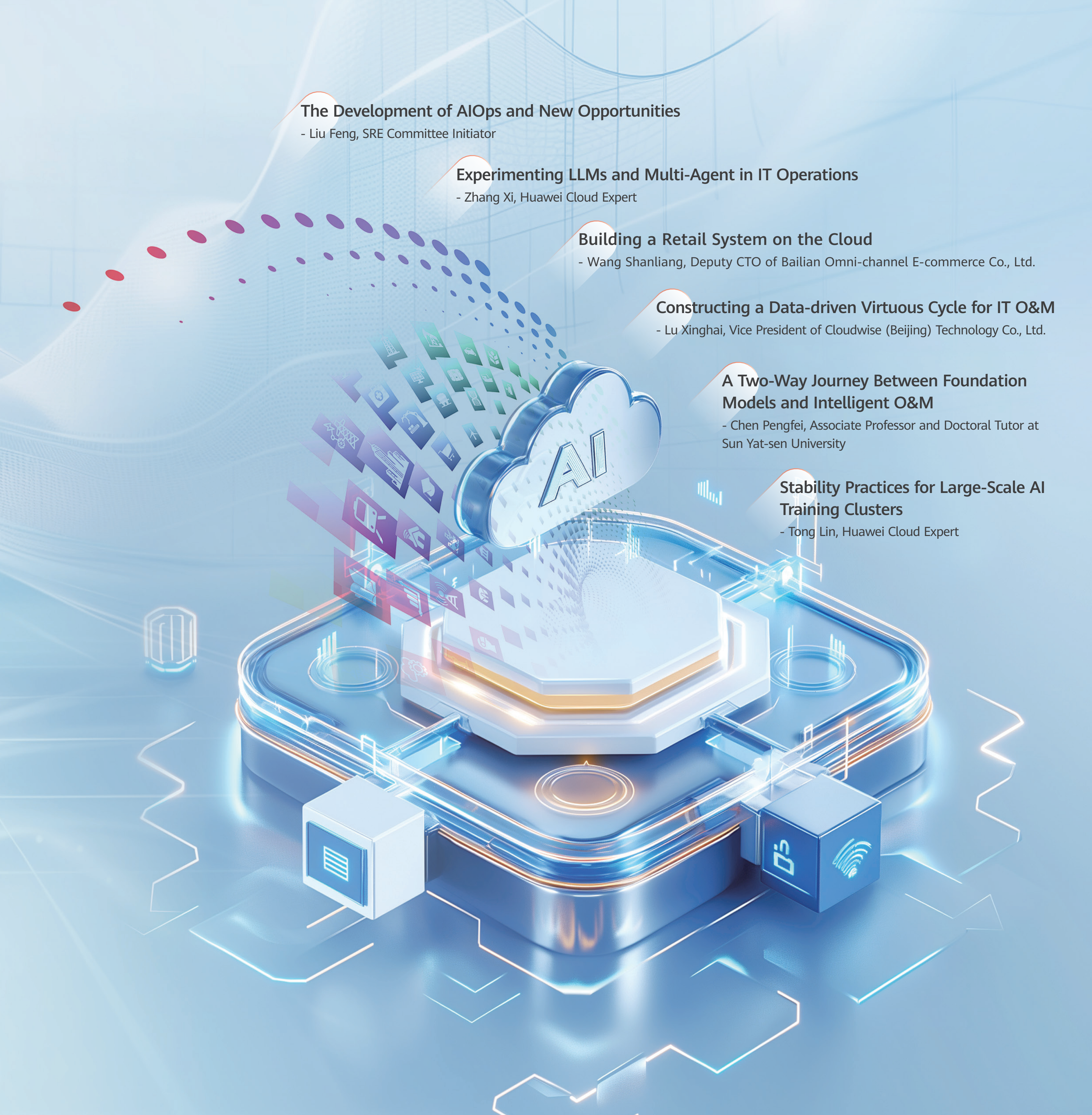
Huawei Cloud Deterministic Operations

Session 5

Internal Release
Free

| Frontier Views | Expert Interview | Technical Innovation | Service Improvement | Operations Milestones |

Intelligent Transformation Powered by Operations



Sponsor SRE Dept of Huawei Cloud Computing Technologies Co., Ltd.

Consultants Gao Jianghai, An Yu, Xue Ying

Editor-in-Chief Lin Huading

Deputy Editor-in-Chief Wang Qingyuan

Contributors Chen Pengfei, Huang Xiao, Jin Biao, Kang Zhen, Li Guoqiang, Liu Feng, Liu, Chunchun, Lu Xinghai, Tong Lin, Wang Shanliang, Wang Fuqiang, Wu Jie, Yang Jinquan, Yin Gezhen, Zhang Xi, Zheng Lei (sorted by alphabetical order)

Editors He Liping, Han Jianghui, Li Chaoqun, Wu Wenhui, Zhang Fei (sorted by alphabetical order)

Contact To obtain an electronic version, to request access to these articles, to contribute an article, or to provide suggestions or opinions, please contact the SRE Dept of Huawei Cloud Computing Technologies Co., Ltd.
Email: snzx1@huawei.com

Address Xiliubeipo Village, No. 9, Huanhu Road, High-tech Industrial Development Zone, Songshan Lake campus, Dongguan, China

Postcode 523830

General Disclaimer The content of this magazine is for reference only. Huawei Cloud Computing Technologies Co., Ltd. does not provide any express or implied warranty on the content of this magazine, including but not limited to a warranty of merchantability or use for a specific purpose. To the extent permitted by applicable law, in no case shall Huawei Cloud Computing Technologies Co., Ltd. be liable for any special, indirect, or consequential damages, or lost profits, data, business goodwill, or anticipated savings arising out of or in connection with any use of this magazine. (Internal release)

Copyright Copyright © Huawei Cloud Computing Technologies Co., Ltd. All rights reserved.
No part of this document may be reproduced or transferred in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Contents

Recommendations

Huang Qiangyuan

Vice President of Luckin

AI technologies have been evolving. How to maintain cluster stability for model training has become a challenge. The article, Stability Practices for Large-Scale AI Training Clusters, describes in detail how to use AI technologies to address complex service problems. It provides capabilities and practical experience for constructing AI clusters, including an effective system that enables E2E exception detection, diagnosis, and fault self-healing to ensure cloud system stability and reliability. Additionally, this article explores various technical methods to address stability issues. Implementing these approaches enables enterprises to not only advance intelligent technologies but also propel digital transformation. I hope this article will be a valuable resource for enterprises seeking to enhance their competitiveness through AI technologies.

Mi Penghui

Director of Huawei GTS Delivery Application Development & SRE Dept

Digital transformation is progressing in every industry, resulting in increasingly complex IT systems. Establishing reliable operational capabilities for IT systems in the mist of uncertainty has been a significant challenge in the SRE field.

The special issues of Huawei Cloud Deterministic Operations offer a wealth of experience from Huawei Cloud in managing large, complex systems. These issues also incorporate insights from O&M experts and exceptional enterprises across different industries. Whether seeking theoretical guidance or practical expertise, enterprises will be well-placed to implement Deterministic Operations. These issues provide a valuable wealth of O&M knowledge that no IT practitioners should overlook. We look forward to working with industry experts to explore a more effective approach to Deterministic Operations, safeguarding the digital transformation of enterprises.

Zeng Huashan

O&M director of Zhuhai Kingsoft Office Software Co., Ltd.

In a volatile, uncertain, complex, and ambiguous (VUCA) environment, characterized by Internet popularization and economic globalization, reliability and stability are vital to service quality and have become major challenges facing enterprises.

Deterministic Operations provides a range of effective strategies and advanced technologies, such as HA architecture design, emergency drills, AIOps, and more to keep services reliable and stable in an uncertain environment. This issue presents numerous cases that demonstrate how to navigate the uncertain path towards intelligent transformation and create more certainty.

With the power of AI, Deterministic Operations is poised for a brighter future.

The foreword writers are sorted in alphabetical order of their names.

Preface

Gao Jianghai
President of Huawei Public Cloud Business Dept

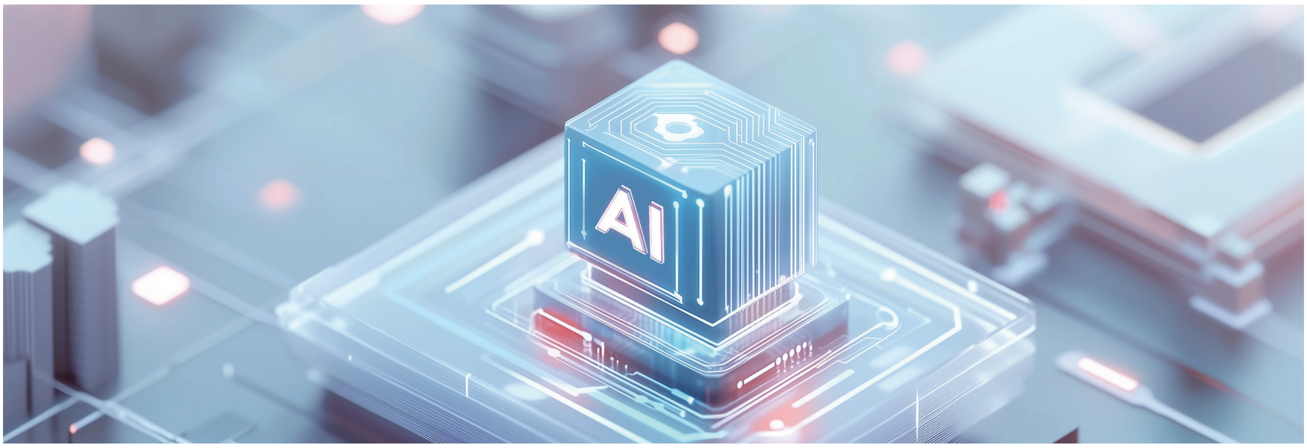
Driving Intelligent Transofrmation with Operations

Cloud computing, enabling universal access to the technologies, has become an engine for digital transformation. More and more enterprises are moving their core services to the cloud. Huawei public cloud services are growing rapidly. However, our customers come from diverse industries and vary in their abilities using the cloud. Ensuring secure, stable, high-quality, and user-friendly cloud services is crucial yet challenging. It places strict demands on the cloud infrastructure. We must implement a range of effective measures to minimize risk and address any issues that may arise. How to effectively manage the cloud and continuously drive service development has been a key focus for enterprises.

We prioritize security and trustworthiness, with a focus on stability and reliability. We also keep service agility front and center with an eye on controlling costs.

Deterministic Operations is the core of Huawei Cloud O&M. We use it to ensure design, development, deployment, monitoring, and O&M quality. We use it to lower fault rate, to minimize blast radii, and to recover from them as quickly as possible based on minimal and grid-based management. We use it to help customers transition from traditional O&M to platform-based O&M from the aspects of processes, awareness, quality culture, appraisal, and tools. We have been enhancing our abilities to provide greater certainty and meet service level objectives (SLOs) for fast growing services.

In September 2023, we officially launched the Deterministic Operations Elite Club, a platform for global customers and industry experts to discuss new technologies, ideas, and post-migration best practices and innovative solutions. We also share insights through special issues, white papers, and case studies. Together, we can build a secure, reliable world of Deterministic Operations.



The Development of AIOps and New Opportunities

Background

Major breakthroughs and rapid advances in next-gen AI technology, such as LLMs and AI foundation models, are driving new advances and creating new opportunities in site reliability engineering (SRE) and artificial intelligence for IT operations (AIOps). This article explains what AIOps is, how LLMs have developed over time, the synergy between AI and SRE for AIOps, and the changes caused by them.

Liu Feng

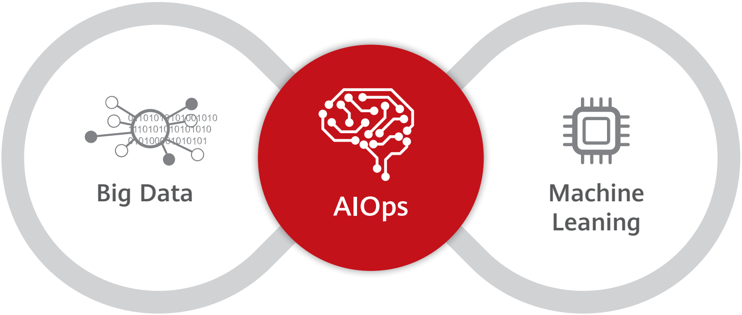
Initiator of the SRE Committee (SRE Community), one of the earliest SRE trainer in China, GitLab gold medal trainer, SRE evangelist.

What's AIOps

AIOps refers to the use of AI, analytics, and other technologies to automate IT operations.

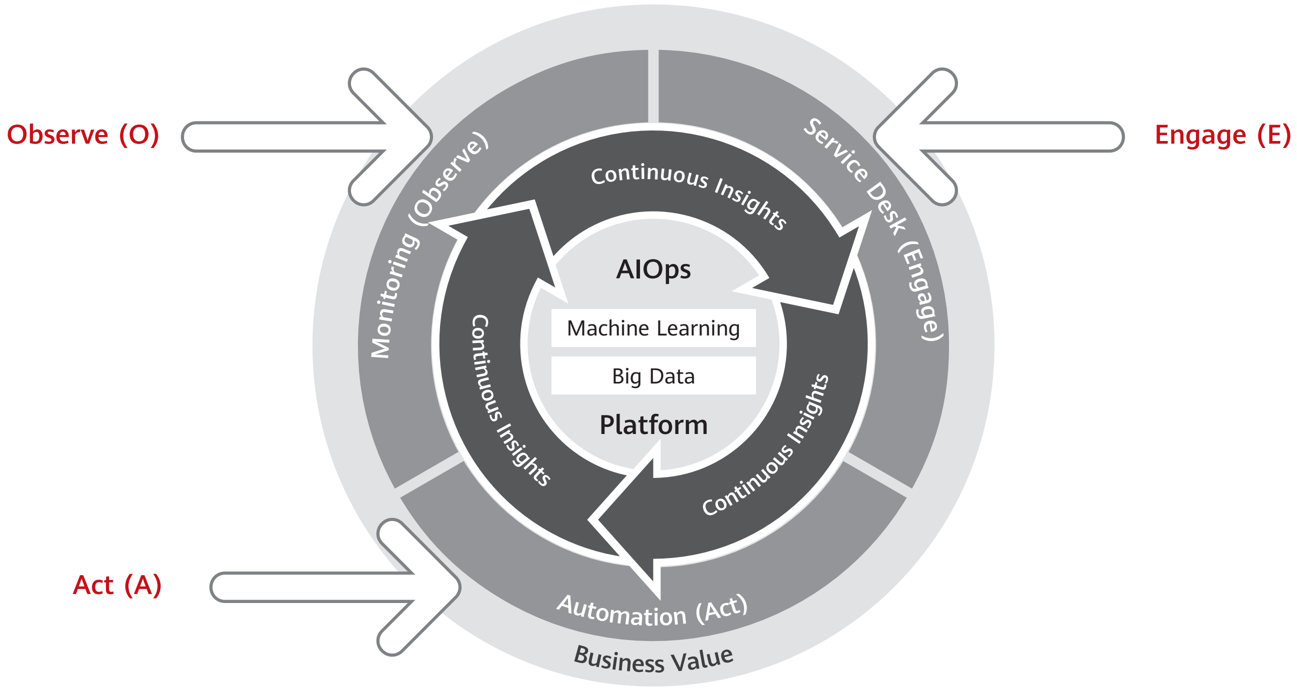
The term "AIOps" was coined by Gartner in 2016. It was originally used to describe an emerging industry where machine learning is used to tackle some of the challenges in operating ultra-large cloud infrastructure. AIOps is closely related to IT operations analytics (ITOA).

In practice, AIOps combines big data and machine learning to accelerate, automate, and simplify all tasks carried out to ensure the performance and reliability of IT systems.



Note: ITOA uses big data analytics alone, while AIOps combines AI and big data.

The Observe, Engage, and Act (OEA) Loop



The OEA loop is a decision-making model proposed for AIOps. This model represents the closed-loop process of interaction between AI systems and their environment, and it is often used in the context of AI-driven automation and decision-making systems.

The following describes each phase of the OEA loop:

1. Observe

During the observation phase, AI systems collect data and information from their environment. This may involve various data sources, sensors, logs, user interactions, and other relevant inputs. AI systems observe and analyze available data to gain insights into the current state of the environment and try to detect any significant issues.

2. Engage

Engagement follows observation. During the engagement phase, AI systems interact with their environment or stakeholders based on their analytical and decision-making capabilities. They can then provide suggestions, generate alerts or notifications, or start operations tasks according to predefined rules, policies, or machine learning models.

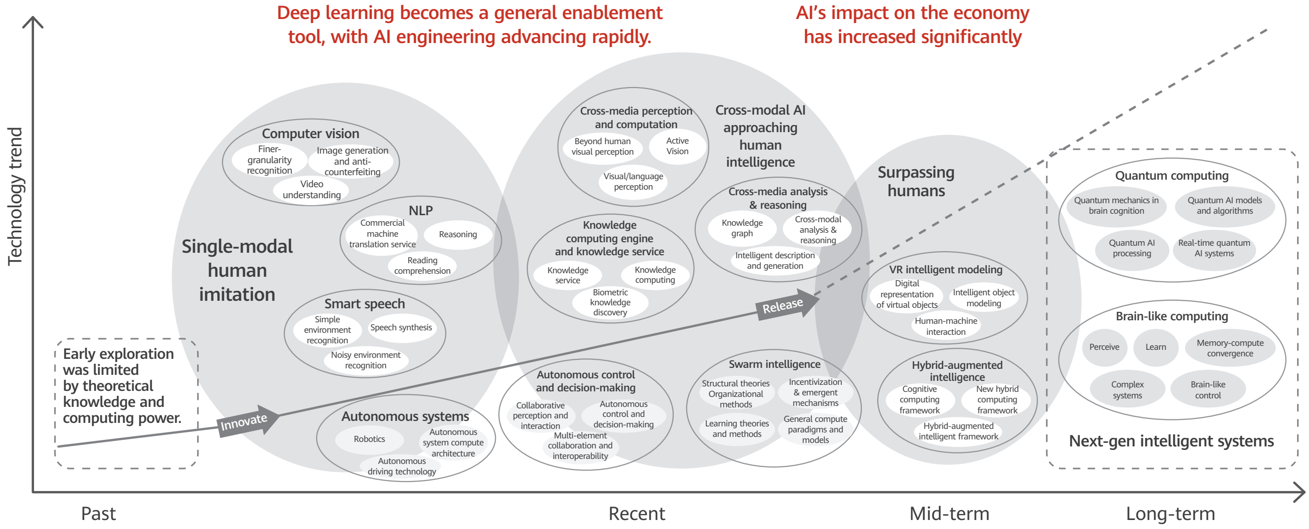
3. Act

During the action phase, actions are taken or decisions are made based on the insights and recommendations generated during the previous phase. AI systems take specific actions, execute operations tasks, or trigger predefined processes to accomplish the intended goals. These tasks can range from simple, automated tasks to complex workflows consisting of multiple steps and multiple rounds of interactions. After the action phase, the whole loop starts all over again, enabling AI systems to continuously adapt to the changing environment and improve their responses and become more automated.

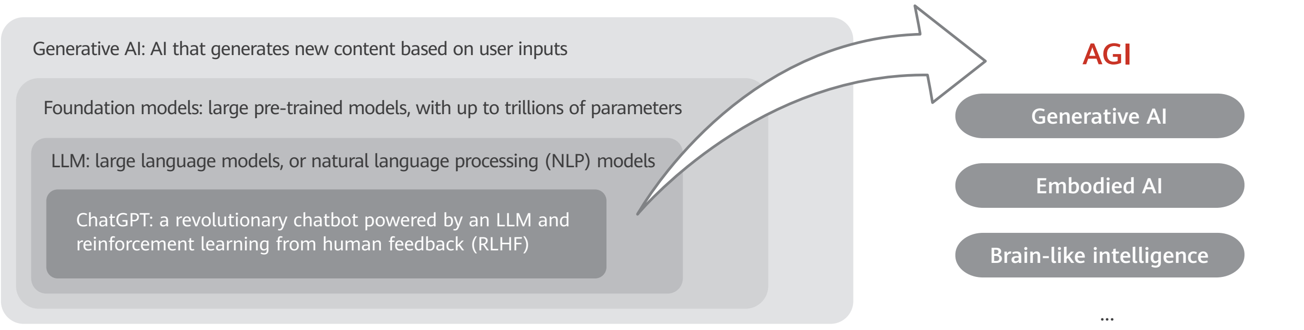
The OEA loop is typically used where real-time or near-real-time interaction is needed, such as in autonomous systems, IoT (Internet of Things) applications, customer service chatbots, or AIOps platforms. It enables AI systems to effectively collect and analyze data and information, take appropriate actions, and support automation and intelligent decision-making in a dynamic environment.

(Source: https://research.aimultiple.com/aiops/)

The Race to Artificial General Intelligence (AGI) Is Accelerating


Deep learning becomes a general enablement tool, with AI engineering advancing rapidly.
AI's impact on the economy has increased significantly
Technology trend
Early exploration was limited by theoretical knowledge and computing power.
Innovate
Single-modal human imitation
Computer vision
Finer-granularity recognition
Image generation and anti-counterfeiting
Video understanding
NLP
Commercial machine translation service
Reasoning
Reading comprehension
Smart speech
Simple environment recognition
Speech synthesis
Noisy environment recognition
Autonomous systems
Robotics
Autonomous system compute architecture
Autonomous driving technology
Cross-modal AI approaching human intelligence
Cross-media perception and computation
Beyond human visual perception
Active Vision
Visual/language perception
Knowledge computing engine and knowledge service
Knowledge service
Knowledge computing
Biometric knowledge discovery
Cross-media analysis & reasoning
Knowledge graph
Cross-modal analysis & reasoning
Intelligent description and generation
Release
Autonomous control and decision-making
Collaborative perception and interaction
Autonomous control and decision-making
Multi-element collaboration and interoperability
Swarm intelligence
Structural theories Organizational methods
Incentivization & emergent mechanisms
Learning theories and methods
General compute paradigms and models
Surpassing humans
VR intelligent modeling
Digital representation of virtual objects
Intelligent object modeling
Human-machine interaction
Hybrid-augmented intelligence
Cognitive computing framework
New hybrid computing framework
Hybrid-augmented intelligent framework
Quantum computing
Quantum mechanics in brain cognition
Quantum AI models and algorithms
Quantum AI processing
Real-time quantum AI systems
Brain-like computing
Perceive
Learn
Memory-compute convergence
Complex systems
Brain-like control
Next-gen intelligent systems
Past
Recent
Mid-term
Long-term

(Source: East China Branch of China Academy of Information and Communications Technology)



Generative AI: AI that generates new content based on user inputs

Foundation models: large pre-trained models, with up to trillions of parameters

LLM: large language models, or natural language processing (NLP) models

ChatGPT: a revolutionary chatbot powered by an LLM and reinforcement learning from human feedback (RLHF)

AGI

Generative AI

Embodied AI

Brain-like intelligence

...

(Source: East China Branch of China Academy of Information and Communications Technology)

The Three Phases of AI

Phase 1: Narrow AI

» Highly dependent on supervised learning algorithms, requiring large amounts of manually labeled data.

» Poor generalization performance: new models need to be developed from scratch for new tasks.

» We know exactly what the AI can or cannot do.

Phase 2: Broad AI

» Self-supervision: no need for supervised learning on labeled data.

» Better generalization: one model can be used for multiple tasks.

» Discriminative AI -> Generative assistant

» LLMs: emergent abilities and zero-shot learning

Phase 3: Artificial General Intelligence

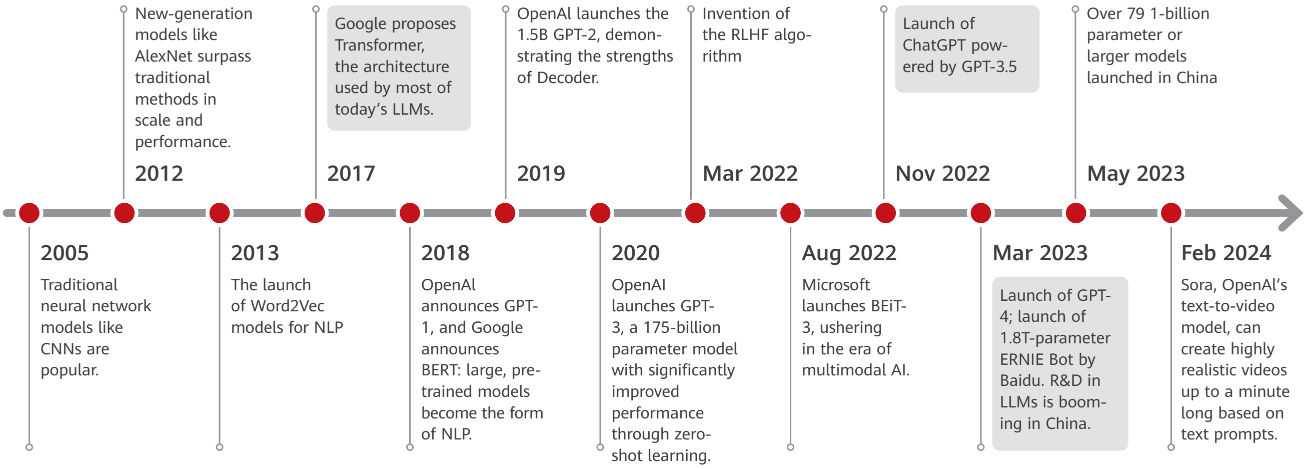
» Smarter than humans, and getting smarter

» Independent, autonomous...

» Better governance and oversight needed

LLMs and Trends

A history of LLMs



2005
Traditional neural network models like CNNs are popular.

2012
New-generation models like AlexNet surpass traditional methods in scale and performance.

2013
The launch of Word2Vec models for NLP

2017
Google proposes Transformer, the architecture used by most of today's LLMs.

2018
OpenAI announces GPT-1, and Google announces BERT: large, pre-trained models become the form of NLP.

2019
OpenAI launches the 1.5B GPT-2, demonstrating the strengths of Decoder.

2020
OpenAI launches GPT-3, a 175-billion parameter model with significantly improved performance through zero-shot learning.

Mar 2022
Invention of the RLHF algorithm

Aug 2022
Microsoft launches BEiT-3, ushering in the era of multimodal AI.

Nov 2022
Launch of ChatGPT powered by GPT-3.5

Mar 2023
Launch of GPT-4; launch of 1.8T-parameter ERNIE Bot by Baidu. R&D in LLMs is booming in China.

May 2023
Over 79 1-billion parameter or larger models launched in China

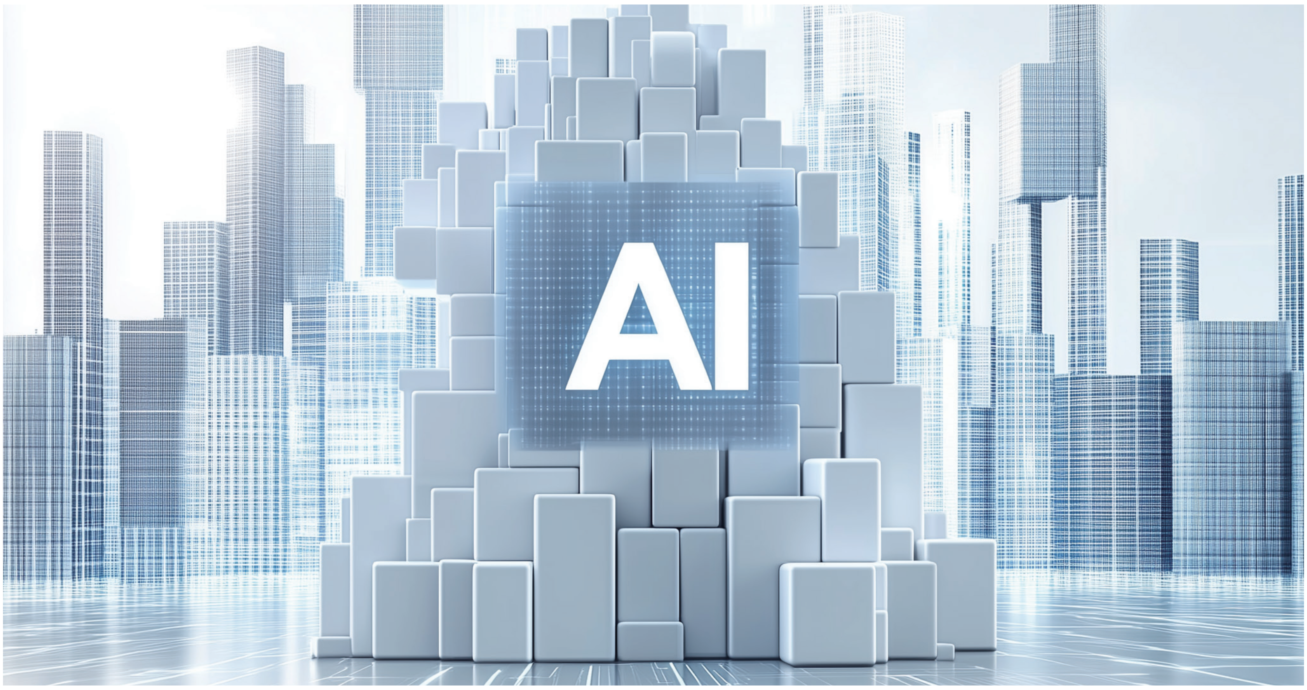
Feb 2024
Sora, OpenAI's text-to-video model, can create highly realistic videos up to a minute long based on text prompts.

(Source: East China Branch of China Academy of Information and Communications Technology)

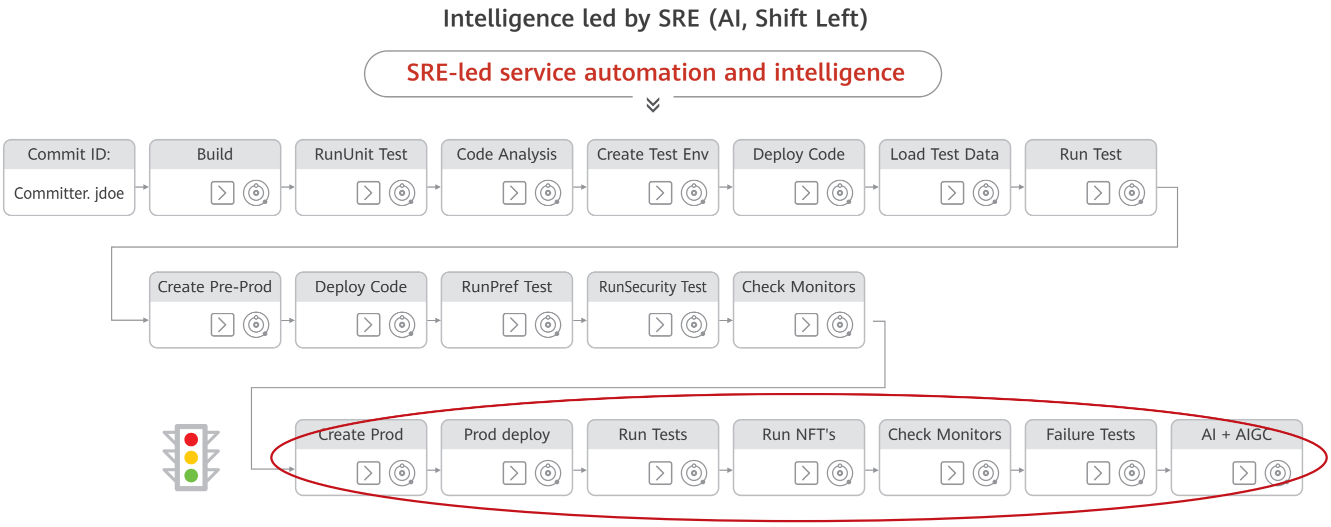
Large pre-trained models are deep learning models that have been pre-trained on large datasets and can be quickly adapted to a wide range of downstream tasks. Today, most of such models are LLMs. LLMs have the following characteristics:

- **Emergent abilities:** As the model reaches a certain size, the model's performance improves drastically, and often unexpectedly, at certain tasks.
- **Large models:** The model has at least 1 billion parameters, and in many cases, over 10 billion.
- **Generalization:** The model can be adapted to a wide range of downstream tasks through prompting and fine-tuning.

Based on the trends and development of LLMs and related technologies, the SRE community proposed an updated OEA model: comprehensive observability + LLM + OEA, so the industry can make full use of new technologies such as LLMs to power AIOps.



SRE + AIGC for AIOps



In the era of LLMs, SRE focuses on AI applications in production environments (Prod) — combining SRE and AI, experimenting and verifying machine learning algorithms and LLMs in a range of operations tasks to drive AIOps, and continuously improving the reliability of the SRE services.

Discussions on AIOps in the SRE Community

What's the state of AIOps today?

AIOps refers to the use of analytics, machine learning and deep learning algorithms, and other AI technologies to enhance and automate various aspects of IT operations. Data determines the upper limit of a model's performance, while algorithms and techniques are just the technical means we use to get near to this upper limit. This is why data governance is so important for AI applications.

AIGC has seen widespread adoption over the past couple of years, transforming our daily lives and work. Today, some LLM-powered AI tools are capable of a certain amount of logical reasoning.

Despite rapid progresses, today the mainstream AIGC models are generally considered unreliable due to their probabilistic nature. In many cases, they face problems such as hallucinations and the uncertainty of results, particularly when they are used to perform tasks bound by specific predefined rules.

What does AIGC mean for AIOps?

In what way will AIGC benefit AIOps? AIGC models offer general knowledge and logical reasoning capabilities. In the past, fault analysis heavily depended on data correlations, which were used to predict outages. AIGC not only collects these hidden correlations, but also combines white box information (for example, known code and service characteristics) to discover hidden issues in a wider range. This way, AIGC helps us quickly build an AIOps system.

We are already seeing AIGC boost productivity in some sub-domains, with immensely more use cases remaining to be explored. The former, such as code generation, image generation, and intelligent customer service with chatbots, already has some deterministic quality. For the latter, examples include multi-AI collaboration and personalized partners/assistants.

In the meantime, AIGC has helped with many breakthroughs and innovations in real-world applications. They include, but are not limited to, automated root cause analysis, automated generation of knowledge bases and handling suggestions, and automated fix execution in ITOps. In addition, it helps R&D engineers write code and implement software projects, develop testing solutions and cases, and more.

For AIOps, current LLMs are generative models that are based probabilistic models. This means that they will generate new things when making predictions or giving suggestions. For example, they are likely to expand existing failure types and show possible abnormal metrics and other anomalies. When using AIGC to generate documents

or solutions, it is important to note that the generated content is likely to contain a lot of useful information, as well as redundant or even incorrect information. This is why human engineers need to filter the information, rather than using it directly.

What Is the Impact of AIGC on ITOps or R&D for ITOps?

There are two main potential impacts of AIGC on ITOps or R&D for ITOps. One is the refactoring of existing processes. In this case, models like CodeLlama and StarCoder can be used to help with code generation, and models can be fine-tuned to enable automatic test case writing, vulnerability discovery, and enriching of submitted information. The other potential impact is more significant. Through the use of multi-agent collaboration in software engineering, models like ChatDev and Autogen can enable all-AI software engineering. The results so far are not yet ideal in complex scenarios, but the potential is immense.

For simpler tasks, such as disk alarms or disk write failures, AIGC can boost productivity significantly. In practice, it is necessary to understand the limits of the capabilities of LLMs, so we can objectively evaluate their application potential in different scenarios.

AIGC can undoubtedly be used for code generation in general domains. Other more specific examples include testing code generation, code expansion (e.g., for game developers), code summarization, invalid code check, and code consistency check.

We have faith that AIGC will deliver good results if used in R&D for ITOps. In other more general tasks, such as helping people create documents or providing enterprise knowledge bases, the results of LLMs have already been proven beyond doubt.

References:
1. What is AIOps, Top 3 Use Cases & Best Tools? in 2024 ; Written by Cem Dilmegani https://research.aimultiple.com/aiops/
2. Information collected by the East China Branch of China Academy of Information and Communications Technology (CAICT)

Experimenting LLMs and Multi-Agent in IT Operations

Abstract

This article starts by discussing the challenges faced by AIOps and goes on to introduce some practices in utilizing AIGC to enhance AIOps. It proposes an LLM-centric AIOps solution powered by multi-agent collaboration. It also offers some thoughts on the next-generation AIOps.

Zhang Xi

Huawei Cloud AI expert, PhD in Statistics from Utah State University, research directions: AI for Data, AI for BI, AIOps, and time series analysis; extensive experience operationalizing AI through enterprise applications, covering marketing, sales, services, supply, procurement, production, and R&D; supported multiple business application + AI projects at Huawei, and led teams to successfully tackle 5+ major technical challenges; and took a major role in launching multiple AI services.

The launch of ChatGPT has been hailed as the "iPhone Moment" of AI. Over the past couple of years, LLMs have been impacting a wide range of areas, such as office collaboration, finance, advertising, and marketing. They are also offering new ideas to tackle the long-standing challenges in IT operations. Our research has found that LLMs are creating new opportunities for AIOps. Cloud vendors are using LLMs to locate the root causes of outages and incidents and generate mitigation suggestions. Nearly 70% of the operations team are satisfied with the results generated by LLMs. We believe AIOps now faces the following challenges:

Challenge 1: Quick acquisition of massive amounts of knowledge, AI-assisted diagnosis and failure analysis

With the fast-growing popularity of LLMs and the generative AI applications powered by them, many ITOps professionals are exploring how to leverage LLMs to facilitate daily operations. This includes using LLMs to access operations knowledge, analyze failures, and suggest possible fixes or remediations. The key strength of our proposed AIOps solution is its ability to combine the extensive general knowledge built into LLMs with specialized IT operations expertise, and on the basis of

this, the ability to give accurate diagnosis and recommend plausible solutions or handling suggestions.

Challenge 2: Fast and accurate anomaly detection based on multimodal data

The ITOps domain has many profit and loss (P&L) data sources and various data types (or data modalities), such as metrics, logs, and traces (call chain). Metrics are time series data that shows service status and machine performance. Logs are unstructured text printed by programs or output of executed code. A trace is a call chain that shows the calling relationships between services in the process of executing a request. Such multimodal operations data can give you a comprehensive view of the system status, but the key is finding the best algorithms for different types of data, so that accurate fault diagnosis can be made.

Challenge 3: Quick root cause analysis based on complex operational data from various sources

Multimodal operations data can provide comprehensive information about the system, but it may be more difficult to decipher than single-modal data. Another important capability enabled by our solution is the ability to efficiently and accurately analyze alarm details contained in logs, metrics, and traces. It generates a real-time topology based on system traces, and identifies the root of a problem by excluding nodes that are just in the fault propagation path.

In reality, when and what types of failures will occur in an IT system are hard to predict. Most of the time we can only make an educated guess based on our experience and the symptoms we observe. By training an LLM on ITOps-specific knowledge using a few-shot learning method, you get a fine-tuned model that can offer more accurate fault diagnosis and classification suggestions. A model like this can help the operations team locate the root cause.

To address these challenges, we propose the solution illustrated below.

Solution Details

We propose an AIOps solution that is LLM-centric and leverages multi-agent collaboration for autonomous decision-making and automated recovery. Figure 1 shows the architecture of this solution. With this solution, we emphasize three key technologies.

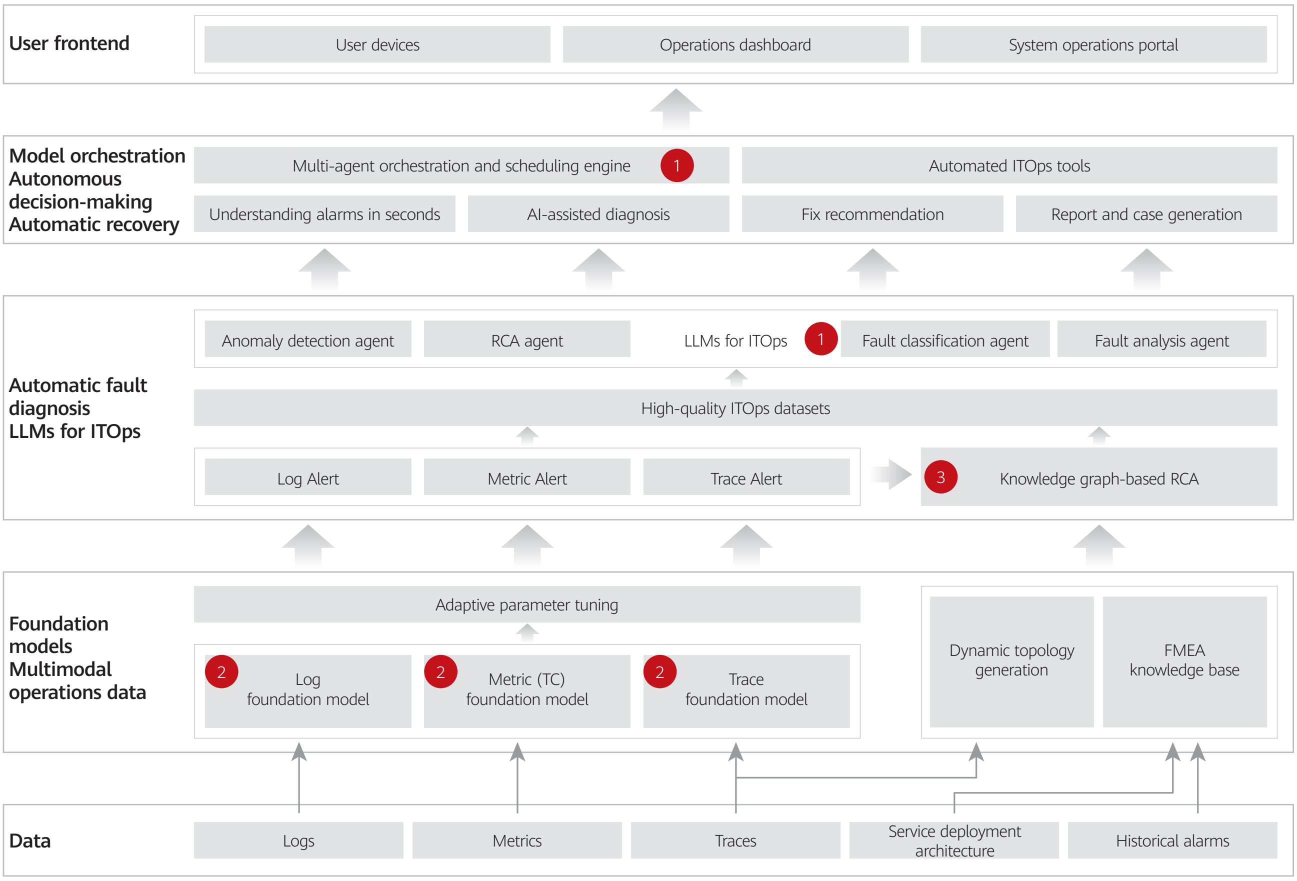


Figure 1: An AIOps solution powered by LLMs and multi-agent collaboration, capable of autonomous decision-making and automated recovery

Key technology 1: An LLM that embeds IT operations knowledge and experience

- Develop a high-quality dataset for the ITOps domain that includes failure and alarm information extracted from past incidents as well as information about failure and outage patterns. Then give the LLM the ability to diagnose and locate IT failures by fine-tuning it on this dataset or by connecting to an ITOps knowledge base.
- Use a master agent to coordinate multiple sub-domain agents, each of which is developed for a specialized task. This improves efficiency thanks to smooth collaboration between these agents.

Key technology 2: More powerful foundation models for anomaly detection based on multimodal data

- Aggregate hundreds of metrics calculated at different granularities and temporal intervals by applying a dimensionality reduction algorithm, thus accelerating anomaly detection. Once an anomaly is detected, analyze the most relevant metrics in detail.
- Parse semi-structured and unstructured logs in a targeted manner. For the semi-structured logs, focus on information extraction using templates, and then use uADR or sADR for anomaly detection. For unstructured logs, focus on semantic understanding and use a pre-trained Biglog model with Deep SVDD+SAD for anomaly detection.
- For structured trace data, extract call relationships and define key nodes; effectively identify abnormal nodes by converting time series data; and predict possible root causes and fault propagation directions based on topological relationship analysis.

Key technology 3: Root cause analysis based on a knowledge graph

- Use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to cluster data by time and identify anomaly events.
- Generate real-time topology views based on real-time traces.
- Analyze the fault propagation path with the help of a Failure Mode and Effects Analysis (FMEA) knowledge base.

We have built three anomaly detection models for different data sources: logs, metrics, and traces. When the system receives an anomaly, it notifies the LLM-powered master agent. The master agent makes decisions on anomaly handling and coordinates multiple sub-domain agents, which then collaborate together to diagnose the anomaly autonomously and jointly. Operational efficiency is improved through multi-task orchestration.

The first model is a metric anomaly detection model. We group operations data collected from different objects at different temporal intervals, and then extract differential features in different time windows based on changing metric curves. Then, we aggregate metrics across different periods and use multiple anomaly detectors to perform anomaly detection and temporal clustering. Once an anomaly is detected, the master agent is notified, which then instructs task-specific agents for specific actions. For example, upon receiving detailed instructions from the master agent, the detection agent calls multiple anomaly detectors to perform detailed anomaly detection.

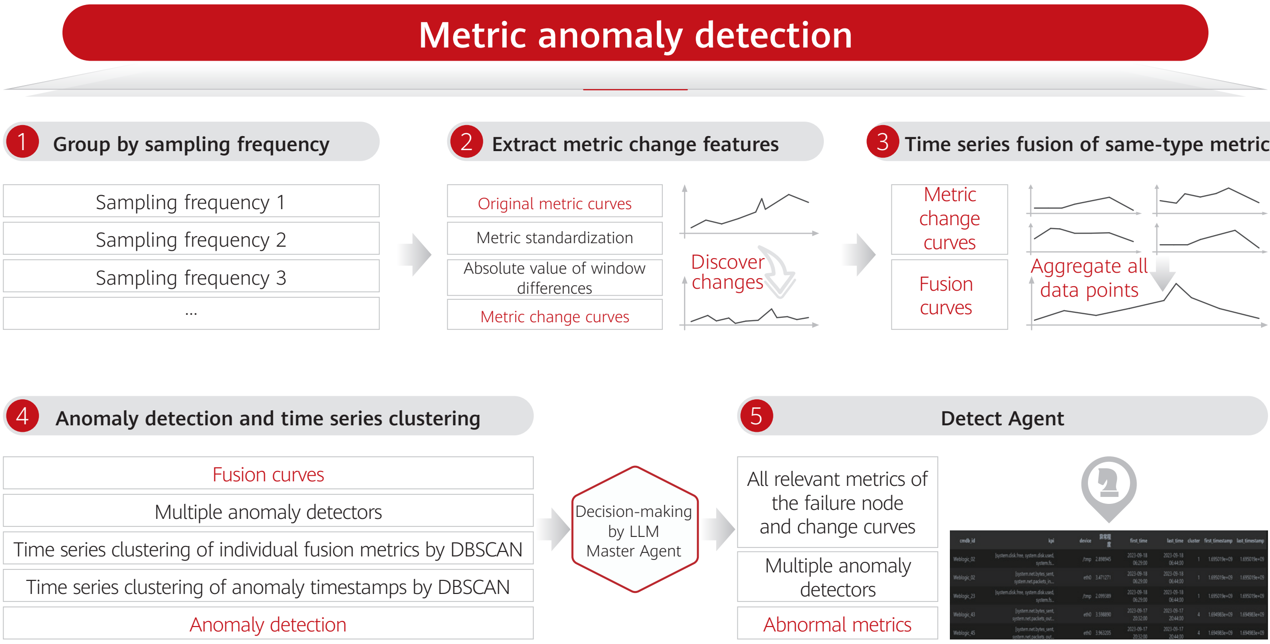
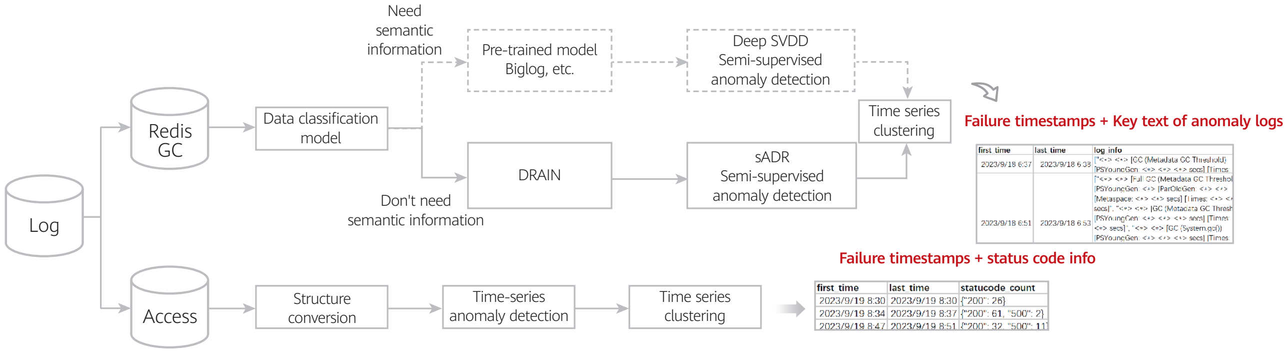

Metric anomaly detection
1 Group by sampling frequency
Sampling frequency 1
Sampling frequency 2
Sampling frequency 3
...
2 Extract metric change features
Original metric curves
Metric standardization
Absolute value of window differences
Metric change curves
Discover changes
3 Time series fusion of same-type metrics
Metric change curves
Fusion curves
Aggregate all data points
4 Anomaly detection and time series clustering
Fusion curves
Multiple anomaly detectors
Time series clustering of individual fusion metrics by DBSCAN
Time series clustering of anomaly timestamps by DBSCAN
Anomaly detection
Decision-making by LLM Master Agent
5 Detect Agent
All relevant metrics of the failure node and change curves
Multiple anomaly detectors
Abnormal metrics

Figure 2 Multimodal anomaly detection - metric anomaly detection model


Log
Redis GC
Access
Data classification model
Need semantic information
Pre-trained model Biglog, etc.
Deep SVDD Semi-supervised anomaly detection
Don't need semantic information
DRAIN
sADR Semi-supervised anomaly detection
Time series clustering
Structure conversion
Time-series anomaly detection
Time series clustering

Failure timestamps + Key text of anomaly logs

first_time	last_time	log_info
2023/9/18 6:37	2023/9/18 6:38	["<*> <*> [GC (Metadata GC Threshold) [PSYoungGen: <*> <*> <*> secs] [Times:
2023/9/18 6:51	2023/9/18 6:53	["<*> <*> [Full GC (Metadata GC Threshol [PSYoungGen: <*> [ParOldGen: <*> <*> [Metaspace: <*> <*> secs] [Times: <*> <* secs]", "<*> <*> [GC (Metadata GC Thresh [PSYoungGen: <*> <*> <*> secs] [Times: <*> secs]", "<*> <*> [GC (System.gc()) [PSYoungGen: <*> <*> <*> secs] [Times:

Failure timestamps + status code info

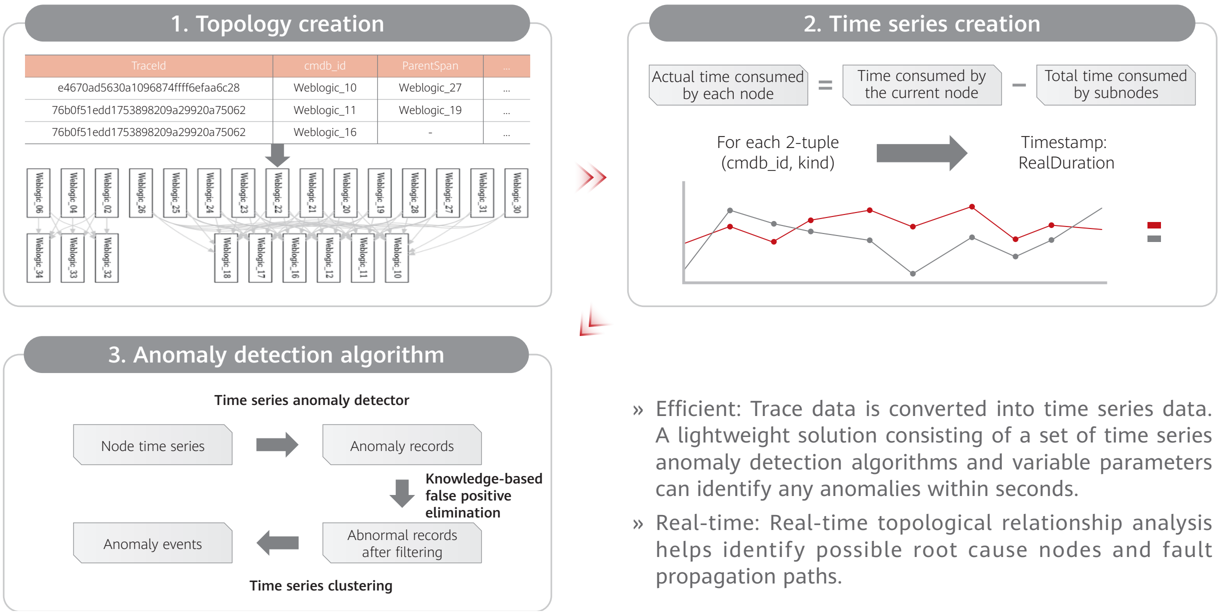
first_time	last_time	statucode_count
2023/9/19 8:30	2023/9/19 8:30	{"200": 26}
2023/9/19 8:34	2023/9/19 8:37	{"200": 61, "500": 2}
2023/9/19 8:47	2023/9/19 8:51	{"200": 32, "500": 11}

- » **Full coverage of log types:** automatic identification of log structures; adaptive time sequence matching for structured, semi-structured, as well as unstructured logs; semi-supervised anomaly detection based on a log parsing and extraction template; and a semi-supervised anomaly detection algorithm based on a pre-trained Biglog model and semantic understanding.
- » **Efficient real-time detection** based on real-time log streams and real-time model update.

Figure 3 Multimodal anomaly detection - log anomaly detection model

The second is a log anomaly detection model. Logs are sorted into two main categories: Redis GC and Access. For Redis GC logs, a structure classification model is used to check whether semantic information is required. If such information is required, a pre-trained LLM tailored to the ITOps domain, such as Biglog, is used. Then, Deep SVDD is used for semi-supervised anomaly detection. For the part that does not require semantic information, the DRAIN algorithm is used for template-based information extraction. Then, sADR is used to perform semi-supervised anomaly detection on the remaining part. Finally, temporal clustering is performed to output the fault occurrence time and key log text that records the anomaly. For Access logs, the time sequence is extracted from structured data. After temporal clustering, the fault occurrence time and status codes are generated.

For traces, the output consists of two parts: one is a real-time topology view generated based on the traces, which is important input for root cause analysis; the other is the call chain time sequence information, which is useful for anomaly detection.



TraceId	cmdb_id	ParentSpan	...
e4670ad5630a1096874ffff6efaa6c28	Weblogic_10	Weblogic_27	...
76b0f51edd1753898209a29920a75062	Weblogic_11	Weblogic_19	...
76b0f51edd1753898209a29920a75062	Weblogic_16	-	...

- » Efficient: Trace data is converted into time series data. A lightweight solution consisting of a set of time series anomaly detection algorithms and variable parameters can identify any anomalies within seconds.
- » Real-time: Real-time topological relationship analysis helps identify possible root cause nodes and fault propagation paths.

Figure 4 Multimodal anomaly detection - trace anomaly detection model

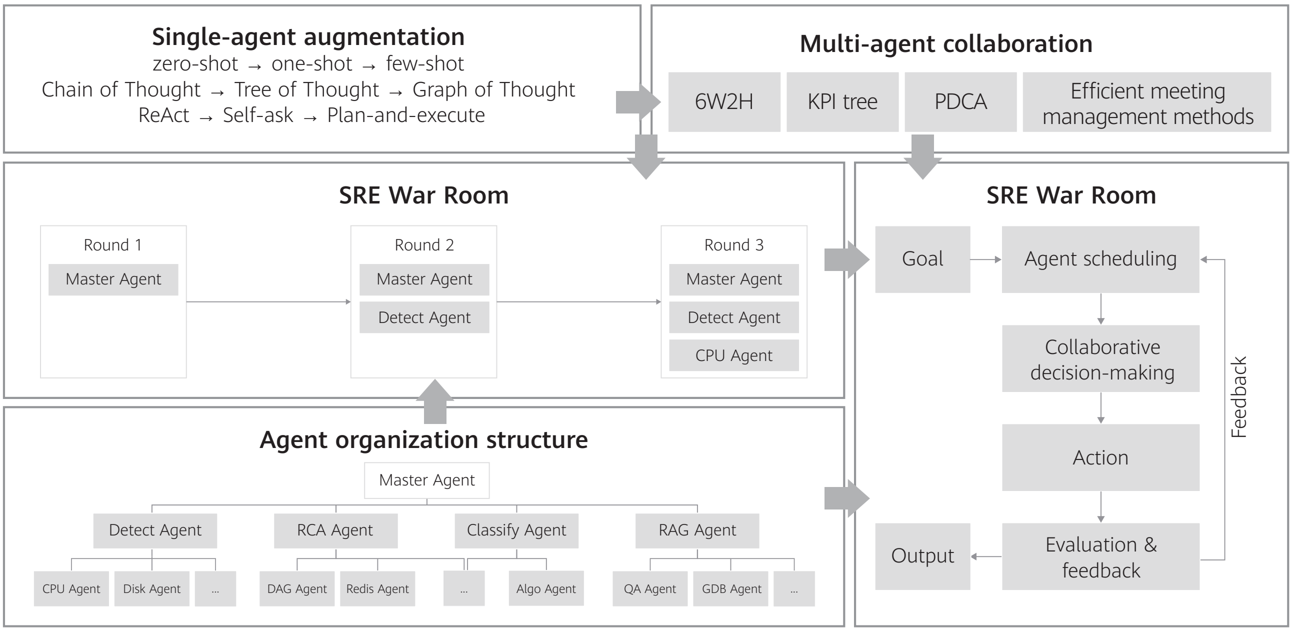
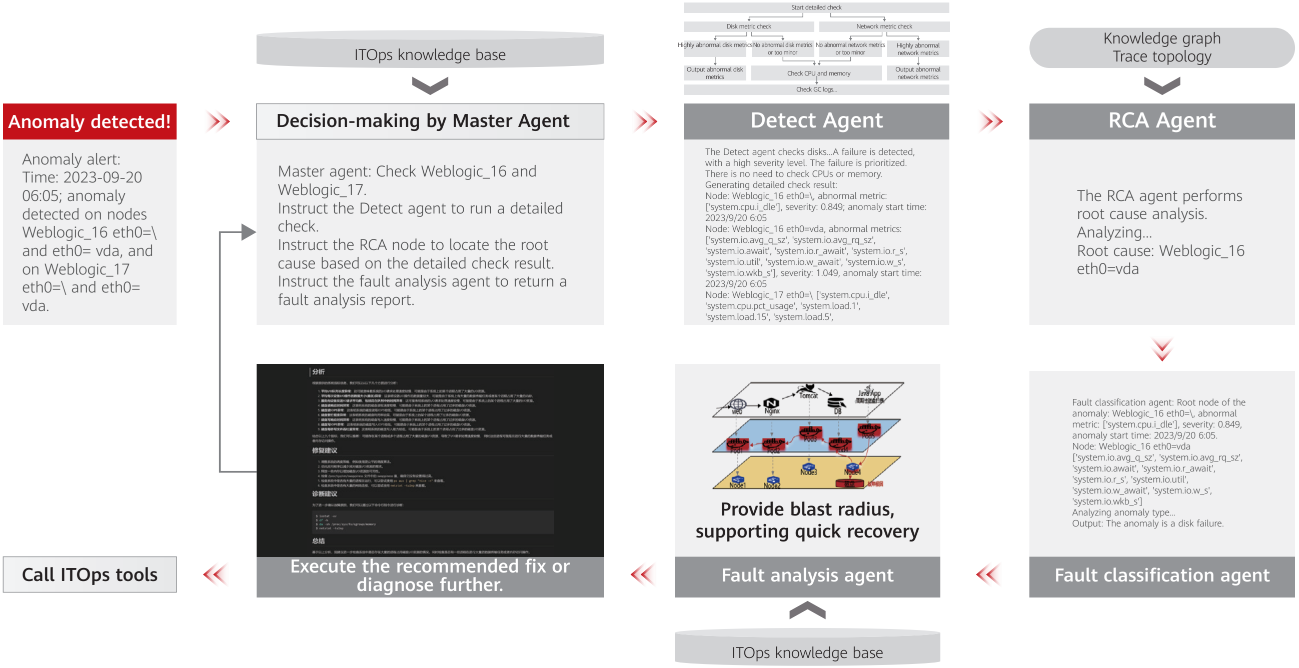


Figure 5 Multi-agent collaboration framework: efficient collaborative diagnosis by agents managed using a typical enterprise organization structure

Against the backdrop of the fast evolution of LLM and foundation model technologies, from zero-shot to one-shot, to few-shot, from chain of thought to tree of thought and on to graph of thought, from single-agent to multi-agent collaboration, multi-agent systems are driving new capabilities for LLMs, but they also present new technical challenges. Recent examples show that a multi-agent architecture, in the absence of proper organization management and collaboration mechanisms, performs even poorer than a single-agent architecture. We believe that in our quest for AGI and better LLMs, agents need to be treated as if they are real humans instead of tools. By this, we mean we need to use a more human-like approach to manage the organization and collaboration of agents in multi-agent scenarios. We hope to develop a multi-agent collaboration framework akin to our modern enterprise organization management systems. For example, we can use a KPI tree to break down tasks; use a Plan–Do–Check–Act (PDCA) cycle to improve the work efficiency of agents; and use an enterprise meeting management system to properly organize the entry and exit of agents in a War Room scenario. All these help to improve the efficiency of multi-agent collaboration.



Anomaly detected!

Anomaly alert:
Time: 2023-09-20 06:05; anomaly detected on nodes Weblogic_16 eth0=\ and eth0= vda, and on Weblogic_17 eth0=\ and eth0= vda.

Call ITOps tools

ITOps knowledge base

Decision-making by Master Agent

Master agent: Check Weblogic_16 and Weblogic_17.
Instruct the Detect agent to run a detailed check.
Instruct the RCA node to locate the root cause based on the detailed check result.
Instruct the fault analysis agent to return a fault analysis report.

分析

[illegible]

修复建议

[illegible]

诊断建议

[illegible]

```
$ iostat -xv
$ df -h
$ du -sh /proc/sys/fs/cgroup/memory
$ netstat -tulnp
```

总结

[illegible]

Execute the recommended fix or diagnose further.

Start detailed check
Disk metric check
Network metric check
Highly abnormal disk metrics
No abnormal disk metrics or too minor
No abnormal network metrics or too minor
Highly abnormal network metrics
Output abnormal disk metrics
Check CPU and memory
Output abnormal network metrics
Check GC logs...

Detect Agent

The Detect agent checks disks...A failure is detected, with a high severity level. The failure is prioritized. There is no need to check CPUs or memory.
Generating detailed check result:
Node: Weblogic_16 eth0=\, abnormal metric: ['system.cpu.i_dle'], severity: 0.849; anomaly start time: 2023/9/20 6:05
Node: Weblogic_16 eth0=vda, abnormal metrics: ['system.io.avg_q_sz', 'system.io.avg_rq_sz', 'system.io.await', 'system.io.r_await', 'system.io.r_s', 'system.io.util', 'system.io.w_await', 'system.io.w_s', 'system.io.wkb_s'], severity: 1.049, anomaly start time: 2023/9/20 6:05
Node: Weblogic_17 eth0=\ ['system.cpu.i_dle', 'system.cpu.pct_usage', 'system.load.1', 'system.load.15', 'system.load.5',

Provide blast radius, supporting quick recovery

Fault analysis agent

ITOps knowledge base

Knowledge graph
Trace topology

RCA Agent

The RCA agent performs root cause analysis.
Analyzing...
Root cause: Weblogic_16 eth0=vda

Fault classification agent: Root node of the anomaly: Weblogic_16 eth0=\, abnormal metric: ['system.cpu.i_dle'], severity: 0.849, anomaly start time: 2023/9/20 6:05.
Node: Weblogic_16 eth0=vda ['system.io.avg_q_sz', 'system.io.avg_rq_sz', 'system.io.await', 'system.io.r_await', 'system.io.r_s', 'system.io.util', 'system.io.w_await', 'system.io.w_s', 'system.io.wkb_s']
Analyzing anomaly type...
Output: The anomaly is a disk failure.

Fault classification agent

Figure 6 Fault diagnosis through multi-agent collaboration

Here, we use a flowchart to describe how multiple agents collaborate. At some point, the system detects that Weblogic_16 and Weblogic_17 are abnormal. Upon being notified of this anomaly, the master agent obtains the two nodes from the enterprise knowledge base and instructs the anomaly detection agent, root cause analysis agent, and fault analysis agent to get to work. The anomaly detection agent first performs a disk check based on the fault knowledge tree, and detects highly abnormal disk metrics. It then concludes that CPU check should be skipped and a high priority is assigned to disk check. The detection agent outputs all abnormal metrics, anomaly occurrence time, and the degree of abnormality of the two nodes. Based on the output of the anomaly detection agent, the root cause analysis agent identifies Weblogic_16 as the root cause node. The fault classification agent determines that a disk fault occurred based on the root cause node and abnormal metrics. The fault analysis agent provides a detailed fault analysis report, including a description of the blast radius and rectification suggestions.

Strengths of the Proposed Solution: Innovative, General, and Practical

Innovative

Automating complex operations tasks through multi-agent collaboration

- The multi-agent collaboration framework draws on modern enterprise organization and management methods, making it possible to perform complex operations tasks autonomously and efficiently.
- Multi-agent collaboration covers the operations process from end to end: anomaly detection -> root cause analysis -> fault classification -> fault analysis -> fix recommendation.

General

Addressing common challenges facing ITOps

- We offer three different models for anomaly detection on multimodal data: traces, metrics, and logs. These models are easy to deploy and ready out-of-the-box.
- The framework and algorithms do not rely on specific application scenarios, so they have a remarkable ability to generalize across a range of different tasks.

Practical

Fault recovery prioritized
Loose coupling of modules

- The diagnosis reports provide explainable blast radiuses of failures, providing an important reference for fast fault recovery in production systems.
- Different modules are loosely coupled and able to plug-and-play, making it possible to quickly respond to emergent, unpredictable requirements during faulty recovery.

Finally, our solution proposes a multi-agent collaboration framework that imitates the typical structure of an SRE organization. Multiple agents can collaborate together to handle complex operations tasks both autonomously and efficiently. Multi-agent collaboration covers the operations process from end to end: anomaly detection -> root cause analysis -> fault classification -> fault analysis -> fix recommendation. We also offer three different models for anomaly detection on multimodal data, including traces, metrics, and logs. The models are easy to deploy and are ready out-of-the-box. The framework and algorithms do not rely on specific application scenarios, so they have a remarkable ability to generalize across a range of different tasks. The diagnosis reports provide explainable blast radiuses of failures, providing an important reference for fast fault recovery in production systems. The modules are loosely coupled and able to plug-and-play, making it possible to quickly respond to emergent, unpredictable requirements during faulty recovery.

References

[1] Zhang S, Pan Z, Liu H, et al. Efficient and Robust Trace Anomaly Detection for Large-Scale Microservice Systems. ISSRE, 2023.

[2] Li D, Zhang S, Sun Y, et al. An Empirical Analysis of Anomaly Detection Methods for Multivariate Time Series. ISSRE, 2023.

[3] Wang Z, Liu Z, Zhang Y, et al. RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models. arXiv, 2023.

[4] Jin P, Zhang S, Ma M, et al. Assess and Summarize: Improve Outage Understanding with Large Language Models. ESEC/FSE, 2023.

[5] Chen Y, Xie H, Ma M, et al. Empowering Practical Root Cause Analysis by Large Language Models for Cloud Incidents. arXiv, 2023.

[6] Zhou X, Li G, Sun Z, et al. D-Bot: Database Diagnosis System using Large Language Models. arXiv, 2023.

[7] Zhou X, Li G, Liu Z. Llm as dba. arXiv, 2023.

[8] Wen Q, Gao J, Song X, et al. RobustSTL: A robust seasonal-trend decomposition algorithm for long time series. AAAI, 2019.

[9] Liu Y, Tao S, Meng W, et al. LogPrompt: Prompt Engineering Towards Zero-Shot and Interpretable Log Analysis. arXiv, 2023.

[10] Tao S, Liu Y, Meng W, et al. Biglog: Unsupervised large-scale pre-training for a unified log representation. IWQoS, 2023.

[11] Ma L, Yang W, Xu B, et al. KnowLog: Knowledge Enhanced Pre-trained Language Model for Log Understanding. ICSE, 2023.

[12] Zhong Z, Fan Q, Zhang J, et al. A Survey of Time Series Anomaly Detection Methods in the AIOps Domain. arXiv, 2023.

[13] Wu H, Hu T, Liu Y, et al. Timesnet: Temporal 2d-variation modeling for general time series analysis. ICLR, 2023.

[14] Yu G, Chen P, Li P, et al. Logreducer: Identify and reduce log hotspots in kernel on the fly. ICSE, 2023.

Empowering Industries with Professional Services for Huawei Pangu Models

Professional services for Pangu models

Consulting service on industrial AI | AI enablement, optimization, and improvement | Industry AI competitions and training service | Huawei Cloud AI training service

Core competitive strengths

High-quality data | Rich selection of industry-specific models | Easy-to-use tools | Security and compliance

Building a Retail System on the Cloud

– Bailian Group Moved Service Middle-Ends to the Cloud

Abstract

Bailian Group has moved service middle-ends of all channels to the cloud. It was a challenging process that involves various services, complex architecture, and the intricacies of cloud migration. To ensure a smooth transition, various effective measures, including a pilot project, traffic switchover with grayscale deployment, testing, and drills, were implemented. New digital systems powered by cloud computing technologies are ready to create a brighter future for retail.

Wang Shanliang

Deputy general manager of Bailian Omni-channel E-commerce Co., Ltd. He oversees O&M, ensures data security, and manages IT service desks.

Digital Transformation of Bailian Group

Bailian Group was founded in April 2003 through the merger of Shanghai 100 (Group) Co., Ltd., Hualian (Group) Co., Ltd., Shanghai Friendship (Group) Co., Ltd., and Shanghai Materials (Group) Head Office. Bailian was initially base out of Shanghai and later expanded into the Yangtze River Delta region. It has now established a nationwide presence, operating a diverse range of retail establishments, including department stores, shopping centers, large stores, standard supermarkets, convenience stores, and specialized stores. These stores cater to various sectors, such as car trade, e-commerce, warehousing and logistics, consumer services, electronic information, and more. By the end of 2023, Bailian had established about 4,700 retail outlets across China. The Yangtze River Delta region accounted for over 80% of these outlets, while Shanghai alone accounted for approximately 50%. The number and total area of department stores, shopping malls, outlet stores, and number of supermarkets all rank first in Shanghai. Additionally, Qingpu Outlet has consistently held the top position in national sales for many years. In 2015, Bailian Group launched Bailian Omni-Channel E-Commerce Co., Ltd. with the aim of advancing the group's e-commerce initiatives. This subsidiary focuses on two key areas: e-commerce development and the transformation of traditional services.

On May 19, 2016, the iBailian platform was officially launched. This platform simplifies the membership and bonus points system across all their branches. Customers can use the same membership card for all stores. Since its launch, the iBailian platform has evolved from digitizing stores and establishing electronic business platforms to a series of digital transformation projects at the group level. By August 16, 2024, all iBailian service middle-ends had been moved to the cloud.

Cloud migration is a significant part of the "1+2+N" digital transformation strategy of Bailian Group. A successful migration sets a strong groundwork for future cloud-based innovations, ultimately leading to Bailian's achievement of digital and intelligent transformation.

Challenge Analysis for Cloud Migration

Bailian Group has diverse and intricate service architecture. The iBailian middle-ends consist of both group-level functions like memberships, payments, and promotions, as well as specialized services for subsidiary companies. For example, it handles customer attraction for Bailian Co., Ltd. and home delivery services for Lianhua Supermarket and First Pharmaceutical Co., Ltd. Moreover, the platform also interacts with systems from various external partners.

Bailian Group faced challenges in three main areas during cloud migration: diverse group services, a complex architecture, and the intricacies of the migration process.

» **Service challenges:** The migration involves various branches such as department stores, shopping malls, supermarkets, logistics, and pharmacies. Some faults could have significant economic and social impacts on people's lives.

» **Technical challenges:** Complex call chains between microservices make it difficult to implement multiple switchovers. They can also result in long periods of downtime during a one-time switchover. iBailian service middle-ends involve more than 4,000 operating systems, 100 databases, 70 middleware resources, 600 microservices, 1,300 scheduled tasks, and 2,000 internal domain names. The workload of cloud migration, including adaptation and reconstruction tasks, is immense.

» **Management challenges:** The cloud migration involved 9 subsidiary companies, more than 10 level-1 departments, and nearly 300 third-party partners. The coordination and communication are quite challenging. It is particularly difficult to determine a unified downtime window that could minimize possible impacts. The cloud migration process involves adapting and reconstructing cloud systems, automating and reconstructing the O&M platform, designing downgrade solutions, conducting downgrade drills, performing cloud migration drills, and carrying out cloud verification. More than 120 formal meetings were held before the cloud migration solution was finalized, and nearly 200+ people were on site for assurance at the night of the cloud migration.

Complex Business

Online	Offline
All channels covered Membership center Payment center Points center ...	20+ parking lots 4,000 physical stores ...

Various business types

9+ subsidiary companies

Multiple third-party platforms

Multiple platforms from 300 third-party suppliers

- The migration involves various branches such as department stores, shopping malls, supermarkets, logistics, and pharmacies. Some faults could have significant economic and social impacts on people's lives.

Complex Technologies

Numerous IT resources

4,000 operating systems
100 databases
XX middleware resources

Numerous microservices

600+ microservices
1,300+ jobs
2,000+ internal domain names
4,000+ test cases

Complex reconstruction and cloud migration

200+ runbook tasks

- Complex call chains between services increased the risk associated with both multiple switchovers and one-time switchover.
- Immense adaptation and reconstruction workload

Challenging Management

Multiple departments

9+ subsidiary companies
10+ level-1 departments
300+ third party suppliers

Difficult personnel and project management

1,000+ project tasks
120+ formal meetings
200+ personnel for assurance during the switchover
5 times of drills

- It was difficult to determine a downtime window when various business types and multiple suppliers were involved.
- Heavy coordination across departments

Three-Step Solution

After thoroughly analyzing the challenges of cloud migration, the Bailian team worked with the Huawei Cloud team to design a three-step solution. This solution includes a pilot project, traffic switchover with grayscale deployment, and testing and drills.

Pilot Project

Given the complex systems involved and the size of the potential great impacts, we designed a pilot project to validate the feasibility of the migration process and solution. This allowed us to identify any issues or risks and ensure a smooth migration.

The pilot project helped:

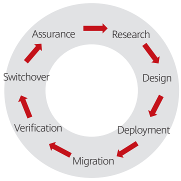
» Verify the cloud migration solution.
» Improve team collaboration.
» Identify migration risks.
» Offer hands-on experience.

The multiple small-scale attempts, summary, and optimization established a strong basis for a successful cloud migration.

Traffic Switchover with Grayscale Deployment

After the cloud environment was set up, we enabled applications to access the databases in the original equipment room. In addition, we redirected a small portion of the traffic to the Huawei Cloud production environment to validate the deployment from the network layer to the application layer. The application layer was gradually migrated. After confirming that there were no issues, we switched all the application layer and data layer to Huawei Cloud.

Pilot Project

• Solution validation • Team collaboration improvement • Risk identification • Hands-on experience	• Relatively independent applications with only a few dependencies • Non-core services with minimized impacts • The pilot project could validate a solution or demonstrate the advantages of cloud migration.	 Assurance Research Design Deployment Migration Verification Switchover	• Problem identification • Impact analysis • Improvement measures • Experience summary
Confirming objectives 1	Selecting applications 2	Executing the migration process 3	Summarize the pilot project 4

Pilot Project Examples

	Objective	Description
Solution validation	Switch plan	Nginx traffic switchover with grayscale deployment (1%, 10%, 50%) Cloud-based middleware and database access Web Application Firewall (WAF) traffic switchover
	Automatic deployment	CMDB-based, automatic application deployment
	Test plan	Gray environment access through internal Wi-Fi
Team collaboration improvement	Communication methods	Six communication methods
	Operational regulations	Switchover, reporting, review, and decision-making regulations
Risk identification	Cloud-based access latency	The impact of latency in cloud-based access on user experience during the traffic switchover
	Whitelisted IP addresses	The impact of IP address changes on business and third-party services
Hands-on experience	Migration tools	MongoDB and MySQL migration using Data Replication Service (DRS)
	Cloud service integration	Application adaptation and reconstruction for accessing Object Storage Service (OBS) CMDB integration with Huawei Cloud

Testing and Drill Solution

The cloud migration involves many steps, including function testing, performance testing, and switchover drills. To simulate the cloud migration, a drill environment was established to clone the production environment using Huawei Cloud best practices.

The drill environment was set up to:

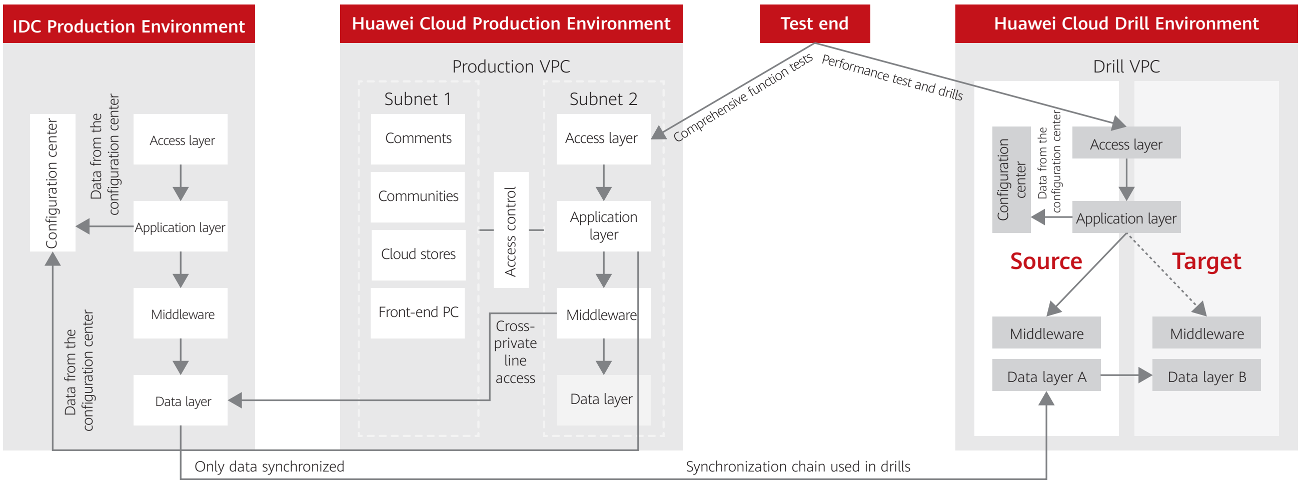
- Verify the steps, scripts, and automatic system release to ensure a smooth setup of the production environment.
- Conduct cloud migration drills.
- Perform pressure testing to identify risks.

The drills helped:

- Familiarize the team with the cloud migration process and minimize the downtime.
- Ensure effective teamwork.
- Optimize the runbook to ensure it is accurate and comprehensive. Some steps were automated using scripts and tools to improve operational accuracy and efficiency.
- Identify problems to reduce risks on the night of the switchover.
- Improve technologies and capabilities to efficiently address problems.

Minimize problems and reduce the time of cloud migration to six hours through four rounds of drill-based verification and one rollback

Given the intricacy of the cloud migration project, we created a checklist and runbook following Huawei Cloud's best practices. The checklist guaranteed that all necessary



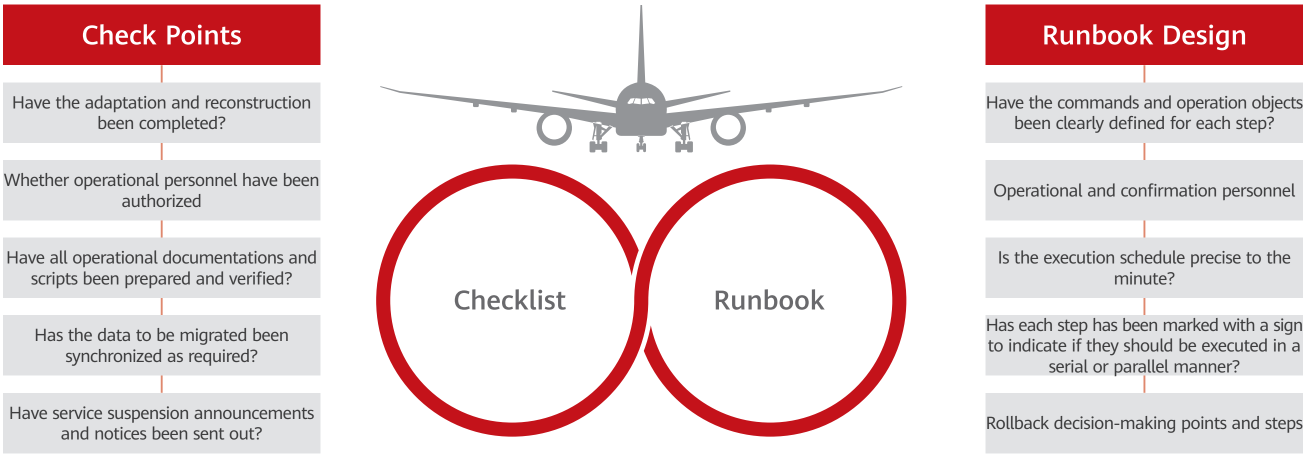
Function tests
The access layer and application layer are set up on the Huawei Cloud production VPC to access the IDC production data layer.

Performance tests
1:1 production environment simulation

Switchovers
From the source end to the target

preparations were done prior to the switchover. The runbook ensured that the on-site team of over 200 individuals could smoothly carry out operations according to the agreed-upon steps during the switch. The runbook provided clear details for each operation, including the operator, confirmer, command script, estimated duration, serial/ parallel indication, and rollback decision point. The checklist encompassed 67 checkpoints, with over 300 executed during the five drills. The official runbook includes 236 tasks, and over 1,000 tasks were executed during the drills.

The sufficient preparation accelerated the cloud migration and helped ensure system stability after the migration.



Check Points

- Have the adaptation and reconstruction been completed?
- Whether operational personnel have been authorized
- Have all operational documentations and scripts been prepared and verified?
- Has the data to be migrated been synchronized as required?
- Have service suspension announcements and notices been sent out?

Checklist

Runbook

Runbook Design

- Have the commands and operation objects been clearly defined for each step?
- Operational and confirmation personnel
- Is the execution schedule precise to the minute?
- Has each step has been marked with a sign to indicate if they should be executed in a serial or parallel manner?
- Rollback decision-making points and steps

Summary of Experience and Value

There are six key factors that contribute to the successful migration of large-scale, complex service systems to the cloud:

» **Clear strategic objectives** for decision-making and resource allocation. Huawei Cloud consulting services supported Bailian in creating a cloud transformation blueprint and a strategic plan that included three steps: establishing a foundation, developing capabilities, and consolidating and optimizing.

» **Mature methodology**. The Huawei Cloud migration methodology served as a framework to ensure a smooth and organized cloud migration.

» **A holistic perspective from day one**. Cloud migration requires a holistic perspective in the early stages. Being prepared guarantees success, while inadequate readiness leads to failure. At the early stages of the project, the project team had started to design milestones for the entire cloud migration. This allowed them to determine the time line and necessary tasks, while also providing a comprehensive overview of the project delivery.

» **Robust organizational support**. As a highly significant project for the Bailian Group, we received extensive backing from leaders at the group level, as well as from various business branches, subsidiary companies, and colleagues across departments. Clear roles and responsibilities were established among the project teams.

» **Streamlined team collaboration** was achieved through effective communication, coordination, and decision-making. The project teams of Huawei Cloud and Bailian Group approached the project in a serious and rigorous manner. We established open communication channels, followed up on work breakdown structures (WBS), and fostered a culture that encourages critical thinking and proposing solutions.

» **A comprehensive solution**. To ensure that the solution could be carried out smoothly, we employed several measures, such as breaking down tasks, prioritizing solutions, iterating through a pilot project, conducting drills, reviewing and summarizing, and gaining valuable experience.

This cloud migration has delivered three benefits:

- DR became easier with VPC peering connecting multiple equipment rooms
- System availability improved with applications deployed in multiple equipment rooms. Systems can be protected from equipment room faults caused by natural disasters and unforeseen circumstances. Prior to migrating to the cloud, cross-region disaster recovery (DR) was implemented for databases, but high availability was not achieved for the applications. Migrating to the cloud, however, with databases deployed in two equipment rooms, resolved this problem.
- Faster service innovation was enabled. Mature PaaS products combined with faster resource request and deployment improved rollout efficiency and drove service innovation.

Easier DR

- No equipment rooms or private lines are required. A high-speed free intranet links AZs in the same city.
- No extra DR configuration required. Cloud services, such as Relational Database Service (RDS), Object Storage Service (OBS), Distributed Cache Service (DCS), and Elastic Load Balance (ELB) all support DR.
- DR construction became easier.

Higher System Availability

- Instead of having a single Internet data center (IDC), Bailian now has multiple equipment rooms spread across different Availability Zones (AZs) in the same city. This significantly enhanced the overall availability.
- Unexpected downtime was minimized.

Faster Service Innovation

- New technologies, such as SaaS, APIs, and SDKs, has accelerated service innovation.
- Global cloud infrastructure allows for rapid deployment and rollouts close to customers.
- New features can be launched to the market and deployed faster.

Constructing a Data-driven Virtuous Cycle for IT O&M

–Methods and Practices of Data-driven O&M

Preface

In recent years, the industry has placed significant emphasis on data-driven O&M. However, there is still a need to understand how data actually influences O&M, the methods involved, and how to effectively implement data-driven O&M. Currently, there is a lack of established solutions in this area. The methodology for data-driven operations (DDOps) is presented in this article. We have created a data-driven virtuous cycle for O&M, offering customers a comprehensive solution that combines theoretical guidance, strategies, and practical methods. Our aim is to provide valuable insights and assistance on utilizing data to enhance O&M practices in the industry.

Huang Xiao

Chief consultant of Cloudwise. He worked for a prominent, publicly listed software company in China and has been engaged in IT service management since 2006. He is one of the earliest experts in advocating and implementing ITIL in China. Additionally, he is a member of the Information Technology Service Standards (ITSS) expert team and has played a key role in formulating numerous international and industry standards for IT service management. He has diverse experience, covering product R&D, marketing, pre-sales promotion, solution design, business consulting, project implementation, and contributing to the development of international standards in the IT field. He has led a team to release several independent intellectual property products in the IT domain. He has also played a leading role in overseeing the delivery and implementation of numerous IT O&M projects worth tens of millions of dollars.

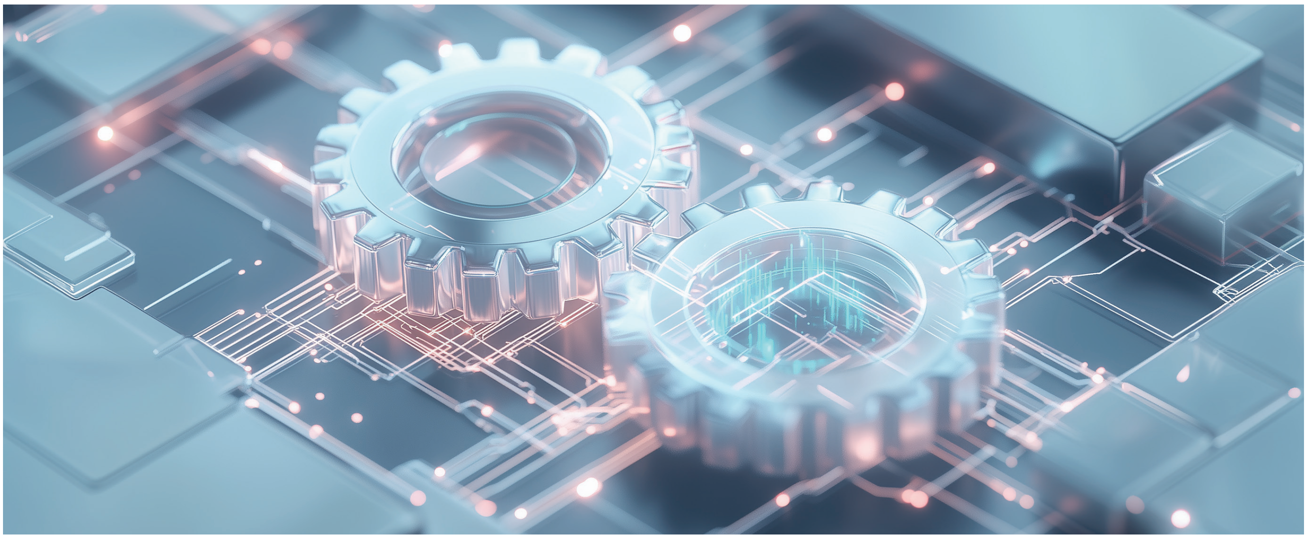
Lu Xinghai

Vice President of Cloudwise. He is now in charge of consulting services. He has 15 years of experience in product planning, design, and research and development in the fields of internet, informatization, and operations and maintenance. He is one of the early pioneers and experts in IT service domains in China. Additionally, he is a member of the intelligent O&M international standards compilation team, and also a member of the Information Technology Application Innovation Working Committee expert team.

When he was in charge of Cloudwise products, he successfully spearheaded the implementation, marketing, and promotion of over a dozen operation-related core products, winning recognition from key customers in multiple industries. Additionally, he holds six invention patents and has authored the book "O&M Data Governance: Building a Cornerstone for Intelligent O&M". He has also co-translated the book "Digital Economy 2.0: Detonating the Ecological Dividends of Big Data".

Wu Jie

Vice President of Cloudwise. He has been engaged in information consulting for 10 years. He has developed a wealth of expertise in government digital transformation, enterprise IT planning, IT service management, and information security system construction. He has offered consulting services to numerous users both within and outside of China, spanning various industries including finance, communications, transportation, government and enterprise, and the Internet. Additionally, he is an expert in intelligent O&M compliance evaluation and an independent evaluator for ITSS.



Developing a Date-Driven Virtuous Cycle for O&M

The key of data-driven O&M is to create a data-driven virtuous cycle, rather than simply managing and applying data in isolation. (Figure 1)

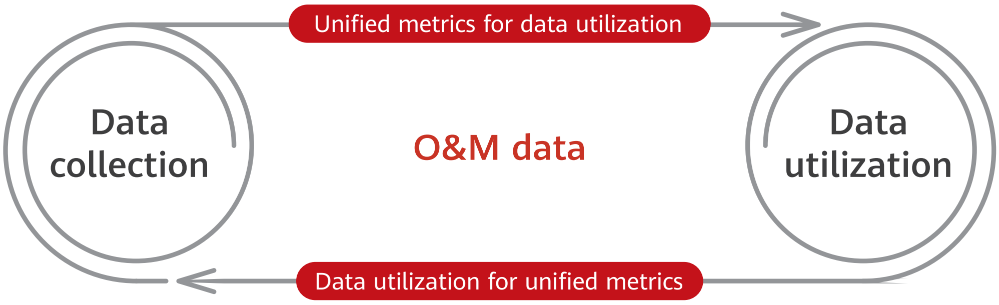

Unified metrics for data utilization
Data collection
O&M data
Data utilization
Data utilization for unified metrics

Figure 1 A data-driven virtuous cycle

IT O&M data forms the basis for data-driven O&M. These twin drivers play a crucial role in the virtuous cycle of data-driven O&M. (Figure 2)

- The first driver is objective-focused and value-driven. It directs IT O&M practices, measures O&M value with metrics, verifies the value of O&M data, and establishes a system for ongoing improvement in IT O&M.
- The second driver emphasizes construction. It encourages IT O&M construction and drives evolution of processes, platforms, scenarios, and O&M models.

The overall framework of the data-driven methodology centers around three value-oriented objectives: enhancing governance, ensuring availability, and improving efficiency. To ensure availability and efficient operations management, various methods are employed, including pre-incident prevention, fast recovery, post-incident review, and specialized approaches for new business branches. To guide the development and improvement of O&M organizations, processes, and tool platforms, a data-driven policy was designed. This policy emphasizes both O&M data governance and the service measurement system, following the principle of beginning with the end in mind.

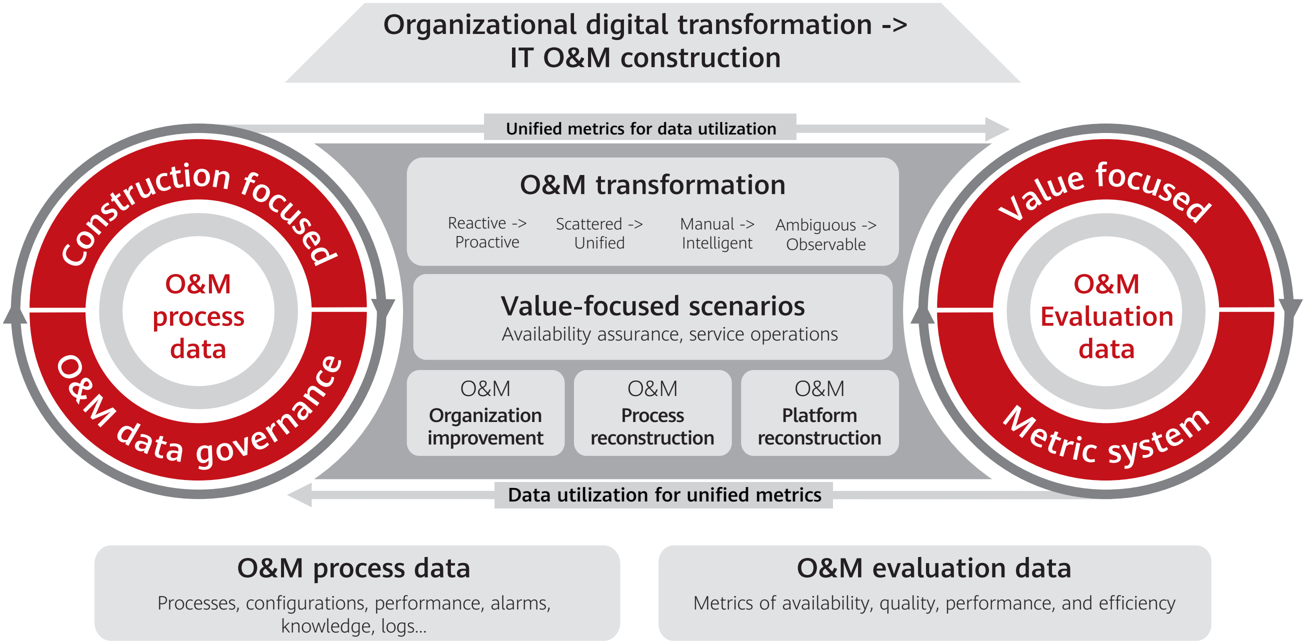

Organizational digital transformation -> IT O&M construction
Unified metrics for data utilization
Construction focused
O&M process data
O&M data governance
O&M transformation
Reactive -> Proactive
Scattered -> Unified
Manual -> Intelligent
Ambiguous -> Observable
Value-focused scenarios
Availability assurance, service operations
O&M Organization improvement
O&M Process reconstruction
O&M Platform reconstruction
Value focused
O&M Evaluation data
Metric system
Data utilization for unified metrics
O&M process data
Processes, configurations, performance, alarms, knowledge, logs...
O&M evaluation data
Metrics of availability, quality, performance, and efficiency

Figure 2 Twin drivers for data-driven O&M

The First Driver: Continuous Object-focused Optimization

We streamlined E2E metrics covering services, scenarios, processes, and resources with a focus on IT service value. We also established a data-driven system and developed capabilities to continuously improve IT O&M. (Figure 3)

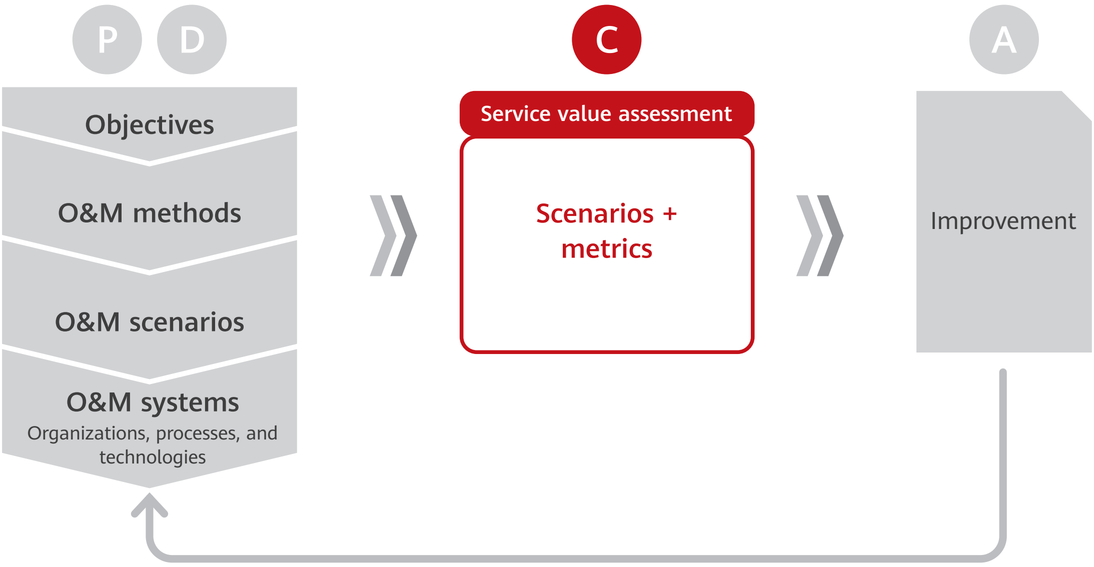


Figure 3 First driver: Continuous object-focused optimization

The first driver focuses on IT service value. To assess IT service value in terms of quality, efficiency, and benefits, related data must be collected and analyzed. The point of measuring IT service value is that it enables organizations to comprehend and evaluate the benefits and value of their IT services. We often analyze IT service metrics and other related data to gain insights into service availability, MTTR, and customer satisfaction. This analysis helps us comprehend the strengths and weaknesses of IT services, enabling us to make informed decisions and drive improvements. Data-driven O&M emphasizes data-based decision-making and practices. Data analysis provides us with objective and quantified information, preventing adverse outcomes that may arise from relying solely on subjective opinions or intuitions. For example, when we need to make a decision related to fault recovery, we can analyze historical data to understand the effects and costs of various solutions and make a more informed choice. Analyzing relevant data is essential for assessing the value of IT services. This analysis, in turn, guides us in implementing data-driven O&M. Integrating IT service value measurement with data-driven O&M enables organizations to efficiently manage and optimize their IT services, enhancing quality, cost-effectiveness, and minimizing risks.

After reviewing research conducted both domestically and internationally, we have developed a framework for establishing a system to measure O&M value. This framework encompasses five dimensions of value, covers all aspects of management throughout the entire lifecycle, provides methods for creating a metric system, and the methods of value assessment.

Figure 4 depicts the framework for establishing a system to measure O&M value.

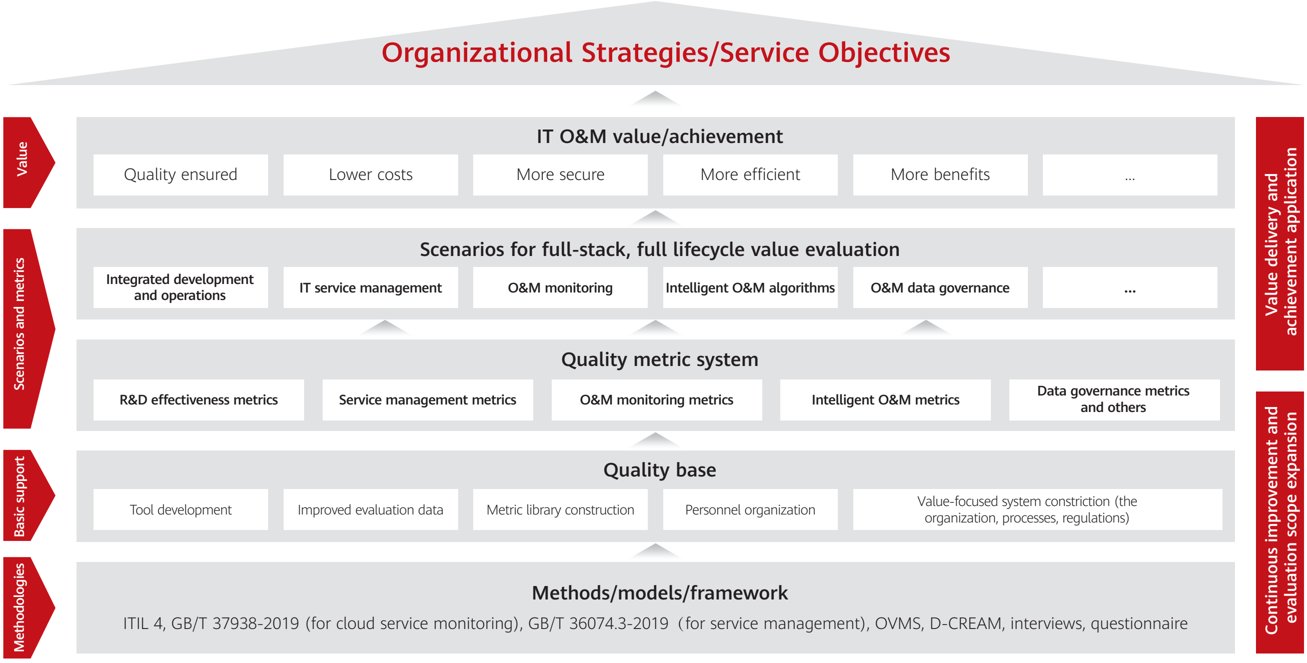

Organizational Strategies/Service Objectives
Value
IT O&M value/achievement
Quality ensured
Lower costs
More secure
More efficient
More benefits
...
Scenarios and metrics
Scenarios for full-stack, full lifecycle value evaluation
Integrated development and operations
IT service management
O&M monitoring
Intelligent O&M algorithms
O&M data governance
...
Quality metric system
R&D effectiveness metrics
Service management metrics
O&M monitoring metrics
Intelligent O&M metrics
Data governance metrics and others
Basic support
Quality base
Tool development
Improved evaluation data
Metric library construction
Personnel organization
Value-focused system constriction (the organization, processes, regulations)
Methodologies
Methods/models/framework
ITIL 4, GB/T 37938-2019 (for cloud service monitoring), GB/T 36074.3-2019 (for service management), OVMS, D-CREAM, interviews, questionnaire
Value delivery and achievement application
Continuous improvement and evaluation scope expansion

Figure 4 Framework of Operations Value Measurement System

This framework aims to help organizations develop business strategies and create business value. It is developed incrementally, incorporating established theories and models from both national and international sources, while also considering the specific O&M requirements. To learn more about the O&M value measurement system, refer to the article "Approaches to O&M Value Measurement" in Huawei Cloud Deterministic SRE Issue 2.

The Second Driver for Five O&M Transformations

Integration of IT and Services Leading to O&M Transformation

Data transforms IT O&M from reactive to proactive, from scattered to unified, from manual to intelligent, and from ambiguous to observable. IT transformations:

From Reactive to Proactive

Proactive O&M represents a move away from the traditional reactive style. Instead of constantly putting out fires, we focus on a more preventive approach. The main goals of proactive O&M are to prevent major faults and reduce the occurrences of faults. Proactive measures involve constantly improving system availability, managing emergencies based on risk levels, and predicting faults.

Logic: Proactive O&M depends on high-quality data. Let's look at an example of how you can improve system availability. Enhancing system availability involves the following steps:

» Identifying system vulnerabilities
» Designing a solution
» Implementing the solution
» Verifying system availability

Good data is crucial for the first and fourth steps.

To identify system vulnerabilities, we need related data to analyze system architecture, deployment, node configurations, performance, and workload.

To verify system availability, we require related monitoring data to assess whether systems are performing as expected

From Scattered to Unified

By consolidating data, we can create a scenario-based integrated O&M system. This allows different teams to work closely together as a comprehensive team. Integrated O&M includes integrated monitoring, development and operations, and supervision and control.

By integrating data, O&M scenarios, processes, and tool platforms can be reconstructed, leading to a transformation of how IT O&M organizations operate. These organizations can move away from traditional functional divisions and implement service objective-oriented cooperation. Cross-functional collaboration allows organizations to prioritize the value of IT O&M.

Data convergence also enables O&M teams to prioritize value-oriented converged services instead of solely focusing on their individual responsibilities. This involves sharing work interfaces and O&M data across teams to transform the way O&M teams communicate with each other.

From Manual to Intelligent

The main difference between intelligent O&M and a traditional manual approach is automation and intelligence. The goal is to achieve better quality, efficiency, and lower costs. Related measures involve predicting faults and automating fault recovery, release, and deployment.

Logic: Intelligent O&M encompasses several core elements. There are organizations, processes, resources, technologies, data, algorithms, and knowledge. Of these, data is the foundation of the other six. Intelligent O&M cannot be implemented without data

From Ambiguous to Observable

Observable IT O&M goes beyond simply visualizing the states and performance of O&M objects through monitoring. It also involves visualizing the impacts of different O&M scenarios, process efficiency, and organizational performance in other areas of O&M management.

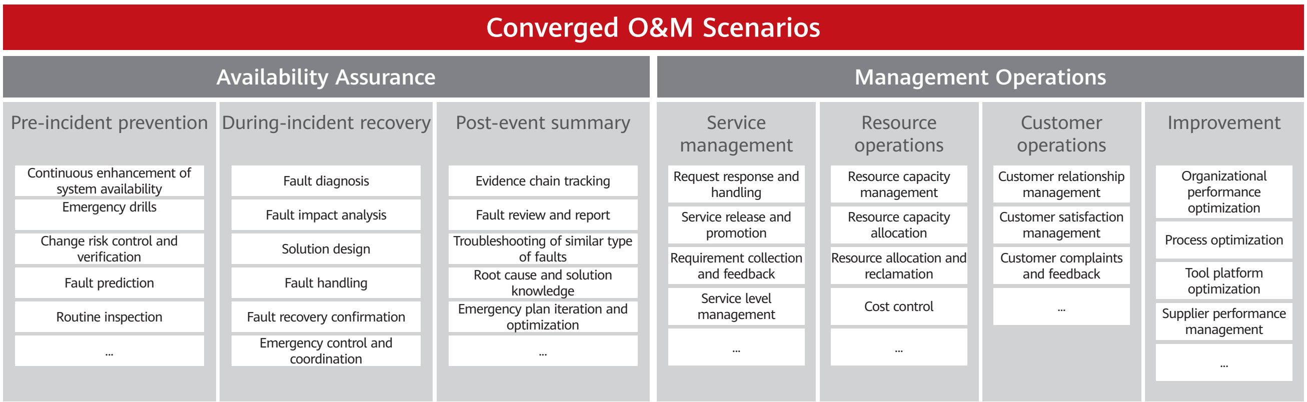
Logic: The core of observable O&M is to present observers with visualized data depicting the status of target objects. The data is collected from resources in different domains, including both raw and processed data.

Enhanced, Innovative Scenarios for Maximized Value of IT O&M

Data-driven O&M not only improves traditional comparatively simple scenarios but also introduces innovative, converged scenarios. It drives more value and address complex O&M problems. This data-driven O&M methodology sorts O&M into two broad categories: availability assurance and management operations, as depicted in Figure 5.

Figure 6 describes IT team responsibilities in both separated and converged scenarios.

Separated scenarios: By integrating data from different domains, independent scenarios can benefit from increased data availability, leading to greater insights into services. For example in basic monitoring scenarios (except for unified alarm



Converged O&M Scenarios						
Availability Assurance			Management Operations			
Pre-incident prevention	During-incident recovery	Post-event summary	Service management	Resource operations	Customer operations	Improvement
Continuous enhancement of system availability	Fault diagnosis	Evidence chain tracking	Request response and handling	Resource capacity management	Customer relationship management	Organizational performance optimization
Emergency drills	Fault impact analysis	Fault review and report	Service release and promotion	Resource capacity allocation	Customer satisfaction management	Process optimization
Change risk control and verification	Solution design	Troubleshooting of similar type of faults	Requirement collection and feedback	Resource allocation and reclamation	Customer complaints and feedback	Tool platform optimization
Fault prediction	Fault handling	Root cause and solution knowledge	Service level management	Cost control	...	Supplier performance management
Routine inspection	Fault recovery confirmation	Emergency plan iteration and optimization	...	...		...
...	Emergency control and coordination	...				

Figure 5 O&M scenarios

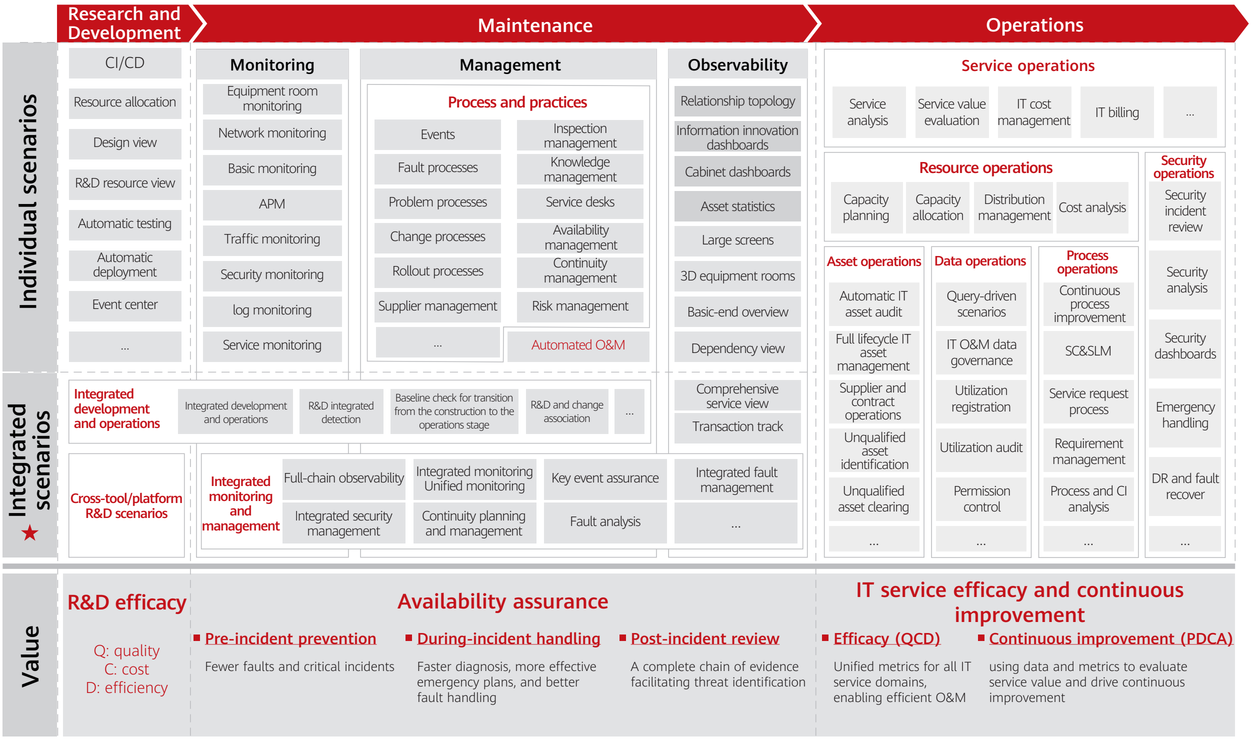


Figure 6 IT O&M scenarios and value (from the IT team responsibility perspective)

scenarios), isolated monitoring tools may fail to send alarm notifications due to a lack of channels to dynamically obtain contact information (such as mobile numbers, email addresses, and IM numbers) of recipients based on association information (such as ownership, maintainable personnel, and personnel in charge) of the monitored objects. When there are changes in an organizational structure, the contact information becomes invalid, and the maintenance of alarm notification settings is often overlooked However, with data convergence, even without unified monitoring services, each monitoring tool can still obtain associated information of monitored objects to enable dynamic setting of alarm notification rules.

Converged scenarios: Integrated data forms the foundation of converged O&M scenarios. Converged scenarios often involve multiple platforms and domains. These scenarios are commonly found in end-to-end observability, integrated monitoring, unified alarms, and O&M support for ensuring key events. To achieve upper-layer service convergence, data from different domains, tools, and teams are required. For example, in integrated monitoring scenarios, data, such as changes and logs, from various layers (network, device, middleware, database, and transaction) are consolidated into a single monitoring platform. This platform offers a unified interface for cross-functional teams. Data standardization, convergence, and application form the foundation of a unified monitoring system.

Scenario-based IT O&M Process Reconstruction

Process reconstruction does not necessarily require tearing down the existing processes. Instead, you can take a scenario-based viewpoint and focus on extracting more value when reviewing, reshaping, and converging the processes that are already established.

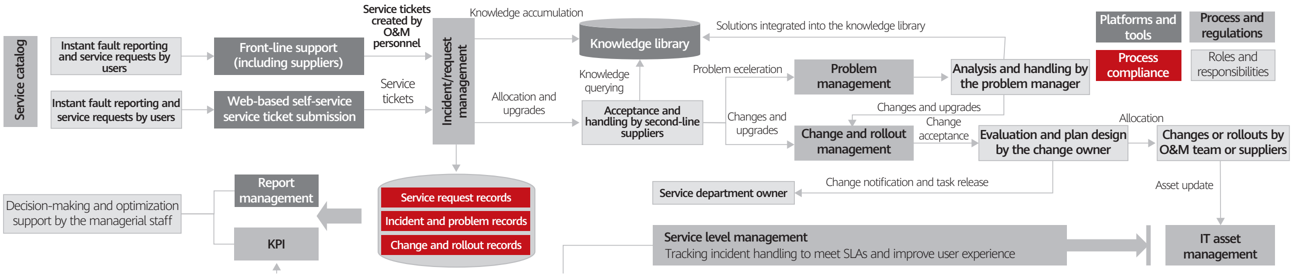

Service catalog
Instant fault reporting and service requests by users
Instant fault reporting and service requests by users
Front-line support (including suppliers)
Web-based self-service service ticket submission
Service tickets created by O&M personnel
Service tickets
Incident/request management
Knowledge accumulation
Knowledge library
Knowledge querying
Allocation and upgrades
Acceptance and handling by second-line suppliers
Solutions integrated into the knowledge library
Problem eceleration
Problem management
Analysis and handling by the problem manager
Changes and upgrades
Changes and upgrades
Change acceptance
Change and rollout management
Evaluation and plan design by the change owner
Allocation
Changes or rollouts by O&M team or suppliers
Platforms and tools
Process and regulations
Process compliance
Roles and responsibilities
Change notification and task release
Service department owner
Asset update
Decision-making and optimization support by the managerial staff
Report management
KPI
Service request records
Incident and problem records
Change and rollout records
Service level management
Tracking incident handling to meet SLAs and improve user experience
IT asset management

Figure 7 Traditional O&M process example

With the data-driven O&M methodology, we can convert the traditional network-like O&M processes into scenario-based ones. Data convergence is the foundation and driving force behind the transformation, which is one of the core aspects of ITIL 4. The network-like traditional O&M processes are depicted in Figure 7. These processes lack a global perspective. They miss the big picture value and scenario-based viewpoints.

A scenario-based practical model combines multiple processes, personnel, organizations with a unified value-centered goal. It also incorporates a unified value chain that relies on data flows.

Here is an example of how the process of ensuring availability can be reconstructed. Process reconstruction does not necessarily require tearing down the existing processes. Instead, you can take a scenario-based viewpoint and focus on extracting more value when reviewing, reshaping, and converging the processes that are already established.

Objectives of availability assurance include:

» Pre-incident: minimized fault rate and zero major faults

» During-incident: improved fault recovery efficiency

» Post-incident: fewer repeated faults, experience and knowledge integrated into the system

In traditional processes, these objectives are only part of the fault handling process. However, in scenario-based processes, these objectives are integrated into a blueprint, as depicted in Figure 8.

This blueprint involves evolution of fault management, continuity management, change management, emergency management, and monitoring and alarms processes.

Cross-scenario Processes The fault process depicted in Figure 7 can encompass both during- and post-incident scenarios. In a cross-scenario fault process, additional elements can be incorporated. You can add more than just flowcharts and roles. These elements may include stakeholder concerns, desired outcomes, information connections, and information to be shared across organizations.

A Single Scenario Containing Multiple Processes A pre-incident situation can involve various processes, including change management, continuity management, SRE processes, and processes that enable the transition from the construction to the operations stage. However, if we stick to the traditional approach of focusing on individual processes, we may miss out on important areas that should be included in the same scenario. This prevents us from achieving our objectives.

Process Association Extends Beyond Data Sharing In a traditional process-centered approach, a change process only interacts with other processes by sharing data. This can result in overlooking important factors such as the reasons and timing for process interactions, the necessary data to be transferred, and the nature of the interaction process. For instance, when making changes to the architecture of an information system, it is important to synchronize the existing risk library, emergency plans, and planned routine assurance tasks. This involves assigning someone to handle the synchronization work and using effective methods to confirm the results. These considerations are taken into account in scenario-oriented process design, but often overlooked in process-centric design.

IT Organizational Architecture Optimization

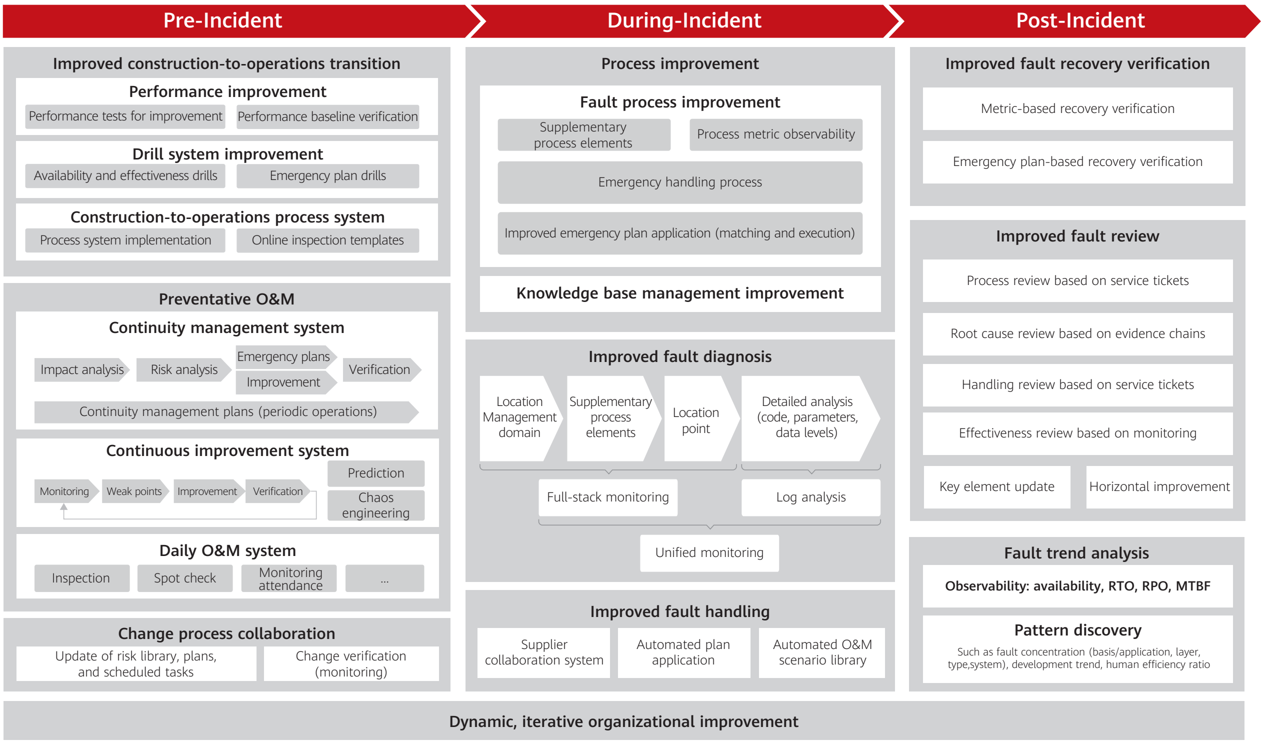

Pre-Incident
During-Incident
Post-Incident
Improved construction-to-operations transition
Performance improvement
Performance tests for improvement
Performance baseline verification
Drill system improvement
Availability and effectiveness drills
Emergency plan drills
Construction-to-operations process system
Process system implementation
Online inspection templates
Preventative O&M
Continuity management system
Impact analysis
Risk analysis
Emergency plans
Improvement
Verification
Continuity management plans (periodic operations)
Continuous improvement system
Monitoring
Weak points
Improvement
Verification
Prediction
Chaos engineering
Daily O&M system
Inspection
Spot check
Monitoring attendance
...
Change process collaboration
Update of risk library, plans, and scheduled tasks
Change verification (monitoring)
Process improvement
Fault process improvement
Supplementary process elements
Process metric observability
Emergency handling process
Improved emergency plan application (matching and execution)
Knowledge base management improvement
Improved fault diagnosis
Location Management domain
Supplementary process elements
Location point
Detailed analysis (code, parameters, data levels)
Full-stack monitoring
Log analysis
Unified monitoring
Improved fault handling
Supplier collaboration system
Automated plan application
Automated O&M scenario library
Improved fault recovery verification
Metric-based recovery verification
Emergency plan-based recovery verification
Improved fault review
Process review based on service tickets
Root cause review based on evidence chains
Handling review based on service tickets
Effectiveness review based on monitoring
Key element update
Horizontal improvement
Fault trend analysis
Observability: availability, RTO, RPO, MTBF
Pattern discovery
Such as fault concentration (basis/application, layer, type,system), development trend, human efficiency ratio
Dynamic, iterative organizational improvement

Figure 8 Scenario-based availability assurance blueprint

Data-driven O&M, along with effective data governance and intelligent O&M, transform collaboration and management practices. This also allows us to optimize the IT organizational structure.

First, to implement data-driven O&M, personnel are needed to perform various tasks, such as data governance, metric system design, and more. The entire IT team will gradually enhance their comprehension and awareness of data, while simultaneously broadening their capabilities.

Second, data-driven O&M redefines team responsibilities, boundaries, and collaboration, resulting in better organized human resources.

Data-driven O&M drives technological transformation, leading to enhanced efficiency and quality in O&M processes. This transformation also necessitates changes in job responsibilities and labor allocation for improved efficiency. Data plays a crucial role too when in evaluating and implementing these changes.

Last be not least, as data-driven O&M continue to evolve, more versatile personnel will be cultivated to drive the transformation of the IT O&M organizational structure.

Data Convergence Helps Reshape IT O&M Platforms

IT O&M data governance achieves centralization and convergence of data, which is essential for converged IT O&M scenarios. This drives the continued convergence and transformation of IT O&M platforms. (Figure 9)

Scattered software tools are also integrated into unified O&M platforms. This unification is reflected in a number of different ways. A unified O&M portal streamlines user management, login and tool integration. Unified data collection and control enable the collection and management of data from multiple O&M tools. A unified data management platform centrally stores and analyzes O&M data, facilitating the establishment and maintenance of metric systems, data quality management, and addressing security issues. Unified incident management makes it possible to collect incident notifications from various monitoring tools, minimize incidents, and standardize incident management. Monitoring management is also unified, providing data analysis, association analysis, and efficient decision-making through diverse dashboards and reports. Unified O&M service management helps establish standardized, efficient process and service systems to deliver high-quality and efficient IT services.

A unified O&M portal centralizes organization, user, and permissions management. It allows for single login access to multiple systems and provides capabilities for data collection, processing, analysis, and display. Additionally, a unified identity management center is provided to facilitate the management of users, permissions, authentication, and audit. It enables full lifecycle management of organizations and users.

Unified data collection and control requires related abilities to centrally collect raw data, performance metrics, incident information, and service ticket details from O&M tools or systems using different protocols or methods.

Unified data management involves filtering, cleaning, processing, storing, and modeling data, and releasing APIs. It allows for the unified management of data assets to satisfy the needs of agile development, intelligent analysis, data lineage, data map, data quality, and other data services. This also helps with data development for complex services and contributes to computing and storage engine services.

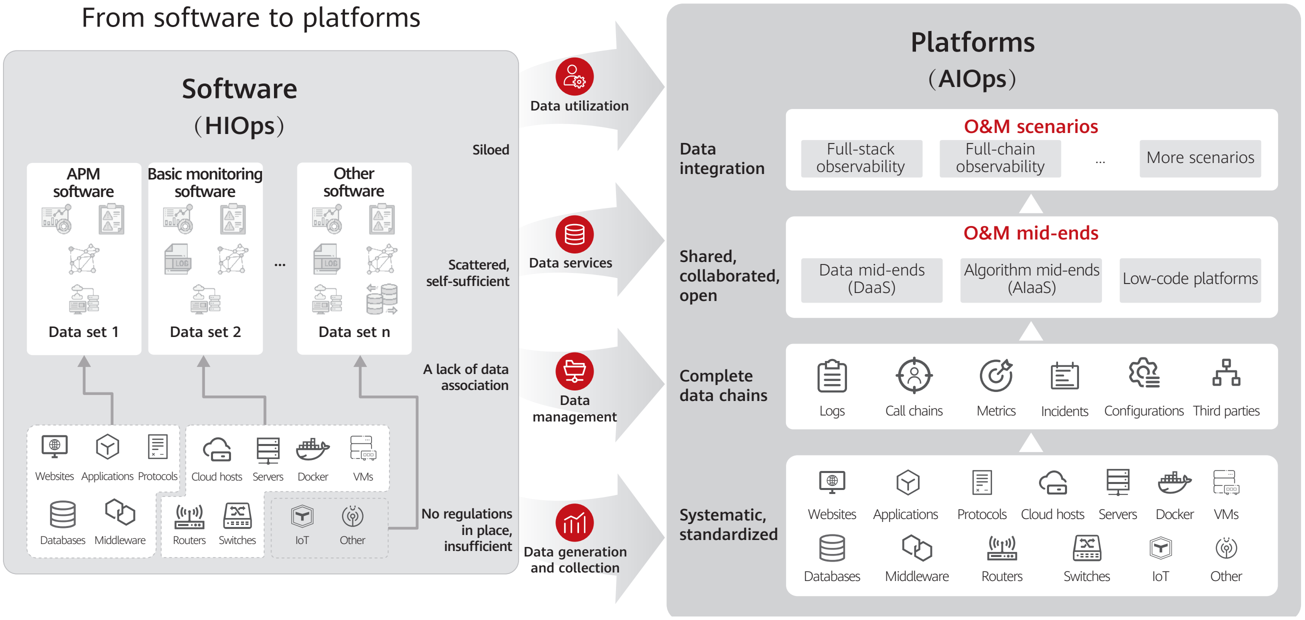


Figure 9 A transformation from isolated O&M to integrated O&M

Unified incident management enables centralized access and handling of alarms and metrics from various monitoring systems. This involves incident filtering, notification, response, handling, rating, tracing, and multi-dimensional analysis. Moreover, various algorithms are used to implement intelligent O&M based specific scenarios. These algorithms help minimize incidents, detect exceptions, and analyze causes, ultimately enabling full-lifecycle incident management and control.

Unified monitoring and management use big data analysis and AI capabilities to monitor network devices, servers, application systems, virtualization, storage, power and environment, and security devices. This is implemented through collecting service logs, metrics, and service chains to construct the service topology. Then association analysis is conducted based on metrics, logs, alarms, attributes, and changes. Various algorithms are also utilized to monitor and analyze business systems, applications, services, and infrastructure in a unified manner. This approach helps efficiently identify and locate faults.

Unified visualization management offers a range of tools, such as dashboards and reports, to present system status, performance metrics, and exceptions. This enables O&M personnel better monitor and manage systems, quickly detect and solve problems, and enhance system availability and stability.

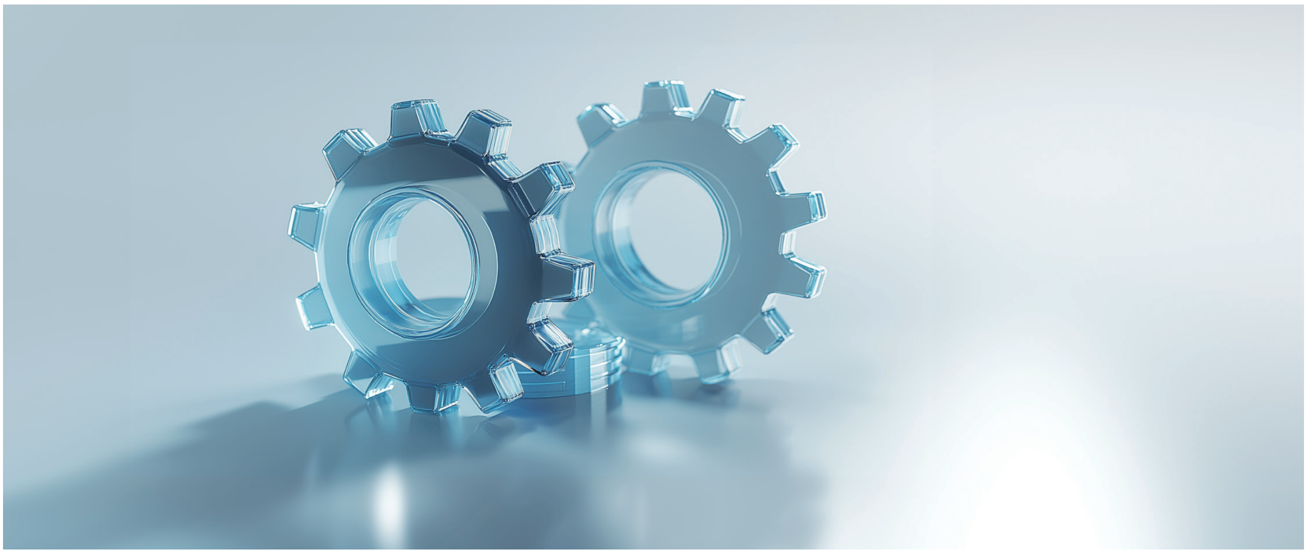
Unified O&M service management delivers high-quality and efficient IT services to organizations. It integrates all IT resources, including hardware, software, network, and security resources to provide IT services in a standardized, process-based manner.

The Second Driver: O&M Data Governance

O&M data governance is essential for digital O&M. It should incorporate established theories and methods of data governance while developing a framework specific to the unique characteristics of O&M data. Additionally, a robust system must be established to ensure the availability of high-quality and comprehensive O&M data for scenario-based digital O&M implementation.

Data governance is a challenging and resource-intensive engineering task. O&M data governance should not revolve around governance. Instead, it needs to prioritize O&M value, such as managing IT risks, enhancing delivery speed, improving customer experience, and enhancing the quality of IT services. We place value creation at the core of our data governance goal: creating more precise and easier-to-use O&M data assets. There are three key aspects of this goal. First, precise data forms the foundation of intelligent O&M, as imprecise data can limit the availability of intelligent O&M scenarios. Second, easy-to-use data promotes the application of intelligent O&M, which is still a relatively new operational model. Data application will in turn enhance data accuracy. Third, O&M data comes in various types, but to fully utilize its potential, it must be elevated to the level of an information asset.

To convert O&M data into information assets, continuous improvement of O&M data governance is required. This improvement should be based on effective governance methods, streamlined governance processes, and advanced technical platforms. To develop data governance methods, we need to focus on master data management, which is represented by O&M metric systems, and generalized meta data management, which is represented by configuration management database (CMDB). Additionally, key governance tasks should be designed based on data standards, quality management, and security management. While implementing data governance, we need to incorporate PDCA, IT governance, and lean innovation throughout the three key cycles: strategy design, construction, and operations. When developing tools for data governance, it is beneficial to take advantages from the existing tools. These tools include O&M data platforms, metric systems, CMDB, monitoring tools, and data portals.



A Two-Way Journey Between Foundation Models and Intelligent O&M

Background

This article describes how various foundation models help summarize faults and find out root causes for systems. This involves evaluating system status, backtracking, and analyzing root causes. The ultimate goal is to improve fault detection and O&M efficiency.

Chen Pengfei

Professor and Doctoral Advisor at
Sun Yat-sen University

Technologies are constantly evolving. Cloud native technologies and foundation models are reshaping various fields at an unprecedented rate. As the amount of data to be handled continues to take off like a rocket, and as requirements become more complex, AI technologies have been advancing, and service scenarios have continued to expand. Foundation models, powered by cloud native technologies, have become a driving force for technical innovation and industrial advancement. Given this context, intelligent O&M capabilities are particularly important.

Cloud native technologies are used to train models effectively and enable precise prediction. When implementing intelligent O&M, enterprises can understand and manage these complex systems better.

Intelligent O&M includes automated problem detection, root cause locating, and system recovery. It provides E2E solutions to enhance the stability, performance, and availability of IT systems.

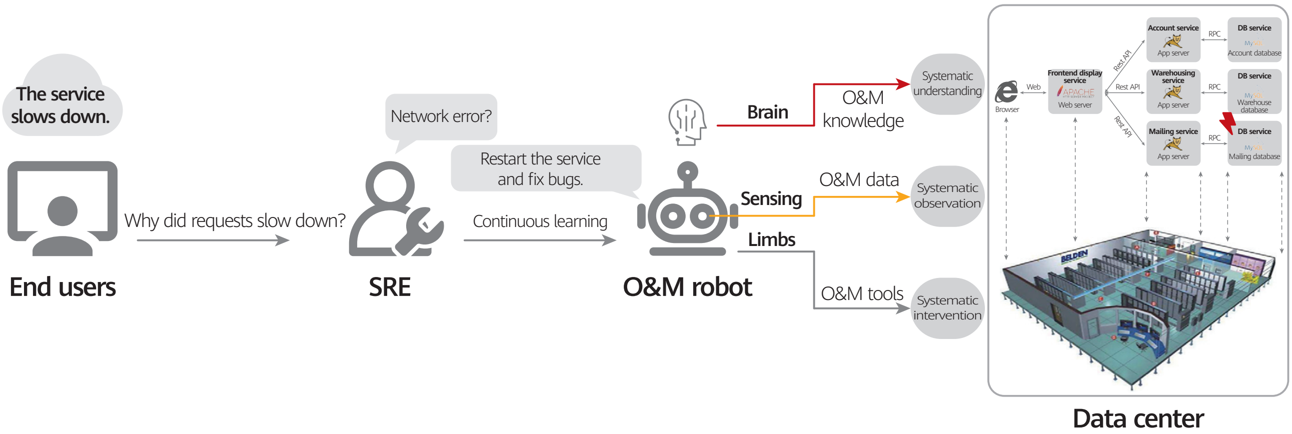
Foundation models use deep learning technologies to automatically identify O&M exceptions. Intelligent O&M can then quickly produce and implement solutions based on the analysis of these models to quickly address problems. Foundation models and intelligent O&M mutually benefit each other, creating a virtuous cycle. Foundation models enhance the intelligence of intelligent O&M, while intelligent O&M contributes practical experience and data to foundation models. Together, they drive the advancement of O&M towards greater automation and intelligence.

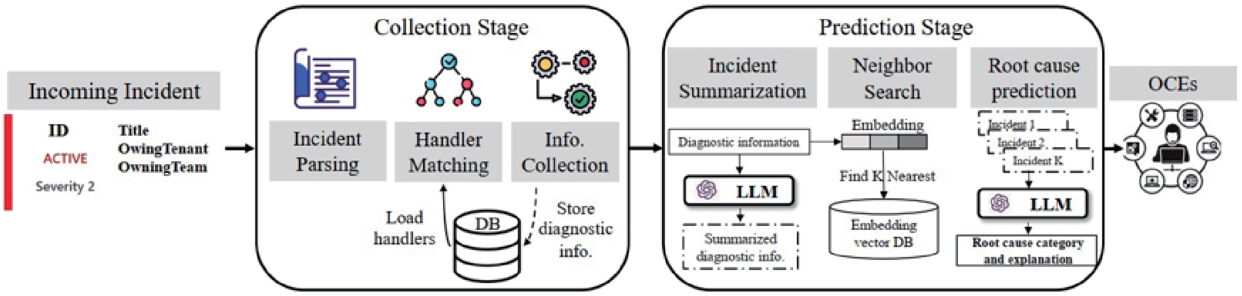
LLM-enabled AIOps

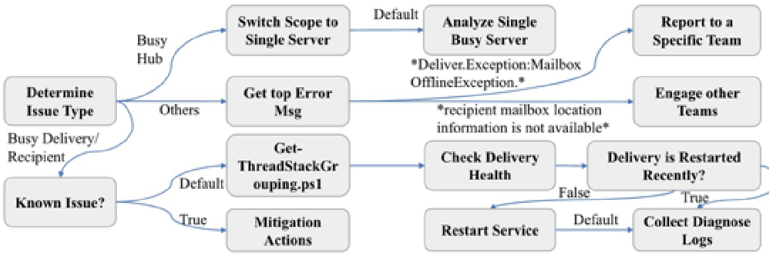
LLMs: Command Centers for AIOps in Harmony with Intelligent Twins

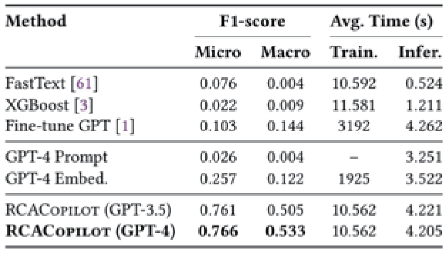
This collaboration involves end users, SREs, O&M robots, and data centers, each playing their respective roles. This synergy not only improves operational efficiency but also implements higher-level automation, resulting in significant economic benefits for enterprises.

- **End users can interact with O&M robots to monitor the system in real time and rectify faults as they occur.** LLMs analyze user behavior and monitor service status to predict and diagnose potential problems. For example, if a user reports a slow service, an LLM can investigate why the request is slow. LLMs report service requirements directly to SREs, who then analyze the requirements and determine the necessary actions on model capabilities.
- **The SREs continually improve the system through in-depth understanding and optimization.** By monitoring and analyzing system performance and availability in real-time, LLMs can detect and resolve faults and eliminate bottlenecks, improving system reliability and stability. This ensures service continuity and sustainable development.
- **O&M robots utilize LLMs as their "brain", and they use various monitoring tools as their "sensing organs" to sense the environment and make decisions.** When a system fault occurs, robots gather information with their "sensory" tools and make decisions based on foundation models. They can also understand O&M knowledge, observe data, and use tools to automate system operations.
- **LLMs analyze and predict data center performance to optimize and improve systems.** By applying deep learning and analysis to large datasets, models can forecast future requirements and trends. This analysis helps optimize systems, enhancing the efficiency and stability of data centers.






Determine Issue Type
Busy Hub
Others
Busy Delivery/ Recipient
Switch Scope to Single Server
Default
Analyze Single Busy Server
Report to a Specific Team
Get top Error Msg
Deliver.Exception:Mailbox OfflineException.
recipient mailbox location information is not available
Engage other Teams
Known Issue?
Default
True
Get-ThreadStackGrouping.ps1
Check Delivery Health
Delivery is Restarted Recently?
False
True
Mitigation Actions
Restart Service
Default
Collect Diagnose Logs



Method	F1-score		Avg. Time (s)	
	Micro	Macro	Train.	Infer.
FastText [61]	0.076	0.004	10.592	0.524
XGBoost [3]	0.022	0.009	11.581	1.211
Fine-tune GPT [1]	0.103	0.144	3192	4.262
GPT-4 Prompt	0.026	0.004	–	3.251
GPT-4 Embed.	0.257	0.122	1925	3.522
RCACopilot (GPT-3.5)	0.761	0.505	10.562	4.221
RCACopilot (GPT-4)	**0.766**	**0.533**	10.562	4.205

Root Cause Analysis of Cloud Faults Using LLMs

How do we leverage LLMs for automatic cloud event analysis?

» **Data collection and integration:** We gather and preprocess various data sources in the cloud environment, including server logs, monitoring metrics, configuration files, application logs, and user operation records. This ensures that the data is effectively readable and understandable by LLMs.

» **Model selection and training:** Choosing the right large language model architecture for sequence and text data is important. The model then needs to be trained using labeled historical cloud event data so that it can learn the relationships between event features and their causes.

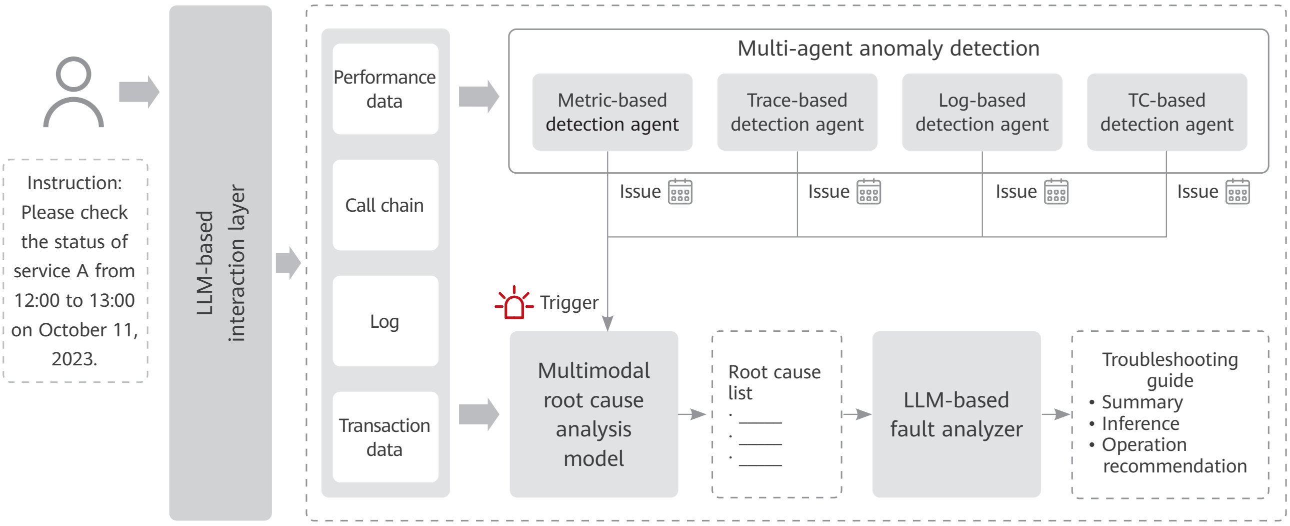
» **Attributional inference:** New cloud event data is fed into the trained model. The model predicts and analyzes possible causes of events based on its learned knowledge and patterns.

» **Results interpretation and verification:** The model's attribution results are interpreted and evaluated to determine their accuracy and reasonableness. The results are then verified and corrected based on domain knowledge and manual expertise.

» **Continuous optimization:** Model parameters are continuously optimized to improve feature engineering based on new data and feedback, with a goal of enhancing the accuracy and reliability of the attribution analysis.

Throughout the entire process, data quality and labeling accuracy are crucial for optimal model performance. Additionally, manual intervention and expert verification can significantly improve the effectiveness and reliability of the attribution analysis.

Multi-Agent Multimodal Data Convergence for Root Cause Analysis


Instruction: Please check the status of service A from 12:00 to 13:00 on October 11, 2023.
LLM-based interaction layer
Performance data
Call chain
Log
Transaction data
Multi-agent anomaly detection
Metric-based detection agent
Trace-based detection agent
Log-based detection agent
TC-based detection agent
Issue
Issue
Issue
Issue
Trigger
Multimodal root cause analysis model
Root cause list
LLM-based fault analyzer
Troubleshooting guide
• Summary
• Inference
• Operation recommendation

The architecture consists of four main components:

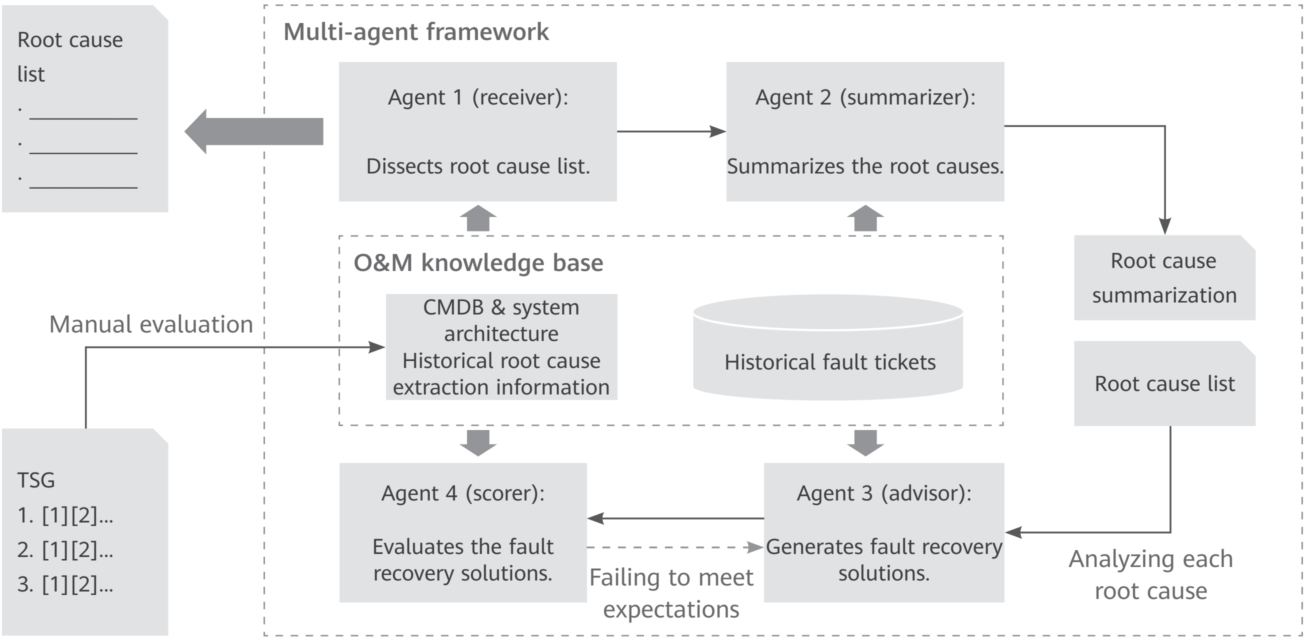
» An LLM-based interaction layer: This component interprets user queries and extracts essential tasks and parameters. It utilizes ChatGLM2 as the foundational model, enhanced with self-consistency, Chain of Thought (CoT), and in-context learning logic to better understand scenarios and provide accurate responses.

» Multi-agent anomaly detection: Due to the involvement of multimodal data sources, achieving high accuracy with a single detection module can be challenging, leading to potential false positives. Hence, a multi-agent detection solution is employed, with dedicated anomaly detection agents designed for trace, log, and metric modal data.

» A root cause analysis model based on multimodal data convergence: This algorithm converts multimodal data, such as call chains, logs, and metrics, into unified event expressions. It employs an unsupervised frequent itemset mining method to identify fault modes and pinpoint fine-grained root causes at the resource and code block levels. Additionally, this method explains faults by comparing mode changes from before and after the occurrence of faults.

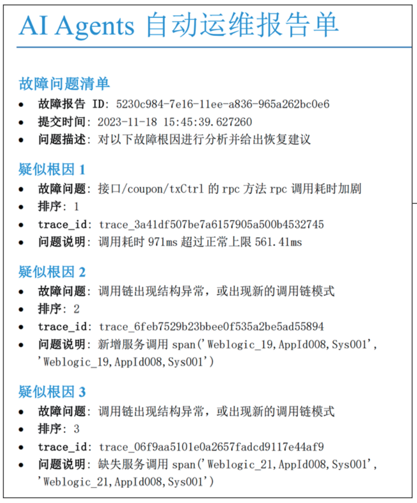
» LLM-based fault analyzers: Multiple LLM agents are used to answer questions, with different LLM sessions functioning in various roles to generate fault report tickets.

These agents identify the root cause among numerous exceptions and produce a fault report with recommended recovery actions.



We simultaneously initiate four LLM sessions, each serving as a different agent with distinct roles: The first agent, acting as the receiver, accepts a root cause list and dissects this list into individual root causes described in multiple natural languages. The second agent, the summarizer, consolidates these root cause descriptions into a comprehensive summary. The third agent, the advisor, analyzes each root cause and generates corresponding fault recovery solutions. The fourth agent, the scorer, evaluates the proposed solutions. If a solution fails to meet expectations, the process iterates for improvement.

In the end, a report is produced, summarizing the root cause and recovery solutions.



AI Agents 自动运维报告单

故障问题清单

- **故障报告 ID**：5230c984-7e16-11ee-a836-965a262bc0e6
- **提交时间**：2023-11-18 15:45:39.627260
- **问题描述**：对以下故障根因进行分析并给出恢复建议

疑似根因 1

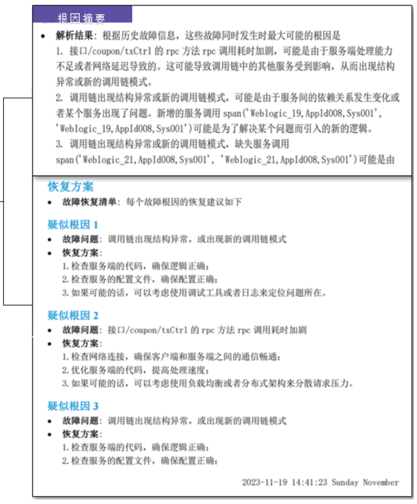
- **故障问题**：接口/coupon/txCtrl 的 rpc 方法 rpc 调用耗时加剧
- **排序**：1
- **trace_id**：trace_3a41df507be7a6157905a500b4532745
- **问题说明**：调用耗时 971ms 超过正常上限 561.41ms

疑似根因 2

- **故障问题**：调用链出现结构异常，或出现新的调用链模式
- **排序**：2
- **trace_id**：trace_6feb7529b23bbee0f535a2be5ad55894
- **问题说明**：新增服务调用 span('Weblogic_19,AppId008,Sys001', 'Weblogic_19,AppId008,Sys001')

疑似根因 3

- **故障问题**：调用链出现结构异常，或出现新的调用链模式
- **排序**：3
- **trace_id**：trace_06f9aa5101e0a2657fadcd9117e44af9
- **问题说明**：缺失服务调用 span('Weblogic_21,AppId008,Sys001', 'Weblogic_21,AppId008,Sys001')



根因摘要

- **解析结果**：根据历史故障信息，这些故障同时发生时最大可能的根因是
  1. 接口/coupon/txCtrl 的 rpc 方法 rpc 调用耗时加剧，可能是由于服务端处理能力不足或者网络延迟导致的。这可能导致调用链中的其他服务受到影响，从而出现结构异常或新的调用链模式。
  2. 调用链出现结构异常或新的调用链模式，可能是由于服务间的依赖关系发生变化或者某个服务出现了问题。新增的服务调用 span('Weblogic_19,AppId008,Sys001', 'Weblogic_19,AppId008,Sys001')可能是为了解决某个问题而引入的新的逻辑。
  3. 调用链出现结构异常或新的调用链模式，缺失服务调用 span('Weblogic_21,AppId008,Sys001', 'Weblogic_21,AppId008,Sys001')可能是由

恢复方案

- **故障恢复清单**：每个故障根因的恢复建议如下

疑似根因 1

- **故障问题**：调用链出现结构异常，或出现新的调用链模式
- **恢复方案**：
  1. 检查服务端的代码，确保逻辑正确；
  2. 检查服务的配置文件，确保配置正确；
  3. 如果可能的话，可以考虑使用调试工具或者日志来定位问题所在。

疑似根因 2

- **故障问题**：接口/coupon/txCtrl 的 rpc 方法 rpc 调用耗时加剧
- **恢复方案**：
  1. 检查网络连接，确保客户端和服务端之间的通信畅通；
  2. 优化服务端的代码，提高处理速度；
  3. 如果可能的话，可以考虑使用负载均衡或者分布式架构来分散请求压力。

疑似根因 3

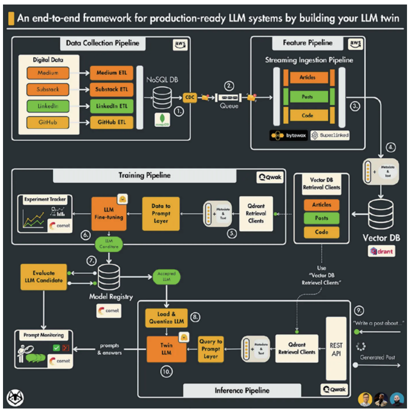
- **故障问题**：调用链出现结构异常，或出现新的调用链模式
- **恢复方案**：
  1. 检查服务端的代码，确保逻辑正确；
  2. 检查服务的配置文件，确保配置正确；

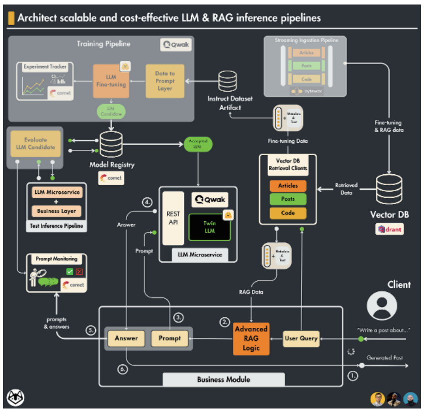
2023-11-19 14:41:23 Sunday November

LLM-enabled AIOps

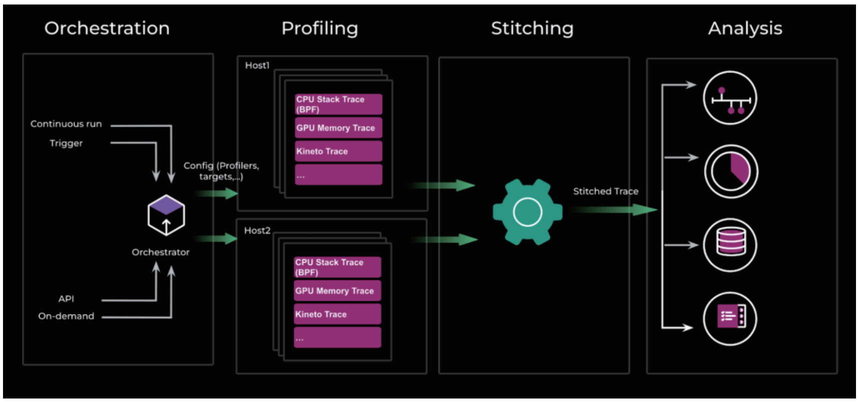
LLM Observability

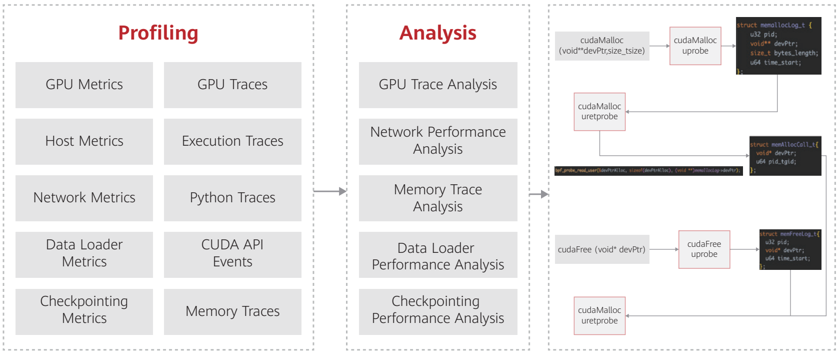
- **LLM observability is crucial due to the complexity of software stacks involved in LLM training and inference.** Effective management and monitoring of these components and modules require robust observability methods. This allows for better understanding, tracking, faster debugging, and problem resolution.
- **A startup has identified five key dimensions for LLM observability: evaluation, call chains, prompt engineering, search and query, and fine-tuning.** Observing LLMs is a multifaceted process that demands comprehensive evaluation, continuous tracking, and prompt verification. Here's a streamlined approach: Evaluation: Assess the model to determine its performance and effectiveness. Tracing: Monitor the model's status to detect and resolve issues promptly. Search and query: Gather additional information and data to enhance understanding and optimization of the model. Fine-tuning: Adjust the model to improve observability and adapt to varying environments and requirements.





- **Meta has proposed a profiling-based multimodal data association method, which links data from different sensors or sources, such as images, voice, and text, for deeper analysis and understanding.** This method has significant practical value and can be applied across various fields, including healthcare, finance, and social media.
- **Meta also introduces eBPF-based cross-layer and cross-node LLM request tracing**. By combining profiling-based multimodal data association with CUDA monitoring using eBPF uprobe, this method provides real-time GPU status and performance metrics. This detailed analysis helps you understand system performance better and identify bottlenecks. It offers substantial support for system optimization. This approach holds great promise for high-performance computing and AI applications.





```
struct memallocLog_t {
    u32 pid;
    void** devPtr;
    size_t bytes_length;
    u64 time_start;
};
```

```
struct memAllocCall_t{
  void* devPtr;
  u64 pid_tgid;
};
```

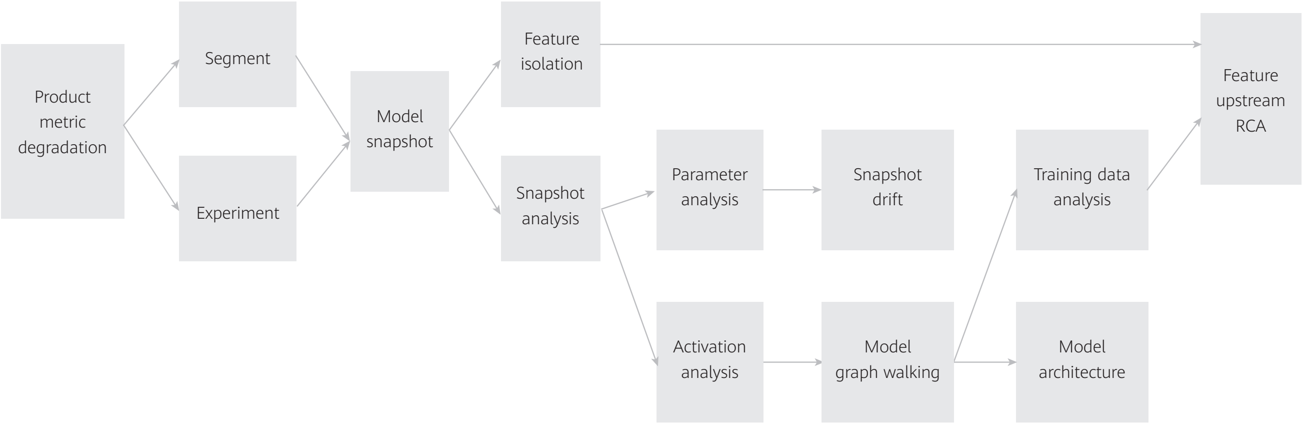
```
bpf_probe_read_user(&devPtrAlloc, sizeof(devPtrAlloc), (void **)memallocLog->devPtr);
```

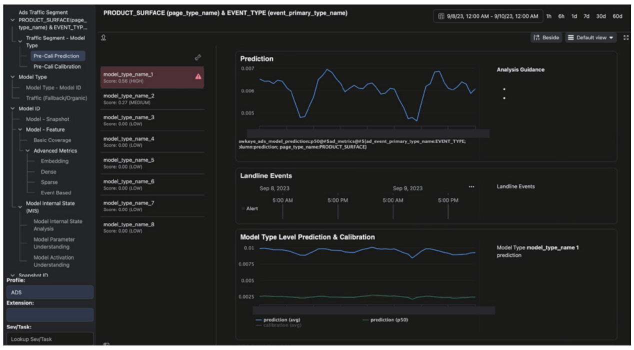
```
struct memFreeLog_t{
  u32 pid;
  void* devPtr;
  u64 time_start;
};
```

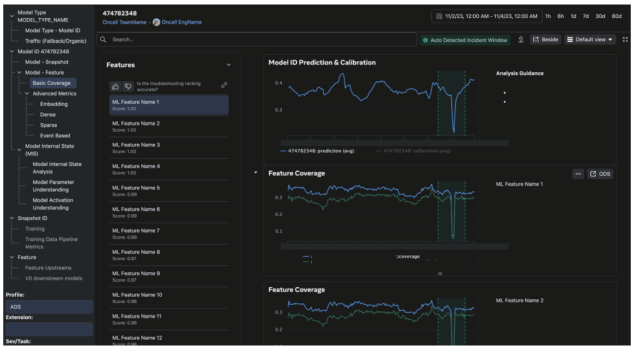
Root Cause Analysis for LLM Faults

Progressive root cause analysis powered by knowledge graphs. This advanced technology leverages the power of knowledge graphs to identify root causes. By using step-by-step reasoning and analysis, it extracts relevant information from the knowledge graph and pinpoints the root cause based on the gathered data. The key advantage of this method is its ability to uncover the underlying causes of complex problems, not just superficial symptoms. Knowledge graphs provide a comprehensive understanding of the issue, including its background, related entities, events, possible causes, and impacts. This comprehensive approach allows for more accurate identification and effective resolution of problems.

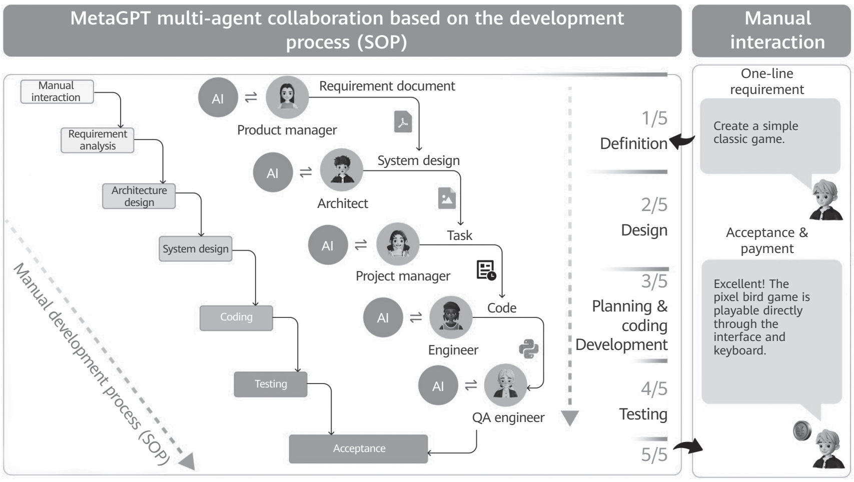
Multi-agent collaboration for AIOps. The MetaGPT framework is a robust multi-agent collaboration framework that enhances the efficiency of building multi-agent systems and offers various collaboration modes. Developers can use metaprogramming within the MetaGPT framework to dynamically generate agent code, facilitating collaboration between different agents.

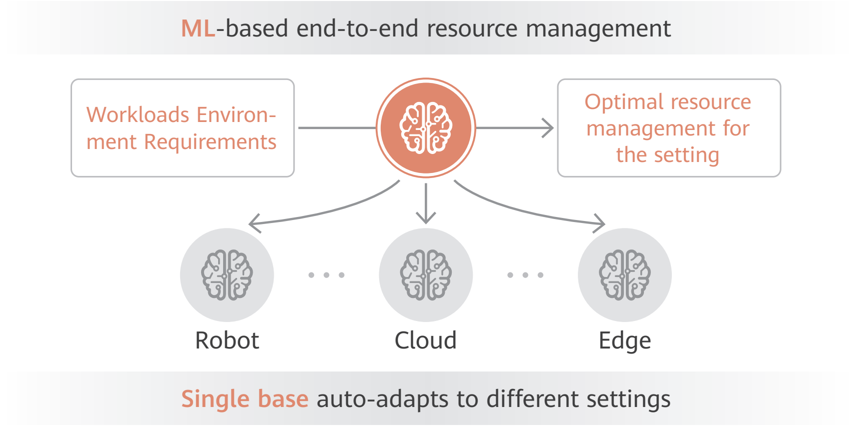



PRODUCT_SURFACE (page_type_name) & EVENT_TYPE (event_primary_type_name)
Prediction
Analysis Guidance
Landline Events
Model Type Level Prediction & Calibration



LLM-driven operating system. The base OS is the cornerstone of a computer system, managing hardware resources and providing essential system functions. With the increasing complexity of modern computer systems, researchers are turning to machine learning technologies for end-to-end resource management. Machine learning can be used to automatically adjust and optimize resource allocation based on workload environment requirements. This approach optimizes resource management for robots, clouds, and edge environments, adapting seamlessly to different settings from a single base. This innovation significantly improves system performance and efficiency.




ML-based end-to-end resource management
Workloads Environment Requirements
Optimal resource management for the setting
Robot
Cloud
Edge
Single base auto-adapts to different settings

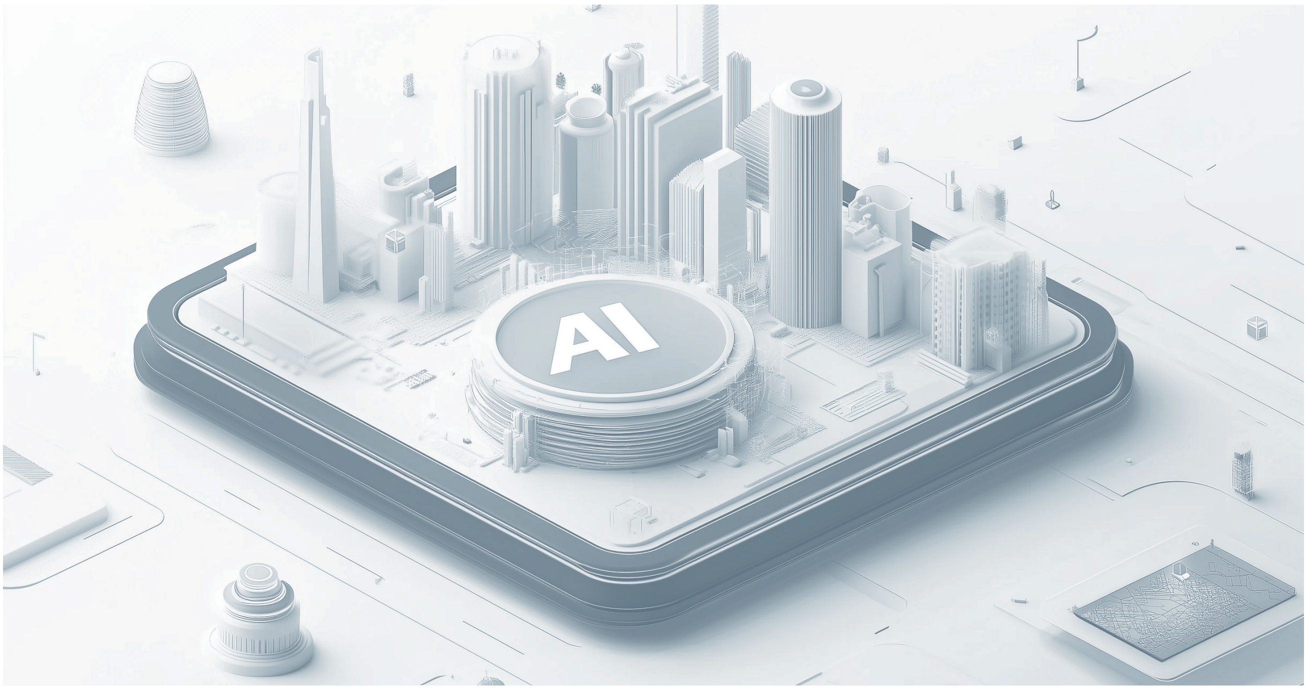
A Two-Way Journey Between LLM and AIOps Based on Deterministic Operations

In the collaboration between LLMs and AIOps, we integrate both large and small models. LLMs handle tasks that require human-like analysis, such as decision-making and problem-solving. SLMs focus on specialized tasks for vertical issues like network fault prediction and performance optimization. This integration leverages the broad knowledge of LLMs while mitigating issues caused by LLM hallucinations.

Applying LLMs creates challenges that require careful consideration and response. To fully utilize their advantages, we must continuously optimize and adjust LLMs during training so they can adapt to changing environments.

The collaboration between LLMs and AIOps is a complex and challenging process. We must address and overcome the problems and challenges brought by LLMs to achieve successful outcomes. By building on deterministic operations practices, we can achieve reliable and consistent results. This approach will enable the effective application of LLMs in AIOps, bringing greater value to enterprises.

Stability Practices for Large-Scale AI Training Clusters



Tong Lin

Huawei Cloud SRE senior expert

Abstract

This document analyzes the stability challenges of AI foundation model training clusters and introduces the comprehensive closed-loop processes and practical experience of Huawei Cloud in AI clusters, ranging from exception detection and diagnosis to self-healing. It describes an effective assurance system for the stability and reliability of AI clusters, accelerating the transition to intelligent systems.

Stability Challenges for Large-Scale AI Training Clusters

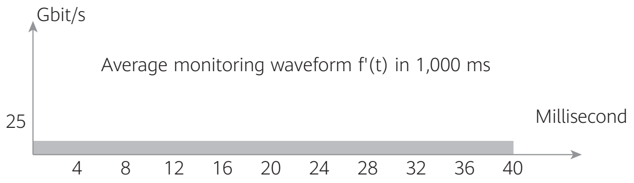
From an industry perspective, the stability of AI foundation models is becoming increasingly important. The communication flow of training tasks is complex and sensitive to network latency and bandwidth. The reliability of AI training clusters faces several challenges:

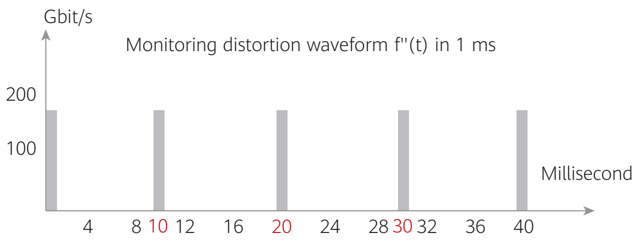
- As the number of model parameters increases, the size of the cluster grows, which make component faults more likely, and damage to the network more likely. The probability of training faults increases exponentially.
- The traffic model of foundation model training tasks is intricate, with long communication links and the involvement of multiple types of infrastructure and cloud services across domains. Fault types can include slow training due to insufficient bandwidth, loss divergence, GPU loss, training task execution failures, suspension, and slowdown. The various causes and types of faults make demarcation and localization difficult.
- After a fault occurs, training recovery is slow, especially, checkpoint loading, which typically takes several hours. The MTBF of a large-scale cluster in the industry is only several hours, but faults can last for 10 hours or more.
- When faults occur, training tasks may be repeatedly rolled back and restarted, resulting in low resource utilization and wasted compute. The average resource utilization of AI training clusters in the industry is only about 30% to 40%.

These faults significantly reduce the efficiency and increase the cost of AI training tasks. So, enhancing the stability and reliability of foundation model clusters will significantly help enterprises address both of these issues.

Challenge 1: Insufficient Monitoring Precision and High Dependency of Training Tasks on Networks

The strong dependency of training tasks on network performance amplifies the impacts of suboptimal optical modules

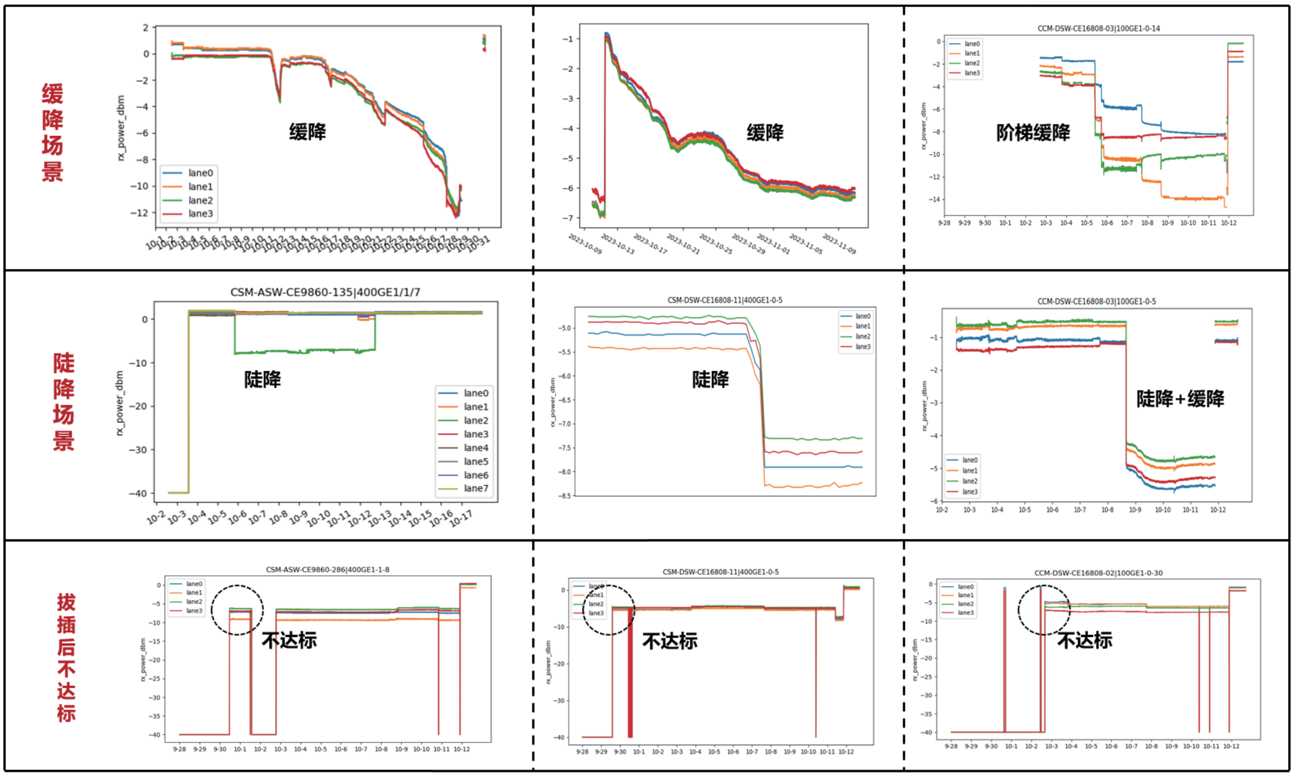


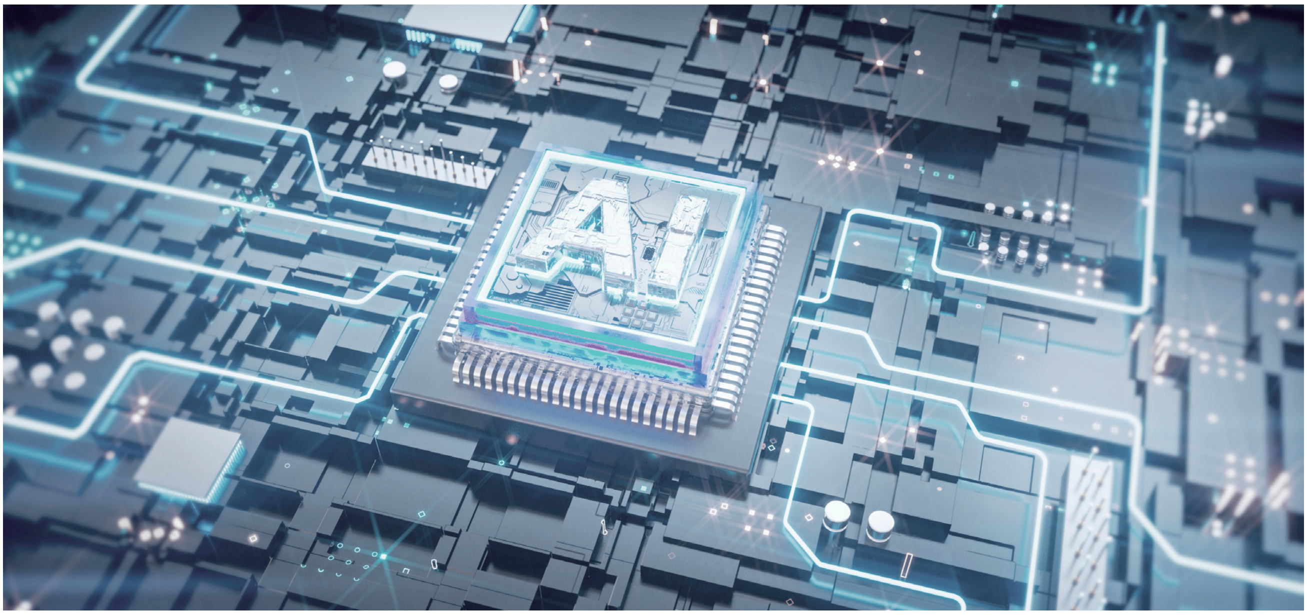


In the observability field of AI clusters, collection precision is insufficient, and network quality and performance detection requirements are high.

» Traditional network device traffic monitoring, mainly using SNMP, is accurate only to the second. Now, foundation model training occurs in two phases: computing and communication. Traffic waveform changes must be collected in milliseconds. Traditional monitoring and collection counters have a software error margin of 10 to 20 ms. Anything beyond this results in distorted and misplaced results. This inadequacy prevents accurate measurement of the actual traffic characteristics of AI cluster training tasks.

» Traditional O&M can only detect optical module faults. As optical modules age, their performance deteriorates, placing them in a suboptimal state. Additionally, fault characteristics vary widely. With increasing AI cluster scale, the impact of suboptimal optical modules on AI training tasks is magnified. Therefore, suboptimal optical module detection and fault prevention are more crucial in AI training clusters than in traditional application clusters.





Challenge 2: Difficult Fault Locating and Demarcation and Slow Fault Recovery

When services are affected, especially when hardware and network components are suboptimal, it is difficult to demarcate and locate AI cluster faults.

- Complex and diverse fault modes: Faults such as slow nodes, slow networks, slow computing, and slow communications result in training service interruptions, frame freezing, and failures. Some faults lack obvious white-box alarms, making quick demarcation difficult.
- Long faulty chains: Large-scale training tasks involve many-to-many GPU communications, encompassing servers, devices, ports, links, and optical modules across multiple network elements (NEs). Association analysis is required for various objects.
- Inefficient diagnosis: Traditional diagnosis relies on analysis of extensive training logs, which takes time. The absence of real-time network traffic topology necessitates analyzing numerous monitoring metrics and data.

Diverse fault modes

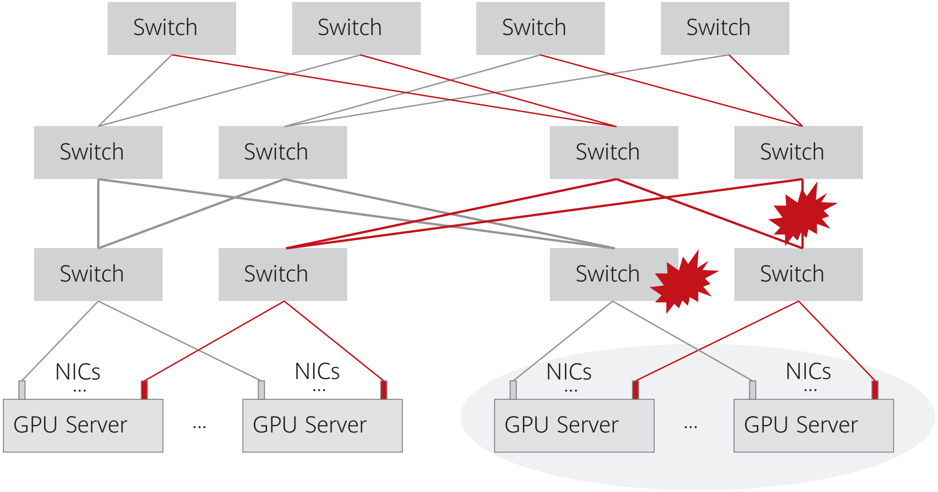
Slow nodes, slow networks, and optical module performance deterioration

Long faulty links

Communication links encompassing servers, devices, ports, and optical modules

Inefficient diagnosis

A large volume of logs and monitoring metrics



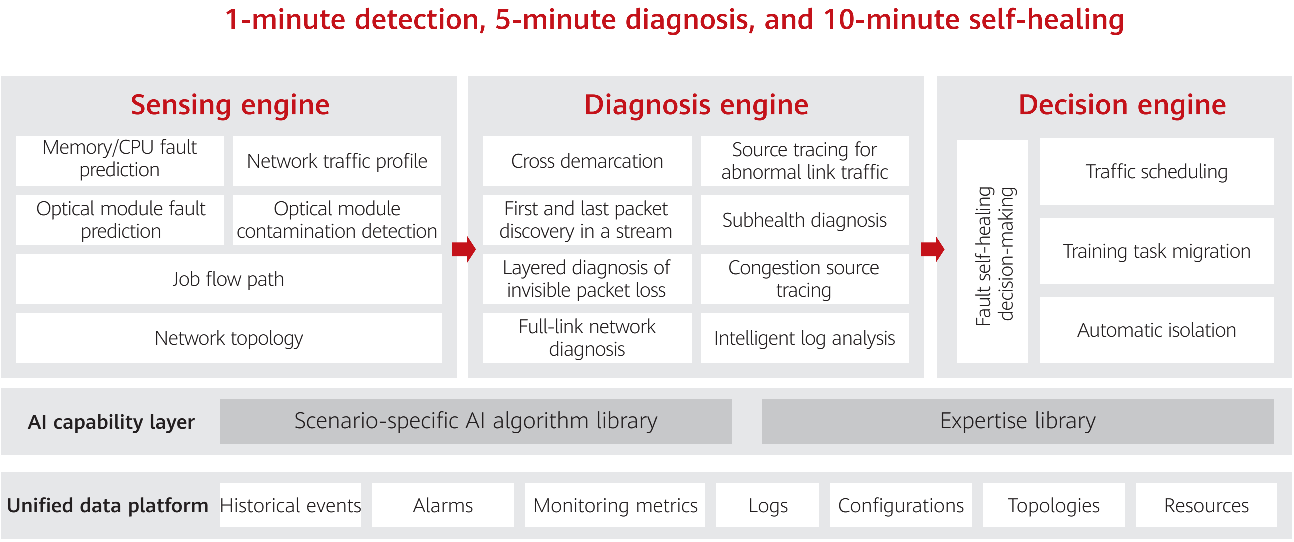
Ensuring Robust AI Cluster Stability

To address these challenges, Huawei Cloud has developed a comprehensive closed-loop system for fault detection, diagnosis, and recovery in AI clusters. This system is built upon a mature internal O&M platform and provides a panoramic view of AI cluster stability.

» Fault detection: Faults are detected and predicted swiftly, well before customers notice. Huawei has created fault prediction algorithms for components like memory, hard disks, and optical modules to preemptively identify most hardware risks and subhealth faults. The detection system continuously and automatically collects network connection topologies and traffic path changes, supplying comprehensive data support for diagnosis.

» Fault diagnosis: There is E2E topological restoration. The service topology is automatically drawn, in real time. Network congestion issues are quickly identified based on real-time changes. Additionally, root cause tracing and recommendations are efficiently performed using indicators, logs, and alarms of associated objects.

» Fault recovery: The decision-making system analyzes and calculates the scope of any impacts (including affected nodes and customers) in real-time, using a graph configuration library. It then employs the least-cost and most efficient policy to quickly restore services based on a preset fault contingency plan model.


1-minute detection, 5-minute diagnosis, and 10-minute self-healing
Sensing engine
Memory/CPU fault prediction
Network traffic profile
Optical module fault prediction
Optical module contamination detection
Job flow path
Network topology
Diagnosis engine
Cross demarcation
Source tracing for abnormal link traffic
First and last packet discovery in a stream
Subhealth diagnosis
Layered diagnosis of invisible packet loss
Congestion source tracing
Full-link network diagnosis
Intelligent log analysis
Decision engine
Fault self-healing decision-making
Traffic scheduling
Training task migration
Automatic isolation
AI capability layer
Scenario-specific AI algorithm library
Expertise library
Unified data platform
Historical events
Alarms
Monitoring metrics
Logs
Configurations
Topologies
Resources

Zero Monitoring Blind Spots with Combined White-Box and Black-Box Observability

Leveraging Huawei Cloud's big data platform, a monitoring system is established to collect, store, compute, and analyze metrics in milliseconds, ensuring comprehensive observability of AI clusters. This system enhances traditional monitoring methods with three key capabilities:

» High-precision metrics: Beyond traditional millisecond-level monitoring of interface rates and packet loss, telemetry collection tracks critical metrics, such as PFC packet sending and receiving, packet loss from queue congestion, and port bandwidth usage. This capability allows rapid detection of network microbursts.

» Real-time network packet loss detection: Network chips come with built-in packet loss detection, which can be used to proactively identify flow-level packet loss exceptions and causes. Distributed probes collect, compress, and aggregate data, which is then reported to servers to form events, which are then, in turn, transmitted to a monitoring platform for alarm reporting.

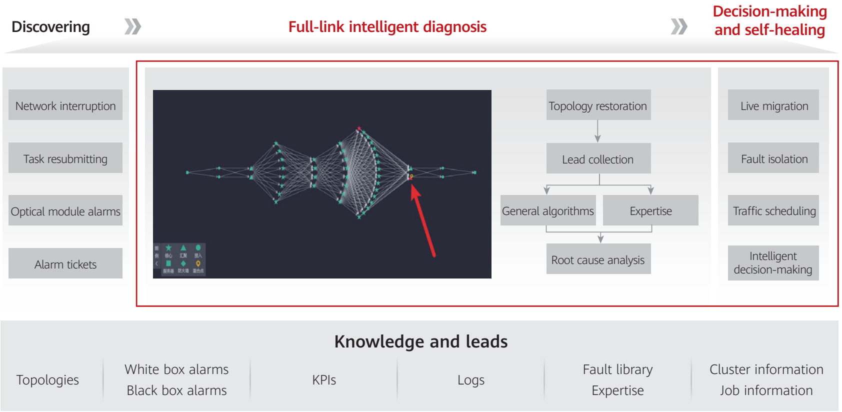
» Black-box monitoring: Unlike traditional pingmesh, which uses standard protocols like ICMP, TCP, and UDP, a link-level full-coverage black-box dialing test system based on the Huawei Cluster Communication Library (HCCL) is implemented. This system detects faults, such as slow networks, within seconds to prevent invisible packet loss in AI clusters.

AI-Driven Optical Module Subhealth Prediction: Proactively Preventing Issues

For optical module fault detection and prediction, spatial, temporal, and multi-dimensional metrics are employed to enhance fault detection capabilities. Historical fault characteristics (vendor, model, and cause), log characteristics, metrics (current, voltage, temperature, and optical power), and the length of time an optical module has been in use, are analyzed comprehensively. This includes analyzing similar metrics on the local and peer ends of the optical module topology.

Using multi-dimensional data, feature engineering, and regression analysis AI algorithms, faulty optical modules and those whose metrics are degrading (including dirty and loose optical modules) are identified. The system scores different fault profiles and makes service life predictions. It generates reports, automatically creates alarms and submits repair tickets based on predefined thresholds. This proactive approach allows for faulty optical modules to be replaced before they affect services.

Rapid Root Cause Diagnosis with Full Path Restoration



For rapid fault demarcation of AI clusters, a self-developed full-link diagnosis system and a comprehensive configuration database quickly calculate the path of any traffic between sources and destination hosts involved in a training task. The path covers all involved NEs. It includes the servers, GPUs, ports, links, boards, and switches.

Based on the metrics, logs, and alarms of objects in the real-time topology, the diagnosis and analysis module uses a knowledge graph to identify root causes of faults and identify faulty units within minutes. This technology, widely used in Huawei Cloud, ensures that fault demarcation takes less than five minutes.

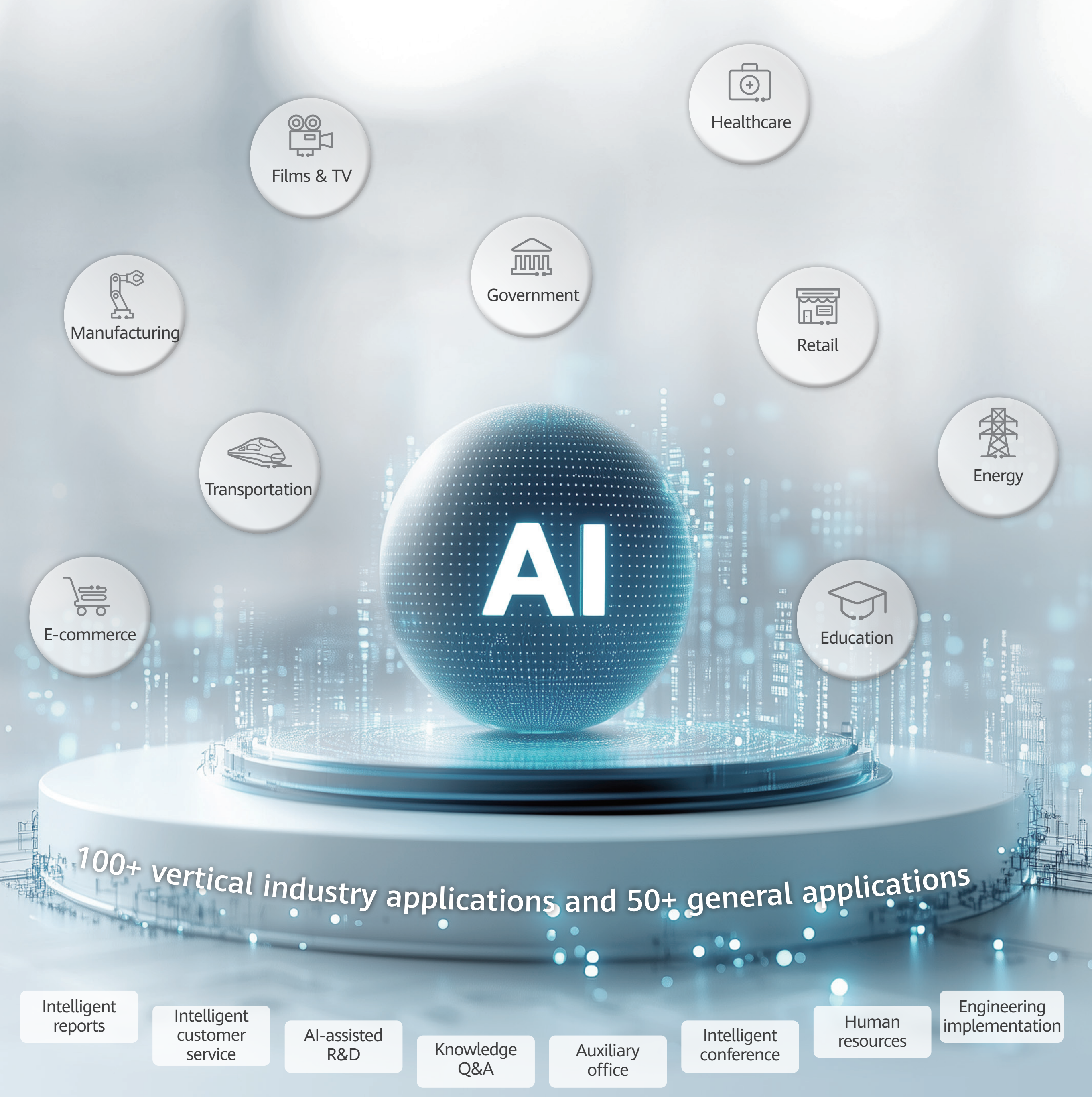
The decision-making and recovery module provides automatic fault contingency plan processing capabilities, supporting port fault isolation, traffic scheduling, and task migration to quickly restore services.

This comprehensive system for cluster stability during AI foundation model training includes both white- and black-box monitoring to detect and prevent faults within seconds. Faults are diagnosed to specific NEs based on the full-link topology and rectified using three methods, providing an effective assurance system for the stability and reliability of AI clusters. This ensures the accelerated development of AI services.

Transforming Industries and Driving Intelligent Evolution

Industry-specific solutions using an AI-native application engine for one-stop enterprise AI adoption challenges

Empower enterprises to achieve intelligent transformation, foster innovative growth, optimize decision-making, reduce costs, and enhance efficiency



100+ vertical industry applications and 50+ general applications

Intelligent reports

Intelligent customer service

AI-assisted R&D

Knowledge Q&A

Auxiliary office

Intelligent conference

Human resources

Engineering implementation

Exploration and Practice of Application Observability Solutions in the Era of Foundation Models

– Huawei GTS and Tingyun's Future-oriented Application Observability Solution

Author: Wang Fuqiang, Yang Jinquan

Abstract

This document explores practice used in Huawei GTS and Tingyun's application observability solutions based on foundation model technologies. It analyzes the background, implementation strategies, and outcomes, using innovative technologies to enhance application observability, ensuring enterprises maintain competitiveness in complex and evolving technical environments.

Wang Fuqiang

An SRE expert at Huawei Cloud services, he has vast experience in SRE, having led the migration of over 180 tenants from other clouds to Huawei Cloud and built an end-to-end operations process and specification system for cloud services.

Yang Jinquan

As Tingyun's CTO, he has extensive R&D and commercialization experience in the intelligent observation platform field. As a pioneer of early commercial APM tools, he offers deep insights into this field.

Background

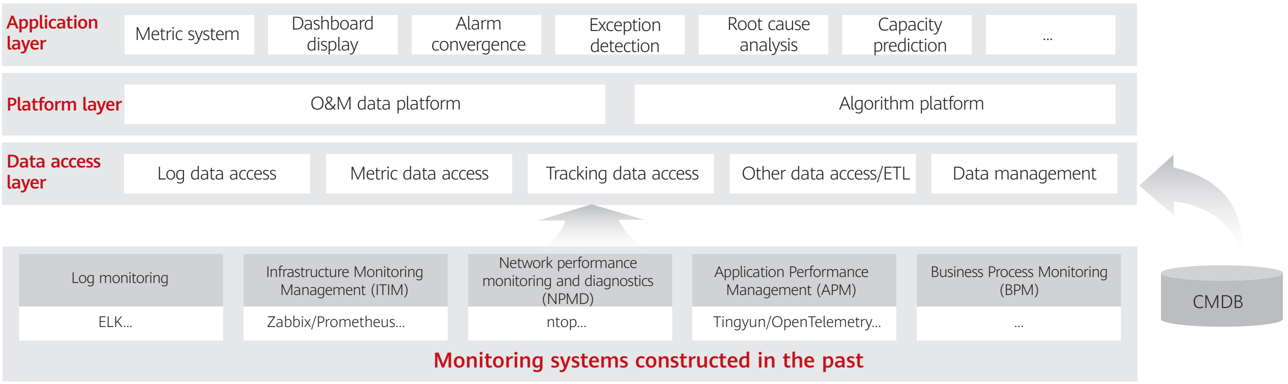
In today's age of digital transformation, enterprises increasingly rely on complex applications and services to support their operations. These applications demand rapid responses, high availability, and reliability. To ensure their performance and stability, application performance management (APM) and observability tools are critical.

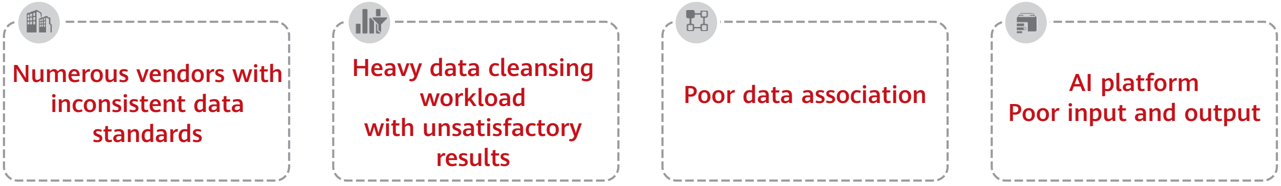
Traditional APM tools focus on monitoring performance metrics, such as response times, throughput, and error rates. However, rapid technological advancements and changing service requirements have made performance monitoring alone insufficient for modern applications. The rise of cloud computing, microservice architecture, containerization, DevOps, AI, IoT, edge computing, and 5G technologies has driven the innovation and development of application observability solutions. While these technologies offer flexible resource management and service deployment capabilities, they also introduce new monitoring challenges. Enterprises need advanced, comprehensive monitoring systems to manage dynamic cloud environments, complex microservices, transient containers, and rapid DevOps iterations. Advances in AI and machine learning provide new tools for in-depth monitoring data analysis, and the growth of 5G and IoT has further expanded the scope and depth of the monitoring.

Modern applications require observability, integrating and analyzing data sources such as logs, metrics, and distributed

tracing, beyond traditional performance monitoring. Traditional methods cannot fully reflect the health and performance of applications.

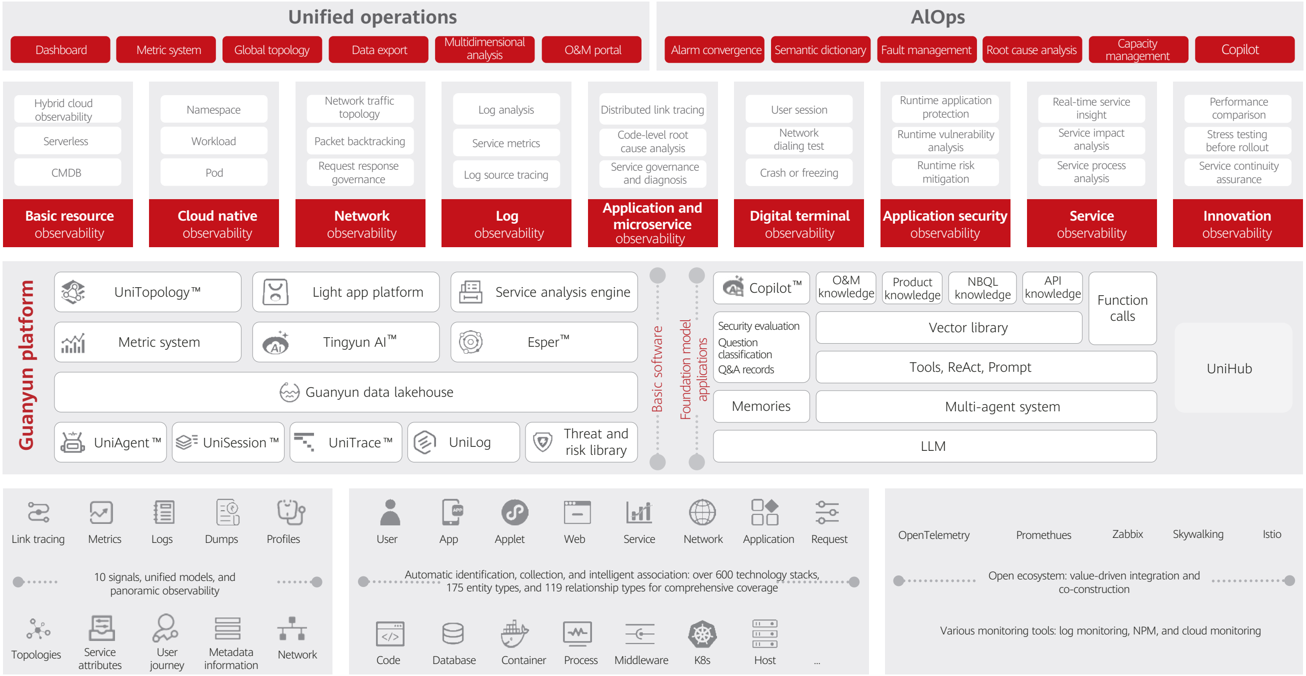
In GTS, all applications have transitioned to modern, cloud-native applications on Huawei Cloud. To ensure their stability, numerous systems and tools have been developed, establishing an integrated monitoring and AIOps system. Despite these advancements, some issues persist.




Numerous vendors with inconsistent data standards
Heavy data cleansing workload with unsatisfactory results
Poor data association
AI platform Poor input and output

Lessons learned

Observability Solution Based on Foundation Model Technologies



Panorama of observability in the foundation model era

In the era of foundation models, observability platforms must evolve to remain relevant. Foundation model technology enables intelligent fault detection and warning, automatic root cause analysis, and accurate capacity prediction. By using foundation models to analyze and process massive data, and integrating and associating different data sources, the observability platform not only enhances operations efficiency but also provides accurate and timely decision-making support, improving system stability and reliability.

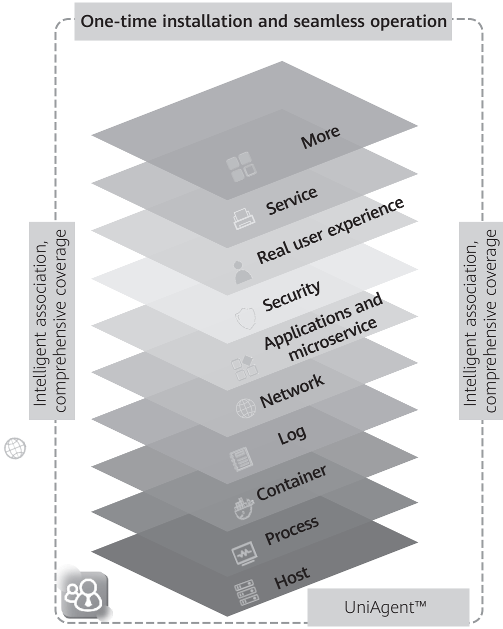
Building on this comprehensive approach, GTS and Tingyun have proposed an observability solution for the foundation model era, which includes the following aspects:

1 Establishing an Observability Data Model

A systematic observability data model is crucial for improving application observability. This model includes several core elements: Model classification adds signals such as topology, service attributes, user journey, metadata, and network based on the five signals of CNCF, forming 10 comprehensive data models. Data source definitions include the types and structures of various data sources, including traditional monitoring data and data generated by emerging foundation models. Data classification and labeling facilitate subsequent query and analysis. Data association establishes relationships between different data sources, providing comprehensive data views and supporting in-depth analysis across models, systems, and applications. Data standardization defines unified data formats and standards to ensure seamless integration and analysis of data from different sources. This comprehensive model significantly improves application observability, helping enterprises use data more efficiently and optimize system performance.

2 Automatic Management of O&M Objects and Full Data Collection and Convergence

Automatic management uses automation tools and intelligent platforms to automatically discover and manage O&M objects such as servers, applications, and services, ensuring that newly deployed resources are monitored automatically. Full data collection uses distributed technologies to ensure the integrity and accuracy of all key data, such as logs, metrics, and distributed tracing. Data convergence integrates data from different sources to form a single, unified view that enables cross-system data integration and analysis. It provides more comprehensive observability. These technologies enhance operations automation and intelligence, laying a solid foundation for efficient system management.

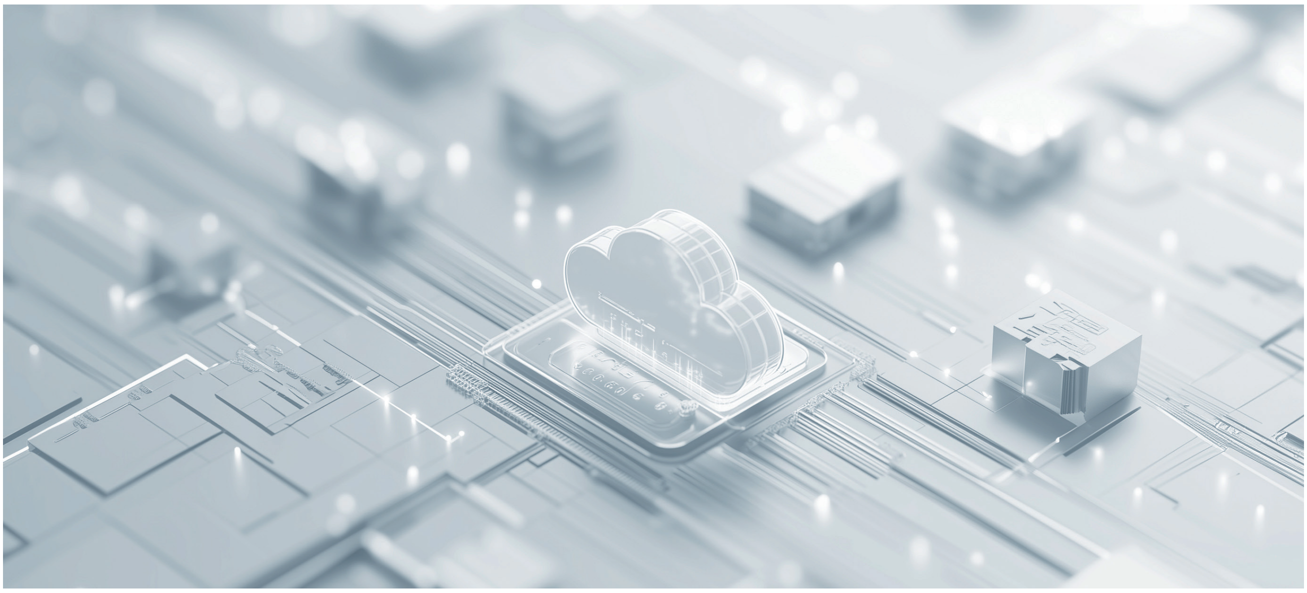

One-time installation and seamless operation
Intelligent association, comprehensive coverage
More
Service
Real user experience
Security
Applications and microservice
Network
Log
Container
Process
Host
Intelligent association, comprehensive coverage
UniAgent™

- Collection as governance: Connects data, models, objects, and AI to implement a panoramic context.
- Automatic end-to-end data collection: Facilitates automatic identification and collection, covering over 600 technology stacks.
- Unified data model: Integrates five signals and five extended signal models from data collection, ensuring a consistent "single source of truth" without tool switching or data calibration.
- Future-oriented scalability: Supports host or Kubernetes-based deployment, easy upgrades, flexible configuration, and on-demand orchestration and loading.
- Unified management: Manages a diverse range of collection objects, including users, services, hybrid apps, web pages, applets, networks, applications, security, requests, code, processes, containers, middleware, and databases.
- Multi-source technology stack: Incorporates Preload, eBPF, Instrumentation, Hook, and more.
- Ecosystem compatibility: Fully compatible with the open-source standard OpenTelemetry.

Install it and forget it – it handles everything

UniAgent Collection as Governance

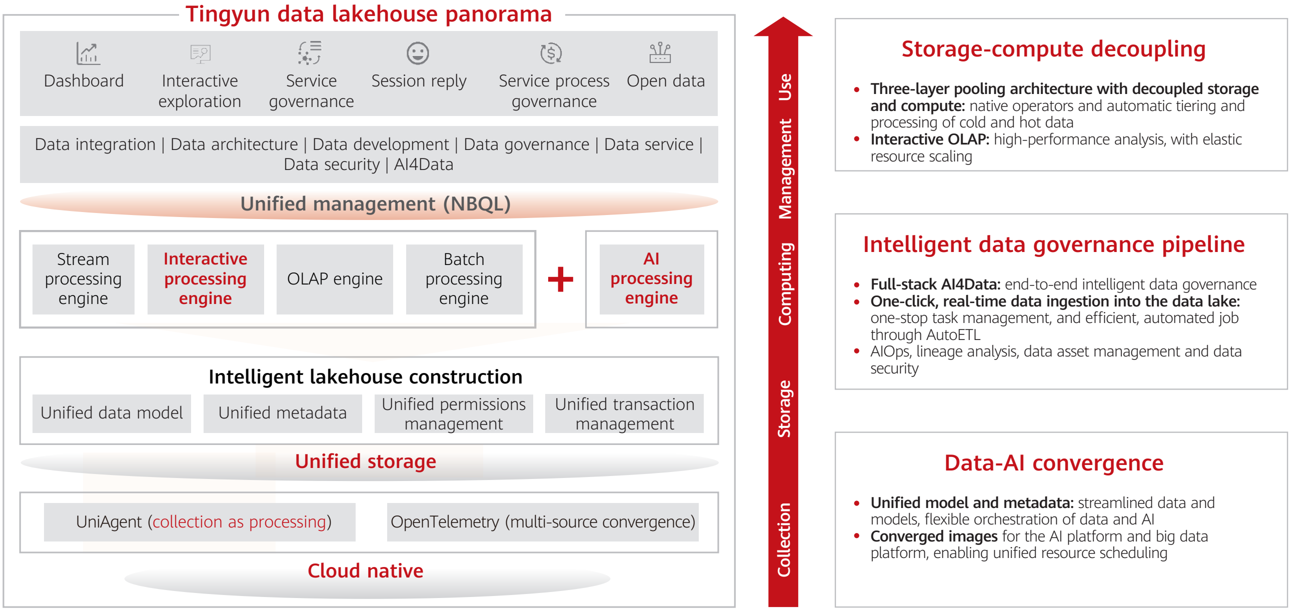
Collection as governance is the core of this solution. It provides high-quality observable data for AI through the governance of data models, data computing standards, and data associations.



3 Observability Basic Software Platform Construction

The platform provides high-availability and high-performance data processing and storage capabilities to support real-time processing and analysis of large-scale data. In a future-oriented observability system, a data lakehouse is a critical part of the observability platform.

Data lakehouses combine the advantages of data lakes and data warehouses to provide a unified data storage and management platform. The data lake supports the storage and processing of various types of raw data, while the data warehouse optimizes the query and analysis of structured data. By integrating the two, enterprises can store, manage, and analyze data on a single platform, improving data processing efficiency and flexibility. Data lakehouses use distributed and columnar storage systems to support efficient storage and query of massive data. Metadata management and data governance tools ensure data quality and consistency. The data lakehouse platform integrates big data processing frameworks and interactive query tools for batch processing, stream processing, analysis, and visualization of data. Its scalability and high availability ensure stable system operation through distributed architecture and multi-copy storage, meeting increasing data demands. This technology has become a core of many modern observability platforms, making enterprise decision making efficient, so they can operate effectively in the big data era.


Tingyun data lakehouse panorama
Dashboard
Interactive exploration
Service governance
Session reply
Service process governance
Open data
Data integration | Data architecture | Data development | Data governance | Data service | Data security | AI4Data
Unified management (NBQL)
Stream processing engine
Interactive processing engine
OLAP engine
Batch processing engine
+
AI processing engine
Intelligent lakehouse construction
Unified data model
Unified metadata
Unified permissions management
Unified transaction management
Unified storage
UniAgent (collection as processing)
OpenTelemetry (multi-source convergence)
Cloud native
Use
Management
Computing
Storage
Collection

Storage-compute decoupling

- **Three-layer pooling architecture with decoupled storage and compute:** native operators and automatic tiering and processing of cold and hot data
- **Interactive OLAP:** high-performance analysis, with elastic resource scaling

Intelligent data governance pipeline

- **Full-stack AI4Data:** end-to-end intelligent data governance
- **One-click, real-time data ingestion into the data lake:** one-stop task management, and efficient, automated job through AutoETL
- AIOps, lineage analysis, data asset management and data security

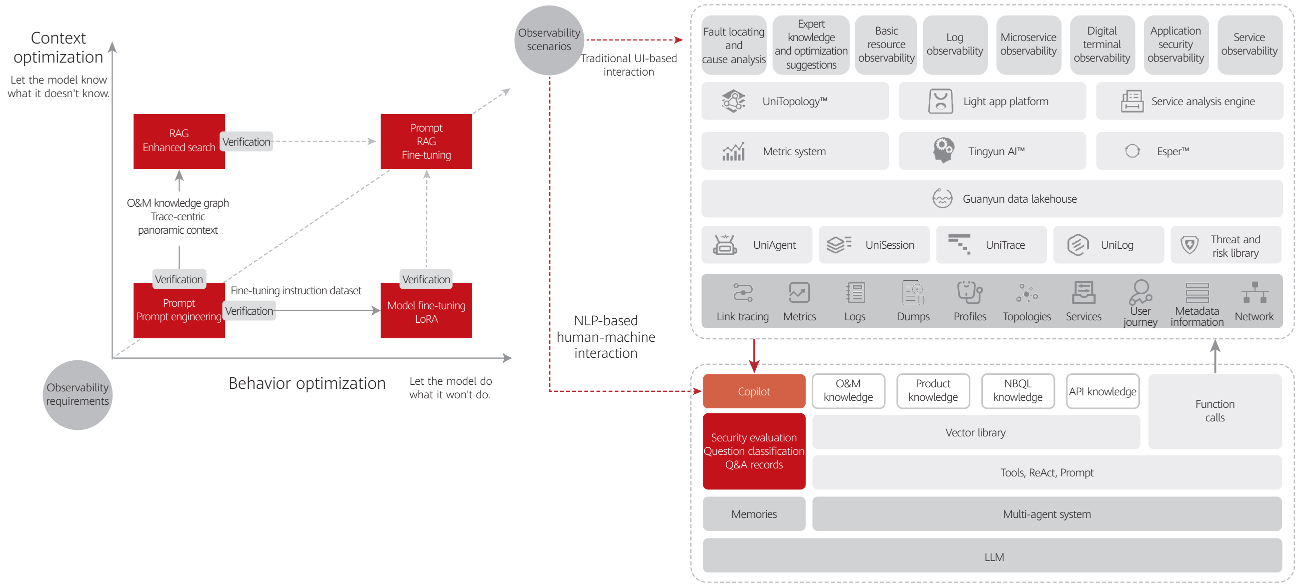
Data-AI convergence

- **Unified model and metadata:** streamlined data and models, flexible orchestration of data and AI
- **Converged images** for the AI platform and big data platform, enabling unified resource scheduling

Data lakehouse: full-stack observability big data infrastructure

4 Model Application Construction

Foundation model technology introduces a new dimension of intelligent analysis for observability. Leveraging big data platforms and machine learning frameworks, foundation models are optimized for anomaly detection, root cause analysis, and predictive analytics. These models automatically analyze logs, metrics, and trace data, identify abnormal patterns, locate root causes of faults, and provide optimization suggestions. Additionally, foundation models support multi-source data convergence and in-depth data mining, helping identify potential problems and optimization opportunities.



Foundation Model-based observability platform architecture

5 Unified Operations System Construction

The construction of a unified operations system is crucial for ensuring efficient operations and cross-department collaboration. The system includes four core capabilities: dashboard, metric systems, multidimensional analysis, and global topology.

» Dashboard: Provides a centralized, visual interface for real-time monitoring of system health and performance, supporting user-defined views and dynamic interactive analysis.

» Metric systems: Various system performance and service metrics need to be defined, including key performance indicators (KPIs) and alarm thresholds, to measure system stability.

» Multidimensional analysis: Data slicing, trend analysis, and root cause analysis are used to explore system behavior and performance changes.

» Global topology: Relationships between system components and services are displayed visually. There are dynamic topology views, fault propagation path analysis, dependency management, and automatic topology updates.

This system provides a comprehensive operations perspective, ensuring efficient and stable system operation for enterprises.

6 AIOps System Construction

GTS and Tingyun focus on key modules such as exception detection, alarm convergence, causal relationship-based deterministic root cause analysis, fault management, and capacity management to build a more intelligent and efficient operations ecosystem.

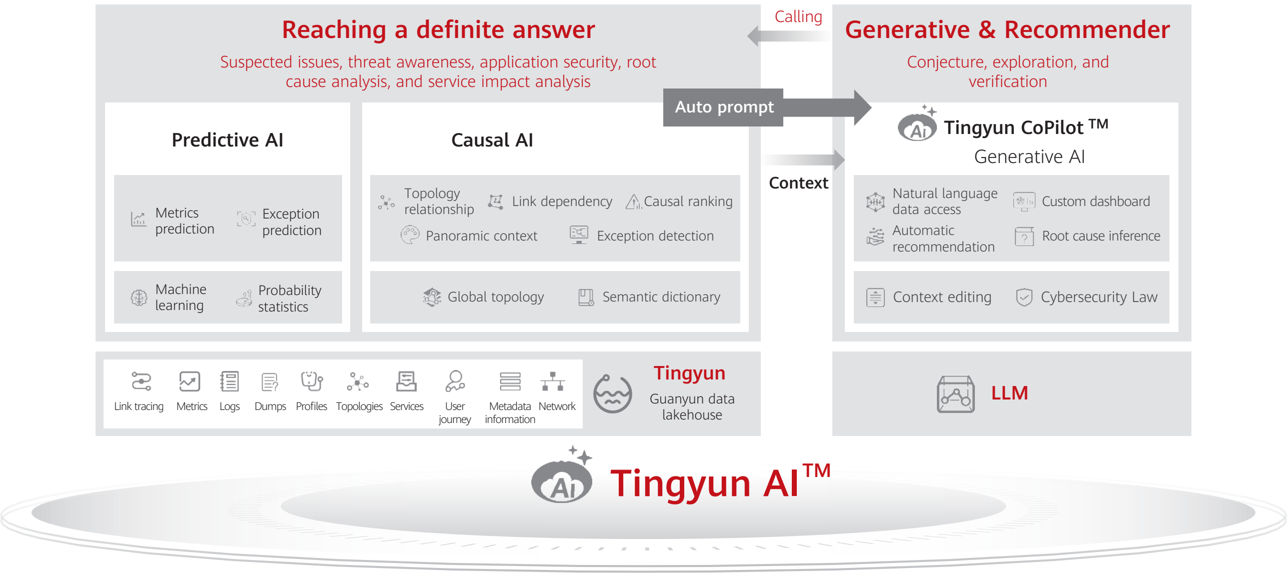
» Anomaly detection uses multidimensional data analysis and machine learning models to identify abnormal behavior in real-time, automatically adjusting thresholds and triggering alarms.

» Alarm convergence reduces alarm noise through aggregation, priority sorting, and intelligent association analysis, ensuring operations personnel focus on key issues.

» Deterministic root cause analysis uses causal relationship models to match historical data, quickly locating root causes and providing real-time analysis.

» Fault management includes automatic fault detection and rectification. The entire process is recorded and traced, promoting continuous system improvement.

Integrating these modules creates an intelligent and efficient operations system, significantly improving system reliability and response speed.



AIOps system architecture

Summary and Prospects

In the context of digital transformation, enterprise application systems are becoming more complex, challenging traditional operations methods. By introducing the AIOps system, GTS and Tingyun have developed a future-oriented observability solution. Advanced AI and automation technologies enhance operations efficiency, system stability, and reliability.

The AIOps system enables enterprises to respond quickly and accurately to system issues, reduce service interruptions, and optimize resource utilization. The application of large-scale data analysis and machine learning models allows the AIOps system to automatically identify problems, optimize resource configuration, and operate efficiently in complex environments, delivering significant operational value to enterprises.

Looking ahead, the further development of foundation model technologies and AI will make AIOps more automated and proactive. The AIOps system will not just monitor and respond reactively. It will predict and prevent problems, enabling self-healing and self-optimization. With the emergence of technologies such as 5G, IoT, and edge computing, operations observability will expand to cover more scenarios and data sources, providing comprehensive, in-depth operational support for enterprises. Future operations systems will be more intelligent, automated, and adaptive, offering robust technical support for digital transformation and promoting continuous service innovation and development. In this process, Huawei GTS and Tingyun will continue to lead the industry, exploring and practicing advanced observability solutions, and helping enterprises achieve operational excellence and business growth.

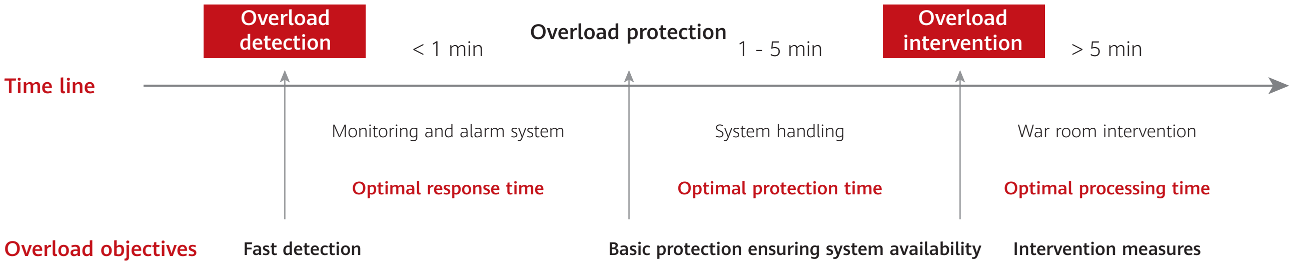
Three Steps of Cloud Service Overload Control

Author: Zheng Lei

Abstract

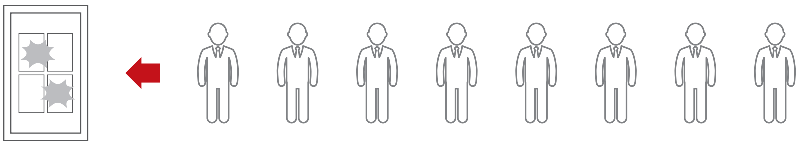
There are many uncertainties involving cloud services. These uncertainties can arise from customer applications, public network requests, and system workloads. In this article, we focus on how to achieve high availability at the cloud service, microservice, and tenant levels to prevent service overload and minimize related impacts on tenant services.

Overload control involves three crucial steps: detection, protection, and intervention. Essential capabilities need to be carefully planned and developed based on these steps. This article specifically emphasizes the practical aspects of overload awareness and overload intervention.

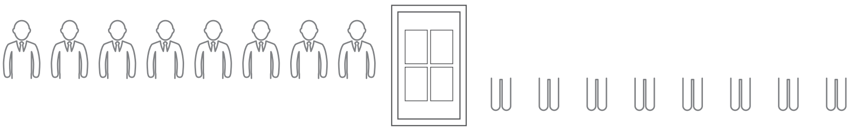


Symptoms

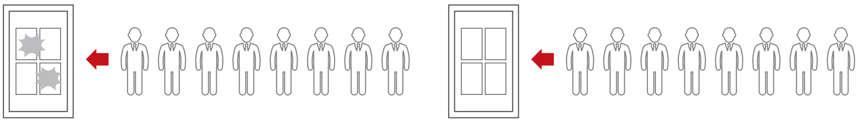
Symptom 1: A critical node becomes overloaded and is unable to handle all incoming requests. Typical example: Hackers initiate excessive queries to overwhelm network services and exhaust CPU resources, preventing other customers' requests from being processed.



Symptom 2: Unintended resource occupation reduces service performance for all tenants or requests Typical example: When there are no QoS policies configured, requests from a single tenant cloud overwhelm the storage pool and use up xxx Gb/s of bandwidth. This can lead to delayed responses and affect all users.



Symptom 3: Dependencies are over loaded. Typical example: Tens of thousands of VMs simultaneously request DNS services, exceeding the threshold of the security system. In such cases, the security system prevents the DNS from being overwhelmed by excessive requests.



Overload Detection

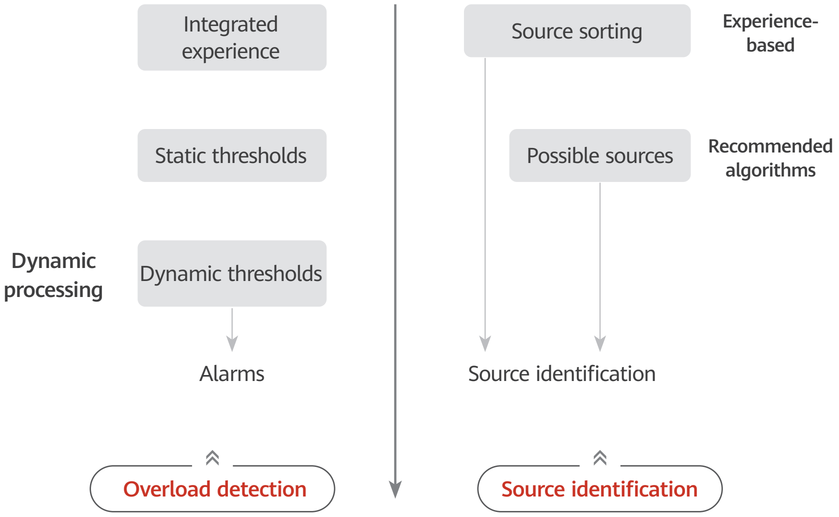
Key Capabilities

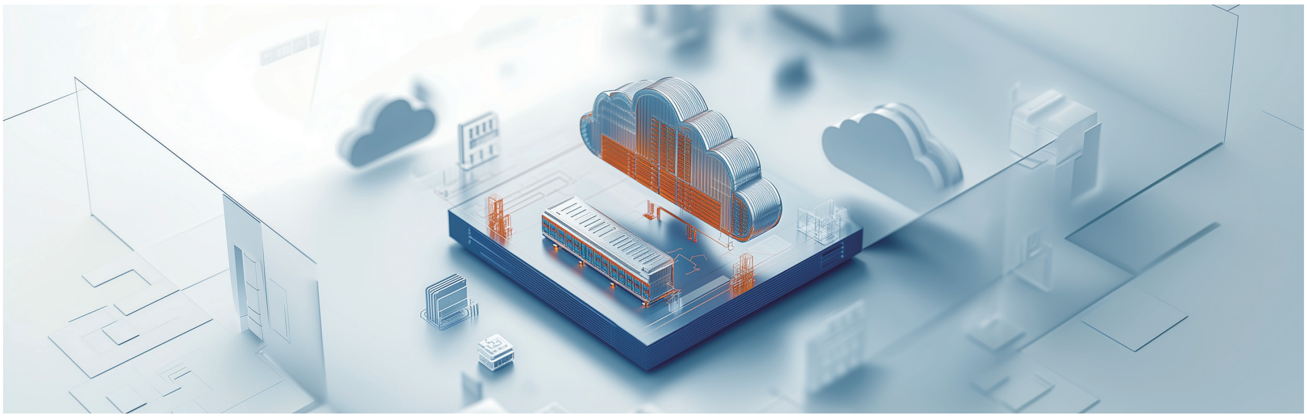
Detecting overloads is crucial for effective control. It is important to detect overloads as early as possible, so we can promptly respond to related incidents and address the sources of overload. We have accelerated this process to where it only takes a few minutes.

Detecting Overload and Identifying Sources

Overload detection: We depend on the monitoring and alarm system to detect overloads. We mainly use two methods: setting thresholds and using recommended algorithms.

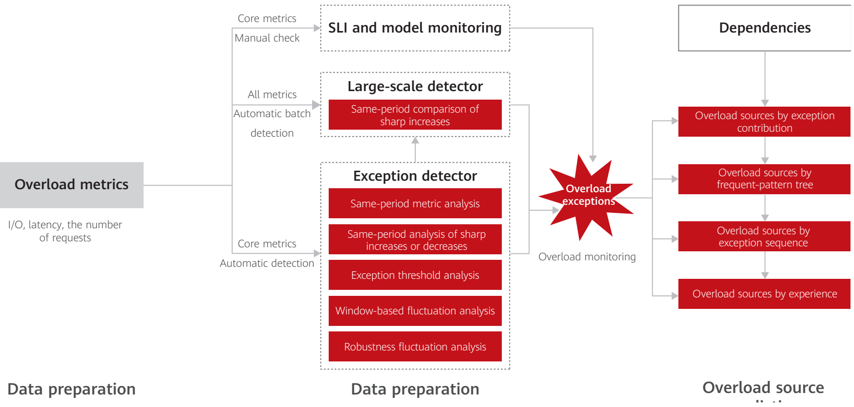
Source identification: We depend on the monitoring and alarm system to identify the sources of overload. We identify and sort sources of overload based on experience. We also use recommended algorithms to dynamically detect sources and identify associated objects.





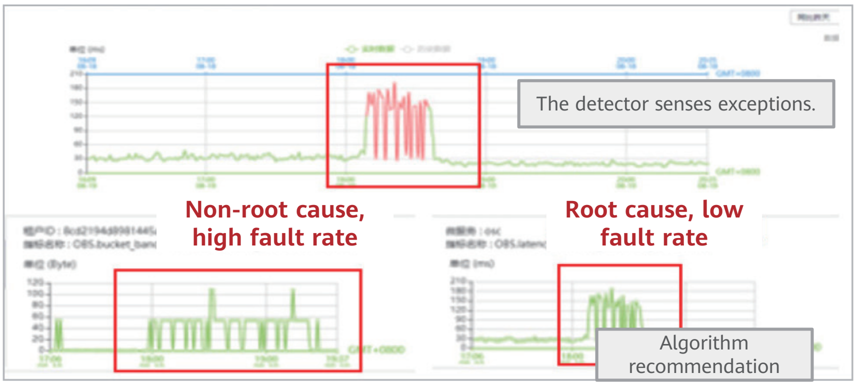
Capability Architecture

We have established a complete service monitoring system to track the health of service processes. This system enables us to accurately detect and quickly locate faults, identify overload in minutes, and pinpoint the sources of overload. It eliminates the need to manually filter data or switch between systems. Using workload data, we designed algorithms that can quickly detect overload and identify its sources.

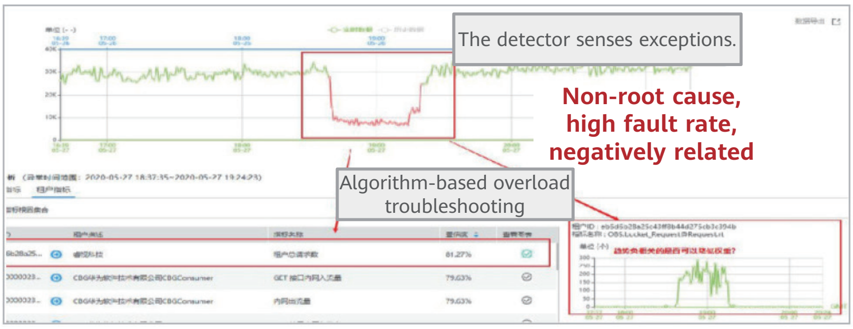

Overload metrics
I/O, latency, the number of requests
Core metrics
Manual check
All metrics
Automatic batch detection
Core metrics
Automatic detection
SLI and model monitoring
Large-scale detector
Same-period comparison of sharp increases
Exception detector
Same-period metric analysis
Same-period analysis of sharp increases or decreases
Exception threshold analysis
Window-based fluctuation analysis
Robustness fluctuation analysis
Overload exceptions
Overload monitoring
Dependencies
Overload sources by exception contribution
Overload sources by frequent-pattern tree
Overload sources by exception sequence
Overload sources by experience
Data preparation
Data preparation
Overload source

Practices

Addressing common false positives


The detector senses exceptions.
Non-root cause, high fault rate
Root cause, low fault rate
Algorithm recommendation

Minimizing the interference caused by associated metrics when sorting the sources


The detector senses exceptions.
Non-root cause, high fault rate, negatively related
Algorithm-based overload troubleshooting

Overload Intervention

Overload intervention is based on service statuses. It involves several layers, including the access, capability, service, and data layers. If an overload is detected, the overload control service takes proactive measures to ensure service stability and limit the impact scope.

Key Capabilities:

Access Layer

Overload is controlled based on APIs, tenant levels, tenant names, URLs, and domain names.

Capability Layer

Multiple measures can be taken, including automatic scaling; throttling based on APIs, users, and domain names, downgrading services based on APIs, users, and domain names; circuit breakers based on APIs and users, and isolation of clusters, nodes, and microservices.

Service Layer

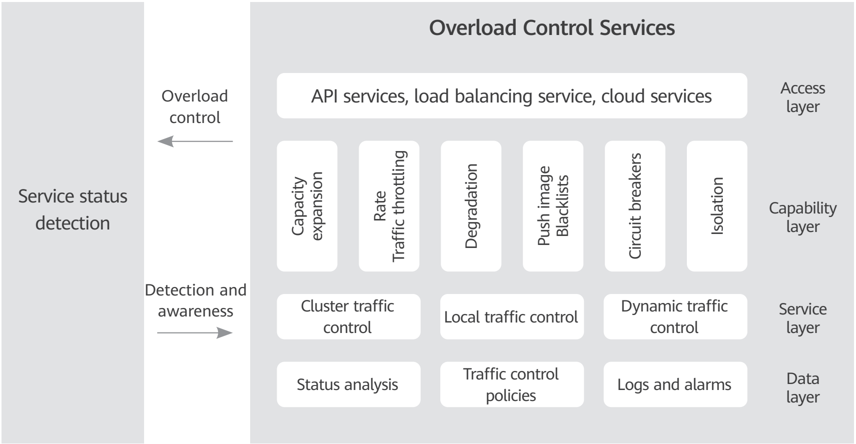
Dynamic overload control can be conducted at the cluster, node, and service levels

Data Layer

» We can centrally deliver traffic control policies to throttle requests based on their priority, latency levels, and workload levels.

» We can also analyze overload control results, including the timeliness, false positives, and false negatives.

» Logs and alarms allow us to detect overloads, analyze the timeliness of overload control, and evaluate the effectiveness of related measures.



Typical architecture of overload control

Summary

To summarize our overload control approaches, we draw inspiration from the Dujiangyan Irrigation System. In ancient times, people who lived in Sichuan Basin were plagued by flooding for many years. Instead of building a dam, they established a system to redirect the water and control the flood while also allowing the water to flow through naturally. This system took advantage from the geographic condition that the land was high in the northwest and descended to the southwest and utilized the river's natural flow, specific topography at the outlet of the river, as well as the river's natural patterns and tributaries. The irrigation system consisted of a weir, a levee, and a channel bottleneck.

» The Fish Mouth Levee divided the water in to inner and outer streams. It not only provided irrigation and drinking water during the dry season but also reduced the river flow, which reduced flooding during the rainy season.

» The Flying Sand Weir was responsible for flood control. It allowed the natural swirling flow of the water to drain out excess water from the inner to outer streams. The swirl also drained out silt and sediment.

» The Bottle-Neck Channel was a narrow gate that could also work to control excess water.

When developing or use cloud services, we may experience overload or excessively long job queues caused by various factors, such as service promotions, public network storms, and large-scale attack or security scanning conducted by hackers. In this article, we described how overload detection and intervention can be used to analyze and address overload issues of cloud services and applications in an organized and systematic manner. We hope that these practices and capabilities can provide insights for overload control challenges in developing cloud services and applications. We also encourage more individuals to join in the sharing and discussion of overload control technical challenges, new technologies, and key capability evolution.

Architecture Design and Switchover Capability Improvement for Government Clouds Service Security, Stability, and Quality Ensured

Author: Kang Zhen, Li Guoqiang, Liu Chunchun, Yin Gezhen

Abstract

This article explores how Huawei's Deterministic Operations capability system can help government customers accelerate their digital transformation. It provides solutions for HA architecture, emergency drills, and cloud adoption practices to ensure the security, stability, and quality of services.

Background

The National Data Administration, along with other government departments have issued the Guidelines on Deepening Smart City Development and Promoting All-Domain Digital Transformation of Cities. These guidelines outline the direction for comprehensive digital transformation of cities and set requirements for key areas, including infrastructure construction, digital economy cultivation, precise city governance, city environment optimization, and city security assurance. Government clouds are expected to go beyond enhancing service efficiency and innovating governance models. They must also prioritize security, reliability, resource efficiency, and service agility.

In the Huawei Cloud Deterministic Operations Issue 4, the SRE team shared methods for improving the observability, security, and capacity health of Government clouds. This issue will share the methods and best practices to optimize architectures and emergency switchover capabilities for Government clouds.

Challenges

Government digital transformation has gone through the early stage of government informatization to the e-Government phase, and has now arrived at the cloud-based digital government stage. More government services, including operations, social services, social governance, and economic development, are being moved to the cloud. Government clouds benefit from the value of data but also face increased threats and attacks. In recent years, there have been many incidents worldwide, such as critical system shutdowns, cloud hacks leading to data loss, and failures causing websites to crash. This poses great challenges to the continuity and security of government cloud services. So let's look at some of the challenges involved in constructing government clouds.

First, government clouds often prioritize the construction of cloud infrastructure over operations. This means that they may lack the necessary systems for ensuring platform stability, reliability, and security compliance. Additionally, emerging technologies like AI are not effectively integrated with operations and maintenance (O&M), and traditional O&M methods are no longer sufficient for maintaining service stability in digitalized scenarios. The shortcomings of traditional O&M, such as weak availability management, reactive responses, complex processes, a lack of fault management capabilities, and misalignment between O&M tools and service requirements, are all impeding the progress of government digitalization.

Second, Government clouds emphasize development over resilience. As the digitalization of government clouds advances, their cloud systems become more complex. Moreover, there are increased data silos and rollbacks during version upgrades. Enhancing the abilities to prevent system faults and quickly recover services has become critical.

To address these challenges, a government cloud SRE team took advantage of Deterministic Operations capability system and solutions. They adopted various cloud practices to improve system availability, optimize architectures, and quickly identify system threats. This helped government customers improve their O&M systems, increase O&M efficiency, and ultimately achieve digital transformation.

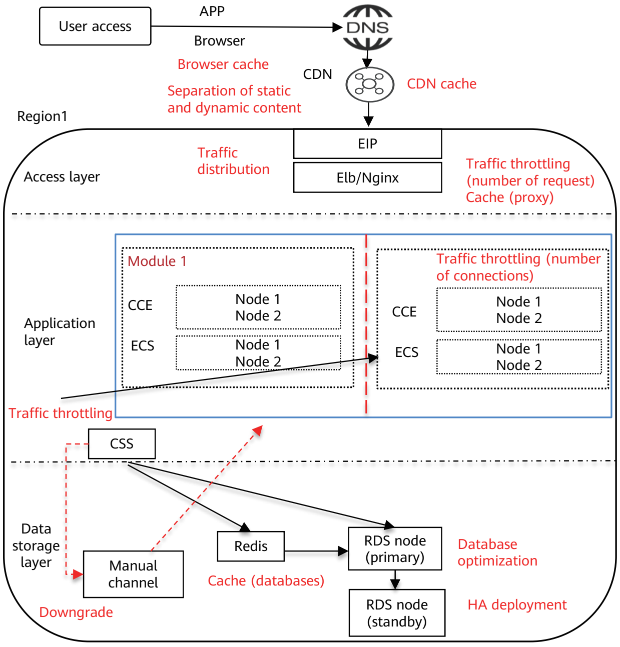
Designing Architectures to Improve Reliability

We analyzed service, application, and technical architectures with a focus on establishing HA capabilities to identify risks in architectures and service systems and provide optimization suggestions for different service scenarios.

- Application architecture optimization: We analyzed performance and scalability issues of each service module and subsystems. We then offered optimization solutions that addressed various scenarios such as separating dynamic and static content, distributing traffic, isolating and decoupling, throttling traffic, and downgrading.
- Data optimization: After analyzing the volume of service requests and the direction of data flow, we offered cache and database solutions to speed up data access.
- Deployment optimization: We eliminated single points of failure (SPOFs) and enabled data backup. Additionally, we invested the best possible resources to help our government customers reconstruct architectures and improve service stability and availability.

Let's use a government website as an example to illustrate how we analyze, optimize, and derive value from architectures. This covers the access, application, and data storage layers.

No.	Architecture Layer	Scenario	Benefit
1	Access layer	Browser cache	Performance: improved website access speed
2		Separation of static and dynamic content	Performance: faster page loading
3		CDN cache	Performance: reduced access latency and less pressure on source servers
4		Traffic distribution	Scalability: auto scaling enabled for better system processing
5		Traffic throttling	Performance: systems protected from excessive requests
6		Proxy cache	Performance: faster access and less pressure on source servers
7	Application layer	Isolation	Scalability and high availability: modular and component-based services, faults isolated to only individual services
8		Traffic throttling	Performance: systems protected from excessive requests
9	Data storage layer	Downgrade	High availability: improved availability of key services
10		Cache (Redis)	Performance: reduced database load and increased throughput
11		Database	Performance: Improved Read and Write speed
12	HA deployment		High availability: improved availability of key nodes

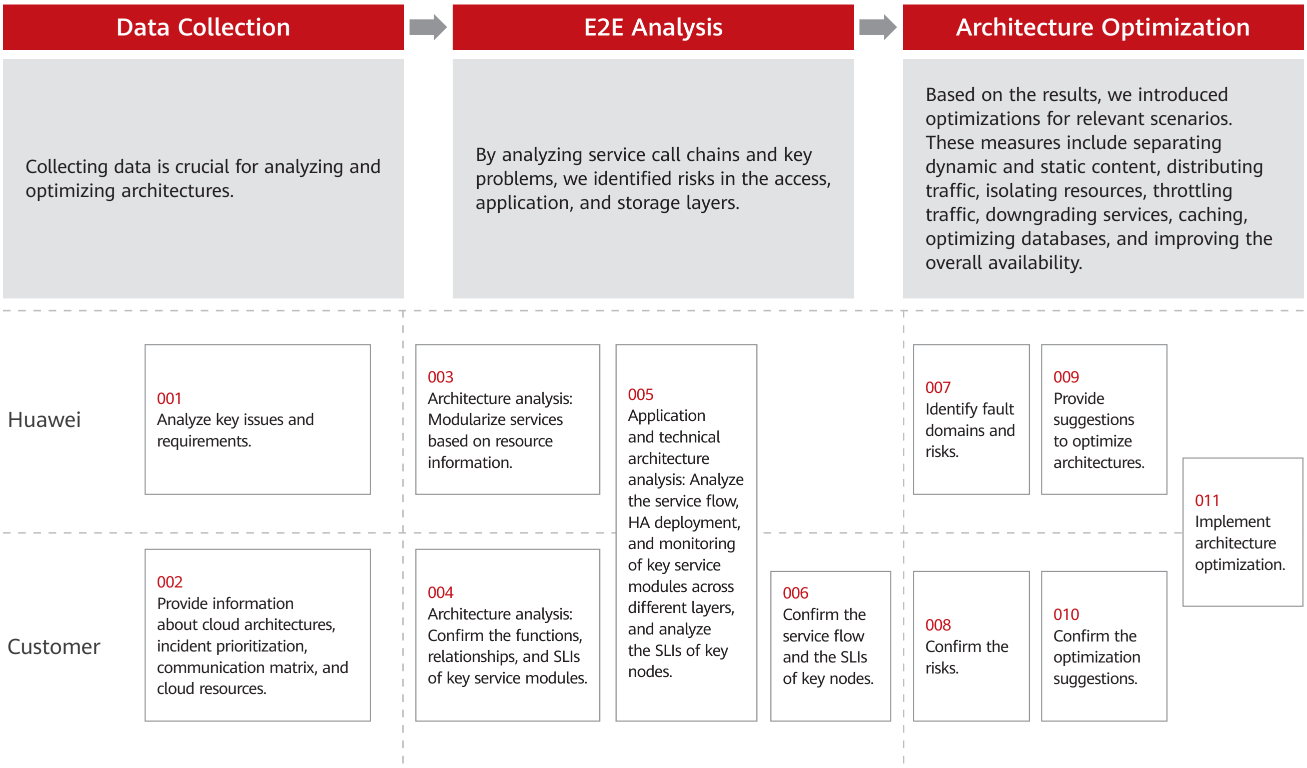


The overall service analysis and optimization consists of three steps:

- Data collection: Collecting data is crucial for analyzing and optimizing architectures. The data collected includes key requirements and issues related to migrating to or using the cloud. This helped us identify service architecture risks. We also collect information about the communications matrix, incident prioritizing criteria, cloud resources, and cloud architectures.
- E2E analysis: To identify risks in the access, application, and storage layers, we analyzed service call chains and examined the key problems involved. Additionally, we modularized service architectures to analyze various aspects such as functions, key SLIs, peak hours, traffic changes (sudden spikes or drops), closely related modules, and incident prioritization for each module.
- Architecture optimization: We analyzed live network metrics and key SLIs, pinpointed fault domains and risks in architectures, identified improvement scenarios, and then determined and implemented optimization measures. These measures addressed various scenarios, including separating dynamic and static content, distributing traffic, isolating resources, throttling traffic, downgrading services, caching, improving databases, and enhancing availability.

Improving Emergency Drill Capabilities

To improve emergency drill capabilities of government clouds, we applied chaos engineering principles. We also drew from the practical experience of the Huawei Cloud SRE team, specifically their expertise in emergency drills. By simulating real-world environments across multiple



Data Collection

Collecting data is crucial for analyzing and optimizing architectures.

E2E Analysis

By analyzing service call chains and key problems, we identified risks in the access, application, and storage layers.

Architecture Optimization

Based on the results, we introduced optimizations for relevant scenarios. These measures include separating dynamic and static content, distributing traffic, isolating resources, throttling traffic, downgrading services, caching, optimizing databases, and improving the overall availability.

Huawei

001 Analyze key issues and requirements.

003 Architecture analysis: Modularize services based on resource information.

005 Application and technical architecture analysis: Analyze the service flow, HA deployment, and monitoring of key service modules across different layers, and analyze the SLIs of key nodes.

007 Identify fault domains and risks.

009 Provide suggestions to optimize architectures.

011 Implement architecture optimization.

Customer

002 Provide information about cloud architectures, incident prioritization, communication matrix, and cloud resources.

004 Architecture analysis: Confirm the functions, relationships, and SLIs of key service modules.

006 Confirm the service flow and the SLIs of key nodes.

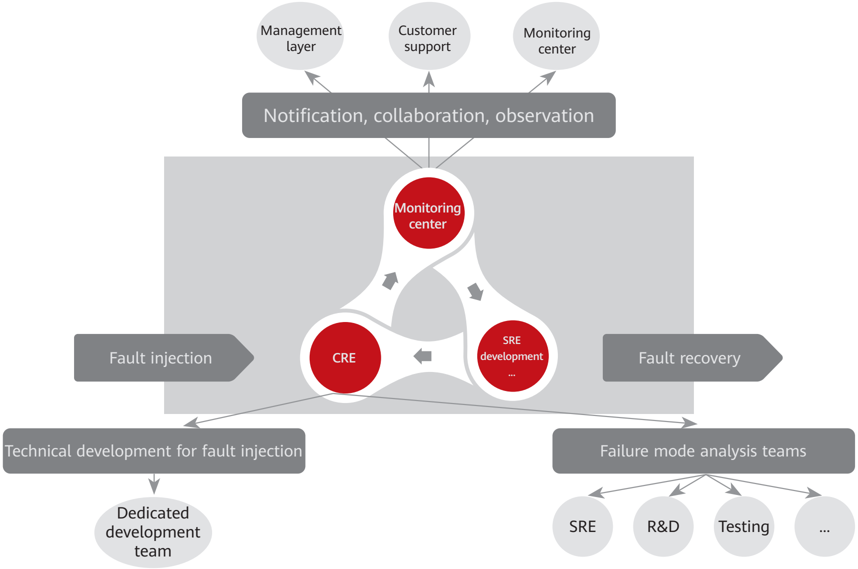
008 Confirm the risks.

010 Confirm the optimization suggestions.

domains, we offered various drill scenarios and emergency plans. Our ultimate goal was to help government customers gain confidence in their cloud service systems.

Expected Benefits

- We can identify places where we can improve the collaboration, communication, and O&M skills of the O&M personnel.
- We can identify issues in the emergency plan, ensure the emergency plans are practical, and confirm the readiness.
- We improve confidence in the system's ability to handle unpredictable and challenging conditions in the production environment.
- We can verify the effectiveness of the cloud service HA solutions, including redundancy design, data protection, and DR measures.


Management layer
Customer support
Monitoring center
Notification, collaboration, observation
Monitoring center
Fault injection
CRE
SRE development ...
Fault recovery
Technical development for fault injection
Failure mode analysis teams
Dedicated development team
SRE
R&D
Testing
...

Drill Solution

Drill Planning

After analyzing customer requirements, the SRE cloud adoption capability center for government clouds designed a drill plan. We then formed an emergency drill team compromising Huawei frontline account managers and O&M engineers and members of the government cloud SRE team. This team was responsible for communicating with customers and discussing solutions with them.

Preparation

Conducting fault drills is crucial for key service scenarios, such as cloud migration, service promotion, version update, and key events. These drills can help identify risks in key services, provide suggestions for optimizing them, and ultimately enhance the system's ability to withstand faults.

- Designing the drill plan: The Huawei government cloud SRE team collaborated with the customer's drill team to determine the scope, scenarios, and schedule for the drill.
- Designing the drill solution and emergency plan: The government cloud SRE team determined the drill objectives for each scenario. The cloud adoption optimization center produced a draft drill solution, monitoring solution, and emergency plan. The Huawei drill team reviewed the solutions and plan and submitted them to the customer for review. Once approved, they would serve as the final drill solutions and emergency plan.

Implementation

- Before the drill, the government cloud SRE team confirmed that all necessary preparations were completed to ensure a smooth drill.
- The customer initiated the drill, and the SRE engineers collaborated with them to conduct the drill following the solutions.
- The government cloud SRE team and the customer recorded the time and outcomes of each step, while also promptly verifying the accuracy of alarms.

Review and Summary

The government cloud SRE team and the customer reviewed the drill records to determine if expected objectives were met, and the customer reviewed the drill results.

The SRE team summarized the drill process and results and generated a drill summary report summarizing the following aspects of the drill.

- Were HA solutions or configurations effective? If not, why, and how could they be improved?
- Were RPO and RTO requirements met? If not, why, and how cloud related solutions be improved?
- Was the emergency plan effective, and are there areas for improvement?

Summary

Digital transformation is an ongoing journey without a definitive end. Likewise, there is no limit to enhancing the security and resilience of digitalized cities. The digital era is already here. Government clouds are essential for the digitalization of cities. It is crucial to maintain stable, reliable, secure, and efficient government clouds to ensure the continuous and stable development of cities. After almost a decade of dedicated effort, the government cloud SRE team has consistently enhanced their capabilities. They have gained valuable experience in optimizing system architectures and enhancing cloud adoption. They have helped numerous government customers develop secure, stable, high-quality services, ultimately leading to successful digital transformation. Furthermore, the government cloud SRE team has added their capabilities in the course of over 800 successful government cloud projects, offering a range of professional service products and solutions. These services encompass consulting and planning, cloud migration and implementation, O&M and management, as well as optimization and enhancement. The team plays a vital role in building resilient and secure cities. They work tirelessly to protect digital cities around the clock, actively contributing to the government's digital transformation efforts and paving the way for a brighter digital future.

Enhancing Architecture Resilience For Stable, Reliable Cloud Migration and Adoption

By Jin Biao

Abstract

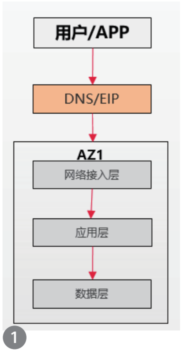
In this article, we will explore how the customer reliability engineering (CRE) team helps customers satisfy SLAs, address risks, and ensure service stability through a range of measures. These measures involve deploying HA architectures, elastic scaling, managing overload, safeguarding services from faults, and minimizing blast radii.

Huawei Cloud is developing rapidly, and enterprises are advancing their cloud journey. However, when designing cloud system architectures, customers encounter various challenges, including issues with system availability, limited monitoring capabilities, slow fault recovery, and risks associated with service operations. To address challenges, such as unknown infrastructure and software, overload, and other exceptions, the Huawei Cloud CRE team has developed key technologies to enhance the resilience of service architectures. These technologies include implementing HA architectures, auto scaling, managing overload, E2E fault escape, minimizing blast radii. They can effectively satisfy SLAs and ensure stable, reliable cloud migration and adoption.

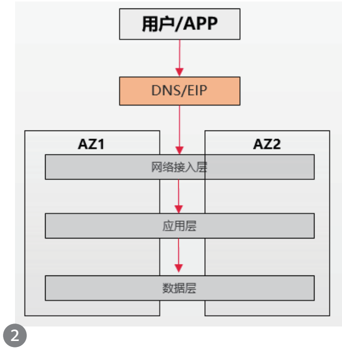
HA Architecture Deployment

During different stages of development, tenants may iterate their service architectures based on costs and environmental factors. There are three key metrics of architecture availability: the blast radius, the recovery time, and the maximum tolerable data loss. Different levels of reliability come with different costs. Public cloud customers should implement HA architectures that align with their service development. Tenant architectures typically follow a certain evolutionary path: single-AZ cluster → cross-AZ architecture → cross-AZ, dual-cluster architecture → cross-AZ active-active architecture → multi-region multi-active deployment.

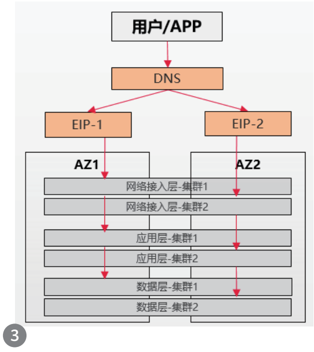
High availability (HA) architectures can effectively mitigate hardware and software faults, ensuring a service availability of 99.99%. The most popular HA architectures on Huawei Cloud are cross-AZ, followed by cross-AZ, dual-cluster, cross-AZ active-active, and multi-region multi-active.



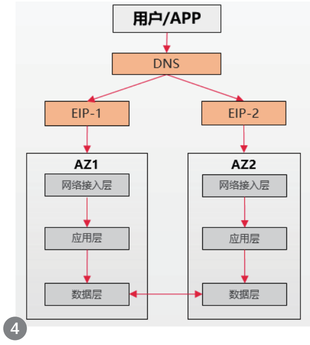
1



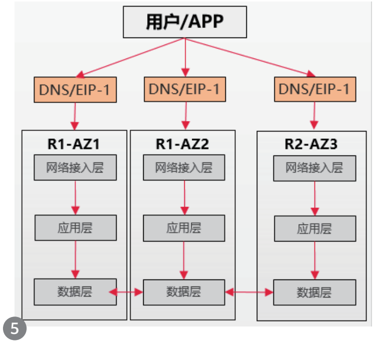
2



3



4



5

Auto Scaling

Controlling cloud costs is crucial for tenants using public cloud services. When it comes to system redundancy, it is essential to carefully manage capacity to avoid excessive idle resources. In the event of a major fault at the AZ level, performing a switchover on active-active clusters could lead to capacity or performance issues. This is where auto scaling comes into play.

Auto scaling dynamically increases or decreases instances in response to service loads and the policies configured to optimize cloud resource usage, performance, and costs.

Capacity can be scaled automatically or manually. Automatic scaling involves setting up rules in advance. When these rules are met, the defined scaling policies are automatically triggered. Manual scaling requires you to manually add or remove instances through certain user interfaces.

Auto scaling supports alarm-based, scheduled, and periodic policies.

» Alarm-based policies enable real-time scaling by configuring alarm rules for resources, such as CPU, memory, disks, and network resources. They are suitable for services with irregular load changes.

» Scheduled policies be precise down to the second. They are ideal for services that have regular daily peaks or troughs, or experience sudden load changes at specific scheduled times.

» Periodic policies can be daily, weekly, or monthly. They are designed for services that experience regular, prolonged peaks or troughs.

There are two types of auto scaling: vertical and horizontal. Vertical scaling involves upgrading or downgrading specifications, while horizontal scaling involves increasing or decreasing the number of nodes.

Auto scaling can be applied to various objects such as EIP bandwidth, ELB backend nodes, CCE workload, ECSs, and nodes in advanced service clusters.

Elastic scheduling is achieved through several capabilities. These capabilities include scheduling control and rule orchestration; metric collection using Application Operations Management (AOM), Application Performance Management (APM), and Cloud Eye; and elastic scheduling based on capacity and performance evaluation.

Container Auto Scaling Practices

Monitoring: We monitor CPU usage, memory usage, custom metrics, timer triggers, and AI triggers.

Calculation: We set thresholds for each scaling condition, such as a period during which a threshold needs to be consistently hit. For example, we trigger scale-out when the CPU usage reaches 80% or memory usage reaches 85%, and scale-in when both CPU and memory usage decrease to 50%.

Decision making: We control step

adjustments, how many nodes or pods to be removed or added at one time.

Execution: Auto scaling needs to be implemented gracefully to minimize impacts on services.

Example of Container Auto Scaling under Faults

Imagine a tenant deployed two Cloud Container Engine (CCE) clusters, one in AZ 1 and one in AZ 2. And there is a scheduling system working at the network layer to evenly distribute traffic between the two clusters. At one point, the node 4 and node 5 in AZ 1 become faulty. The traffic was then redirected to the pods of the remaining normal nodes, but this created an excessive load on those pods. To balance the load, the team decided to add two more nodes. However, the attempt to scale out in AZ 1 failed for some reason, so the two nodes were added to AZ 2 instead. Unfortunately, this did not effectively decrease the load on each pod, so two more nodes were added to AZ 2 after the failed scale-out in AZ 1. As a result of two consecutive scale-outs, the service traffic was distributed in a ratio of 3:7 between the two availability zones, controlled by the upper-layer scheduling system. This example demonstrates how auto scaling makes it easier to automatically redirect traffic between different AZs when there are faulty nodes.

Other Scaling Capabilities

Node Scaling

Auto scaling creates new VMs during scale-out, which can take a few minutes. AS is ideal for services that have regular daily peaks and troughs in demand.

- Alarm-based scaling works in real-time based on alarms you can configure for resources, such as CPU, memory, disks, and network.
- Scheduled scaling is precise down to the second.
- Periodic scaling can be configured on a daily, weekly, or monthly basis.
- Horizontal Pod Autoscaling (HPA) enables horizontal pod scaling for Cloud Container Engine (CCE).

Public Network Bandwidth Scaling

Bandwidth auto scaling is implemented based on Elastic IPs (EIPs). We can also associate elastic load balancers with EIPs to balance workloads. In addition, we can modify EIP bandwidth specifications on Huawei Cloud whenever needed.

Scaling for Elastic Load Balance (ELB)

We can modify the specifications of elastic load balancers.

Scaling for Distributed Cache Service (DCS)

Replica scaling can be implemented on active/standby, cluster, and read/write splitting instances.

Shard scaling is supported by both Proxy Cluster instances and Cluster instances.

Scaling for Distributed Message Service for Kafka

- Broker instances can be scaled out, but not scaled in.
- The storage space of Broker nodes can be scaled up, instead of being scaled down.
- Service specifications can be upgraded or downgraded.

Scaling for Relational Database Service (RDS)

- RDS supports primary/secondary deployment and read/write splitting. Read replicas can be scaled in or out.
- We can adjust the specifications of database instances, such as upgrading or downgrading CPUs, memory, and disks.
- Storage autoscaling is also supported.

Performance Protection

Benefits

- Detection: Our monitoring system provides information about system overload and reports alarms to help identify overload sources.
- Traffic control: Services are allowed to reject traffic if performance thresholds are surpassed to prevent service breakdowns. It is important to block excessive traffic when systems are already overloaded and unable to process more demand
- Intervention: We can monitor traffic at different levels, including tenant, function, resource pool, and cluster levels. In case of excessive or sudden changes in traffic, we can manually or automatically control the traffic.
- Self-healing: After the traffic returns to a normal level, service functions can quickly and automatically recover.

Key Technologies

- Load shedding refers to controlling how much load a system needs to process at a given time frame. The goal is to keep an overloaded system operational. This means when the load in a system reaches a certain threshold, excessive requests are rejected. Load shedding can be implemented on the client or server end.
- Graceful degradation ensures that the core services remain functional even if secondary features perform poorly.
- Circuit breakers interrupt traffic flows to restore services. If a dependent component malfunctions, a circuit breaker isolates the component to prevent any related faults from spreading. Overload circuit breakers are more commonly used.

Fault Tolerance and Blast Radius Control

Load shedding, graceful degradation, and circuit breakers are critical in all fault scenarios. These measures allow us to implement a recovery plan without having to locate faults. They can help us reduce the time to recovery and satisfying SLAs. In addition, we have enabled comprehensive fault tolerance for baseline scenarios. We also closely monitor faults on each service flow to prevent them from affecting the entire service chain and causing service breakdowns.

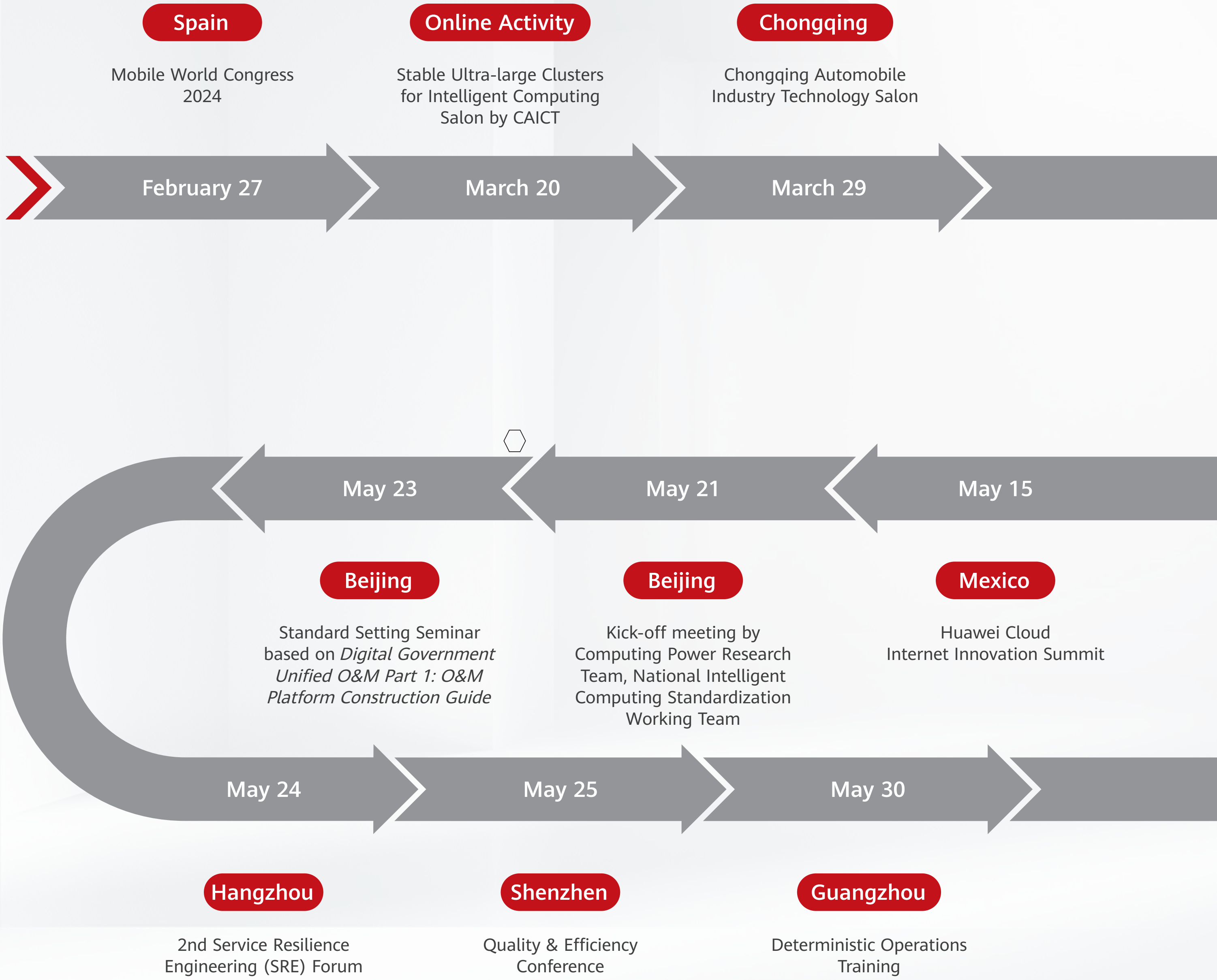
The following capabilities are required for comprehensive fault tolerance:

- Responding to emergencies: Emergency capabilities can be established through simulating faults. By imagining faults for every service module, we can proactively create and execute emergency plans to practice addressing faults, rather than simply reacting to them when they occur.
- Minimizing blast radii: To reduce the impact of failures, it is essential to divide an application into separate components, considering both performance and fault isolation principles.
- Resolving faults: To minimize downtime, it is important to rely less on the expertise and skills of O&M personnel and instead develop proactive measures for handling faults.
- Recovering services: We must be able to use three key measures of fault recovery: load shedding, graceful degradation, and circuit breakers. We also need to continuously improve emergency plans to enable fast service recovery.
- Designing emergency plans: To effectively handle emergencies, we have to design response processes that align with service requirements and establish a collaborative team. Additionally, we need to establish an intelligent platform to facilitate emergency handling.
- Developing prevention measures: To effectively prevent faults, we need to focus on improving visual monitoring, optimizing traffic control strategies, and validating the effectiveness of emergency drills.

Summary

The CRE team has developed key technologies for enhancing architecture resilience. They employ best practices and effective methodologies to help customers design HA architectures that meet requirements of various service scenarios. They specialize in designing HA architectures, precisely managing capacity to minimize costs, and ensuring best possible service performance. They also employ three key measures to recover services, design predicative solutions to control risks, and help satisfy SLAs. They are dedicated to ensuring secure, reliable cloud migration and adoption.

Key Operations Events in 2024



February 27

Spain

Mobile World Congress 2024

March 20

Online Activity

Stable Ultra-large Clusters for Intelligent Computing Salon by CAICT

March 29

Chongqing

Chongqing Automobile Industry Technology Salon

May 15

Mexico

Huawei Cloud Internet Innovation Summit

May 21

Beijing

Kick-off meeting by Computing Power Research Team, National Intelligent Computing Standardization Working Team

May 23

Beijing

Standard Setting Seminar based on *Digital Government Unified O&M Part 1: O&M Platform Construction Guide*

May 24

Hangzhou

2nd Service Resilience Engineering (SRE) Forum

May 25

Shenzhen

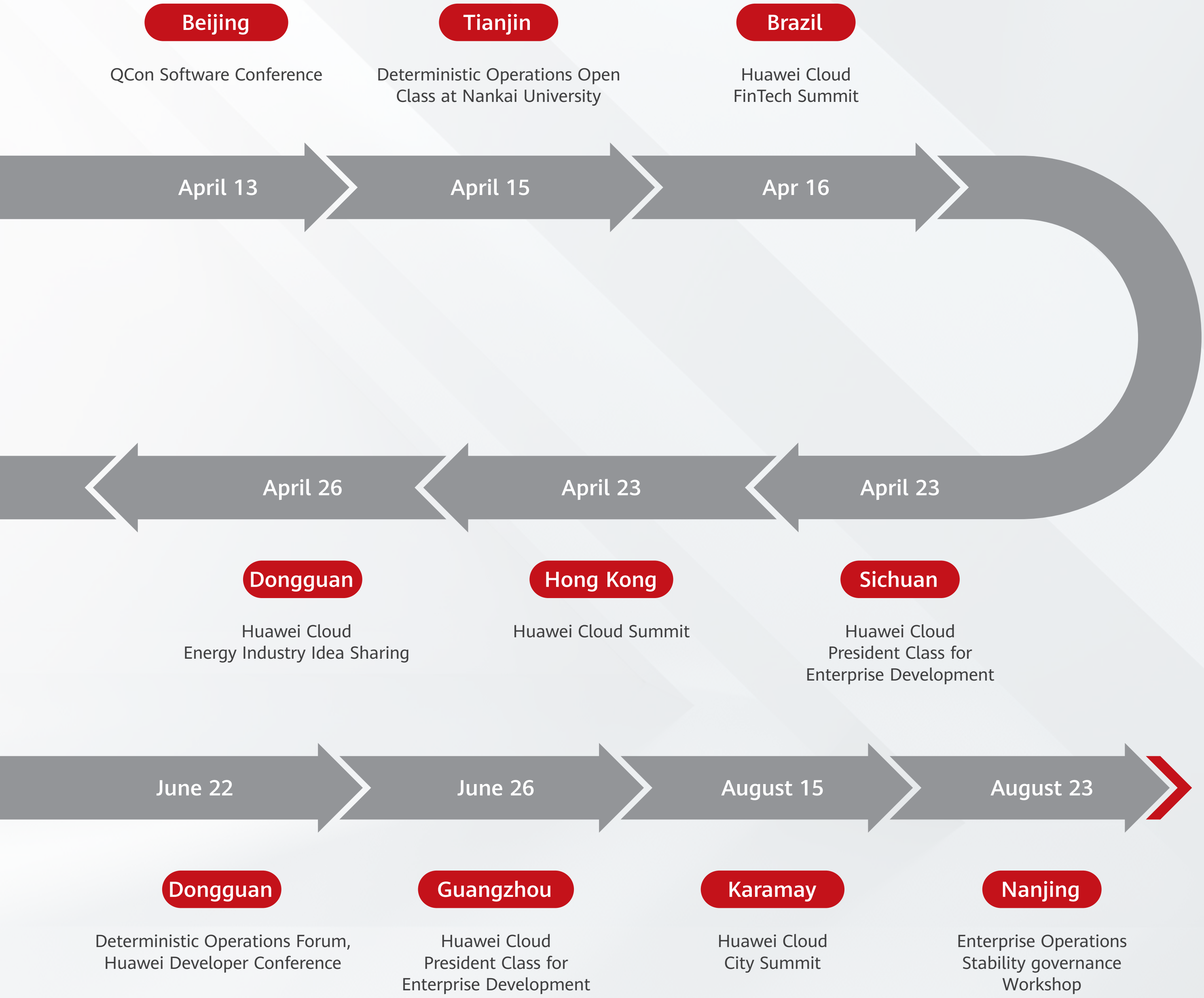
Quality & Efficiency Conference

May 30

Guangzhou

Deterministic Operations Training

In 2024, the Deterministic Operations Elite Club held several salons where industry experts came together to discuss the innovative application of AI technologies. We also explored ideas and practices inspired by Huawei Cloud Deterministic Operations practices. Through the Deterministic Operations system and solutions, we encourage service innovation on the cloud and facilitate operations transformation for enterprises. Our goal is to help enterprises better manage the cloud and safeguard stable, sustainable development.



April 13 – Beijing: QCon Software Conference

April 15 – Tianjin: Deterministic Operations Open Class at Nankai University

Apr 16 – Brazil: Huawei Cloud FinTech Summit

April 23 – Sichuan: Huawei Cloud President Class for Enterprise Development

April 23 – Hong Kong: Huawei Cloud Summit

April 26 – Dongguan: Huawei Cloud Energy Industry Idea Sharing

June 22 – Dongguan: Deterministic Operations Forum, Huawei Developer Conference

June 26 – Guangzhou: Huawei Cloud President Class for Enterprise Development

August 15 – Karamay: Huawei Cloud City Summit

August 23 – Nanjing: Enterprise Operations Stability governance Workshop

We offer a range of measures to meet the requirements and address pain points of the automotive sector. These measures include HA architecture design, observability diagnosis, enhancing capabilities for emergency recovery, traffic control, and capacity management to safeguard Internet of Vehicles (IoV) services. We also provide experience and technologies on establishing Deterministic Operations platforms and transforming from the traditional maintenance to IT platform-based operations.

In the Deterministic Operations open class at Nankai University, we delved into how to establish O&M systems and capabilities. We explored advanced technologies of intelligent O&M, discussed how to cultivate skilled SRE personnel, and shared a goal of cultivating talent with universities.

- "Establishing Deterministic Operations System and Capabilities" by Li Kaiguang, Huawei Cloud SRE intelligent O&M architect, on April 15, in Tianjing

Automobile Industry Technology Salon in Chongqing

Deterministic Operations Open Class at Nankai University

We share case studies where AI generated content (AIGC) was used in the O&M domain. We also draw from practical experience with Deterministic Operations and large language models (LLMs) to promote a multi-agent collaboration O&M solution. We want to inspire ideas and exploration of intelligent O&M in the era of foundation models.

- "LLM and Multi-Agent Collaboration for the O&M Domain" by Zhang Xi, Huawei Cloud SRE AI enablement expert, on April, 13 in Beijing

Deterministic Operations integrates capabilities of foundation models to comprehensively ensure high-quality development of enterprises with a one-stop dedicated cloud O&M management system. We offer observability solutions to help quickly troubleshoot problems. Our goal is to enable various industries to better manage the cloud and accelerate digital transformation.

- "Practical Experience Ensuring Secure, Stable, High-quality Services" by Li Wei, Huawei Cloud professional service expert, on April 23 in Chengdu

QCon Software Development Conference 2024 in Beijing

Huawei Cloud President Class in Sichuan

The Deterministic Operations system combines Huawei Cloud's expertise with O&M hands-on experience. By using predictive recovery and verification, we successfully transformed the O&M team from a cost-focused department to one focused on. This transformation has enabled a transformation from firefighters reacting to emergencies as they happen to proactive strategists, ensuring the secure, stable, and high-quality development of cloud services.

- "Achieve Higher Availability with Deterministic Operations" by Lin Huading, Director of Huawei Cloud O&M Enablement Center, on April 26, in Dongguan

This forum attracted hundreds of developers from around the world. We discussed how digital transformation reshapes productivity and why Deterministic Operations was important to digital transformation. We also explored how AI capabilities can be used to drive the transformation. We would like to express our gratitude to the experts from academia and enterprises. They shared the latest achievements and methods in the O&M field that we have integrated into *Deterministic Operations White Paper – Stability and Reliability 2.0*. We are committed to collaborating with every industry to foster continuous innovation.

- Deterministic Operations Forum, Huawei Developer Conference 2024, on June 22 in Dongguan

Release of Deterministic Operations white paper

Deterministic Operations forum

Huawei Cloud Deterministic Operations offers a security compliance framework to ensure the stability, data security, quality, reliability, and operational trustworthiness of our services. This framework helps us comply with local laws and regulations, making data processing more secure and facilitating service expansion across borders. In addition, Deterministic Operations provides hands-on experience and application hosting services to help improve application reliability, helping services going global in a secure, efficient manner.

- "Compliance and Reliability Practices of Chinese Companies Going Global" by He Yuan, Huawei Cloud SRE Senior Expert, on May 30 in Guanzhou

Deterministic Operations Training

Huawei Cloud works with customers to innovate using AI technologies. We leverage AI native engines to drive innovation in Deterministic Operations. By deeply integrating AI technologies into O&M, we provide enterprises with a new and efficient O&M system that caters to the diverse needs of different industries. We are dedicated to helping our customers accelerate intelligent O&M decision-making and digital transformation.

- President Class for Enterprise Development on June 26, in Guangzhou

President Class in Guangzhou

We acquired hands-on experience and make some impressive achievements ensuring cloud security and stability for Karamay. Our Deterministic Operations capability system and solutions, combined with a stronger focus on prevention, as well as various measures, including both proactive prevention and effective recovery help ensure secure, stable digitalization of Karamay.

- "Deterministic Operations and Security Services Ensure Security and Stability for Karamay" by Kang Zhen, Huawei government cloud O&M director, on August 15, in Karamay

电力
业务深度云化
挑战平台可用性及业务稳
架构可靠平台稳定，不出故障
关键业务越来越多，业务稳定性要求越来越高
自然灾难、网络过载、维护操作等各种因素
社会后果
加速智能升级 共建美好新疆
2024华为云城市峰会 · 克拉玛依

Huawei Cloud City Summit in Karamay

To help customers address challenges caused by fast service growing in various scenarios, such as smart stores and service going global, we offer methods that have been proven by a wealth of Deterministic Operations hands-on experience to improve fault management capabilities. These methods include identifying failure modes, establishing fast recovery capabilities, and effective verification. They can help customers improve system resilience and fault management capabilities.

- "Fault Management Practices" by Wang Hong, Huawei Cloud Deterministic Operations fault management expert, on August 23 in Nanjing

O&M stability workshop

Highlight Moments

Li Guo, Huawei Cloud SRE expert, at Mobile World Congress 2024

Wang Zhangyu, Director of Huawei Cloud Latin America SRE, at FinTech Summit in Brazil

Tong Lin, Huawei Cloud SRE senior expert, at Huawei Cloud Summit in Hong Kong

Zheng Lei, Huawei Cloud Application Service SRE Director, at the kick-off meeting by Computing Power Research Team, National Intelligent Computing Standardization Working Team in Beijing

Liu Tao, Huawei Cloud SRE senior architect, at the second Service Resilience Engineering (SRE) Forum in Hangzhou

Quality & Efficiency Conference in Shenzhen

Wang Hong, Huawei Cloud Deterministic Operations fault management expert, at the O&M stability workshop

Wang Yongheng, Huawei Cloud AI native engine architect, at O&M stability workshop

Secure | Reliable | Intelligent | Efficient | Agile

Huawei Cloud

Deterministic Operations Website

Deterministic Operations Elite Salon