Reliability-Driven AIOps for Cloud Resilience

Prof. Michael R. Lyu
The Chinese University of Hong Kong
• Modern software systems are serving many aspects of our life
**Cloud Computing**

- Cloud adoption rising

- Cloud revenue growing

<table>
<thead>
<tr>
<th></th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
<th>2022</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud Business Process Services (BPaaS)</td>
<td>41.7</td>
<td>43.7</td>
<td>46.9</td>
<td>50.2</td>
<td>53.8</td>
</tr>
<tr>
<td>Cloud Application Infrastructure Services (PaaS)</td>
<td>26.4</td>
<td>32.2</td>
<td>39.7</td>
<td>48.3</td>
<td>58.0</td>
</tr>
<tr>
<td>Cloud Application Services (SaaS)</td>
<td>85.7</td>
<td>99.5</td>
<td>116.0</td>
<td>133.0</td>
<td>151.1</td>
</tr>
<tr>
<td>Cloud Management and Security Services</td>
<td>10.5</td>
<td>12.0</td>
<td>13.8</td>
<td>15.7</td>
<td>17.6</td>
</tr>
<tr>
<td>Cloud System Infrastructure Services (IaaS)</td>
<td>32.4</td>
<td>40.3</td>
<td>50.0</td>
<td>61.3</td>
<td>74.1</td>
</tr>
<tr>
<td><strong>Total Market</strong></td>
<td><strong>196.7</strong></td>
<td><strong>227.8</strong></td>
<td><strong>266.4</strong></td>
<td><strong>308.5</strong></td>
<td><strong>354.6</strong></td>
</tr>
</tbody>
</table>

BPaaS = business process as a service; IaaS = infrastructure as a service; PaaS = platform as a service; SaaS = software as a service

**Worldwide Public Cloud Service Revenue Forecast (Billions of U.S. Dollars)**
Microsoft Azure Global Network

60+ regions  
100 Gbps bandwidth  
130,000 miles of fiber optics

Real-World Revenue Loss

Lloyd's Estimates the Impact of a U.S. Cloud Outage at $19 Billion

By: Sean Michael Kerner | January 24, 2018

A joint research report from insurance provider Lloyd's of London and the American Institutes for Research (AIR), looks at the potential costs related to a major public cloud outage in the U.S.

As organizations around the world increasingly rely on the cloud, the impact of a public cloud failure is something that insurance companies are now concerned about. A 67-page report released on Jan. 23 from Lloyd's of London and AIR Worldwide provides some insight and estimates on the potential losses from a major cloud services outage—and the numbers are large.

According to the report, a cyber-incident that impacted the operations of one of the top three public cloud providers in the U.S. for three to six days, could result in total losses of up to $19 billion. Of those loses, only $1.1 to $3.5 billion would be insured, leaving organizations

Cloud Resilience Is Very Crucial!

• State-of-the-art cloud reliability
  • Service Level Agreement (SLA)
  • 5-6 9s’ availability
  • High degree of automation

• Cloud reliability issues
  • Tough cloud failures take a long time to mitigate
  • Impose large revenue loss
  • Harm customer trust and enterprise reputation
Site Reliability Engineering (SRE)

Reliability $R(t) = e^{-\int_0^t \lambda(x) dx}$

Fault Avoidance $\rightarrow$ Fault Removal $\rightarrow$ Fault Tolerance $\rightarrow$ Fault Prediction
Data-Driven AI Applications

Data Models/Paradigms Tasks

- Image classification
- Image localization
- Object detection
- Semantic segmentation

- Machine translation
- Information retrieval
- Question answering
- Sentiment Analysis
- Natural language understanding

- Code summarization
- Code clone detection
- Code suggestion
- API recommendation
- Bug localization
- Semantic parsing
## Cloud Generates a Variety of Data

### Application Layer
- **Application**
- **Microservice**
- **Function**

### Platform Layer
- **Container**
- **Orchestration**
- **Database**

### Infrastructure Layer
- **Compute**
- **Networking**
- **Storage**
- **Virtual Machine**
- **Physical Machine**

### Support Services
- **Customer Service**
- **On-call Engineer**

### Key Features
- **Users**
- **Log**
- **Meter Data**
- **Topology**
- **Alert**
- **Incident Ticket**

24/7 support ensures seamless service delivery.
Challenges of Resilient Cloud Operations

• Current Status:
  • Incidents are highly-correlated, but separately resolved

• Reasons:
  • New DevOps paradigm, complex service dependency, load balance, backup and restore

Humans are not good at solving this large-scale complex problem, but AI is
AIOps for Cloud Resilience

**Log**
- Abnormal
- Network int
- Traffic burs

**Meter Data**

**Topology**

**Alert**

**Incident Ticket**
- Low
- High
- Medium

**Fault Avoidance**

**Fault Prediction**

**Anomaly Detection**

**Fault Removal**

**Fault Tolerance**

**Log Meter Data Topology Alert Incident Ticket**

**Raw Log Messages**

<table>
<thead>
<tr>
<th>Log Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008-11-11 03:40:58 BLOCK* NameSystem.allocateBlock: /user/root/randtxt4/_temporary/_task_200811101024_0010_m_00011_0/part-00011.blk_904791815409399662</td>
</tr>
<tr>
<td>2008-11-11 03:41:48 PacketResponder 0 for block blk_904791815409399662 terminating</td>
</tr>
<tr>
<td>2008-11-11 03:41:48 Received block blk_904791815409399662 of size 67108864 from 10.250.18.114</td>
</tr>
<tr>
<td>2008-11-11 03:41:48 PacketResponder 1 for block blk_904791815409399662 terminating</td>
</tr>
<tr>
<td>2008-11-11 03:41:48 Received block blk_904791815409399662 of size 67108864 from 10.251.43.210</td>
</tr>
<tr>
<td>2008-11-11 03:41:48 BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.43.210:50010 is added to blk_904791815409399662 size 67108864</td>
</tr>
<tr>
<td>2008-11-11 03:41:48 BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.250.18.114:50010 is added to blk_904791815409399662 size 67108864</td>
</tr>
<tr>
<td>2008-11-11 08:30:54 Verification succeeded for blk_904791815409399662</td>
</tr>
</tbody>
</table>
Main Contents in This Talk

- Outage Prediction
  - Incident Management
    - Alert Aggregation
    - Empirical Study
  - Service Dependency
    - Correlation Mining
    - Root Cause Analysis
  - KPI Analysis
    - Multivariate Analysis
    - Problem Identification
  - Log Analysis
    - Log Parsing
    - Log Anomaly Detection
    - Log-based Failure Diagnosis

High-level

Low-level

Easy

Hard
AIOps: Log Analysis

Log

Meter Data

Topology

Alert

Incident Ticket

Anomaly Detection

Failure Diagnosis

Root Cause Analysis

Failure Prediction
Log Parsing: Preprocessing of Log Data

- **Objective**
  - transform raw log data to structural data

- **Key problem to solve**
  - extract event type and variables in log messages

---

### Log Events

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event 1</td>
<td>BLOCK* NameSystem.allocateBlock: *</td>
</tr>
<tr>
<td>Event 2</td>
<td>Receiving block * src: * dest: *</td>
</tr>
<tr>
<td>Event 3</td>
<td>PacketResponder * for block * terminating</td>
</tr>
<tr>
<td>Event 4</td>
<td>Received block * of size * from *</td>
</tr>
<tr>
<td>Event 5</td>
<td>BLOCK* NameSystem.addStoredBlock: blockMap updated: * is added to * size *</td>
</tr>
<tr>
<td>Event 6</td>
<td>Verification succeeded for *</td>
</tr>
</tbody>
</table>

### Structured Logs

<table>
<thead>
<tr>
<th>Log</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>blk_904791815409399662</td>
<td>Event 1</td>
</tr>
<tr>
<td>blk_904791815409399662</td>
<td>Event 2</td>
</tr>
<tr>
<td>blk_904791815409399662</td>
<td>Event 3</td>
</tr>
<tr>
<td>blk_904791815409399662</td>
<td>Event 4</td>
</tr>
<tr>
<td>blk_904791815409399662</td>
<td>Event 5</td>
</tr>
<tr>
<td>blk_904791815409399662</td>
<td>Event 6</td>
</tr>
</tbody>
</table>

---

Log Anomaly Detection

• Feature Engineering

Parsed logs

Log Partition

Fixed windows

Sliding windows

Identifier partition

Δt

Δt

Δt

Δt

Feature Extraction

Model Training

Anomaly detection

Identifier no.

1 1 1 1

2 2 1 1

2 2 1 1

Identifiers
## Log Anomaly Detection

<table>
<thead>
<tr>
<th>Methods</th>
<th>Algorithm/Model</th>
<th>Feature</th>
<th>Unsupervised</th>
<th>Online</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al. [180]</td>
<td>PCA</td>
<td>★ †</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Lin et al. [108]</td>
<td>Clustering</td>
<td>*</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>He et al. [75]</td>
<td>Clustering</td>
<td>* ★</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Liang et al. [104]</td>
<td>SVM</td>
<td>‡</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Kimura et al. [91]</td>
<td>SVM</td>
<td>‡</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Xu et al. [179]</td>
<td>Frequent pattern mining</td>
<td>★ ★</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Shang et al. [161]</td>
<td>Frequent pattern mining</td>
<td>★</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Lou et al. [125]</td>
<td>Frequent pattern mining</td>
<td>★</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Farshchi et al. [54]</td>
<td>Frequent pattern mining</td>
<td>★</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Nandi et al. [145]</td>
<td>Graph mining</td>
<td>¶</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Lou et al. [124]</td>
<td>Graph mining</td>
<td>¶</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yamanishi et al. [181]</td>
<td>Statistical model</td>
<td>★</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>He et al. [76]</td>
<td>Logistic regression</td>
<td>★</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Du et al. [46]</td>
<td>LSTM model</td>
<td>* †</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Zhang et al. [196]</td>
<td>LSTM classification model</td>
<td>*</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Meng et al. [136]</td>
<td>LSTM model</td>
<td>* ★</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Xia et al. [177]</td>
<td>LSTM-based GAN model</td>
<td>*</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Lu et al. [128]</td>
<td>CNN model</td>
<td>*</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Liu et al. [109]</td>
<td>Graph embedding model</td>
<td>¶</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

* Log event sequence, ★ Log event count vector, † Parameter value vector
‡ Ad hoc features, ¶ Graphical feature

Log-based Failure Diagnosis for Cloud System

- Log is the major source for failure diagnosis
Failure Diagnosis: Ranking Buggy Functions

- PCA algorithm to find abnormal components

T. Zaman, X. Han, T. Yu, “SCMiner: Localizing System-Level Concurrency Faults from Large System Call Traces”, ASE 2019
AIOps: KPIs Analysis

- Log
- Meter Data
- Topology
- Alert
- Incident Ticket

Anomaly Detection
Failure Diagnosis
Root Cause Analysis
Failure Prediction
Key Performance Indicators (KPIs)

network traffic
response delay
CPU usage

monitor runtime information
understand health status
anomaly detection

system anomaly

CPU LOAD
ETH INFLOW
Multivariate KPIs Analysis

- Should capture dependency of multivariate KPIs
- Unsupervised anomaly detection

![Graph showing time series of various performance metrics with anomaly detection]

- CPU LOAD
- ETH INFLOW
- MEMORY USAGE
- DISK WRITE

anomaly

normal
Machine Learning Algorithms

- **Training:**
  - KPIs
  - Sliding windows
  - Model
  - Prediction
  - Minimize
  - Ground truth

- **Detection:**
  - Model
  - Prediction
  - Difference
  - Anomaly
  - Normal

KPIs

Prediction

Minimize

Ground truth

Difference

Anomaly

Normal

Normal

Anomaly

Entity Anomaly Score

Threshold

Observation

Prediction
AIOps: Correlation between Logs and KPIs

- Log
- Meter Data
- Topology
- Alert
- Incident Ticket

- Anomaly Detection
- Failure Diagnosis
- Root Cause Analysis
- Failure Prediction
Two Automated Log Analysis Tasks

Anomaly Detection (binary classification)

Problem Identification (multiclass classification)

- Normal
- Anomalous

- Normal
- Different types of problem
Efficient Multi-class Classification / Clustering

• Efficient and effective cascading clustering

Hierarchical Clustering

Relation between Log and KPI

KPIs

Time Intervals

Logs
Problem Identification

- **Impactful problems:**
  - Can lead to the degradation of KPI.

- **Target:**
  - Identify clusters that are highly correlated with KPI’s changes.

- **Method:**
  - Model the relation between cluster sizes and KPI values

---

Problem Identification

- Evaluation on real Microsoft Azure data

<table>
<thead>
<tr>
<th>Data</th>
<th>Snapshot starts</th>
<th>#Log Seq (Size)</th>
<th>#Events</th>
<th>#Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>Sept 5th 10:50</td>
<td>359,843 (722MB)</td>
<td>365</td>
<td>16</td>
</tr>
<tr>
<td>Data 2</td>
<td>Oct 5th 04:30</td>
<td>472,399 (996MB)</td>
<td>526</td>
<td>21</td>
</tr>
<tr>
<td>Data 3</td>
<td>Nov 5th 18:50</td>
<td>184,751 (407MB)</td>
<td>409</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 1: Summary of Service X Log Data

<table>
<thead>
<tr>
<th>Data</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>0.465</td>
<td>0.946</td>
<td>0.623</td>
<td>0.142</td>
<td>0.834</td>
<td>0.242</td>
<td>0.207</td>
<td>0.922</td>
<td>0.338</td>
</tr>
<tr>
<td>Invariants Mining</td>
<td>0.604</td>
<td>1</td>
<td>0.753</td>
<td>0.160</td>
<td>0.847</td>
<td>0.269</td>
<td>0.168</td>
<td>0.704</td>
<td>0.271</td>
</tr>
<tr>
<td>Log3C</td>
<td><strong>0.900</strong></td>
<td>0.920</td>
<td><strong>0.910</strong></td>
<td><strong>0.897</strong></td>
<td>0.826</td>
<td><strong>0.860</strong></td>
<td><strong>0.834</strong></td>
<td>0.903</td>
<td><strong>0.868</strong></td>
</tr>
</tbody>
</table>

Table 2: Accuracy of Problem Detection on Service X Data

AIOps: Service Dependency

- Log
- Meter Data
- Topology
- Alert
- Incident Ticket

- Anomaly Detection
- Failure Diagnosis
- Root Cause Analysis
- Failure Prediction
From Correlation to Root Cause Investigation

W. Ping, J. Xu, M. Ma, W. Lin, D. Pan, Y. Wang, and P. Chen. 'CloudRanger: Root Cause Identification for Cloud Native Systems'. CCGRID 2018
Root Cause Analysis: Service Call Graph

- Metric data: response time, error counts, queries per seconds
- Anomaly propagation chains
- Rank candidate root causes based on correlation analysis

AIOps: Alert Aggregation

- Log
- Meter Data
- Topology
- Alert
- Incident Ticket

Anomaly Detection
Failure Diagnosis
Root Cause Analysis
Failure Prediction
Objectives

- Alert aggregation
  - Group alerts associated the same failure
  - Narrow down the problem scope

- Root cause recommendation
  - Recommend culprit incidents
  - Speed up fault localization

![System topology](image1)
A failure occurs to service A
Cascading effect of the failure
Graph Representation Learning

- Fine-grained cloud monitoring data to auto-complete the graphs
- Temporal and topological relationship to learn the alert representation vector
Graph Representation Learning

- Fine-grained cloud monitoring data to auto-complete the graphs
- Temporal and topological relationship to learn the alert representation vector

<table>
<thead>
<tr>
<th></th>
<th>NMI</th>
<th>TF-IDF</th>
<th>Zhao's approach</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Online Incident Aggregation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP-Growth</td>
<td>0.42</td>
<td>N/A</td>
<td>0.61</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Root Cause Recommendation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>N/A</td>
<td>0.73</td>
<td>0.81</td>
<td>0.91</td>
</tr>
<tr>
<td>Recall</td>
<td>N/A</td>
<td>0.77</td>
<td>0.88</td>
<td>0.93</td>
</tr>
<tr>
<td>F1 score</td>
<td>N/A</td>
<td>0.75</td>
<td>0.85</td>
<td>0.92</td>
</tr>
</tbody>
</table>

A real case in a top public cloud
AIOps: Incident Management

- Log
- Meter Data
- Topology
- Alert
- Incident Ticket

Anomaly Detection

Failure Diagnosis

Root Cause Analysis

Failure Prediction
Inefficient and Error-prone Workflow

• Significant delays
  o Critical incident detection
  o Impact scope identification
  o Root cause analysis
  o etc.

• Complicated root causes
  o Multi-location
  o Multi-source
  o Multi-layer
  o etc.
Incident Management

Incident management procedure

- Incident reporting
  - Time to detect (TTD)

- Incident triage
  - Time to engage (TTE)

- Incident mitigation
  - Time to mitigate (TTM)
Incident Mitigation

• Incident mitigation is important yet challenging
  • Large volume of incidents
  • Cross-region failures
  • Cloud system complexity
  • etc.
Characteristics of Incidents

• Incident severity
  • Low + Medium incidents > 90%
  • High incidents from 1.21% (Network) to 5.48% (DCM)
  • Critical incidents < 0.5%

<table>
<thead>
<tr>
<th></th>
<th>DCM</th>
<th>Network</th>
<th>Storage</th>
<th>Compute</th>
<th>Database</th>
<th>WS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.31%</td>
<td>0.40%</td>
<td>0.07%</td>
</tr>
<tr>
<td>High</td>
<td>5.48%</td>
<td>1.21%</td>
<td>2.57%</td>
<td>5.27%</td>
<td>4.32%</td>
<td>3.33%</td>
</tr>
<tr>
<td>Medium</td>
<td>86.65%</td>
<td>46.90%</td>
<td>43.32%</td>
<td>74.19%</td>
<td>63.93%</td>
<td>84.52%</td>
</tr>
<tr>
<td>Low</td>
<td>7.86%</td>
<td>51.88%</td>
<td>54.10%</td>
<td>20.23%</td>
<td>31.35%</td>
<td>12.08%</td>
</tr>
</tbody>
</table>

Distribution of incident severity

Chen et al., ‘Towards Intelligent Incident Management: Why We Need It and How We Make It’. FSE 2020
Characteristics of Incidents

• Incident fixing time
  • Time to fix (TTF) = TTD+TTE+TTM
  • TTF of Low & Medium incidents > TTF of High incidents
  • TTF of Critical is the largest

<table>
<thead>
<tr>
<th></th>
<th>DCM</th>
<th>Network</th>
<th>Storage</th>
<th>Compute</th>
<th>Database</th>
<th>WS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>38.33x</td>
<td>8.46x</td>
<td>10.06x</td>
<td>142.05x</td>
<td>209.97x</td>
<td>286.6x</td>
</tr>
<tr>
<td>High</td>
<td>19.25x</td>
<td>3.18x</td>
<td>2.52x</td>
<td>2.56x</td>
<td>5.75x</td>
<td>3.56x</td>
</tr>
<tr>
<td>Medium</td>
<td>1x</td>
<td>9.8x</td>
<td>7.09x</td>
<td>2.95x</td>
<td>25.28x</td>
<td>12.93x</td>
</tr>
<tr>
<td>Low</td>
<td>3.01x</td>
<td>5.49x</td>
<td>1.09x</td>
<td>11.65x</td>
<td>2.41x</td>
<td>144.79x</td>
</tr>
</tbody>
</table>

Distribution of incident fixing time

Chen et al., ‘Towards Intelligent Incident Management: Why We Need It and How We Make It’. FSE 2020
Characteristics of Incidents

• Root Cause:
  • Network Issue
  • Human Error
  • Deployment Issue
  • External Issue
  • Capacity Issue
  • Others

<table>
<thead>
<tr>
<th>Root Cause</th>
<th>Dist.</th>
<th>Root Cause</th>
<th>Dist.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network (Hardware)</td>
<td>22.95%</td>
<td>Human Error (Code Defect)</td>
<td>19.23%</td>
</tr>
<tr>
<td>Network (Connectivity)</td>
<td>2.24%</td>
<td>Human Error (Config.)</td>
<td>7.45%</td>
</tr>
<tr>
<td>Network (Config.)</td>
<td>0.89%</td>
<td>Human Error (Design Flaw)</td>
<td>5.66%</td>
</tr>
<tr>
<td>Network (Other)</td>
<td>4.47%</td>
<td>Human Error (Integration)</td>
<td>2.09%</td>
</tr>
<tr>
<td>Deployment (Upgrade)</td>
<td>5.22%</td>
<td>Human Error (Other)</td>
<td>2.83%</td>
</tr>
<tr>
<td>Deployment (Config.)</td>
<td>3.87%</td>
<td>External Issue (Partner)</td>
<td>2.83%</td>
</tr>
<tr>
<td>Deployment (Other)</td>
<td>1.19%</td>
<td>External Issue (Other)</td>
<td>1.64%</td>
</tr>
<tr>
<td>Capacity Issue</td>
<td>6.56%</td>
<td>Others</td>
<td>10.88%</td>
</tr>
</tbody>
</table>

Distribution of incident root causes

Chen et al., ‘Towards Intelligent Incident Management: Why We Need It and How We Make It’. FSE 2020
AIOps: Outage Prediction

Log

Meter Data

Topology

Alert

Incident Ticket

Anomaly Detection

Failure Diagnosis

Root Cause Analysis

Failure Prediction
Alerts vs Outage

Fault-tolerance Architecture

Outage happens!!
Causal Relationship between Alerts and Outage

- Historical failure statistics
  - Build dependency among alert signals
  - Train classification model to predict outage

Bayesian Network

Classification models to link alerts and outages

Historical failure statistics

Build dependency among alert signals

Train classification model to predict outage

# Causal Relationship between Alerts and Outage

Table 1: Comparison of different methods for component-level outage prediction.

<table>
<thead>
<tr>
<th>Method</th>
<th>Outage (Storage Location)</th>
<th>Outage (Physical Networking)</th>
<th>Outage (Storage Streaming)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1-score</td>
</tr>
<tr>
<td>Simple Spike</td>
<td>61.65</td>
<td>100.00</td>
<td>76.28</td>
</tr>
<tr>
<td>PLR</td>
<td>70.02</td>
<td>92.71</td>
<td>79.78</td>
</tr>
<tr>
<td>SVM</td>
<td>65.65</td>
<td>95.83</td>
<td>77.92</td>
</tr>
<tr>
<td>AirAlert Related</td>
<td>65.31</td>
<td>100.00</td>
<td>79.01</td>
</tr>
<tr>
<td>AirAlert Full</td>
<td>71.11</td>
<td>100.00</td>
<td>83.17</td>
</tr>
</tbody>
</table>

Table 2: Comparison of different methods for service-level outage prediction.

<table>
<thead>
<tr>
<th>Method</th>
<th>Outage (Website Application)</th>
<th>Outage (Cloud Network)</th>
<th>Outage (Microsoft Cloud System Operation)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1-score</td>
</tr>
<tr>
<td>Simple Spike</td>
<td>5.73</td>
<td>11.83</td>
<td>7.72</td>
</tr>
<tr>
<td>PLR</td>
<td>61.18</td>
<td>54.17</td>
<td>57.46</td>
</tr>
<tr>
<td>SVM</td>
<td>66.41</td>
<td>88.54</td>
<td>75.89</td>
</tr>
<tr>
<td>AirAlert Related</td>
<td>92.18</td>
<td>85.63</td>
<td>88.78</td>
</tr>
<tr>
<td>AirAlert Full</td>
<td>82.75</td>
<td>76.74</td>
<td>79.63</td>
</tr>
</tbody>
</table>

Conclusions

• Why cloud resilience needs AIOps?
  • Endless pursuit of reliability
  • From automatic to intelligent, from reactive to proactive
  • Important data sources: log, meter data, topology, alert and incident ticket

• How AIOps achieves reliability goals?
  • Endless pursuit of advanced algorithms
  • From anomaly detection, failure diagnosis, root cause analysis to failure prediction
  • Intelligent algorithms designed with human experts’ experiences

• What’s the next?
  • How to integrate human knowledge with algorithms automatically and comprehensively?
  • Further investigations on AI and Software Engineering
Thank you!

ICSE21 Workshop on Cloud Intelligence
In conjunction with the 43rd International Conference on Software Engineering

Schedule: 11:00am - 7:30pm CET on May 29th, 2021