Auto-PIP: Real-time Identification of Critical Performance Inflection Points in Software Stress Testing

Shenglin Zhang[1], Xiao Xiong[1], Mengyao Li[1], Yongqian Sun[1*], Yongxin Zhao[1],
Xia Chen[2], Bowen Deng[2], Dan Pei[3]

[1]Nankai University, [2]Huawei, [3]Tsinghua University

Table of Content

Background

Design

Evaluation

Deployment

Conclusion

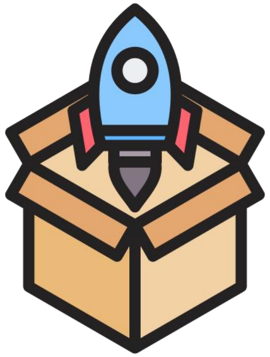
Stress Testing Is Vital to Special Business Events

- The scale and complexity of software systems continue to increase
- Performance problems are becoming more and more significant, especially in some business events
- Lead to service interruption or performance degradation, impacting user experience and bringing losses to companies

Black Friday


Major Sporting Events

Tax Filing Deadlines

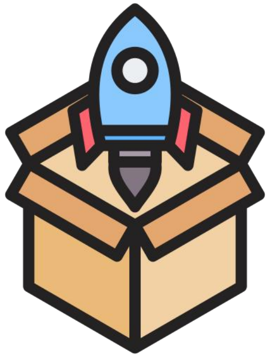

Product Launches

Stress Testing Is Vital to Special Business Events

Black Friday

Major Sporting Events

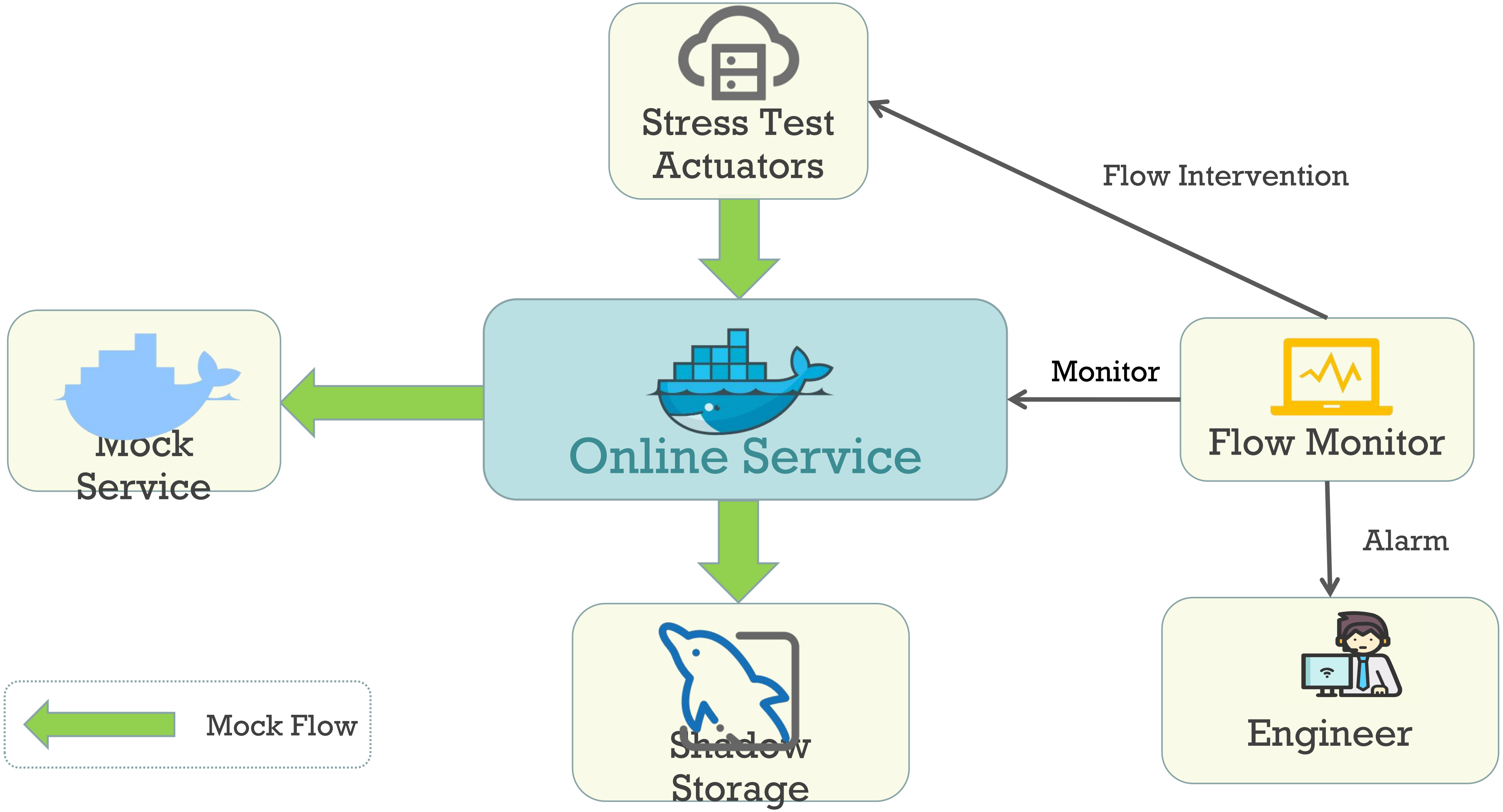
Tax Filing Deadlines



Product Launches

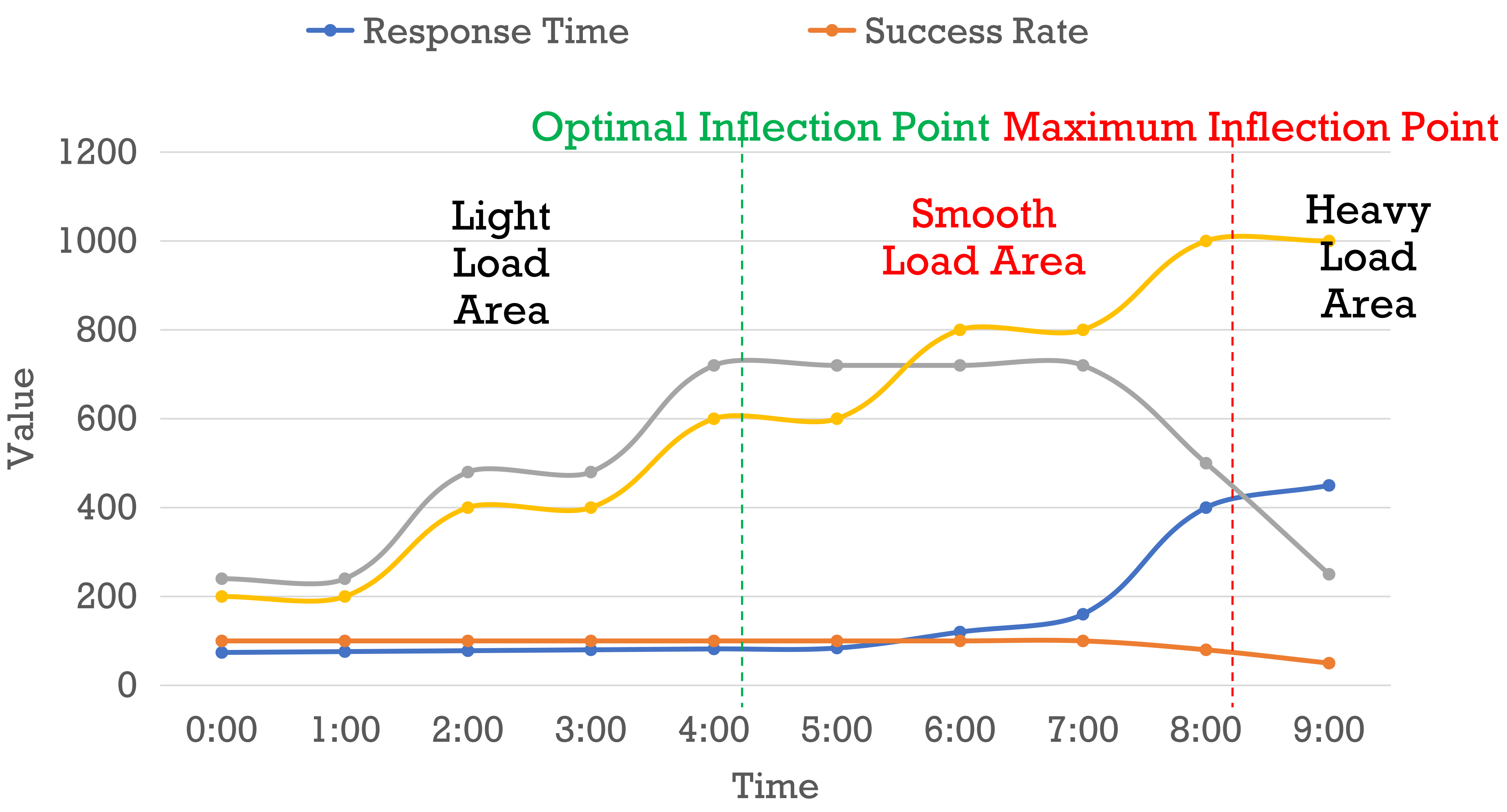
- Through stress testing, engineers can identify bottlenecks and potential issues in the system under high load conditions, optimizing resource allocation and system architecture
- Important to these special business events

Stress Testing to Estimate The Maximum System Capacity



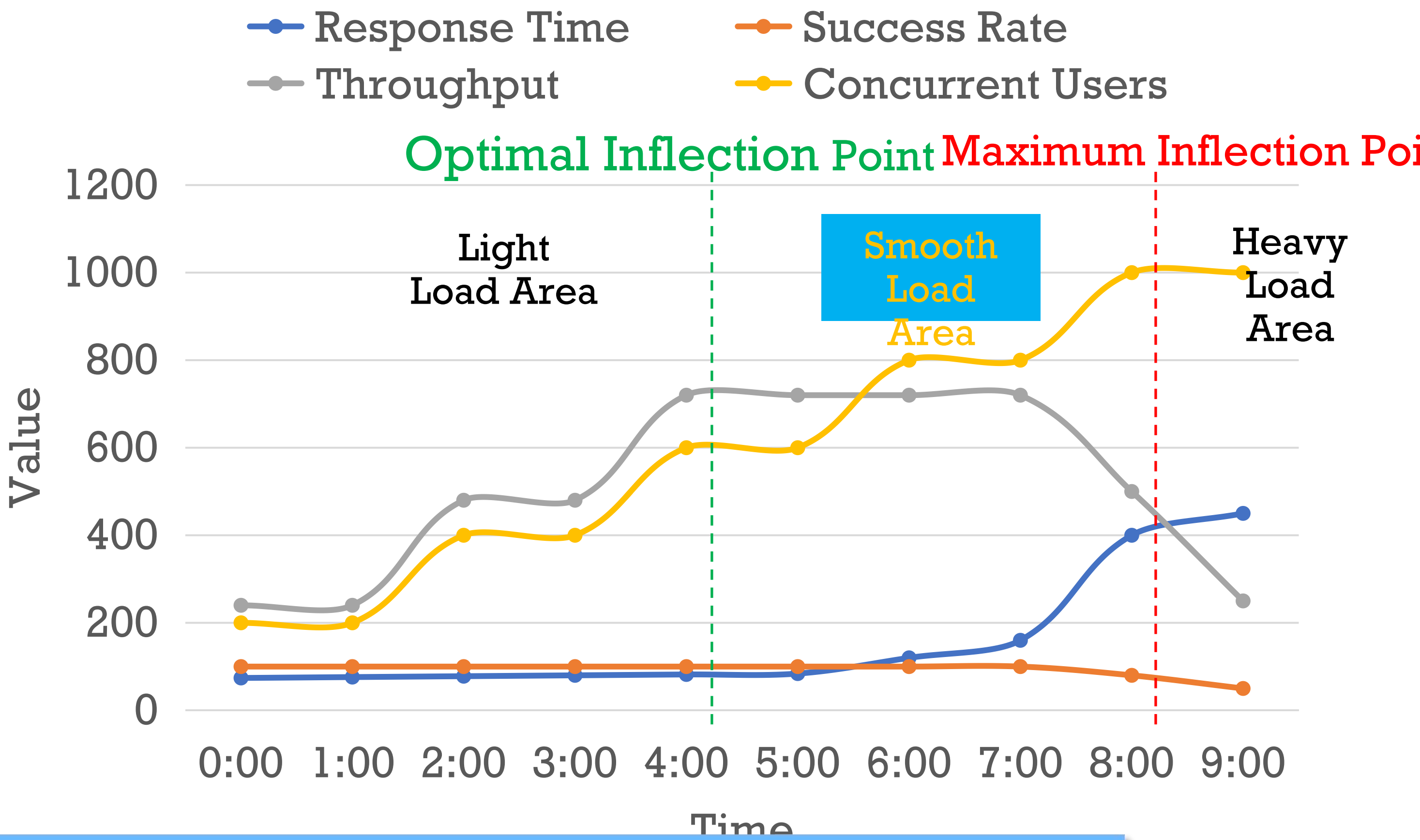
Performance Inflection Point Model

- Engineers determine the operating status of a software system by observing Key performance indicators (KPIs)
  - Response time
  - Success Rate
  - Throughput
- Stress test areas
  - Light load area
  - Smooth load area (SLA)
  - Heavy load area



SLA is Important

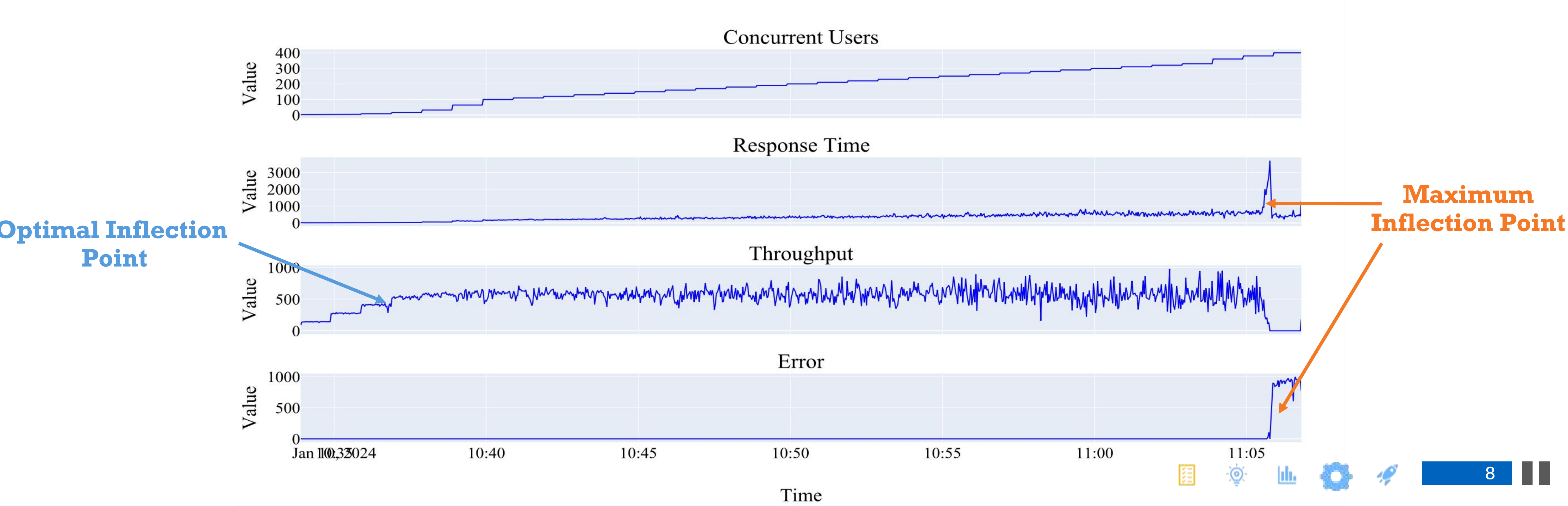
- SLA is the most important
  - Represents the optimal range where the system can handle increasing loads without performance degradation
- SLA is decided by optimal inflection point and maximum inflection point
  - Identifying the optimal inflection point helps reduce the waste of computing resources while ensuring user experience
  - Identifying the maximum inflection



Our goal is real-time identification of optimal and maximum inflection points during software stress testing

Key Ideas

- Optimal inflection point detection
  - As concurrency increases, throughput no longer increases → KPI trend detection
- Maximum inflection point detection
  - As concurrency increases, response time deteriorates rapidly and success rate drops sharply → KPI anomaly detection



Challenges

Challenge 1: Low quality of KPIs

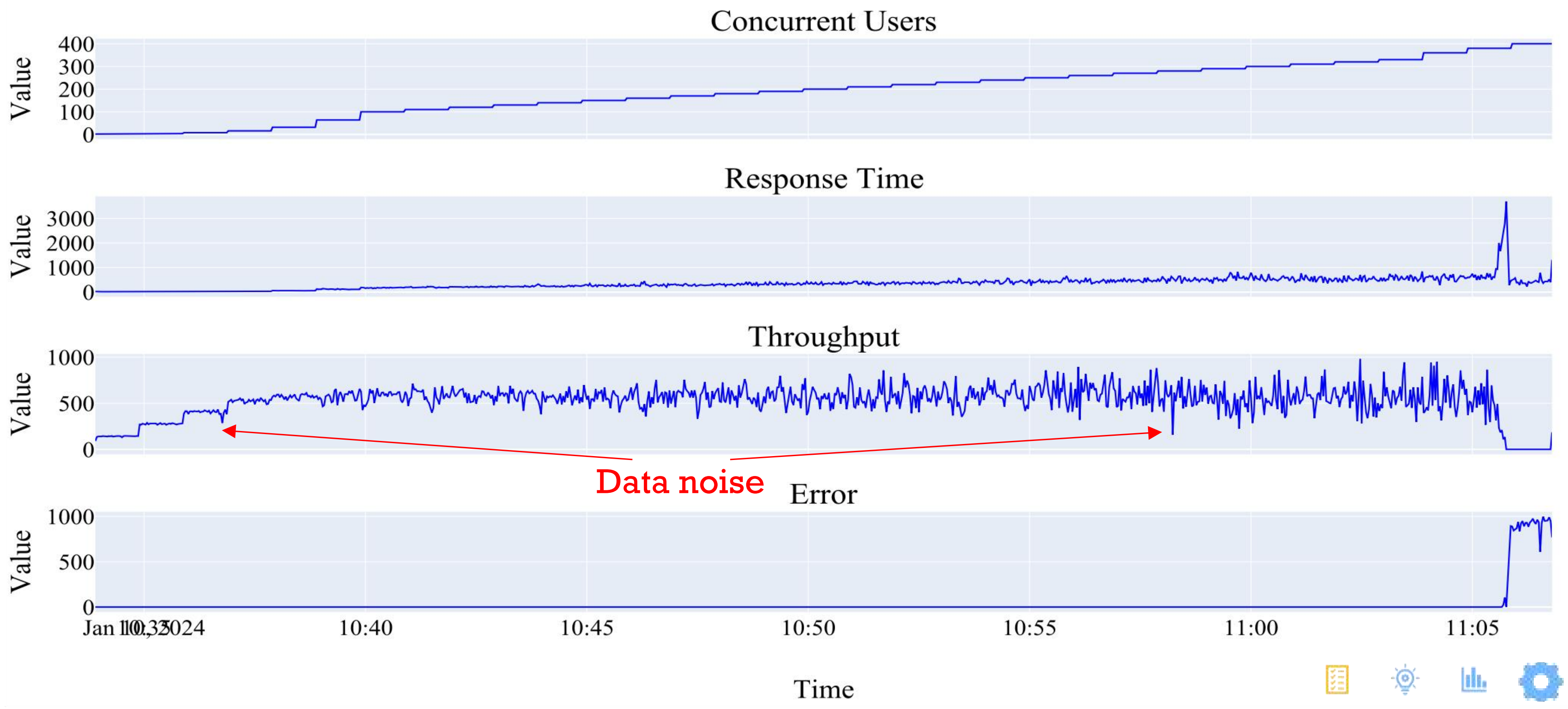
Challenge 2: Short period of KPIs

Challenge 3: Setting KPI thresholds

Challenges

Challenge 1: Low quality of KPIs

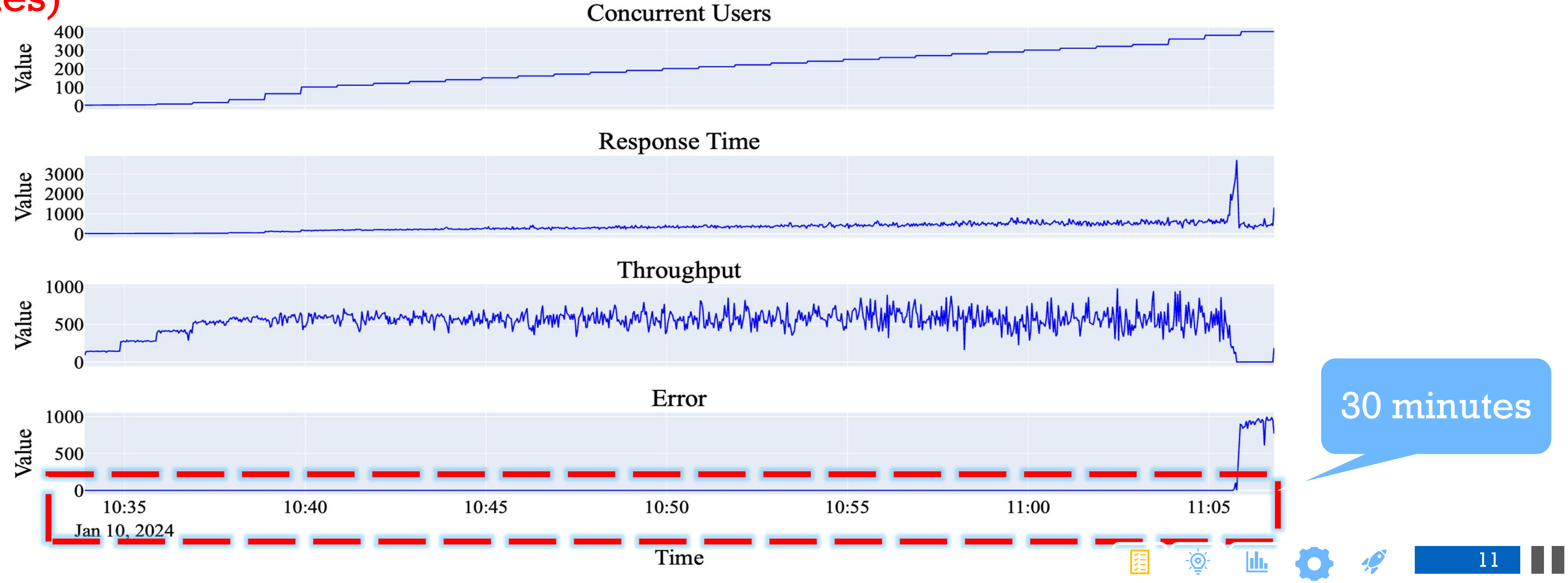
- Errors during the collection and recording of KPIs introduce data noise which can degrade the performance of tread detection and anomaly detection methods



Challenges

Challenge 2: Short period of KPIs

- Advanced anomaly detection methods based on deep learning require long period of training data (e.g., weeks)
- The duration of individual software system stress test is typically short (e.g., tens of minutes)

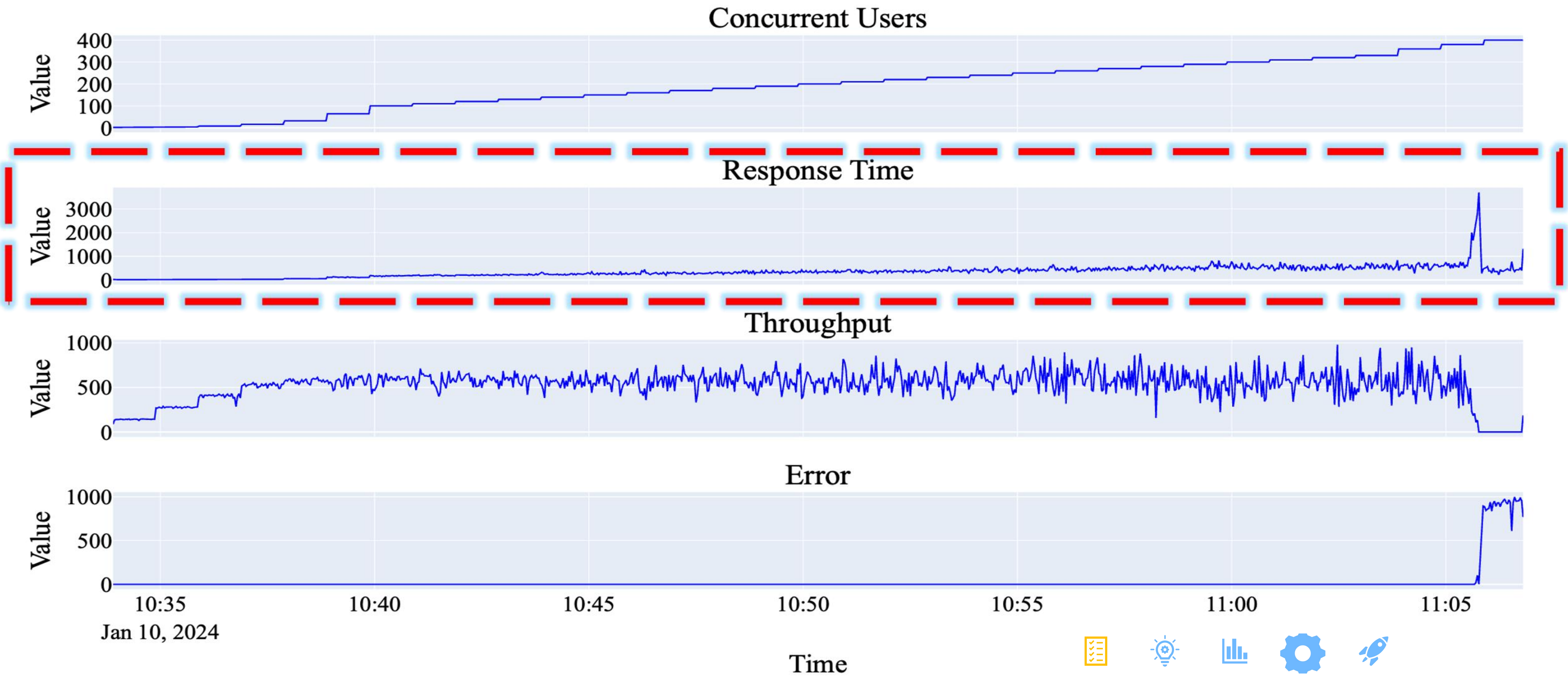


Challenges

Challenge 3: Setting KPI thresholds

- Engineers often need to manually set the threshold for each KPI according to traditional methods, which is time-consuming, labor-intensive, and prone to errors

Different KPIs have different thresholds

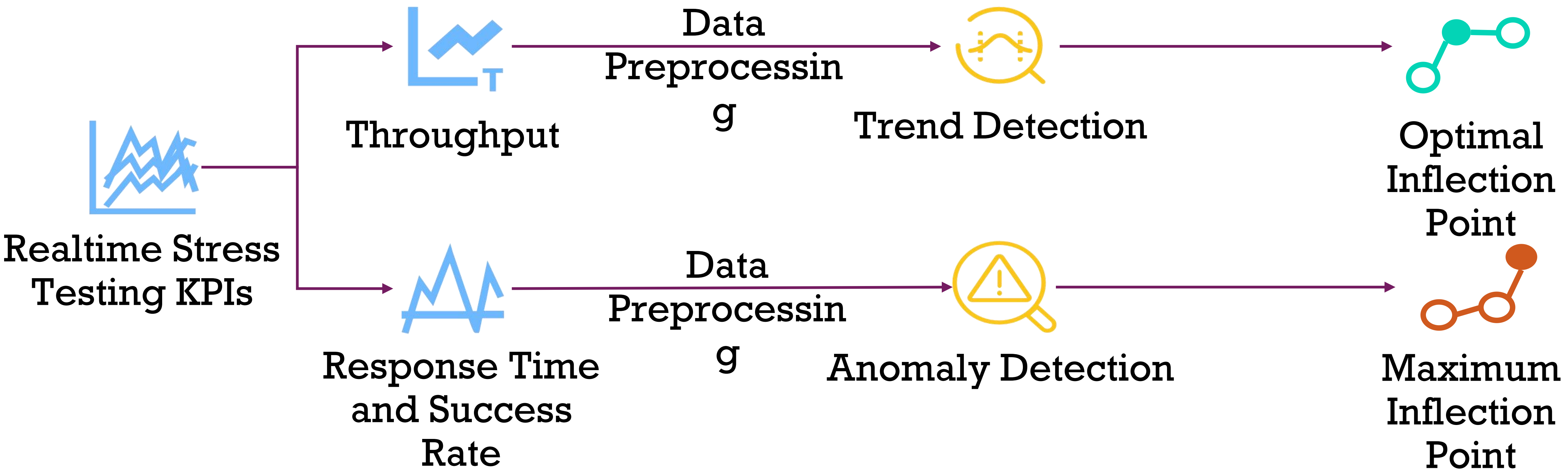


Auto-PIP

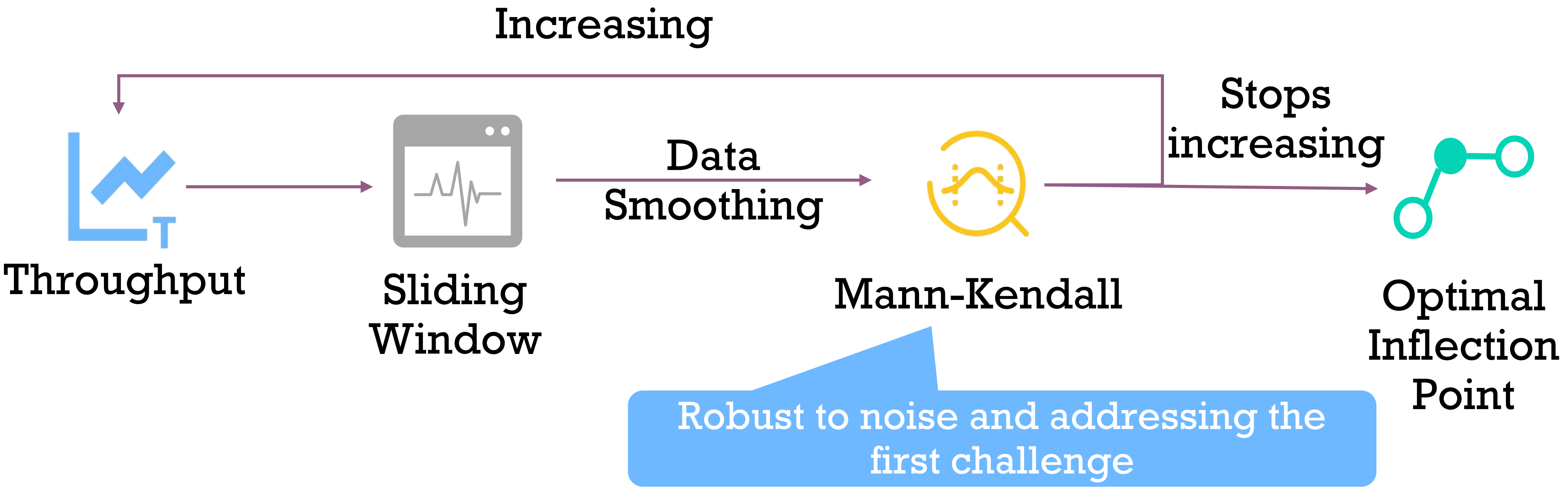
Real-time identification of
critical performance inflection
points
in software stress testing

Table of Content

Design Overview



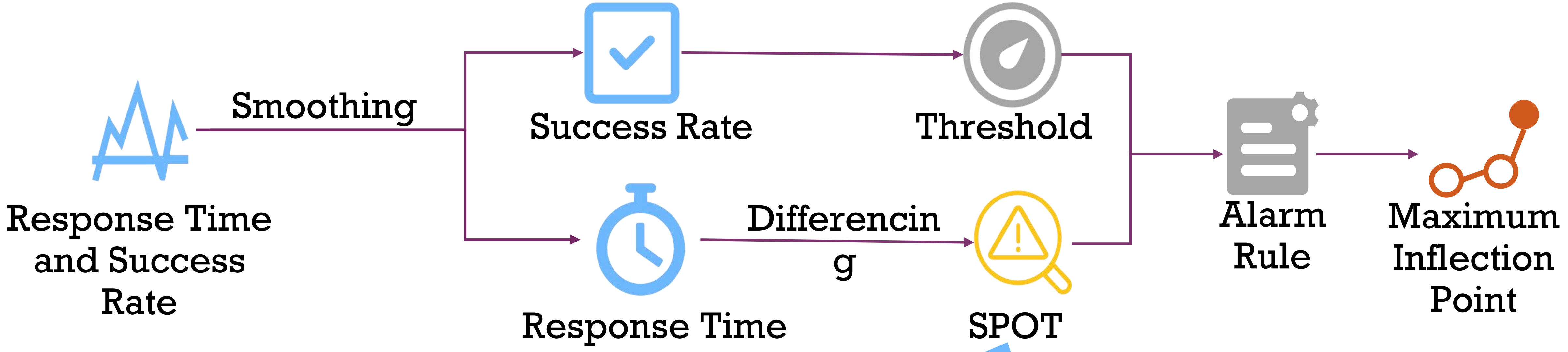
Optimal Inflection Point Identification



Mann-Kendall Test

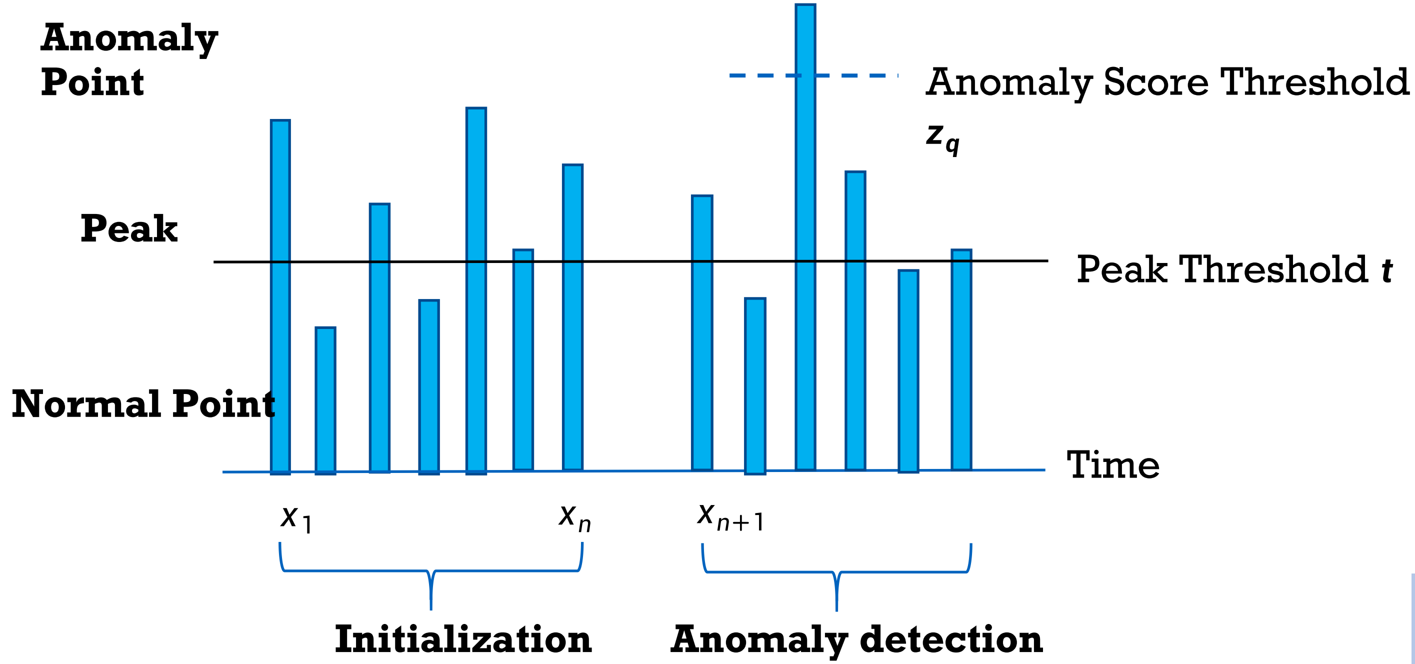
- A classical hypothesis test method in trend detection
- Null hypothesis H0: time series data $(X_1, X_2, \ldots, X_n)$ is a sample of n independent and identically distributed random variables
- Alternative hypothesis H1: there is a monotonic trend
- If the null hypothesis is unacceptable, that is, p-value < significance level α (α=0.05), there is a clear upward or downward trend in time series data

Maximum Inflection Point Identification

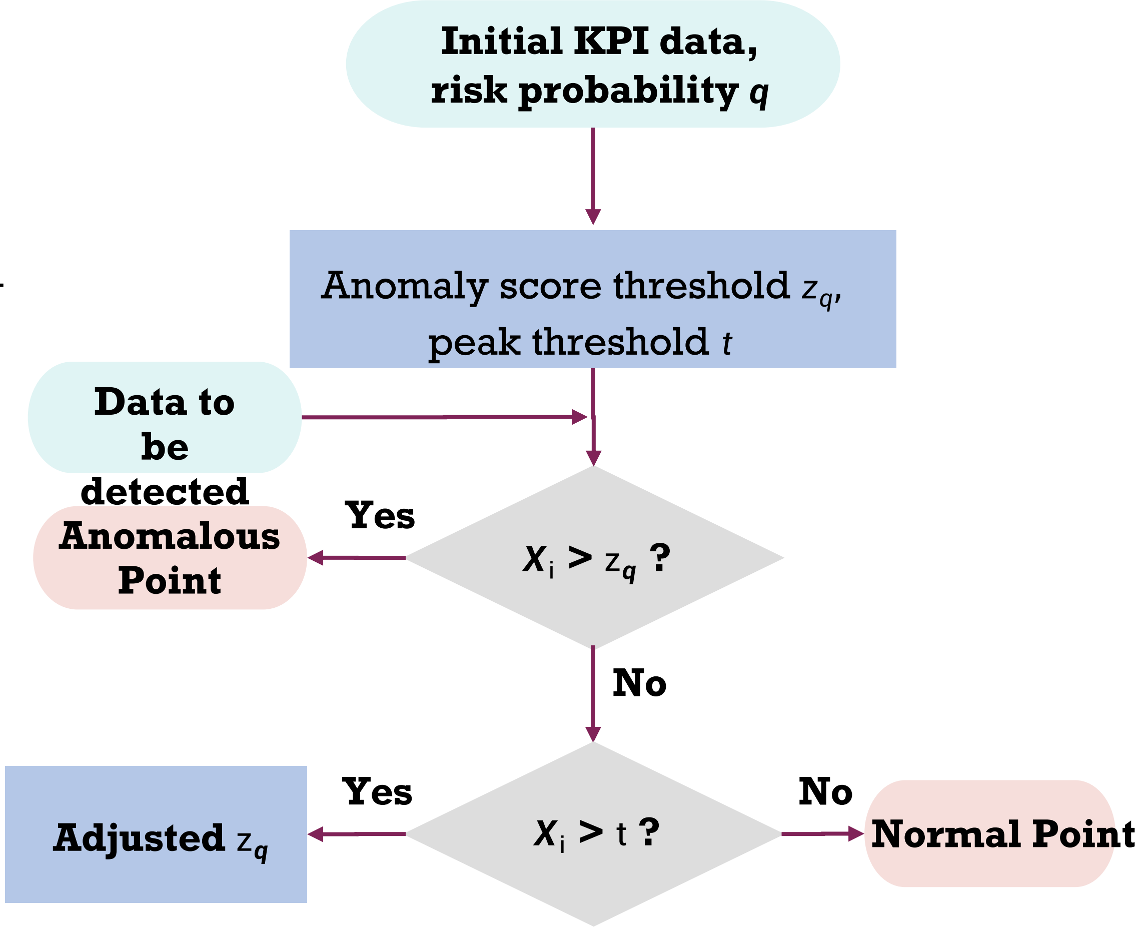


Does not need long-period data to initialize, and automatically adjusts the threshold,
addressing the second and third challenge

SPOT



Anomaly Point
Anomaly Score Threshold z_q
Peak
Peak Threshold t
Normal Point
Time
x_1
x_n
x_{n+1}
Initialization
Anomaly detection



Initial KPI data, risk probability q
Anomaly score threshold z_q, peak threshold t
Data to be detected
$X_i > z_q$?
Yes
Anomalous Point
No
$X_i > t$?
Yes
Adjusted z_q
No
Normal Point

Alarm Rule

- We use queues to maintain suspected maximum inflection points that have been detected recently
- When the proportion of suspicious maximum inflection point in the latest period of **window** reaches the threshold **k**
  - The maximum inflection point has been found
  - The signal to stop the stress test is returned

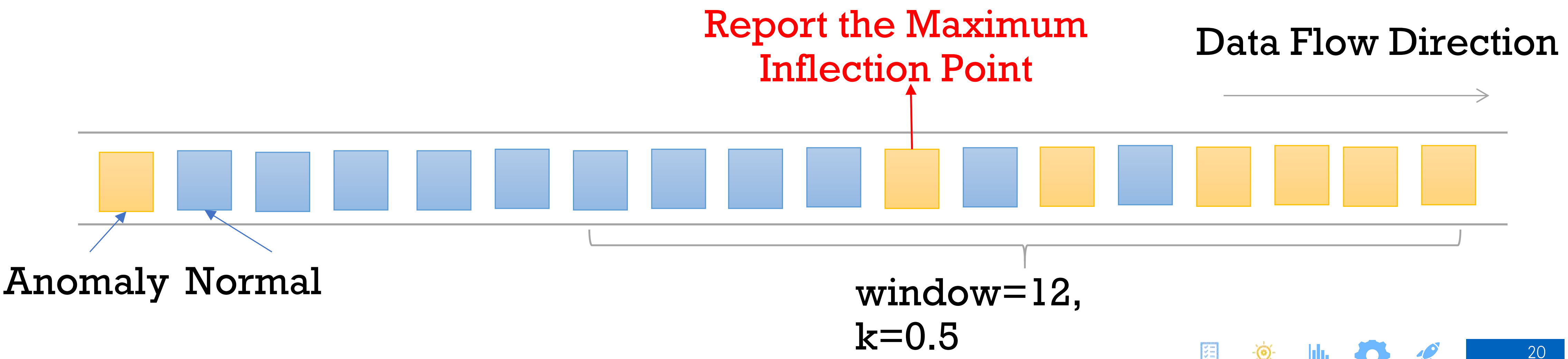


Table of Content

Experiment Setup

- Research questions (RQs)
  - RQ1: How does the performance of Auto-PIP compare to the baseline methods?
  - RQ2: Does each component of Auto-PIP significantly contribute to its performance?
- Dataset
  - A total of 128 stress test cases
  - From Huawei Cloud industrial environment
  - Labeled by experienced test experts

Experiment Setup

- Baseline methods
  - Optimal inflection point: slope method, Cox-Stuart test
  - Maximum inflection point: k-sigma, box plot, Bagel
- Evaluation metric
  - $Accuracy = \frac{1}{|A|}\sum_{a \in A} \mathbb{I}(f(a) \in Y_a)$
  - A is the test case collection
  - Y_a indicates the range of concurrent users corresponding to the inflection point of the test case a
  - $f(a)$ is the number of concurrent users predicted by Auto-PIP

Auto-PIP vs. Baseline Methods (RQ1)

Compared with baseline methods, Auto-PIP is indeed effective and computationally efficient in inflection point identification

Detection Type	Method	Accuracy	Efficiency
Optimal	Slope method [12]	53.3%	0.231s
	Cox-Stuart test [13]	58.3%	**0.220s**
	Auto-PIP	**100%**	0.228s
Maximum	K-sigma [16]	73.2%	2.085s
	Box plot [17]	21.4%	**1.398s**
	Bagel [19]	66.1%	5.830s
	Auto-PIP	**83.9%**	2.229s

Contribution of Key Components (RQ2)

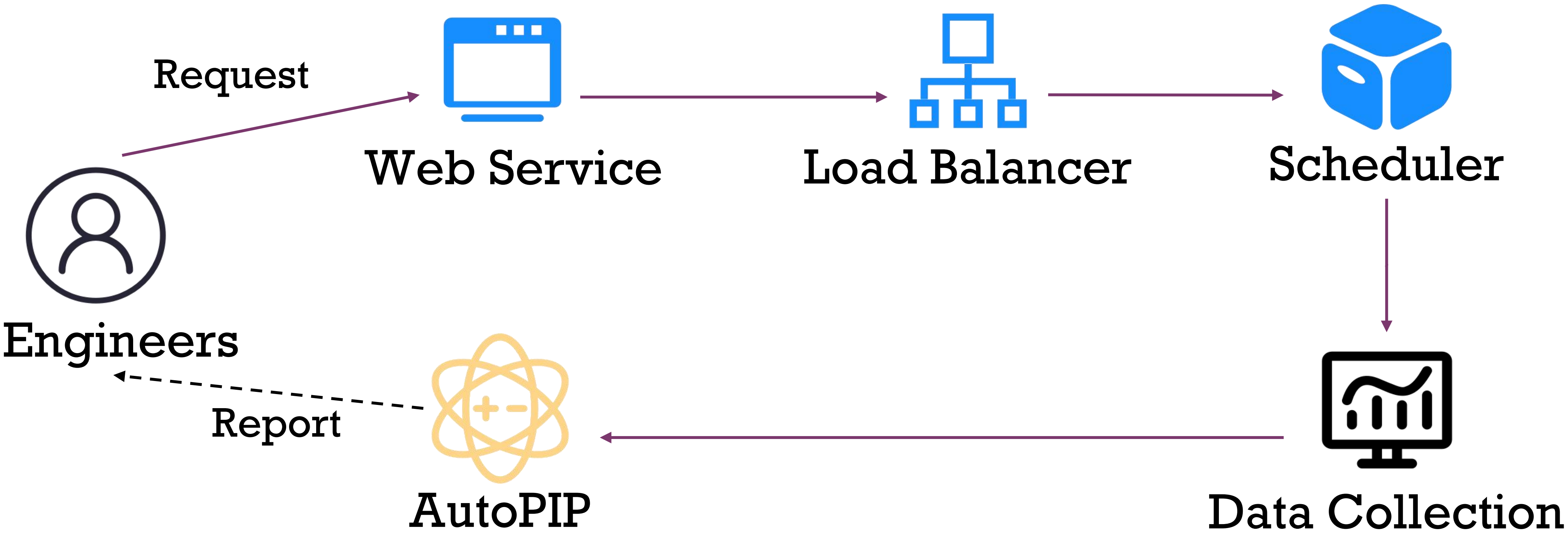
Both the SPOT algorithm and the success rate threshold are critical for identifying the maximum inflection point.

Model	Accuracy	Efficiency
Auto-PIP	**83.9%**	2.229s
Auto-PIP w/o SPOT	60.7%	**0.359s**
Auto-PIP w/o success rate	21.4%	2.790s

Table of Content

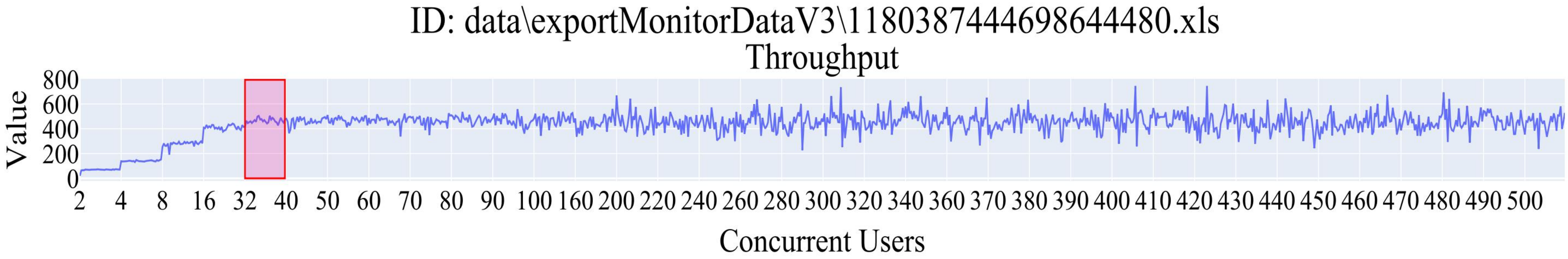
Workflow of Auto-PIP's Deployment

Auto-PIP has been successfully deployed in Huawei Cloud's industrial environment



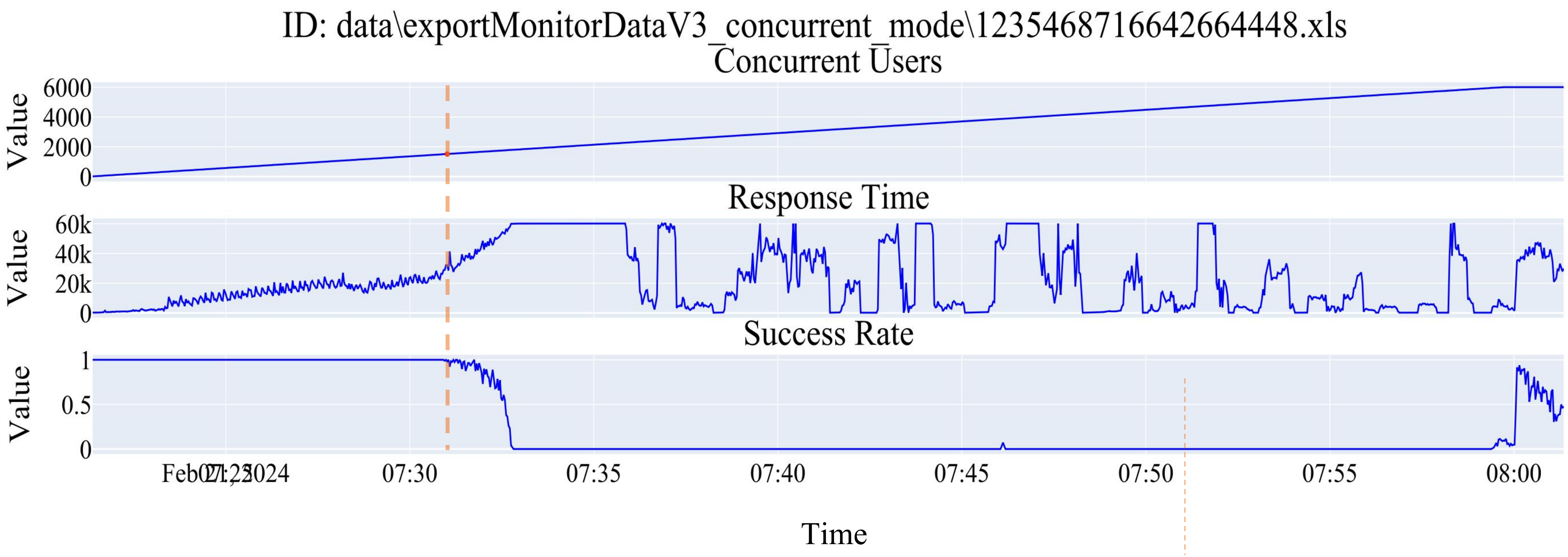
Case Study #1

Identified the optimal inflection point when throughput stopped increasing



Case Study #2

Detected the maximum inflection point when the success rate dropped sharply as the system reached its capacity



Case Study #3

Identified a sudden surge in response time, signaling that the system was becoming overloaded.

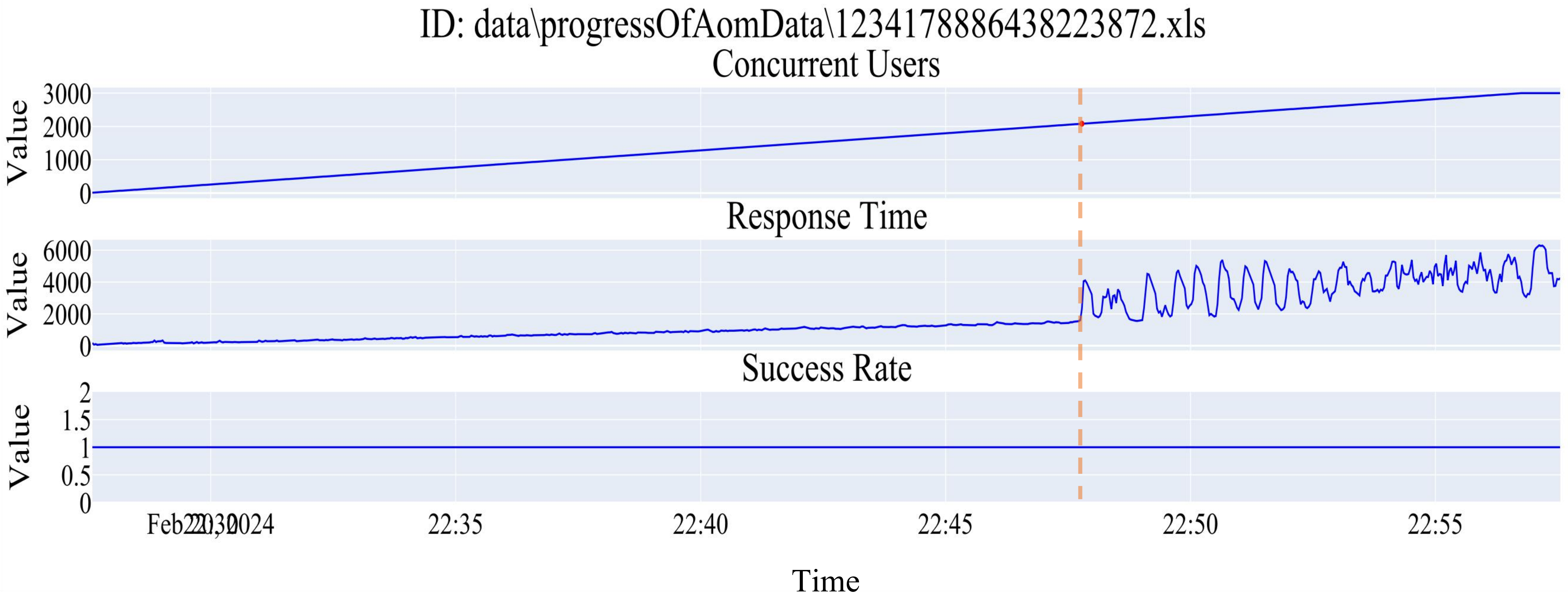


Table of Content

Background

Design

Evaluation

Deployment

Conclusion

Conclusion

Auto-PIP: real-timely identifies critical performance inflection points

- Mann-Kendall test → address the challenge of low quality of KPIs
- SPOT → address the challenges of short period of KPIs and setting KPI thresholds

Evaluation experiments and industrial deployment

- Evaluation conducted on the dataset collected from Huawei Cloud demonstrate the effectiveness and efficiency of Auto-PIP
- Auto-PIP has been successfully deployed in Huawei Cloud to demonstrate its practicability
- We have released our labeled dataset at https://doi.org/10.5281/zenodo.13337204

Thanks

Q&A