

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

AI-driven Operations: Better Quality, More Efficiency

Intelligent O&M Models Driven by Both Data and Mechanisms

Professor Qin Tao, Department of Computer Science, Xi'an Jiaotong University

One-Stop AI Applications: Starting a New Chapter in Intelligent Industry Transformation

Wang Ju, Big Data Director, New Hope Real Estate


One-Stop E2E IDC Cloud Transformation

Hu Jianhua, Infrastructure and O&M Director, Shanghai Ximalaya Technology Co., Ltd.

Strategic Development of Predictive Operations Capabilities in China's Retail Sector: a Case Study of a Leading Convenience Store Brand

Wu Hongqin, Data IT System Operations Director and Technical Director, Meiyijia Holdings Co., Ltd.





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	Ding Xiaohong, Hu Jianhua, Li Heqing, Ma Tao, Qin Tao, Tang Runhong, Wang Ju, Wu Hongqin (sorted by alphabetical order)
Editors	Liu Jiarui, 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 >>>



Frontier Views

Intelligent O&M Models Driven by Both Data and Mechanisms	01
---	----



Expert Interview

One-Stop AI Applications: Starting a New Chapter in Intelligent Industry Transformation	05
One-Stop E2E IDC Cloud Transformation	12
Strategic Development of Predictive Operations Capabilities in China's Retail Sector: a Case Study of a Leading Convenience Store Brand	18



Technical Innovation

Staying Ahead in O&M with an Observability System	25
Empowering IT and Application System Operations with DeepSeek	30



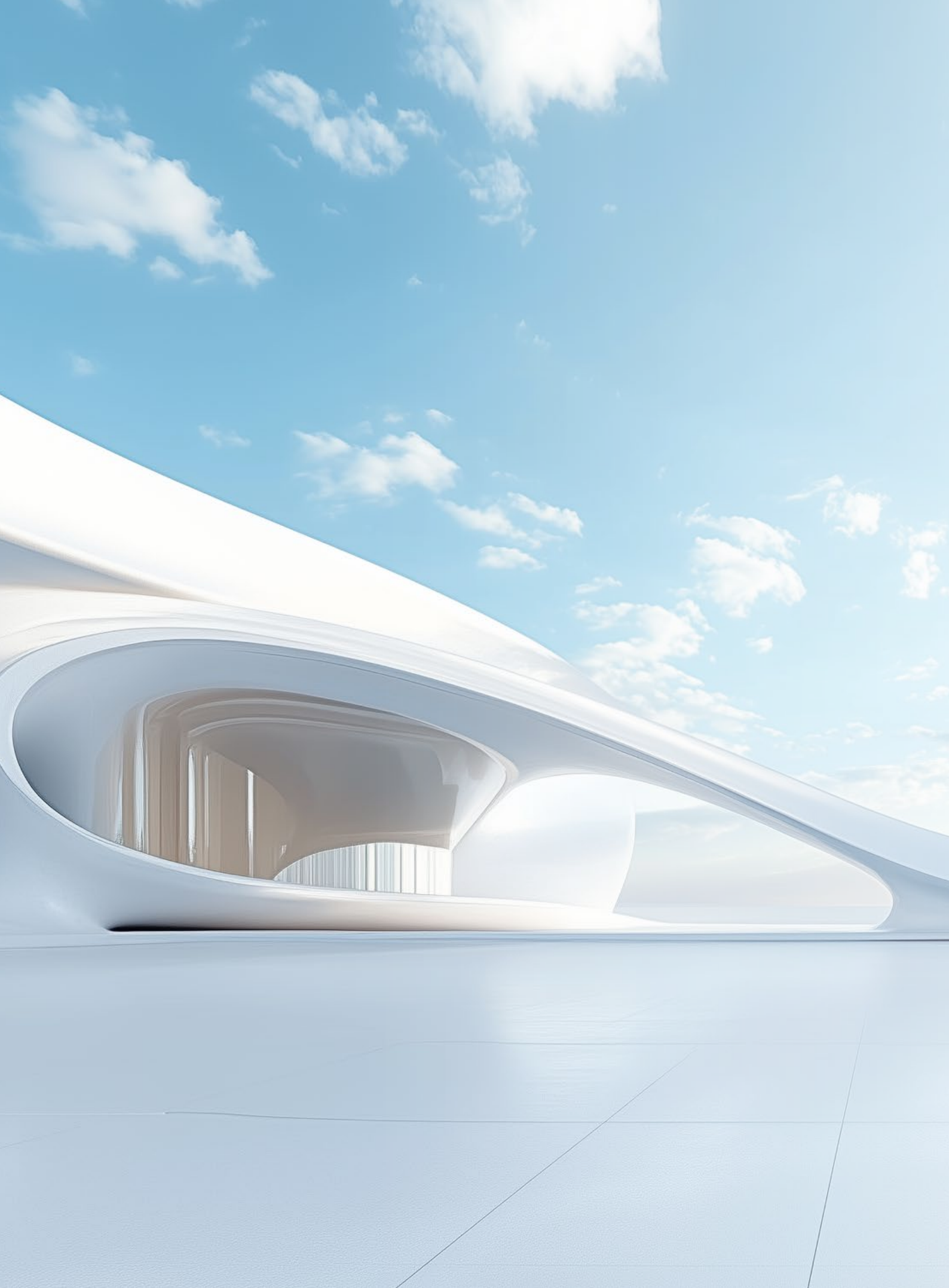
Service Improvement

Cloudification of Enterprises' Core Services: Experience in Availability Assurance	35
Digital Resilience in Chile's Massive Blackout – The Technology Behind Navigating a Nationwide Power Crisis	48



Operations Milestones

Huawei Cloud Credence Club: Key Events	51
--	----



“ Preface



Gao Jianghai
President of Huawei Public
Cloud Business Dept

Driving Intelligent Transformation 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 Credence 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.





Intelligent O&M Models Driven by Both Data and Mechanisms



Prof. Qin Tao
Department of
Computer Science, Xi'an
Jiaotong University

Abstract

Complex operations and maintenance (O&M) tends to drive up costs and reduce efficiency. To address this issue, the current trend is towards more intelligence. This article explores the key roles that data and mechanisms play in O&M model development. It highlights methodologies for optimizing O&M data retrieval and for depicting model mechanisms, while also addressing fundamental principles of model lightweighting.

01

New Characteristics of O&M Models in the Age of Intelligence

We have stepped into an era of intelligence, marked by the explosive growth of smart technologies — robots, autonomous vehicles, surveillance systems, and more. This rapid evolution relies on huge, high-performance computing systems composed of diverse, heterogeneous physical devices. To sustain this momentum, intelligent O&M must prioritize the stability of these huge systems, minimize their construction expenses, and pinpoint critical insights through comprehensive analysis. Enhancing the sophistication and recognizing the inherent value of O&M systems will be pivotal in driving this transformative wave forward.

When viewed through the lens of systems engineering, intelligent O&M stands apart from conventional approaches in three distinct ways:

1. Enhanced autonomy in system control behavior

Over the past two decades, researchers and industries have focused on improving the adaptability and generalization capabilities of complex systems. But today's intelligent applications demand even higher performance. Today's systems need proactive sensing capabilities, the ability to interpret dynamic and unstructured environments, and the agility and intelligence to make rapid, human-like decisions in the face of conflicting objectives.

2. Deeper engagement with all sorts of users

In modern IT environments, users are the primary drivers of system operations. To meet their diverse needs, intelligent O&M systems must cater to both untrained operators and skilled engineers. First, these systems should facilitate seamless communications with non-technical users by adopting natural and intuitive interfaces, while also leveraging advanced capabilities like intent recognition to anticipate unexpressed user requirements. Second, they must enable highly accurate retrieval from vast repositories of dynamic O&M data, ensuring streamlined and efficient interactions for developers and technical professionals.

3. Expanded integration into real society

As digital transformation accelerates, intelligent applications are growing increasingly intertwined with digital infrastructures such as social networks and electronic payment systems. These platforms replicate the intricacies of the physical world, blurring the boundaries between virtual and tangible realms. As these platforms grow in scale and influence, it becomes imperative for intelligent O&M strategies to not only optimize functionality but also uphold significant social responsibilities. This dual focus is essential for fostering social harmony, driving economic growth, and promoting collective well-being.



02 Challenges to Data & Mechanism Driven O&M

O&M data — comprising audio-visual content, images, and text — serves as a repository of historical insights for diverse systems, brimming with operational knowledge and serving as the bedrock for developing robust machine learning models. By leveraging data-driven approaches, these models can uncover potential patterns hidden in vast datasets, enabling accurate simulations of intricate environments and system dynamics. However, the sheer scale of O&M data, and its multimodal nature, make unlocking its

full potential and maximizing its utility a pressing challenge for intelligent O&M.

In this context, mechanisms generally refer to the system mechanism data derived from service constraints, deployment constraints, functional interdependencies, and related factors. These mechanisms serve as the source of operational rules. Rule constraints and augmentation are key approaches for refining machine learning models and boosting their adaptability across varied scenarios. Crafting operational models that seamlessly

integrate with these mechanisms remains a significant hurdle in advancing intelligent O&M practices.

Finally, lightweight implementation is key to deploying models for practical applications. Efficient deployment strategies — encompassing techniques such as input data compression, model size optimization, and model deployment acceleration — are essential for ensuring the practicality of these models in operational environments for on-the-ground use.

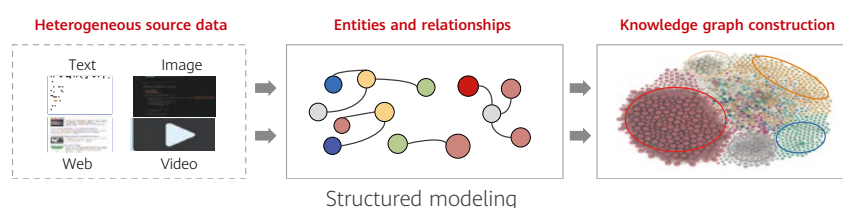
03 Data-Driven: Better Knowledge Utilization with Enhanced Data Retrieval

Data drives intelligent O&M models by providing valuable insights. Beyond the classic 4Vs — Volume, Velocity, Variety, and Veracity — this data embodies complexity as it often comes from multiple sources, changes over time, and reflects inherent heterogeneity. This data can be used to construct high-quality domain-specific knowledge bases for O&M. When combined with generation techniques powered by external knowledge retrieval, the data enables autonomous design of operational strategies, elevating O&M intelligence to new heights. However, the diversity of O&M data presents challenges. O&M data spans structured formats like databases, logs, and reports to unstructured, multimodal inputs such as audio-visual content and images. Processing this data requires structuring it first and then retrieving relevant information.

1. Structured modeling

O&M data comprises analyzed and validated expertise. It serves as a critical foundation for resolving recurring issues. Nevertheless, such knowledge often exists in fragmented forms, requiring systematic structuring and correlation analysis to better guide O&M decisions. As efficient O&M increasingly hinges on advanced data analytics, there is growing demand for algorithms capable of performing structured modeling and correlation analysis on extensive, fragmented expert knowledge.

During the knowledge extraction phase, it is essential to address both inherent data characteristics — such as multi-sourced origins, temporal variability, and heterogeneity — and challenges posed by limited sample sizes. Common strategies include entity alignment for multi-source integration, incremental or continuous learning for handling dynamic changes, cross-modal learning for managing heterogeneity, and various techniques tailored to small-sample scenarios. With their exceptional content aggregation and generation capabilities, large models have emerged as potent tools for overcoming these complexities.



In the structured modeling phase, knowledge graphs are often used to systematically organize the extracted knowledge, thereby improving data usability. Given the sheer volume of O&M experiential data, knowledge graphs should adopt a progressive presentation strategy to accommodate diverse user requirements across varied levels of expertise. Structuring knowledge at three distinct granularities — points, lines, and surfaces — enables more effective modeling. Points concentrate on contextual information surrounding individual knowledge elements; lines illustrate the causes and effects of events; and surfaces explore intricate connections between multiple events.

2. Knowledge retrieval

External knowledge bases are essential for enhancing the performance of machine learning models in vertical domains, particularly in specific scenarios where retrieval tasks need to handle diverse forms of data. However, current large model-based data retrieval systems suffer from poor indexing between different data modalities and inaccurate semantic matching during queries, often

resulting in critical semantic omissions or redundant information in associative retrievals. Consequently, achieving precision when querying vast O&M datasets is another challenge in effective data utilization. A novel retrieval approach integrating semantic segmentation, multi-dimensional image feature storage, and small-to-big retrieval can address the innate drawbacks of conventional Retrieval-Augmented Generation (RAG) methodologies, which are constrained by fixed text chunk sizes and suboptimal sorting. Semantic segmentation ensures intra-chunk semantic consistency while maintaining distinct inter-chunk boundaries, thereby mitigating contextual semantic degradation during segmentation. Meanwhile, multi-dimensional image feature storage facilitates multimodal data processing, harmonizing semantic coherence with visual fidelity across disparate modalities during retrieval operations. Additionally, small-to-big retrieval ensures high recall rates for O&M knowledge retrievals. Collectively, these advancements enable more accurate knowledge retrieval and significantly enhance O&M efficiency.



Efficient utilization of multi-modal O&M data

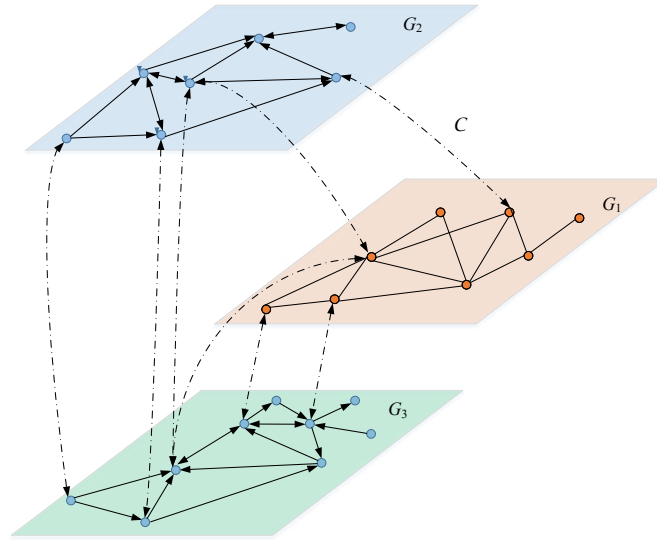
04 Mechanism-Driven: Multi-Layer Coupled System Operations

As huge systems grow more and more complex, subjects start to demonstrate distinct modular and hierarchical evolutionary patterns. Effective coordination across the various modules and layers serves as a foundational approach to optimizing system performance. Meanwhile, constructing an abstract system model has become a prerequisite for gaining deeper insights into huge systems and developing efficient O&M models. In contrast to monolithic software systems, huge systems often incorporate diverse operational mechanisms dispersed across multiple modules or layers. Given the structural and device heterogeneity among these components, the interactions and dependency strengths between modules and layers exhibit significant variability. To address this issue, such systems can be effectively represented using a multi-layered coupled abstraction model, where macro objectives are defined across layers and micro goals are delineated within each individual layer.

Intra-layer dependency characteristics can be effectively modeled as directed graphs, where nodes represent individual functional modules or corresponding microservices, and edges denote their mutual dependencies. The exchange of information or data among

these nodes reflects different dependency patterns, which are primarily manifested through latency metrics that indicate various evolutionary modes. These intra-layer dependencies can further be quantified using matrices, facilitating evolutionary mode identification via matrix analysis. Similarly,

inter-layer dependency characteristics can also be represented using directed graphs, with the same graph sketching methods as those used within layers. Notably, established theoretical frameworks from graph theory offer robust analytical instruments for comprehensive mechanism analysis.



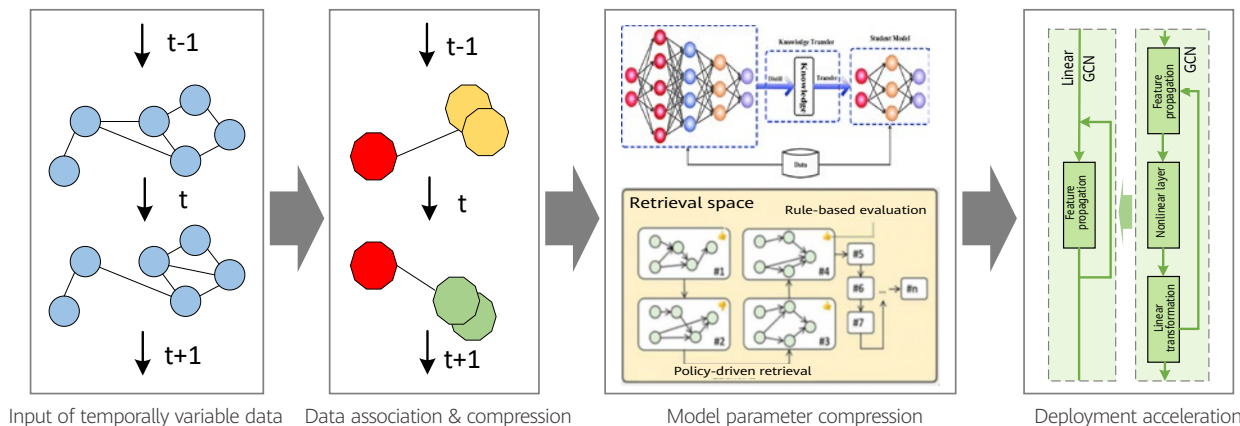
Multi-layer coupled system modeling

05 Model Lightweighting: Prerequisite for Practical Application

Model lightweighting enables efficient computation of results for complex multimodal data or dependency graphs. It specifically tackles challenges such as speeding up the processing of dynamically updated datasets and large-scale graph structures, thereby enhancing the real-time operational stability of O&M systems.

Model lightweighting requires three types of optimizations: the model itself, model inputs, and model deployment. For models with excessive parameters and structural complexity, parameter compression techniques should be integrated during design. To mitigate issues arising from sparse

and expansive directed graphs, it is essential to consider dependencies between different types of data during computation, enabling effective input data compression. Finally, the characteristics of models' computation processes can be leveraged to accelerate model deployment.



Model lightweighting



06 Summary

In the field of intelligent O&M, new characteristics represented by increased control autonomy, enhanced user interaction, and wider societal integration require that O&M system design integrates the advantages of both data-driven and mechanism-driven approaches. On the data-driven side, structured modeling and knowledge retrieval can address challenges posed by the diverse sources, temporal variability, and inherent heterogeneity of O&M data, ultimately optimizing the data

utilization. On the mechanism-driven side, leveraging the multi-layer coupled system operation mechanisms allows for analyzing dependencies between complex system modules and layers. It fosters a deeper understanding of system operational rules. Model lightweighting is also fundamental for practical application, which can be approached from three directions: the model itself, the inputs, and the deployment. This process ensures the computational efficiency and real-time stability of O&M systems.

The fusion of multimodal O&M data with the operational mechanisms of huge systems represents a viable solution for improving the intelligent O&M of these complex systems. As technological innovation progresses, this integrated methodology is poised for refinement and broader adoption across academic research and industrial applications. Such advancements will propel the evolution of intelligent O&M systems, empowering diverse sectors to achieve transformative O&M milestones and significantly enhancing O&M efficacy and performance.

One-Stop AI Applications: Starting a New Chapter in Intelligent Industry Transformation



Wang Ju

Big Data Director, New
Hope Real Estate

Abstract

Enterprise AI transformation faces challenges in asset sharing and reuse and the re-enablement of legacy systems. To address them, we built a central AI management platform. Aided by this platform, we developed an intelligent Q&A app and a data analytics & reporting app. They automate data queries and analytics, improving efficiency and decision-making quality. Additionally, we acquired valuable expertise in building high-quality datasets and choosing the right models. They solidified our determination to become a future AI leader, and to facilitate the intelligent transformation and diversified development of different business units within the Group.

01 Background

AI is becoming an important driving force behind enterprises' digital transformation. Over the past few years, we have tried to integrate AI technology into various business units at New Hope, and have given each business unit full autonomy to choose their own AI models and applications. However, as time goes by, we are facing the following challenges:

1. AI asset sharing

- » Each business unit has accumulated their own AI assets.
- » All business units will need to pool their assets together for better sharing and reuse, and to accelerate AI application development.

2. Re-enablement of legacy systems

- » The Group spent a lot of money building up digital systems that are currently in use. We cannot simply tear them down and build new AI systems from scratch.
- » AI needs to empower existing systems and enhance their capabilities.

To address these challenges, we decided to count and consolidate AI assets across the Group and build a central platform to manage them, ensuring that we can effectively utilize AI technology while keeping the cost under control.



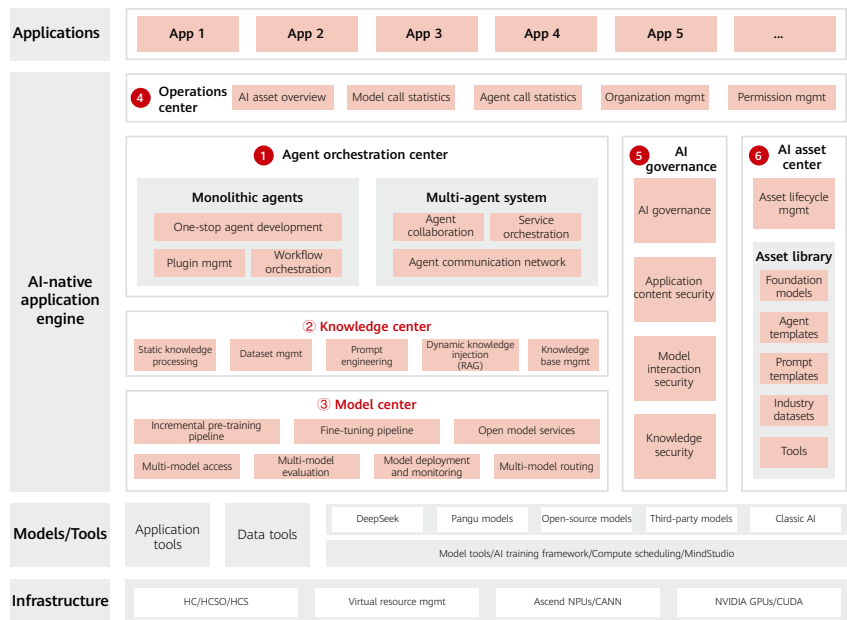
02 Solution

1. Pilot selection: real estate first

The real estate business already has laid a solid digital foundation, as most of its services are already online, and it has accumulated considerable knowledge assets. For this reason, we selected the real estate business as the pilot for our AI project. The project included an AI management platform along with two key applications: an intelligent Q&A system and a business data analytics & reporting tool. Leveraging Huawei Cloud's AI native application engine, we designed our own AI management platform and integrated the latest DeepSeek models.

The AI platform consists of an agent orchestration center, knowledge center, model center, operations center, AI governance framework, and AI asset center.

- » Agent orchestration center: Coordinates the deployment, collaboration, and management of AI agent applications, with an architecture capable of handling millions of agents, allowing users to quickly build agent-powered applications; improves collaboration between AI agents and human employees to unleash the full potential of LLMs.
- » Knowledge center: Efficiently builds standard datasets through processes such as dataset management, static data processing, and document parsing;

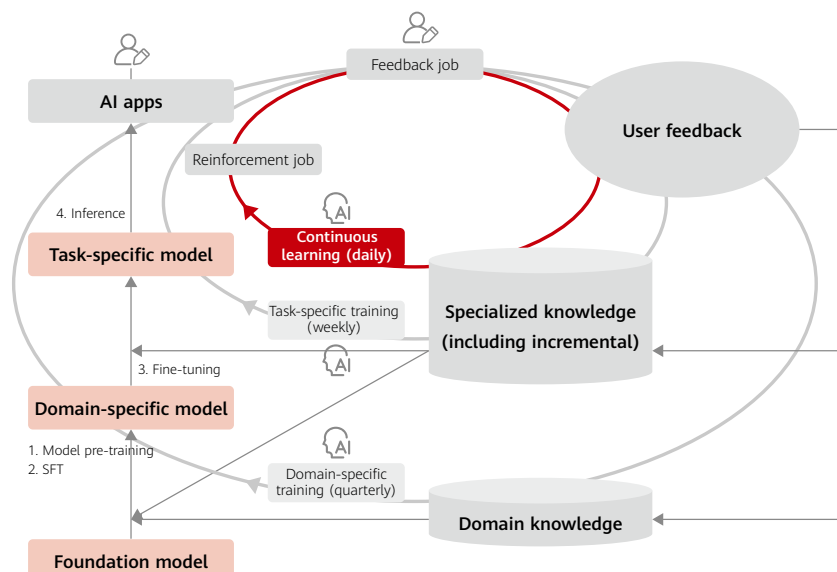


designs, refines, and optimizes prompts through prompt engineering; integrates domain-specific knowledge through retrieval augmented generation (RAG); enhances LLM output reliability by enabling access to multiple knowledge bases via a flexible search framework, coupled with a high-performance vector search engine. builds enterprise knowledge bases; creates a virtuous cycle of knowledge, models, and applications, and continuously improves AI performance.

» Model center: Presets 40+ of the world's most popular LLMs. All the models have been thoroughly evaluated based on industry benchmarks and datasets across diverse use cases, with joint validation by AI and domain experts. Model routing ensures that the most suitable model is selected for each task, ensuring optimal performance.

» Operations center: Intuitively shows the status of AI applications (datasets, models, services, and agents) by organization, as well as their resource consumption and costs (external tools, model APIs, tool APIs, and internal computational power). Sets upper limits and alarm policies for expenditures to control costs. Creates a permission management system to facilitate AI platform operations.

» AI governance: Builds an application-centric, multi-layer security system that encompasses data, models, and content. Enhances dataset security through data asset security management; controls the model interaction process to prevent data theft; filters model output to ensure content security. These measures build a "wall" between core enterprise applications and AI applications, lowering the risk of data breaches and compliance violations.



» AI asset center: Integrates AI assets from both internal and external partners to enable better sharing and reuse.

The application of AI technology is guided by practical use cases, and the delivery of AI use cases relies on the AI management platform. In practice, we focused on two AI apps that are particularly useful for the real estate business: intelligent Q&A and data analytics & reporting.

2. Apps: intelligent Q&A and data analytics & reporting

Intelligent Q&A system: intelligent query of real estate sales data

The intelligent Q&A system enables users to query sales data, such as conversion rate, contract amounts, and the number of first visits and revisits, in natural language. It now covers more than 200 reports across various business domains. Its key functions include:

» Natural language parsing and SQL generation: Supports single-table query, multi-table association query and complex condition filtering, meeting

diverse analytical requirements from different departments.

» Statistical analysis and visualization: Displays query results in bar charts, pie charts, and trend charts. Accurate in-context understanding and multi-turn dialog enable a seamless interactive experience.

On the intelligent Q&A system, users can ask questions in natural language and get intuitive responses from the system in no time. To improve accuracy, we embedded high-quality data dictionaries into this system. They describe database tables, fields, their relationships, business rules, and data samples in detail. We have made the following enhancements:

» Intelligent Q&A on basic data tables from functional departments (sales, finance, operations, etc.): Simplifies data query and analysis for frontline teams. Basic capabilities such as single-table query, multi-table association query, and data grouping and statistics are supported.

» Trend and risk analysis for functional departments: To enable functional

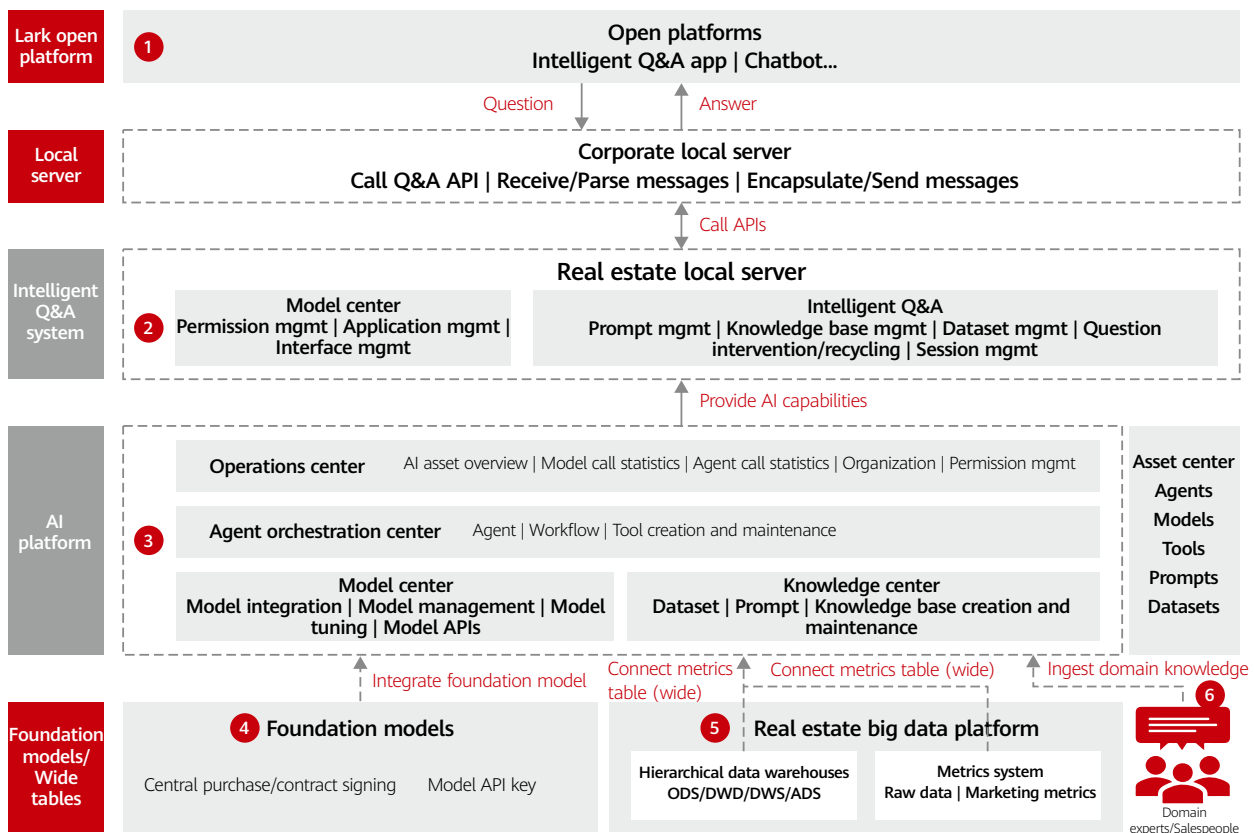
departments to make accurate, informed decisions, the system generates custom analytical reports periodically. For example, for the sales department, the report may provide numbers about visitors and visits, which provide useful information on sales trends and risks.

» Database access: The system is connected to multiple databases to perform query, analytical, and reporting tasks.

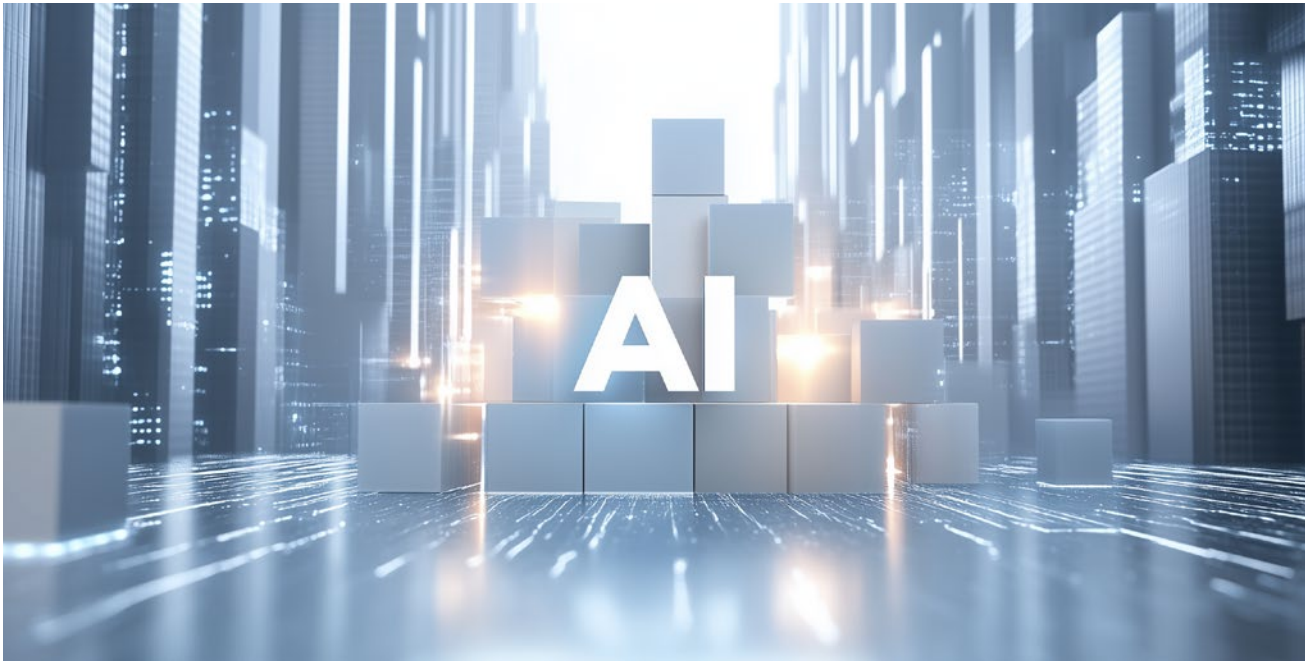
» System integration: Open APIs that support multiple sessions, multi-turn dialog, OBS-based reporting are provided, along with the necessary access control and authentication mechanisms.

» Permission management: Determine the scope of data query and analysis based on data access permission and control policies, preventing unauthorized data access.

To improve performance and accuracy, we have integrated the expertise and experience of domain experts and sales personnel, as well as external AI models



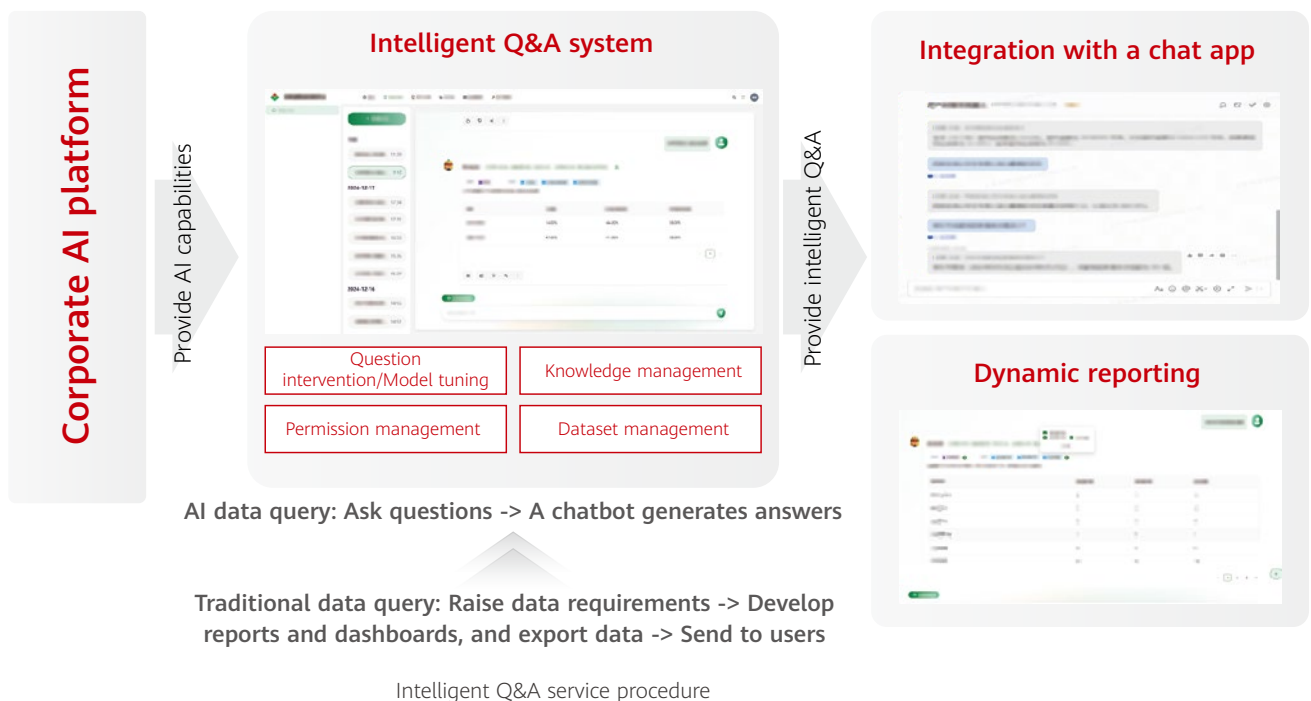
Intelligent Q&A system architecture

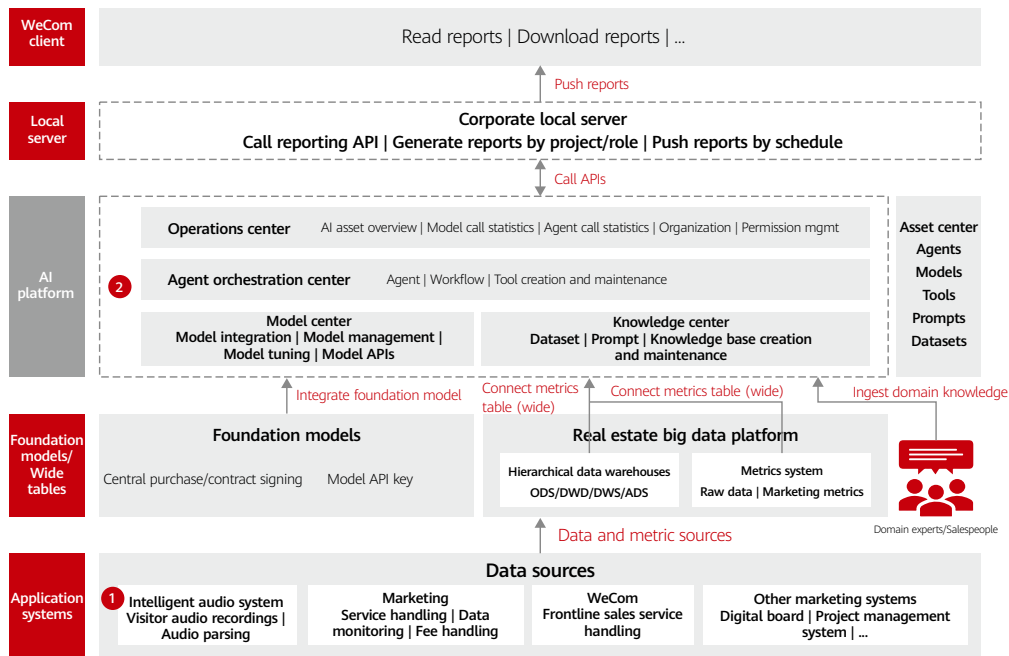


and datasets into our AI management platform. Through knowledge and data management, we have integrated the intelligent Q&A system with our workplace chat app. The intelligent Q&A bot can empower frontline personnel by predicting sales and generating custom reports.

The traditional data query process is complex. Typically, after a business user submits a request, colleagues in charge of the data platform or backend developers need to develop the required report, export data, and create data dashboards before delivering them to that business user. Today, business users simply ask the intelligent Q&A bot, for

example, "how many units are sold this month," and the bot gives an answer in real time. Additionally, the bot can generate dynamic, custom reports based on specified metrics and data dimensions. This architecture is lightweight and efficient. It fully leverages the capabilities of the AI platform to boost data query and analytical efficiency.





Data analytics & reporting solution

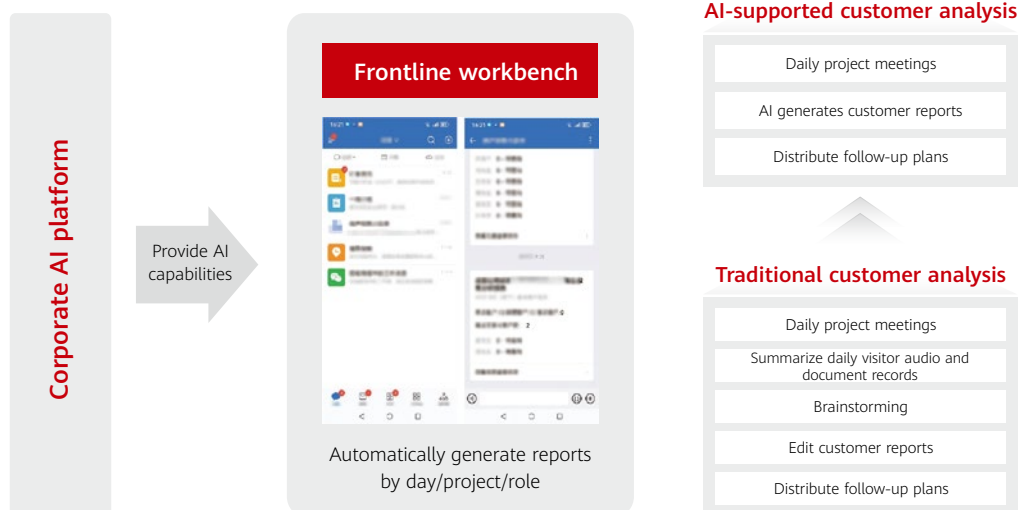
Data analytics and reporting - intelligent customer analysis

The data analytics & reporting app generates daily customer analysis reports for property agents and pushes the reports through the enterprise sales workbench app. The report covers information such as the number of visitors, first-time visitors, revisits, identified high-intent customers, and follow-up suggestions.

Core process:

- » Ingest and consolidate customer behavior data and transaction records from different systems.
- » Create a model for predicting high-intent customers to accurately narrow down target customer groups.
- » Automatically generate reports that contain customer segmentation, customer behavior features, and marketing & promotion suggestions.

A range of data visualizations, such as distribution charts, funnel charts, and customer journey maps, deliver actionable insights for marketing decision-making. Reports can be automatically generated and downloaded, enabling efficient query and analysis. These features significantly improve data utilization. They also lay a solid foundation for more future use cases.



Data analytics & reporting procedure



In the past, sales teams had to manually analyze visitors' audio recordings and document records every day. Now, AI automates this process by extracting key insights from audio recordings, identifies

high-intent customers and critical transaction milestones, and sends daily reports to property agents.

This significantly improves sales efficiency. This AI use case is highly specialized—

automating a single yet time-consuming task, but it is not insignificant, as it significantly improves sales efficiency and performance.

03 Takeaways

While implementing the AI management platform and use cases above, we learned a few valuable lessons and gained significant operational insights. The key takeaways are as follows:

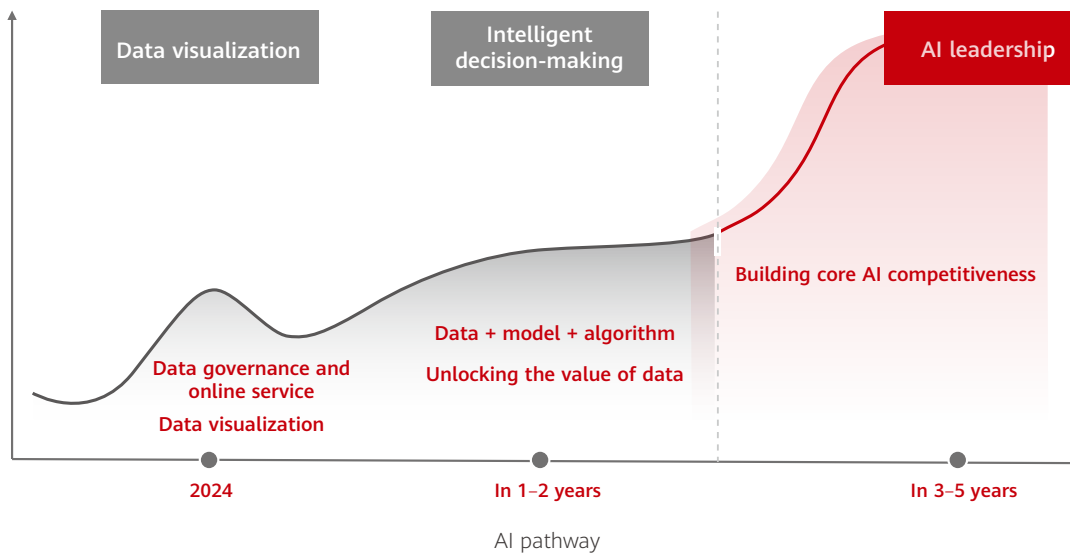
- » Ensure data quality: AI systems are only as good as the data they're trained on. Clean data—accurate, consistent, and properly structured—is the foundation of reliable AI performance. Dirty data, on the other hand, will make it difficult to reuse workflows. This is why significant efforts need to be devoted into normalizing knowledge bases and datasets, as it lays a solid foundation for efficient model training and tuning.
- » Select the right models: A single AI model is unlikely to meet diverse service needs. Take intelligent Q&A as an example. To build this app, you need the voice Q&A model and metric calculation model to work together, and you need to balance the performance

and costs of these models. While some models can achieve reasonably high accuracy, they need 20 to 30 seconds to generate a response. This makes them unsuitable for real-time chat applications where immediate interaction is critical.

- » Involve all stakeholders in model evaluation: To ensure that models are evaluated comprehensively and objectively, both the model developers and end-users must participate in the evaluation process. This ensures that models can be further optimized and reinforced based on unbiased user feedback.
- » Prioritize compatibility: Before launching an AI application, fully evaluate its compatibility with the existing digital infrastructure. If compatibility can be achieved through a system upgrade or refactoring, you should go ahead. Otherwise, consider

integrating this application using other less intrusive methods, for example, attaching it as an external system.

- » Calculate costs and benefits accurately: If the return on investment (ROI) for AI models, compute, storage, and development is low—meaning the cost won't be recovered for years, the project's viability must be re-evaluated.
- » Build foolproof applications: To guarantee adoption, it is crucial to minimize the learning curve for AI applications. A good user experience is also highly desirable.
- » Define the model's capability boundaries: Align the model's capabilities with the target use cases. For example, in the case of the intelligent Q&A app, there is no need to pursue capabilities beyond this use case, as long as the model, enhanced by RAG, can give accurate answers to user questions.



04 Future Prospects

On our quest for enterprise AI, we have set a clear goal: to become an AI leader in three to five years. To achieve this goal, we

have developed a detailed roadmap, and based on scenario- and task-specific AI requirements we gather continuously, we

are constantly enriching this roadmap to ensure that we develop AI that is aligned with genuine business needs.

Section	Total	Function						Business												
		Finance	HR	Administrative	Legal	IT	Procurement	Products	Marketing	Operations	Production	Logistics	Warehousing	Projects	Channels	Customer services	Security	XX	XX	Planning
Section 1	24	1				1	1		3	15			2		1					
Section 2	15									6				6		3				
Real estate	8		1		1			1	3	1				1						
Section 3	7	2	1			1			1									2		
Section 4	15	2	1			2			3	6					1					
Section 5	7					1					4	2								
Section 6	14	3						2	3		5									1
Section 7	6									4									2	
HQs	2		1			1														

05 Conclusion

AI is an end-to-end transformation that covers business processes, organizations, data, and IT. Today, it has become a powerful productivity booster. At New Hope, we are integrating general models with smaller, specialized models to accelerate the intelligent transformation of all business units.

Looking forward, we plan to develop AI into one of our core competitive strengths and explore new AI use cases across all business domains. At present, we will focus on data visualization, data governance, and online data services. By forging a deep synergy between data,

models, and algorithms, we will try to leverage AI to create real value, support decision-making, while ensuring an acceptable ROI. In the end, our goal is to transform New Hope Group by embracing "All Intelligence".

One-Stop E2E IDC Cloud Transformation



Hu Jianhua

Infrastructure and O&M
Director, Shanghai Ximalaya
Technology Co., Ltd.

Abstract

This section highlights the one-stop E2E IDC cloud transformation, covering Ximalaya's journey to build cloud, go cloud, and manage cloud, how they overcame challenges of stability and scalability, and how this solution realizes deterministic operations.

01

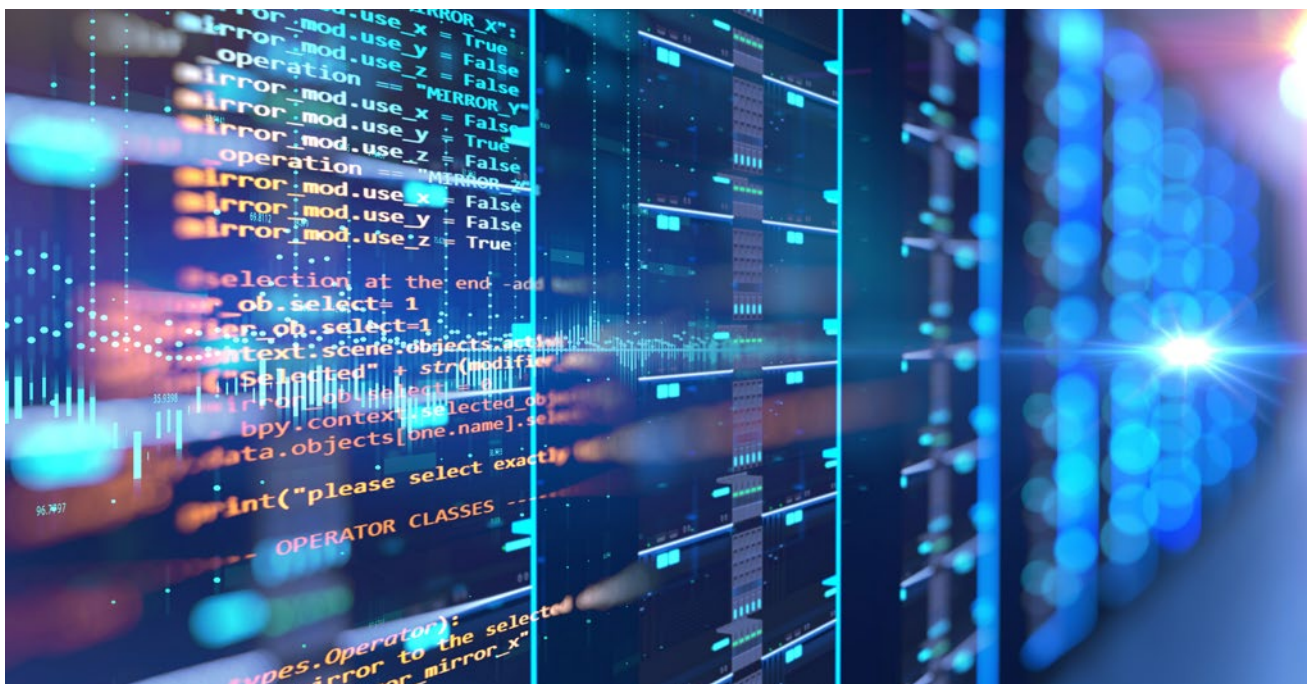
Overview to Ximalaya IDC Cloud Transformation

Ximalaya is an audio sharing platform that has attracted hundreds of millions users thanks to its high-quality content and excellent user experience. The core of its business is 'The Sound of Everything', a concept that fuels the company to deliver content tailored to users of different ages. By 2024, the company registered over 2.9 million active creators and 300 million monthly active users. The platform hosts

a total of 480 million audio files, owing to its high user engagement and wide market appeal.

To support the robust growth of its extensive business, Ximalaya has hosted over XX servers in professional IDC data centers running in active and standby mode, with equipment provided by cloud service vendors A and B, respectively.

However, as the company expanded, this traditional IDC setup had become a bottleneck for the company's sustainable growth. For this reason, Ximalaya invested CNY0.XX million to upgrade its major infrastructure and migrate to Huawei Cloud's innovative cloud-based hybrid hosting model, CloudDC. With the new cloud-based IDC model, the company began its first step to a "Cloud-based Ximalaya 2.0" age.



02 Challenges and Pain Points of Traditional IDCs

Demand for high-stability IDC equipment rooms

- One major cause of poor stability of Ximalaya services was power outages in the equipment rooms, which impacted user experience and significantly reduced the lifespan of servers. Another worry was cybersecurity, as existing firewalls struggled to defend against DDoS attacks for security issues and policy control.

Equipment room network expansion

- Due to the non-standard network architecture, it was difficult to expand the private lines in equipment rooms.
- Further, the company's flagship audio services need low network latency and high throughput, which was not always provided by legacy infrastructure.
- Indeed, a major headache and one cause of poor user experience stemmed from unstable networks and low definition.

Expensive O&M and rigid IT architecture

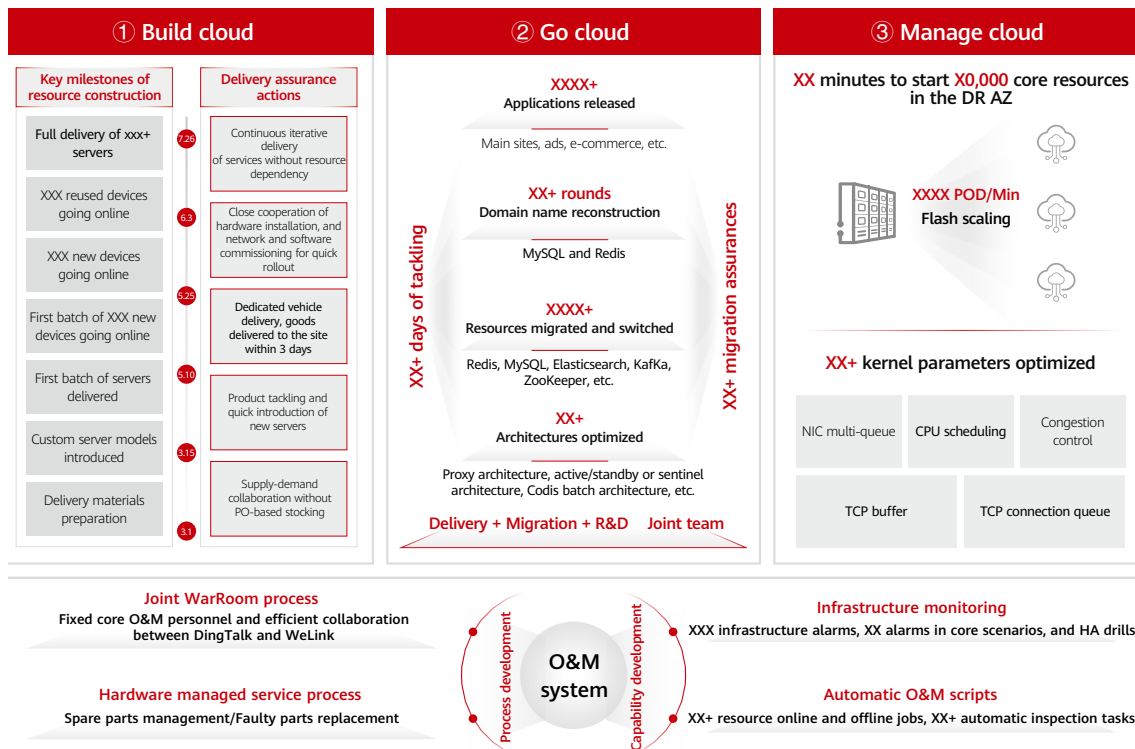
- The existing setup did not allow for auto scaling during peak hours, which needed to be addressed to improve cost-efficiency. In addition, the unbalanced and fragmented use of power, cooling, and space resources in equipment rooms also led to O&M difficulties.

System O&M

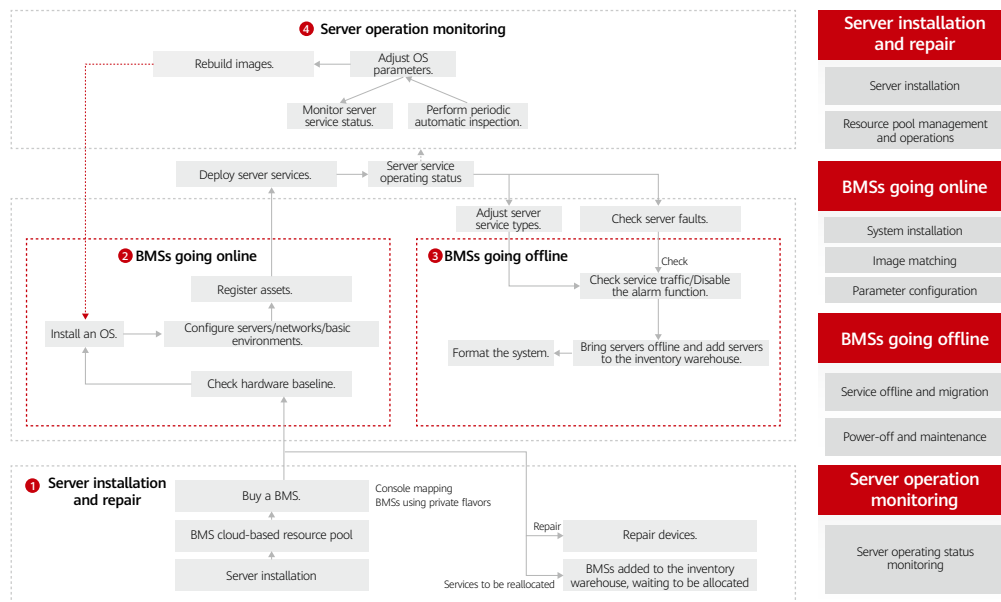
- As the company grew, more services were launched, but in doing so, system O&M became more difficult. A solution was needed to transform O&M into a streamlined, easy-to-maintain process.

03 E2E Optimized IDC Cloud Transformation in Three Phases

Faced with many challenges in IDC hosting, Ximalaya actively explored a new IDC model with Huawei Cloud, who proposed a three-phase plan designed to build, go, and manage cloud.



E2E optimized solution in three phases



Building cloud: automatic pipeline solution

1. Building Cloud: Innovative Automatic Pipeline for Delivery

The original active/standby IDC equipment rooms of Ximalaya housed more than XXX physical devices, but the company wanted unified setups to prevent faults caused by invalid configurations. Based on the project and cloud migration requirements, Huawei provided a cloud-based unified O&M platform with an automatic pipeline. This solution provides server lifecycle management and real-time monitoring of OSs and other core metrics.

The cloud building phase focused on server installation and repair, BMS online and offline, and device operations. To ensure the cloud development stayed on track, a cloud building team was set up to oversee the process and provide standardized response to various emergencies.

An HA architecture was used for the IDC equipment rooms to comply with international and national level-A regulations. Ximalaya established a cloud-based HA unified O&M platform to implement full lifecycle management of existing and new BMS resource pools. The entire management process is divided into three steps.

(1) Standardized Processes

A comprehensive and standardized management process was set up, covering resource configuration, provisioning, and usage.

(2) Automatic Implementation

All configurations are automatically completed to reduce misoperations.

(3) Real-Time Runtime Monitoring and Dynamic Adjustment

The system runtime is monitored in real

time, so that, if a potential exception is detected, the system automatically triggers the inspection process by taking BMSs offline, performing server installation and repair, and bringing BMSs online based on the inspection results.

2. Going Cloud: Zero Downtime Upon Migration

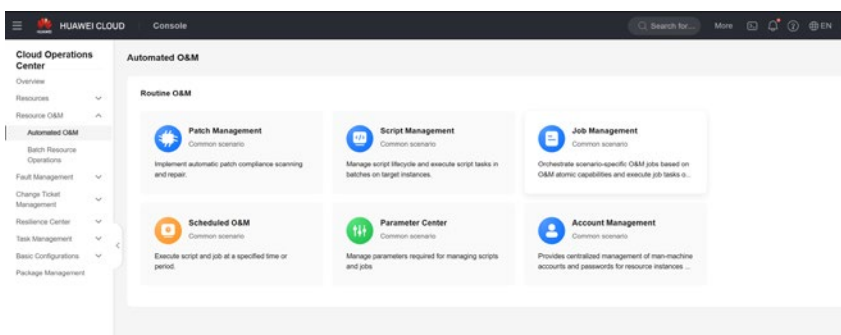
Ximalaya boasts excellent user engagement and market influence, with wide business coverage, expert systems, and tight system interconnections in the audio industry. Taking services offline for migration was a non-starter. That means during cloud migration or switchover, the company prioritized service continuity and no downtime. The company formulated a plan that ensured seamless cloud migration with zero downtime. It consists of two key phases: layered migration policy and ordered switchover.

(1) Layered Migration Policy: Migrating the Application Layer Before the Data Layer

Custom policy for the private line capacity expansion (X.X TB) was completed beforehand.

» During the migration at the application layer, gray release was performed before full release. Rollback was performed if any issue occurred.

» The data layer was centrally migrated within three days to reduce the intermediate state period and private line faults.



High-availability unified O&M platform



In this migration, custom policies ensure the stable and reliable migration and minimize service impact. Private line capacity expansion must be completed beforehand to ensure stable network transmission. During switchover at the application layer, the company initially started with a gray policy before transitioning to a full policy. This allows smooth rollback in case of a failure. Then, the data layer switchover was completed within three days to reduce the intermediate state period and the risks of private line faults.

(2) Sequenced Process: Data Synchronization → Application Release

→ Access to Gray Switchover → Resource Layer Switchover

The switchover process is as follows:

- » Data synchronization
- » Application release, which comprises RPC, Web, and tasks for single instance
- » Gray traffic switchover at the access layer, with 100% traffic used
- » Resource layer switchover

3. Managing Cloud: Service Assurance, Stable Live Network Operations, and a 1-5-10 Recovery Model

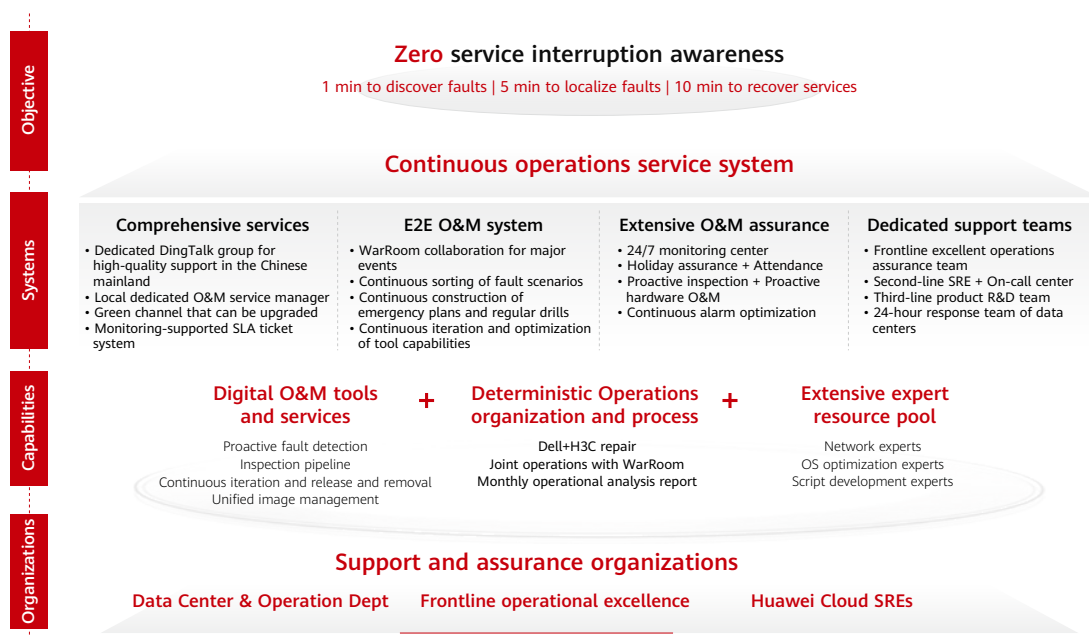
Building cloud and going cloud are the initial first steps of the Ximalaya's

O&M transformation. Subsequent cloud management, such as that for stable online services, is more challenging. Specifically, there is a need for three core capabilities: high availability (HA) architecture, risk governance, and rapid recovery. To better respond to extreme or unexpected situations, we have developed X x X fast disaster solutions.

(1) HA Architecture

High availability standards, including evaluation and objectives, were tailored to match Ximalaya IDC sites to effectively mitigate system risks.

» Then HA objectives, metrics, and



Managing cloud: E2E continuous service assurance for stable operations of the live network

methods were determined and availability research and assessment were performed on the live network in terms of the impact on the architecture, after which, an evaluation was made for each component

- » HA objectives and metrics were determined for each module.

(2) Risk Governance

Chaos engineering risk management in testing and quasi-production environments can actively seek out system risks.

- » The pilot scenarios of chaos engineering (regional power failure, network disconnection, and high temperature in the equipment rooms) were specified to establish the drill baseline and process specifications.
- » In the pilot scenarios, the blast radius control of drills, fault scenario simulation, and drill operations can evaluate and detect risks.

Huawei provides high reliability drills for data centers in four scenarios.

- | | |
|----------------|---|
| Drill 1 | Equipment room temperature fluctuation drill |
| Drill 2 | Single-rack single-route power failure |
| Drill 3 | Dual-source power supply & Diesel generator load test drill |
| Drill 4 | Active/Standby AZ network HA drill |

(3) Quick Recovery and "1-5-10"

A comprehensive monitoring system, covering applications, cloud platforms, middleware, and full-stack networking, was built to locate E2E faults and enhance fault demarcation efficiency. Based on Huawei Cloud's product capabilities and existing monitoring and O&M tools, we have planned, designed, and implemented E2E full-link monitoring, fault location, isolation, and recovery solutions, spanning from applications and the cloud to the networks. These solutions have been verified in actual network environments.

A continuous operations service system was built by focusing on identification, coverage, and drills to achieve the

objectives. This system aims to ensure that faults are detected within 1 minute, demarcated within 5 minutes, and rectified jointly within 10 minutes.

Fault detection within 1 minute:

- » All faults can be reported timely, and period tests and drills were performed. A multi-level alarm reporting mechanism was set up, which was tailored to various alarm levels and stages. The alarm rules for the four lines of defense were optimized to effectively address the challenges during peak hours.
- » Category- and level-based monitoring was enabled for alarms generated during batch fault and threshold detection to quickly respond to and handle the alarms. In addition, the potential risk alarm reporting mechanism was continuously optimized to ensure the efficiency and accuracy of the monitoring system.

Fault demarcation within 5 minutes

Information about hardware and software resources was fixed, such as servers, switches, and cloud resources. The fault locating methods were optimized, and probe functions were added for high-risk scenarios, such as private lines.

- » The core O&M team members remain stable, including technical support managers (TAMs), level-2 O&M engineers, and other key position personnel, with backups assigned for all.
- » Through continuous network observation, explanation training, and

war room exercises, the teams can quickly acquire and apply these skills.

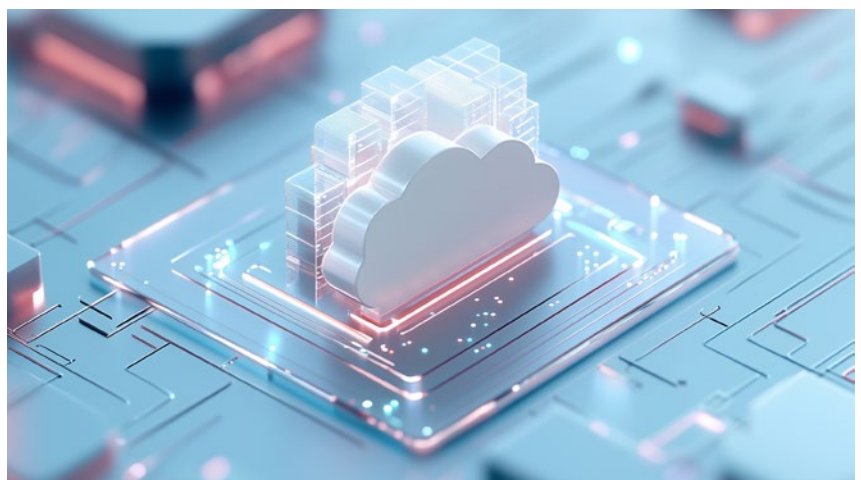
Service recovery within 10 minutes

- » Five types of common fault mode libraries were established, covering devices in equipment rooms, CCE clusters (highly automated and managed Kubernetes services), networks (physical and virtual), private lines, and distributed denial of service (DDoS) attacks.
- » Four types of high availability architectures were analyzed: redundancy and disaster recovery, overload control, fault management, and access paths.
- » Two escape solutions were formulated: network address translation (NAT) and NG (an emergency escape solution configured for firewalls or other network devices to ensure rapid restoration of service traffic during network failures).
- » All drill solutions must have a 100% success rate.

(4) X x X Disaster Response and Ultra-Fast Recovery

X0,000-core resources started within 5 minutes

In the event of a geological disaster or other emergencies, X0,000-core resources can be started within 5 minutes by combining Ximalaya IDCs into a unified IDC equipment room. This significantly improves resource flexibility and availability. Dual availability zones (AZs) ensure seamless switchover during migration.



04 Key Experience Summary

By working with Huawei Cloud project team, Ximalaya developed an E2E IDC cloud transformation solution to complete the IDC cloud transformation in three phases.

1. Building Cloud: Introducing an Automatic Pipeline

- » Standardized equipment, configurations, and automation are essential to streamline the Ximalaya IDC environment, server lifecycle management, and O&M.
- » Elastic resource management and dynamic reallocation policies enable O&M personnel to adjust the resource pools on demand to fit service demand, improving resource utilization.
- » The HA architecture is designed to comply with international and national Class A regulations, while also establishing a robust foundation for future cloud expansions.

2. Going Cloud: Service Migration with Zero Downtime

- » Layered migration policy: Application layer is migrated before the data layer, with gray switchover and quick rollback available to ensure smooth migrations.
- » Quick resource startup thanks to Huawei Cloud, which can start X0,000-core resources and connect resources from the old and new IDC environments to ensure service continuity.
- » A sequenced switchover ensures an efficient and stable migration, starting with data synchronization, through to application release and gray traffic switchover, and finally resource switchover.

3. Managing Cloud: Deterministic Operations

- » The solution uses an HA architecture design, which is designed based on high availability evaluation criteria and objectives, effectively mitigating system risks before they impact ongoing services.

- » Risk governance is achieved through measures like chaos engineering, which actively detects risks, and high-reliability drills in simulated fault scenarios, improving system resilience.
- » In terms of quick recovery, the O&M service system is equipped with 1-minute fault detection, 5-minute fault demarcation, and 10-minute service recovery, delivering much faster fault response and recovery.

Through this three-phase project of building, going, and managing cloud, Ximalaya successfully addressed issues of stability, availability, scalability, security, TCO, and O&M of IDC equipment rooms. The modern infrastructure and streamlined management further improved service efficiency, market competitiveness, and overall O&M efficiency.

05 Values

The cloud-based IDC solution empowers Ximalaya with best-in-class services and enhanced technical value.

1. Service Benefits

- » Lower Costs: The company saves up to CNYX.XX million per month (in racks, bandwidth, servers, switches, etc.), with total savings exceeding CNYXX.XX million in three years.
- » Improved O&M efficiency: The platform enables automated periodic inspections and one-stop hardware reports providing fast and accurate information.

- » High availability: The solution provides a mature disaster recovery and backup design, using dual-AZs that provide 99.99% availability and enable fast, elastic scaling of X0,000 cores.

2. Technical Benefits

- » Applications optimization thanks to domain name reconstruction, which resolves IP address hard-coding issues, and containerization that enhances service stability
- » Specifications optimization by unifying kernel parameters and upgrading

middleware components and unifying their versions

- » Architecture optimization by upgrading services and middleware for HA dual AZs and splitting and isolating ultra-large databases
- » Security optimization to meet the requirements of DJCP level-3 protection standards. Further, DDoS cleans north-south traffic and CFW protects east-west traffic, with Top N IP tracing and rapid locating and rapid blockage also provided.

06 Summary

By working closely with Huawei Cloud, Ximalaya now deploys its servers in Huawei equipment rooms and uses Huawei Cloud technologies. The

partnership with Huawei Cloud was a key milestone for the company's long-term growth. In addition to improving resource usage and cutting its leasing and

O&M costs, the reliable infrastructure and technical support provided by Huawei Cloud were a major factor in improving the reliability of Ximalaya services.

Strategic Development of Predictive Operations Capabilities in China's Retail Sector: a Case Study of a Leading Convenience Store Brand



Wu Hongqin

Data IT System Operations
Director and Technical
Director, Meiyijia Holdings
Co., Ltd.

Abstract

The article expands on how Meiyijia — a leading convenience store brand in China — has developed and implemented predictive and proactive operations capabilities to address challenges arising from rapid business growth, cloud-based operations, and accelerated digital transformation. By embracing the site reliability engineering (SRE) culture and implementing proactive measures like rapid fault rectification, chaos drills, and release management grounded in deterministic operations principles, Meiyijia aims to maximize system uptime and deliver exceptional user experiences.

01

About Meiyijia

Meiyijia Holdings Co., Ltd. was established in 1997, with China's first chain supermarket, Meijia Supermarket, as its foundation. It operates as a commercial

distribution company under Dongguan Sugar & Wine Group. Over the past two decades, Meiyijia has focused on product

research and development and has grown to have more stores than any other convenience store brand in China.



87 million+
registered online members



250 million+
customers served monthly in
physical stores



5 million+
customers served monthly via
convenient services





02 O&M Challenges amid Business Expansion and Digital Transformation

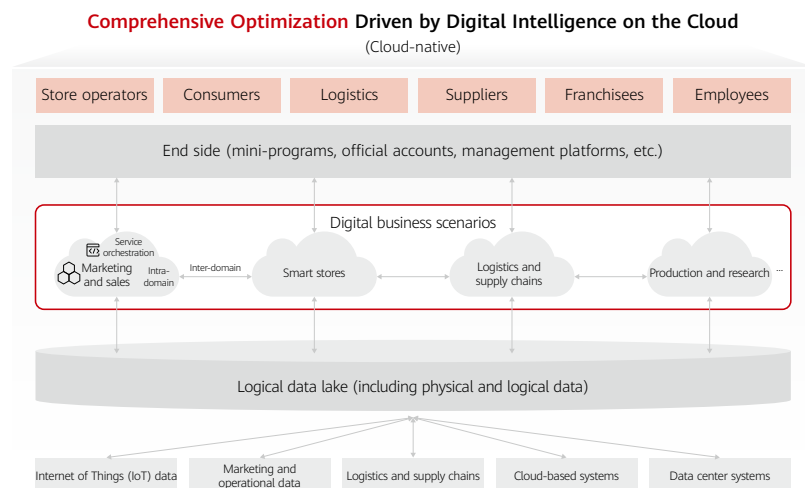
Meijijia operates over 38,000 stores. As it expands, its daily transactions and transaction volumes grow rapidly. Such swift development introduces increasingly complex business scenarios and diverse service objects, posing significant challenges to O&M:

- » The growing number of application systems and O&M objects substantially increases O&M workload demands.
- » Cloud migration has complicated the system architecture, raising the skill requirements for O&M personnel.
- » Rapidly evolving business needs shorten version iteration cycles, while frequent releases heighten the risk of live network issues.
- » The prolonged fault recovery duration negatively impacts user experience.

To address these challenges and enhance its competitiveness, Meijijia has initiated comprehensive digital transformation in a range of areas. Key initiatives include:

» Deploying advanced IT technologies to facilitate the migration of legacy applications to the cloud, thereby improving business stability and user experience.

» Leveraging cloud-native technologies to develop innovative solutions, including smart stores and cloud-driven sales systems, as well as smart logistics and supply chains, driving significant business evolution.



03

During our operations transformation, Meiyijia has been closely cooperating with Huawei and learning from Huawei's deterministic operations practices. By utilizing the deterministic operations maturity assessment model, we have systematically evaluated and analyzed our operations status, pinpointed key challenges, and formulated improvement plans.

1. Operations Maturity Assessment

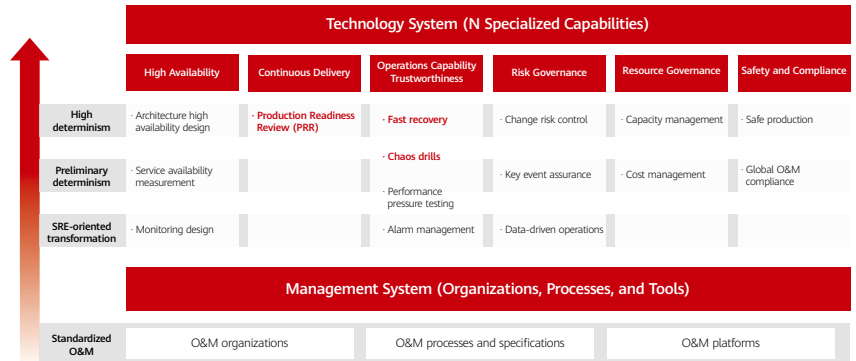
Drawing upon the deterministic operations maturity model and the "1+N" framework (illustrated in the accompanying figure as an integrated structure comprising one management system and multiple specialized technological capabilities), we evaluated our operational management and technology systems. At the time of assessment, both systems operated at the standardized operations level. The management system required additional refinement, and the technical system exhibited deficiencies in its proactive operations capabilities.

2. Three-Stage Enhancement of IT Operations Capabilities

Meiyijia's core business objectives include reducing network fault occurrences and ensuring swift service restoration when incidents arise. To address these needs and enhance operations maturity, improvements must focus on proactive operations. Firstly, operations capabilities should extend into early-stage business management by implementing a release management mechanism to preemptively identify and mitigate quality risks. Secondly, robust rapid recovery mechanisms must be established to minimize the impact of faults through timely resolution.

Based on these considerations, Meiyijia has outlined a phased approach to advance its operations capabilities over the coming years:

Stage 1: Implement a quick-win strategy targeting immediate operations challenges, focusing on developing key capabilities such as release management, rapid fault rectification, and chaos drills.



Stage 2: Optimize organizational structures, processes, and tools. Establish a service reliability metrics framework, foster expertise in high-availability architecture design, and enhance observability to improve fault detection, isolation, and locating.

Stage 3: Prioritize operations risk management and resource governance to drive continuous improvement in business availability.

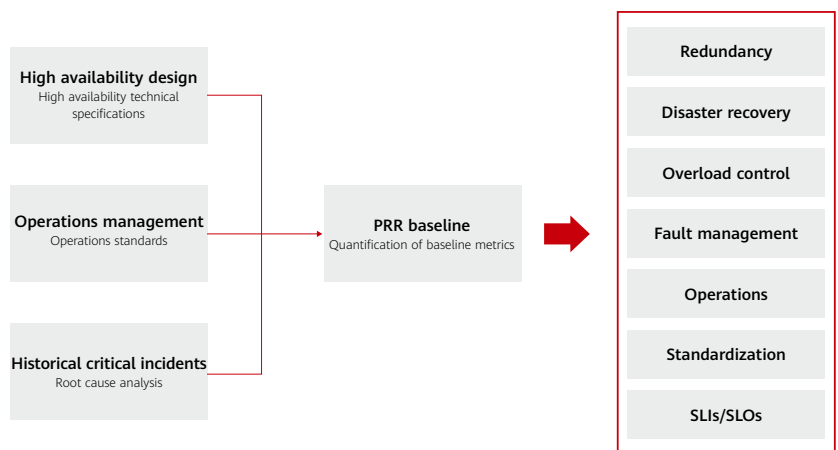
3. Release Management and Non-functional Requirement Reviews

As mentioned, integrating operations capabilities early in business management constitutes a key part of Meiyijia's operations transformation strategy. To execute this approach, we established release management capabilities through Production Readiness Reviews (PRRs). By

leveraging PRRs, we have embedded operations considerations into the design and review processes for non-functional system requirements. This integration enhances business system quality, minimizes post-release failure risks, and improves overall business availability.

(1) PRR Baseline

To establish a systematic approach for evaluating non-functional requirements, we developed a PRR baseline for release management by leveraging both our internal expertise and Huawei's PRR framework. Our internal contributions encompassed high-availability design principles, operations management guidelines, and insights from analyzing significant past incidents.



(2) Review Mechanism

Non-functional requirements need to be regularly evaluated via reviews throughout the business system's lifecycle to ensure timely identification and elimination of potential risks.

(3) PRRs

The operations team is responsible for organizing PRRs. This team includes architects, product managers, developers, and quality specialists. For each application, various issues can be uncovered during a PRR. These may include ambiguous rate-limiting strategies, suboptimal system concurrency limits, improper utilization of caching mechanisms, and insufficient monitoring. For each issue identified in the PRR, a remediation timeline is established. Architects propose solutions and offer technical guidance, developers address product-related issues, and the operations team resolves operational deficiencies such as inadequate monitoring coverage.

(4) Continuous Implementation

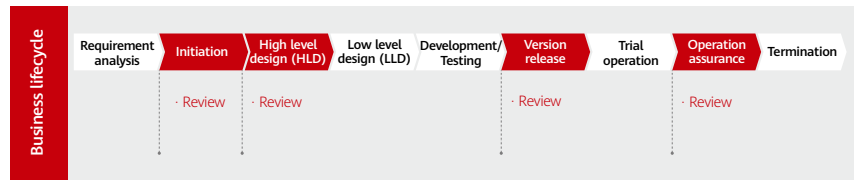
- » For externally sourced business systems, PRRs are conducted during the project initiation, HLD, and release phases.
- » For in-house business systems, PRRs are performed during the HLD and release phases.
- » For core business systems, regular PRRs are executed after the system goes live.

(5) Value

- » Release management has been enhanced through standardization, normalization, and alignment with an efficient workflow for deploying business systems.
- » The high-availability technical specifications have been uniformly defined.
- » Potential risks in business systems are proactively identified and promptly addressed, improving system reliability, reducing fault occurrences in production environments, and increasing service uptime.

4. Practices for Shortening Recovery Time

Rapid fault rectification capabilities include building a failure mode library,



developing contingency plans, and conducting trustworthiness verification drills. Utilizing resources from the Huawei Cloud Operations Center (COC), these measures are applied in practical scenarios to significantly minimize downtime.

(1) Building a Failure Mode Library

From a fault tolerance perspective, we employed fault tree (FT) as well as failure mode and effects analysis (FMEA) to perform a layered examination of the business architecture. With this as the foundation, we conducted both forward and backward analyses, and ultimately developed a comprehensive failure mode library. This library categorizes failures into five distinct types: redundancy, disaster recovery, overload, dependency, and configuration. To date, it covers failure modes spanning more than 80 scenarios.

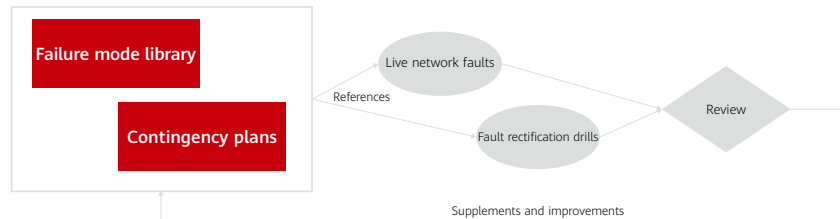
(2) Developing Contingency Plans

To ensure rapid service recovery in known fault scenarios, we have developed specific contingency plans for each scenario and conduct regular drills to verify their effectiveness. The plans that prove effective through drills undergo further reviews by our internal team before being documented for use in actual network issue resolution. Additionally, ongoing efforts will focus on refining and disseminating these plans.

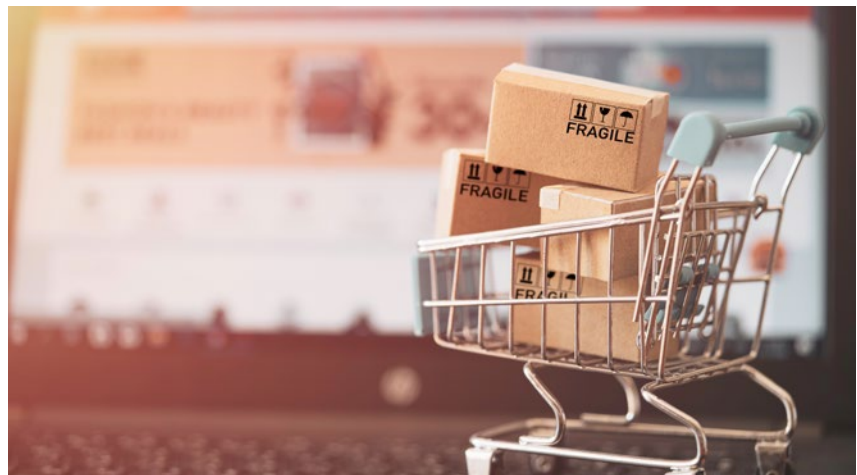
Methodologies for Developing Contingency Plans

Forward development: Establish a failure mode library utilizing FT-FMEA. Subsequently, develop corresponding contingency plans tailored to these identified failure modes.

Backward improvement: Refine the contingency plans based on real fault scenarios and post-incident reviews.



Continuous improvement of contingency plans



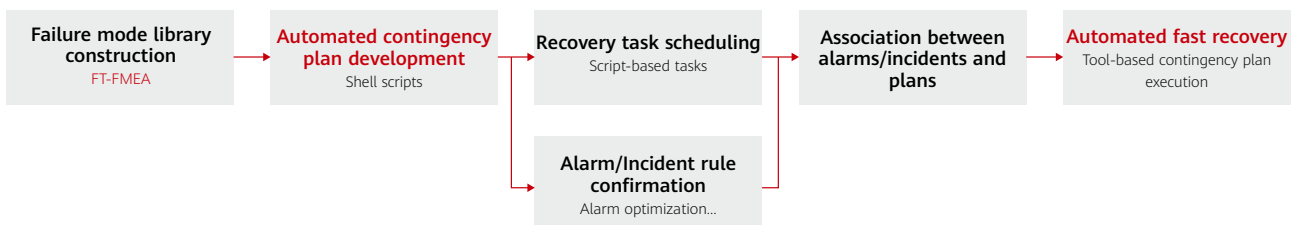


(3) Trustworthiness Verification Through Drills

Drills are systematically designed and executed according to the predefined failure modes to consistently verify the reliability of contingency plans. Post-drill reviews are subsequently carried out to pinpoint deficiencies in both failure modes and the corresponding plans. Continuous enhancements are implemented to address these issues, with updates integrated into the related libraries.

Rapid Fault Rectification

The failure mode library and contingency plan library are integrated within Huawei Cloud COC. Alarms are associated with corresponding contingency plans, whose reliability is validated through systematic drills to ensure effective fault resolution. It is critical to incorporate drill-based verification into standard operational practices, so as to enable swift identification and deployment of appropriate contingency plans when faults occur on the live network.



Benefits

- » The development and implementation of rapid fault rectification capabilities have improved the operations efficiency and reduced the Mean Time To Repair (MTTR).
- » Specialized personnel skills have been developed.

5. Chaos Drills for Higher System Availability

Chaos drills are executed across different business failure scenarios to identify potential risks and deficiencies in businesses, while simultaneously improving the emergency response proficiency of the operations team.

(1) Failover Drills

Drills are performed on a designated application cluster to validate its automatic failover, backup, and recovery capabilities. By doing this, the operations personnel can be familiarized with the

standardized drill procedures, and the organization mechanism of standard drills can be clarified.

(a) Drill Plan

Prior to conducting a drill, it is essential to analyze potential fault scenarios for the target application and evaluate their impacts. Based on the fault resolution mechanisms, contingency plans must be developed for each identified failure point.

During the drill, faults are injected using Huawei Cloud COC. Operations personnel then monitor the system, execute recovery procedures in accordance with the contingency plan, and thoroughly document the entire process.

Following the drill, a debriefing session is conducted involving all participants to review the process, analyze encountered issues, and designate responsible personnel to oversee the implementation of corrective actions until all these issues are resolved.



(b) Drill Process and Benefits

Drill Process



Benefits

The cluster serves as a critical component of Meijijia's infrastructure. Prolonged downtime could disrupt store operations, resulting in potential customer dissatisfaction. Through drills, weaknesses in cluster failover mechanisms can be identified, and the effectiveness of backup and recovery policies can be evaluated.

(2) Chaos Drills for Store System Switching

After the store operation-related platforms and application systems are migrated to the cloud, the individual store systems need to be switched to the updated infrastructure in batches. The viability of this approach has been verified through drills involving 100 stores.

(a) System Switching Plan

Before the switching, it is essential to confirm the information of the target stores. A scheduled task then needs to be configured on the tool platform to push upgrade instructions in batches.

The upgrade process consists of three key steps:

- » Synchronize foundational data to the new platform before the upgrade instructions are officially pushed for execution.
- » Upgrade the POS systems at target stores. If the upgrade fails, the systems will automatically roll back to the old version.
- » Migrate store operational data to the updated systems.

(b) Tool-based Migration Drill

The following takes a drill in the User Acceptance Test (UAT) environment as an example. In this drill, the production environment data was synchronized to the UAT environment, where the data for 20, 50, 100, and 400 stores was migrated in separate batches. The PerfTest tool was used to call APIs to simulate POS upgrades at the front end. Once the upgrade was completed, backend ticket data migration was initiated automatically.

(c) Drill Process and Conclusion

Drill Process

Conclusion

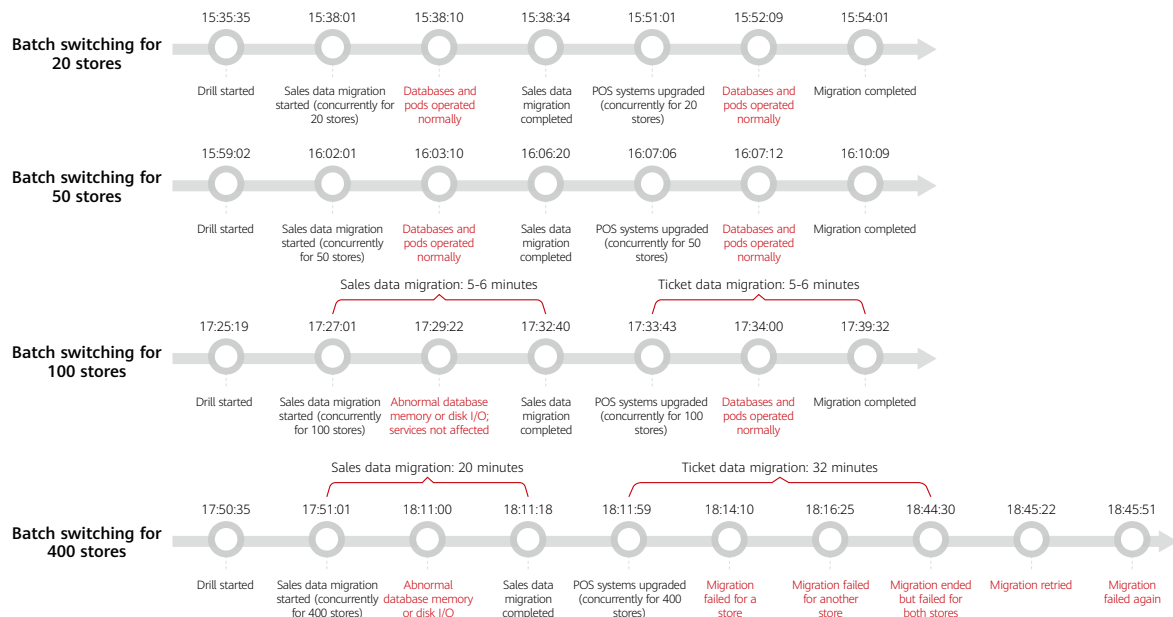
In the UAT environment with defined resource configurations, the tool can be used to migrate data for 100 stores simultaneously as expected.

(d) Results

- » The capability of simultaneous system switching for 100 stores has been verified.
- » The service functions of the migration tool have been verified.
- » The architect, R&D, testing, and operations teams are all able to apply the theories of chaos drills through practices.

6. Transferring Skills Through "Golden Seeds"

During the initial phase of developing proactive operations capabilities, we identified key individuals — referred to as "golden seeds" — and provided them with classroom-based training followed by practical application to build expertise. Given that deterministic operations effectively address Meijijia's diverse requirements, we opted to fully implement it. Subsequently, these trained personnel were tasked with conducting organization-wide IT training sessions and overseeing practice exercises, ensuring the transfer of proactive operations skills across all business teams.



04 Summary

1. Success Factors

The success of Meiyijia's deterministic operations practices stems from three key factors:

- » Well-defined objectives for operations transformation: Aiming at transitioning towards proactive operations, we focus on enhancing expertise in release management, rapid fault rectification, and chaos drills. This approach minimizes network failure risks, reduces downtime, and ensures the stability and reliability of the production environment.
- » Integration of Huawei's proven methodologies: By collaborating closely with Huawei, we continuously adopt and adapt their SRE best practices and deterministic operations methodologies. These insights are systematically applied to develop Meiyijia's distinctive deterministic operations capabilities.
- » Effective cross-functional teamwork: Leveraging a core group of high-

performing "golden seed" employees, we designed robust plans for building deterministic operations capabilities. These individuals spearhead collaborative efforts across diverse teams, enabling the practical application and deployment of these methodologies in real-world business contexts.

2. Outcomes

- » Accelerated development of proactive operations capabilities: Key proactive operations capabilities — including rapid fault rectification, chaos drills, and release management — have been rapidly established and implemented. A systematic operations solution has been set up, complemented by the training of high-potential "golden seed" personnel. All of this propels the continuous advancement of Meiyijia's operations system.
- » Enhanced system availability and user experiences: Leveraging the expertise of "golden seed" teams, the three core proactive operations

capabilities have been progressively integrated into new business systems. This integration reduces fault occurrence and downtime while enabling early detection and resolution of potential issues, ultimately enhancing system reliability and user satisfaction.

- » Expedited digital transformation: The advancements in operations are accelerating Meiyijia's broader digital transformation journey, fostering business innovation and speeding up organizational growth.

Moving forward, we aim to implement deterministic operations practices across the data IT system and expand our specialized capabilities enterprise-wide, encompassing all critical business functions. Additionally, we will further refine the operations organization, build a robust proactive operations management system, and broaden proactive operations scope. These initiatives will enhance product quality, elevate operations efficiency, and strengthen enterprise competitiveness.



Staying Ahead in O&M with an Observability System



Ma Tao

aPaaS and application
SRE expert, Huawei Cloud

Abstract

Huawei product "P" once struggled with fault detection and location during its cloud-native SaaS transition. To tackle these challenges, a smart O&M data warehouse was created. It brought together logs, metrics, and trace data for a complete view of the system from IaaS, PaaS, and SaaS. This data warehouse makes it possible to detect faults in just 1 minute, locate their causes in 5, and fix them in 10.

01 Background

Huawei product "P" is a SaaS platform built on cloud-native technologies. As it is an externally facing service with XX millions of users, hitting a 99.99% uptime is key. But faults are hard to detect and fix. It needed an observability system to make that easier.

Challenge 1: Complex Architecture and Data Silos

Microservices are exploding in volume, and the calls between them are getting tangled, further complicated by the wider use of distributed cloud-native technologies. Any application fault necessitates a full check on the entire path: application > container > PaaS instance > VM > virtual network. Logs, metrics, and trace data are scattered in different tools, making it really hard to detect and fix faults quickly.

Challenge 2: Mass Invalid Alarms

Cloud services could generate tons of alarms, most of which are often false alarms or deny easy handling. O&M teams get buried in all these alarms and could only respond reactively upon customers' reporting.



02 Solution

1. Building an End-to-End Observability System

Huawei product "P" is all about creating a top-notch monitoring system that keeps an eye on each application and everything it uses, such as VMs, Docker, middleware, and databases. The system tracks the whole process, so it can detect faults in just 1 minute, locate their causes in 5, and fix them in 10, and even make some faults fix themselves.

End-to-end monitoring is essential if you want to:

- » Proactively detect faults before they can wreak havoc on your system. Proactive fault detection keeps everything running smoothly and reliably.
- » Zoom in on faults and fix them quickly with deep data insights.
- » Monitor your system and your data 24/7, use real-time data to make smart moves, and optimize system performance and resource use.

Huawei product "P"'s metric system zeroes in on the service level-data. It monitors and analyzes key metrics to show how services are running and how users feel. By visualizing the user journey and interactions, it helps SRE engineers quickly detect and fix faults. This keeps SRE engineers focused on keeping critical services stable.

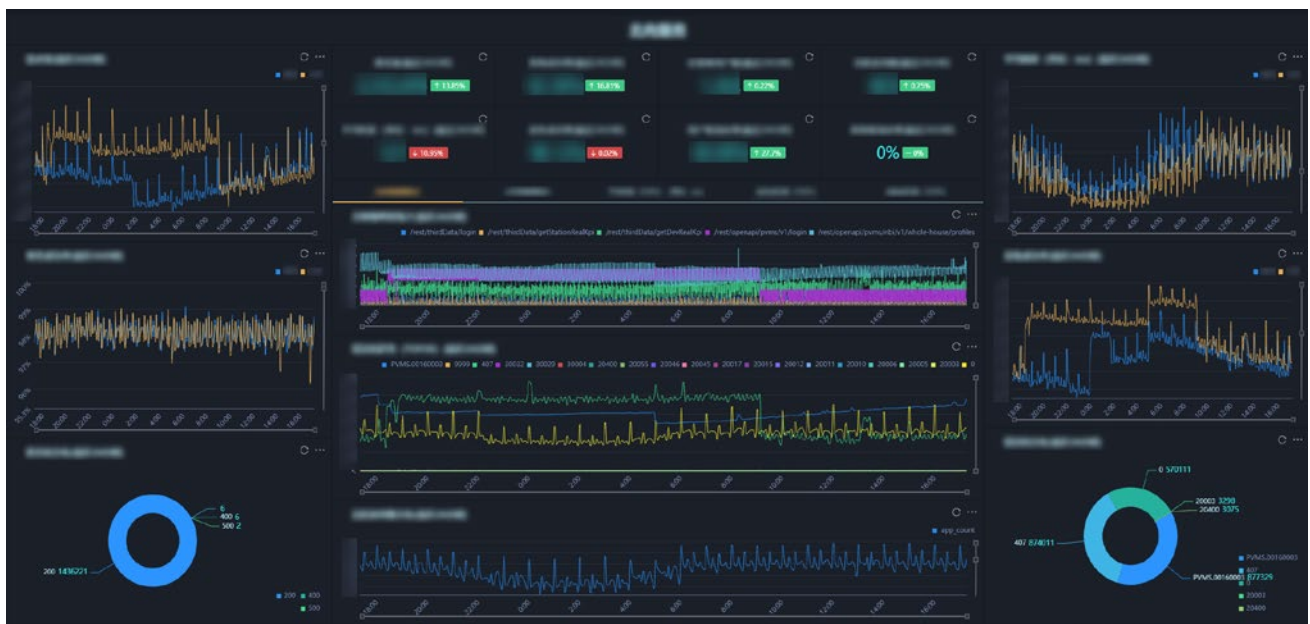
1) The team behind the metrics

To create observability metrics, you need to know the service inside and out. To accomplish this, we rely an "iron triangle" comprising developers, testers, and O&M personnel. Developers are responsible for the system's logic and add monitoring and logging. Testers think like users, designing test cases and spotting key metrics. O&M personnel then fine-tune these metrics based on what they see in real use. Together, they craft an observability system.

2) The process of designing metrics

We can divide the process into four steps:

- » Data survey and service breakdown: Metric designers break down the service system to identify its main functions and the data they create, such as structured data, logs, and metrics.
- » Conceptual model and bus matrix: Developers map out the service processes for each core function, noting important data points and their details. This creates a bus matrix, which is like a blueprint for the data.
- » Logical model design: Based on the bus matrix, designers create a logical model for the data warehouse. Here are the key concepts and components involved:
 - a. Fact table: This is the main table in a data warehouse. It holds numbers and metrics related to business activities. Each row represents a business event and links to one or more dimension tables. In practice, all measurements from the same business process should be stored in one dimension model.





- b. **Dimension:** A dimension is a feature that describes a business process. It helps classify and group facts.
- c. **Dimension table:** This table provides context for the numbers in the fact table. It describes events related to "who, what, where, when, how, and why", and helps group and filter facts.
- d. **Hierarchy:** Dimensions can have levels. For example, the time dimension can include years, quarters, and months.
- e. **Measure/Atomic metric:** These are the numbers in the fact table that show how a business process is

doing. They are the main metrics users look at in the data warehouse.

- » Physical model development and rollout: This step is about filling the logical models with data. Data from databases, logs, traces, and metrics is cleaned and organized into physical tables. Once the data is ready, metric designers can work on finalizing the metrics and bringing them online.

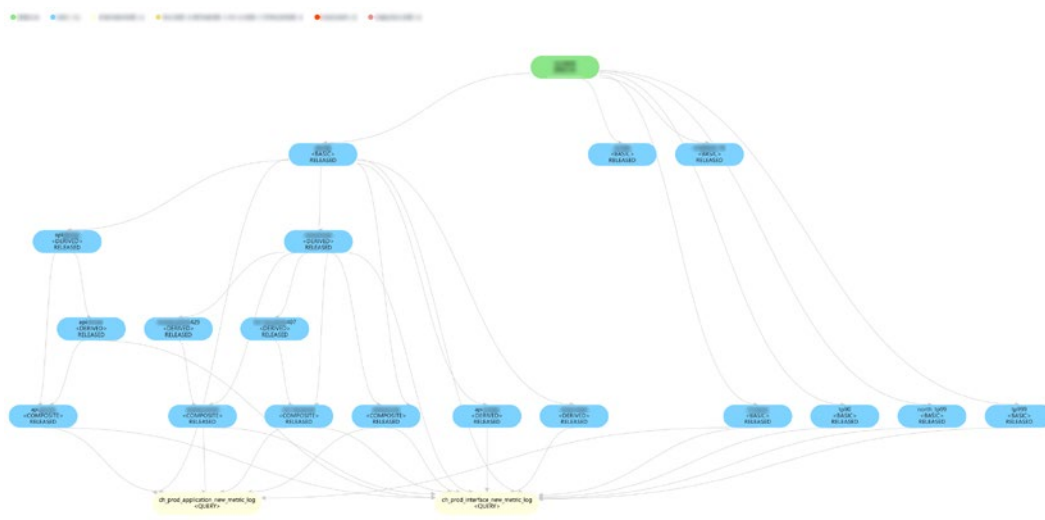
2. Building an Intelligent O&M Data Warehouse

To get your end-to-end observability system up and running, you will need a data warehouse that is flexible, scalable, and compatible. Huawei product "P"

steps in and help you quickly build a solid, observable data foundation using Huawei Cloud AppStage, a one-stop O&M platform that makes it all possible.

Huawei product "P" generates a lot of O&M data, including devices, network dialing tests, instance metrics, logs, and traces. All this data needs to be stored in a unified O&M data warehouse. This warehouse includes components like data integration, ETL, a data lake, MPPDB, and data applications to do these jobs:

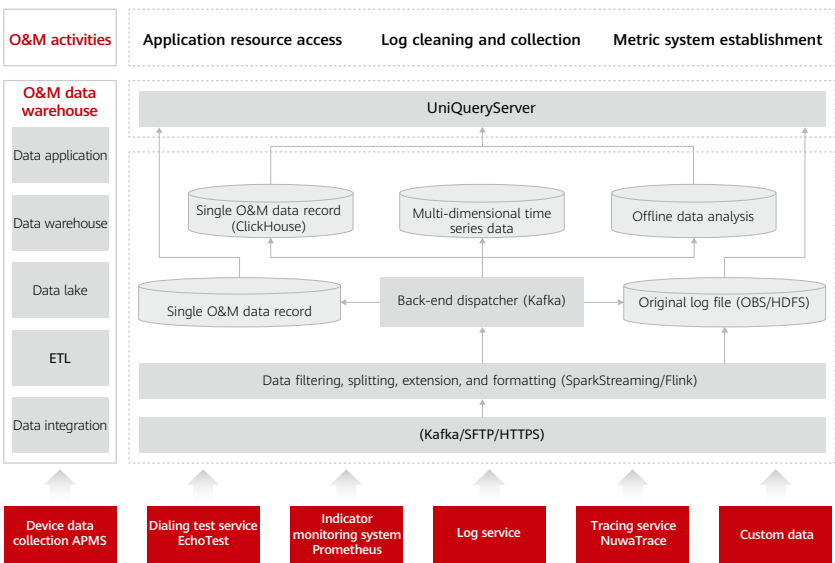
Data integration: Since not all data is neatly organized, we need a platform that can bring in different kinds of O&M data, like from message queues, APIs, and SFTP.



Data extraction: Once the data is in, ETL steps in to clean it up—filtering, splitting, and formatting it—then puts it in a message queue.

Data lake: Different parts of the system take data from the message queue and store it in different places. For example, raw logs go to OBS, detailed data goes to databases or ClickHouse, and complex time-series data goes into MPPDB tables based on specific rules.

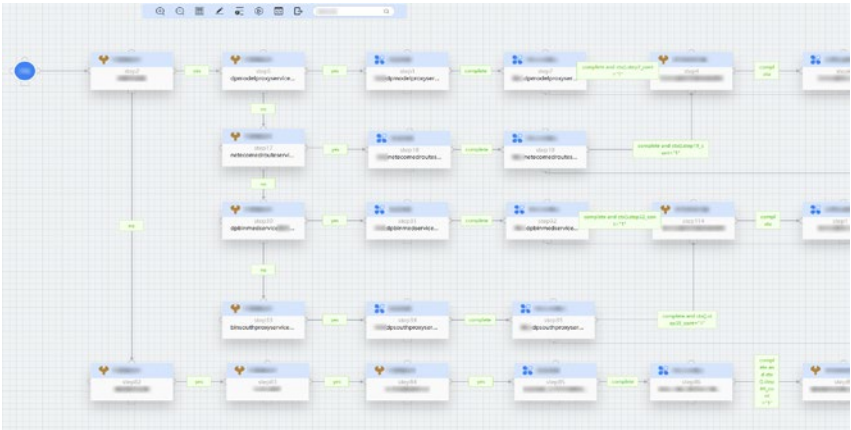
Data application: After the data is all sorted out, data applications turn it into APIs for unified query.



3. Enabling Automated Fault Recovery

If we know why a service is broken, we can fix it faster. This might mean switching traffic, isolating bad nodes, or restarting them. But new folks might need help, and even seasoned SRE engineers can struggle if the system is huge. A good O&M team should be able to fix faults on their own.

Figuring out and fixing faults takes time and experience. Huawei product "P" uses the EAP process from AppStage and data like application monitoring, alarms, and diagnostics to set up a smooth fault recovery process. When something goes wrong, this process kicks in quickly. Even beginners can fix faults fast.



service is by keeping an eye on its APIs. Huawei product "P" figures out which APIs are most important, sets success

rate targets, and uses AppStage's data orchestration to create a solid real-time SLO monitoring system.

4. Building an Online SLO Monitoring System

The end-to-end observability system can monitor all service APIs. This means we can check how available a





03 Innovative Solutions

1. Breaking Data Silos and Establishing an Application-Oriented Data Observability System

As applications become cloud native, many supporting tools and technologies come to life. This scatters data, slowing down O&M efficiency.

Huawei product "P" uses AppStage to build an intelligent O&M platform that

connects all the dots: logs, tools, traces, and cloud resource monitoring metrics. This gives you a comprehensive view of the system, making it easier for SRE engineers to quickly find and fix faults.

2. Building an Online Availability Monitoring System

An end-to-end observability system creates a wide range of metrics. These help track call success rates of key APIs, as well as figuring out their weights and

unavailability thresholds based on service logic. This information is then used in a unified, abstract computation model for real-time SLO monitoring.

3. Automating Fault Recovery with Process Orchestration

The system has pre-set ways to fix common faults. When an alarm is triggered, the system quickly starts the recovery process, cutting down the Mean Time to Repair (MTTR).

04 Achievements

High SLO: The product hits a 99.99% SLO. In core scenarios, the product team can detect faults in 1 minute, locate their causes in 5, and fix them in 10. They broke down data silos in logs,

metrics, and trace data, generated xxx monitoring reports on xxx key metrics by service, and created a top dashboard for SaaS applications for real-time SLO monitoring.

Fast fixes: AppStage helps detect issues, find the causes, and fix them automatically with RPA capabilities. This enabled faster issue resolution, ultimately saving CNYXXXX.

05 Summary and Improvement

Our observability system, built with AppStage, works great for detecting, locating, and fixing faults. It also helps with gray releases and chaos engineering, which are our next steps.

Gray releases: During a traffic switchover, a dialing test helps compare the performance of both the new and

old versions. However, these tests do not always reflect reality. With an end-to-end observability system, you can see the key metrics of both versions on the same chart, a great way to keep a close eye on the new version. The next step could be linking the new version to the observability system to get the most out of the latter.

Chaos engineering: During fault injection drills, observability metrics can be collected and displayed in one unified report where you can see data changes and expected vs. actual results. This enables O&M personnel to detect any possible faults that were caused by the drill but would have otherwise been missed.

Empowering IT and Application System Operations with DeepSeek



Tang Runhong

Database administrator
(DBA) at a well-known
joint-stock commercial bank

Abstract

DeepSeek is sending shockwaves around the world. There are many articles and guides about it, including instructions on how to use, deploy, and achieve optimal cost-efficiency with DeepSeek. Business and organizations around the world across all sectors are announcing they have integrated DeepSeek models. In this article, I will explain how to use DeepSeek to empower IT and application system operations and maintenance and the likely feasible use cases in this domain.

01 Why DeepSeek?

1. Advantages of DeepSeek Models

In my opinion, the huge popularity of DeepSeek V3 and R1 is due to the following reasons: they are open-source and free; they can be deployed on-premises and accessed through open APIs; their training and inference costs are significantly lower than other models with comparable performance; and they demonstrate outstanding performance in semantic understanding and in-context reasoning.

1) Open source and free

Most large language models on the market are closed-source or only partially open-source. In contrast, DeepSeek R1 is free under an MIT license, allowing users to freely use it for commercial purposes, modify it as they wish, and develop other products on top of it. In a real sense, this openness challenges the monopoly of traditional closed-source models and lowers the barriers to wider AI adoption. Small and medium

businesses, as well as independent developers, can build upon DeepSeek R1 without incurring high royalty fees. Furthermore, DeepSeek offers a full series of open-source models (1.5B to 70B parameters), which can be adapted to multiple hardware architectures (e.g., NVIDIA PTX programming and storage-compute integrated chips). They can be deployed locally, even on a regular laptop. So far, major cloud vendors, including Alibaba Cloud, Huawei Cloud, Tencent Cloud, AWS, and Microsoft Azure



have announced their own DeepSeek R1 hosting services. More than 160 enterprises, both within and outside China, have joined or plan to join the DeepSeek ecosystem, spanning AI chips, cloud computing, and end-user applications.

2) Significantly low training and inference costs but on-par performance

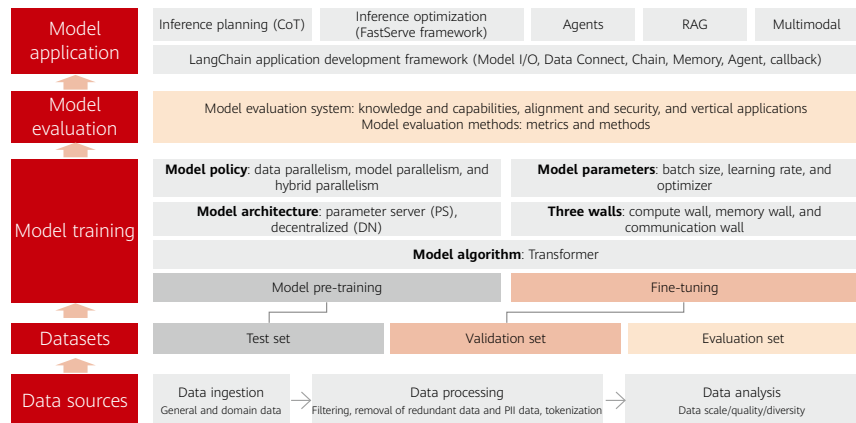
Through optimized algorithms (e.g., reinforcement learning and the mixture of experts (MoE) architecture) and training processes, DeepSeek R1 significantly reduces computational loads for both training and inference. DeepSeek R1 demonstrates superb performance in tasks such as mathematical and logical reasoning, code generation, and physical simulation. Most notably, it achieves this performance with only a fraction of the training and inference costs of comparable models. This makes it economically viable to deploy world-leading LLMs on-premises and tailor them to domain-specific use cases.

3) Reinforcement learning and reasoning

Compared with other Chinese language models, DeepSeek R1 can more reliably parse complex Chinese sentences and accurately understand complex reference relationships and implicit meanings in them. Its open chain-of-thought reasoning process is similar to the human thinking process. It is even capable of self-reflection and deduction.

2. Tailoring DeepSeek to IT and Application System Operations

Building a specialized model from a general foundation model is a complex project that involves the collection,



cleaning, and processing of domain-specific data; model fine-tuning, training, evaluation, and validation; and model application development and promotion.

- » Data collection and cleaning: Develop SRE knowledge bases by integrating application system O&M logs and monitoring data, troubleshooting case studies, O&M operation manuals, emergency response manuals, official documents and maintenance manuals for software products (such as Oracle manuals and Kylin system maintenance manuals), CMDB data of application and device instances, and topology relationship data.
- » Supervised fine-tuning (SFT): Fine-tune DeepSeek R1 on SRE data to enhance its understanding of SRE terms, processes, and scenarios. Use it to generate in-depth reasoning data for simulated SRE scenarios (such as fault diagnosis steps), and on top of it, create high-quality SFT datasets through manual labeling.

» Reinforcement learning: Build a reward model based on SRE metrics and knowledge accuracy metrics, and couple it with reinforcement learning algorithms such as PPO to optimize the model's performance in complex operations decision-making.

» Model deployment and application: Create a local SRE knowledge base, and connect the model to this database. Allow access to the model via APIs to enable functions such as real-time fault query and automatic script generation. The model supports natural language interaction and multi-modal input.

This custom model training process applies to all LLMs, not just DeepSeek R1. The reason for choosing DeepSeek R1 is because it is open-sourced and delivers performance on par with some of the leading models, despite its significantly lower training and inference costs. Below is a comparison:

Dimension	DeepSeek R1	Traditional AI Solution
Development cycle	2-4 weeks (based on a pre-trained model)	6+ months (train from scratch)
Hardware cost	Edge nodes (16 GB vRAM)	V100 server cluster
Knowledge update efficiency	Incremental training: 1 hour/time	Full training: >1 week
Adaptability	A dynamic MoE architecture supports log analytics, root cause analysis, and script generation.	Multiple specialized models required
Long-tail problem solving	Continuous detection of rare fault modes through reinforcement learning	Relies on a rule library, high missed detection rate

02 SRE Use Cases

The benefit-to-cost ratio (BCR) is a crucial consideration in deploying localized SRE AI models. Developing an SRE AI model takes a lot of time, money, and manpower. If you are uncertain that the model will deliver the expected performance and efficiency in handling complex SRE tasks, you shouldn't rush to start the project. This part of the article will introduce you to some of the most viable LLM use cases in the world of SRE and IT operations.

1. An intelligent Q&A system for SRE

The SRE knowledge base is relatively easy to implement. With capabilities such as text processing and search, and in-context understanding, an LLM can provide users with useful knowledge, such as explanations about domain-specific terms and procedures for handling specific issues. Examples include the change request procedure and emergency handling procedures for specific database exceptions. General models have been proven to be able to handle this type of tasks through natural language interaction. In the case of SRE and IT operations, the model needs to be re-trained or fine-tuned on specialized knowledge bases. The solution is quite mature. The main challenges, however, lie in data preprocessing and cleaning, model training, and model accuracy evaluation.

The following is a simplified description of this process:

1) Phase 1: Data preparation and knowledge base creation

Knowledge collection

- » Integrate data such as O&M documents, service ticket records, and failure modes and case studies. Use Markdown or structured tables.
- » Clean the data, including denoising (such as redundant logs) and marking key entities (such as server IP addresses and error codes).

Knowledge vectorization

- » Use the Embedding API of DeepSeek R1 to convert text into vectors and then apply dynamic blocking (for example, segment by paragraph or semantics).
- » Store the data in a vector database and fine-tune indexing parameters (such as HNSW graph layers) to improve recall performance.

2) Phase 2: Model deployment and tuning

Environment configuration: local deployment

Model enhancement

- » Domain adaptation: Connect the model to SRE knowledge bases and use RAG and prompt engineering (e.g., adding the system prompt "You are a senior DBA") to enable the model to generate reliable, professional answers.
- » Performance enhancement: Use distillation to create a lightweight model, or use INT4 quantization to reduce inference latency while maintaining acceptable accuracy.

3) Phase 3: System integration and function development

Workflow engine development

- » Use Flowise AI or Anything-LLM to configure dialog chains and integrate different modules, including LLM, knowledge search, and context management.
- » Enable multi-turn dialog, memory, and source traceability. Include links to the sources in LLM-generated answers.

Key features

- » Alarm analysis: Connects to the O&M and monitoring system to automatically parse alarm information and trigger knowledge search.
- » Proactive diagnosis: The dynamic chain of thought (CoT) technique guides the model to break down questions or problems (for example, high CPU load -> check processes -> analyze logs).

4) Phase 4: Verification and iteration

Performance evaluation

- » Build test sets that cover all common scenarios (such as slow SQL optimization and DR switchover). Use manual scoring and automatically calculated metrics (BLEU and ROUGE) to quantify model accuracy.
- » Mitigating bad cases: Adjust the blocking policy, enrich knowledge bases, or add policies to reject certain questions.

Continuous iteration

- » Optimization based on feedback: Undesirable or incorrect answers are automatically marked based on user ratings. Their corrections are periodically integrated to fine-tune the model.
- » Dynamic update of knowledge bases: Set a scheduled task to ingest the latest O&M documents and trigger incremental updates of the vector database.

2. Generating and reviewing standard change procedures

The performance of leading LLMs, such as DeepSeek R1, in writing scripts and programs may have already surpassed average developers. For example, in ITOps, DeepSeek R1 can be used to write a standard change procedure or script for a specific function, for example, modifying Kylin OS parameters or upgrading the kernel. It can also automatically check script compliance (e.g., flagging high-risk commands such as rm and drop), correct logical errors, and generate standard operation guides. To ensure safety and accuracy, however, scripts or procedures generated by an LLM should always be manually tested before being deployed in a production system.

3. Generating emergency response plans based on alarms

When multiple alarms (e.g., CPU usage spikes and long lock waits) are generated simultaneously for a system, manual troubleshooting often proves too slow. By associating alarm contexts in real time, including the application system's topology architecture and specific alarm information, DeepSeek can quickly generate an emergency response plan and handling suggestions, including an assessment of the blast radius and specific impact. The O&M team can then review the plan and determine whether to execute it. For example, for a single error code generated for software, the model can generate specific handling suggestions and procedures. In a more complex situation, such as a system-level failure, it must generate a more comprehensive plan, determining whether to switch or throttle traffic and even switch to another database, as well as the specific impact on both upstream and downstream components.

4. Writing incident review reports based on the incident handling procedures and alarms

Based on the recorded incident handling procedures, the timeline (from first alarm to recovery confirmation), relevant log records, and operation records, a pre-trained report generation model can generate intuitive incident review reports. These reports include time sequences and root cause topology, structured as "impact -> handling procedures -> root cause analysis -> improvement measures". LLMs like DeepSeek R1 are good at generating reports and summarization.

The key challenge lies in collecting and analyzing relevant logs and data based on the incident handling procedures, processing them, and arriving at the right conclusions.

5. Enhancing DDL and SQL auditing for databases

The DeepSeek audit plug-in can be integrated into the application deployment process to check user-submitted SQL and DDL statements for security, compliance, performance, and other purposes. It does so based on predefined auditing rules and knowledge about the syntax of various database languages. Specifically, it checks the following: 1) Whether the statements match the target database version; 2) commands for full table scanning and query, to which it generates a warning; 3) plaintext passwords or excessive permissions in the commands; (4) DDL commands that would change the database schemas, in which case, it predicts the impact (such as downtimes) and how long the change would take. The final conclusion (blocking or warning) is sent to each DBA and project team for further review and confirmation.

The following is a simplified description of the implementation process:

1) Core components

- » Rule library: Use DeepSeek R1 to predefine auditing/review rules (such as index specifications and field naming constraints) for each type of database.
- » Semantic parsing layer: Use DeepSeek R1 to parse the semantic meaning of SQL statements and review multiple statements together in context.
- » Static audit engine: Use RAG to search a vector database for rule matching.
- » Dynamic analysis layer: Perform physical validation based on the MySQL metadata and execution plan.
- » Improvement suggestion module: Automatically rewrite SQL statements for better compliance.

2) Rule customization

- » Use DeepSeek R1 to parse documents on database development specifications, automatically generate executable auditing rule templates, and define SQL and DDL auditing rules for each type of database.
- » Build a custom model through fine-tuning to identify specific business patterns and rules (such as account ID rules for the financial industry).

3) Multi-dimensional auditing

- » Static auditing: DeepSeek R1 searches knowledge bases to verify naming and indexing rules.

- » Dynamic verification: Check the presence of database tables and foreign key constraints.
- » Performance prediction: Predict the number of scanned rows and index usage based on previous performance statistics.

4) Tiered conclusion

- » Fatal error (e.g., absence of a primary key): block directly
- » Warning (no index used): generate an optimization suggestion

5) Closed-loop

- » Automatically generate audit reports that contain suggestions for correction or remediation.
- » Automate DDL/DML processes by interconnecting with the service ticket system via an API.
- » Create a learning on feedback mechanism and continuously optimize the auditing rule library.

6. Converting SQL statements during the migration and adaptation of XC databases

A significant challenge during the migration and adaptation of XC databases is ensuring the accurate conversion of SQL and DDL statements across different database systems. This challenge arises from the differences in the syntax and semantics used by different database languages. Based on the official documents and SQL/DDL syntax rules for the source database, DeepSeek can convert and optimize the SQL syntax and table schemas for the target database, enabling more efficient database migration. For example, in the case of table schema migration, after parsing the DDL of the source database, DeepSeek can automatically adjust the data types (e.g., changing NUMBER to DECIMAL) and index policies (e.g., replacing function-based indexes with virtual columns), process empty strings, and generate compatibility solutions for complex structures such as partitioned tables. After the conversion, a differential test can be performed. Test cases are automatically generated and executed to compare query results between the source and destination databases, making sure there is no data or functionality loss.

Database vendors typically integrate this feature into their own database migration tools.

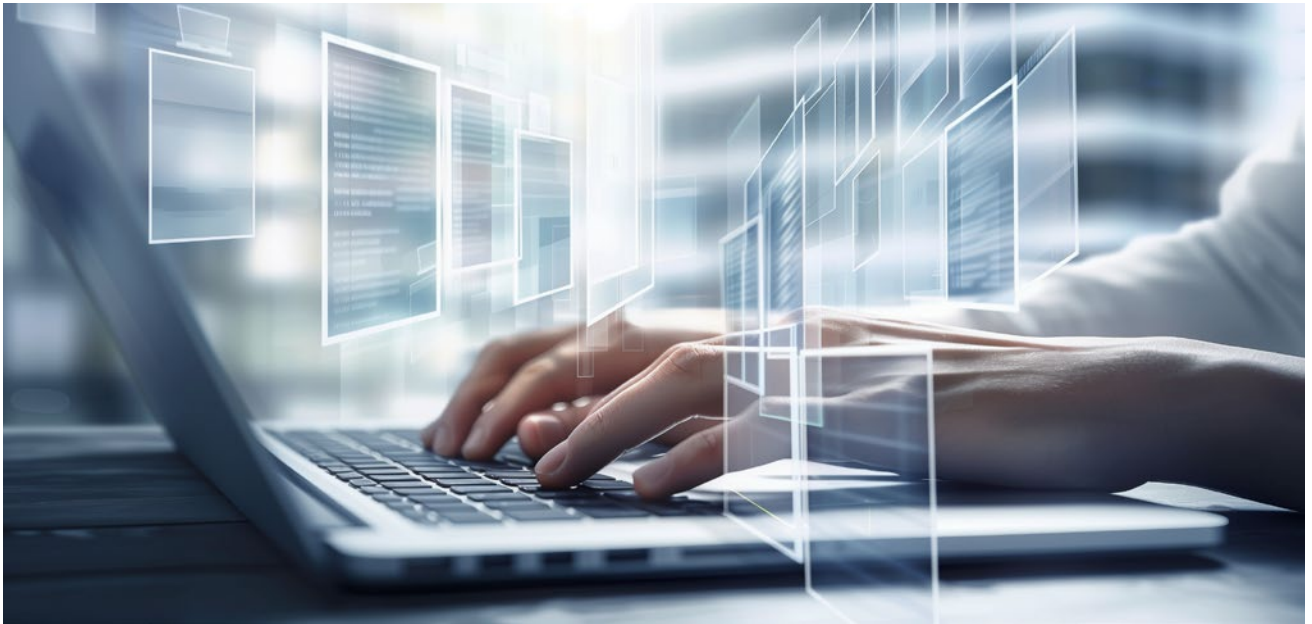
7. Application system performance and capacity evaluation

Train a time series prediction model on existing O&M and monitoring data (e.g., CPU, memory, I/O, and storage) to

simulate resource consumption curves under different load conditions. Use DeepSeek to analyze the dependency chain based on the application topology. The model can, for example, predict that the transactions per second (TPS) at which the ordering service calls the payment service will soon exceed the upper limit supported by the current thread pool. Then, it can calculate the number of pods or server resources that need to be added for capacity expansion. For a storage system, the model can sample and analyze the tables' growth rates and indexing efficiency to predict the disk usage six months later. The final output is an evaluation report containing resource water levels and heatmaps, identified performance bottlenecks, and capacity expansion suggestions. Dynamic alarm thresholds can be configured. Based on the capacity evaluation report and visualized indicators, application systems and servers can be scaled up or down to improve resource pool utilization.

8. Quick fault locating and root cause analysis

When an application system becomes faulty, quick fault localization and root cause analysis are the most essential part of an emergency response and also the most complex steps. They may involve all parts of the system, including software, hardware, applications, systems, networks, and storage. A stream computing engine can be used to aggregate logs, performance metrics, and link tracing data in real time, and then DeepSeek can create a dynamic service dependency graph. When an alarm is generated, a causal inference algorithm is used to locate the root cause. For example, if the transaction durations of an application increase drastically, the upstream and downstream call chains can be analyzed to try to identify the cause. In one example, the cause was an I/O exception on a shard server in the underlying distributed database cluster. By matching these against similar failure modes and cases in the database, a probabilistic conclusion similar to the following can be provided: There is a 90% possibility that the cause is an I/O exception of a database server. Finally, the fault propagation path and scope of impact are highlighted in the application topology view, and emergency handling suggestions (such as database switchover) are offered. The training and inference costs for such a model are quite high, and real-time, accurate performance metrics are also required.



03 Conclusion

In fact, there are a lot more use cases for LLMs like DeepSeek in the world of SRE and IT operations. More examples include using an LLM to anonymize audit logs and analyze client operation logs, and using RAGFlow for workflow orchestration and management. The following are important factors to consider when deciding to deploy an LLM, whether it is DeepSeek or another, to power SRE use cases:

The benefit-cost ratio: If the costs of developing an AI far outweigh its foreseeable benefits, you should not rush to start the project. In the case of developing an SRE AI model, the costs may include those in model procurement,

deployment, and custom development; data processing, maintenance, and integration; risk control, fault tolerance, and compliance. The benefits may include labor cost savings, faster troubleshooting and emergency responses, lower failure rate, lower compliance auditing cost, potentially enhanced operational capabilities, and knowledge accumulation.

All LLMs hallucinate, some more, some less. Reports indicate that DeepSeek R1 has a hallucination rate over 14%, which is significantly higher than other reasoning models. When using an LLM, it is essential to verify its results. If the output falls outside of your domain of expertise, you may need to cross-examine

it with other LLMs or reliable sources. Otherwise, you may end up with the hallucination of some AI, which could lead to unpredictable outcome when applied in real-world scenarios or production systems. For instance, in the case of an operations system, an LLM-generated instruction might worsen the situation instead of resolving it. Therefore, in some of the use cases I described above, LLM outputs can only be used as reference only. It is essential to verify them. For example, any SQL or DDL statements generated by an LLM must be tested before they can be used in a production environment.

— Reposted from the author's social account "The Direction of a Shepherd"

Cloudification of Enterprises' Core Services: Experience in Availability Assurance



Ding Xiaohong

Director, Huawei Cloud Computing
Delivery & Service Dept
Former Director, GTS Software
Service Dept
Founder, GTS Software SRE O&M
Center



Li Heqing

SRE O&M Expert, GTS Software
Service Dept
Owner, Software SRE O&M Center

Abstract

The cloudification of enterprises' core services significantly changes the traditional O&M mode. Ensuring on-cloud system availability poses a critical challenge in every software cloudification project. This article describes GTS Software's practical experience in providing availability assurance after the cloudification of enterprises' core services, serving as a reference for similar products or projects.

01 Background

With the rapid development of public cloud services, an increasing number of enterprises are migrating their core service systems to clouds. Cloud services are redefining the traditional IT architecture at an unprecedented speed. According to Gartner's report, global cloud service revenue is expected to increase by 138% by 2025. At that time, the cloud service market is expected to surpass the traditional IT service

market and become enterprises' primary focus of technology investment.

As a key software provider in the carrier field, GTS Software has developed multiple core service software applications. In traditional scenarios, GTS Software provides software applications to carriers, who then deploy and maintain them in their local equipment rooms to deliver services to external systems. Following the trend of software

cloudification, GTS Software now deploys the system on a public cloud platform and handles the system's operations and maintenance (O&M). Carriers simply subscribe to the services to provide them to their customers. This article describes GTS Software's practices on availability assurance after the cloudification of carriers' core services and explores methods to maintain system stability and efficiency.



02 Changes and Challenges

Compared to traditional scenarios, software services in cloudification scenarios undergo the following changes:

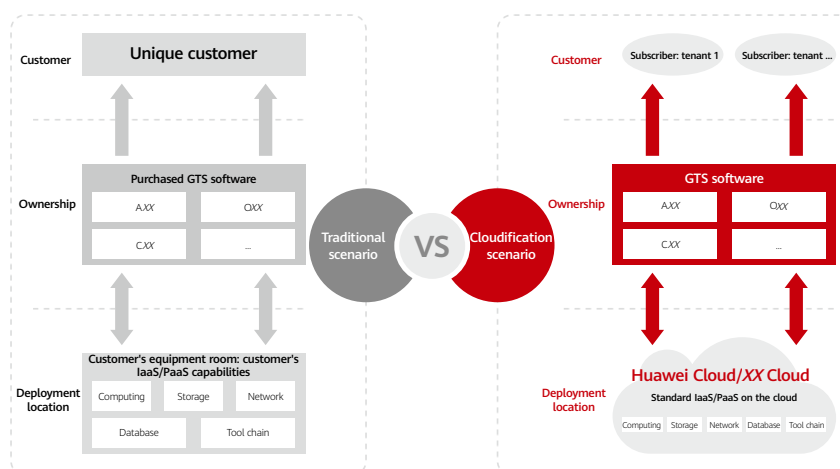
- » System deployment has moved from customers' equipment rooms to clouds, such as Huawei Cloud or clouds provided by other service vendors.
- » Software ownership transitions from customers to software suppliers, such as GTS Software.
- » The customer form evolves from a single customer to multiple customers in multi-tenant mode.

The changes present the following challenges:

- » Challenge 1: Following cloudification, software ownership shifts to software suppliers, who assume responsibility for O&M and become service providers, and the O&M mode transitions from passive to proactive O&M.
- » Challenge 2: The O&M organizational structure becomes complex. After

cloudification, southbound IaaS O&M relies on the cloud SRE, and northbound services depend on the customer's system, necessitating collaboration between the software supplier, cloud SRE, and carrier for end-to-end O&M. This O&M organizational structure is more complex than the bidirectional collaboration in traditional scenarios.

- » Challenge 3: Typical cloudification service flows involve on-cloud IaaS devices, on-cloud software service systems, private network lines, and the customer's third-party systems. The system's external borders become complex, leading to increased challenges in fault demarcation and emergency recovery.



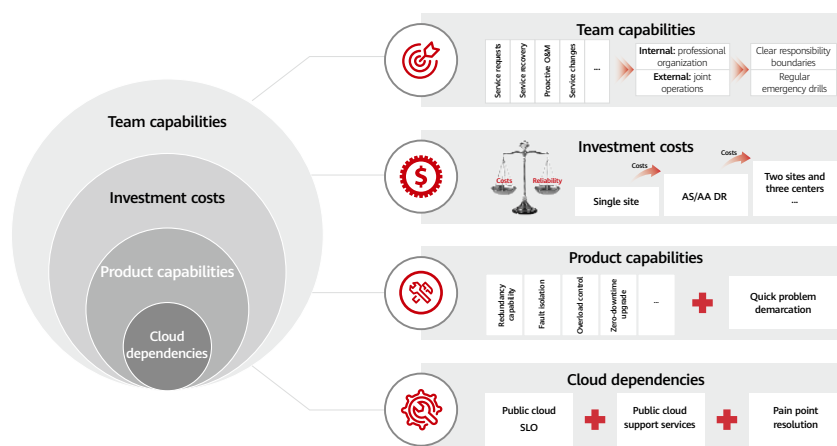
03 Solution

Over the past three years, the GTS Software SRE O&M team has implemented a series of practices on availability assurance following software cloudification. This section describes the practices from the perspectives of strategy, team, cost, product, and cloud.

1. Strategy: Defining a Model to Systematically Develop the Availability Assurance Capability of Software Cloudification

Developing the availability assurance capability of enterprises' core service cloudification is a systematic project. It relies on public cloud services, product capabilities, industry investments, and team capabilities to discover and resolve problems and recover services. To address this, we define a model that describes the availability assurance hierarchy for software cloudification, as illustrated in the figure on the right.

Availability assurance hierarchy model for software cloudification



Availability assurance hierarchy model

- » **Team capabilities:** After service cloudification, a professional assurance team needs to be built to proactively conduct O&M. The team needs to collaborate with the cloud and customer for regular drills to achieve the availability target.
- » **Investment costs:** Investment costs determine the upper limit of availability. Balancing investment costs and availability is crucial, as higher availability typically comes with higher investment costs. Achieving the optimal balance between investment costs and system availability involves a comprehensive analysis of service development and customer requirements.

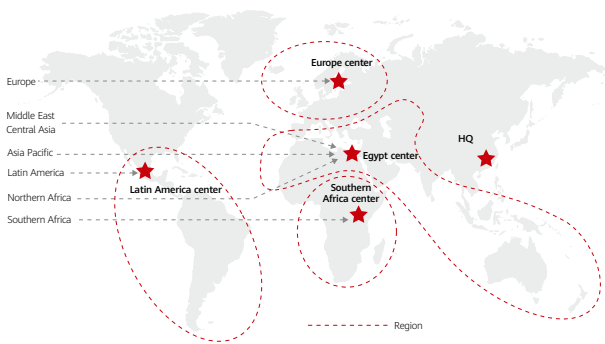
- » **Product capabilities:** Product reliability is the cornerstone for achieving the availability target. Other capabilities, such as rapid fault discovery, demarcation, and recovery, and planned failure duration, form the basis for ensuring availability after cloudification.
- » **Cloud capabilities:** Public cloud capabilities determine the lower limit of availability. After cloudification, infrastructure (IaaS) resources are provided by the cloud service provider. The committed availability service level objective (SLO) and the offered support services establish the foundation for availability assurance.

2. Team: Building an Efficient Assurance Team

Practice 1: Setting Up an SRE O&M Center to Maintain Service Availability and Customer Satisfaction

GTS Software's on-cloud services span various global regions. To maintain high service availability, we have set up the Software SRE O&M Center. After three years of strategic development, we have built a robust global structure, comprising a headquarters (HQ) and four regional centers. This framework enhances the scalability of our worldwide services in compliance with local laws and regulations. Additionally, it effectively overcomes challenges related to time zones and language barriers, providing a strong foundation for seamless and stable service O&M.

Structure of the Software SRE O&M Center



SRE Center	Location	Region
Europe SRE O&M center	Romania	Europe
Southern Africa SRE O&M center	Kenya	Southern Africa
Latin America SRE O&M center	Mexico	Latin America
Egypt SRE O&M center	Egypt	Middle East, Central Asia, Northern Africa, and Asia Pacific

The following table describes the positioning and responsibilities of HQ and regional centers.

Center	Positioning	Responsibility
HQ	Capability center	<ul style="list-style-type: none"> Leads the overall planning and development of tools, processes, and capabilities at SRE O&M centers. Accelerates the improvement of product maintainability and reliability. Introduces capabilities to regional centers.
	Operation center	<ul style="list-style-type: none"> Supports customer satisfaction management, major problem handling, and assurance operations at regional centers.
Regional center	Operation center	<ul style="list-style-type: none"> Responsible for the availability of regional services and ensures the achievement of availability targets at regional centers. Leads localization development by developing SRE capabilities with local talent. Establishes a mechanism for regular communication with key customers, holds regular benchmarking meetings with customers, and conducts emergency drills to enhance customer experience and satisfaction.

Application: SRE O&M centers have been established across six regions outside China, supporting the SRE O&M work of XX projects.

Typical case: The SRE war room engages with customers to enhance customer confidence and satisfaction.

Background: Customer A is the largest enterprise in its industry within a country and is among the world's top 500 enterprises. After its S service is migrated to the cloud, the customer expressed concerns about the lack of in-person contact with the O&M team and the effectiveness of ensuring system stability.

Practice: The Software SRE O&M Center quickly responded to the customer's concerns by organizing remote video communications with the customer's technical owner and expert team. The communication process is carried out through the SRE war room. The communication content includes detailed descriptions of SRE personnel setup, organizational structure, O&M processes, and key capabilities such as proactive monitoring, proactive O&M, change management, and problem handling, all tailored to address the customer's concerns. Through clarification and communication, the customer's concerns were addressed, and their confidence in the system's stability was enhanced.

Effect:

- » The customer appreciated the efficient and professional communication, mentioning that it enabled them to see our systematic approach and professionalism, which alleviated their anxiety.



» In addition, a mechanism for regular communication with the customer was established.

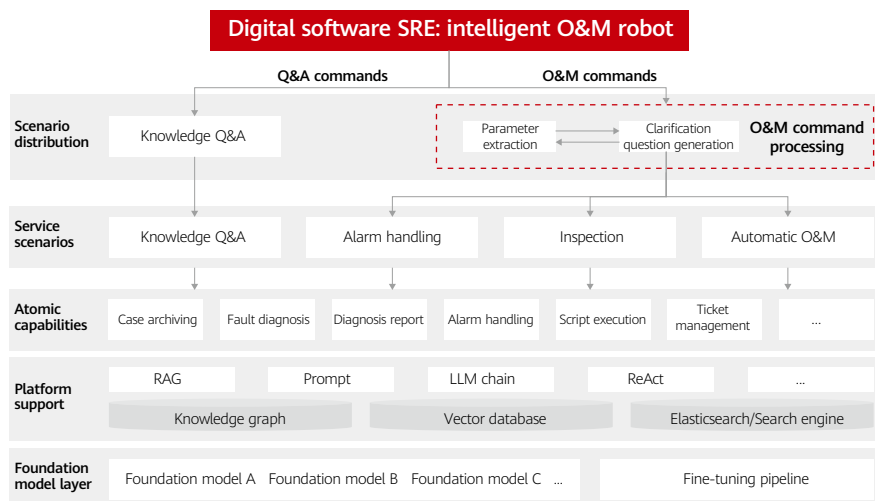
» This practice is being implemented across multiple key customers worldwide.

Practice 2: Building an Intelligent O&M Robot and Leveraging AI to Improve O&M Efficiency

The team members of the Software SRE O&M Center were located in China and four countries (regional centers). Efficiently sharing global O&M knowledge became a significant challenge for organizational development. To tackle this challenge, we

developed an intelligent O&M robot named "digital SRE" using the large language model (LLM) technology to facilitate global O&M knowledge sharing and automate O&M tasks.

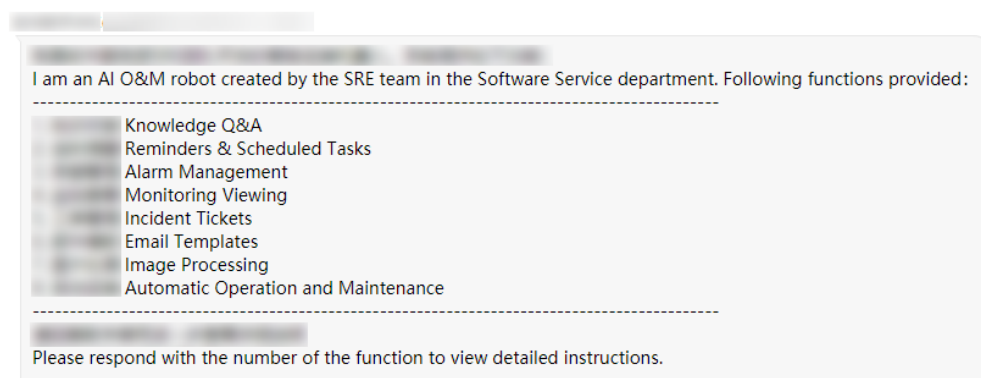
As illustrated in the following figure, when O&M personnel send a message to the digital SRE, the system checks if it is a Q&A or O&M command and processes it accordingly. The digital SRE's bottom layer integrates internal foundation models and a fine-tuning pipeline. Leveraging platform support and atomic capabilities, the digital SRE implements the entire process from knowledge Q&A to automatic O&M.



Typical applications

- **Knowledge Q&A** for alarms, emergency, services, etc.
- **Alarm management**: viewing, analysis, handling, masking, etc.
- **Inspection**: viewing, analysis, handling, email notifications, etc.
- **Ticket management**: creation, dispatch, viewing, reminders, etc.

Architecture and typical applications of the digital software SRE



Intelligent robots are essential assistants in O&M centers aiding in daily tasks.

O&M responsibility matrix in the traditional scenario				O&M responsibility matrix in the cloudification scenario				
		Customer	Software			Customer	Software	Public cloud
Customer-side third-party O&M	Customer-side third-party system maintenance	R		Customer-side third-party O&M	Customer-side third-party system maintenance	R		
	...	R			...	R		
Software service O&M	Service monitoring	R	S	Software service O&M	Service monitoring	S	R	
	Proactive maintenance	R	S		Proactive maintenance	S	R	
	Problem handling	R	S		Problem handling	S	R	
	Change handling	R	S		Change handling	S	R	
	...	R	S		...	S	R	
Device management	OS	R		Cloud tenant management	Cloud tenant management		R	S
	Network management	R			Rental management		R	S
	Storage resources	R			Public cloud service management		R	S
	Computing resources	R			...		R	S
	...	R			OS			R
Equipment room	Ventilation, fire fighting, water supply, and power supply	R		Device management	Network management			R
	...	R			Storage resources			R
	...	R			Computing resources			R
				Equipment room	...			R
					Ventilation, fire extinguishment, moisture-proof, and electricity			R
					...			R

R: Responsibility
S: Support
Change

Practice 3: Defining the Responsibility Matrix and Implementing Collaborative Operations

To address the growing complexity of the O&M organizational structure, we identify differences in responsibilities across scope and organizational dimensions, defining a responsibility matrix for each stakeholder in the cloudification scenario.

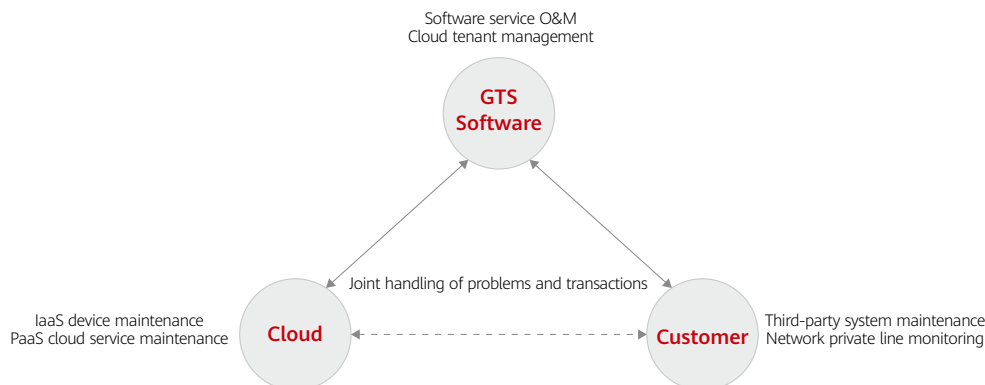
As illustrated in the preceding figure, the changes in the O&M responsibility

matrix in the cloudification scenario are as follows:

- » O&M organizations: A public cloud team is added.
- » O&M scope: Cloud tenant management is added.
- » O&M responsibilities:
 - The customer retains responsibility for customer-side third-party O&M.
 - The software team is responsible for service O&M with customer support.

- The software team is responsible for managing cloud tenants with public cloud support.
- The public cloud team is responsible for device management and other tasks.

Based on this responsibility matrix, we collaborate with cloud service providers and customers in key projects. By now, we have established joint teams with two public cloud service providers across six regions and are conducting joint assurance activities with our customers.



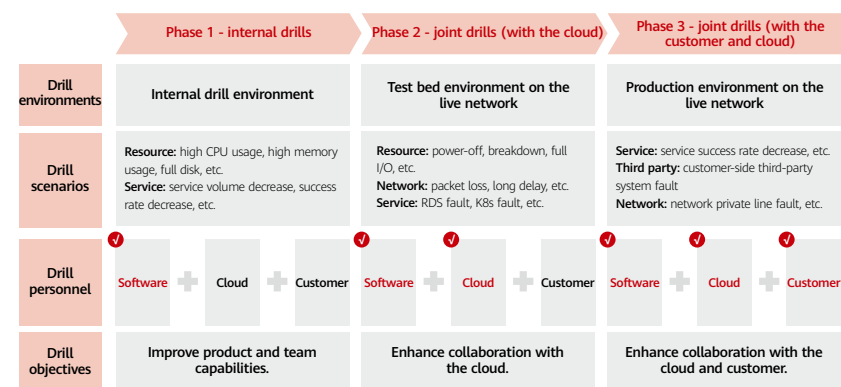
Practice 4: Identifying Key Fault Points and Conducting Regular Cross-Team Drills

Service cloudification increases the complexity of service processes, hinders inter-organization collaboration, and presents challenges in fault demarcation and emergency recovery. To address this, we identify key fault risks from an end-to-end perspective and develop corresponding emergency plans. Additionally, we conduct regular emergency drills in collaboration with cloud service providers and customers to identify potential issues in products, teams, and processes. This enables us to make targeted improvements that continually enhance our emergency recovery capabilities.

Our approach involves defining necessary emergency drills for each product or project, which are categorized into three phases, as illustrated in the figure on the right.

Phase 1 - internal drills: In the internal R&D drill environment, continuously conduct automatic red team/blue team drills to identify and rectify product reliability weaknesses in typical service fault scenarios and improve software emergency recovery capabilities.

Phase 2 - joint drills (with the cloud): In the test bed environment on the live network,



conduct joint drills between the cloud and software to address cloud dependency risks and enhance emergency collaborative emergency recovery capabilities for cloud and cloud-software faults.

Phase 3 - joint drill (with the customer and cloud): In the production environment on the live network, after approval by the customer, conduct joint drills among the customer, cloud, and software to verify the robustness of the production environment and collaborative emergency recovery capabilities for typical service, cloud, and third-party faults.

The following figure shows the typical fault scenarios of a software product, involving

software service faults, customer-side third-party faults, and IaaS cloud faults. The leading and supporting recovery organizations are specified for each scenario to conduct emergency drills addressing these typical faults.

Currently, the Software SRE O&M Center incorporates regular drills jointly conducted with customers and clouds into its daily work. In 2024, a total of XX drills were jointly conducted with customers. According to customers' feedback, the drills made O&M transparent rather than a black box. This has not only garnered customer recognition for the value of O&M but also strengthened the team's emergency response capabilities.

Fault Type	Typical Fault		Fault Scenario	Emergency Response Team		
	Scenario	Sub-scenario		Cloud	Software	Customer
Service fault	Sharp service volume decrease	No service request received	Third-party network fault	S	S	R
			VPN/Private line fault	R	S	R
			Third-party entry fault		S	R
		Service request received	Packet loss	S	R	R
			VPN/Private line fluctuation	S	R	R
			Third-party interface fault		R	R
			Service fault		R	R
			R
Customer-side third-party fault	Third-party fault	Exception on line XX	Exception on XXX response		S	R
		Exception on channel XX	Exception on interface calls of channel XXX		S	R
		Exception on the XX service	Exception on interface calls of the XXX service		S	R
IaaS fault	Exception on public cloud services	ECS	Failed node restart after power-off	S	R	
			Storage fault (OBS or VM SSD storage fault)	S	R	
			Operating system suspension	S	R	
		RDS	Abnormal CPU/content/storage usage	S	R	
			Bucket read/write exception	R	S	
		ELB	Unavailable ELB service	S	R	
			ELB backend exception	S	R	
		S	R	
	Network fault	Abnormal network indicator	Abnormal packet loss rate, delay, and jitter during EIP access from the public network	R	S	S

3. Cost: Striking a Balance Between Investment Costs and Availability

The costs of service cloudification involve investments in hardware, software, personnel, and maintenance, all of which impact system availability.

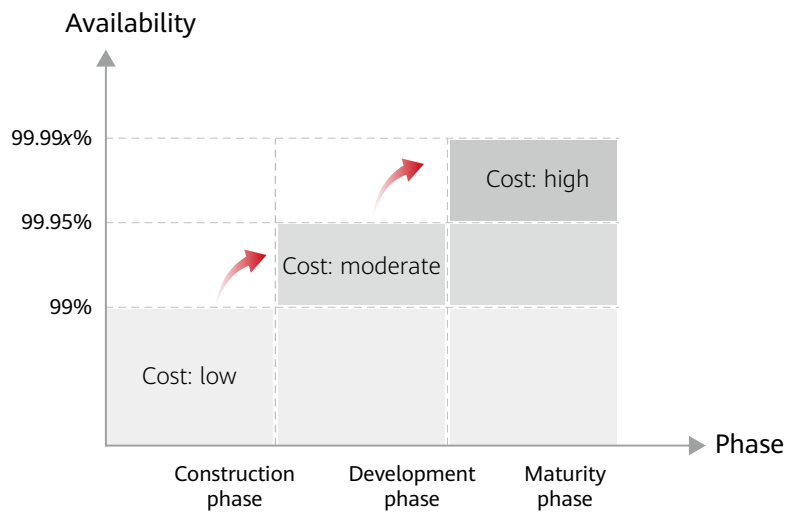
- » Achieving high availability requires additional resources, such as redundant servers, load balancing resources, and backup systems, which can lead to increased hardware costs. Conversely, reducing investments may lower availability and raise the risk of system faults or delays.
- » Different service levels correspond to varying O&M costs. Maintaining a high-availability system requires complex monitoring tools and more technical support personnel, resulting in higher costs. Conversely, lower availability can cut costs but may compromise user experience and service continuity.

Software cloudification needs to strike a balance between costs and availability, continuously optimizing practices based on service development requirements.

Practice 5: Increasing Investment to Improve Availability Based on Service Development Requirements

Selecting an appropriate networking solution based on project progress and customer requirements is a practice for balancing investment costs and availability in software cloudification.

- » Construction phase: Commercial services are not provided (or limited services



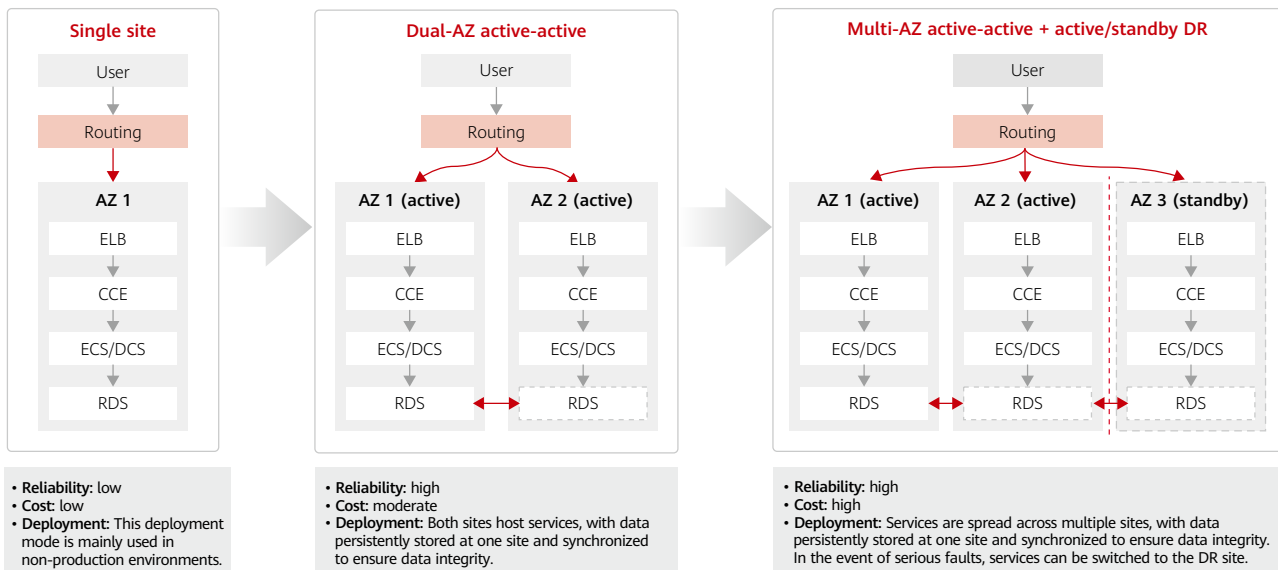
are provided with low availability requirements). This enables a system for user experience with minimal investment, facilitating market development.

- » Development phase: Services with average availability are provided, investment is gradually increased, and business and service commitments are strictly controlled.
- » Maturity phase: Services grow rapidly. The product and team's assurance capabilities have been verified in the early stage, and the project revenue can cover the investment costs.

Take product B, for example, which is deployed on a partner cloud, connects

to multiple customers, supports multiple tenants, and caters to small- and medium-sized carriers. The investment and reliability progression is as follows:

- » Phase 1: The product is in the customer experiencing and trial phase, with low reliability requirements and single-site networking at a low cost.
- » Phase 2: Some tenants with low availability requirements go online. Dual-AZ active-active networking is used at the application layer to improve availability at a moderate cost.
- » Phase 3: Tenants with high availability requirements go online. 3-AZ multi-active networking is used in the production environment at a high cost.





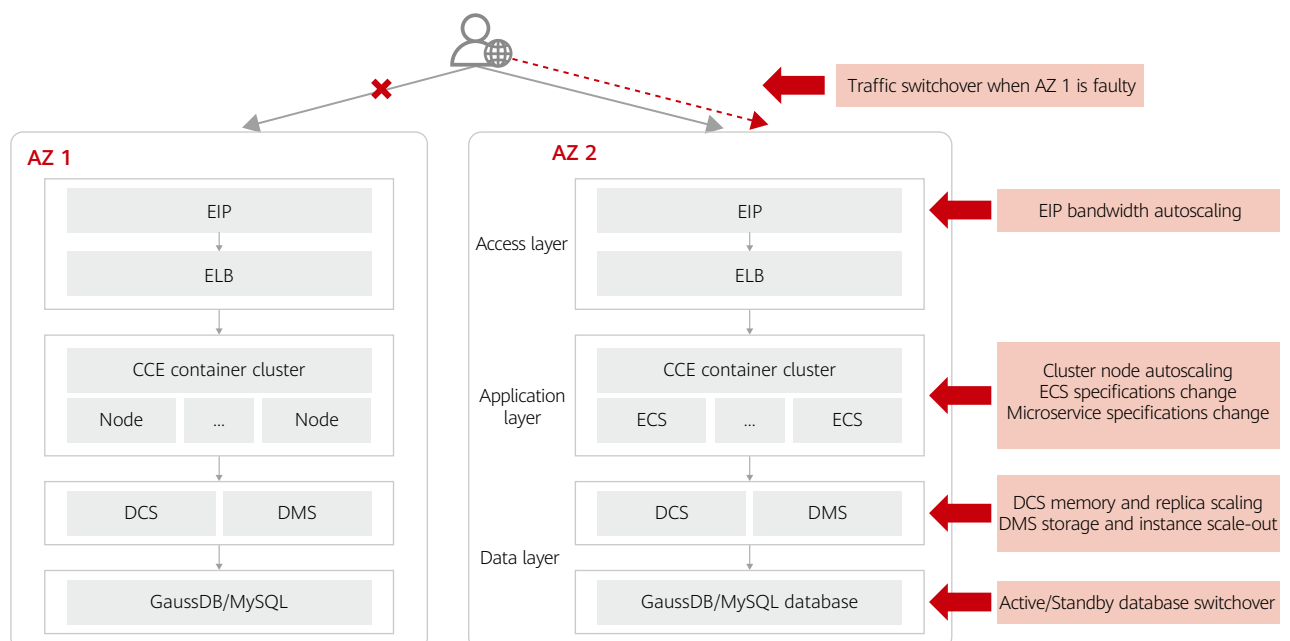
Practice 6: Using the Autoscaling Capability to Adapt to Service Changes

A product's autoscaling capability is crucial for balancing costs and availability in software cloudification. This involves dynamically changing the number of instances based on service loads to efficiently utilize cloud resources and achieve a balance between availability and costs in different phases.

Take product C, for example, which is deployed in dual-AZ active-active mode and features end-to-end autoscaling capabilities at different layers: the access layer (EIP), the application layer (CCE container cluster and ECS computing resources), and the middleware layer (DMS and DCS). The database is deployed in dual-AZ active/standby mode and supports automatic active/standby switchover. In normal scenarios, services run in AZ 1 while AZ 2 maintains minimal resources. When AZ 1 is

faulty, services are switched to AZ 2 through the access layer. After AZ 1 is recovered, services are switched back, and AZ 2 resources are released.

The relationship between costs and availability in software cloudification is directly linked. By selecting the right networking solution for each project phase and leveraging autoscaling capabilities to help the system properly configure resources, customers can achieve a balance and optimize cost efficiency.



4. Product: Enhancing Product Maintainability and Reliability

Software maintainability check scope

Proactive maintenance			Troubleshooting			Change management			Human-caused fault prevention
Alarm	Monitoring	Health check	Log availability	Fault diagnosis	Fault recovery	Upgrade and patch installation	Dynamic balancing	Configuration management	
Validity/Integrity check	Resource monitoring	IaaS inspection	Validity/Integrity check	Information collection	Fault isolation	Pre-upgrade check	Online balancing	Centralized configuration	Configuration foolproof
Alarm locating	Application monitoring	Service health inspection	Log classification and grading	Fault diagnosis	Overload control	Zero-downtime upgrade	Autoscaling	Configuration version management	...
Alarm clearance	Service monitoring	Inspection tool	Storage/Retrieval	Log tracing	System switchover	Automatic upgrade/rollback	
Alarm correlation	Business monitoring	Data clearance	...	Call chain	Backup and restoration	Comparison before and after the change			
Alarm suppression	Monitoring dashboard	...		Service call chain	Autoscaling	...			
...				

Product capabilities are crucial for ensuring system availability. Therefore, it is essential to establish a set of systematic product maintainability standards in the software, while specifying the minimum capability requirements. Each product must comply with these standards to be developed and launched for commercial use.

- » Proactive O&M: Develop proactive alarm, monitoring, and system health check capabilities to detect or predict problems quickly.
- » Troubleshooting: Develop capabilities related to log availability, fault diagnosis, and fault recovery to ensure quick fault demarcation and recovery.
- » Change management: Develop capabilities for zero-downtime version upgrade and patch installation to minimize planned downtime.
- » Human-caused fault prevention: Develop proactive foolproof capabilities to reduce faults resulting from misoperations.

Practice 7: Developing Key Product Reliability Capabilities

(1) Overload control

After cloudification, the system supports multiple tenants, whose service forms vary significantly, leading to potential service surges. To maintain system stability in surge scenarios, it is essential to have comprehensive overload control capabilities in place.

Before overload:

- » Plan and deploy resources in accordance with system capacity and performance specifications, including redundancy.
- » Analyze service flows, architecture, networking, and scenarios to identify bottlenecks and potential overload scenarios.
- » Design services and architecture to prevent heavy traffic, performance bottlenecks, and overload propagation.

During overload:

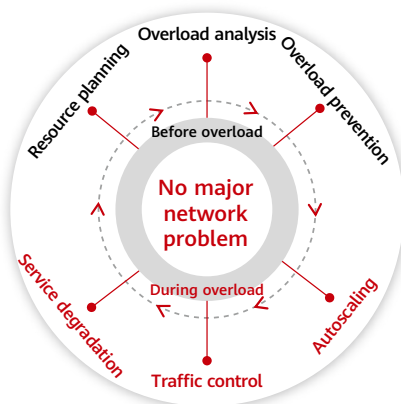
- » Enable autoscaling to dynamically adjust resource allocation online and improve the system's processing capability.
- » Control traffic based on the system's processing capability to adjust the service access volume.
- » Use a circuit breaker to conduct service degradation, prioritizing core services and interrupting non-core services.

After overload:

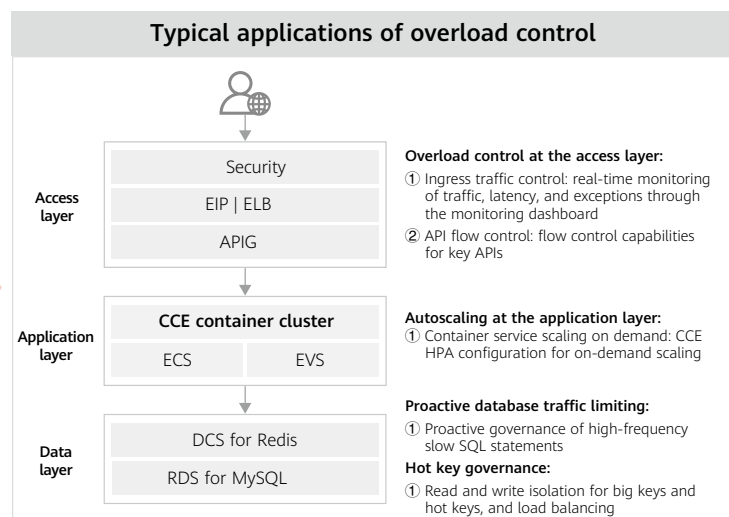
Release new resources added during autoscaling to reduce hardware costs.

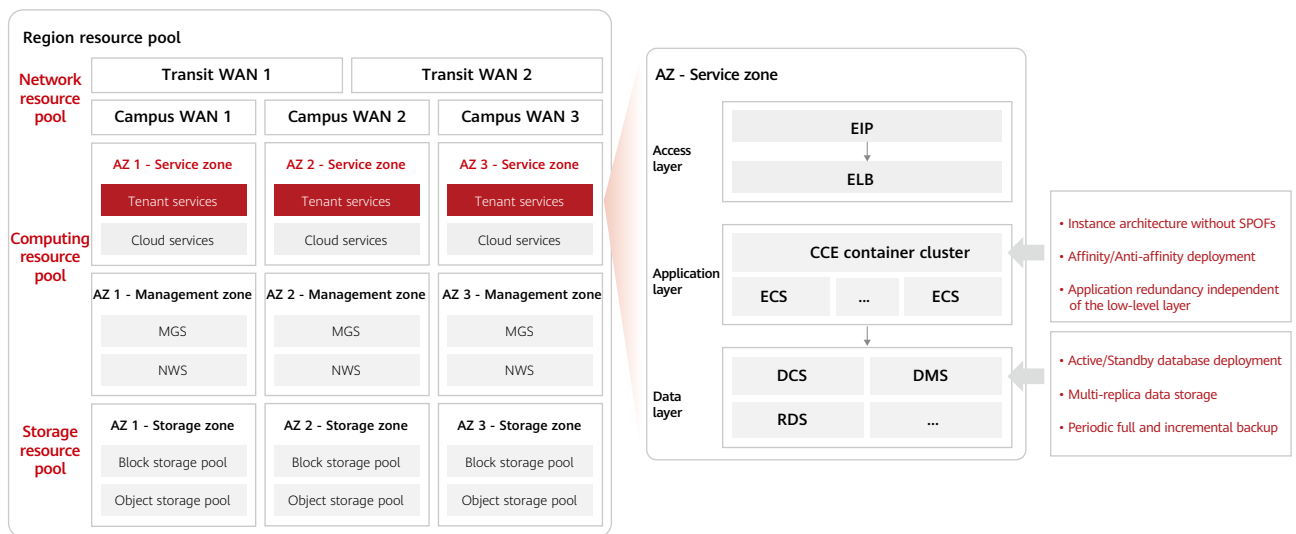
After cloudification, the system should have the capability to expand capacity flexibly. In the event of overload, the system should follow a sequence of autoscaling, service degradation, and at last service discarding.

Overload control policy of the cloud system



Typical applications of overload control





(2) Redundancy

When a single point of failure (SPOF) occurs in the system, redundant units can take over services from the faulty node to ensure that the system or device continues to function properly, improving service availability. Common redundancy technologies in software services include cloud resource pools, load balancing, active/standby deployment, and multi-copy backup.

In a public cloud site reliability solution, redundancy technologies are applied to software services mainly in the following ways:

- » Public cloud site construction: Adhering to the reliability requirements of three available data centers, implement multi-AZ multi-active or disaster recovery (DR) deployment for software adaptation.

» Public cloud resource pools: Divide resources into network, computing, and storage resources to prevent SPOFs in cloud services and provide support for the high availability of upper-layer application services.

- » Cloud-based software: Deploy software in the service zone with redundancy capabilities independent of the low-level layer. Typically, software is deployed in multi-instance, anti-affinity, or active/standby mode to ensure high service availability.

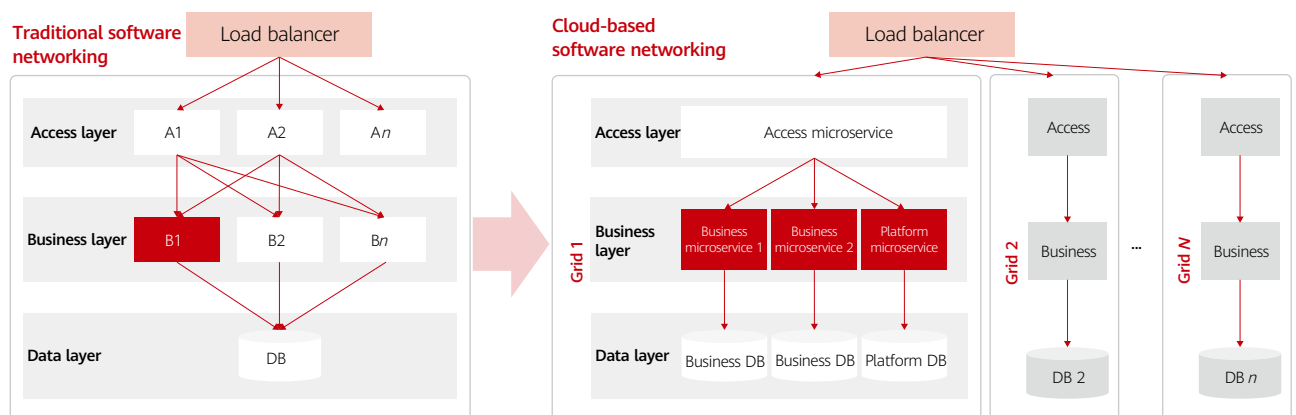
(3) Fault isolation

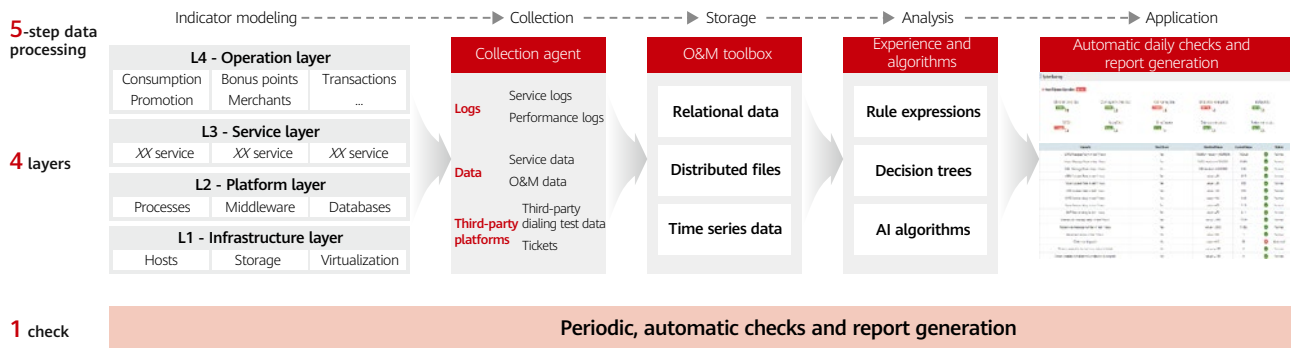
When a unit in the system is faulty, separate the faulty unit to limit the impact of the fault, ensuring that other units in the system continue to function properly.

Compared to traditional networking, the networking of product D after cloudification employs microserviceization and grid design to isolate faults more efficiently and minimize their impact.

- » Service application microserviceization: From an isolation perspective, an application is decomposed into multiple basic and platform services, with each service process representing an independent microservice and database. A service process fault does not affect other basic or platform services.

- » Grid design: Each grid encompasses a complete set of end-to-end service NEs and provides comprehensive system functions. User requests are forwarded to the appropriate grid based on specific routing rules. When a grid is unavailable, the service operation of other grids remains unaffected.





(4) System health check

To help O&M personnel stay informed about the health of the service system, we have developed an automatic system health check function. This function periodically and automatically collects system key performance indicators (KPIs), generates health reports, and pushes these reports via email or SMS message. This enables O&M personnel to monitor system status at any time and from anywhere, allowing them to identify potential risks in advance. Key features of this function include:

- » Comprehensive checks across infrastructure, platform, service, and operation layers

- » Data processing involving indicator modeling, data collection, data storage, data analysis, and report generation
- » Data analysis incorporating methods like performance/alarm analysis, log analysis, health experience rules, and AI-based prediction

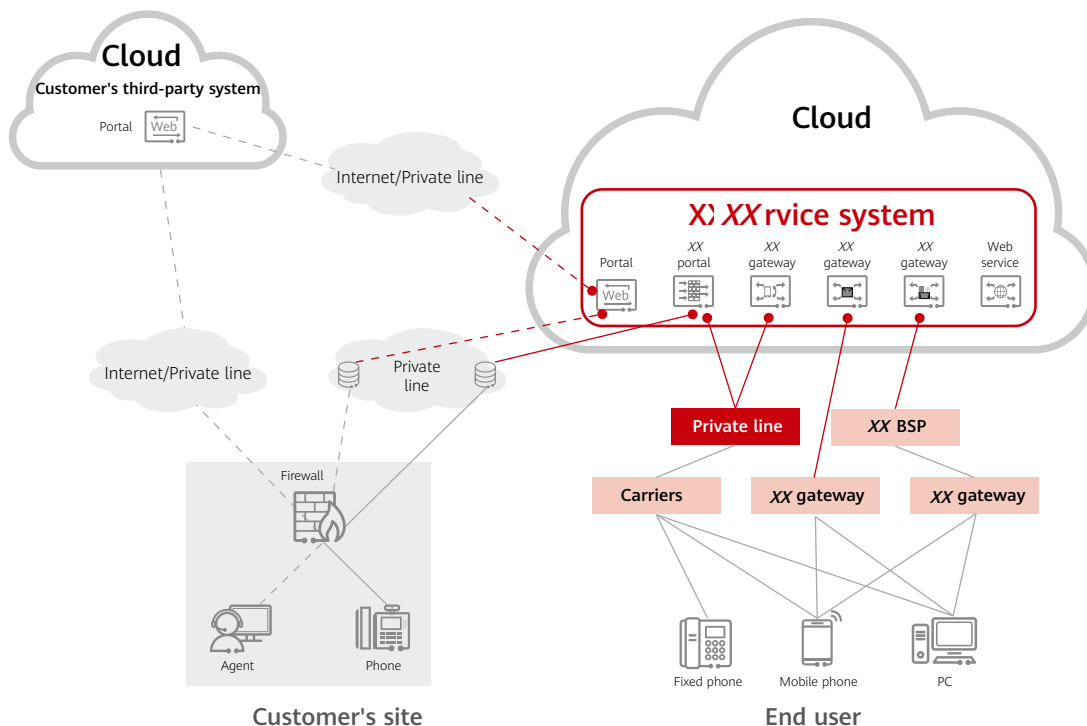
Currently, the system health check function has been implemented across all software service projects.

Practice 8: Improving the System's Fault Demarcation Capability

After the cloudification of a customer's core services, equipment room services must be connected using private network lines, or third-party systems must be integrated through interfaces. This complexity in system networking increases

external borders. When a fault occurs, it becomes a significant challenge for O&M personnel to locate the fault and provide a recovery solution quickly.

The following figure shows the connection between GTS Software's on-cloud services and the customer's other services. The red line represents the external border between the system, on the cloud, and the customer's third-party systems, which is crucial for fault demarcation. We enhance the fault locating capability through improved border monitoring and log recording. Key border faults are categorized into cloud faults and customer or third-party faults. The enhancement has been implemented across all cloudification projects, yielding impressive results.



Border	Monitored Object	Monitored Information
Cloud faults	ECS	20 KPIs, including the CPU, memory, I/O, OS, and SFS
	RDS	12 KPIs, including the CPU, memory, disk, long SQL, connection, and lock
	DCS	28 KPIs, including the CPU, memory, delay, volume, connection, and I/O
	Network	10 KPIs, including the bandwidth, lost rate, latency, and third-party QoS
Customer or third-party faults	Key NEs on the borders outside the DMZ	8 KPIs, including the traffic and latency of NEs like the API Fabric and NSLB
	External service interface	Overall external interface CAPS, TPS, success rate, and 7-day comparison data External interface CAPS, TPS, success rate, and 7-day comparison data by service type External interface CAPS, TPS, success rate, and 7-day comparison data by customer channel
	Network private line	10 KPIs, including the bandwidth, lost rate, latency, and third-party QoS

Case 1: Identify third-party problems ahead of the customer and win the customer's recognition.

Customer X has a dozen third-party channels, but lacks effective monitoring methods. The technical team's awareness of third-party issues lags behind that of the service team. To address this, we designed a third-party channel monitoring dashboard for the customer, helping the O&M team identify more than 20 third-party issues within half a year and garner recognition from the customer.

Case 2: Establish network latency benchmarks to address the customer's concerns.

Project Y operates in a country with outdated infrastructure and poor network conditions, which pose risks to its service operations. To mitigate this, we enhanced network latency monitoring for the project, established clear network latency benchmarks with the customer, and promptly notified them of any deviations from the standards. This proactive approach addressed the customer's concerns and earned their recognition.

5. Cloud: Enhancing the Support for and Collaboration with the Public Cloud

Practice 9: Enhancing Cloud Support Services Through Precise Matching and Insights

(1) **Precise matching: selecting the most suitable cloud service**

After service cloudification, the cloud provides IaaS devices and some PaaS services, with the cloud being responsible for their availability. Service personnel should be familiar with the availability indicators and support services offered by the cloud. They should select services based on their requirements.

- » Select the appropriate SL service with either 99.95% or 99.975% availability, based on the service features.
- » Select the appropriate cloud support service, specify a technical account manager (TAM) contact person, and establish an effective communication mechanism to address live network issues promptly.

For example, Huawei Cloud offers four levels of support services: Basic, Developer, Business, and Enterprise. GTS Software's on-cloud services cater mainly to medium- and large-sized enterprises that require a rapid incident response mechanism. Typically, the Enterprise support service is selected.

(2) **Precise insights: collaborating with the cloud to meet specific requirements**

In cases where cloud resources cannot meet the functional and performance requirements for special services, the service team should proactively communicate with the public cloud to address potential issues beforehand, thereby preventing major problems from occurring after service rollout. The following are some typical scenarios with specific cloud requirements.

Category	Scenario	Solution
Network	Services involve extensive cross-AZ communication with strict latency requirements.	<ul style="list-style-type: none"> Before delivery, proactively define cross-AZ communication indicators with the cloud. During delivery, conduct network quality tests in the cloud environment, using basic test cases.
	Packet loss occurs between two SLB instances during peak hours.	Adjust the physical NIC CT verification configuration on the host machine where the ECS is located to prevent packet loss on the cloud.
Disk	EVS disks on the cloud are hard disk drives (HDDs), whose response time may not meet requirements in certain situations.	Clarify disk performance requirements with the cloud and illustrate the impacts of slow response on services to prompt the cloud to replace HDDs with solid-state drives (SSDs).
Resource	A service requires VMs with GPUs, but GPU resource planning varies across different Huawei Cloud regions.	Communicate requirements with the cloud, understand the cloud's GPU resource planning and the service rollout time, and align service planning accordingly.
Process	Major changes on the cloud result in network failures in AZ 1, causing service unavailability for X minutes.	Establish a joint mechanism for addressing major changes and problems with the cloud to provide joint assurance on these matters.

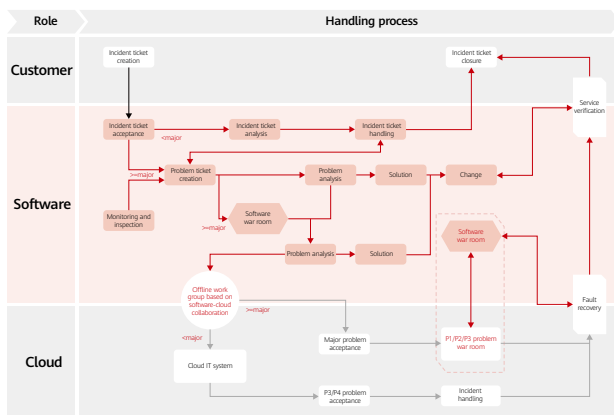
Practice 10: Establishing Collaborative Problem Handling and Change Processes with the Cloud

Note: GTS Software and Huawei Cloud belong to the same company, allowing for collaboration beyond typical commercial contracts. Other vendors cannot follow this practice directly. If necessary, communicate with Huawei Cloud separately.

GTS Software's on-cloud services serve medium- and large-sized enterprises that have high SLA requirements for system availability and major problem resolution. To effectively address major problems and service changes, we partnered with Huawei Cloud to establish regional offline assurance teams, define problem handling and change processes for both parties, and conduct joint assurance activities.

(1) Problem handling and collaborative operations process

- » Software: accepts customer incident tickets in a unified manner, assesses problem impacts, and determines problem severities; synchronizes cloud-related problems with the joint work group and cloud; in the event of major problems, establishes a software war room in accordance with software processes to resolve them.

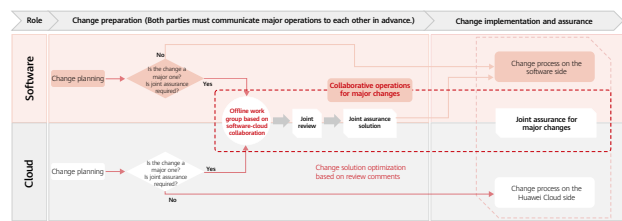


- » Cloud: creates trouble tickets simultaneously; in the event of major cloud faults, initiates war room operations in accordance with predefined rules.

- » War room: allows both parties to collaborate and share information to locate and rectify faults for problem resolution.

(2) Live network change process

- » Change planning: The cloud and software proactively notify the joint work group of planned major changes.
- » Joint assurance: For such major changes, the joint work group organizes both parties to jointly review change solutions, develop joint assurance solutions, and conduct joint assurance activities.
- » Change execution: The party that initiates a change follows its operation process. In the change preparation phase, the party assesses if joint assurance is required. If so, the party initiates joint assurance, and both parties determine the joint change and assurance solution.



04 Summary and Experience Sharing

After three years of construction, GTS Software has successfully deployed the Software SRE O&M Center globally, providing strong support for software service expansion worldwide and ensuring the smooth operations of XXX cloudification projects. After cloudification, system deployment locations change, and service processes are prolonged, making O&M collaborations more complex. Availability assurance becomes a systematic project, focusing on the following four aspects:

- » **Developing team capabilities:** Set up a professional SRE O&M team to ensure service availability on the cloud, collaborate with the cloud and customer, and regularly conduct cross-organization emergency drills.
- » **Balancing costs and availability:** Strike a balance between costs and availability based on industry trends, customer requirements, and product capabilities.
- » **Enhancing product maintainability and reliability:** Continuously develop reliability capabilities, including product redundancy and fault isolation, and improve the system's border monitoring and fault demarcation capabilities.
- » **Enhancing the support for and collaboration with the cloud:** Clearly understand the support services offered by the cloud, and establish a collaborative mechanism for addressing major problems and changes with the cloud.

Digital Resilience in Chile's Massive Blackout – The Technology Behind Navigating a Nationwide Power Crisis

Abstract

At three in the afternoon on February 25, 2025, a transmission line failure in northern Chile triggered a nationwide blackout, plunging over 98% of the population into darkness and paralyzing critical public infrastructure. Amid the chaos, multinational corporations, financial institutions, and public service systems supported by Huawei Cloud had zero interruptions. This resilience was no accident. Thanks to innovative technical architecture and globally distributed infrastructure, Huawei Cloud empowered its customers to navigate three critical challenges.



01 Challenge 1: Sustaining Data Center Operations During Prolonged Power Loss

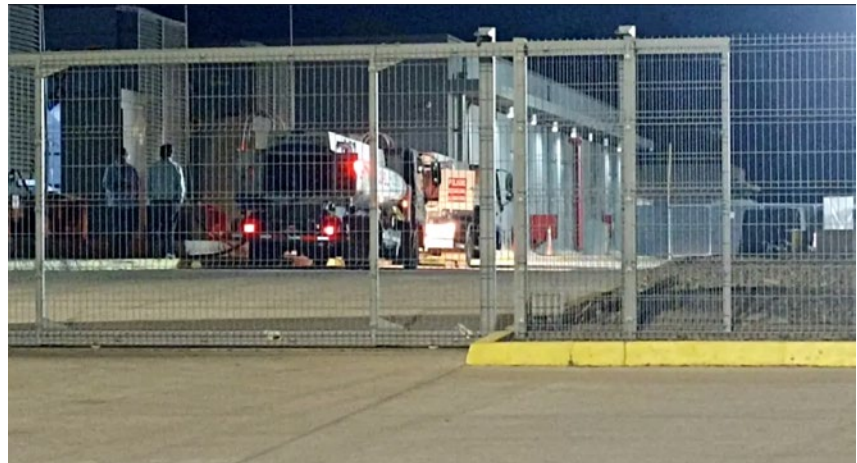
To withstand extreme grid failures, Huawei Cloud's data centers use a multi-layered approach:

High-availability power architecture: This redundant system integrates dual mains power, diesel generators, uninterruptible power supplies (UPS), and intelligent controls. When grid power failed, UPS systems bridged the gap until generators activated, ensuring uninterrupted long-term power

Scenario-based control logic verification: Preemptive simulations of power failure scenarios validated medium- and low-voltage control logic during infrastructure acceptance phases, enabling swift troubleshooting during real-world emergencies.

Proactive UPS performance monitoring: Real-time tracking of key UPS parameters and health status via intelligent infrastructure algorithms allows for timely component replacement.

Consistent system preparedness: Regular generator tests validate long-term power supply readiness, while emergency drills for diverse outage scenarios ensure teams and systems remain ready.



Fuel transportation

02 Challenge 2: Maintaining Customer Confidence Amid Uncertainty

As panic spread, Huawei Cloud's Site Reliability Engineering (SRE) team delivered end-to-end assurance through rapid response and unparalleled visibility:

Rapid war room activation: Huawei Cloud's resource monitoring platform triggered alerts the instant the outage struck. A global team of over 300 experts mobilized within a single minute to coordinate recovery efforts.

Full-link observability: Equipment room management, resource management, and tenant-level platforms provided complete visibility into infrastructure health, allocation and usage of resources, and user-facing exceptions.



Onsite inspection

End-to-end service inspection: Meticulous monitoring of critical metrics ensured consistent service performance.

24/7 key service assurance: Round-the-clock staffing ensured immediate resolution of emerging issues.

03 Challenge 3: Sustaining Support Through to Recovery

Even as Chile's grid began restoring power, Huawei Cloud maintained vigilance to manage post-outage risks.

Continuous monitoring and alerting: Anticipating performance strains from sudden IoT traffic spikes post-recovery, preconfigured emergency plans for services like OBS ensured metric stability.

Seamless service switchover: After power restoration, Huawei Cloud extended assurance for six additional hours, monitoring cloud platforms, WANs, data center networks, and security devices to guarantee traffic stability and zero alarms.

27 hours of nonstop effort during the Chilean blackout was worth it — zero service interruptions, a 1-minute war room response, and 24/7 readiness — to Huawei Cloud's customers.

Deterministic Operations:

Transforming High Availability into Always-On Reliability

This event underscored the power of Huawei Cloud's extensive SRE expertise in transforming the "uncertainty" of digitalization into "deterministic" outcomes. This capability is powered by a 1+N deterministic operations system platform, a framework designed to make risks avoidable, controllable, and manageable.

1 represents the management system encompassing organizations, processes, and tools. Organizational transformation involves realigning human resources, to optimize efficiency, reduce costs, and enhance sustainable competitiveness. Process optimization streamlines workflows across the entire product lifecycle—from request acceptance and change management to availability assurance—ensuring seamless collaboration between technical and business teams. O&M tools



1+N deterministic operations framework

are accelerators for reliability, security, and operational efficiency.

N represents the six proactive capabilities—high availability, continuous delivery, O&M trustworthiness, risk governance, resource governance, and security compliance—to address lifecycle challenges from design to deployment and runtime. Specific capabilities help enterprises resolve specific O&M problems.

These capabilities are distilled into three customer-centric solutions: Operation Enabling Service (OES), Infrastructure Management Service (IMS), and Application Management Service (AMS). OES enables rapid fault recovery, full-link observability, and chaos engineering. IMS provides fault recovery capabilities

for 99.999% availability. AMS provides enhanced infrastructure as code support and one-stop O&M hosting.

Illuminating the Future of Digital Resilience

Chile's blackout served as a testament to the cloud industry's ability to thrive under extreme conditions. For Huawei Cloud, commitment to uninterrupted service remains paramount. While power outages cannot be prevented, the digital world can be illuminated, even in the darkest of times. By balancing quality, cost, and efficiency, Huawei Cloud continues to empower industries with deterministic, future-ready resilience.

- Reposted from Huawei Cloud Official Account





Huawei Cloud Credence Club: Key Events

In 2024, the Huawei Cloud Credence Club hosted a series of salons and activities, uniting industry leaders to delve into the innovative applications of AI technologies. Drawing on Huawei Cloud's deterministic operations experience, participants engaged in deep discussions and shared practical insights. The club aims to leverage deterministic operations systems and solutions to help enterprises drive business innovation on the cloud, accelerate O&M transformation, and reshape their cloud management models, enabling them to advance steadily, swiftly, and sustainably in the digital intelligent world.

Shanghai

Deterministic Operations
Forum at HUAWEI
CONNECT 2024

Shanghai

European Session by the
Cloud Native Elite Club
(CNEC)

September 20

September 21

December 13

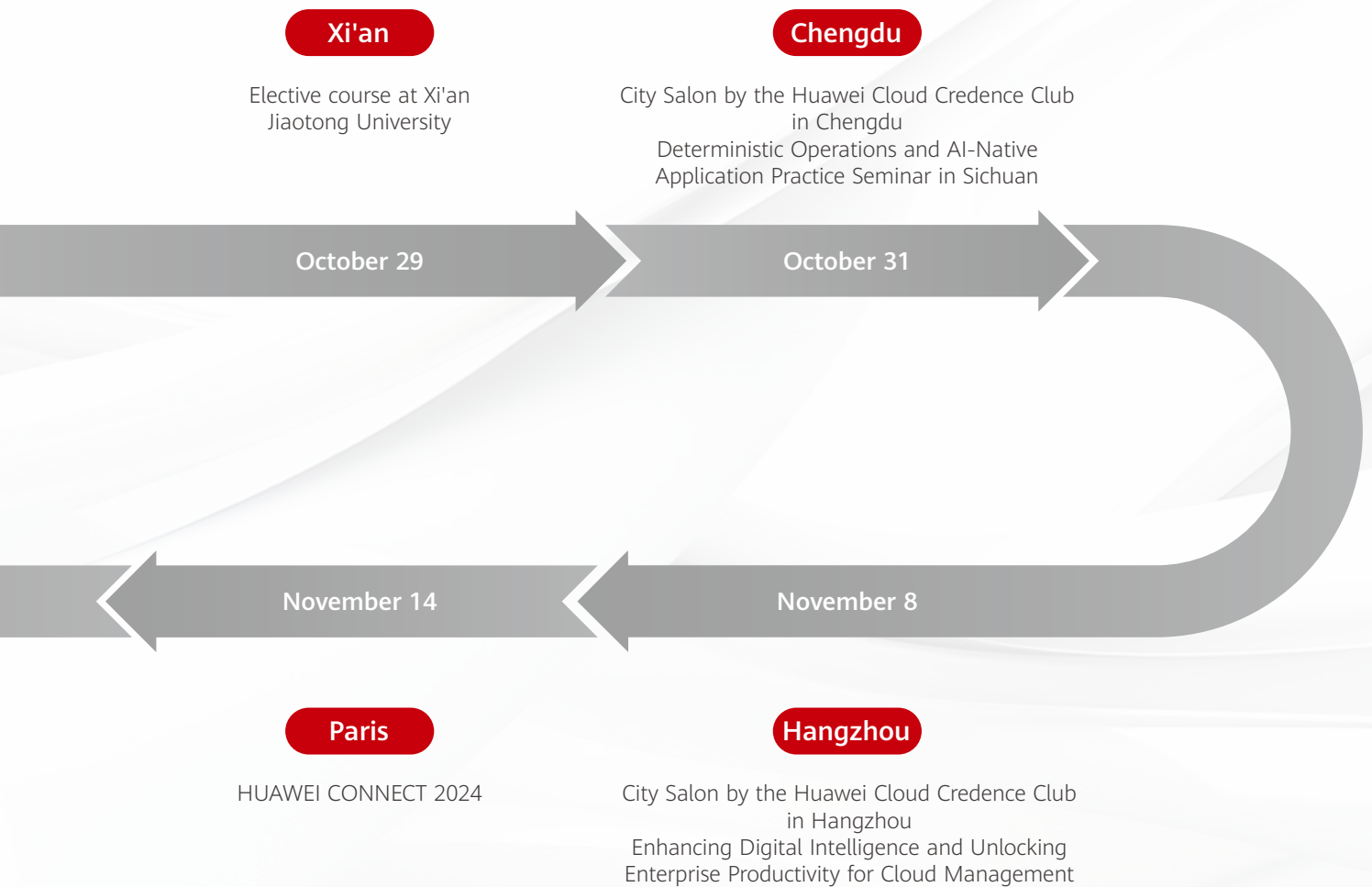
November 26

Wuhan

City Salon by the Huawei Cloud
Credence Club in Wuhan
High-End Operations
Governance Seminar

São Paulo

Huawei Cloud Summit 2024
in Brazil



“

Huawei Cloud consistently leverages its capabilities to help enterprises build robust operations systems, enhancing their end-to-end competencies from strategy formulation to the final implementation. Security and trustworthiness are essential prerequisites for these systems, while stability and reliability provide a solid foundation. Resource efficiency is a critical focus, and service agility lies at the core. Committed to driving enterprise transformation, Huawei Cloud promotes enterprise shifts in mindset, organizational structure, culture, and operations models, fostering their deeper engagement in digital initiatives and enabling sustained growth and innovation in the digital age.

”

—“XoE: Revolutionizing Enterprise Operations and Unleashing Robust Productivity on the Cloud” by Alex An, Director of Huawei Cloud SRE Dept, on September 20 in Shanghai



Deterministic Operations Forum at HUAWEI CONNECT 2024

“

The digital era has heightened the need for skilled O&M professionals. The Deterministic Operations course addresses this by imparting essential SRE capabilities, high-availability system design, live network reliability principles, and software engineering skills. This curriculum prepares learners to effectively navigate challenges within intricate operational environments.

”

—Elective course “Deterministic Operations Concepts and System Design” delivered by Li Huan, a digital twin expert from Huawei Cloud, on October 29 in Xi'an



Elective course on deterministic operations at Xi'an Jiaotong University

“

The “Cloud Day Europe” event concluded successfully at Yuyuan Garden, in Shanghai, on September 21, 2024. Focused on optimizing cloud migration, utilization, and management, the event sought to empower enterprises with enhanced cloud-native capabilities. Representatives from Xiaohongshu, Shanghai Seven-Cat Culture Media, and other organizations presented their cloud-native implementations, while Huawei Cloud's technical teams demonstrated practical applications of relevant technologies. Through the free exchange of ideas, attendees gained actionable insights into technological advancements. The open discussion segment fostered active participation, broadening avenues for communication and collaboration. Serving as a robust platform for cloud-native technology exchange, the event significantly contributed to advancing industry progress.

”

—“Cloud Day Europe” on September 21 in Shanghai



European Session by the CNEC

“

Deterministic operations relies on robust fault management, change management, drills, and disaster recovery capabilities, complemented by effective fault prevention and rapid service recovery. To ensure service stability, enhance system resilience, and strengthen architectural foundations, DevSecOps models can be employed alongside continuous improvements in R&D efficiency, security, and quality. Furthermore, Huawei Cloud collaborates with industry partners to advance the use of deterministic operations tools and refine related standards, such as jointly publishing white papers. These efforts aim to integrate AIGC into operational scenarios, helping enterprises elevate O&M quality to new levels.

”

—Shared by Lin Huading, Director of Huawei Cloud SRE Enabling Center, on October 31 in Chengdu



Deterministic Operations and AI-Native Application Practice Seminar in Sichuan



Drawing from its operational expertise, Huawei Cloud has developed the AppStage platform to enable deterministic operations. The platform prioritizes swift fault detection and recovery, enabling enterprises to maintain stable and efficient IT environments. Through continuous technological innovation and process refinement, Huawei Cloud integrates cutting-edge technologies and tools to enhance the precision and broaden the coverage of fault monitoring and alarm reporting. These advancements foster deterministic recovery capabilities, elevate fault management intelligence, and ensure system reliability and user experience.



—**"Deterministic Operations: Enabling Enterprise Digital Transformation and Ensuring Application Stability in the Cloud-Native Era"** by Chen Xijin, an SRE expert from Huawei Cloud Computing Global Ecosystem Dept, on November 8 in Zhejiang



City Salon by the Huawei Cloud Credence Club in Zhejiang



Huawei Cloud adheres to the principle of "In Europe, For Europe", driving innovation through cutting-edge technologies, fostering trust via localized services, and enriching the regional ecosystem with global experience. It supports European enterprises in expanding their international presence during digital transformation. Through demonstrated capabilities in deterministic operations and ongoing advancements in O&M technologies, Huawei Cloud accelerates cloud-based organization transformation and promotes global growth for European enterprises.



—**"Addressing IT Uncertainties in the Cloud Era with Deterministic Operations"** by Huawei Cloud's chief engineer for observability, on November 14 in Paris



HUAWEI CONNECT 2024



Themed "Empowering Brazil's Intelligence Growth with Advanced Cloud Services", the summit featured thought leaders from Huawei Cloud and key customers discussing technological innovations and industry best practices to drive digital transformation across Brazilian sectors. At the Deterministic Operations exhibit, Huawei Cloud experts showcased their proven cloud utilization and management experience, presenting professional service offerings such as deterministic operations consulting, planning, design, and support plans. Leveraging this expertise, Huawei Cloud enables customers to achieve robust global business expansion.



—**"Deterministic Operations System and Solutions"** by Xu Dianjun and He Yuan, Huawei Cloud experts, on November 26 in São Paulo



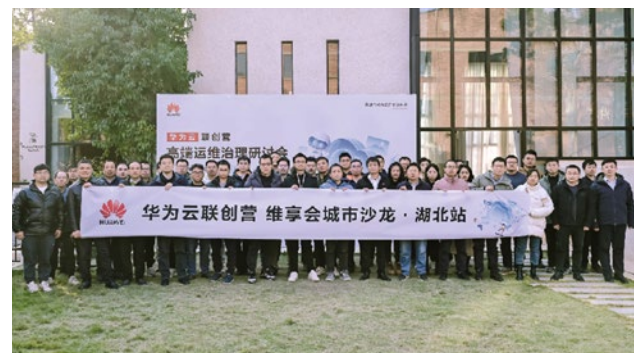
Huawei Cloud Summit 2024 in Brazil



As enterprises advance in their digital and intelligent transformations, IT operations and maintenance will become a critical source of productivity. Given the increasing complexity of business systems and the surge in data traffic, fault management has emerged as the primary challenge in O&M. Enterprises must establish mechanisms for rapid and accurate fault detection and develop efficient, agile capabilities for business recovery to ensure smooth operational continuity.

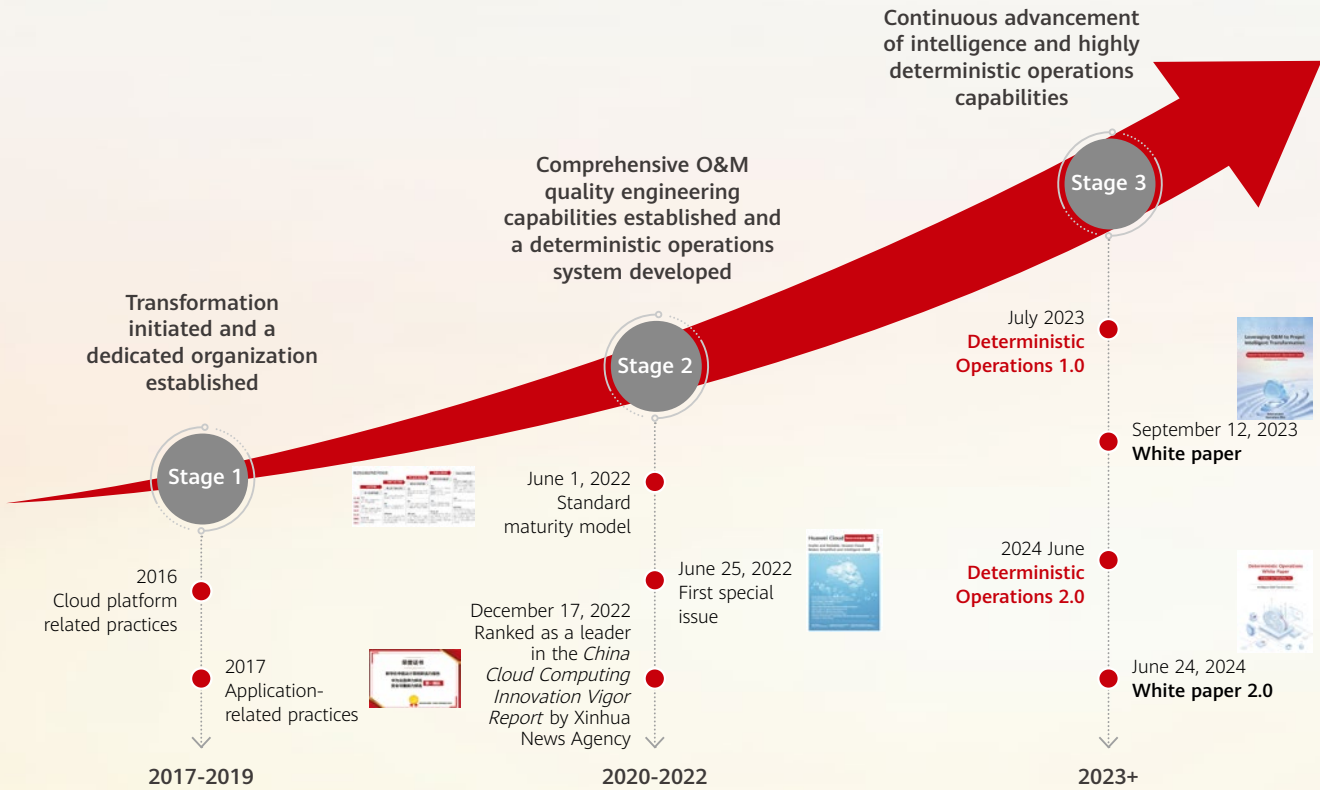


—**"Fault Management in Deterministic Operations: Safeguarding the Business Lifeline"** by Xu Dianjun, Huawei Cloud SRE Senior Architect, on December 13 in Wuhan



City Salon by Huawei Cloud Credence Club in Hubei

Evolution of Deterministic Operations



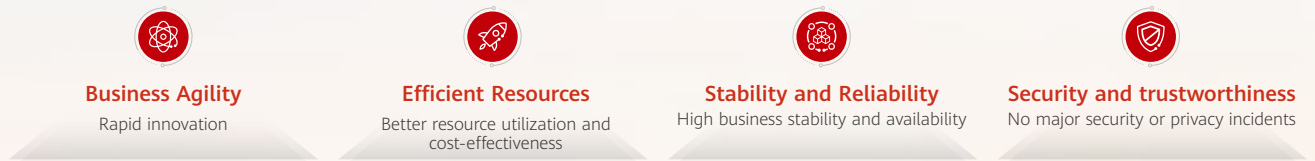
Deterministic Operations Capability System

Quality Culture as the Basis	High-Availability Architecture as the Prerequisite	Dynamic Risk Governance as the Safeguard	Intelligent Operations as the Vision
A culture of deterministic quality	Deterministic failure rate	Trustworthy tasks	Automatic recovery
Shared SLOs for development and SRE teams	Deterministic recovery time	Deterministic recovery	Intelligent alarm reporting
SRE organizational transformation	Deterministic blast radius	Data-based intelligent operations	Smart fault locating
		Proactive operations throughout the entire lifecycle	Data & algorithms

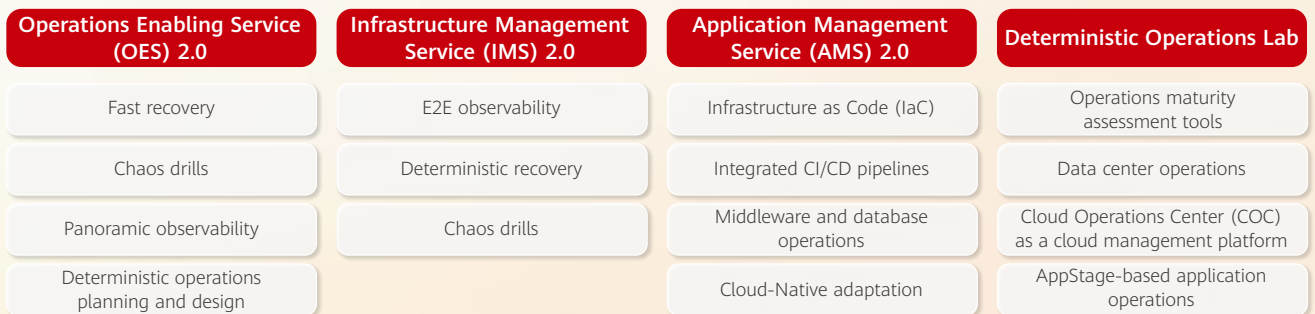
Deterministic Operations 2.0

Enhancing O&M Across Full-Stack Scenarios from Infrastructure to Applications,
Driving Rapid, Sustainable Business Growth for Customers

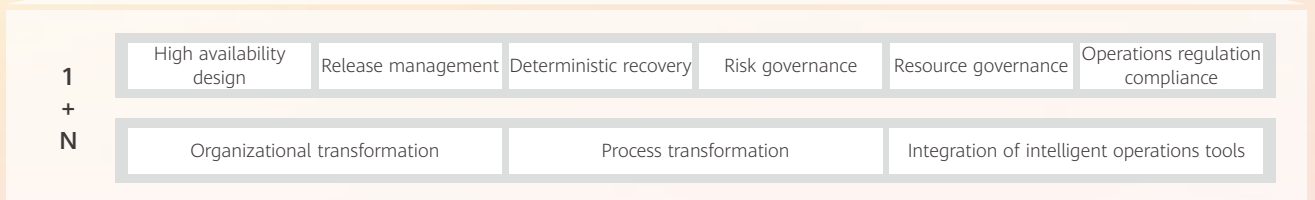
Customer Benefits



Services for Four Key Scenarios



Fundamental Capabilities for Digital Operations Transformation



Deterministic Operations Lab: a Collaborative Innovation Platform for Huawei and Customers

Objective: Equip each lab member with a dedicated cloud host, a suite of development tools, and allocated cloud storage. Centralize the development resources for key core technologies such as those related to Ascend, HarmonyOS, and Kunpeng. Supplement this with practical case studies that streamline the process from coding to application debugging. Empower developers to utilize their personal cloud hosts for seamless access to Huawei's tools and resources.

Deterministic Operations Lab

Intuitive lab
Uis



Maturity diagnosis



Data center operations



Free cloud hosts



Sandbox environments

Deterministic Operations 2.0

Scenario-
oriented
tools and
platforms

Maturity Assessment Tools	Data Center Operations	COC for Cloud Management	AppStage-based Application Operations
Maturity assessment	E2E observability: Server fault monitoring	Fast recovery: CPU overload simulation	Full-stack observability: IaaS observability
Maturity diagnosis reports	E2E observability - Customer asset visibility	Fast recovery: Traffic overload	Full-stack observability: Cloud phones or Kunpeng sandboxes
Optimization suggestions	E2E observability - Network fault monitoring	Fault drills: Primary/standby database switchovers	
	E2E observability - Power/Cooling fault monitoring		





**Secure | Reliable
| Intelligent |
Efficient | Agile**

Huawei Cloud



Deterministic
Operations
Website