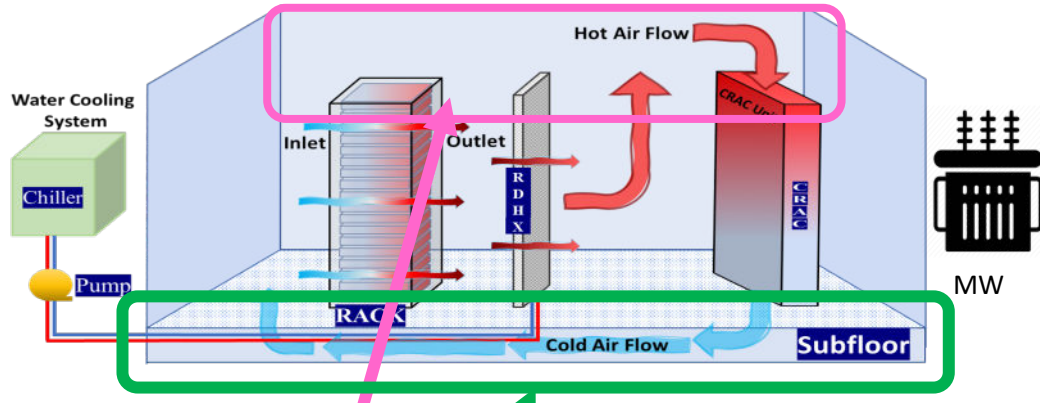


- CINECA is a Large Scale HPC Facility in Europe and a PRACE Tier-0 Hosting Site [1]
- Marconi-100: Ranked 9th in June 2020 Top500 list. [2]
  980 Compute Node each with 2 CPUs and 4 GPUs. (~32 PFlop/s)
  - Hot/Cold Aisle
  - Cooling Systems:
    - Water Cooling System:
      - Rear Door Heat Exchangers (RDHX)
      - The RDHX device is placed in front of the hot outlet airflow of the compute node.
      - All racks are equipped with RDHX
      - RDHX of racks are in the hot aisle.
    - Six Computer Room Air Conditioning (CRAC) units
      - Four CRAC units support the Direct Free Cooling (DFC)

[1] https://www.hpc.cineca.it/hardware/marconi100
[2] https://www.top500.org/
[3] Bartolini, A., et al.: Paving the way toward energy-aware and automated datacentre.ICPP 2019, pp. 8:1–8:8. ACM, New York (2019)

Complex Cooling Systems





- CINECA is a Large Scale HPC Facility in Europe and a PRACE Tier-0 Hosting Site [1]
- Marconi -100: Ranked 9th in June 2020 Top500 list. [2]
  - 980 Compute Node each with 2 CPUs and 4 GPUs. (~32 PFlop/s)
  - Hot/Cold Aisle
  - Cooling Systems:
    - Water Cooling System:
      - Rear Door Heat Exchangers (RDHX)
      - The RDHX device is placed in front of the hot outlet airflow of the compute node.
      - All racks are equipped with RDHX
      - RDHX of racks are in the hot aisle.
    - Six Computer Room Air Conditioning (CRAC) units
      - Four CRAC units support the Direct Free Cooling (DFC)
      - Cold airflow moves under the raised floor
      - Hot air returns to the CRAC units above the raised floor.

[1] https://www.hpc.cineca.it/hardware/marconi100
[2] https://www.top500.org/
[3] Bartolini, A., et al.: Paving the way toward energy-aware and automated datacentre.ICPP 2019, pp. 8:1–8:8. ACM, New York (2019)

Datacenter Thermal Hazards/Anomalies in Practice

[HPC-NEWS] Marconi100: reduced production due to extraordinary heatwave

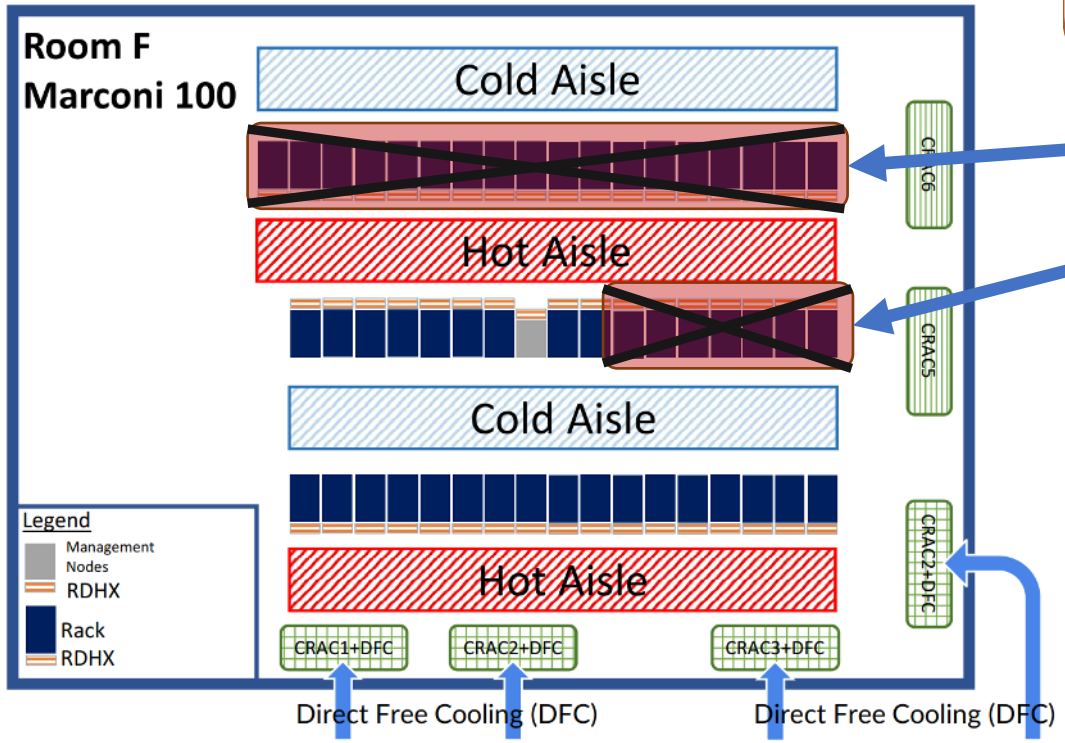
H Hpc-news on behalf of HPC-news
Wed 7/28/2021 6:16 PM
To: hpc-news@list.cineca.it

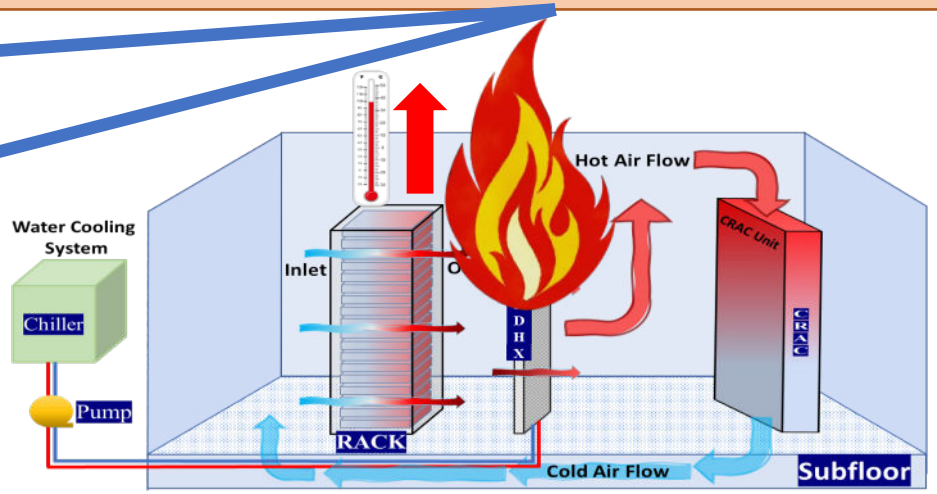
ATT00001.txt
536 bytes

28/07/2021

Reported Real Case

Extraordinary Heatwave
@6PM of 28-07-2021 Marconi100
(9th Most Powerful Computing System (2020))
Cooling Shortage & Thermal Hazards
Reduced 50% of its Computing Capacity
Outage of 380 Nodes





MW

SoA Anomaly Preventive Actions in Datacenter

- Careful design and worst-case design.
  - It has some limitations:
    - A couple of years ago, when the datacenter was designed, the designer did not know about the future.
    - Heat waves and climate change
    - It is not efficient and has more carbon footprints.
- Human in control loop:
  - Datacenter experts constantly (24x7) monitor the datacenter thermal/power conditions to identify the possible signature of the failure/future failure and update the setpoint to prevent the failure. This is expensive and not effective.

- Thermal Anomaly is Severe
- The cooling system is complex and involves many hierarchical sensors on multiple levels (Infrastructures and compute nodes)
- Human in the control loop and worst-case design does not work
- Sysadmins need: a tool to highlight the components which lead to the thermal anomaly

→ Need for Automated Approach

Automated Tools in Literature

SoA Anomaly Detection in Datacenter

Automated Methods
Rule-based
ML-based

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015)
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on World Wide Web. pp. 469–478 (2017)
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2015)
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Symposium(2016)
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Springer (2021)
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM '16, Association for Computing Machinery, New York, NY, USA (2016).
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence 85, 634–644 (2019)
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

SoA Anomaly Detection in Datacenter

Automated Tools in Literature

Water/Air Cooling Facilities,
Compute Nodes
Synthetic Anomalies

Automated Methods	Dataset		
	Scale	Real Failure	Label
Rule-based	Nodes[2] Racks[0,1]	-	-
ML-based	Nodes Racks[3,7,8]	Nodes Cooling Sys[4]	X[7,8]

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015)
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on World Wide Web. pp. 469–478 (2017)
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2015)
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Symposium(2016)
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Springer (2021)
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM '16, Association for Computing Machinery, New York, NY, USA (2016).
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence 85, 634–644 (2019)
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

SoA Anomaly Detection in Datacenter

Automated Tools in Literature

Water/Air Cooling Facilities,
Compute Nodes
Synthetic Anomalies

Node Level
Application Level
Not Datacenter Level

Automated Methods	Dataset			Scalable
	Scale	Real Failure	Label	
Rule-based	Nodes[2] Racks[0,1]	-	-	X
ML-based	Nodes Racks[3,7,8]	Nodes Cooling Sys[4]	X[7,8]	In General No

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015)
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on World Wide Web. pp. 469–478 (2017)
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2015)
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Symposium(2016)
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Springer (2021)
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM '16, Association for Computing Machinery, New York, NY, USA (2016).
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence 85, 634–644 (2019)
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

SoA Anomaly Detection in Datacenter

Automated Tools in Literature

Water/Air Cooling Facilities, Compute Nodes Synthetic Anomalies

Node Level Application Level Not Datacenter Level

HW, Application

Automated Methods	Dataset			Scalable	Study Focus
	Scale	Real Failure	Label		
Rule-based	Nodes[2] Racks[0,1]	-	-	x	SYS[2], App[0,1], HW[2]
ML-based	Nodes Racks[3,7,8]	Nodes Cooling Sys[4]	X[7,8]	In General No	SYS, App[3], HW, Cooling Fail.[4]

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015)
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on World Wide Web. pp. 469–478 (2017)
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2015)
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Symposium(2016)
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Springer (2021)
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM '16, Association for Computing Machinery, New York, NY, USA (2016).
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence 85, 634–644 (2019)
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

SoA Anomaly Detection in Datacenter

Automated Tools in Literature

Water/Air Cooling Facilities, Compute Nodes Synthetic Anomalies

Node Level Application Level Not Datacenter Level

HW, Application

Automated Methods	Dataset			Scalable	Study Focus	Lack of the Methodology	
	Scale	Real Failure	Label			Rules	Thresholds
Rule-based	Nodes[2] Racks[0,1]	-	-	x	SYS[2], App[0,1], HW[2]	Lack of Mthd.[1,2]	Lack of Mthd.[0,1,2]
ML-based	Nodes Racks[3,7,8]	Nodes Cooling Sys[4]	X[7,8]	In General No	SYS, App[3], HW, Cooling Fail.[4]	-	-

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015)
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on World Wide Web. pp. 469–478 (2017)
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2015)
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Symposium(2016)
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Springer (2021)
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM '16, Association for Computing Machinery, New York, NY, USA (2016).
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence 85, 634–644 (2019)
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

SoA Anomaly Detection in Datacenter

Automated Tools in Literature

Water/Air Cooling Facilities, Compute Nodes Synthetic Anomalies

Node Level Application Level Not Datacenter Level

HW, Application

Real Physical Thermal Hazards

Automated Methods	Dataset			Scalable	Study Focus	Lack of the Methodology		Method Validation
	Scale	Real Failure	Label			Rules	Thresholds	
Rule-based	Nodes[2] Racks[0,1]	-	-	X	SYS[2], App[0,1], HW[2]	Lack of Mthd.[1,2]	Lack of Mthd.[0,1,2]	-
ML-based	Nodes Racks[3,7,8]	Nodes Cooling Sys[4]	X[7,8]	In General No	SYS, App[3], HW, Cooling Fail.[4]	-	-	X[7,8]

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015)
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on World Wide Web. pp. 469–478 (2017)
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2015)
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Symposium(2016)
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Springer (2021)
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM '16, Association for Computing Machinery, New York, NY, USA (2016).
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence 85, 634–644 (2019)
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

SoA Anomaly Detection in Datacenter

Automated Tools in Literature

Water/Air Cooling Facilities, Compute Nodes Synthetic Anomalies

Node Level Application Level Not Datacenter Level

HW, Application

Real Physical Thermal Hazards

Automated Methods	Dataset			Scalable	Study Focus	Lack of the Methodology		Method Validation
	Scale	Real Failure	Label			Rules	Thresholds	
Rule-based	Nodes[2] Racks[0,1]	-	-	X	SYS[2], App[0,1], HW[2]	Lack of Mthd.[1,2]	Lack of Mthd.[0,1,2]	-
ML-based	Nodes Racks[3,7,8]	Nodes Cooling Sys[4]	X[7,8]	In General No	SYS, App[3], HW, Cooling Fail.[4]	-	-	X[7,8]

Contributions

[0] Ahad, et al., : Toward autonomic cloud: Automatic anomaly detection and resolution. In: International Conf. on Cloud and Autonomic Computing. pp. 200–203 (2015
[1] Jayathilaka, et al.,: Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th International Conference on Wor
[2] Brandt , et al.,: Enabling advanced operational analysis through multi-subsystem data integration on trinity. Tech. rep., Sandia National Lab.(SNL-CA), Livermore, CA
[3] Ates, et al.,: Application detection through rich monitoring data. In: European Conference on Parallel Processing. pp. 92–105. Springer (2018)
[4] Cong Li : Cooling anomaly detection for servers and datacenters with Naïve ensemble, Annual IEEE Semiconductor Thermal Measurement and Management Sympos
[5] Aksar et al.,: E2ewatch: An end-to-end anomaly diagnosis framework for production hpc systems. In: European Conference on Parallel Processing. pp. 70–85. Spring
[6] Arzani et al.,: Taking the blame game out of data centers operations with netpoirot. In: Proceedings of the 2016 ACM SIGCOMM Conference. p. 440–453. SIGCOMM
[7] Borghesi et al.,: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vo
[8] Borghesi A.et al.,: A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artifi
[9] Shaykhislamov et al.,: An approach for dynamic detection of inefficient supercomputer applications. Procedia Computer Science 136, 35–43 (2018)
[10] Netti et al.,: A machine learning approach to online fault classification in hpc systems. Future Generation Computer Systems 110, 1009–1022 (2020)
[11] Marathe et al., : An empirical survey of performance and energy efficiency variation on intel processors. In: Proceedings of the 5th International Workshop on Energy Efficient Supercomputing. pp. 1–8 (2017)

Outline

1. Introduction
2. **Contributions**
3. Experimental Results
4. Conclusions and Future Works

Contributions

Automated Methods

1

Rule-based Statistical Tools

Contributions

Automated Methods	Dataset		
	Scale	Real Failure	Label

1

Rule-based Statistical Tools

2

Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the:

- normal in-production and real physical thermal failure (Reported Failure)
- Generate thermal hazard labels (ML)

Contributions

Automated Methods	Dataset			Scalable	Study Focus
	Scale	Real Failure	Label		

1

Rule-based Statistical Tools

2

Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the:

- normal in-production and real physical thermal failure (Reported Failure)
- Generate thermal hazard labels (ML)

3

Room/Datacenter Level Thermal Anomaly Detection

Contributions

Flag: Rule violation named as Flag!

Automated Methods	Dataset			Scalable	Study Focus	Lack of the Methodology	
	Scale	Real Failure	Label			Rules	Thresholds

1

Rule-based Statistical Tools

2

Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the:

- normal in-production and real physical thermal failure (Reported Failure)
- Generate thermal hazard labels (ML)

3

Room/Datacenter Level Thermal Anomaly Detection

4

s: based on a study of istic monitoring signals, During a normal production period and reported real physical thermal failure.

Thresholds:

- Statistical Approaches
- Recommendation (ASHARE)

SLTA: Severity Level of the Thermal Anomaly (SLTA) in the datacenter.

Contributions

Flag: Rule violation named as Flag!

Automated Methods	Dataset			Scalable	Study Focus	Lack of the Methodology		Method Validation
	Scale	Real Failure	Label			Rules	Thresholds	

1

Rule-based Statistical Tools

2

Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the:

- normal in-production and real physical thermal failure (Reported Failure)
- Generate thermal hazard labels (ML)

3

Room/Datacenter Level Thermal Anomaly Detection

4

Rules: based on a [illegible]y of holistic monitoring signals, During a normal production period and reported real physical thermal failure.

Thresholds:

- Statistical Approaches
- Recommendation (ASHARE)

SLTA: Severity Level of the Thermal Anomaly (SLTA) in the datacenter.

5

Real Physical Thermal Hazards

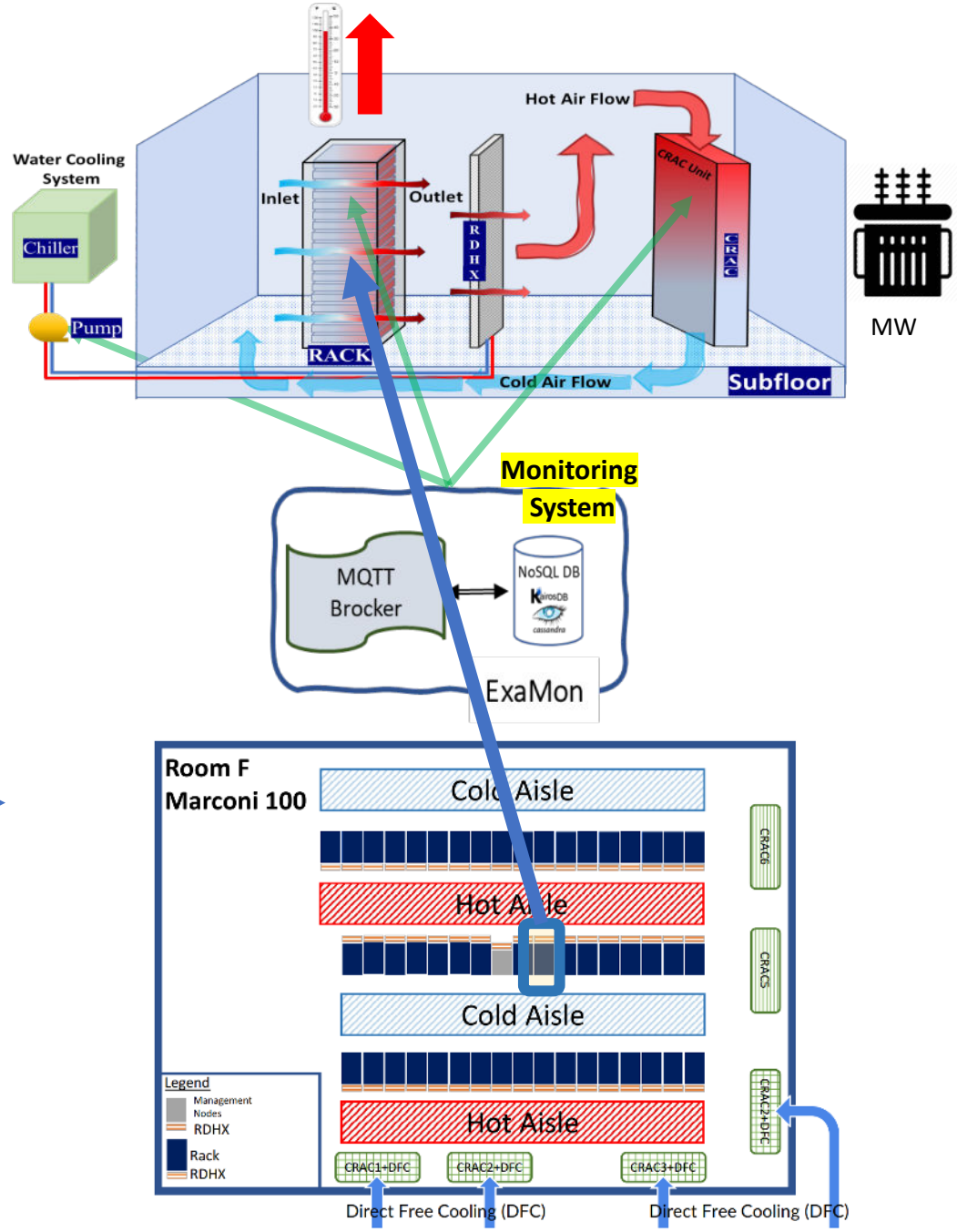
Dataset

1. Dataset: Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the normal in-production and real physical thermal failure (Reported Failure)

- Period of the study: 2021-04-08 to 2021-08-18
- Reported real physical thermal hazard on the 2021-07-28

Reported Real Physical Thermal Hazard
2021-07-28

April 2021 — August 2021

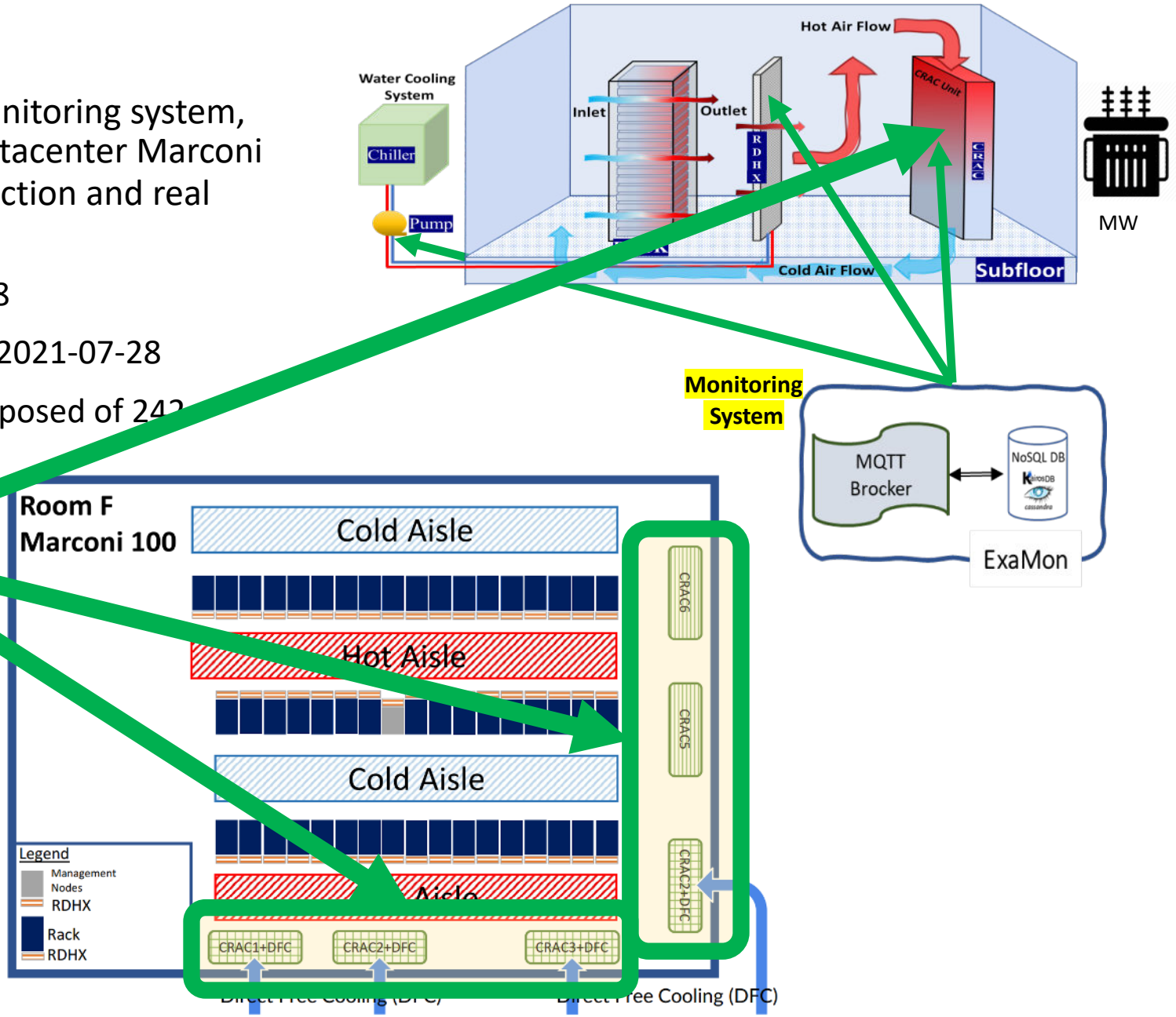


Dataset

1. Dataset: Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the normal in-production and real physical thermal failure (Reported Failure)

- Period of the study: 2021-04-08 to 2021-08-18
- Reported real physical thermal hazard on the 2021-07-28
- Analysis conducted on a reduced dataset composed of 242 parameters
  - Cooling Facilities:
  - Air cooling system CRAC unit

Compressor utilization
Free cooling
Free cooling valve open position
Fan speed
Return, and supply air temperature

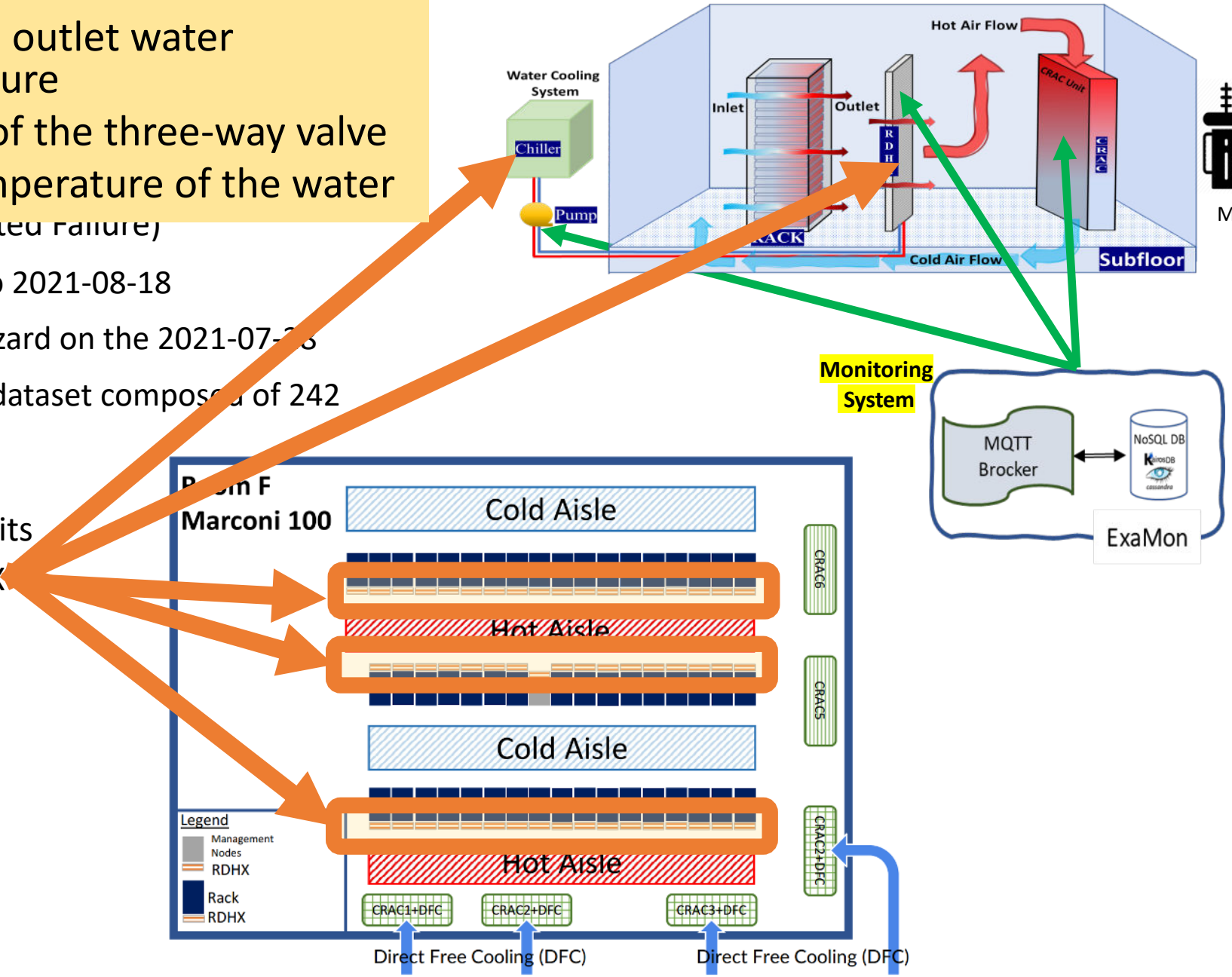


Contribu

1. Dataset: Utili
   we collected a
   100 in CINECA
   physical thermal failure (Reported Failure)

- Period of the study: 2021-04-08 to 2021-08-18
- Reported real physical thermal hazard on the 2021-07-
- Analysis conducted on a reduced dataset compos of 242 parameters
  - Cooling Facilities:
  - Air cooling system CRAC units
  - Water cooling system RDHX

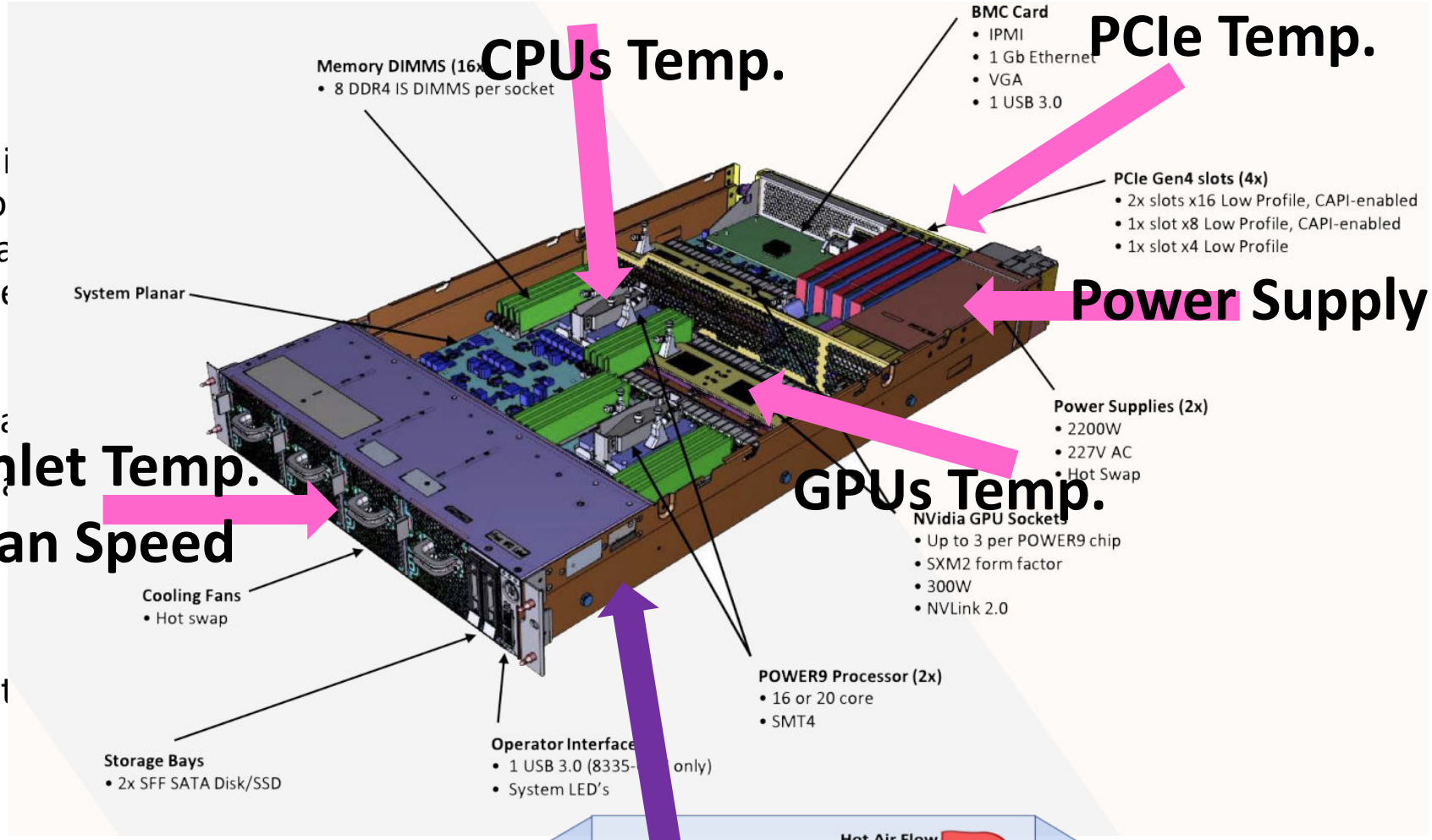
Water flow rate
Inlet, and outlet water temperature
Position of the three-way valve
Delta temperature of the water

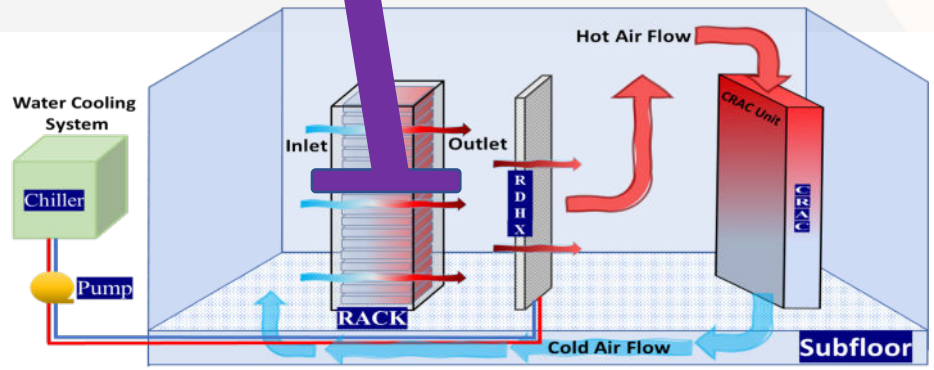


Dataset

1. Dataset: Utilizing EXAMON, whi
   we collected a holistic dataset o
   100 in CINECA during the norma
   physical thermal failure (Reporte

- Period of the study: 2021-04-08 to
- Reported real physical thermal haza
- Analysis conducted on a reduced da
  parameters
  - Cooling Facilities:
  - Water cooling system RDHX
  - Air cooling system CRAC unit
  - One rack with 20 nodes





MW

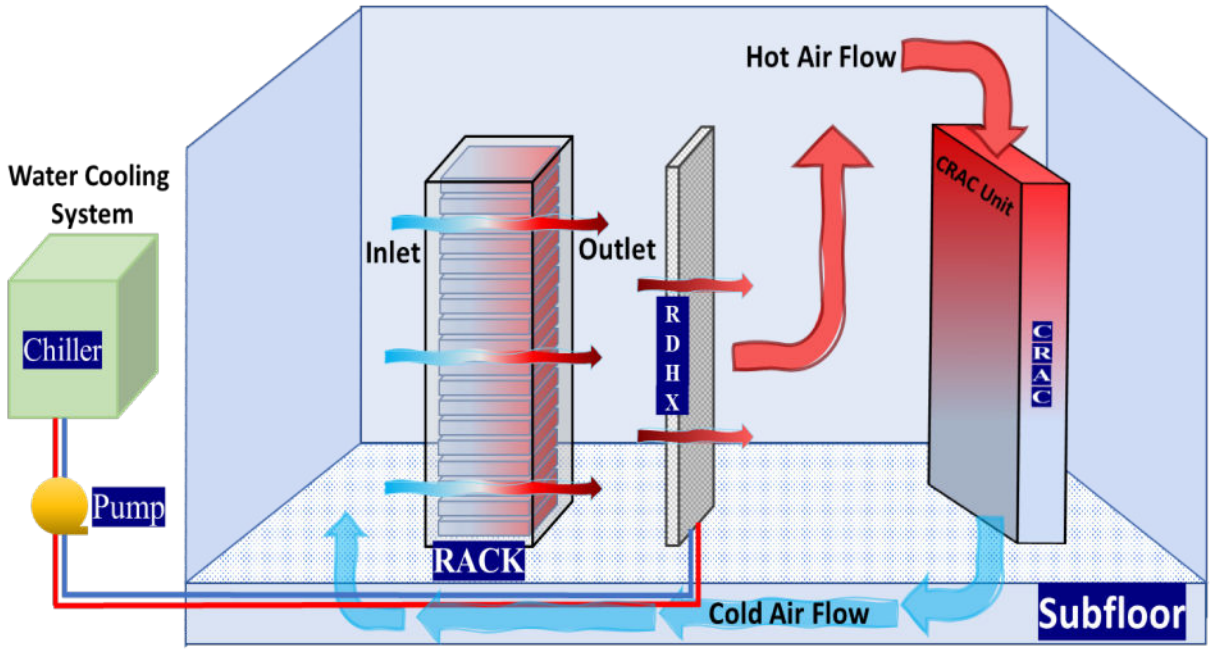
Dataset

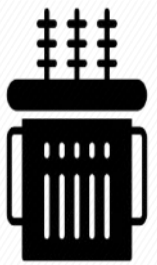
1. Dataset: Utilizing EXAMON, which is a monitoring system, we collected a holistic dataset of Tier-0 datacenter Marconi 100 in CINECA during the normal in-production and real physical thermal failure (Reported Failure)

- Period of the study: 2021-04-08 to 2021-08-18
- Reported real physical thermal hazard on the 2021-07-28
- Analysis conducted on a reduced dataset composed of 242 parameters
  - Cooling Facilities:
  - Water cooling system RDHX
  - Air cooling system CRAC units
  - One rack with 20 nodes
  - Modbus

Main electrical power distributions system (Modbus)

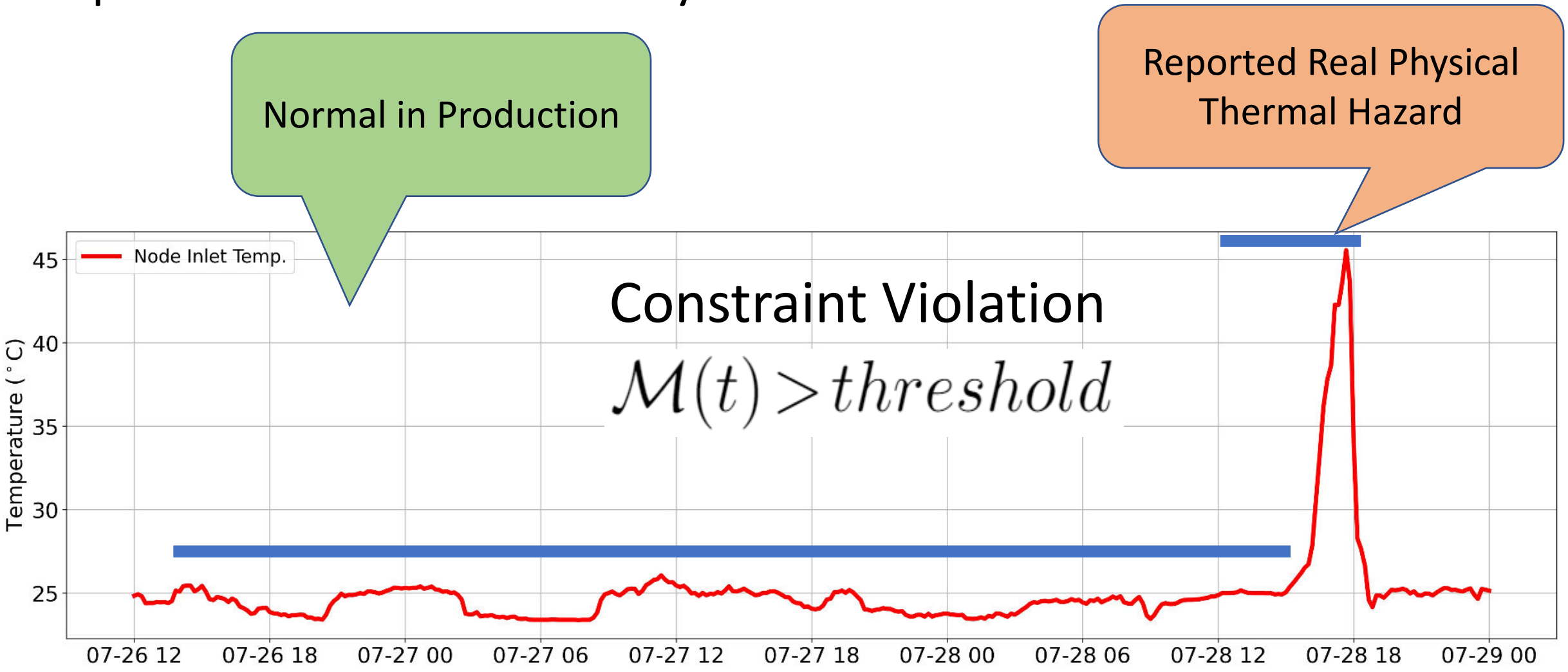
- Total power consumption of ICT
- Total power consumption of RDHX pumps
- Total power consumption of chillers
- Total power consumption of CRAC units
- etc.





MW

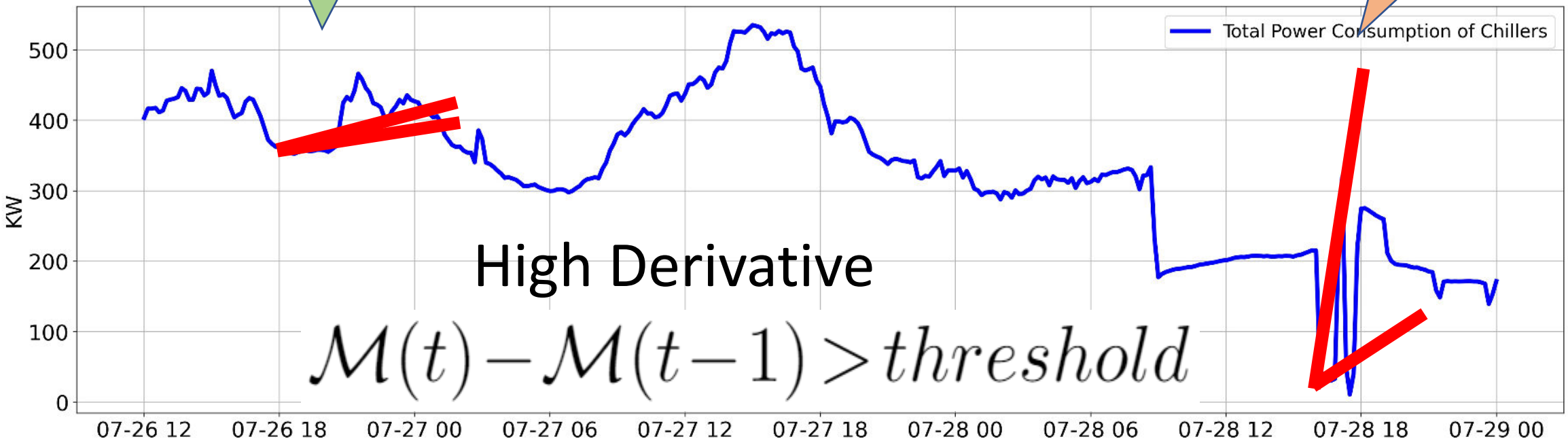
Reported Failure Study



Reported Failure Study

Normal in Production

Reported Real Physical Thermal Hazard

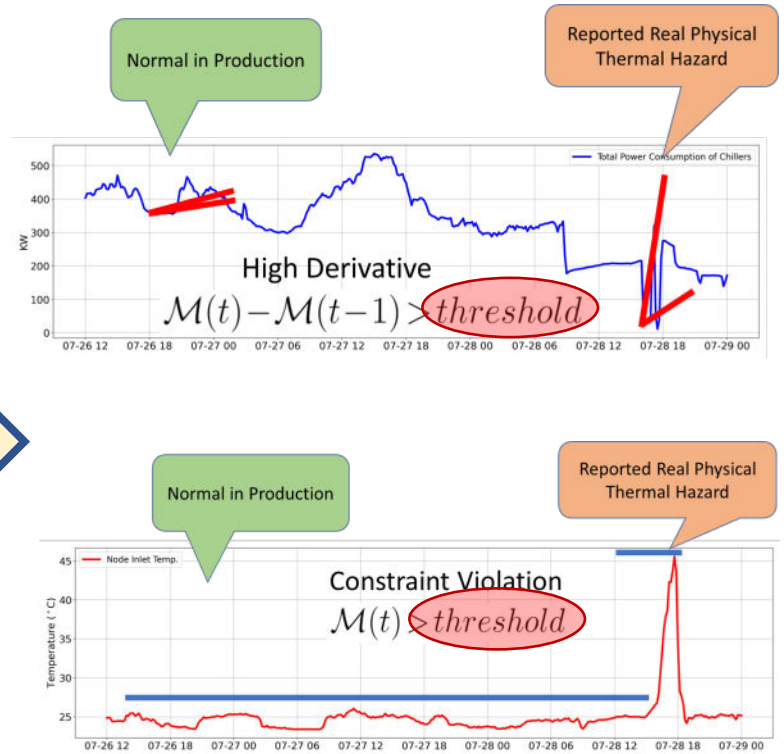


High Derivative

$$\mathcal{M}(t) - \mathcal{M}(t-1) > threshold$$

Methodology

Methodology for Rules Definition

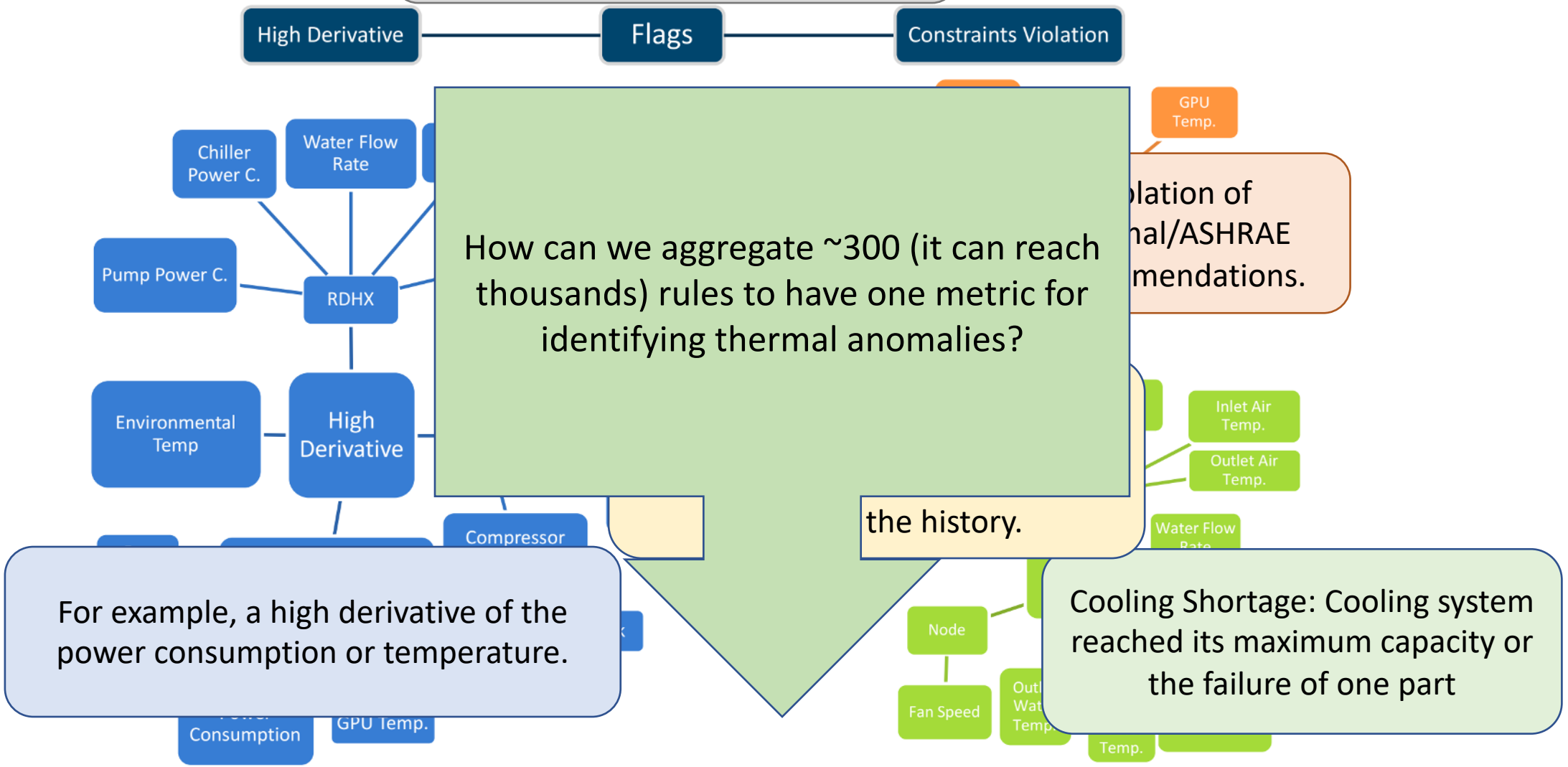


Thresholds:

1. Statistical Approaches
   Quantile of 0.99
2. Recommendations
   ASHRAE

Methodology: F

Two main groups of abnormal patterns
~300 Rules each one can raise a flag



How can we aggregate ~300 (it can reach thousands) rules to have one metric for identifying thermal anomalies?

lation of
nal/ASHRAE
mendations.

the history.

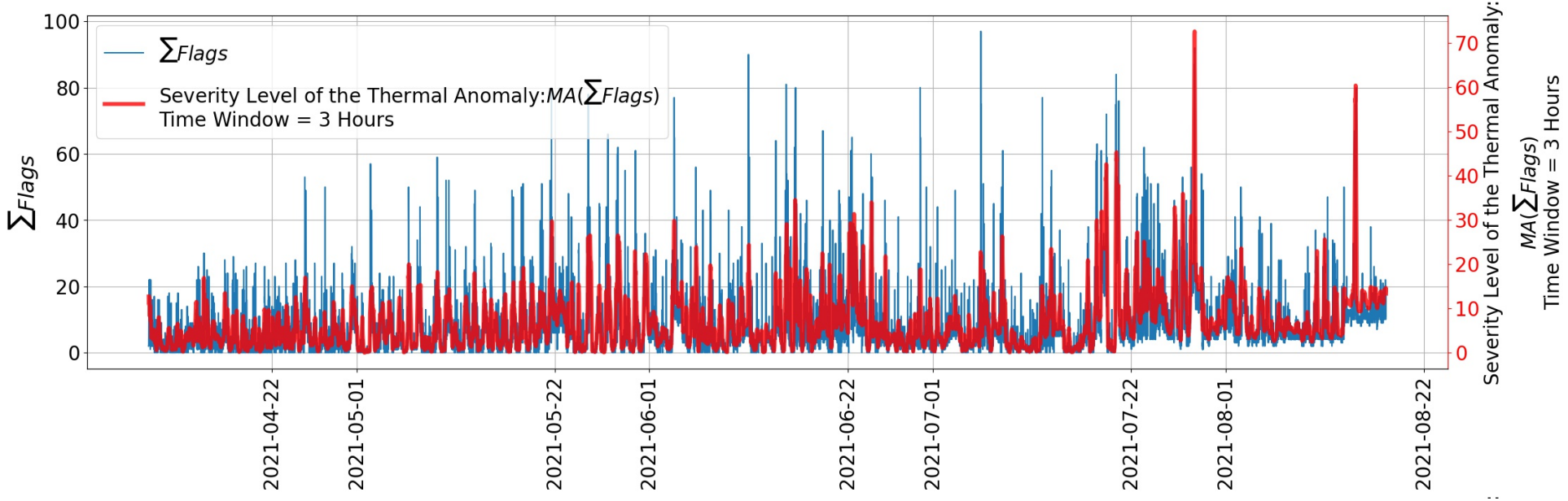
For example, a high derivative of the power consumption or temperature.

Cooling Shortage: Cooling system reached its maximum capacity or the failure of one part

Severity Level of Anomaly MA($\Sigma Flags$)

How can we aggregate ~300 (it can reach thousands) rules to have one metric for identifying thermal anomalies?

- Blue Line:
  - ~300 Rules each one can raise a Flag
  - $\Sigma Flags$
- Red Line:
  - Severity Level of the Thermal Anomaly (SLTA) in the datacenter,
  - Proposed as a new metric calculated by aggregation of rules MovingAverage($\Sigma Flags$)



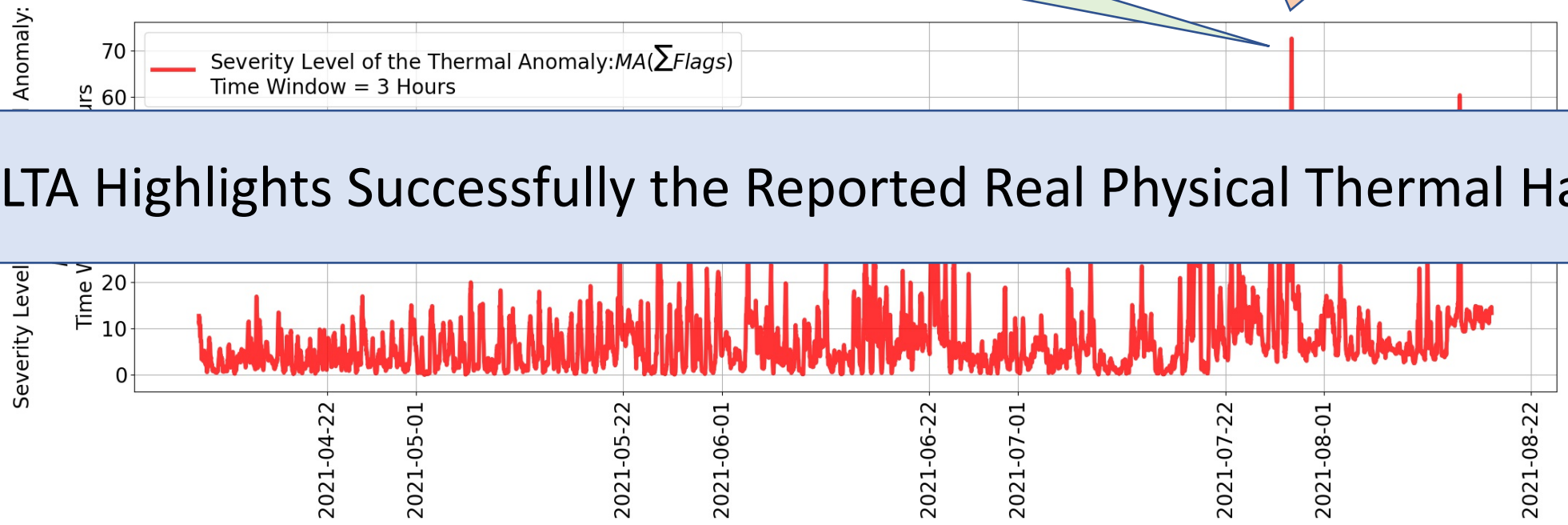
Severity Level of Anomaly MA($\Sigma Flags$)

How can we aggregate ~300 (it can reach thousands) rules to have one metric for identifying thermal anomalies?

- Severity Level of the Thermal Anomaly (SLTA) in the datacenter:
- We applied this met[illegible]aset for 4 months

Peak of SLTA
All Study Period

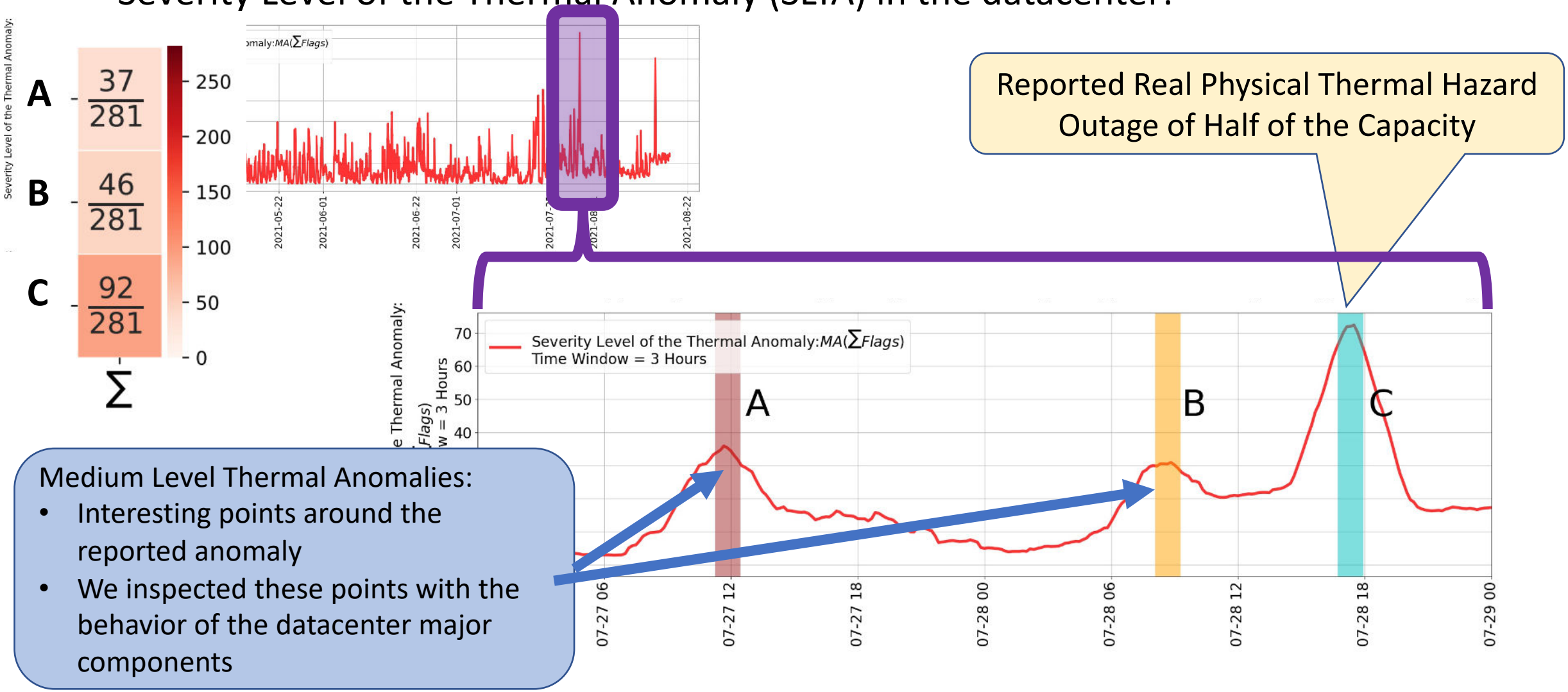
Reported Real Physical Thermal Hazard



SLTA Highlights Successfully the Reported Real Physical Thermal Hazard

Severity Level of Anomaly MA($\Sigma Flags$)

- Severity Level of the Thermal Anomaly (SLTA) in the datacenter:



Medium Level Thermal Anomalies:

- Interesting points around the reported anomaly
- We inspected these points with the behavior of the datacenter major components

Severity Level of Anomaly: $MA(\sum Flags)$

Severity Level of the Thermal Anomaly:

Sysadmins need: a tool to highlight the components which lead to the thermal anomaly

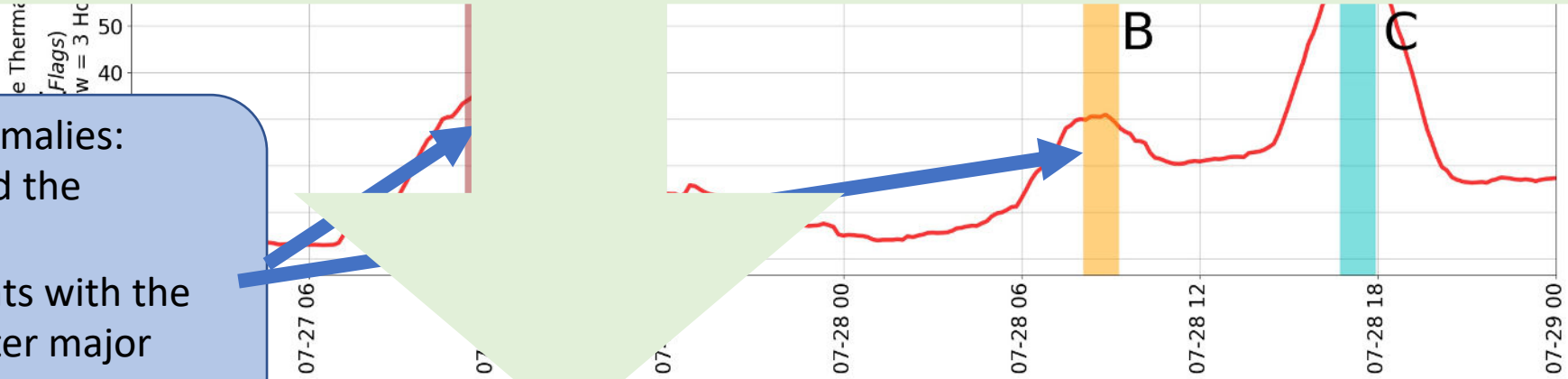
- Visual inspection of each of these three conditions.
- Introduce per components severity level of the anomaly, which can identify the sources of the anomalies

Automated Approach
Per Components

A

B

C

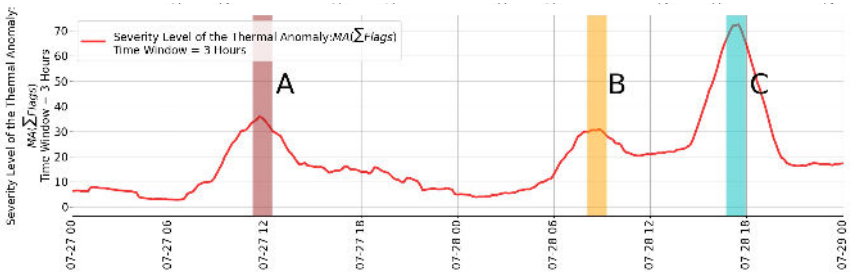


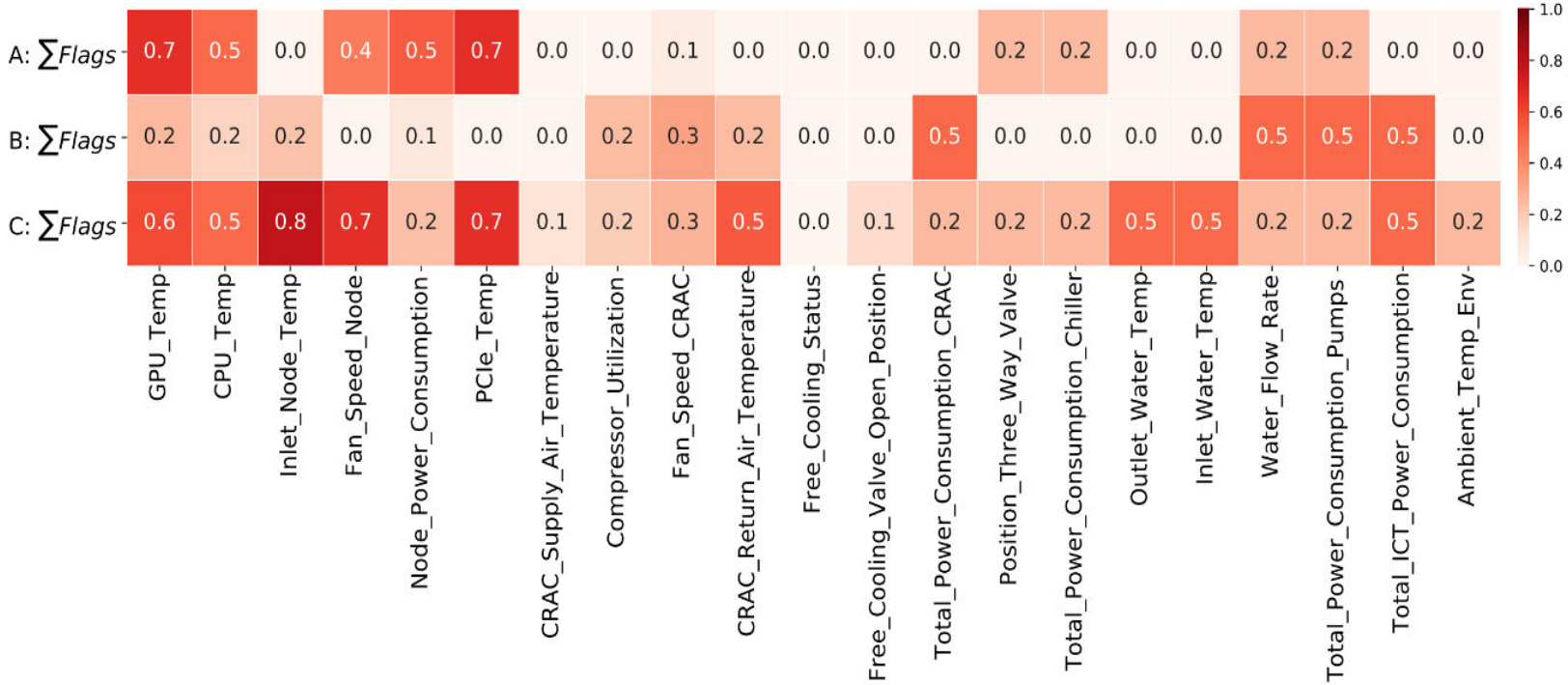
Medium Level Thermal Anomalies:

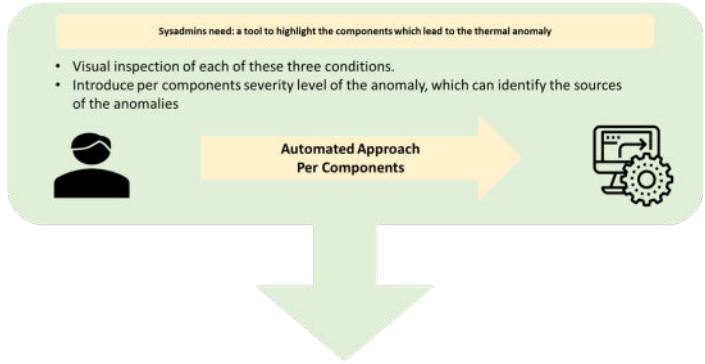
- Interesting points around the reported anomaly
- We inspected these points with the behavior of the datacenter major components

Thermal Anomaly Severity Level Percomponent

Locations of Anomalies (The annotation is a normalized number)







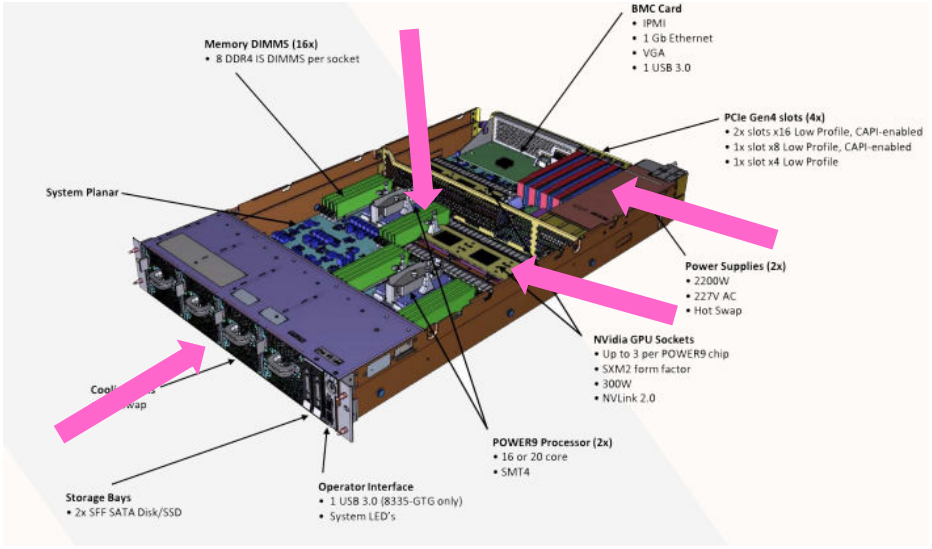
Outline

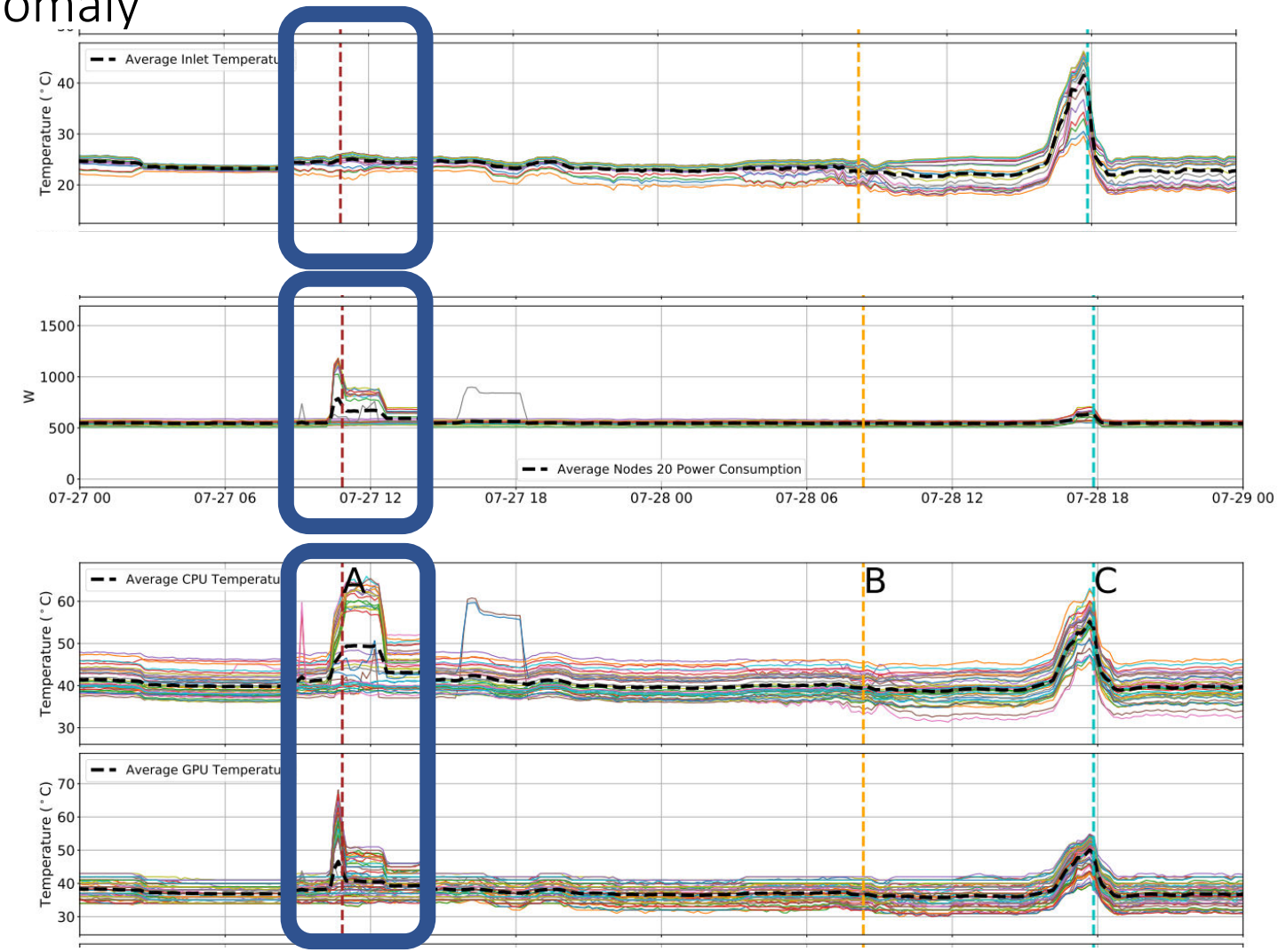
1. Introduction
2. Contributions
3. **Experimental Results**
4. Conclusions and Future Works

Experimental Results

Detailed Study of Real Physical Failure to Understand the Reasons Behind the High Severity Level of Anomaly

- Point A:
  - Nodes' inlet temperatures are normal
  - Computing loads are high
  - Reaction of the cooling systems is not fast enough to support computing load, which turns into high temperature at nodes level.

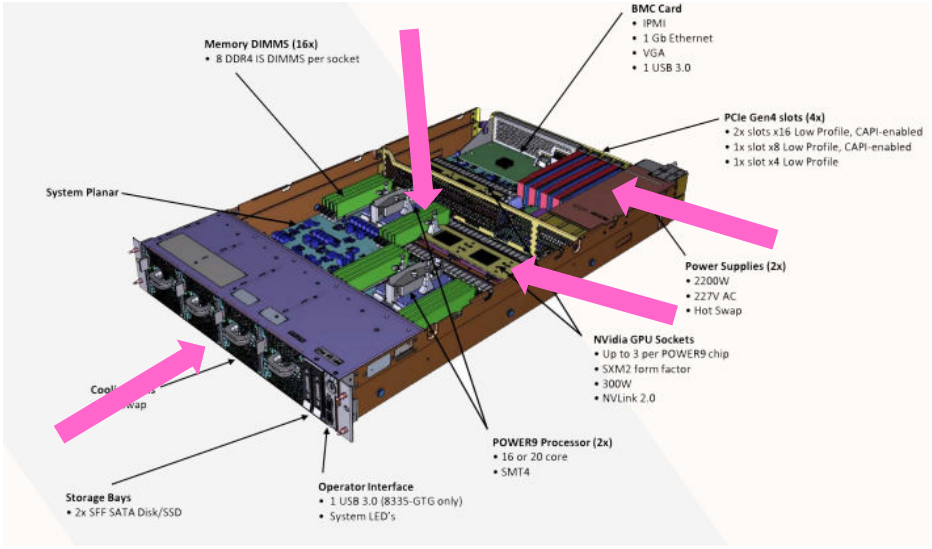


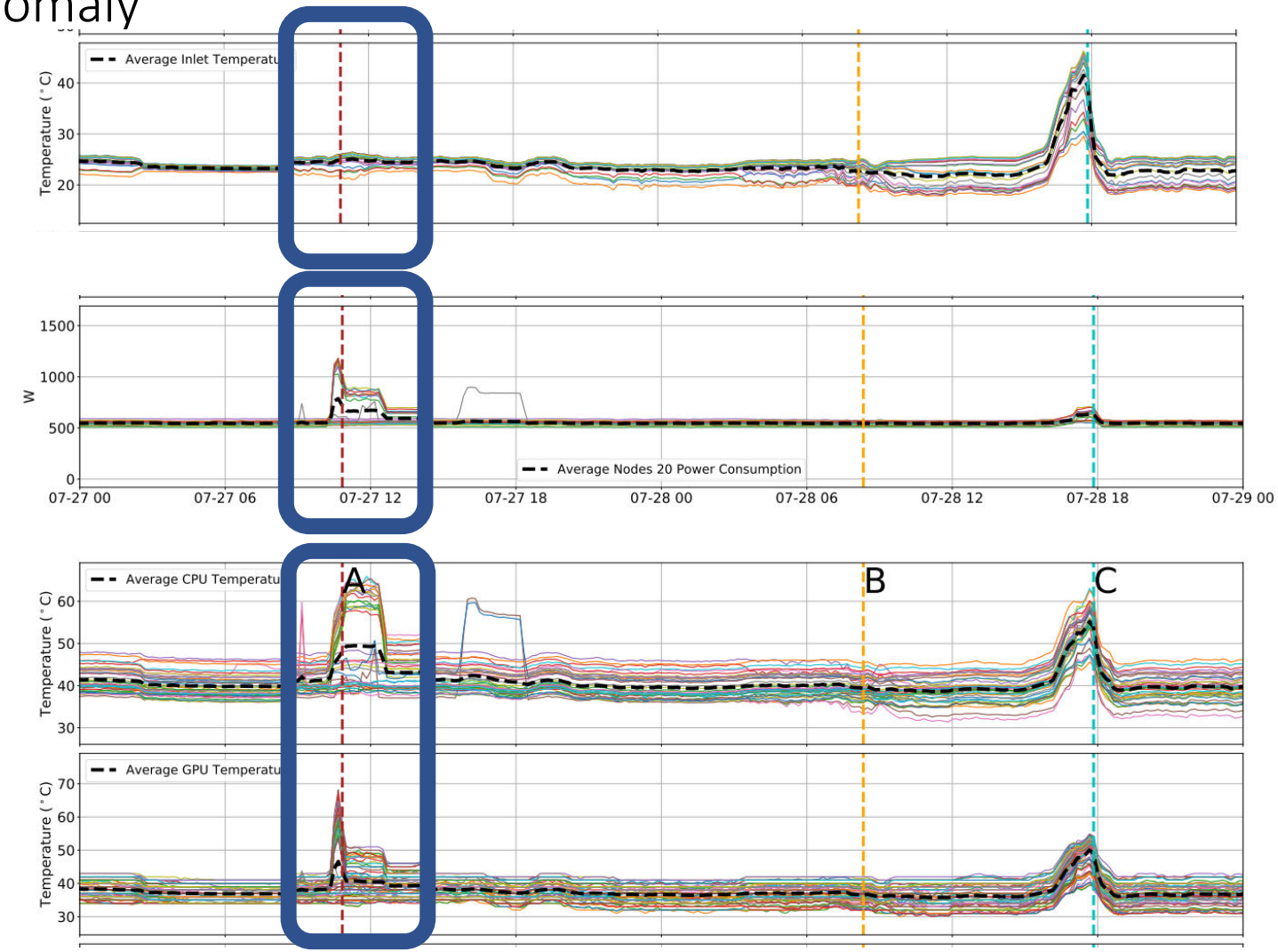


Experimental Results

Detailed Study of Real Physical Failure to Understand the Reasons Behind the High Severity Level of Anomaly

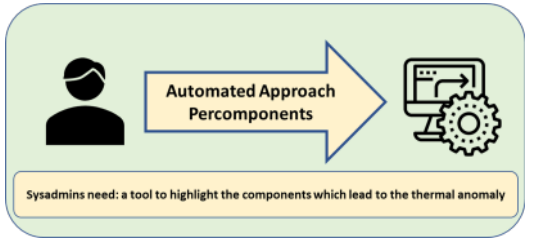
- Point A:
  - Nodes' inlet temperatures are normal
  - Computing loads are high
  - Reaction of the cooling systems is not fast enough to support computing load, which turns into high temperature at nodes level.

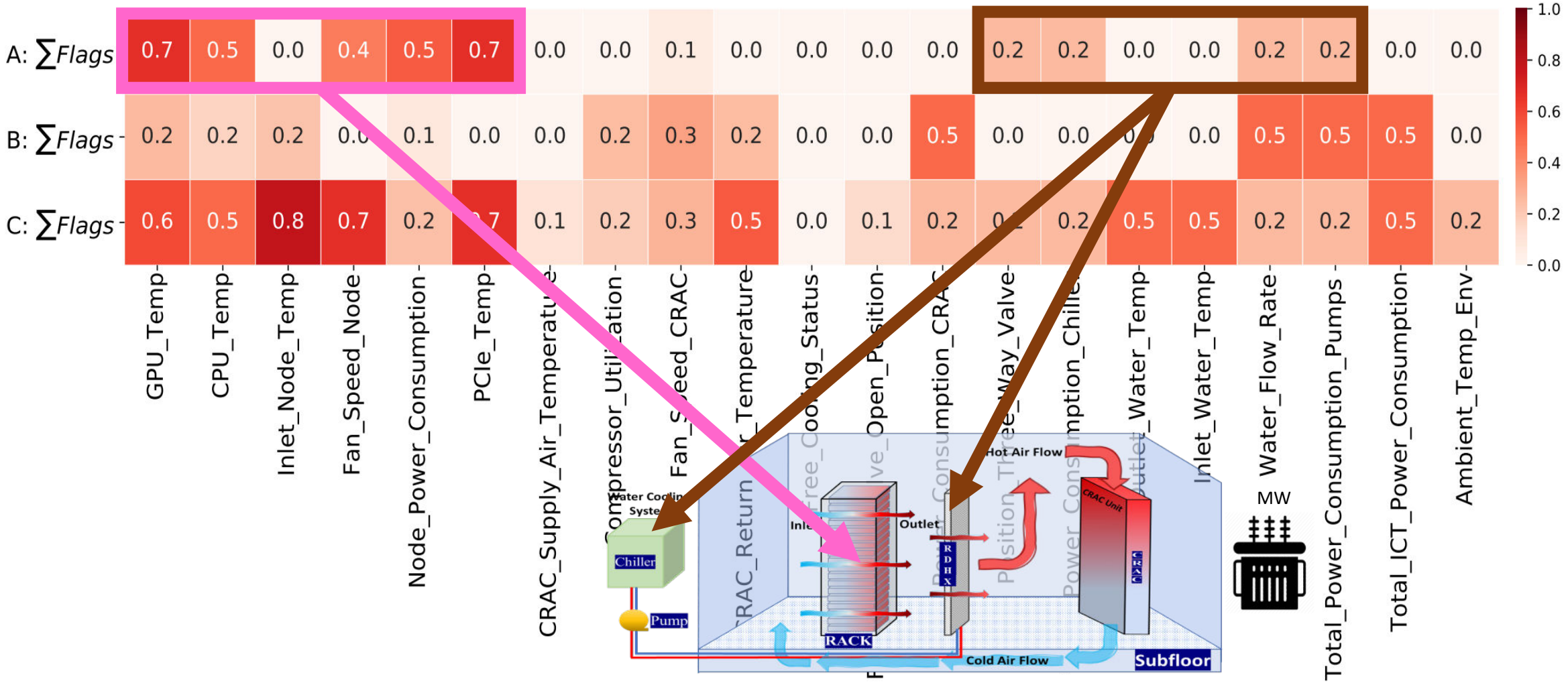




Thermal Anomaly Severity Level Percomponent

Locations of Anomalies (The annotation is a normalized number)

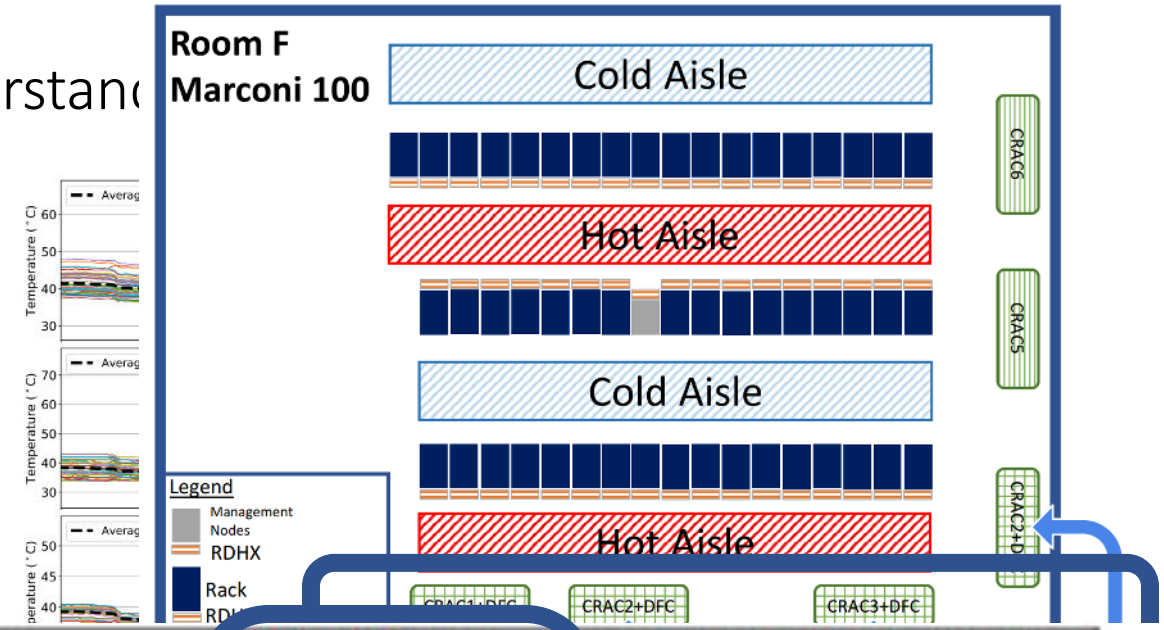


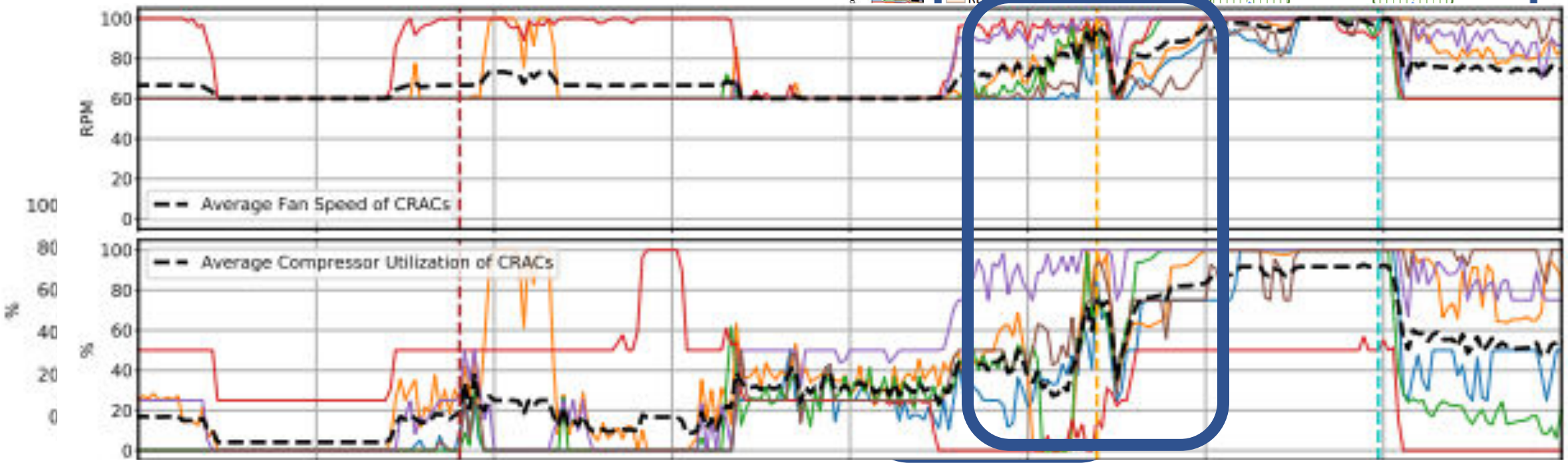


Experimental Results

Detailed Study of Real Physical Failure to Understand Behind the High Severity Level of Anomaly

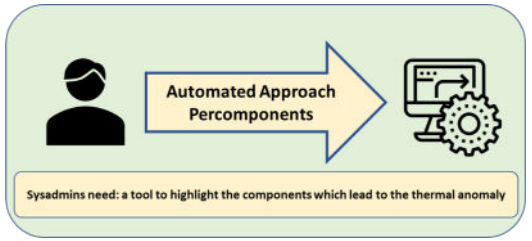
- Point B:
  - Nodes' level parameters are normal
  - Activation of free cooling is the primary source of signals' fluctuations in cooling systems
  - Signals' fluctuations as a suspicious condition.

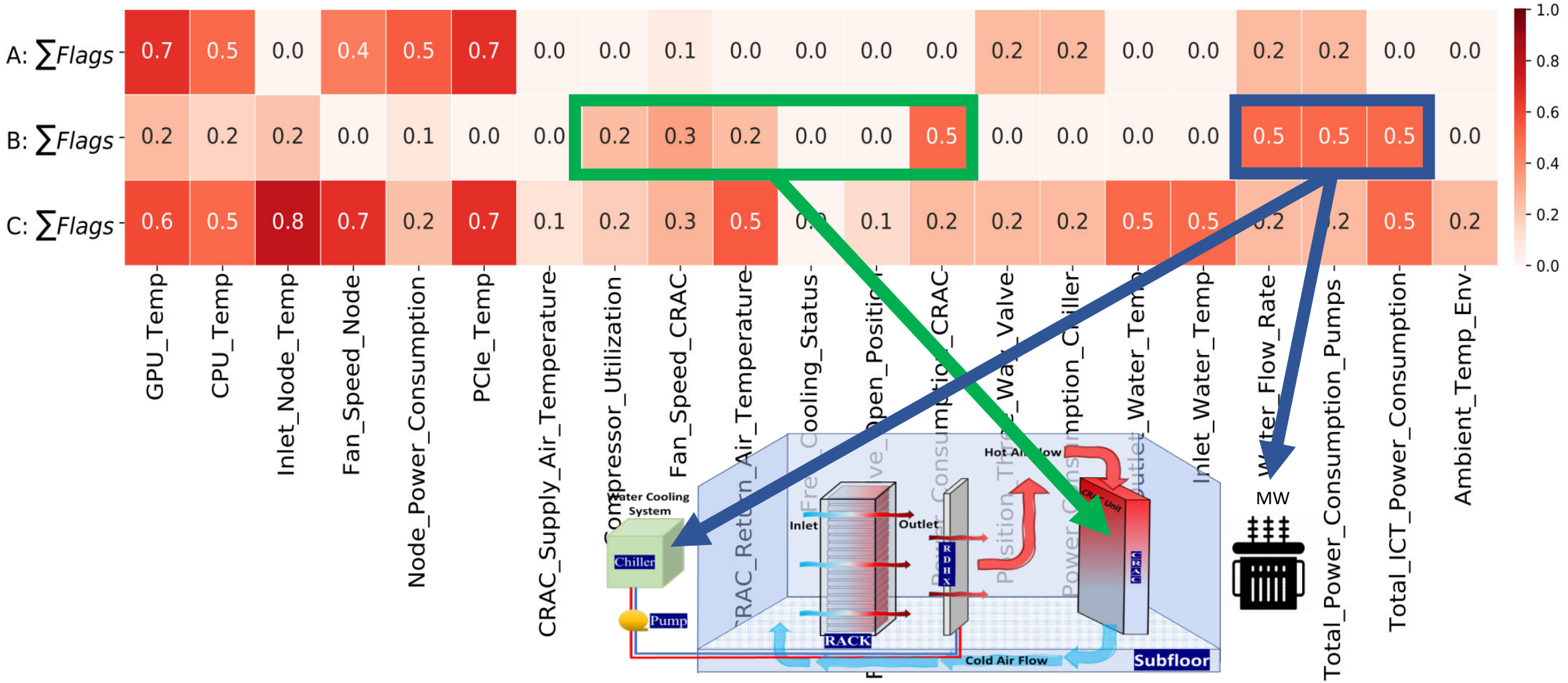




Thermal Anomaly Severity Level Percomponent

Locations of Anomalies (The annotation is a normalized number)

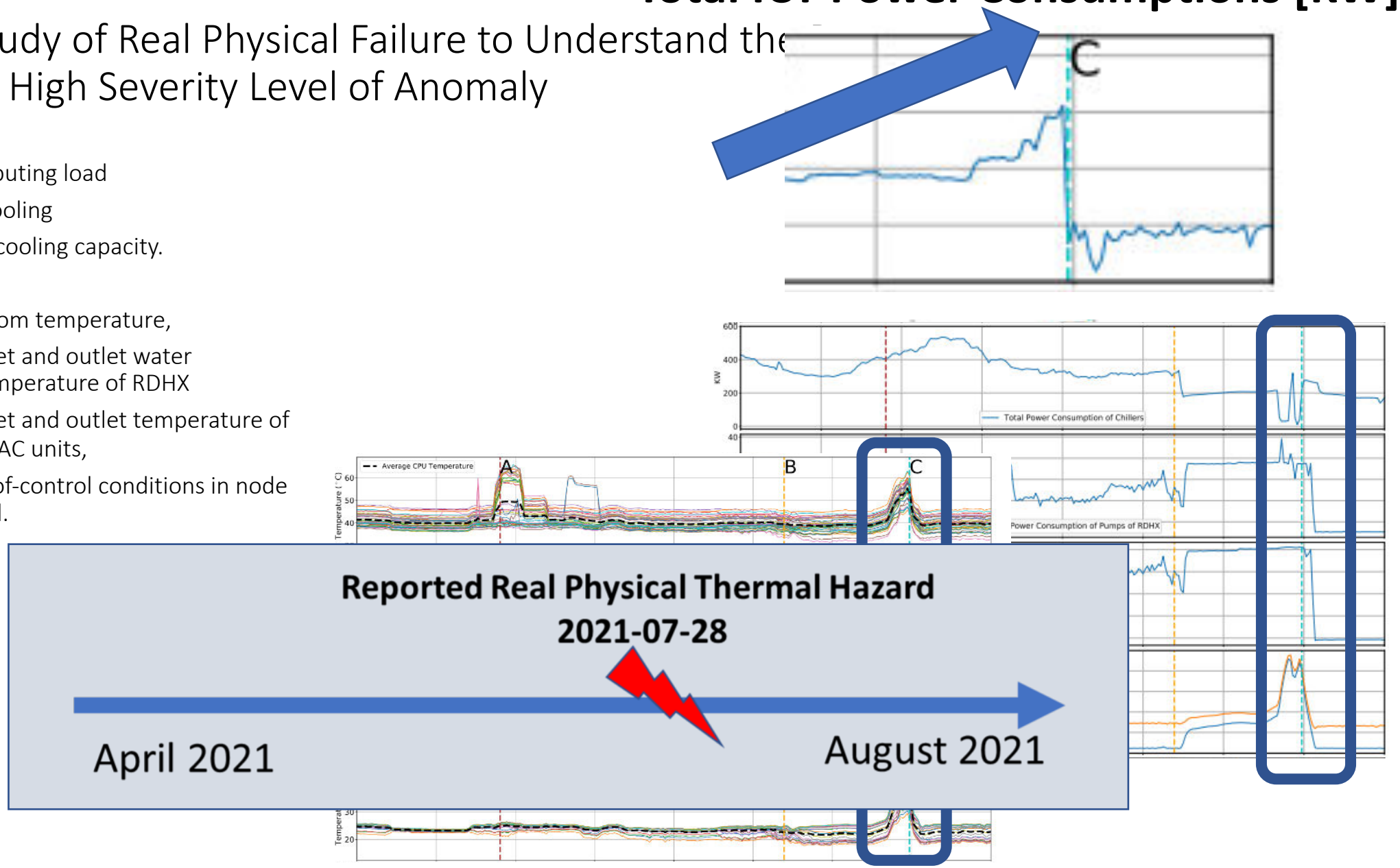




Experimental Results

Detailed Study of Real Physical Failure to Understand the Behind the High Severity Level of Anomaly

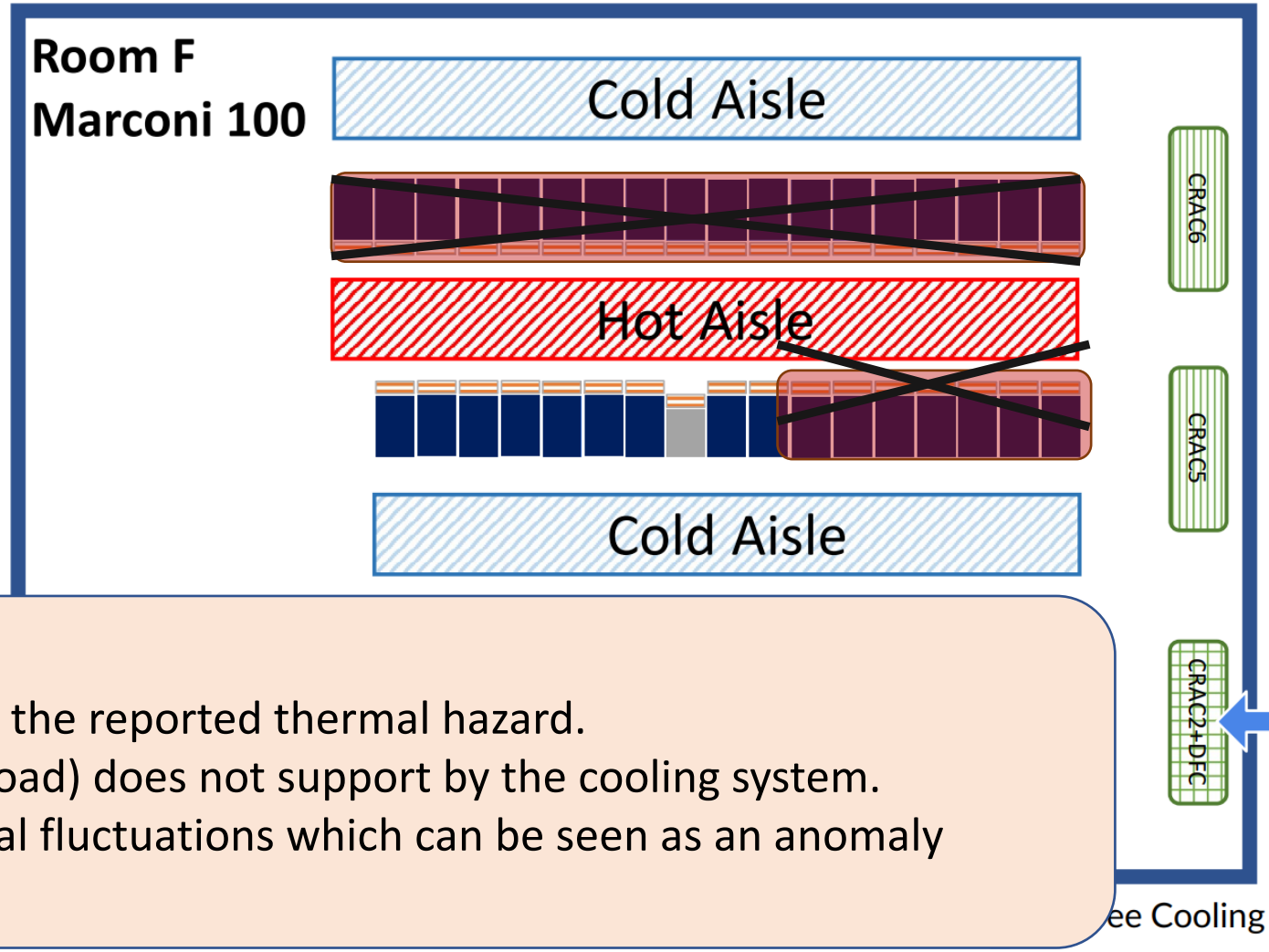
- Point C:
  - Increasing the computing load
  - Activation of free cooling
  - Reduction in RDHX cooling capacity.
  - Which increase:
    1. Room temperature,
    2. Inlet and outlet water temperature of RDHX
    3. Inlet and outlet temperature of CRAC units,
  - which leads to out-of-control conditions in node level and room level.



Experimental Results

Detailed Study of Real Physical Failure to Understand the Reasons Behind the High Severity Level of Thermal Anomaly

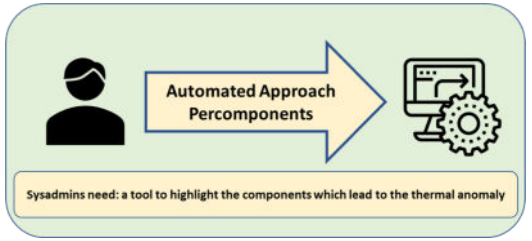
- Point C:
  - Increasing the computing load
  - Activation of free cooling
  - Reduction in RDHX cooling capacity.
  - Which increase:
    1. Room temperature,
    2. Inlet and outlet water temperature of RDHX
    3. Inlet and outlet temperature of CRAC units,
  - which leads to out-of-control conditions in node level and room level.



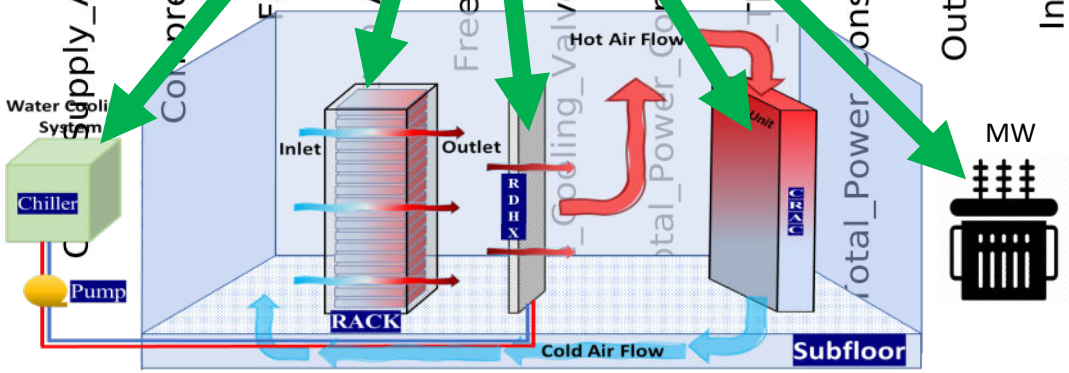
- The SLTA method is practical.
- SLTA's peak during the 4 months is in the reported thermal hazard.
- Fast variation of the IT (Computing Load) does not support by the cooling system.
- Free Cooling System: can create signal fluctuations which can be seen as an anomaly

Thermal Anomaly Severity Level Percomponent

Locations of Anomalies (The annotation is a normalized number)



	GPU_Temp	CPU_Temp	Inlet_Node_Temp	Fan_Speed_Node	Node_Power_Consumption	PCIe_Temp	Supply_Air_Temperature	Compressor_Utilization	Fan_Speed_CRAC	Air_Temperature	Free_Cooling_Status	Cooling_Valve_Open_Position	Power_Consumption_CRAC	Three_Way_Valve	Total_Power_Consumption_Chiller	Outlet_Water_Temp	Inlet_Water_Temp	Water_Flow_Rate	Total_Power_Consumption_Pumps	Total_ICT_Power_Consumption	Ambient_Temp_Env
A: ∑Flags	0.7	0.5	0.0	0.4	0.5	0.7	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.2	0.2	0.0	0.0
B: ∑Flags	0.2	0.2	0.2	0.0	0.1	0.0	0.0	0.2	0.3	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.5	0.5	0.5	0.0
C: ∑Flags	0.6	0.5	0.8	0.7	0.2	0.7	0.1	0.2	0.3	0.5	0.0	0.1	0.2	0.2	0.2	0.5	0.5	0.2	0.2	0.5	0.2



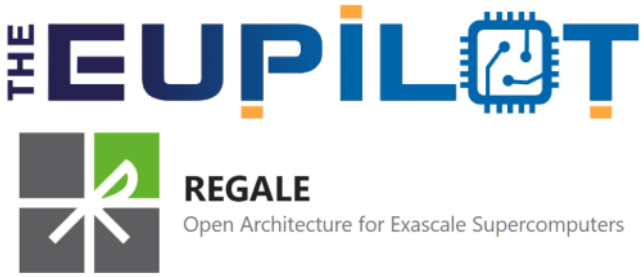
Outline

1. Introduction
2. Contributions
3. Experimental Results
4. Conclusions and Future Works

Conclusions and Future Works

- Complete Dataset: Normal and Abnormal
- Rule-based statistical methods (flags):
  - Explore different metrics at the datacenter, system, sub-system, and compute node level
  - Severity Level of the Thermal Anomaly (SLTA) in the datacenter MA($\Sigma Flags$)
  - Threshold Definition Methodology
  - Method Successful Validation Against Reported Real Physical Thermal Anomaly
  - Method Highlight the location of the anomalies
    - Thermal Anomaly Severity Level Per Component
- High potential in maintenance and troubleshooting.
- Future Work:
  - This method can extend to other kinds of anomalies (like application-level anomaly detection, etc.)
  - For more sophisticated ML classification methods which rely on the label
  - Or in the semi-supervised method, which relies on a normal part of the dataset
  - We are working to remedy flags' weakness in analyzing the complicated correlation of the signals in finding the anomalies or suspicious patterns by employing a semi-supervised ML-based approach to improve anomaly detection performance.

Acknowledgment

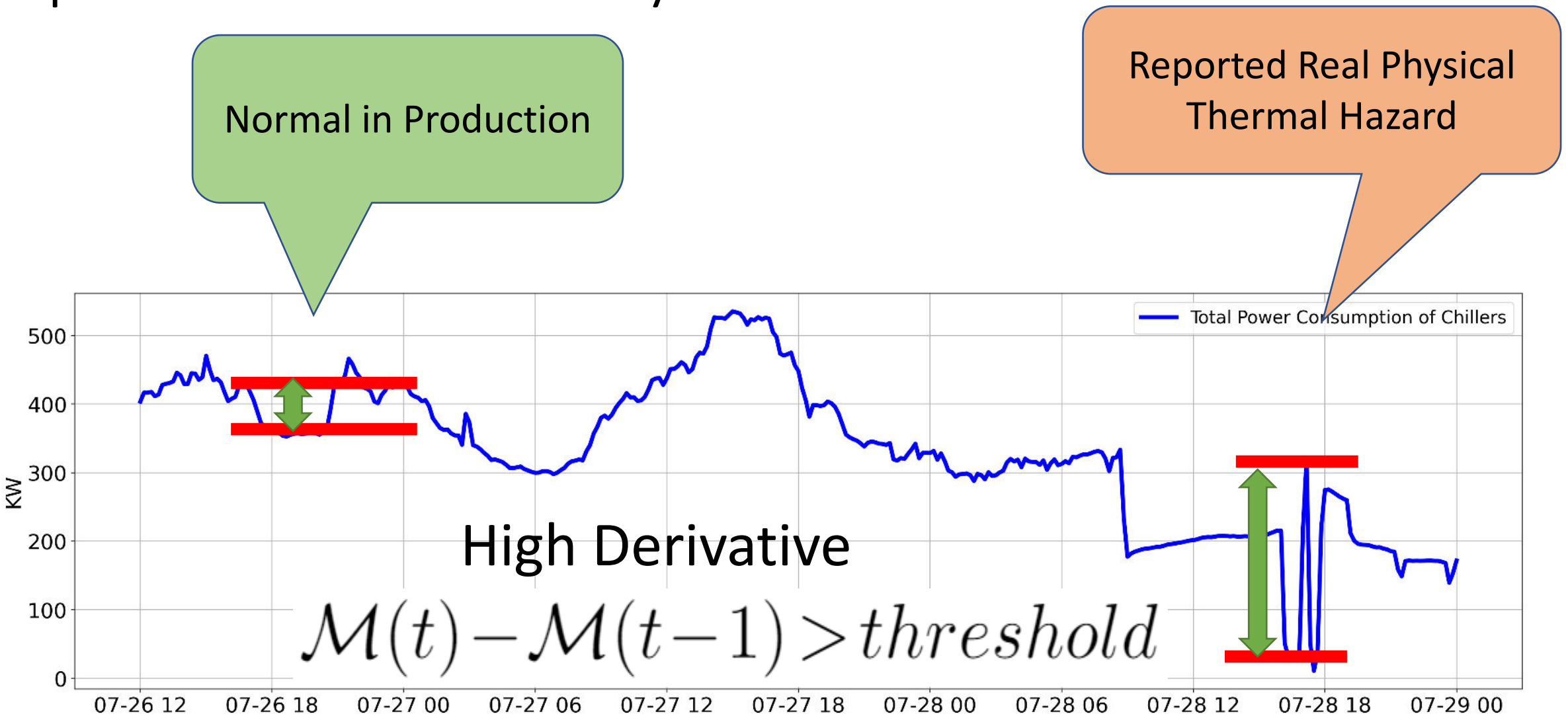


- The European-project initiative has received funding from the European High-Performance Computing Joint Undertaking (JU) under Framework Partnership Agreement No 800928 and Specific Grant Agreement No 101036168 (EPI SGA2). The JU receives support from the European Union's Horizon 2020 research and innovation program and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland.
- The European PILOT project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No.101034126. The JU receives support from the European Union's Horizon 2020 research and innovation program and Spain, Italy, Switzerland, Germany, France, Greece, Sweden, Croatia, and Turkey.
- This REGALE-project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956560. The JU receives support from the European Union's Horizon 2020 research and innovation program and Greece, Germany, France, Spain, Austria, and Italy.

Thank You!
Questions?

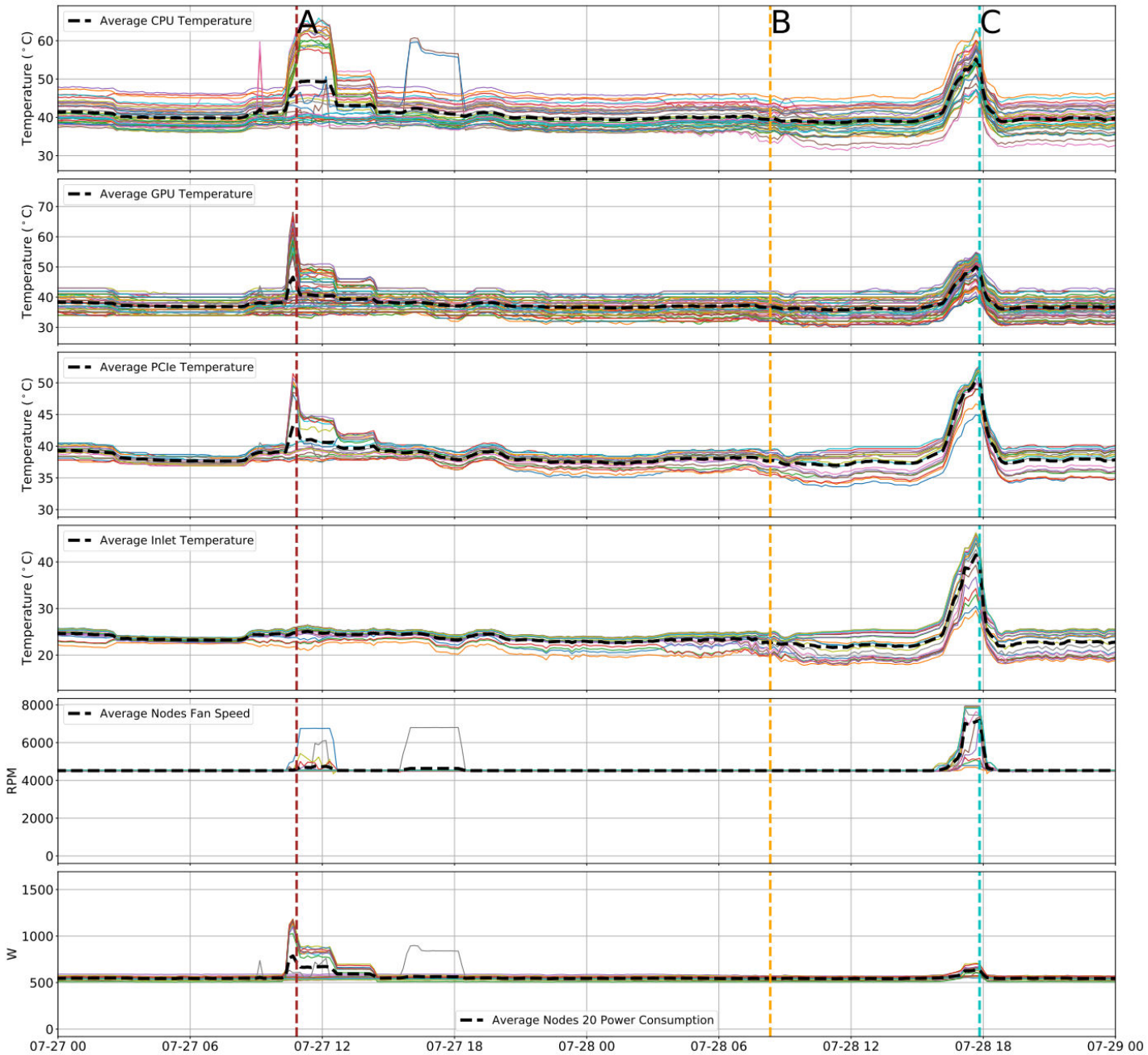
mohsen.seyedkazemi@unibo.it

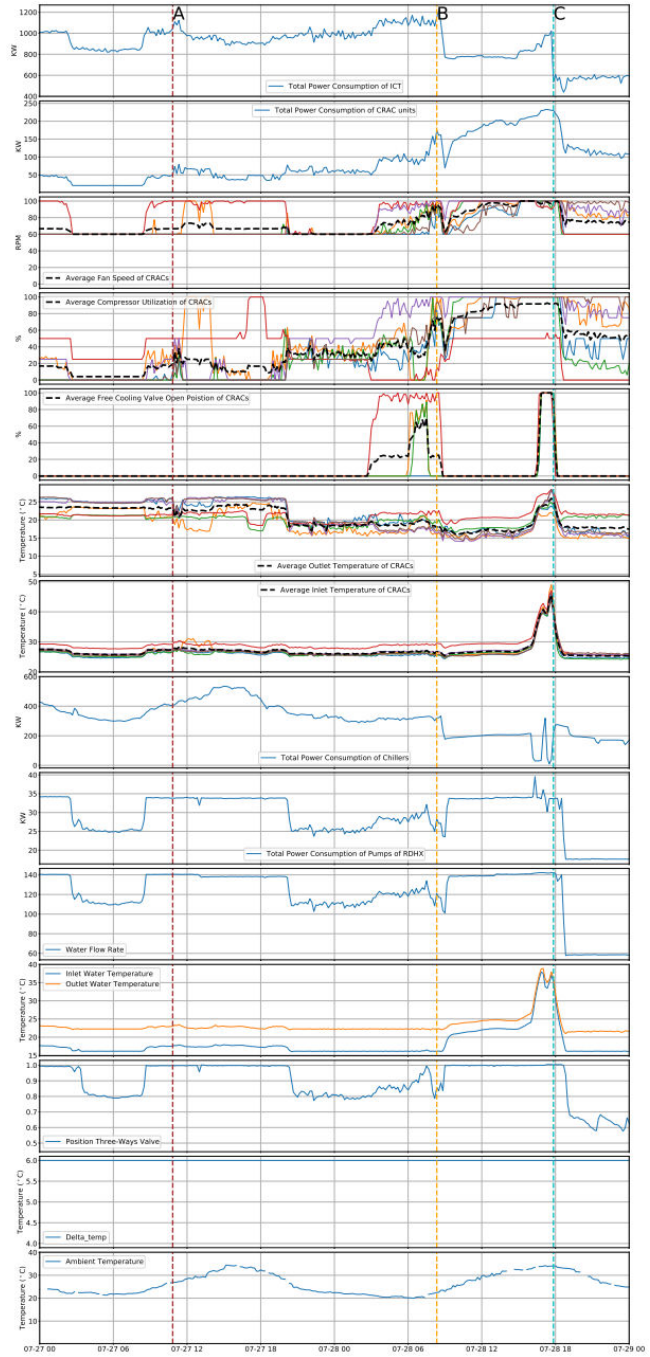
Reported Failure Study



Backup Slides

Name of Parameter	Value
processor	2x16 cores IBM POWER9
Memory Per Node	256 GB
Peak Performance	~32 PFlop/s
Number of Racks	55 total (49 compute)
Number of Chassis Per Rcak	20
Number of Nodes in Room	980
Sampling Rate	20 Second
Accelerators	4 x NVIDIA Volta V100 GPUs
Thermal Hazard	2021-07-28


A
B
C
Average CPU Temperature
Temperature (°C)
Average GPU Temperature
Temperature (°C)
Average PCIe Temperature
Temperature (°C)
Average Inlet Temperature
Temperature (°C)
Average Nodes Fan Speed
RPM
Average Nodes 20 Power Consumption
W
07-27 00
07-27 06
07-27 12
07-27 18
07-28 00
07-28 06
07-28 12
07-28 18
07-29 00


A
B
C
Total Power Consumption of ICT
KW
Total Power Consumption of CRAC units
KW
Average Fan Speed of CRACs
RPM
Average Compressor Utilization of CRACs
%
Average Free Cooling Valve Open Position of CRACs
%
Average Outlet Temperature of CRACs
Temperature (°C)
Average Inlet Temperature of CRACs
Temperature (°C)
Total Power Consumption of Chillers
KW
Total Power Consumption of Pumps of RDHX
KW
Water Flow Rate
Inlet Water Temperature
Outlet Water Temperature
Temperature (°C)
Position Three-Ways Valve
Delta_temp
Temperature (°C)
Ambient Temperature
Temperature (°C)
07-27 00
07-27 06
07-27 12
07-27 18
07-28 00
07-28 06
07-28 12
07-28 18
07-29 00

Sysadmins need: a tool to highlight the components which lead to the thermal anomaly

- Visual inspection of each of these three conditions.
- Introduce per components severity level of the anomaly, which can identify the sources of the anomalies

Automated Approach
Per Components