ICSE 2025
47th International Conference on Software Engineering
Sun 27 April - Sat 3 May 2025
Ottawa, Ontario, Canada

BSODiag: A Global Diagnosis Framework for Batch Servers Outage in Large-scale Cloud Infrastructure Systems

Tao Duan, Runqing Chen, Pinghui Wang*, Junzhou Zhao*,
Jiongzhou Liu, Shujie Han, Yi Liu and Fan Xu
Xi'an Jiaotong University, Alibaba Cloud Intelligence Group

Outline

☞ ☐ **Background**

☐ Empirical Observations & Problem Formulation

☐ Methodology Design

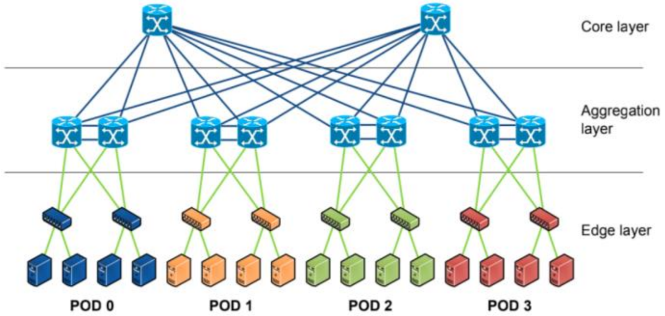
☐ Evaluation

☐ Conclusion

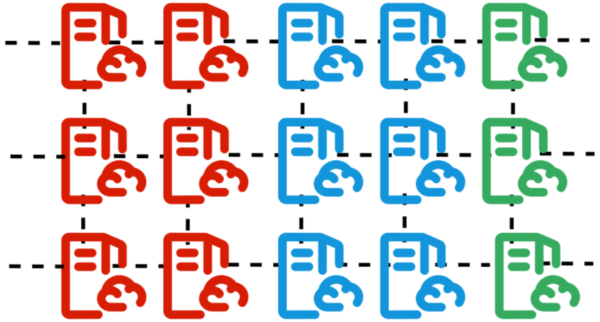
Background

- Cloud Infrastructure Systems (CIS)

Internet Data Center (IDC)



Cloud Networking

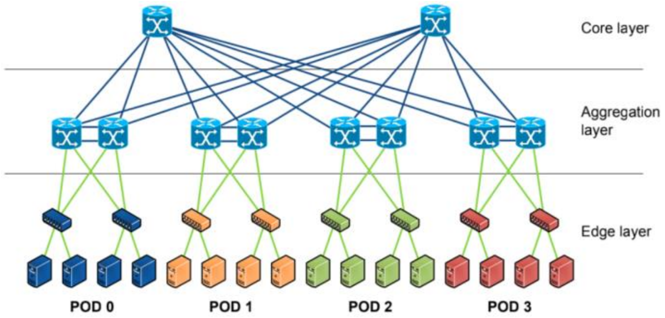


Cloud Servers

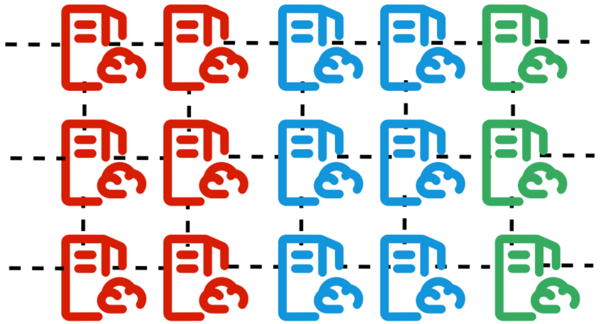
Background

Cloud Infrastructure Systems (CIS)

Internet Data Center (IDC)



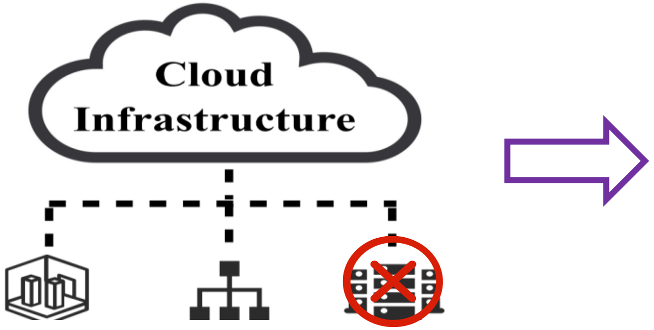
Cloud Networking



Cloud Servers

Batch Servers Outage in CIS

- **Batch Servers Outage**: Simultaneous breakdown of a cluster of related servers

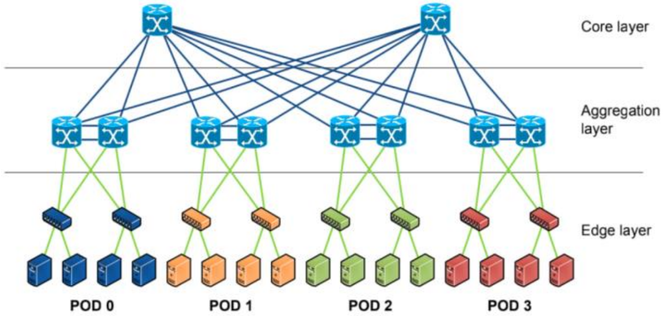


- Services Catastrophic Interruption
- Networking Outage
- …

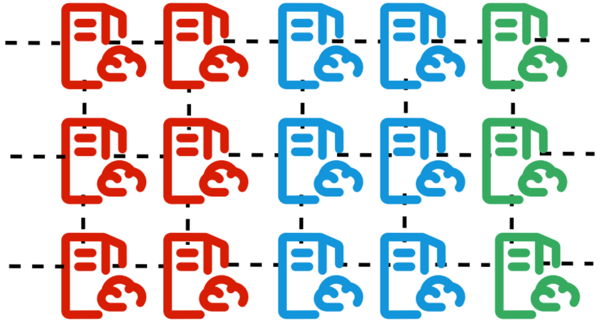
Background

Cloud Infrastructure Systems (CIS)

Internet Data Center (IDC)



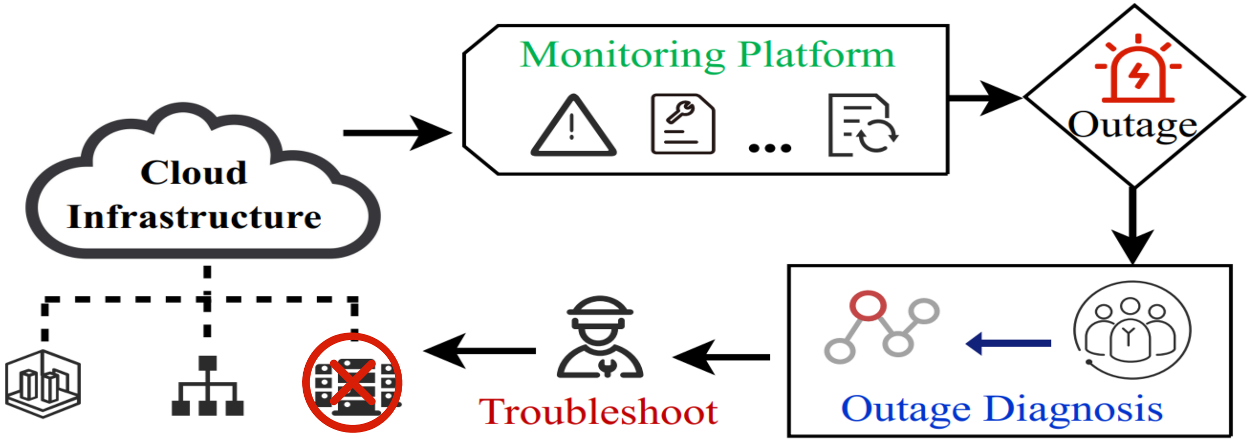
Cloud Networking



Cloud Servers

Batch Servers Outage in CIS

- The life cycle of a batch servers outage diagnosis



Outline

Empirical Observations

- RQ1: Can monitoring data collected in a CIS adequately describe failures, if not, **how to obtain a more comprehensive failure profiling**?

ID	Occurence Time	Device SN	Anomaly Type	Anomaly Content
66	22/03/25 21:12:06	XX_632	server temperature anomaly	temperature: 42.5℃
67	22/03/25 21:12:07	XX_225	circuit group interrupt	partial interrupt
68	22/03/25 21:13:22	XX_764	network device state anomaly	psw offline
69	22/03/25 21:13:22	XX_046	high cpu utilization	cpu utilization: 77%
69	22/03/25 21:13:22	XX_046	high cpu utilization	cpu utilization: 82%

An Alert Sequence

Batch Servers Outage Incident

Incident ID: 1633525
Occurence Time: 22/03/25 21:33:26 - 22/03/25 21:57:48
Location: RACK: R.43-A.22-HZ.116
Relative Devices SN: XX_6389, XX_7200, XX_1573,XX_4532,...
Description: 26 servers are unreachable, suspected to be a batch servers outage failure.

A Servers Outage Incident

AC Refrigerant Replace Change

Change ID: 23665
Operation Time: 22/03/25 18:25:30 - 22/03/25 18:27:55
Location: ROOM-A.22-HZ.116
Relative Devices SN: XX_302,XX_306
Change Content: Replace the refrigerant of the air conditioner
Change Reason:Abnormal cooling of air conditioner

A Refrigerant Replace Change

Empirical Observations

- RQ1: Can monitoring data collected in a CIS adequately describe failures, if not, **how to obtain a more comprehensive failure profiling**?
  - Analysis of monitoring data quality:

Table I: Analysis of monitoring data quality.

Failure Type	Incident	Change	Alert	#Failures
Switch Reboot	✓			4
Temperature Anomaly	✓		✓	126
Refrigerant Replacing		✓		1
PSU Power Outage	✓			2
High CPU Utilization			✓	305
Partial Network Loss			✓	206

- Genuine failures related outage
- Irrelevant failures

Empirical Observations

- RQ1: Can monitoring data collected in a CIS adequately describe failures, if not, **how to obtain a more comprehensive failure profiling**?
  - Analysis of monitoring data quality:

Table I: Analysis of monitoring data quality.

Failure Type	Incident	Change	Alert	#Failures
Switch Reboot	✓			4
Temperature Anomaly	✓		✓	126
Refrigerant Replacing		✓		1
PSU Power Outage	✓			2
High CPU Utilization			✓	305
Partial Network Loss			✓	206

- Alert flooding

Empirical Observations

- RQ1: Can monitoring data collected in a CIS adequately describe failures, if not, **how to obtain a more comprehensive failure profiling**?
  - Analysis of monitoring data quality:

Table I: Analysis of monitoring data quality.

Failure Type	Incident	Change	Alert	#Failures
Switch Reboot	✓			4
Temperature Anomaly	✓		✓	126
Refrigerant Replacing		✓		1
PSU Power Outage	✓			2
High CPU Utilization			✓	305
Partial Network Loss			✓	206

- Repeat report
- Omission report

Empirical Observations

- RQ1: Can monitoring data collected in a CIS adequately describe failures, if not, **how to obtain a more comprehensive failure profiling**?
  - Analysis of monitoring data quality:

Table I: Analysis of monitoring data quality.

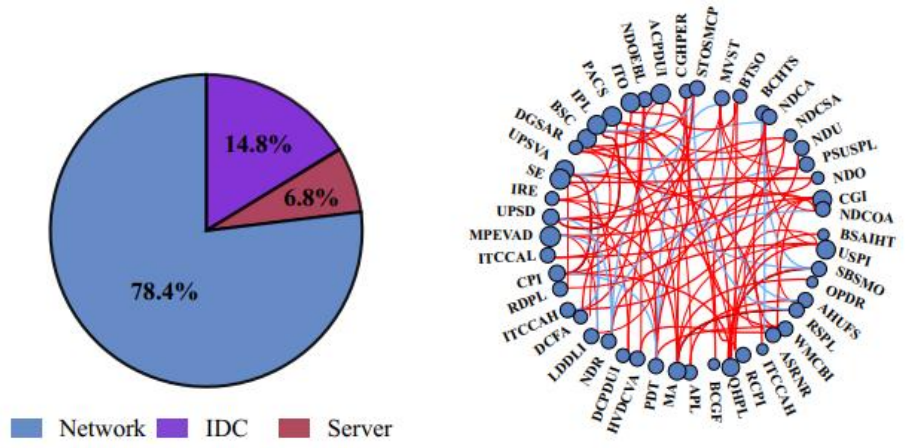
Failure Type	Incident	Change	Alert	#Failures
Switch Reboot	✓			4
Temperature Anomaly	✓		✓	126
Refrigerant Replacing		✓		1
PSU Power Outage	✓			2
High CPU Utilization			✓	305
Partial Network Loss			✓	206

Single-source monitoring data are insufficient to reveal all suspicious failures, synchronous analysis of multi-source monitoring data is imperative.

Empirical Observations

☐ RQ2: What is the cause of batch servers outage, and what is the **correlation mechanism between failures**?

- <u>Analysis of failure correlation:</u>

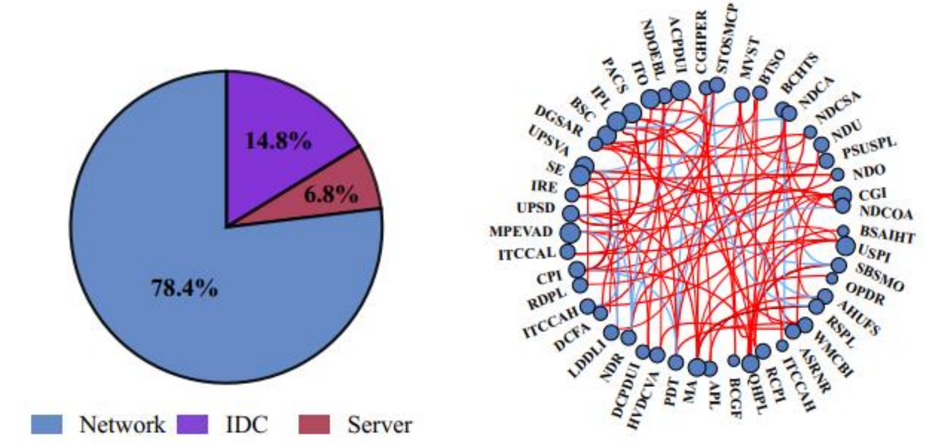


(a) Root Causes Distribution (b) Failure Correlation Patterns

- Cross-domain network failures and IDC failures are the primary root causes.
- Batch servers outage often results from concurrent multi-domains failures.

Empirical Observations

- RQ2: What is the cause of batch servers outage, and what is the **correlation mechanism between failures**?
  - Analysis of failure correlation:



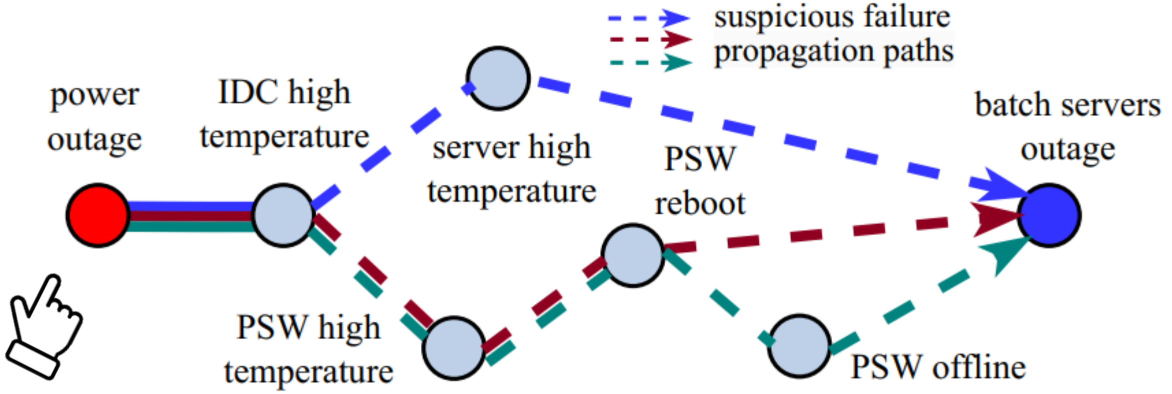
(a) Root Causes Distribution (b) Failure Correlation Patterns

- Cross-domain network failures and IDC failures are the primary root causes.
- Batch servers outage often results from concurrent multi-domains failures.

It is crucial to develop a failure correlation measurement technique that can model failure correlations from a global perspective.

Empirical Observations

- **RQ3:** RQ3: What are the necessary **diagnostic results for real-world applications**?
  - Analysis of Efficient Troubleshooting:
    - Root Case Location:

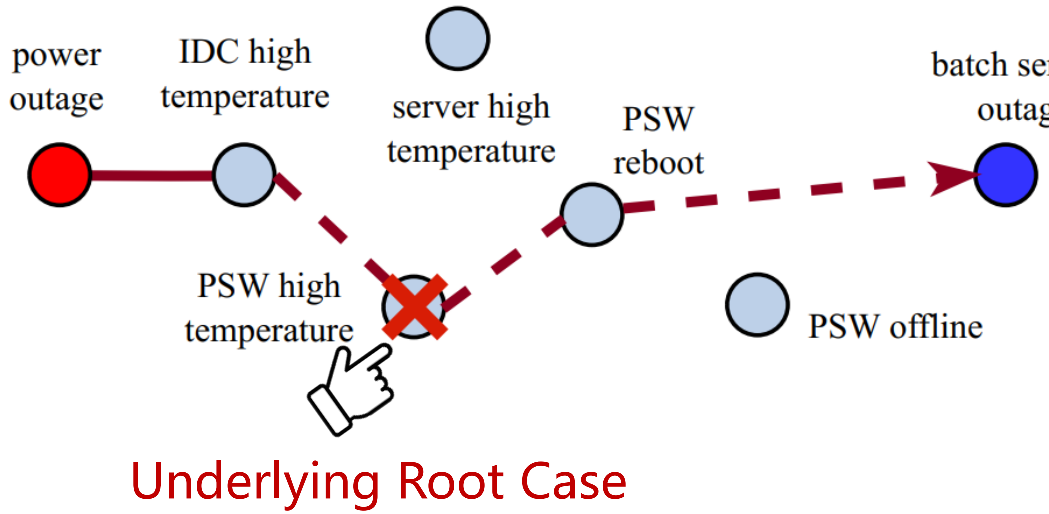


1-th Root Case

- However, solely on pinpointing the root cause is not enough!

Empirical Observations

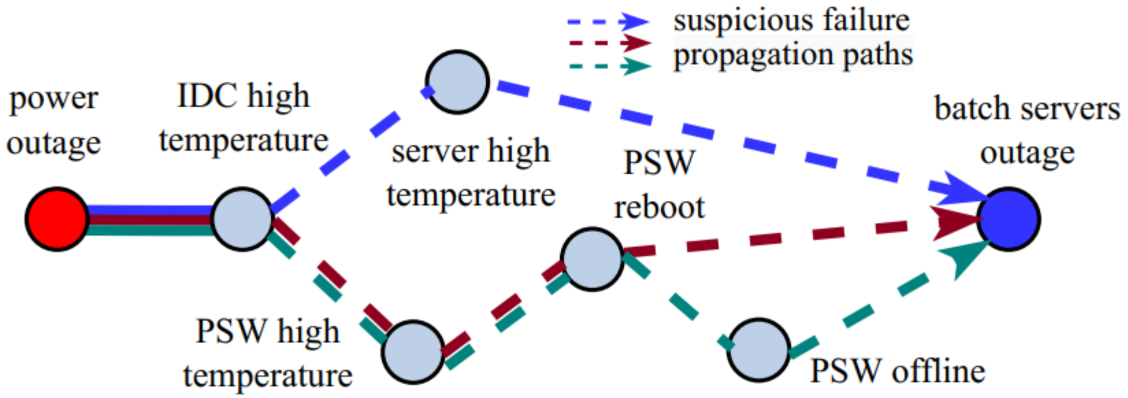
- RQ3: What are the necessary **diagnostic results for real-world applications**?
  - <u>Analysis of Efficient Troubleshooting:</u>
    - Failure propagation path inference:



- In fact, aging of the PSW is another underlying reason for this outage!

Empirical Observations

- RQ3: What are the necessary **diagnostic results for real-world applications**?
  - <u>Analysis of Efficient Troubleshooting:</u>



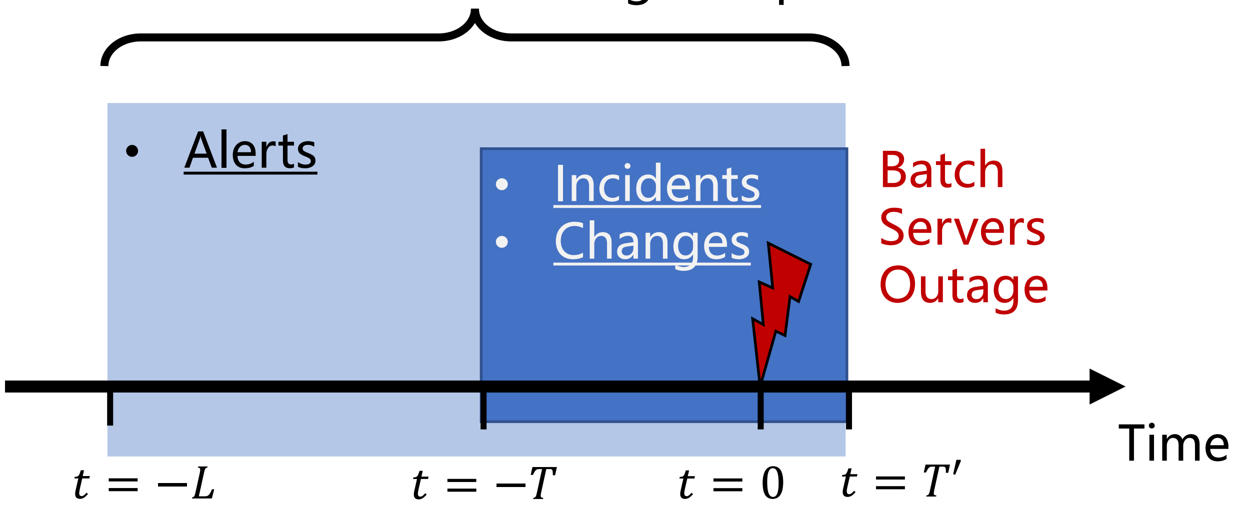
Providing the interpretable diagnosis results that include both root cause failure and failure propagation path is necessary for troubleshooting.

Problem Formulation

◻ The Batch Servers Outage Diagnosis Problem:

- Outage Snapshot:

outage snapshot $U \equiv \{\mathcal{A}_{-L:T'}, \mathcal{I}_{-T:T'}, \mathcal{C}_{-T:T'}\}$



- Failure detection sub-problem takes: detects all outage-related events E in U

$$\mathcal{F}: U \mapsto E = \{e_1, \ldots\}$$

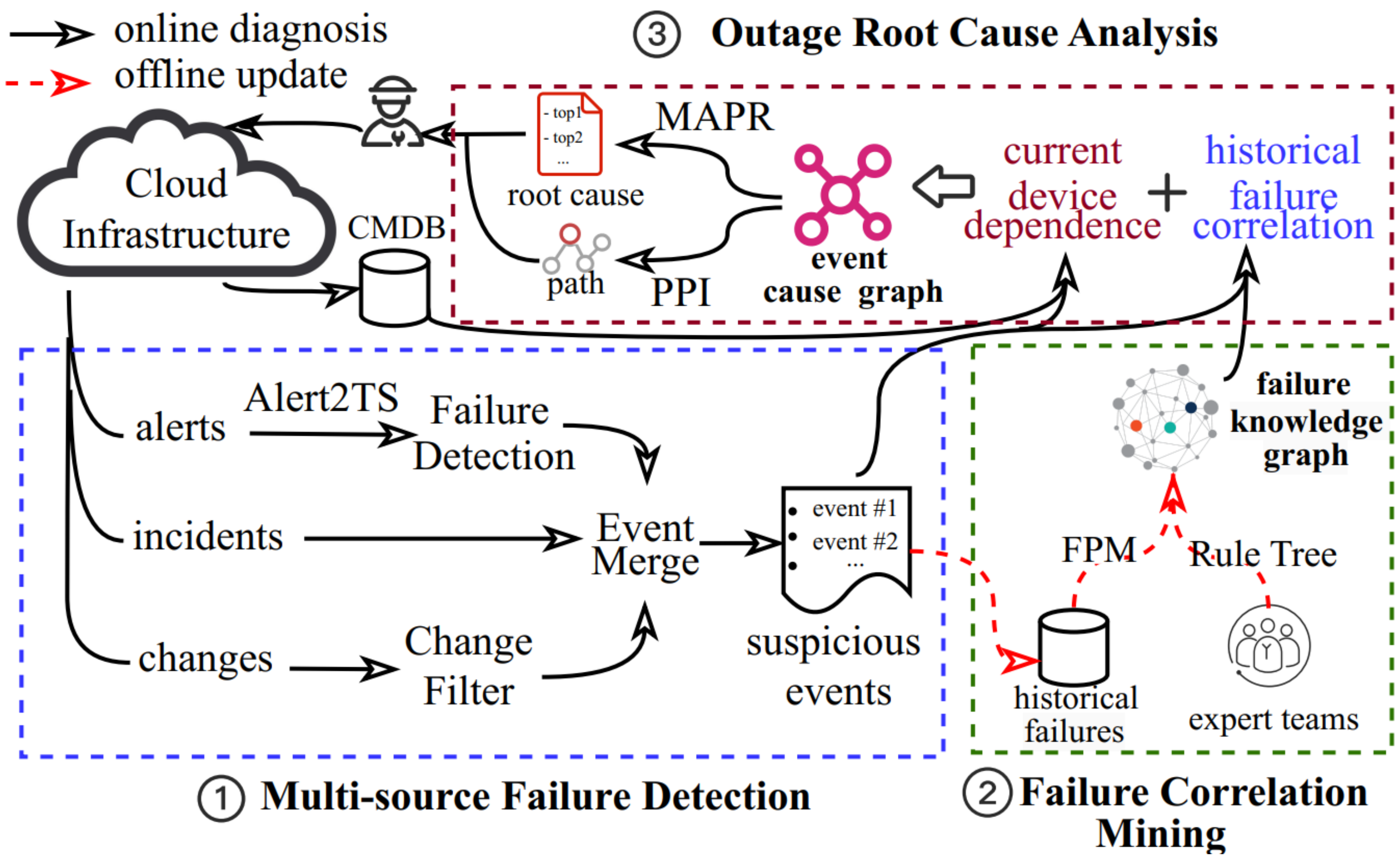
- Outage root cause analysis sub-problem takes: locates the root cause set e_r and infers the failure propagation path p_U

$$\mathcal{M}: E \mapsto \{e_r, p_U\}$$

Outline

- ☐ Background
- ☐ Empirical Observations & Problem Formulation
- ☞ ☐ Methodology Design
- ☐ Evaluation
- ☐ Conclusion

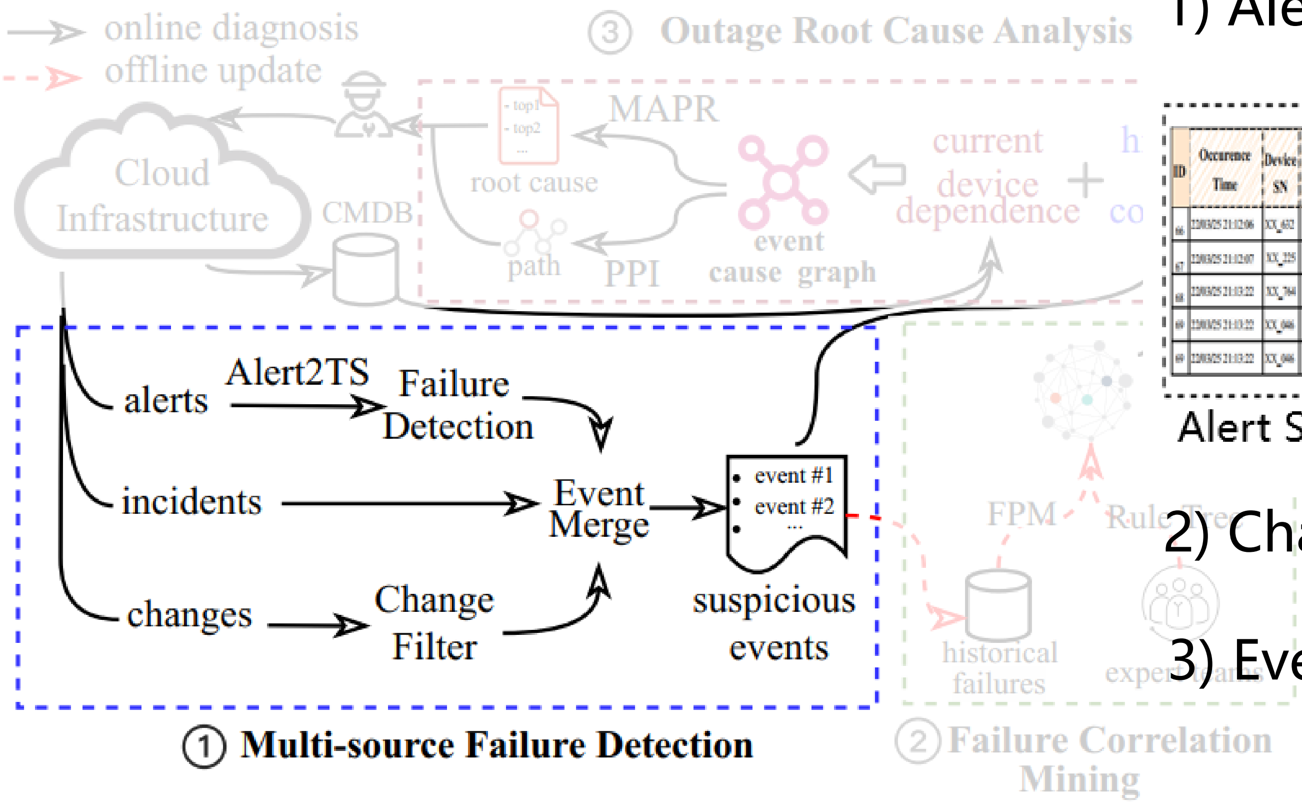
Overview



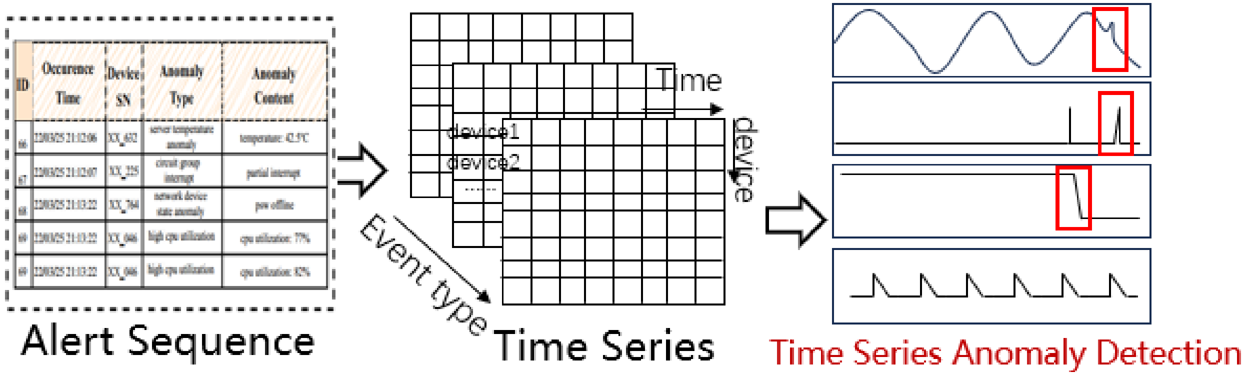
- **Multi-source Failure Detection**: detect outage-related failures from alerts, incidents, and changes.
- **Failure Correlation Mining**: discover the failure correlations reflected in historical data.
- **Outage Root Case Analysis**: delivers interpretable diagnostic results using event cause graph.

Model Detail

☐ Multi-source Failure Detection Module

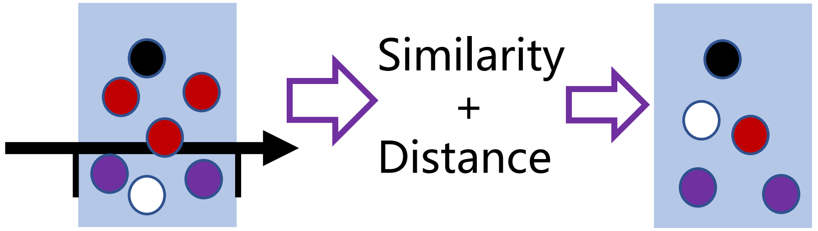


1) Alert2Event



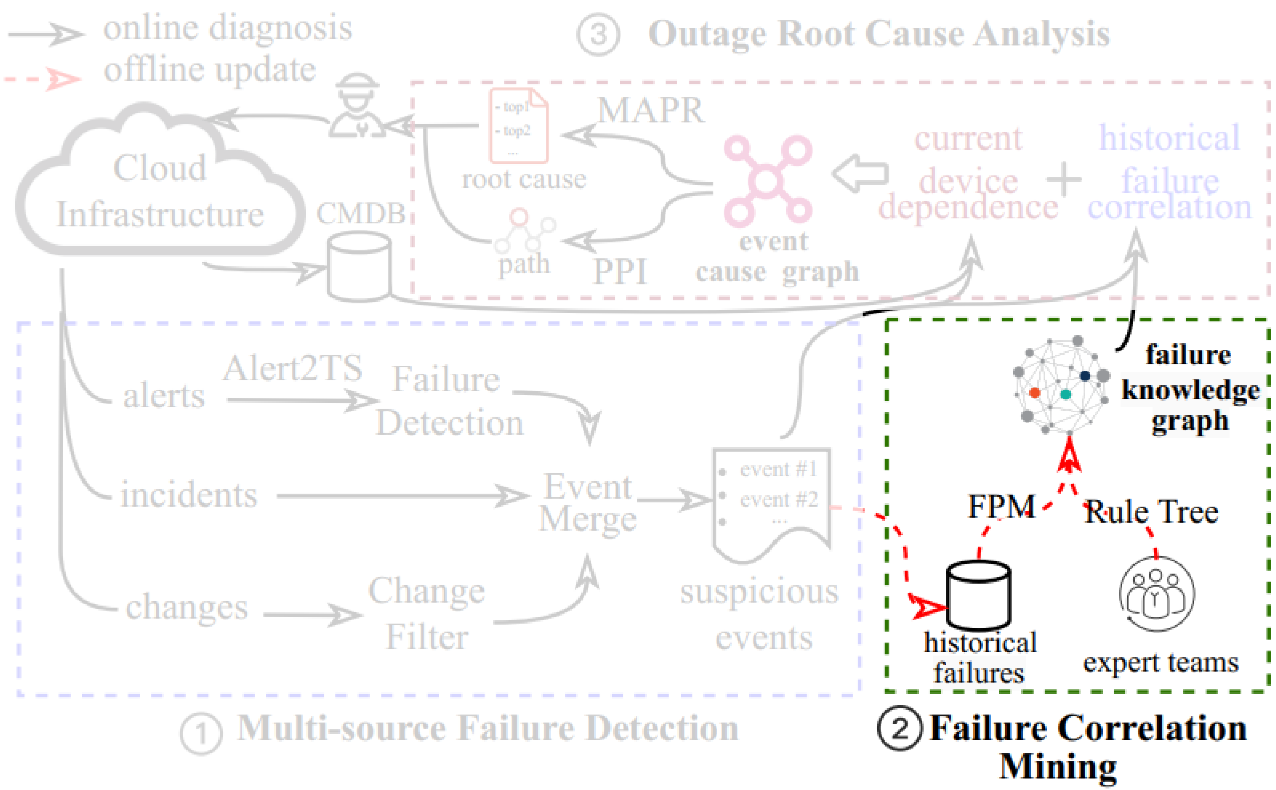
2) Change Filter

3) Event Merge

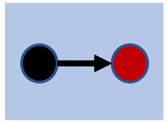


Model Detail

◻ Failure Correlation Mining Module

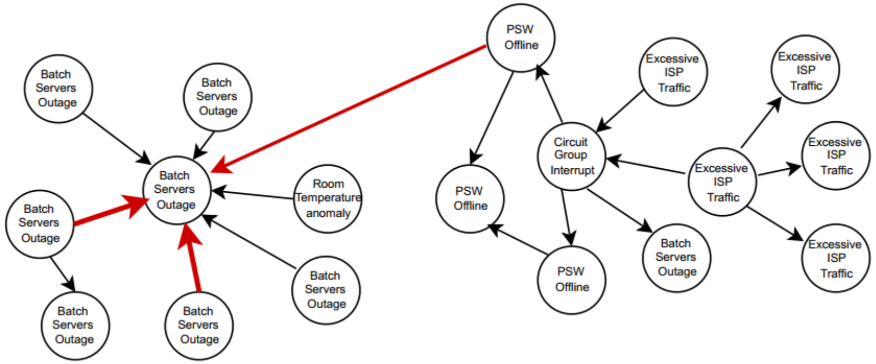


1) Failure Pairs Mining



$$\text{support}(\langle e_a, e_b \rangle)) = P(e_a, e_b) = \frac{num(\langle e_a, e_b \rangle)}{num(failure\ pairs)}$$

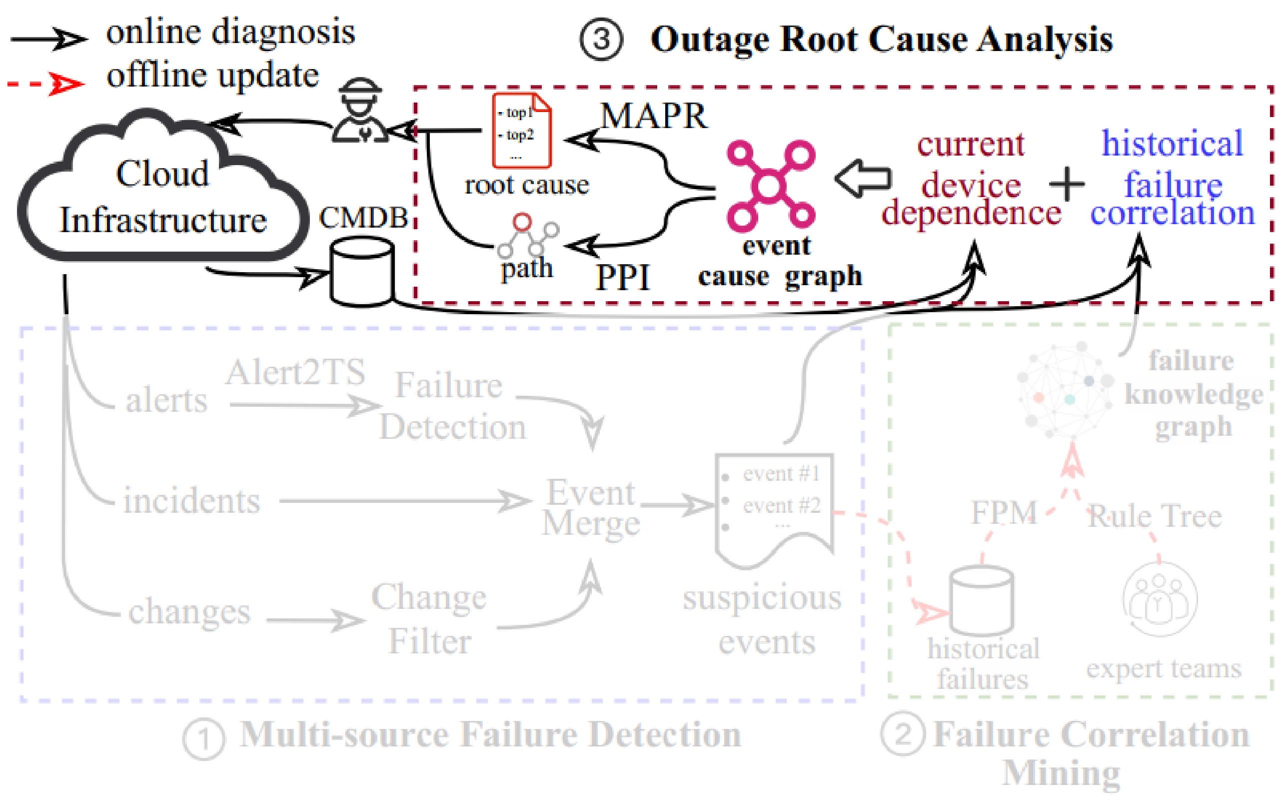
$$\text{confidence}(\langle e_a, e_b \rangle)) = P(e_b | e_a) = \frac{num(\langle e_a, e_b \rangle)}{num(e_a)}$$



Failure Knowledge Graph

Model Detail

Outage Root Cause Analysis Module



1) Event Cause Graph Construction

- Nodes: events
- Edges:

$$w_{ij} = \underbrace{\exp(p_{ij}.conf)}_{\text{Historical failure correlation}} \cdot \underbrace{dist(e_i, e_j)}_{\text{Current device dependence}}$$

2) Outage Root Cause Location

- Node personalization score: $u_i = exp(-t)$
- Failure transition probability: w_{ij}

3) Failure Propagation Path Inference

$$p_U = \arg\max_{p_i \in \mathcal{P}} \mathrm{TransPr}(p_i) = \arg\max_{p_i \in \mathcal{P}} \prod_{j \in |p_i|} \bar{u}_j$$

Outline

- ☐ Background
- ☐ Empirical Observations & Problem Formulation
- ☐ Methodology Design
- ☞ ☐ **Evaluation**
- ☐ Conclusion

Evaluation

- Datasets:
  - We built a large-scale testing platform in **Alibaba CIS** and collected all monitoring data in this testing platform from **January 2022 to December 2023**

Table II: Datasets statistics

Dataset	#Incident	#Alert	#Change	#Failure Types	#Outage Cases
$\mathcal{D}_{init}$	19,020	879,870	3,255	62	27
$\mathcal{D}_{idc}$	173	256,212	1,657	31	19
$\mathcal{D}_{net}$	478	774,638	5,091	44	47
$\mathcal{D}_{all}$	665	1,032,851	7,644	56	68

Evaluation

- Performance in Root Case Location task: BSODiag improved by 9.3%, 8.4%, 10.2%, and 9.3% on the PR@1, PR@2, PR@3, and MAP.

Table III: Comparison of different methods for RCL task

Methods	$\mathcal{D}_{idc}$				$\mathcal{D}_{net}$				$\mathcal{D}_{all}$			
	PR@1	PR@2	PR@3	MAP	PR@1	PR@2	PR@3	MAP	PR@1	PR@2	PR@3	MAP
Random Selection	14.3%	31.2%	42.6%	39.4%	8.7%	23.8%	36.4%	23.0%	10.4%	28.8%	39.6%	25.4%
Hierarchy-First	12.6%	27.5%	46.3%	28.8%	13.9%	37.8%	66.4%	39.4%	12.5%	35.4%	62.5%	36.8%
Time-First	22.5%	42.6%	53.8%	39.6%	41.2%	56.0%	73.7%	57.0%	35.0%	52.0%	70.1%	52.4%
SVM	32.0%	44.6%	62.5%	46.4%	27.4%	44.2%	65.3%	45.6%	27.8%	43.9%	66.1%	45.9%
Random Forest	39.2%	58.8%	72.0%	56.7%	41.4%	57.9%	71.4%	56.9%	42.6%	58.7%	74.3%	58.5%
AirAlert	18.5%	30.9%	41.0%	30.1%	28.0%	43.2%	53.6%	41.6%	24.5%	38.7%	48.8%	37.3%
COT	46.3%	66.0%	82.7%	65.0%	40.8%	57.5%	72.2%	56.8%	44.9%	62.4%	77.3%	61.5%
BSODiag (**ours**)	**56.1%**	**72.9%**	**88.2%**	**72.4%**	**52.4%**	**70.7%**	**86.7%**	**69.9%**	**54.2%**	**70.8%**	**87.5%**	**70.8%**

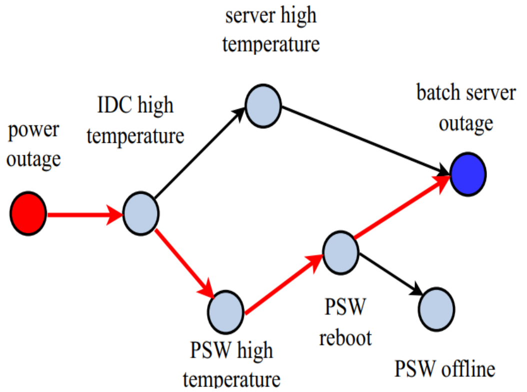
- Proposed unsupervised diagnosis strategy based on the event cause graph is more suitable for the outage diagnosis problem

Evaluation

- **Performance in Failure Propagation Path Inference**: BSODiag achieves 46.3% PCR, showing an improvement of 6.1%, 12.5%, and 3.7% compared to the other baselines.

Table IV: Comparison of different methods for PPI task.

Dataset	Methods			
	DPS	SPS	FHM	BSODiag(**ours**)
$\mathcal{D}_{idc}$	38.0%	35.2%	41.8%	**45.6%**
$\mathcal{D}_{net}$	41.6%	32.4%	43.3%	**46.8%**
$\mathcal{D}_{all}$	40.2%	33.8%	42.6%	**46.3%**



Failures	u_i	r_i
power outage	0.18	0.30
IDC high temperature	0.26	0.22
server high temperature	0.11	0.06
PSW high temperature	0.22	0.26
PSW reboot	0.09	0.12
PSW offline	0.12	0.04

1-st Root Case

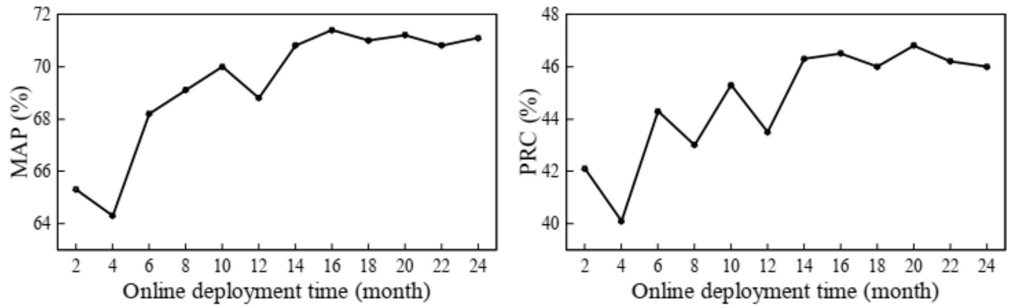
Underlying Root Case

Case study

- BSODiag can provide explainable diagnosis results, which prompts practical Troubleshooting.

Evaluation

Online Deployment Evolution & Ablation Study



(a) MAP (b) PCR

Figure 6: The online deployment performance of BSODiag.

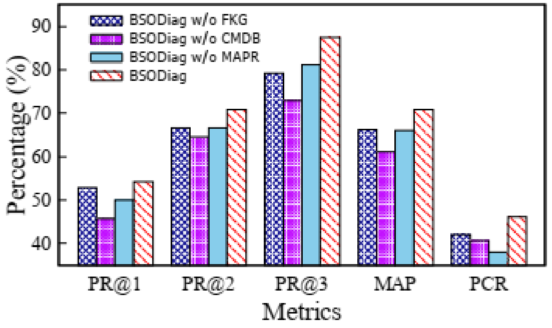


Figure 7: Ablation study

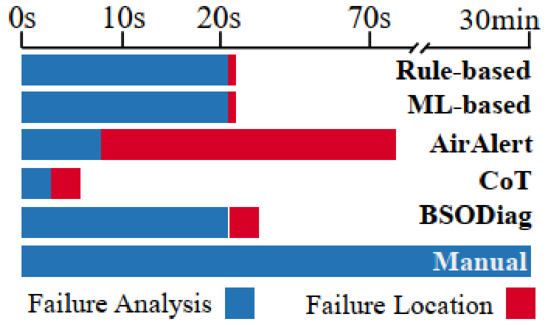


Figure 8: Time Consumption

- In actual online deployment, as more failure data are collected, we can continuously update BSODiag to optimize its performance
- BSODiag achieves a single diagnosis of an outage case in 24.5 seconds, marking a substantial improvement over the traditional manual diagnosis

Outline

- ☐ Background
- ☐ Empirical Observations & Problem Formulation
- ☐ Methodology Design
- ☐ Evaluation
- ☞ ☐ **Conclusion**

Conclusion

- We formulate the batch servers outage diagnosis problem. Our empirical study on a large-scale cloud system uncover the key insights of this problem.
- We propose BSODiag, an unsupervised and lightweight diagnosis framework to address the problem.
- We collected real-world data from Alibaba Cloud infrastructure system and demonstrate that BSODiag outperforms all alternative methods.

Thanks for listening!

Paper Link

duantao@stu.xjtu.edu.cn