FaultProfIT: Hierarchical Fault Profiling of Incident Tickets in Large-scale Cloud Systems

Junjie Huang[1], Jinyang Liu[1], Zhuangbin Chen[2], Zhihan Jiang[1], Yichen Li[1], Jiazhen Gu[1], Cong Feng[3], Zengyin Yang[3], Yongqiang Yang[3], Michael R. Lyu[1]

[1]The Chinese University of Hong Kong

[2]Sun-yat Sen University

[3]Huawei Cloud

ARISE Automated Reliable Intelligent Software Engineering

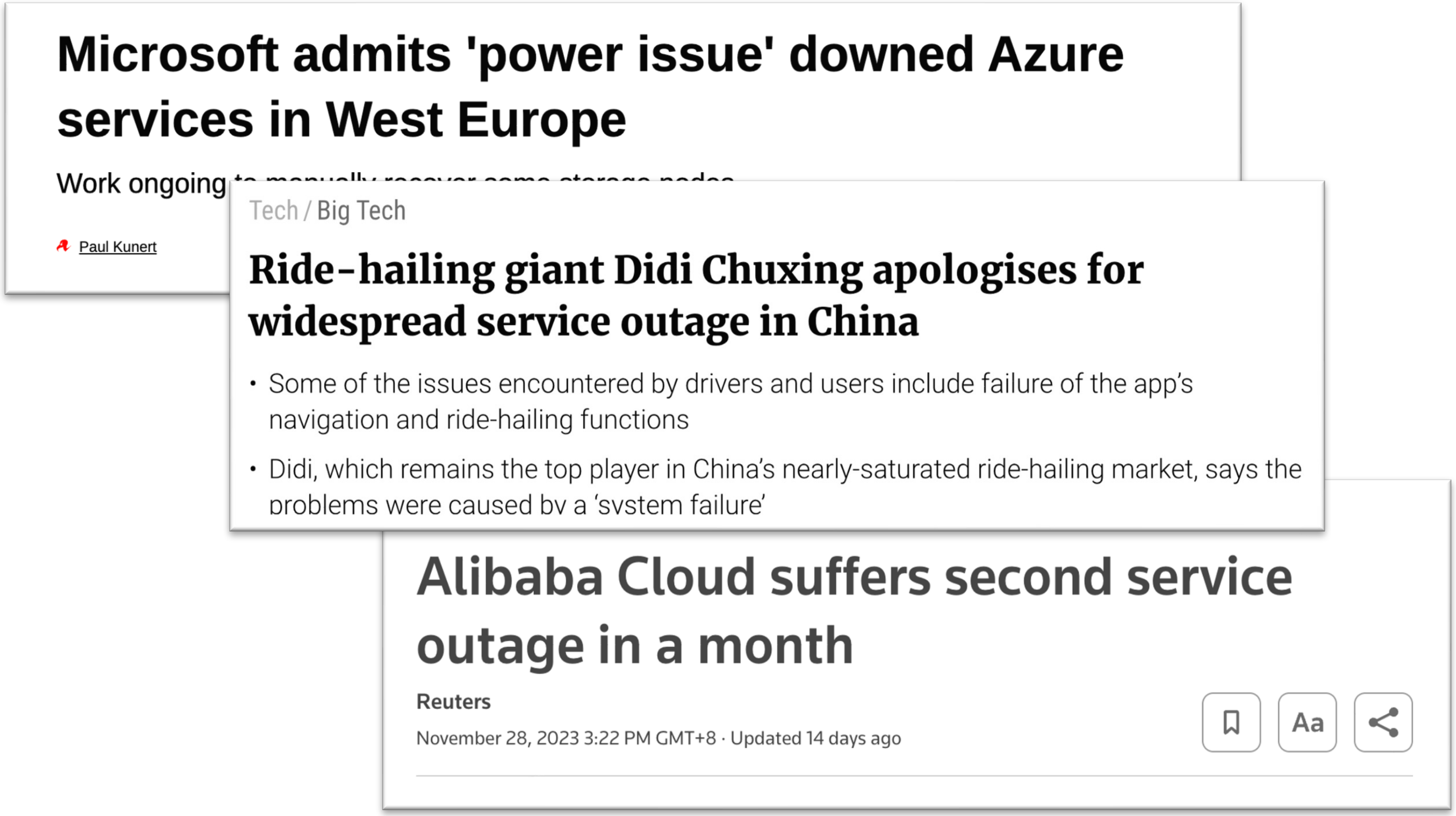
香港中文大學 The Chinese University of Hong Kong

中山大學 SUN YAT-SEN UNIVERSITY

HUAWEI

HUAWEI CLOUD

Ensuring reliability of cloud systems is crucial



Microsoft admits 'power issue' downed Azure services in West Europe

Work ongoing to manually recover some storage nodes

Paul Kunert

Tech / Big Tech

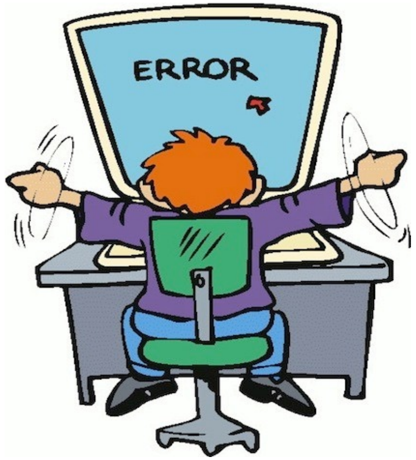
Ride-hailing giant Didi Chuxing apologises for widespread service outage in China

- Some of the issues encountered by drivers and users include failure of the app's navigation and ride-hailing functions
- Didi, which remains the top player in China's nearly-saturated ride-hailing market, says the problems were caused by a 'system failure'

Alibaba Cloud suffers second service outage in a month

Reuters

November 28, 2023 3:22 PM GMT+8 · Updated 14 days ago

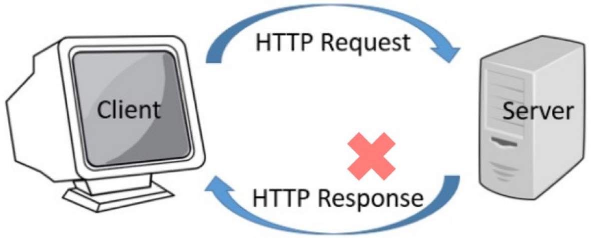


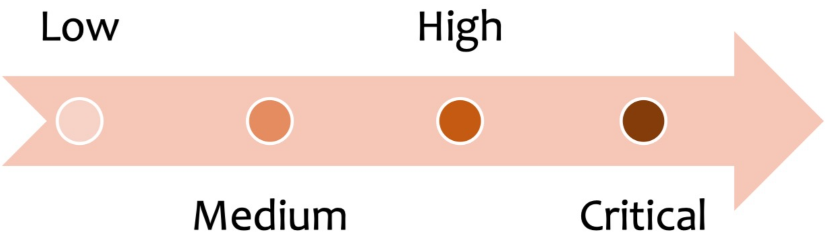
User dissatisfaction

Huge revenue loss

What is incident?

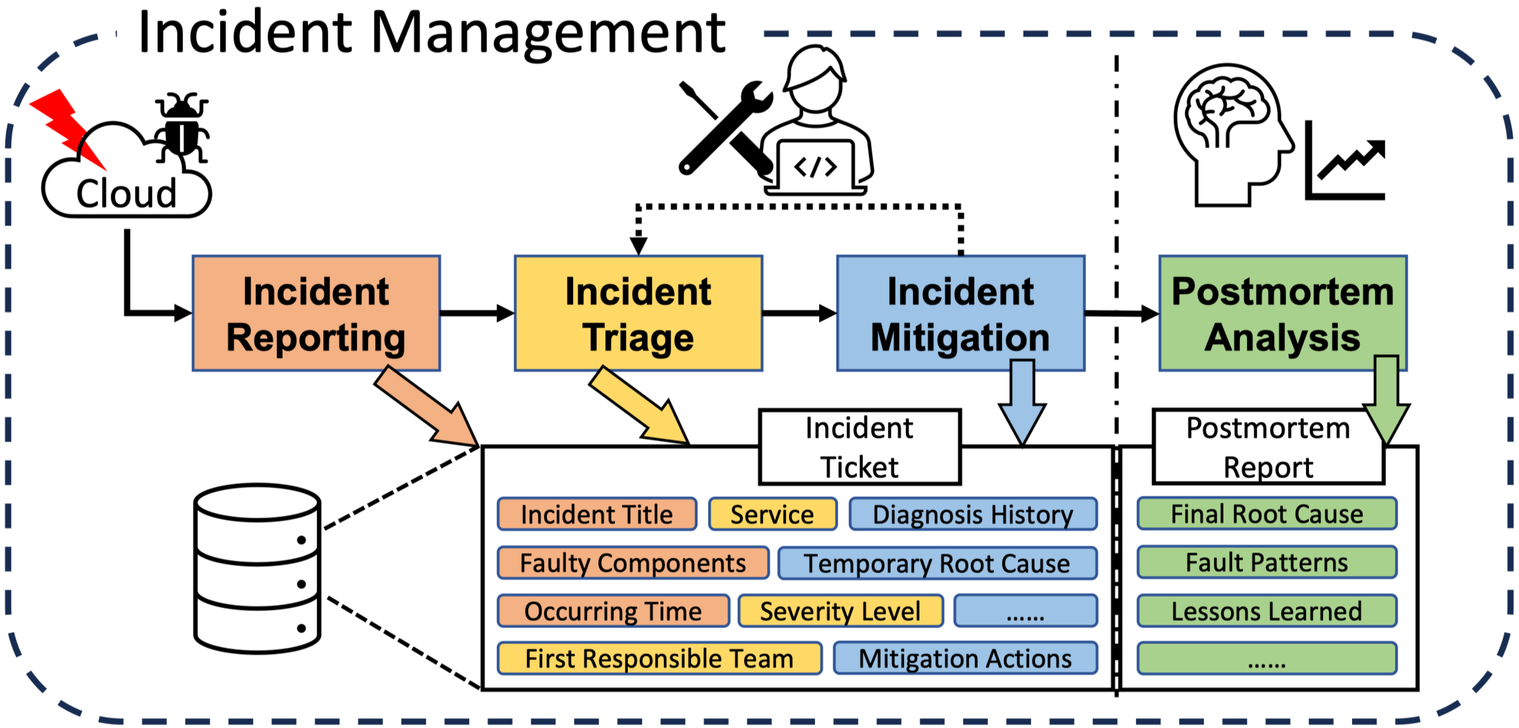
- Unplanned service interruption or performance degradation
  - Can be referred to as *failure*
  - Examples:
    - Bad HTTP requests
    - Power outages
    - User-reported errors





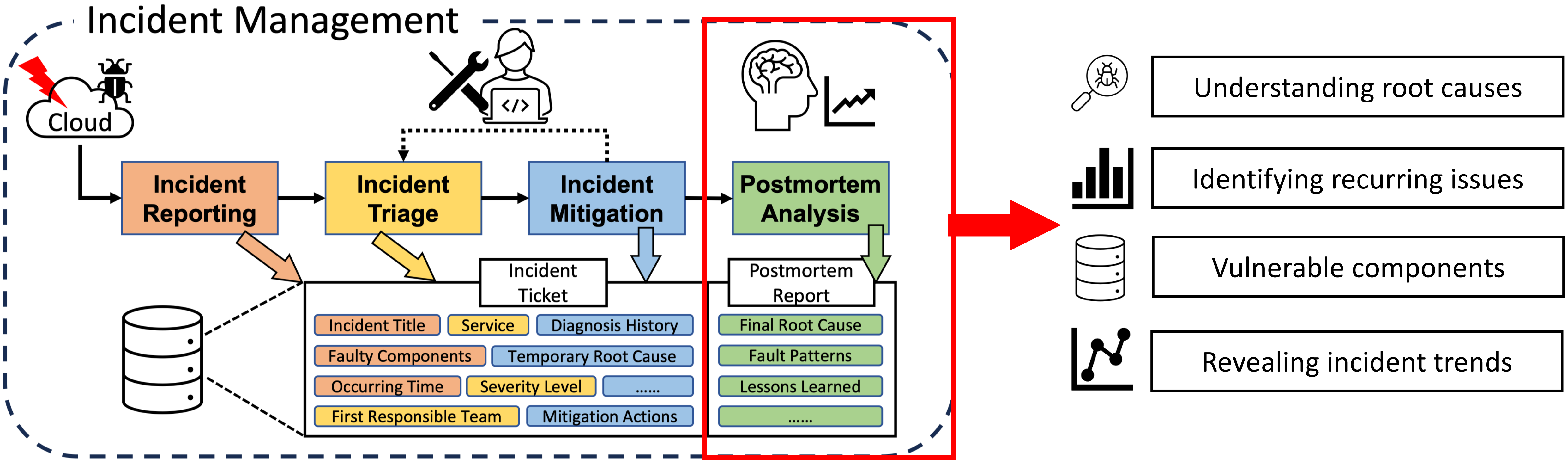
Incident management process

- **Real-time response**
  - **Goal:** mitigate incidents as quickly as possible
- **Postmortem analysis**
  - **Goal:** analyze incident tickets to summarize experience



Incident management process

- **Postmortem analysis**
  - **Goal:** analyze incident tickets to summarize experience
  - It is important in improving cloud reliability



Incident Management

Incident Reporting

Incident Triage

Incident Mitigation

Postmortem Analysis

Incident Ticket

- Incident Title
- Service
- Diagnosis History
- Faulty Components
- Temporary Root Cause
- Occurring Time
- Severity Level
- ……
- First Responsible Team
- Mitigation Actions

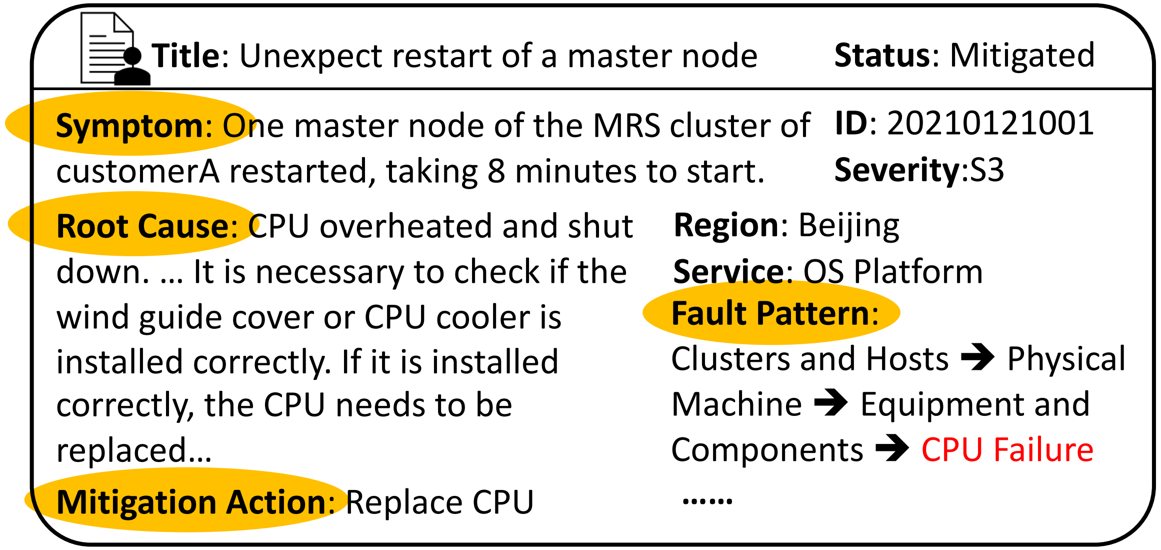
Postmortem Report

- Final Root Cause
- Fault Patterns
- Lessons Learned
- ……

- Understanding root causes
- Identifying recurring issues
- Vulnerable components
- Revealing incident trends

Fault pattern profiling for postmortem analysis

- **Fault Pattern Profiling in CloudA**
  - Categorize faults occurred in each incident into different types
  - E.g., CPU overload, power outage, SSD failure, etc. → Fault Patterns



Title: Unexpect restart of a master node

Status: Mitigated

Symptom: One master node of the MRS cluster of customerA restarted, taking 8 minutes to start.

ID: 20210121001

Severity:S3

Root Cause: CPU overheated and shut down. … It is necessary to check if the wind guide cover or CPU cooler is installed correctly. If it is installed correctly, the CPU needs to be replaced…

Mitigation Action: Replace CPU

Region: Beijing

Service: OS Platform

Fault Pattern: Clusters and Hosts → Physical Machine → Equipment and Components → CPU Failure

……

An example of incident tickets

Symptom: One master node …

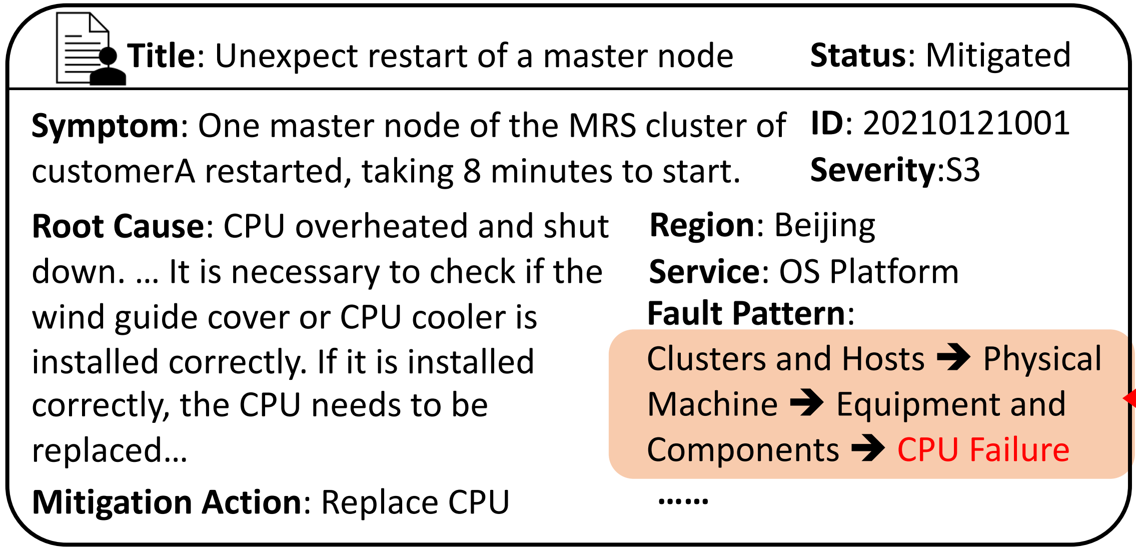
Root Cause: CPU overheated and shut down …

Mitigation Action: Replace CPU

Fault Pattern: Clusters and Hosts → Physical Machine → Equipment and Components → CPU Failure

Fault pattern profiling for postmortem analysis

- **Fault Pattern Profiling**
  - Classifying faults occurred in the incidents according to tickets
  - E.g., CPU overload, power outage, SSD failure, etc. ➔ Fault Patterns
- **Fault Pattern Taxonomy**
  - Hierarchy: 5 levels and 334 fault patterns in total
  - Description: A fault pattern contains symptoms, fault tolerance measures, *etc.*



Title: Unexpect restart of a master node **Status**: Mitigated

Symptom: One master node of the MRS cluster of customerA restarted, taking 8 minutes to start.

ID: 20210121001

Severity:S3

Root Cause: CPU overheated and shut down. … It is necessary to check if the wind guide cover or CPU cooler is installed correctly. If it is installed correctly, the CPU needs to be replaced…

Region: Beijing

Service: OS Platform

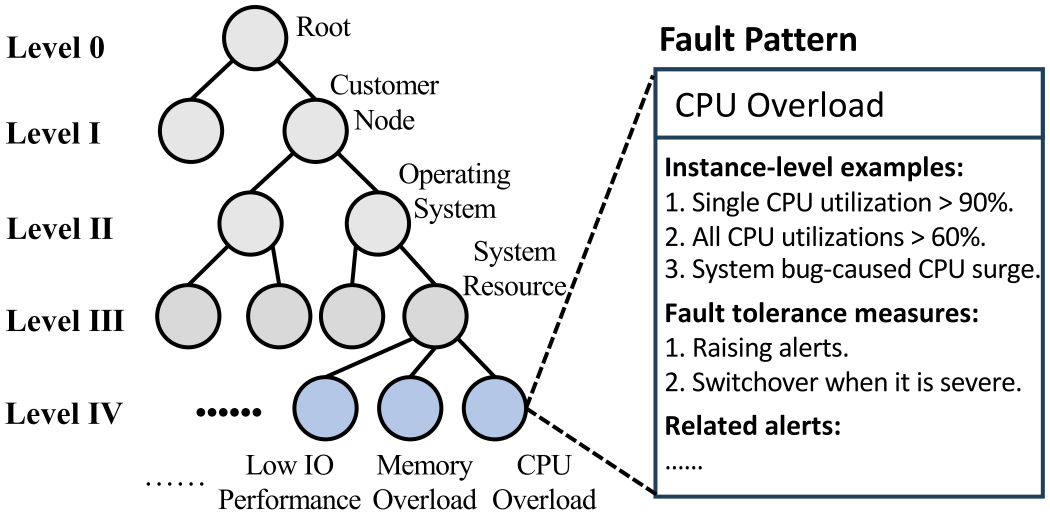
Fault Pattern:

Clusters and Hosts ➔ Physical Machine ➔ Equipment and Components ➔ CPU Failure

……

Mitigation Action: Replace CPU

An example of incident tickets



Fault Pattern

CPU Overload

Instance-level examples:

1. Single CPU utilization > 90%.
2. All CPU utilizations > 60%.
3. System bug-caused CPU surge.

Fault tolerance measures:

1. Raising alerts.
2. Switchover when it is severe.

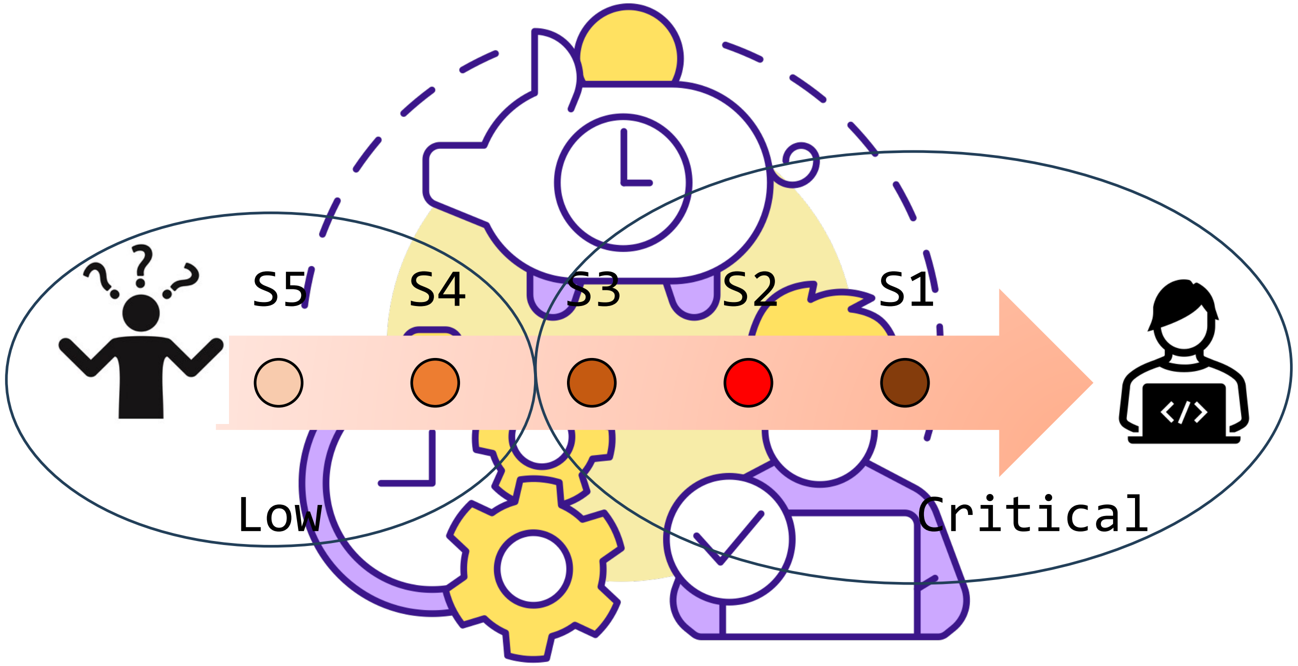
Related alerts:

……

Fault Pattern Taxonomy

Manually fault pattern profiling is challenging

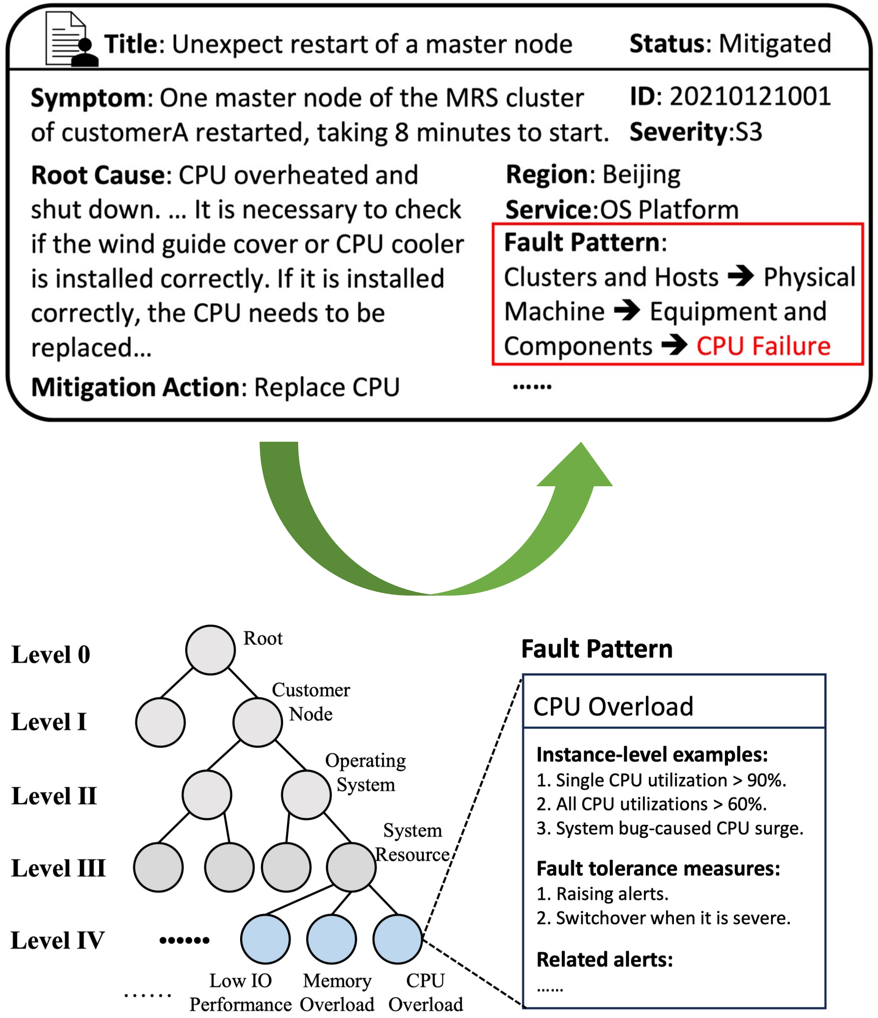
- **Large-scale:**
  - Focusing on S2/S3 level incidents
  - Less efforts for S4/S5 level
  - But they are common and numerous
- **Expensive**:
  - Time-consuming
  - Labor-intensive
  - Domain knowledge
- **Inconsistent Profiling:**
  - Variations in expert knowledge
  - Complex fault pattern taxonomy



We need an automated approach!

FaultProfIT: a hierarchical contrastive learning method

- **Task Definition:**
  - Input:
    - Textual incident tickets
    - Fault pattern taxonomy
  - Output:
    - Fault pattern labels



Title: Unexpect restart of a master node **Status**: Mitigated

Symptom: One master node of the MRS cluster of customerA restarted, taking 8 minutes to start.

ID: 20210121001

Severity:S3

Root Cause: CPU overheated and shut down. … It is necessary to check if the wind guide cover or CPU cooler is installed correctly. If it is installed correctly, the CPU needs to be replaced…

Region: Beijing

Service:OS Platform

Fault Pattern: Clusters and Hosts ➔ Physical Machine ➔ Equipment and Components ➔ CPU Failure

Mitigation Action: Replace CPU

……

Fault Pattern

CPU Overload

Instance-level examples:

1. Single CPU utilization > 90%.
2. All CPU utilizations > 60%.
3. System bug-caused CPU surge.

Fault tolerance measures:

1. Raising alerts.
2. Switchover when it is severe.

Related alerts:

……

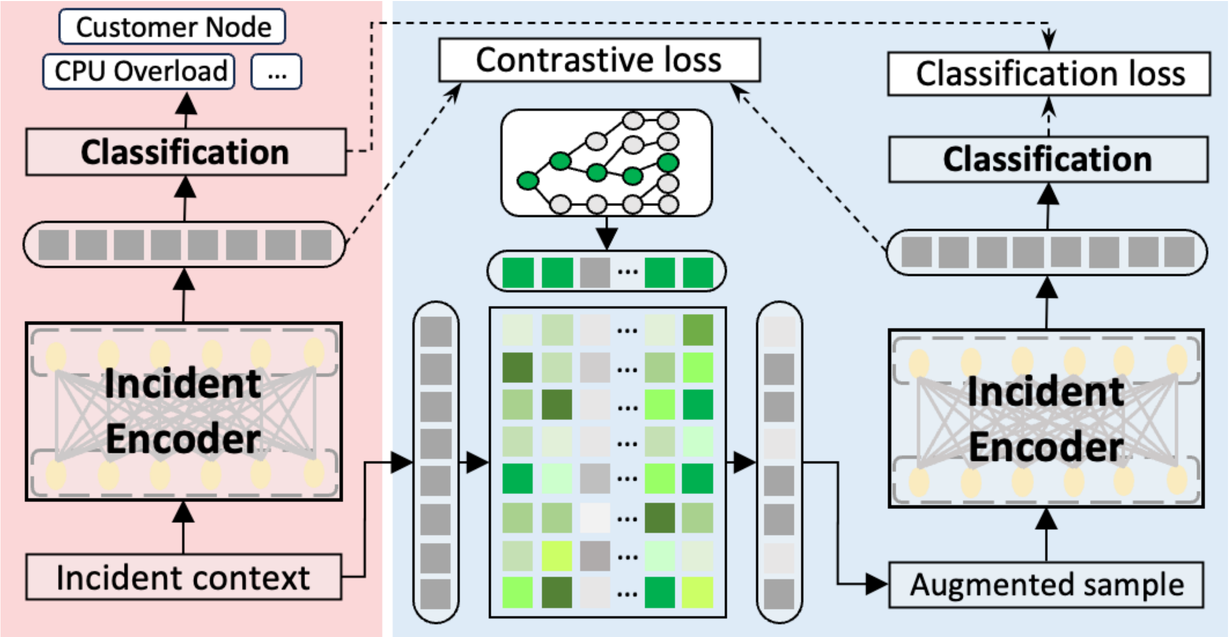
FaultProfIT: a hierarchical contrastive learning method

- **Challenge 1:** Complex fault patterns
  - 5 levels and 334 fault patterns in total
  - Hierarchical and textual information

→ Hierarchical Textual Classification

- **Challenge 2:** Insufficient training data
  - We only have 1463 annotated tickets

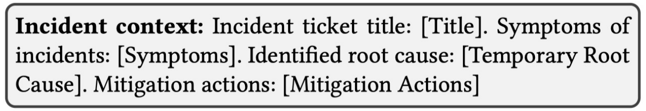
→ Augmented Examples and Contrastive Learning



FaultProfIT: a hierarchical contrastive learning method

Incident Encoder

- Goal: Encode incident context into vectors
- Incident context

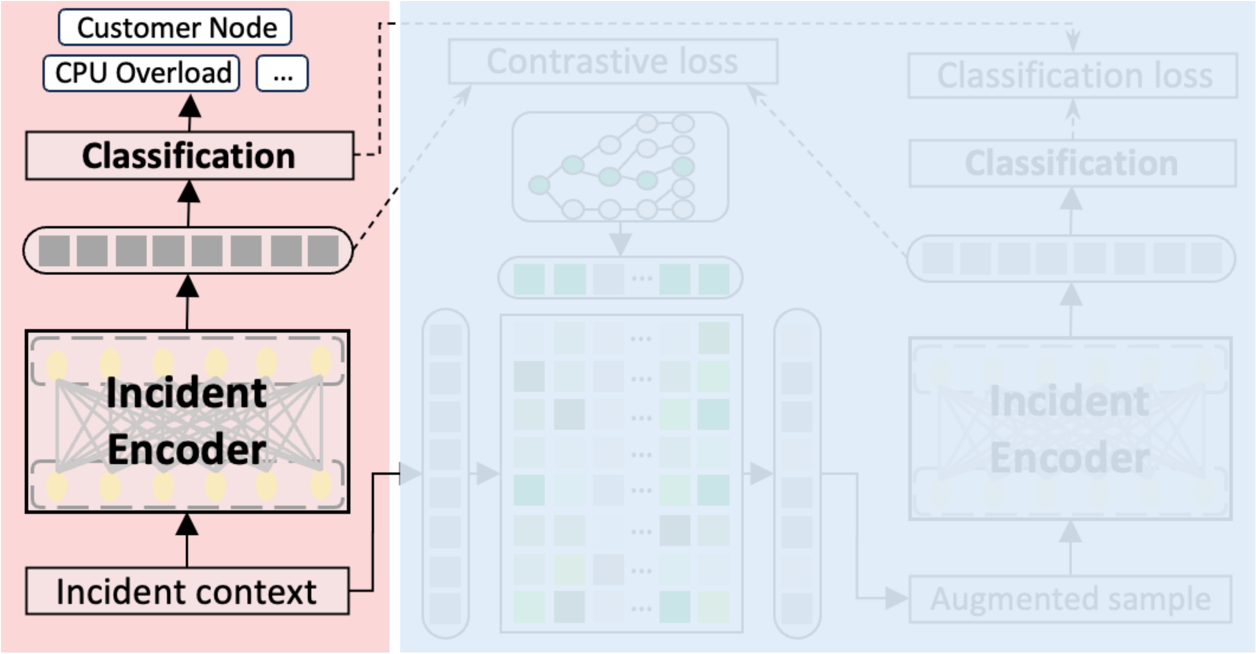


Incident context: Incident ticket title: [Title]. Symptoms of incidents: [Symptoms]. Identified root cause: [Temporary Root Cause]. Mitigation actions: [Mitigation Actions]

- Incident encoder based on MacBERT[1]

$$\mathbf{X} = \mathrm{MacBERT}(x)$$

$$\boldsymbol{x} = \mathrm{X}_{[\mathrm{CLS}]}$$



[1] Cui, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing. (EMNLP'20)

FaultProfIT: a hierarchical contrastive learning method

Fault Pattern Encoder

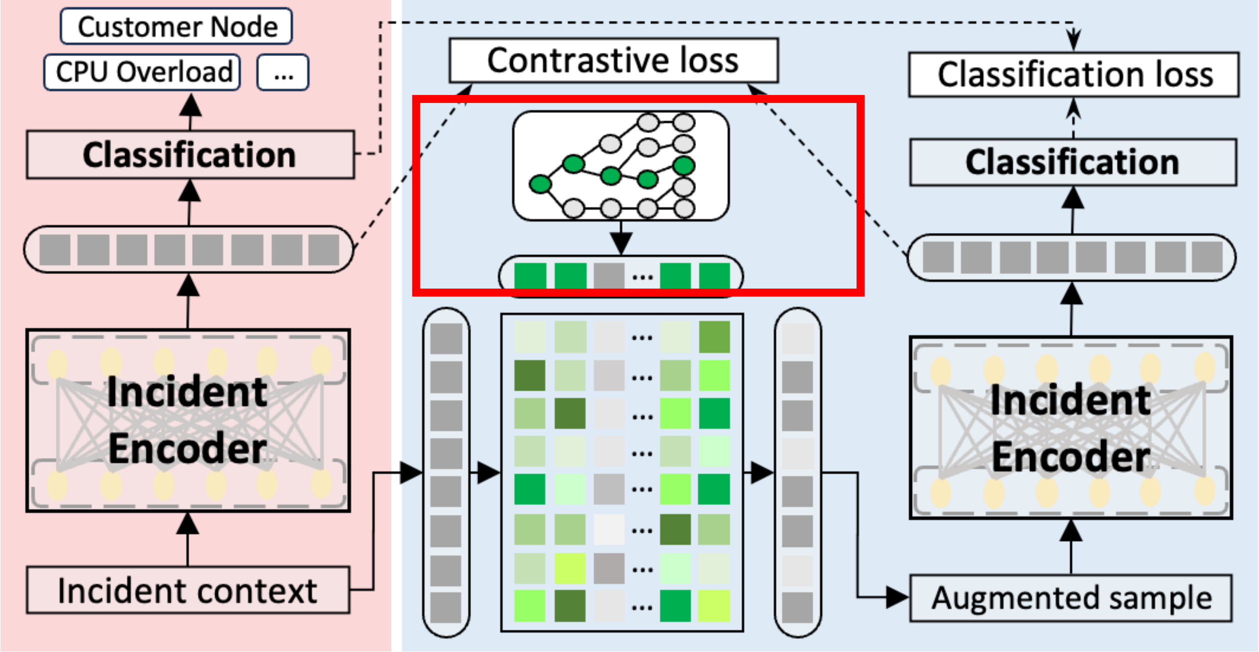
- Goal: Encode each fault pattern f_i into a vector
- Embed fault patterns into vectors

$$\boldsymbol{f_i} = LabelEmbedding(f_i) + DescriptionEmbedding(f_i)$$

- Apply graph encoder to encode $DAG(F, E)$

$$\boldsymbol{H} = Graphormer\ ([\boldsymbol{f_1}, \dots \boldsymbol{f_i}, \dots \boldsymbol{f_k}])$$

- Hierarchy-aware embedding $f_i \rightarrow \boldsymbol{H_i}$



FaultProfIT: a hierarchical contrastive learning method

Original Context

Title: Unexpected restart of a master node in USEast. Symptom of … Root Cause: CPU overheated and shut down. … If it is installed correctly, the CPU needs to be replaced …

Augmented Context

Title: Unexpected restart node. Symptom … Root Cause: CPU overheated and shut down. … CPU be replaced …

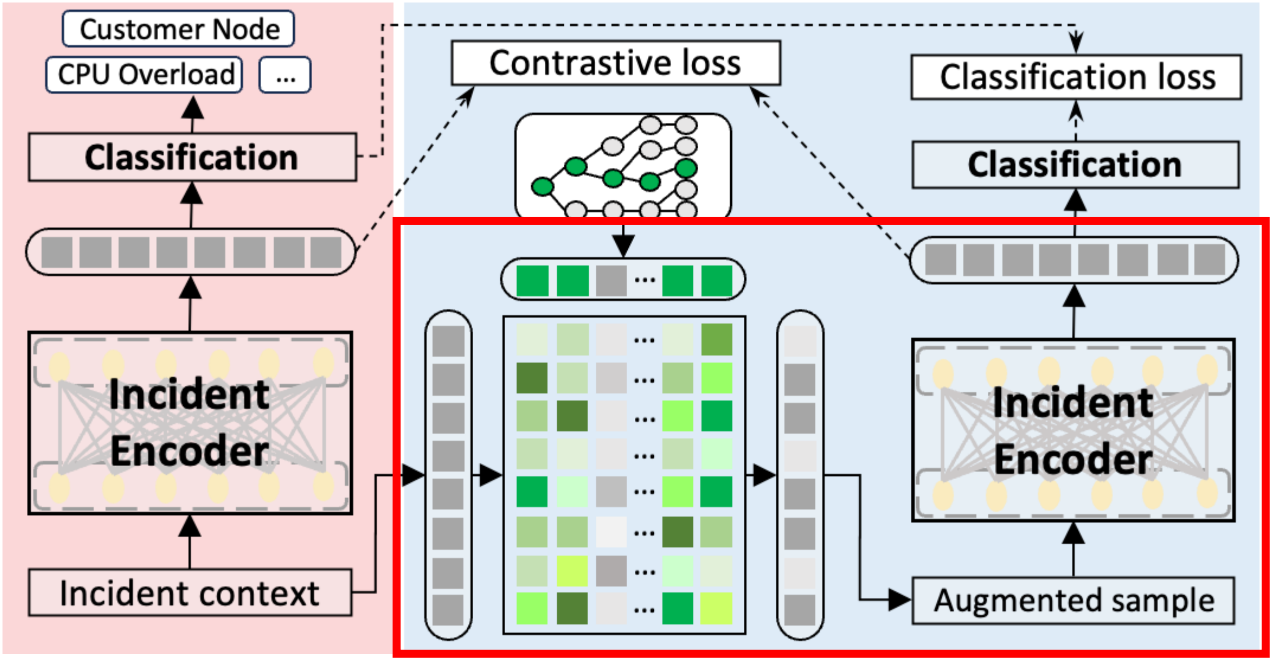
Hierarchy-guided data augmentation

- Idea: Remove unimportant words in incident context so that the ticket keeps the same labels.
- Weight score of x_i to be a keyword of f_j

$$A = scale\_dot\_attention(\mathbf{X}, \mathbf{H}) \qquad P_{ij} = gumbel\_softmax(A_{i1}, A_{i2}, \dots, A_{ik})_j$$

- Embedding of the augmented

$$\hat{x} = \{x_i \text{ if } P_i > \lambda\} \qquad \hat{\mathbf{X}} = \text{MacBERT}(\hat{x}) \qquad \hat{\boldsymbol{x}} = \hat{\mathbf{X}}_{[\text{CLS}]}$$



FaultProfIT: a hierarchical contrastive learning method

Original Context

Title: Unexpected restart of a master node in USEast. Symptom of … Root Cause: CPU overheated and shut down. … If it is installed correctly, the CPU needs to be replaced …

Multi-label Classification

Clusters and Hosts → Physical Machine → Equipment and Components → CPU Failure

Contrastive Learning

Title: Unexpected restart node. Symptom … Root Cause: CPU overheated and shut down. … CPU be replaced …

Augmented Context

Multi-label Classification

Clusters and Hosts → Physical Machine → Equipment and Components → CPU Failure

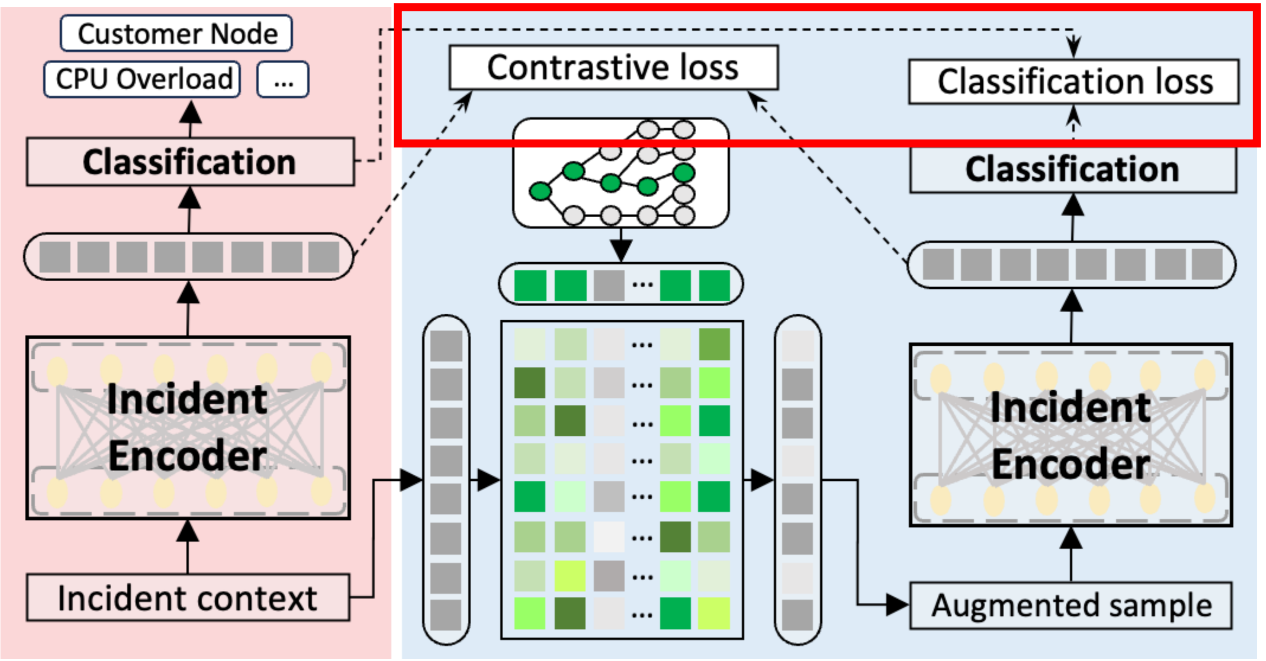
Loss Function:

- Final loss $Loss = Loss^{cls} + \hat{Loss}^{cls} + \alpha Loss^{contra},$
- Multi-label classification loss:

$$Loss^{cls} = -\sum_{i=1}^{N}\sum_{j=1}^{k} \gamma \dot{f}_j^{(i)} \log\left(p_j^{(i)}\right) + \left(1 - f_j^{(i)}\right) \log\left(1 - p_j^{(i)}\right)$$

- Contrastive loss:

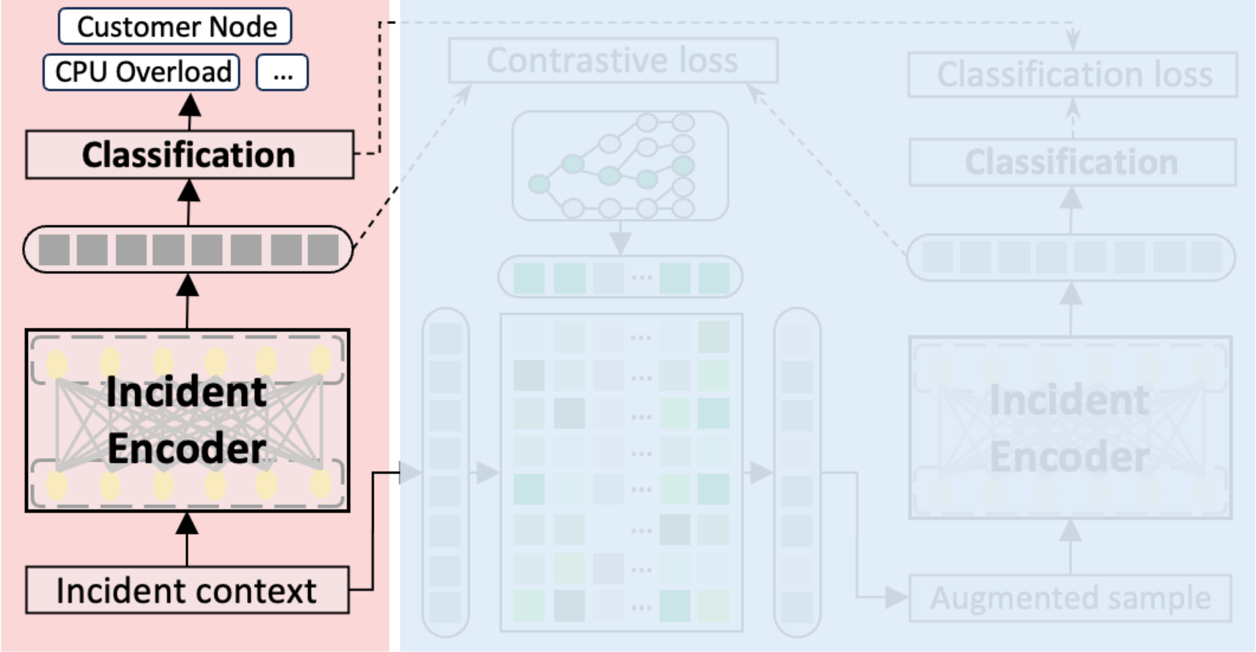
$$Loss^{contra} = -\sum_{i=1}^{2N} \log \frac{e^{\mathrm{cosine}(\mathbf{x}^{(i)},\hat{\mathbf{x}}^{(i)})/\tau}}{\sum_{j=1,j\neq i}^{2N} e^{\mathrm{cosine}(\mathbf{x}^{(i)},\mathbf{x}^{(j)})/\tau}},$$



FaultProfIT: a hierarchical contrastive learning method

Inference Stage

- Constructing incident context
- Encode with the trained incident encoder for fault pattern profiling

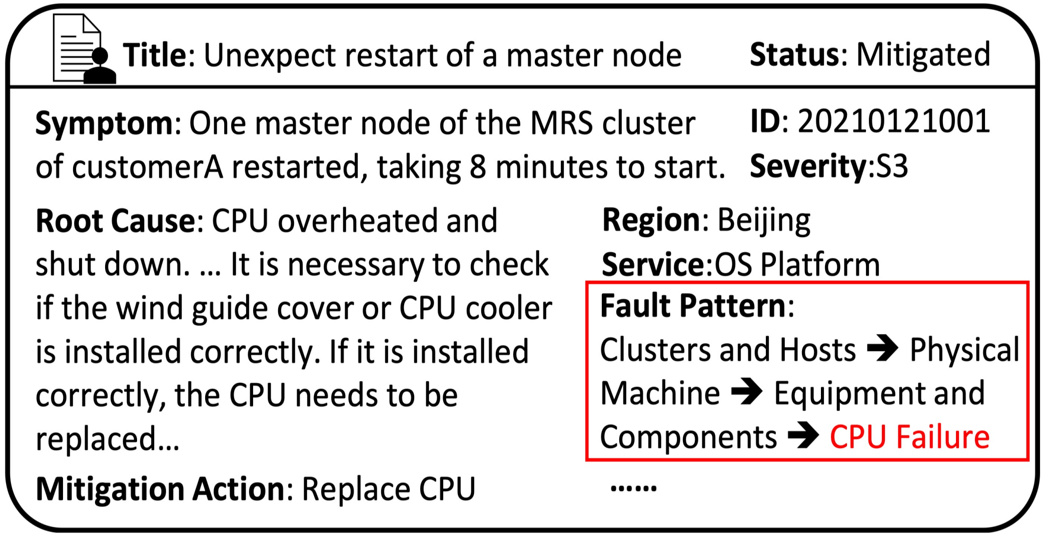


Incident context: Incident ticket title: [Title]. Symptoms of incidents: [Symptoms]. Identified root cause: [Temporary Root Cause]. Mitigation actions: [Mitigation Actions]

Experiment

- **Industrial Dataset**
  - 6 years of incident tickets
  - 22,560 incidents in total
  - 1,463 incidents with annotated labels
  - Train: dev: test = 8:1:1

- **Core services**
  - Elastic Computing Service (ECS)
  - Virtual Private Cloud (VPC)
  - Cloud Container Engine (CCE)
  - OBS, DCS…



Title: Unexpect restart of a master node **Status**: Mitigated

Symptom: One master node of the MRS cluster of customerA restarted, taking 8 minutes to start.

ID: 20210121001

Severity:S3

Root Cause: CPU overheated and shut down. … It is necessary to check if the wind guide cover or CPU cooler is installed correctly. If it is installed correctly, the CPU needs to be replaced…

Region: Beijing

Service:OS Platform

Fault Pattern: Clusters and Hosts ➔ Physical Machine ➔ Equipment and Components ➔ CPU Failure

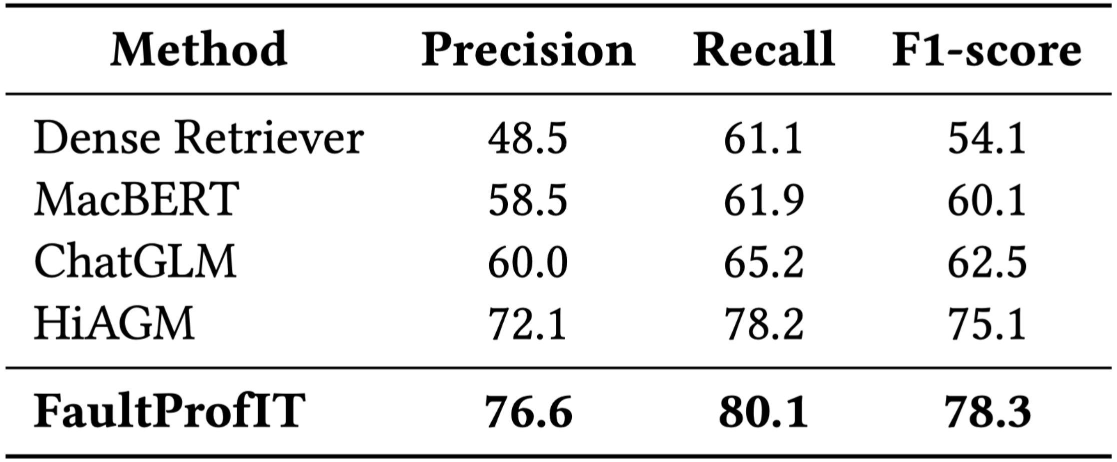
Mitigation Action: Replace CPU

……

An example of incident tickets

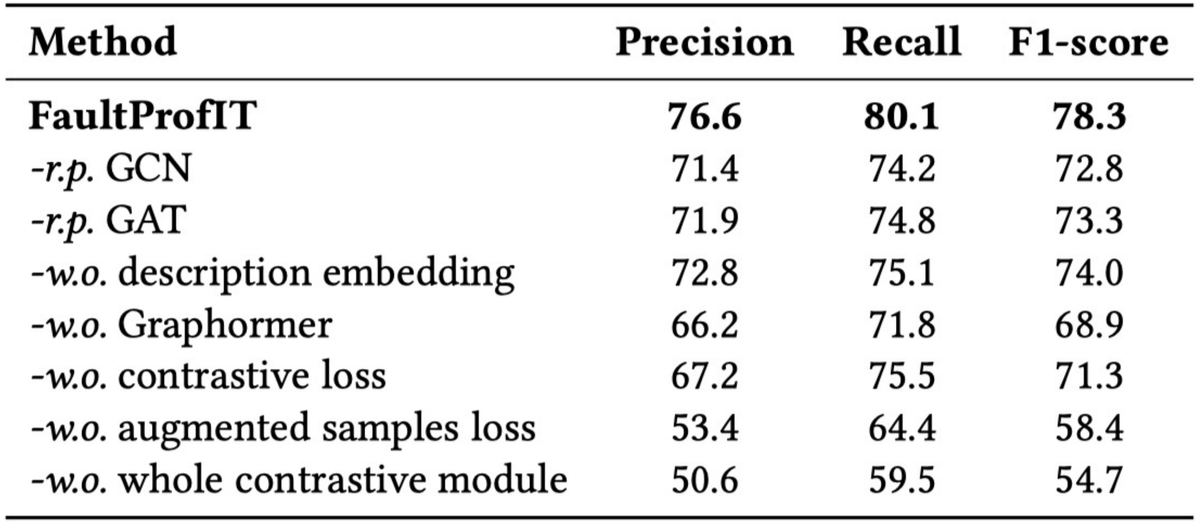
Performance

- FaultProfIT achieves a high degree of accuracy!
- Hierarchical contrastive learning is effective!



Method	Precision	Recall	F1-score
Dense Retriever	48.5	61.1	54.1
MacBERT	58.5	61.9	60.1
ChatGLM	60.0	65.2	62.5
HiAGM	72.1	78.2	75.1
FaultProfIT	**76.6**	**80.1**	**78.3**

Overall performance compared with baselines



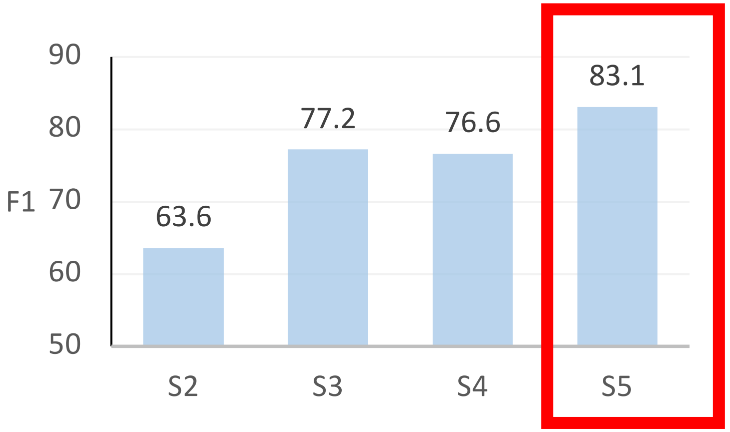
Method	Precision	Recall	F1-score
FaultProfIT	**76.6**	**80.1**	**78.3**
-*r.p.* GCN	71.4	74.2	72.8
-*r.p.* GAT	71.9	74.8	73.3
-*w.o.* description embedding	72.8	75.1	74.0
-*w.o.* Graphormer	66.2	71.8	68.9
-*w.o.* contrastive loss	67.2	75.5	71.3
-*w.o.* augmented samples loss	53.4	64.4	58.4
-*w.o.* whole contrastive module	50.6	59.5	54.7

Ablation study on different components

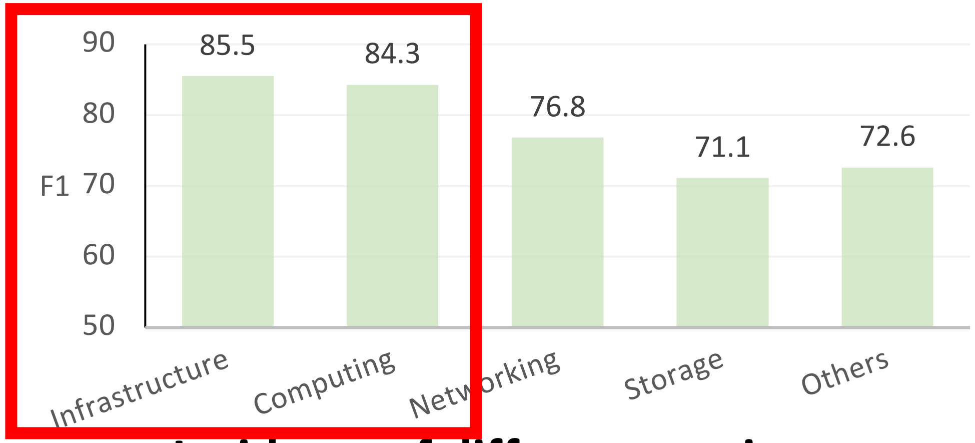
[1] Karpukhin, et al. Dense Passage Retrieval for OpenDomain Question Answering. (EMNLP'20)
[2] Cui, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing. (EMNLP'20)
[3] Du, et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. (ACL'22)
[4] Zhou, et al. Hierarchy-aware global model for hierarchical text classification. (ACL'20)

Effects of severities and services

- **Varying severities**
  - FaultProfIT performs better on less severe incidents.
  - Reason: Severe incidents often have complex and extended contexts
- **Varying services**
  - FaultProfIT performs better on incidents from infrastructure and computing services.
  - Reason: Incidents involving servers and hardware have more explicit descriptions



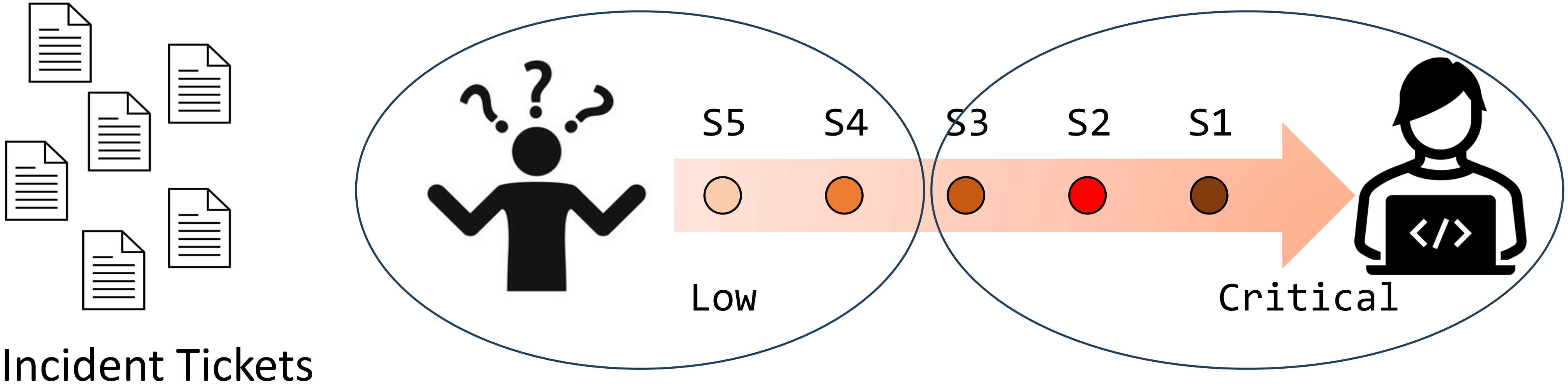
Incidents of different severities



Incidents of different services

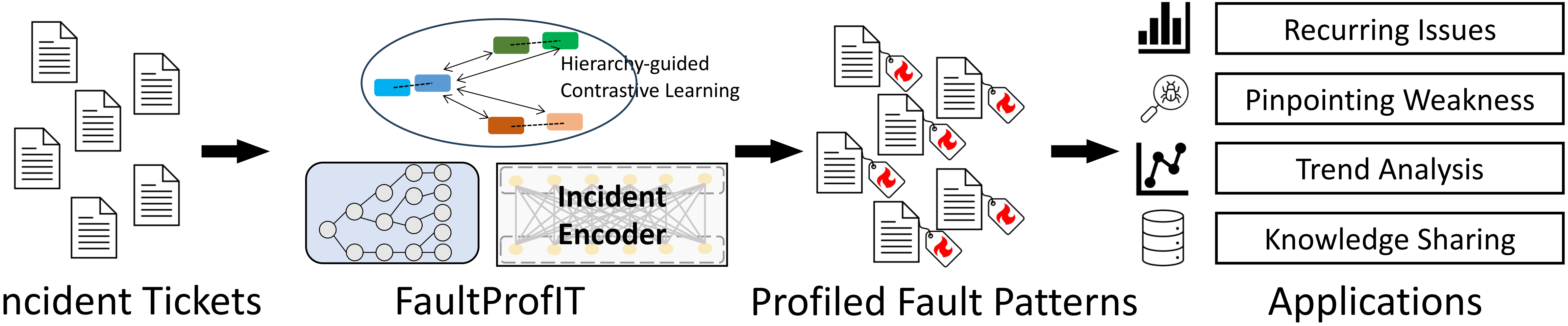
Deployment Experience

- **FaultProfIT has been successfully deployed in CloudA**
  - 10000+ incidents from 30+ services have been analyzed
  - The efficiency and accuracy of fault pattern has been substantially improved
  - Being integrated as a profiling service for internal users



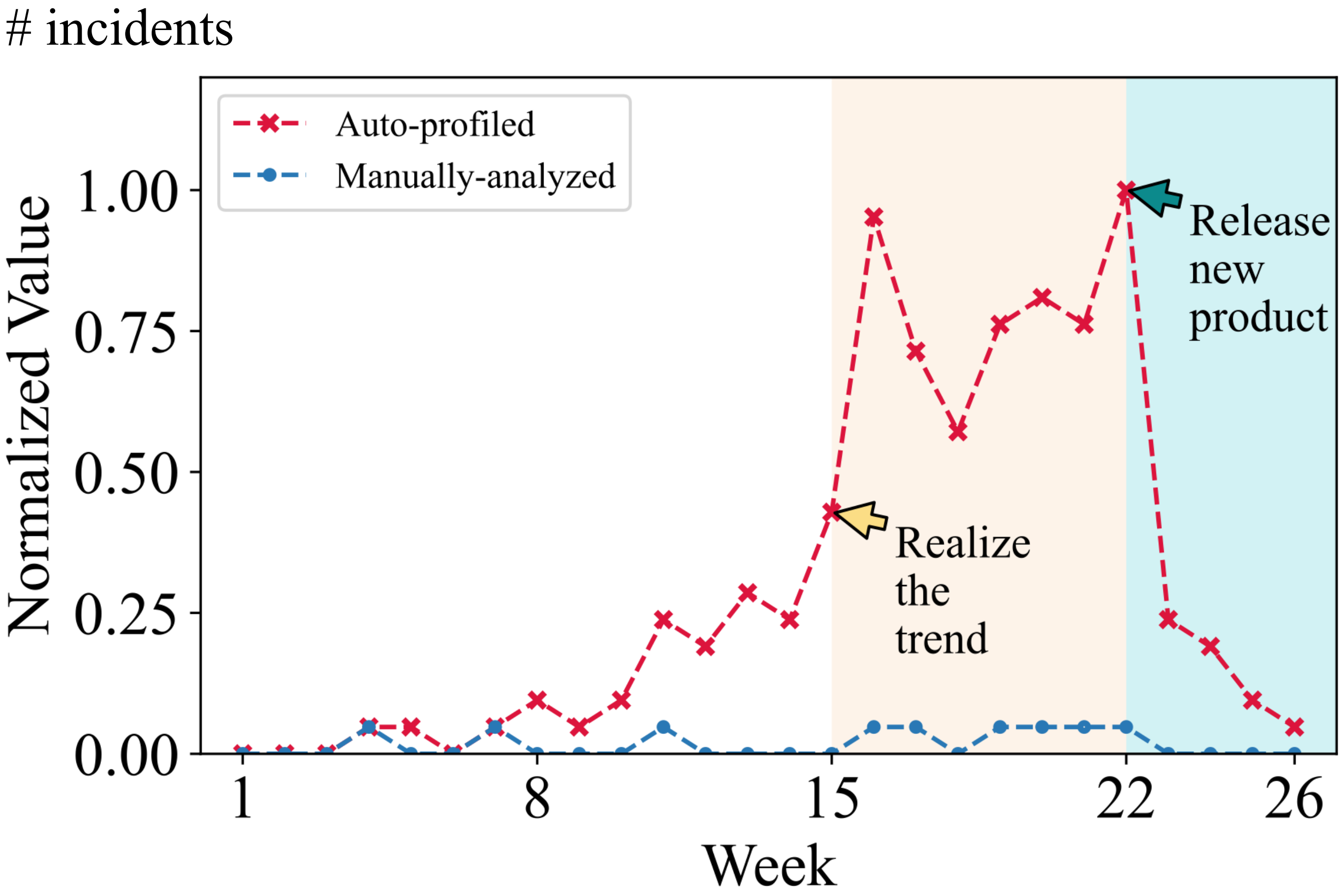
Deployment Experience

- **FaultProfIT has been successfully deployed in Cloud A**
  - 10000+ incidents from 30+ services have been analyzed
  - The efficiency and accuracy of fault pattern has been substantially improved
  - Being integrated as a profiling service for internal users



Deployment Experience

- An example of **trend analysis** for ***memory overload*** fault pattern



Conclusion

- Fault pattern profiling is important in incident postmortem
- We developed FaultProfIT for automatic profiling
  - Inject hierarchy information and mitigate data insufficiency problem
- FaultProfIT is effective in predicting fault patterns
- FaultProfIT has been successfully deployed in Cloud A

Q & A