Giving Every Modality a Voice in Microservice Failure Diagnosis via Multimodal Adaptive Optimization

Lei Tao[1], Shenglin Zhang, Zedong Jia, Jinrui Sun, Minghua Ma,

Zhengdan Li [2], Yongqian Sun, Canqun Yang, Yuzhi Zhang, Dan Pei

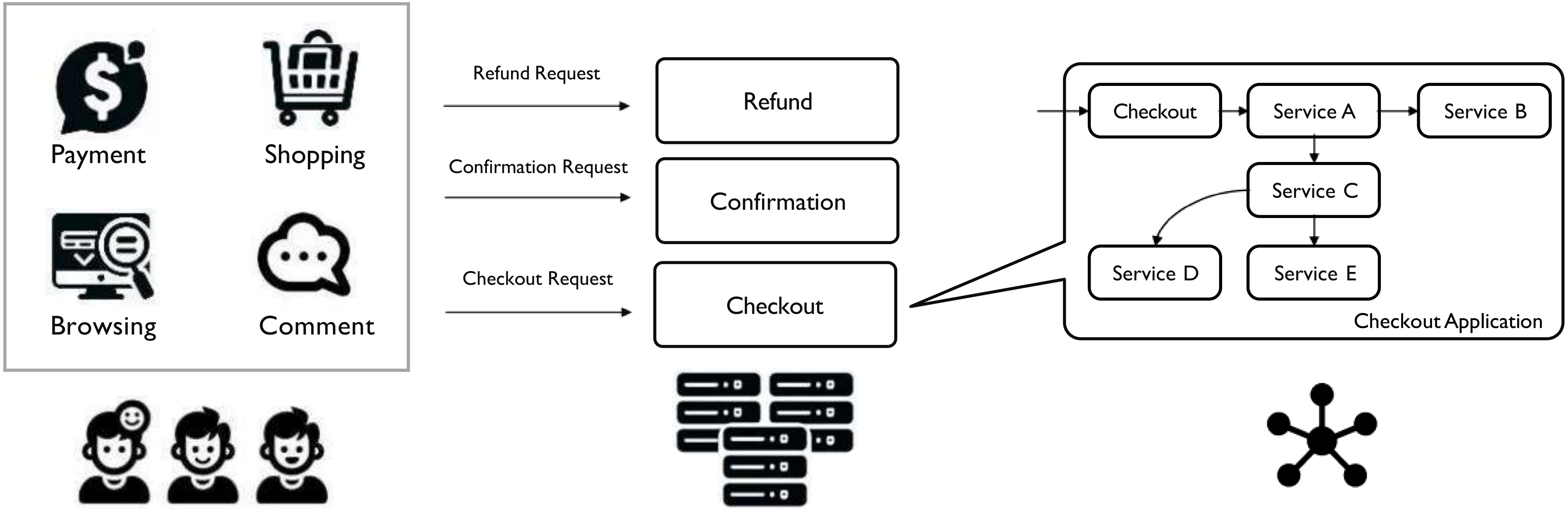
[1] Presenter. Email: leitao@mail.nankai.edu.cn
[2] Corresponding Author

南開大學
Nankai University

Outline

- Background
- Design
- Evaluation
- Conclusion

Microservice Systems



Various User Activities

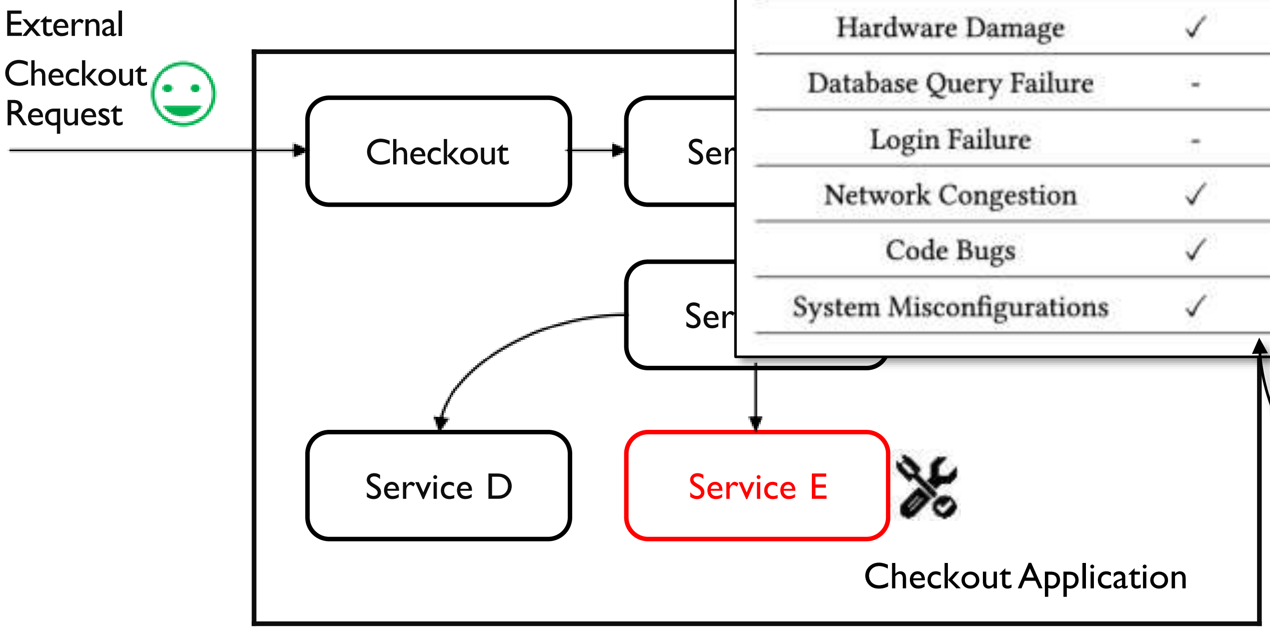
Web Applications

Thousands of Microservices

Microservice Reliability Maintenance

南開大學
Nankai University

Failure Type	Metric	Log	Trace	Remedial Measures
Resource Underprovisioning	✓	-	-	Implement auto-scaling to adjust resources based on demand
Hardware Damage	✓	-	-	Use redundancy and failover systems to ensure continuous operation
Database Query Failure	-	✓	✓	Optimize queries and implement indexing to prevent failures
Login Failure	-	✓	✓	Implement multi-factor authentication to enhance login security
Network Congestion	✓	-	✓	Apply traffic shaping to prioritize critical network traffic efficiently
Code Bugs	✓	-	✓	Use automated testing to detect and fix bugs early
System Misconfigurations	✓	✓	✓	Employ configuration management tools to standardize settings



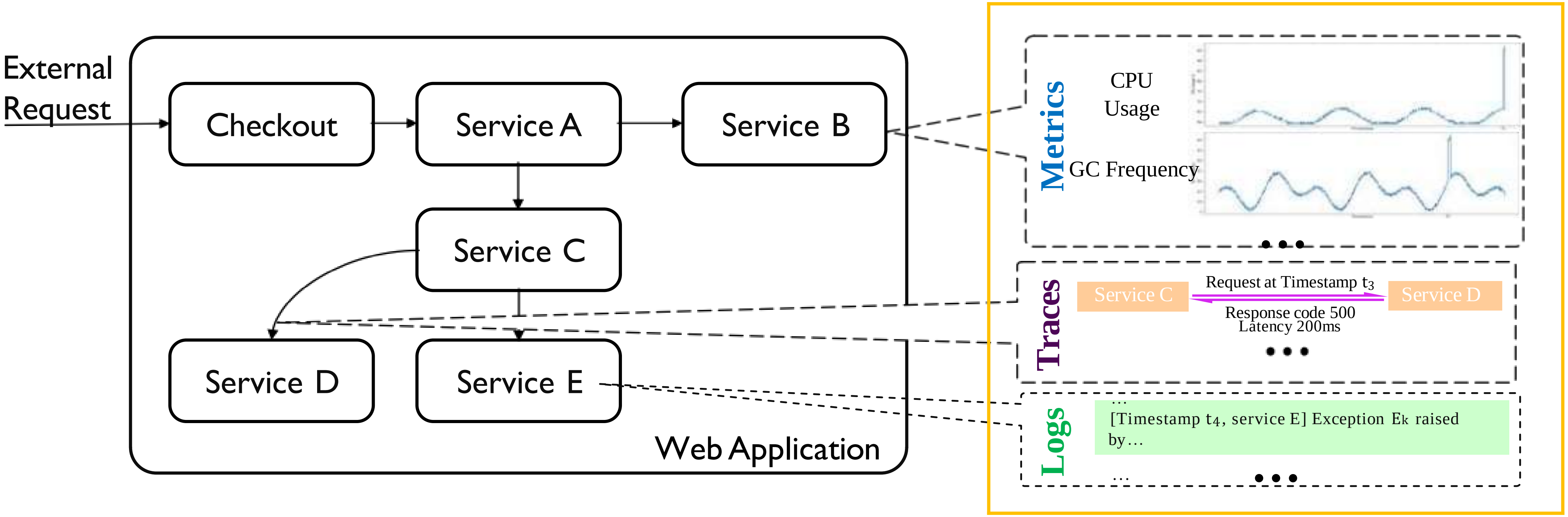
Checkout-related Microservices

Failure Classification

Root Cause Localization

Failure Diagnosis

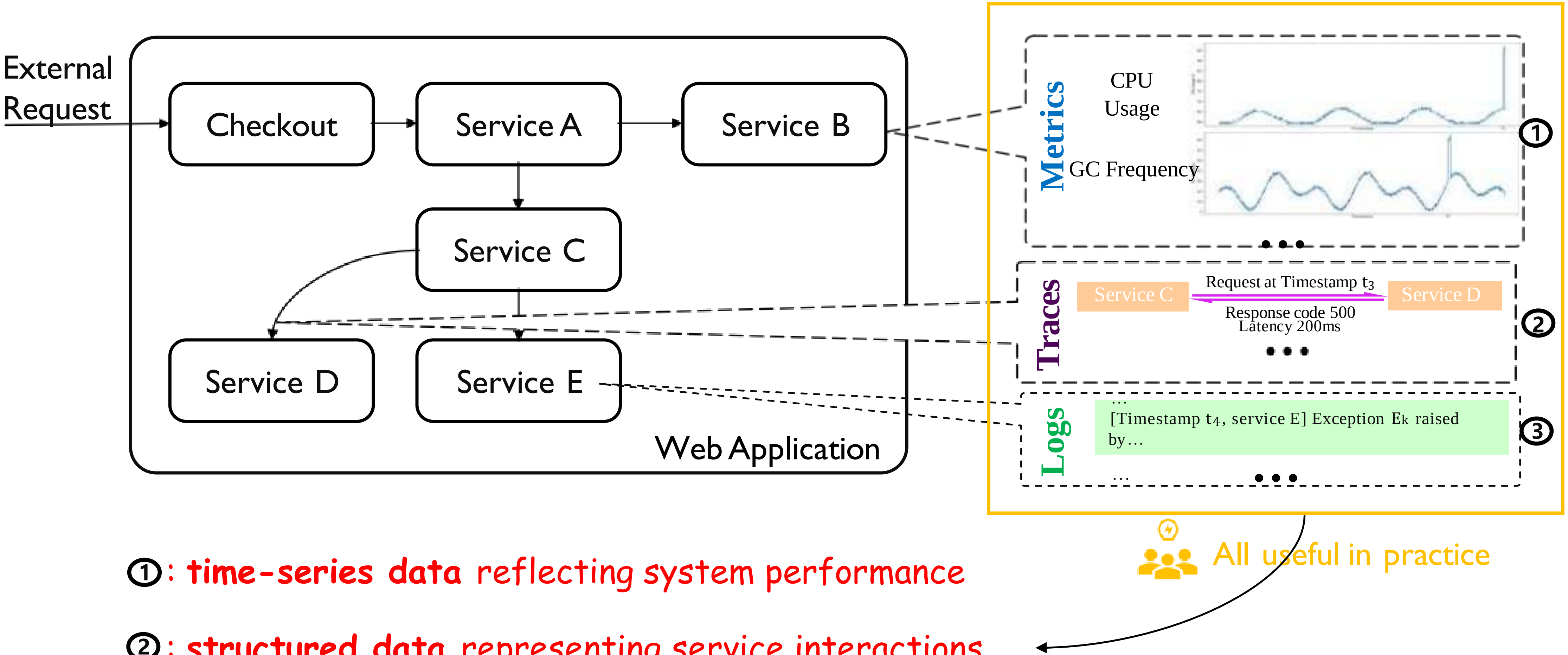
Multimodal Monitoring Data in Microservice Systems



All useful in practice

Multimodal Monitoring Data in Microservice Systems

Nankai University

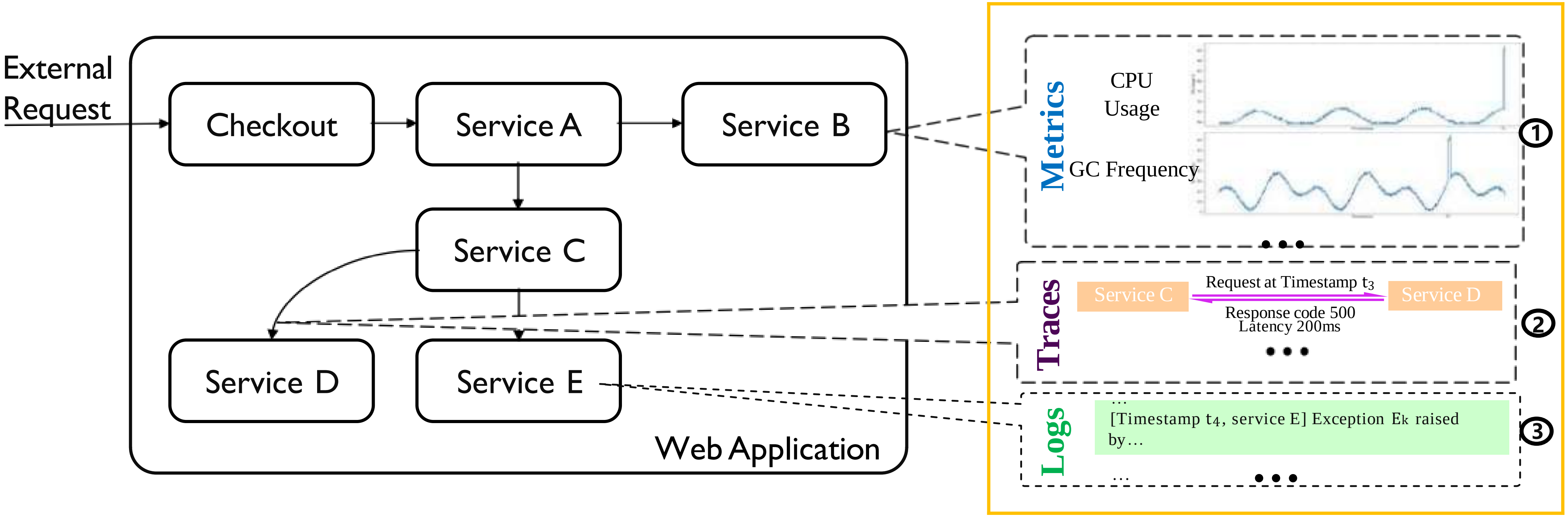


①: **time-series data** reflecting system performance

②: **structured data** representing service interactions

③: **unstructured text** detailing system events

Multimodal Monitoring Data in Microservice Systems



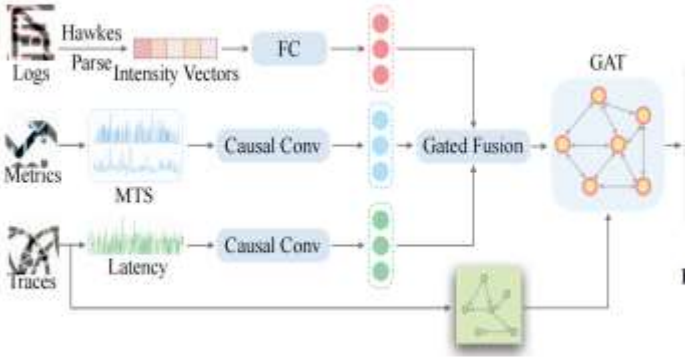
All useful in practice

Challenge 1: How to analyze multimodal data, leveraging information from various observation types?

Utilizing Multimodal Monitoring Data

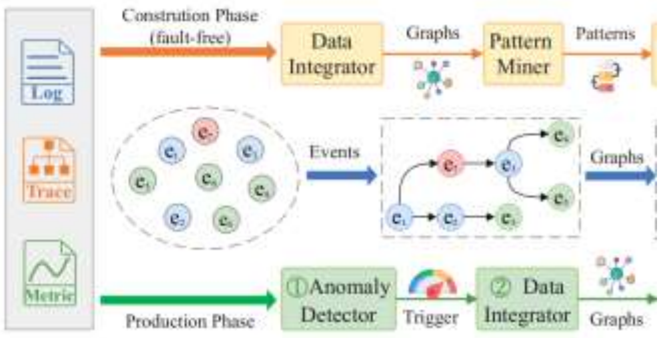
南開大學
Nankai University

GAT-based



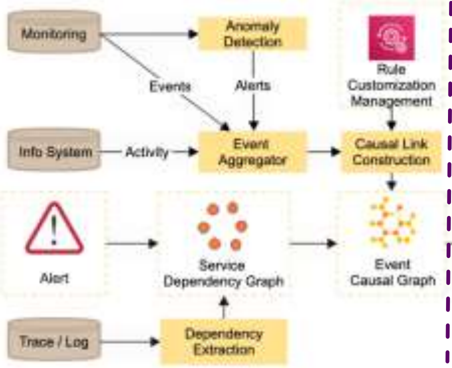
Eadro (ICSE' 23)

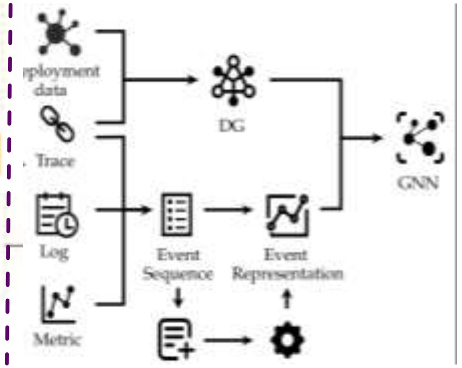
Event Pattern-based



Nezha (ESEC/FSE' 23)

Event Graph-based





Groot (ASE' 21)

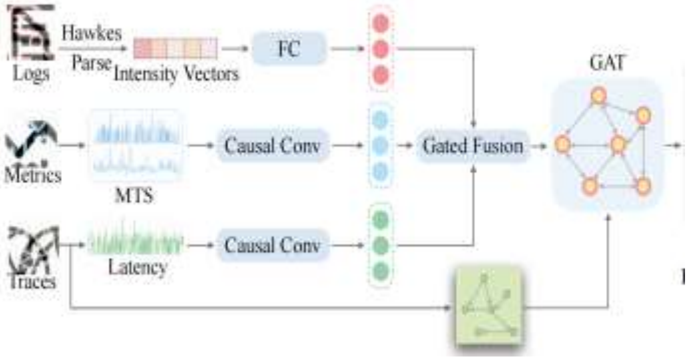
DiagFusion (TSC' 23)

⚠ **Modality coupling: Interaction or interdependence between different modalities**

⚠ **Modality dependence: Relying on multiple modalities simultaneously**

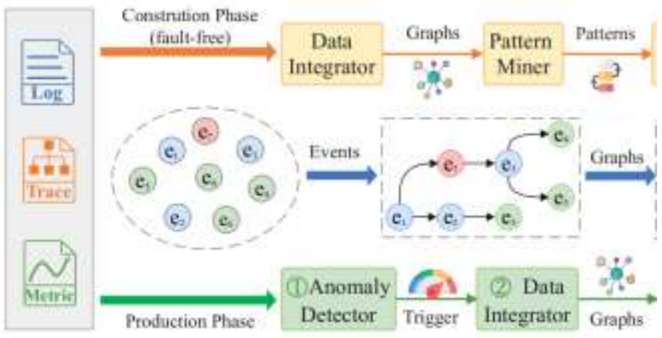
Utilizing Multimodal Monitoring Data

GAT-based



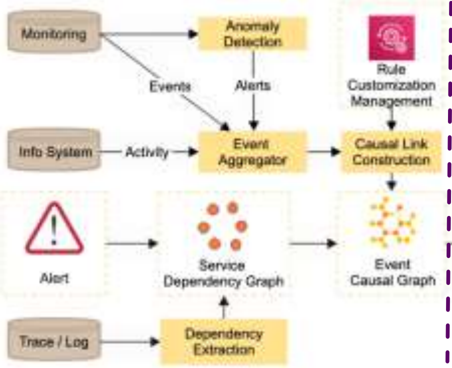
Eadro (ICSE' 23)

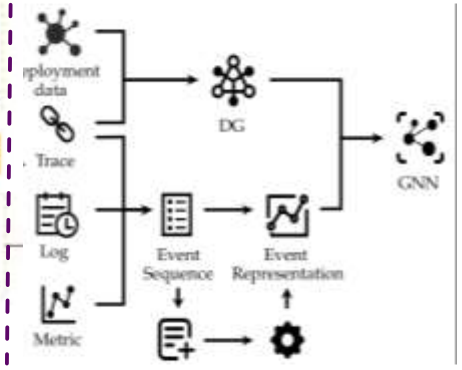
Event Pattern-based



Nezha (ESEC/FSE' 23)

Event Graph-based



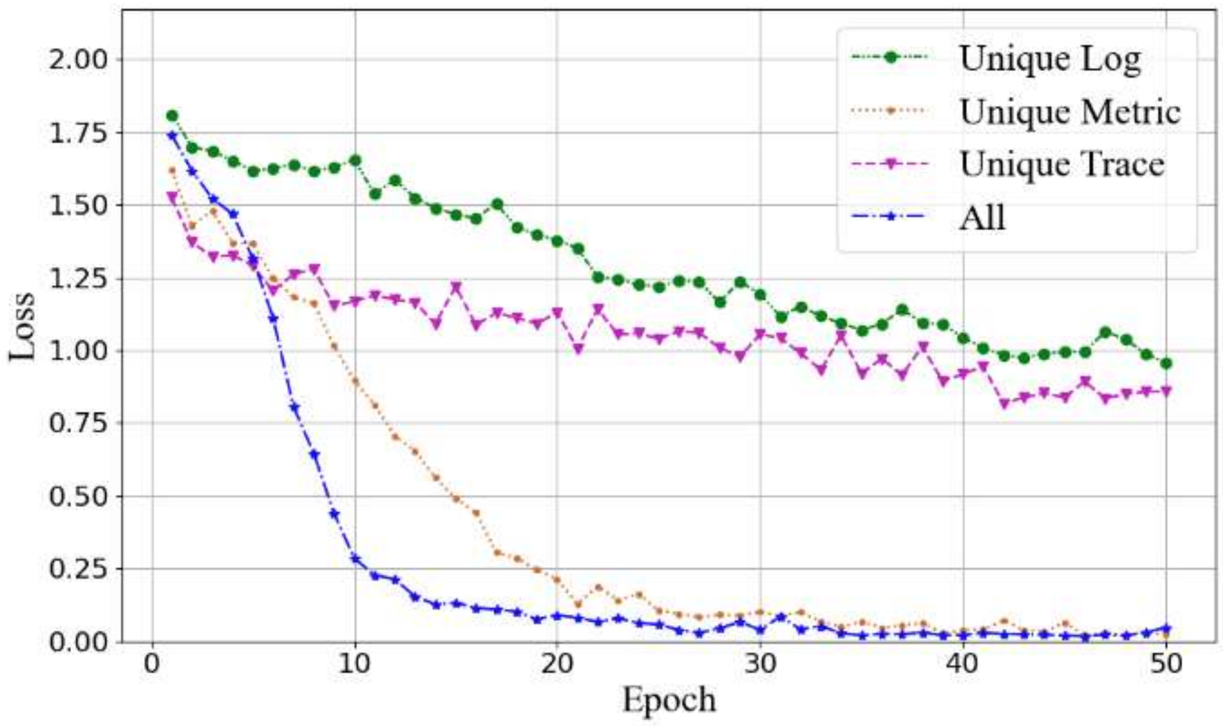


Groot (ASE' 21)

DiagFusion (TSC' 23)

Challenge 2: How to address the substantial performance degradation caused by missing or low-quality data from any modality?

Interference in Modality Optimization



High-Yield Modality

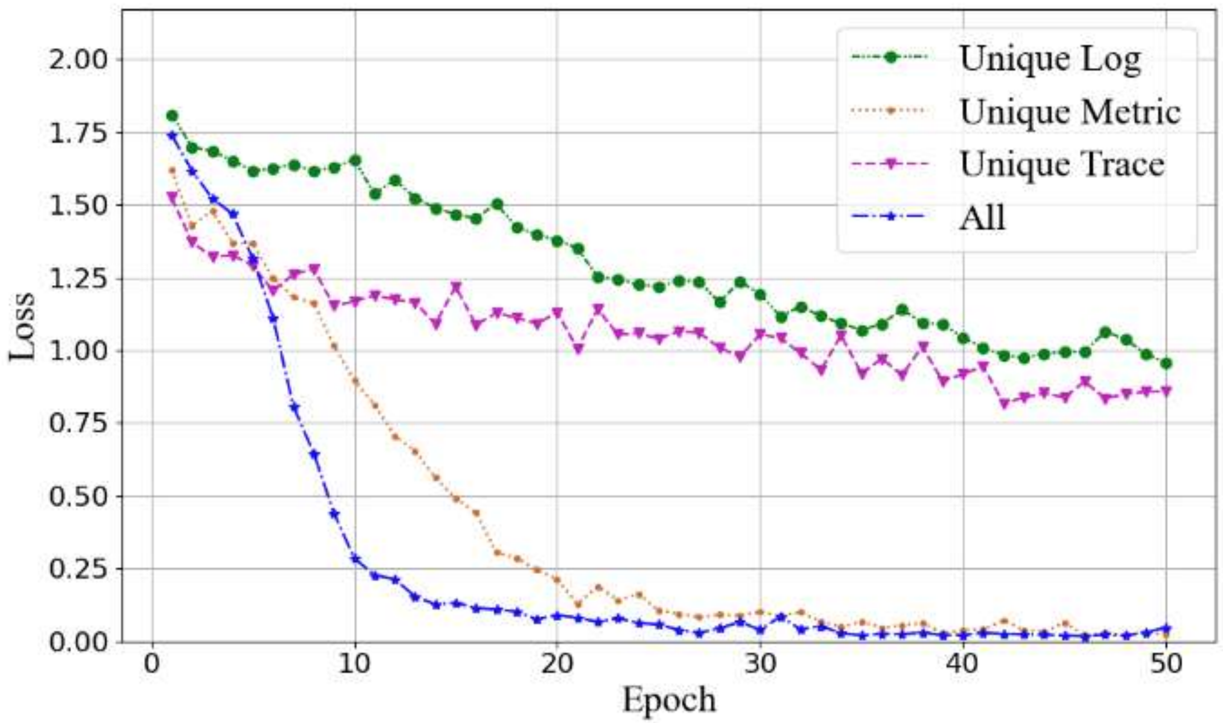
e.g., Metric

Low-Yield Modality

e.g., Log, Trace

The dominant modality may suppress the optimization of other modalities, preventing them from fully utilizing their features

Interference in Modality Optimization



High-Yield Modality

e.g., Metric

Low-Yield Modality

e.g., Log, Trace

Challenge 3: How to reduce the negative impact caused by inconsistent convergence rates and mutual interference between different modalities?

Challenges

Nankai University

Inconsistent Data Formats

- Microservice systems generate diverse operational data, such as metrics, logs, and traces, each with distinct formats and methods of encapsulating information

Incomplete and Low-Quality Data

- In real-world microservice environments, the completeness and quality of multimodal data are often lacking
- Missing or low-quality data from any modality can lead to substantial performance degradation in multimodal failure diagnosis approaches

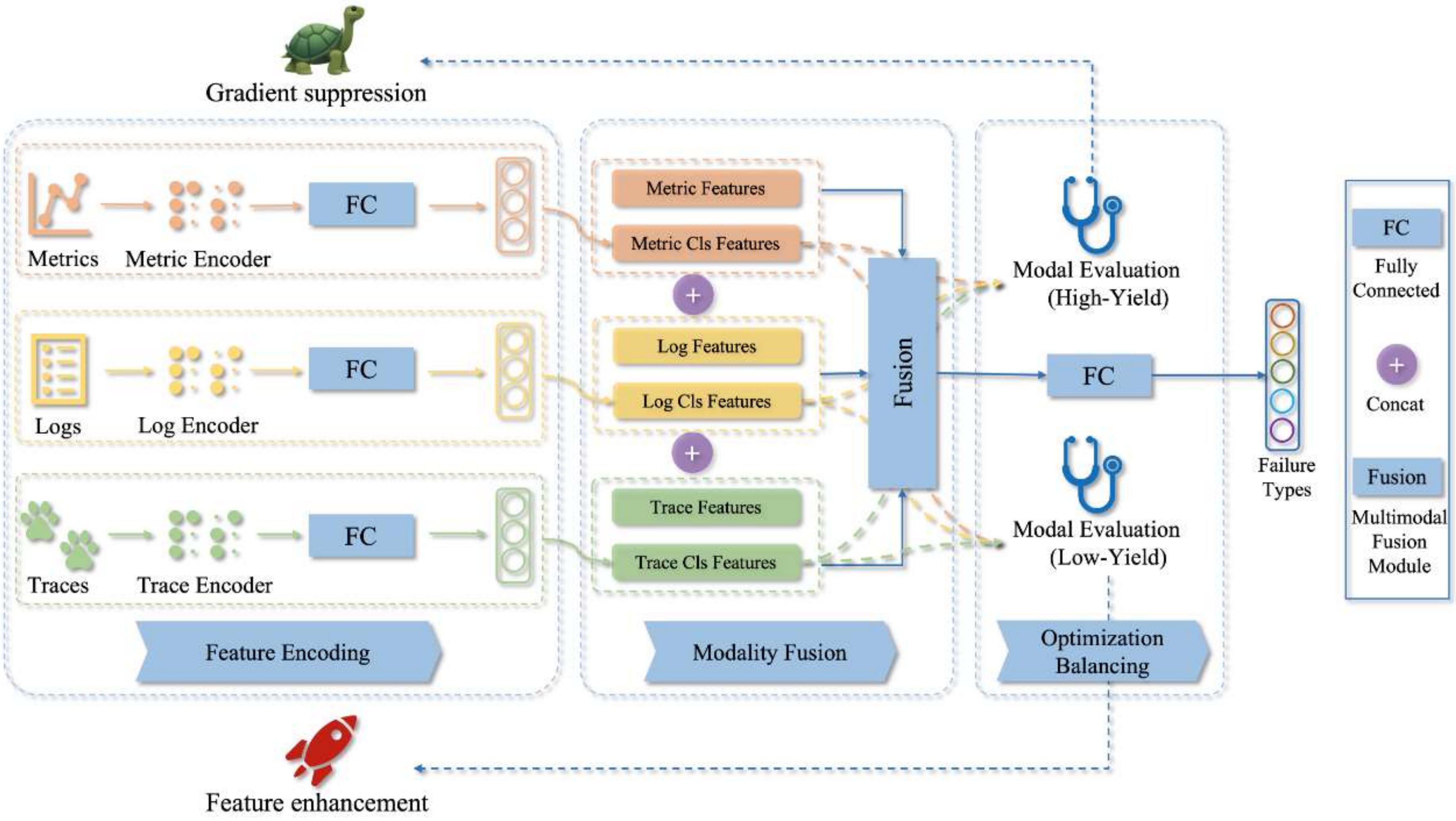
Interference in Modality Optimization

- The dominant modality may suppress the optimization of other modalities, preventing them from fully utilizing their features

Outline

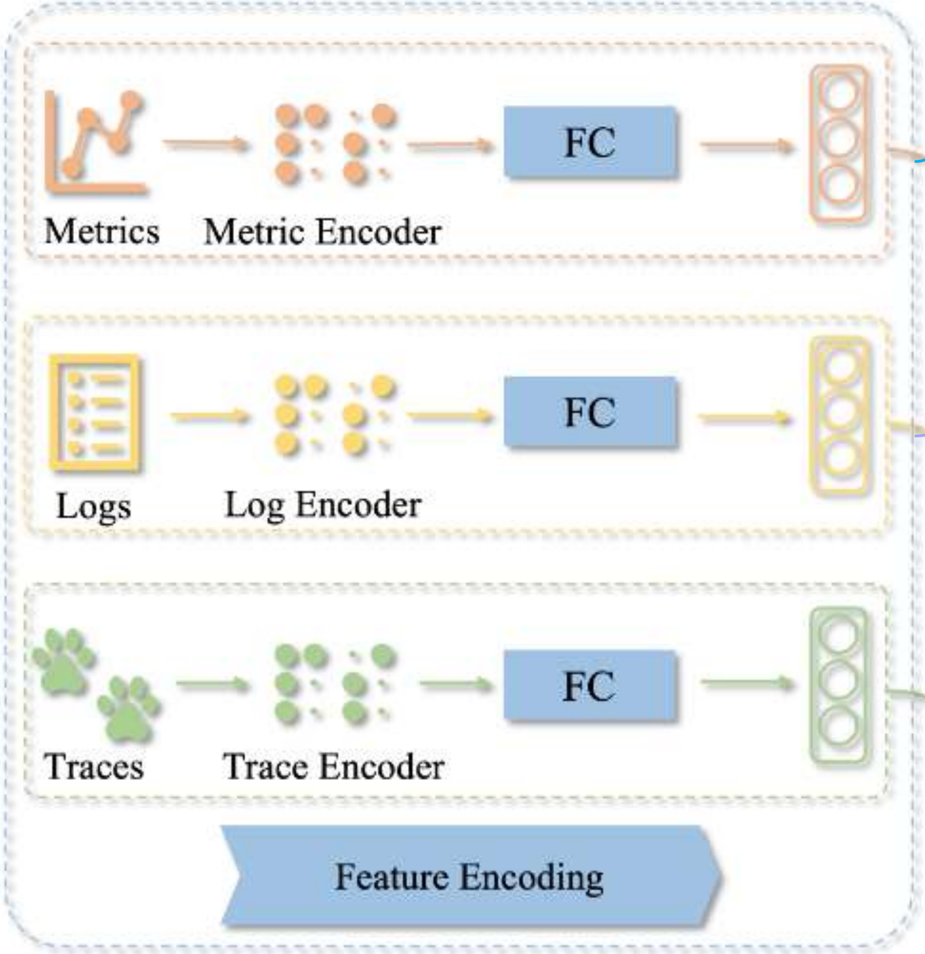
- Background
- Design
- Evaluation
- Conclusion

Medicine Design Overview



Stage I: Feature Encoding

Nankai University



Log Filtering and Template Fusion

Log Modality Fusion

Using statistical methods and Bert to obtain log representations

Using Transformer and global pooling to obtain a high level representation of log modality

Feature Selection and Processing

Metric Modality Fusion

Using the set of metric categories as the feature set characterizing the failure interval

A high level representation of metric modality

Feature Selection and Processing

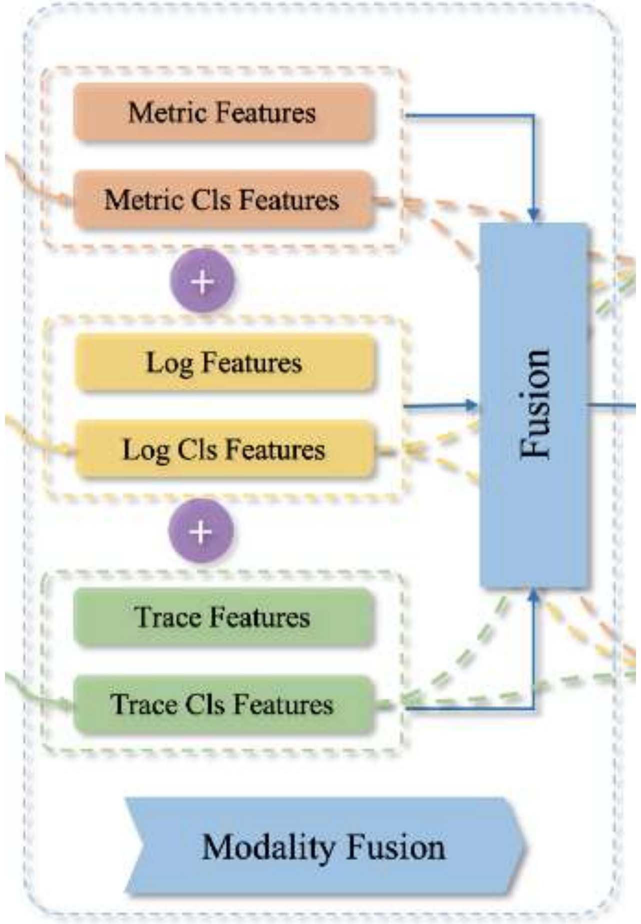
Trace Modality Fusion

Extracting features from the duration data of different types of spans within the time window

A high level representation of trace modality

Stage II: Modality Fusion

Nankai University



(Feature Concatenation) Concatenate the features of different modalities and use a fully connected layer to generate an integrated feature representation

(Modality-specific Linear Transformations) Each modality's features are separately processed through individual linear layers

(Channel Attention Mechanism) The outputs of linear transformations are passed through a sigmoid activation function to produce attention weights

(Feature Stacking and Squeezing) The original and attention-weighted modality features are stacked and processed through adaptive average pooling to reduce dimensionality and focus on key features

(Classification) The pooled feature representation is passed through a fully connected layer to perform failure classification

Stage III: Optimization Balancing

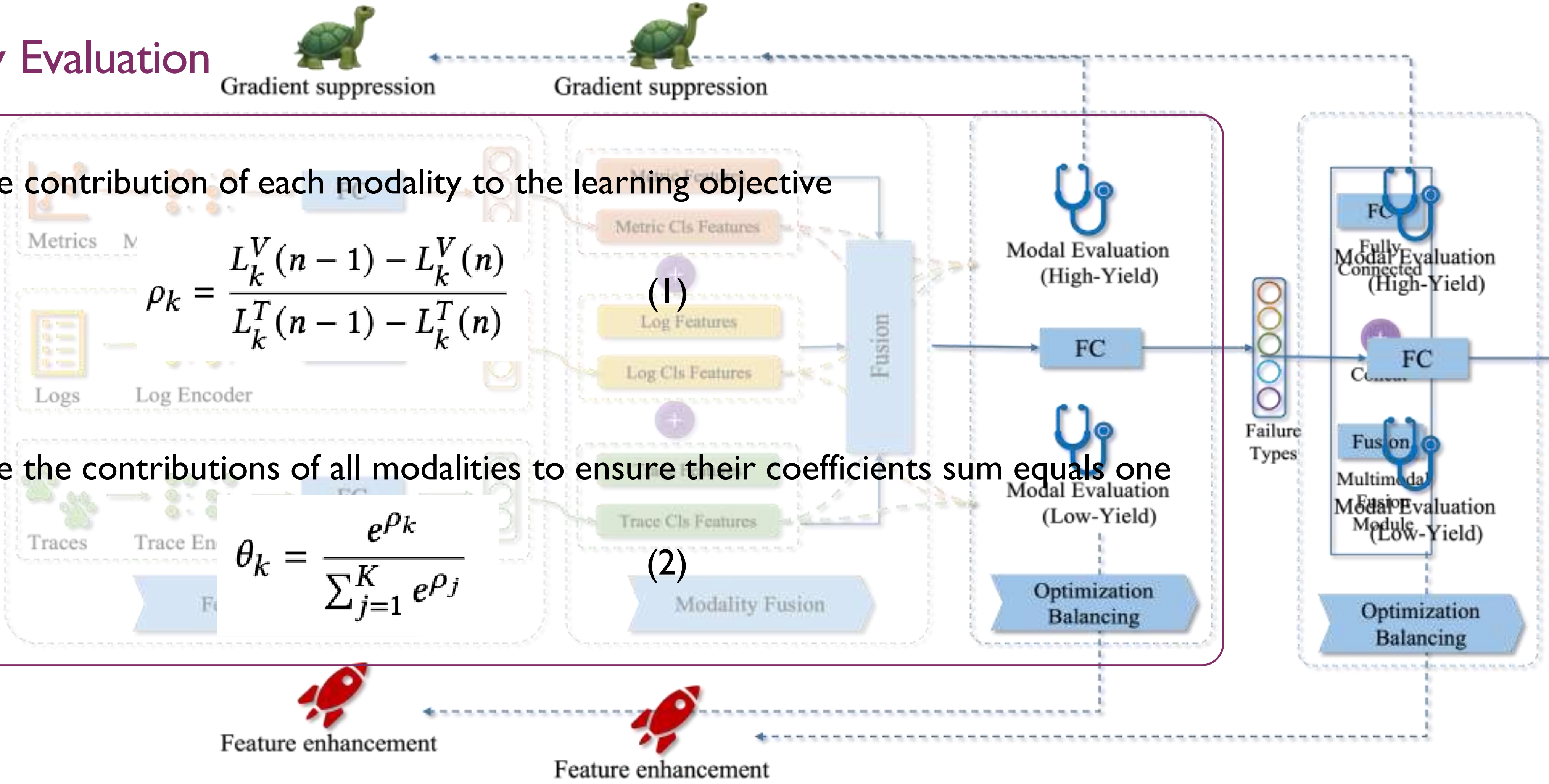
Modality Evaluation

Assess the contribution of each modality to the learning objective

$$\rho_k = \frac{L_k^V(n-1) - L_k^V(n)}{L_k^T(n-1) - L_k^T(n)} \quad (1)$$

Normalize the contributions of all modalities to ensure their coefficients sum equals one

$$\theta_k = \frac{e^{\rho_k}}{\sum_{j=1}^{K} e^{\rho_j}} \quad (2)$$



Stage III: Optimization Balancing

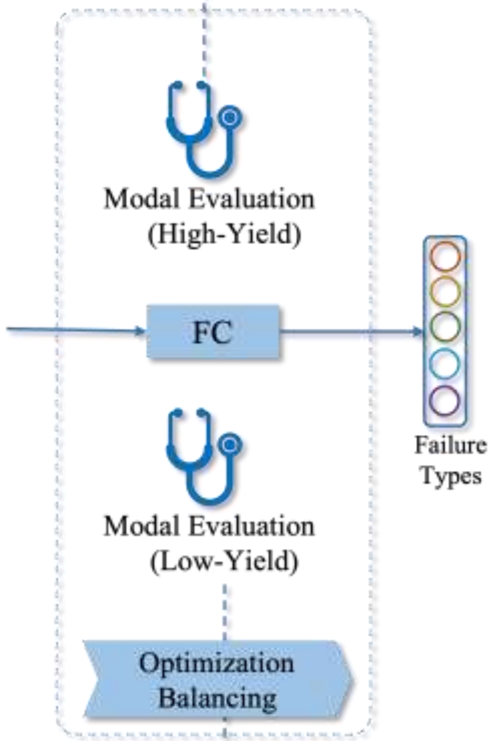
Gradient Suppression

For the dominant modality (the one with the highest θ_k), the gradient is suppressed to prevent it from overwhelming other modalities

$$s_t^k = \begin{cases} 1 - \alpha \cdot \theta_k & \text{if } k = \arg\max(\theta_t^k) \\ 1 & \text{otherwise} \end{cases} \qquad (3)$$

The network parameters are updated as follows:

$$\omega_{t+1}^k = \omega_t^k - \eta \cdot s_t^k \tilde{g}(\omega_t^k) \qquad (4)$$



Stage III: Optimization Balancing

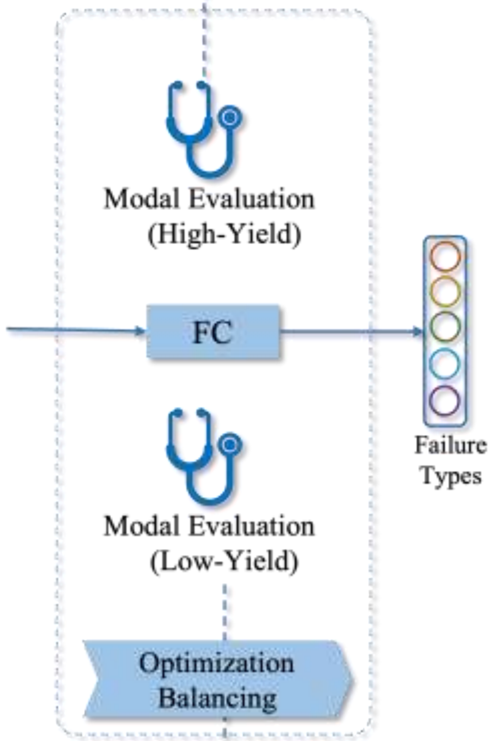
Feature Enhancement

To compensate for the lower contribution of underperforming (low-yield) modalities, the feature enhancement component boosts the features of these modalities.

$$s_t^k = \begin{cases} \beta \cdot \theta_k & \text{if } k = \arg\min(\theta_t^k) \\ 1 & \text{otherwise} \end{cases} \qquad (5)$$

The enhanced feature representation is given by:

$$\tilde{\mathbf{x}}_t^k = \mathbf{F}_{scale}(\mathbf{u}_t^k, s_t^k) = s_t^k \cdot \mathbf{u}_t^k \qquad (6)$$



南開大學
Nankai University

Outline

- Background
- Design
- Evaluation
- Conclusion

Evaluation: Performance of *Medicine*

南開大學 Nankai University

Approach	Modality			D1			D2			D3		
	Metric	Log	Trace	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
DéjàVu [14]	✓			0.4569	0.5526	0.4972	0.4620	0.4820	0.4682	0.5990	0.1852	0.1962
iSQUAD [9]	✓			0.4291	0.5429	0.4750	0.6798	0.6591	0.6457	0.1600	0.2500	0.1857
Cloud19 [15]		✓		0.5082	0.5429	0.5231	0.5703	0.5682	0.5690	0.3602	0.4167	0.3830
LogCluster [10]		✓		0.4867	0.3714	0.3852	0.4522	0.4862	0.4671	0.4128	0.5000	0.4260
MEPFL [16]			✓	0.3286	0.4571	0.3823	0.2321	0.4818	0.3133	0.2946	0.4035	0.3562
CloudRCA [20]	✓	✓		0.2463	0.1370	0.1143	0.0913	0.2174	0.1180	0.3708	0.2630	0.2652
DiagFusion [17]	✓	✓	✓	0.7326	0.6744	0.7015	0.8176	0.7891	0.7895	0.3870	0.2813	0.3165
MicroCBR [22]	✓	✓	✓	0.6286	0.8000	0.6500	0.4630	0.4310	0.4464	0.4626	0.5714	0.4835
Medicine	✓	✓	✓	**0.9714**	**0.9428**	**0.9508**	**0.9152**	**0.9136**	**0.9136**	**0.8358**	**0.8333**	**0.8260**

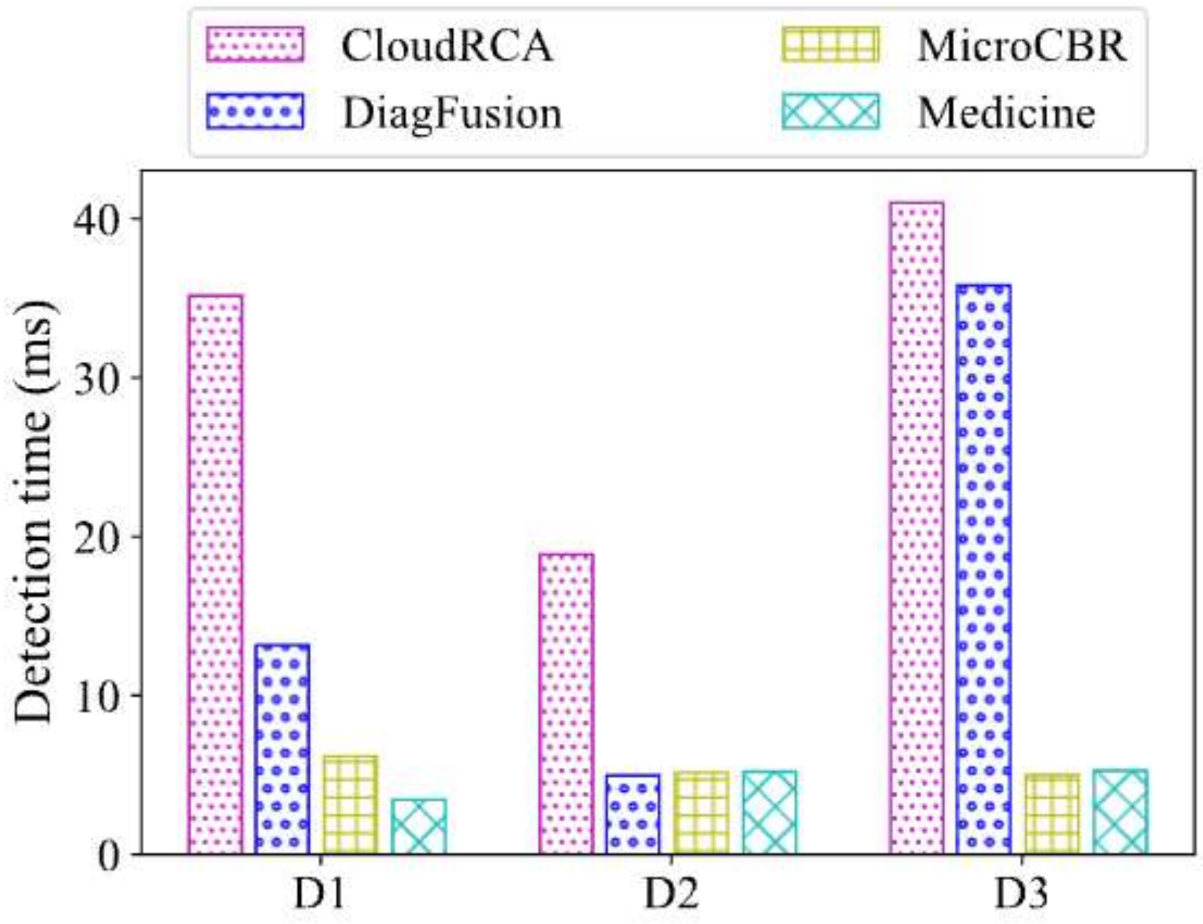
D1: collect from a top-tier global commercial bank

D2: Generic AIOps Atlas (GAIA) dataset from CloudWise

D3:collect from a microservice benchmark, MicroServo

Contrasted with the benchmark multimodal approach, DiagFusion, ***Medicine*** showcased substantial improvements, enhancing F1-score by 35.54% and 15.72% on D1 and D2, respectively.

Evaluation: Efficiency



Medicine demonstrates the shortest detection time on D1, taking only 3.44ms, whereas CloudRCA is the slowest at 35.16ms. On D2 and D3, Medicine's average detection time is comparable to that of MicroCBR, all around 5ms.

Compared with baseline methods, ***Medicine*** is indeed efficient in diagnosising failures.

Evaluation: Ablation Study

Nankai University

Dataset	Approach	Precision	Recall	F1-score
D1	Only Metric	0.8926	0.8857	0.8847
	Only Log	0.3316	0.4571	0.3820
	Only Trace	0.3595	0.4571	0.4020
	w/o MAO	0.9350	0.9143	0.9086
	Medicine	**0.9714**	**0.9428**	**0.9508**
D2	Only Metric	0.7705	0.7909	0.7753
	Only Log	0.8836	0.8500	0.8445
	Only Trace	0.5232	0.5227	0.5139
	w/o MAO	0.8959	0.8955	0.8953
	Medicine	**0.9152**	**0.9136**	**0.9136**
D3	Only Metric	0.6875	0.5000	0.4538
	Only Log	0.4792	0.4583	0.4431
	Only Trace	0.2550	0.2083	0.2179
	w/o MAO	0.7121	0.7083	0.6956
	Medicine	**0.8358**	**0.8333**	**0.8260**

Medicine significantly outperforms unimodal methods.

Our designed **unimodal encoder** can extract useful features based on the characteristics of microservice systems for failure classification.

MAO dynamically adjusted and optimized the weights and interactions between different data modalities, thereby enhancing the overall performance of the model.

南開大學
Nankai University

Outline

- Background
- Design
- Evaluation
- Conclusion

Conclusion

Nankai University

Medicine, a microservice failure diagnosis framework

- Modal-independent failure diagnosis framework based on multimodal adaptive optimization
- Reduce dependence on any single modality
- Balance the optimization process during training by suppressing gradients for high-yield modalities and enhancing features for low-yield ones based on modal evaluation

Key Designs of *Medicine*

- Individually designed unimodal encoder
- Multimodal fusion with channel attention
- Multimodal Adaptive Optimization (Modality Evaluation, Gradient Suppression, Feature Enhancement)
- Proved effectiveness of the key components in ablation study

Open source code

- https://github.com/AIOps-Lab-NKU/Medicine

Thank You !

Giving Every Modality a Voice in Microservice Failure Diagnosis via Multimodal Adaptive Optimization

Paper: https://doi.org/10.1145/3691620.3695489
Code: https://github.com/AIOps-Lab-NKU/Medicine