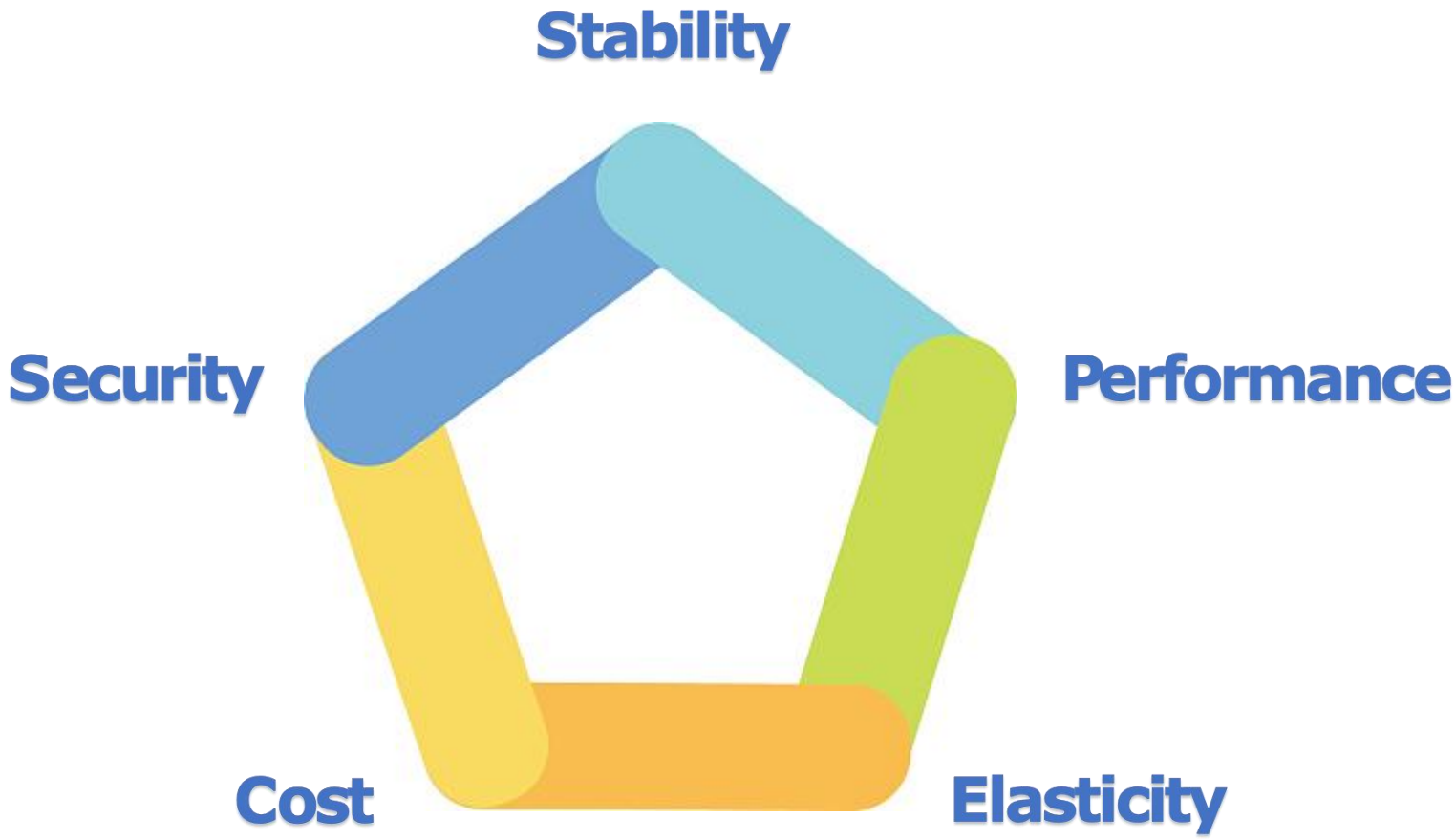
Stability is not Downtime: Comprehensive Stability Evaluation for Large-Scale Cloud Servers in Alibaba Cloud

Haoyu Wang, Zhicheng Liu, Yeliang Qiu, Haozhe Li, Hongke Guo,
Zhaoliang Zhu, You Zhang, Yu Zhou, Xudong Zheng

Alibaba Cloud Computing

ICDE 2025

Why Cloud Service?



Importance of Stability

- Serious faults in cloud computing may bring a very significant impact on society.

Community

Alibaba Cloud tech fault shuts down major websites in Macao

- Problem traced to refrigeration equipment in PCCW-HKT Data Centre; Macao Daily News, Galaxy Entertainment and food delivery platforms among those affected.

ARTICLE BY
Staff Reporter

PUBLISHED
19 Dec 2022

Importance of Stability

- A short period of unavailability may have a catastrophic impact on the businesses of customers.
  - A customer experienced an unexpected memory hardware failure.

Instance type: ecs.g7.2xlarege

Resource: 8 cores + 32 GB Mem

Price: ~1000 CNY / month

VS

Service unavailability: ~10 min

Business loss: 70,000+ CNY

Complaint: 20+

Stability Challenge

- Elastic Compute Service (ECS) in Alibaba Cloud has an enormous scale, facing significant pressures in stability.
  - We built an AIOps system, CloudBot, to maintain stability.

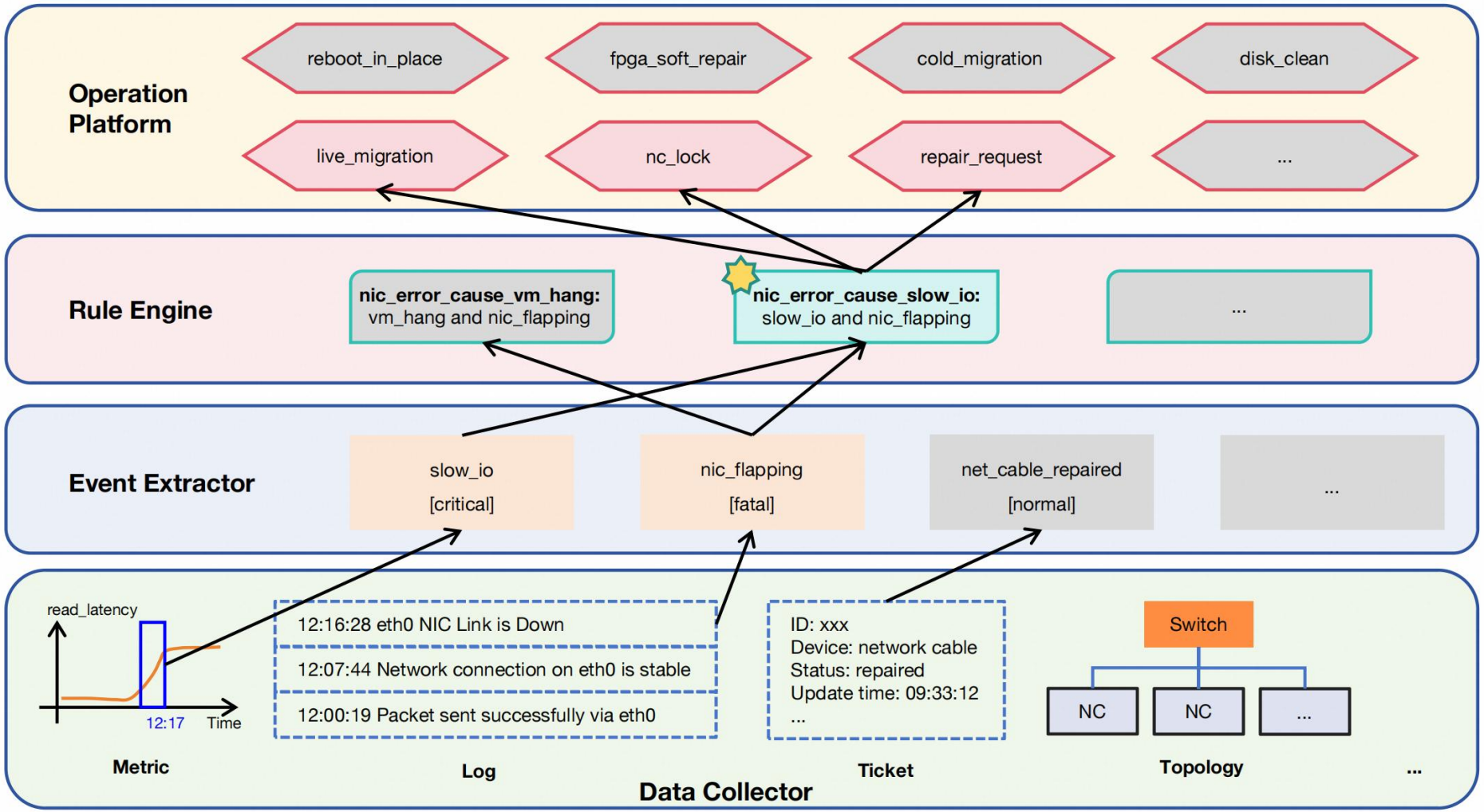
10,000,000+ VMs

1,000,000+ Servers

5,000+ Clusters

29 Regions

CloudBot


Operation Platform
reboot_in_place
fpga_soft_repair
cold_migration
disk_clean
live_migration
nc_lock
repair_request
...
Rule Engine
nic_error_cause_vm_hang:
vm_hang and nic_flapping
nic_error_cause_slow_io:
slow_io and nic_flapping
...
Event Extractor
slow_io
[critical]
nic_flapping
[fatal]
net_cable_repaired
[normal]
...
read_latency
12:17
Time
Metric
12:16:28 eth0 NIC Link is Down
12:07:44 Network connection on eth0 is stable
12:00:19 Packet sent successfully via eth0
Log
ID: xxx
Device: network cable
Status: repaired
Update time: 09:33:12
...
Ticket
Switch
NC
NC
...
Topology
...
Data Collector

Quantitative Stability Evalution

- A quantitative stability evalution is required to drive the evolution of CloudBot.
  - **Interpretability**: an interpretable evaluation is more convincing than a black-box one especailly in cross-team collaborations.
  - **Cost-Effectiveness**: heavyweight evaluation would result in significant extra resource usage due to the vast scale.
  - **Non-Invasiveness**: it is not allowed to operate inside the VM.
  - **Comprehensiveness**: The cloud server is a complex product including computing, storage and networking.

Related Work Comparison

- Benchmark-based evaluation: SuperBench...
- Downtime-based evaluation: Downtime Percentage, Annual Interruption Rate...

Dimension	Benchmark-based	Downtime-based
Interpretability	Yes	Yes
Cost-Effectiveness	No	Yes
Non-Invasiveness	No	Yes
Comprehensiveness	Yes	No

Stability ≠ Downtime

- Complaints about VM performance
  - The performance is impacted by ECS underlying changes during its release.

- Incident on Nov. 12, 2023
  - VMs are not controllable due to faulty logic within the AccessKey system.

[Exception (recovered)] Abnormalities in Alibaba Cloud product console access and API calls

•

Scope of impact

- Some products such as Object Storage Service (OSS), Tablestore, Log Service, and Message Service (MNS) went through partial service disruption, while service provision of most products remained intact, such as Elastic Compute Service (ECS), ApsaraDB RDS, and all networking products.
- The consoles and management APIs of cloud products were affected.

Time

November 12, 2023 17:39 – 19:20 (UTC+8)

Definition & Category

- Definition
  - The stability of cloud servers is defined as its capacity to deliver and **manage** computational resources in a **continuous** and **consistent** manner.
- Category
  - Unavailablity
  - Performance
  - Control-Plane

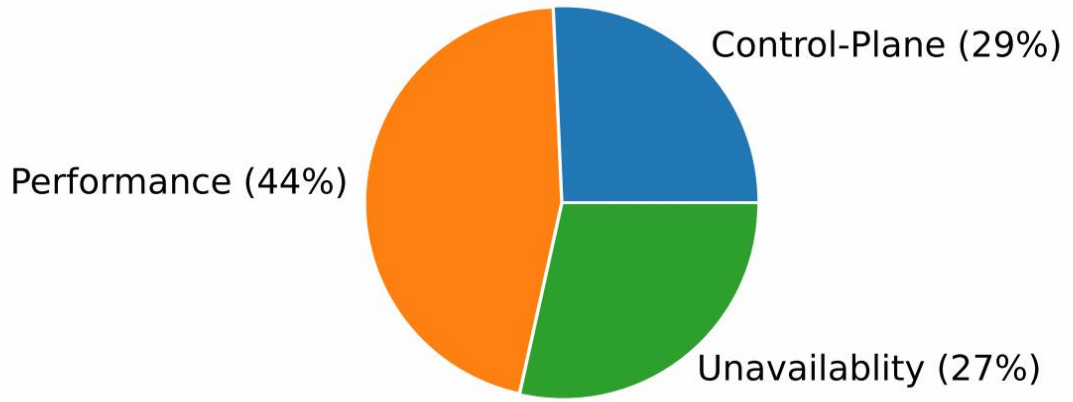
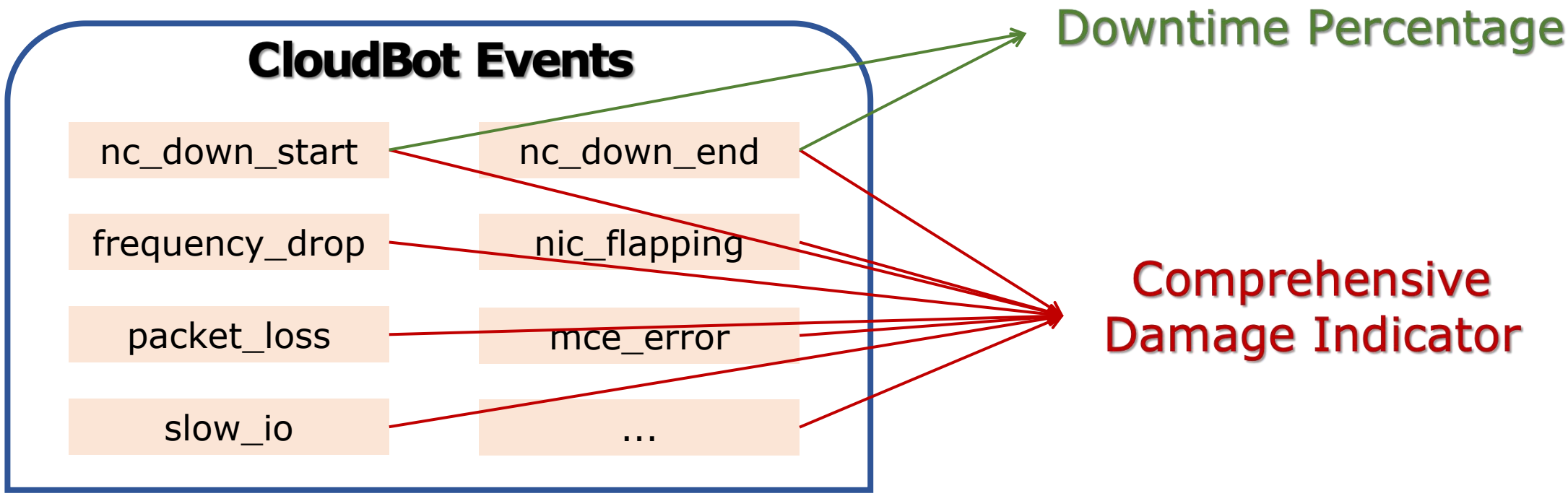


Fig. 2: Distribution of tickets related to ECS stability

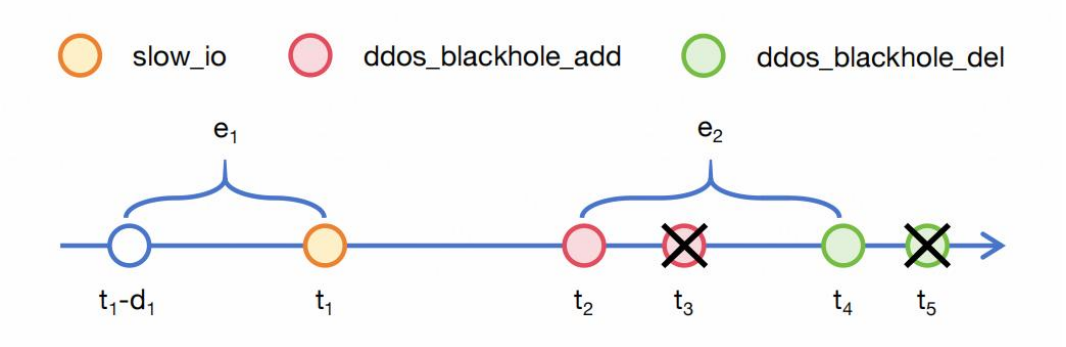
From Downtime to CDI

- Comprehensive Damage Indicator (CDI)
  - Evaluate large-scale cloud server stability based on **interpretable intermediate results**, CloudBot events.



Comprehensive Damage Indicator

- Sub-metrics
  - Unavailability & Performance & Control-Plane Indicator
  - Only difference lies in the specific events they rely on
- Event Period
  - Stateless events -> single
  - Stateful events -> paired



- Event Weight
  - Expert weight: according to experience
  - Customer weight: according to ticket number
  - Analytic Hierarchy Process (AHP) to integrate the weights

Comprehensive Damage Indicator

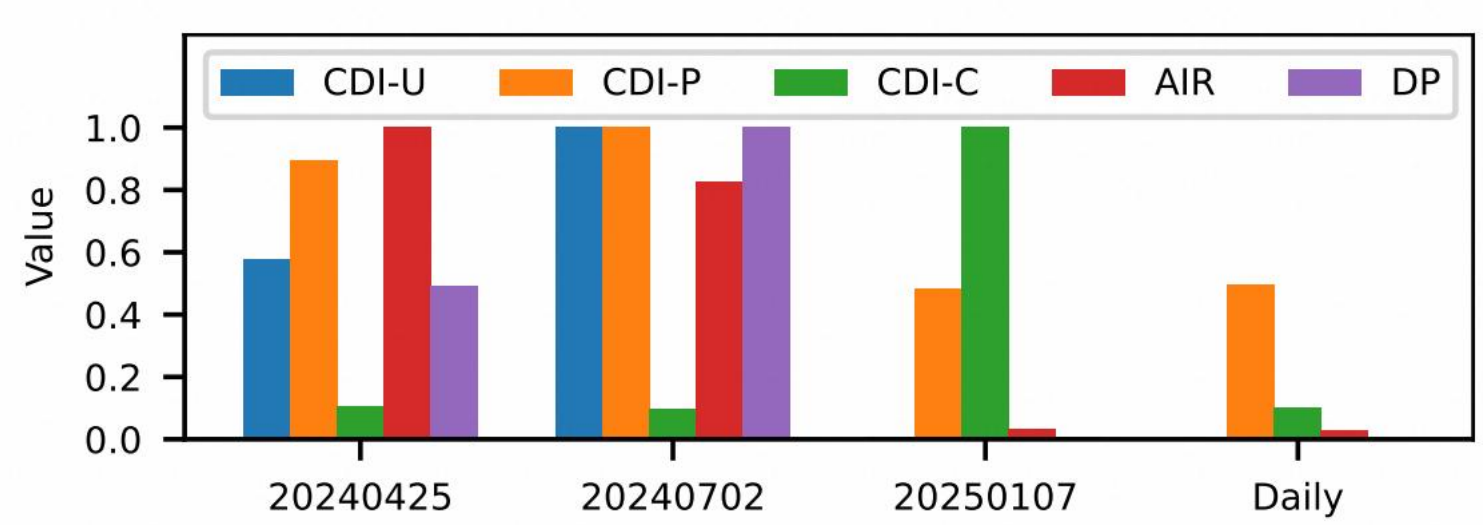
- Example

VM	Service Time	Event	Period	Weight	CDI
1	60min	packet_loss packet_loss	10:08-10:10 10:10-10:12	0.3 0.3	0.020
2	1440min	vcpu_high	13:25-13:30	0.6	0.002
3	1000min	slow_io slow_io vcpu_high	08:08-08:10 08:10-08:12 08:10-08:15	0.5 0.5 0.6	0.004
All	2500min	-	-	-	0.003

- Solution
  - Q_1 = (2*0.3+2*0.3) / 60 = 1.2 / 60 = 0.020
  - Q_2 = (5*0.6) / 1440 = 3 / 1440 = 0.002
  - Q_3 = (2*0.5+5*0.6) / 1000 = 4 / 1000 = 0.004
  - Q_{all} = (0.020*60+0.002*1440+0.004*1000) / (60+1440+1000) = 0.003

Application: Stability Evalution

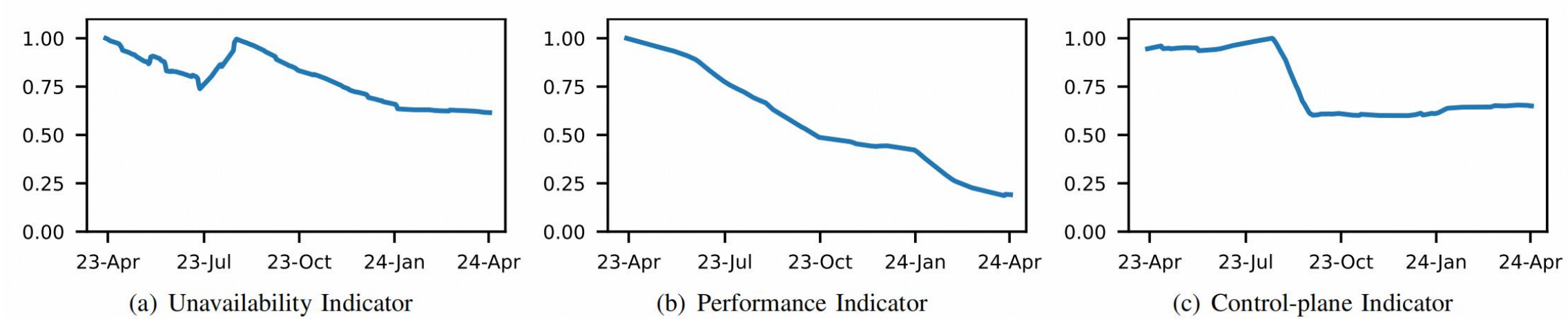
- Stability Evalution on Incidents
  - 20240425: ECS service exception
  - 20240702: network access abnormalities
  - 20250107: abnormalities in purchase and modify



CDI-U: Unavailablity Indicator
CDI-P: Performance Indicator
CDI-C: Control-Plane Indicator
DP: Downtime Percentage
AIR: Annual Interruption Rate

Application: Stability Evalution

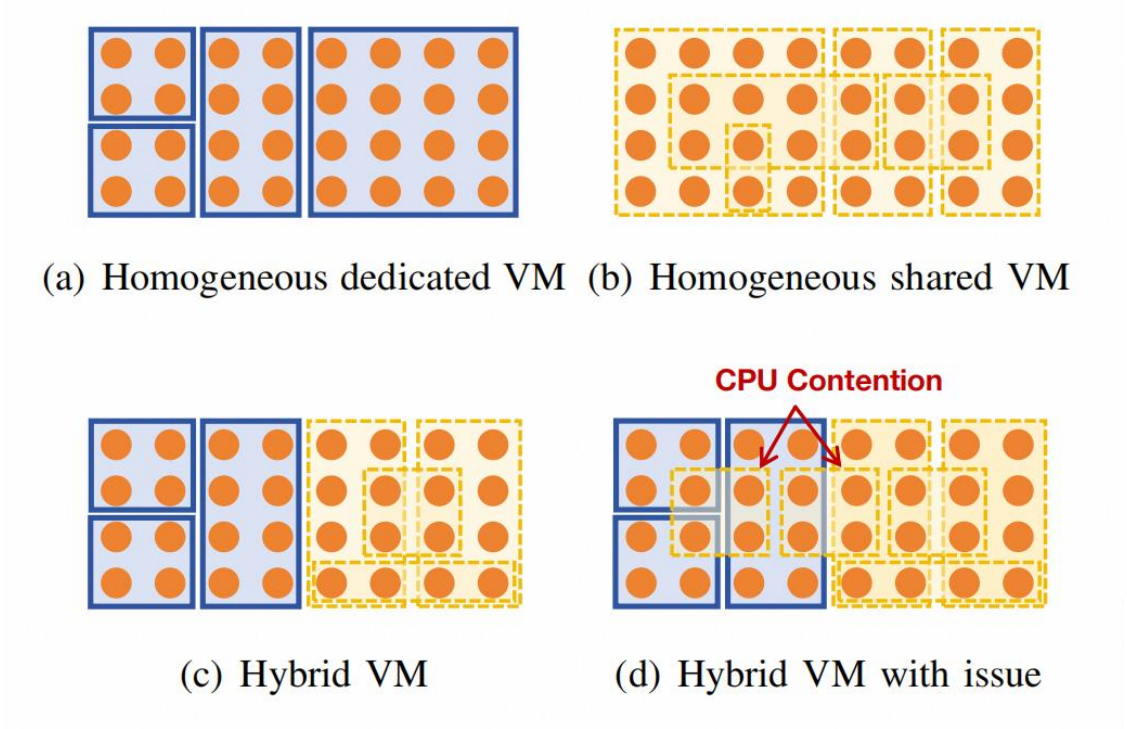
- Daily stability evaluation for FY24 (April 2023 to March 2024)
  - Unavailability Indicator: **40%** decrease
  - Performance Indicator: **80%** decrease
  - Control-plane Indicator: **35%** decrease

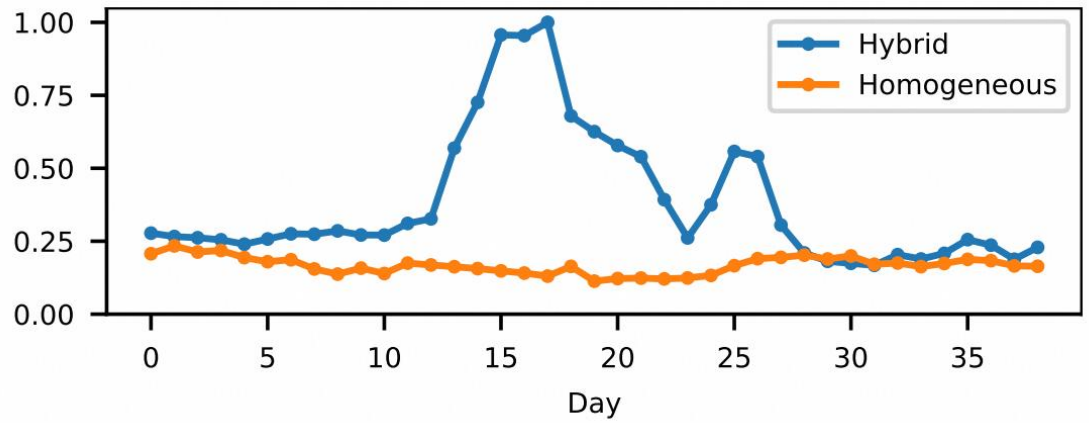


(a) Unavailability Indicator (b) Performance Indicator (c) Control-plane Indicator

Application: Architecture Comparison

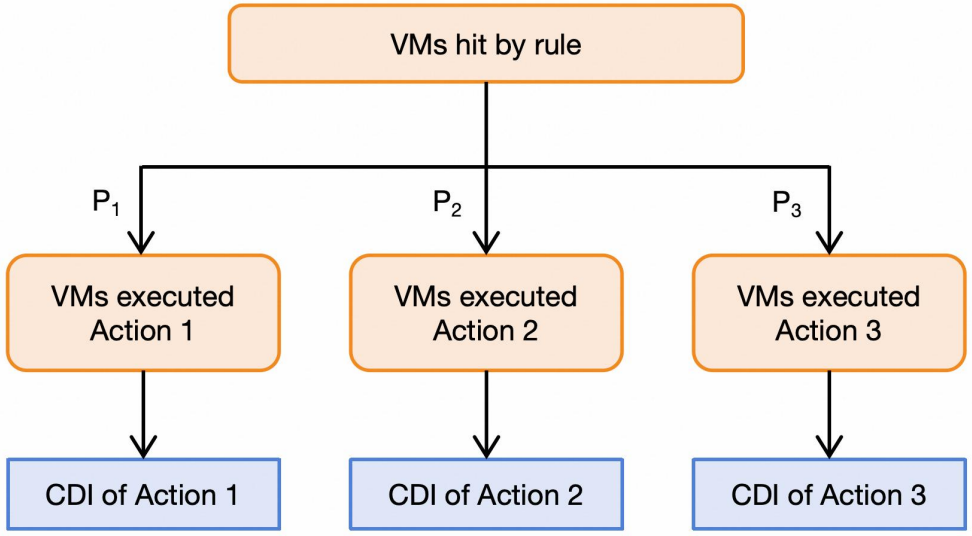
- Homogeneous-deployment to hybrid-deployment
  - Monitor CDI to prevent stability loss from architecture evolution.

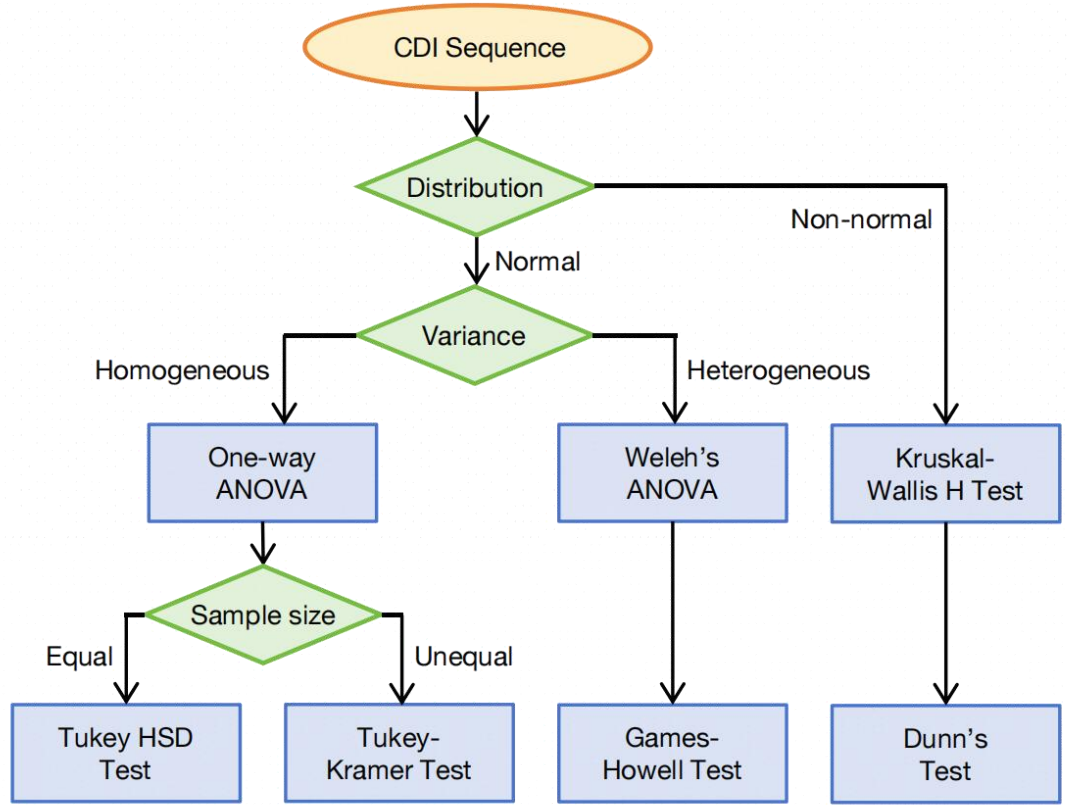




Application: Operation Action Optimization

- Select optimal action from several candidates
  - A/B Test: randomly execute action and calculate its CDI for the next two days
  - Hypothesis testing: omnibus test + post-hoc analysis

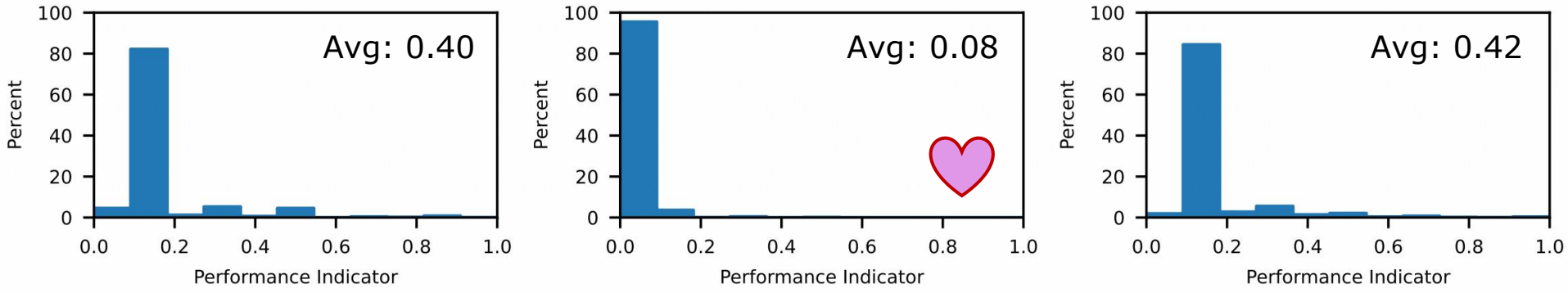




Application: Operation Action Optimization

- Select optimal action from several candidates
  - A/B Test: randomly execute action and calculate its CDI for the next two days
  - Hypothesis testing: omnibus test + post-hoc analysis

Sub-metric	Omnibus Test		Post-hoc Analysis		
	P-value	Sign.	Pair	P-value	Sign.
Unavailability	0.47	False			
Control-plane	0.89	False			
Performance	0	True	A-B	0	True
			A-C	0.03	True
			B-C	0	True



(a) Action A (b) Action B (c) Action C

Conclusion

- In this paper, we propose Comprehensive Damage Indicator (CDI).
  - The **first** comprehensive quantitative stability evaluation metric for large-scale cloud servers, to the best of our knowledge.
  - Extending stability evaluation beyond system unavailability to include **performance** and **control-plane** issues.
  - Deployed in various domains for over two years within Alibaba Cloud ECS, providing support for the evolution of stability.

THANKS FOR YOUR LISTENING

Slides are shown on the personal homepage https://wanghy.pages.dev