Product White Paper for KSManage

Document Version: V3.0

Release Date: January 30, 2026

Environmental Protection

Please dispose of product packaging by recycling at a local recycling center for a greener planet.

Trademarks Statement

All trademarks or registered trademarks mentioned herein may be the property of their respective holders.

Security Statement

The Company's products do not proactively collect or utilise users' personal data. Certain personal data (such as email addresses or IP addresses) may only be collected or utilised during business operations or fault diagnosis when you consent to specific features or services. Throughout the entire lifecycle of personal data processing

activities—including collection, storage, use, transmission, and deletion—necessary

security safeguards have been implemented within the product's functionalities. Concurrently, you are obligated to establish appropriate user privacy policies in accordance with applicable national or regional laws and regulations, and to implement sufficient measures to ensure adequate protection of users' personal data.

Inspur Information places high importance on product data security. Throughout the entire lifecycle of system operation and security data processing activities, the company's products have implemented necessary security safeguards in product functionality in strict compliance with relevant laws, regulations, and regulatory requirements. As the processor of system operation and security data, you are obligated to formulate necessary data security policies in accordance with the laws and regulations of the applicable country or region and to take sufficient measures to ensure that system operation and security data are adequately protected.

Inspur Information will continue to rigorously monitor the security of its products and solutions, striving to provide customers with increasingly satisfactory services. The company has comprehensively established emergency response and handling mechanisms for product security vulnerabilities, ensuring prompt resolution of any security issues. Should you encounter any security concerns during the use of this product, or require necessary support regarding product security vulnerabilities, please contact Inspur Information customer service personnel directly.

Protocol Use Statement

- The product supports LDAP authentication. LDAP over SSL (LDAPS) can be used to implement encrypted data transmission. We recommend using port 636 for LDAPS security authentication.
- The product supports log dumping through the Syslog protocol. Syslog over SSL can be used to implement encrypted data transmission. To ensure the security of log data transmission, we recommend dumping logs by using Syslog over SSL.
- The product supports device detection through SNMP. Three versions of SNMP are available: SNMPv1, SNMPv2c, and SNMPv3. There are potential security risks when SNMPv1 and SNMPv2c are used to detect devices; therefore, SNMPv3 is recommended.

Updating and Patching Statement

Before the product version update or patch installation, it is recommended that you check the product hash value or digital signature and verify the legitimacy of the upgraded software, thus avoiding unauthorized tampering or replacement of the software that may bring security risks.

Security Response Statement

We have established emergency response procedures and action plans for security vulnerabilities so that product safety issues can be dealt with in a timely manner. Please contact us if you find any security issues or need support on security vulnerabilities when using our products.

We will remain committed to the safety of our products and solutions to achieve better customer satisfaction.

Disclaimer

The purchased products, services, and features shall be bound by the contract made between us and the customer. All or part of the products, services, and features described herein may not be within your purchase or usage scope. Unless otherwise agreed in the contract, we make no express or implied statement or warranty on the contents herein. Images provided herein are for reference only and may contain information or features that do not apply to your purchased model. This document is only used as a guide. We shall not be liable for any damage, including but not limited to loss of profits, loss of information, interruption of business, personal injury, or any consequential damage incurred before, during, or after the use of our products. We assume you have sufficient knowledge of servers and are well-trained in protecting yourself from personal injury or preventing product damage during operation and maintenance. The information in this document is subject to change without notice. We shall not be liable for technical or editorial errors or omissions contained in this document.

Technical Support

Global Service Hotline: +1 800 611 8899 / +65 6611 8899

Address: 150 Beach Road, #14-05/08, Gateway West, Singapore 189720

KAYTUS PTE. LTD.

Preface

Abstract

This document describes the main functions, basic operations, and frequently asked questions of KSManage.

Intended Audience

This document is intended for:

- Technical support engineers
- Product maintenance engineers

It is recommended that server installation, configuration, or maintenance be performed only by experienced technicians who know our servers inside and out.

Notices

- If your purchases do not include our on-site installation service, make sure that you inspect the shipping cartons before unpacking the device. If a shipping carton appears severely damaged, waterlogged, or the seal or pressure-sensitive adhesive tape (PSA) is broken, report this based on your purchase channel. If you purchased from a third-party supplier, contact your supplier directly; if you purchased through our direct sales stores, call our service hotline +1 800 611 8899 / +65 6611 8899 for technical support.
- For your safety, please do not arbitrarily install or remove the server's components, extend the configuration, or connect other peripherals. Contact us for our authorization and support.
- Before installing or removing the server's components, please be sure to disconnect all the cables connected to the server.
- Please install a product-compatible operating system and use the driver coming with the server or provided by us. You can go to our official site and find the correct driver of your product based on the prompt. An incompatible operating system or a driver which has not been validated by us may cause compatibility issues and affect the normal use of the product. We will not assume any responsibility or liability for this.
- BIOS and BMC setup is critical in configuring your server properly. Do not alter the default settings unless you are familiar with the options and aware of the effect

your changes will have on performance. The first time you log in to the BMC, please change the user password.

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
 DANGER	A potential for serious injury or even death if not properly handled
 WARNING	A potential for minor or moderate injury if not properly handled
 CAUTION	A potential loss of data or damage to device if not properly handled
 IMPORTANT	Operations or information that requires special attention to ensure successful installation or configuration
 NOTE	Supplementary description of document information

Revision History

Version	Date	Description of Changes
V3.0	2026/1/30	The third official release
V2.0	2025/06/30	The second official release
V1.0	2024/12/30	The first official release

Table of Contents

Introduction

Definition

Data centers, as key infrastructure, play an important role during the critical period of digital transformation. A data center is a facility for centralized storage, management, and processing of mass data, which provides high-performance computing and storage resources to meet the huge data processing demands required for digital transformation.

With the continuous growth of data volume, the scale of data centers is expanding, making infrastructure management increasingly challenging. Data center infrastructure manager refers to the comprehensive management of data center infrastructure such as computing devices, network devices, storage devices, micro modules, and security facilities. It includes functions such as resourcs management, monitor management, alert management, configuration management, knowledge base management, power efficiency management, statistical analysis, remote management, agent and system management.

Application Scenarios

KSManage can be applied to the infrastructure management of data centers of different scales.

Enterprise data center: For large enterprises, the key to an infrastructure management platform lies in monitor and managing a vast number of servers and network devices to ensure the efficiency and security of data processing.

Cloud service provider: Cloud service providers optimize resource allocation with infrastructure management platform to improve service reliability and resource utilization.

Small and medium-sized enterprises: Small and medium-sized enterprises manage IT resources in a cost-effective way through the platform, ensuring the stable operation and effective monitor of device, and improving operation and maintenance efficiency and response speed.

Current Industry Status

With the rapid growth of data volume and data center scale, the market for infrastructure management software is expanding rapidly. The manufacturers have launched a variety of data center infrastructure management software. However, due to the diversity and complexity of data center infrastructure, the existing management software faces a series of challenges, such as the lack of standardization, integration, and

intelligence, which leads to the difficulty of collaboration between software and affects management efficiency. In addition, existing technologies have limitations in meeting the specific needs of other industries, restricting the application scope.

Management Value

The presence of numerous infrastructure management software in the market results in a lack of standardization, integration, and intelligence in existing management tools. This poses challenges to operating costs, stability, maintainability, and scalability of data centers. The infrastructure management system is designed to meet the digitalization demands of data center infrastructure. Through efficient data collection and storage, it integrates functions such as resource management, monitor management, alert management, configuration management, knowledge base management, energy efficiency management, statistical analysis management,, and system management, achieving comprehensive and unified management of data center infrastructure.

1 Overview

1.1 Product Positioning

KSManage is an integrated infrastructure management platform designed for data centers in finance, telecommunications, cloud services, and other industries. It helps user achieve the unified intelligent management of basic devices, including servers, storage devices, network devices in data centers.

This platform is applicable to multi-brand IT devices on the market and provides functions such as unified online and offline asset management, large-scale real-time alert monitor, AI-based fault prediction of drives and memory, performance prediction, energy consumption management, report statistics, topology display, and dual-channel automated site deployment delivery. KSManage enables the unified management of heterogeneous devices such as servers, storage devices, network devices, and security devices. It truly promotes the intelligent management of data centers, helping customers build unattended data centers, improve O&M efficiency, reduce O&M costs, and ensure the safe, reliable, and stable operation of data centers.

KSManage can be widely used in public clouds, private clouds, and data centers of telecom and enterprise customers. It can be deployed in a variety of scenarios, such as AI, HPC, cloud service, and smart city. Moreover, it provides interfaces such as RESTful, SNMP, and Prometheus for easy integration and interfacing.

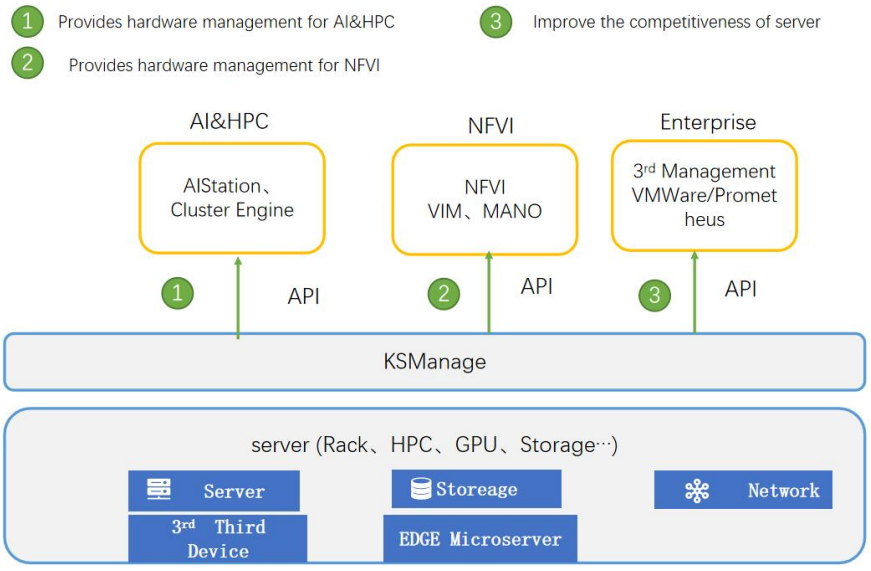

Figure 1- 1 Positioning of KSManage

1.2 Product Features

AI intelligent Q&A and root cause analysis of faults

By analyzing user questions, integrating the knowledge base and context correlation, the Operation Assistant can provide more accurate answers to users. Through the 24-hour self-service question answering service, reliance on human customer service is reduced. When an alert occurs, the Operation Assistant automatically associates the alert information, quickly locates the root cause, conducts in-depth analysis and reveals the fundamental cause of the alert, and further infers and provides repair suggestions.

Lightweight deployment in multiple scenarios and full lifecycle management

KSManage can be deployed on multiple devices and adapts to virtual machines (KVM/VMware) and bare-metal deployment scenarios. It can provide full lifecycle management of network-wide device such as servers for small-, medium-, and large-sized enterprises.

High reliability and on-demand 1-N data collection and analysis node expansion

KSManage can meet the requirements of various business scenarios, providing high reliability capabilities and the ability to smoothly expand the number of collection and analysis nodes from 1 to N, to cope with scenarios of capacity expansion and multiple data centers without affecting the original monitor business.

Unified view and scheduling of computing resources, performance monitor

KSManage provides a visual view of the global computing resource capacity, utilization rate, and job count. It uniformly manages and schedules various computing resources of the cluster, such as CPU and memory, to enhance the utilization rate of resources. At the same time, it monitors the performance metrics of computing resources in real time, such as utilization rate and load.

Comprehensive monitor and fault early warning for the overall control of services

KSManage provides comprehensive alert monitor and fault warning services. These services, combined with advanced AI technology, realize fault prediction of memory and hard disk server and AI-based maintenance, ensuring the efficient and stable operation of enterprise infrastructure.

Second-level performance monitor and intelligent prediction for health status control of devices

The seamless interconnection between the KSManage and KSManage Driver systems enables second-level real-time collection of performance to ensure real-time receipt of device performance metrics. With self-developed core components for performance analysis, KSManage supports second-level performance data monitor and alerts for large-scale servers. Meanwhile, KSManage combines performance analysis of data center devices and comprehensive monitor and analysis of multiple key metrics, providing effective O&M decision-making support for administrators to achieve efficient management of data centers.

Unified view and scheduling of computing power resources, performance monitor

KSManage provides a visual view of global computing resource capacity, utilization, and job count, and uniformly manages and schedules various cluster computing resources, such as CPU and memory, to improve resource utilization. At the same time, the performance metrics of computing power resources, such as utilization and load, are monitored in real time.

Precise energy consumption monitor, energy efficiency analysis and optimization

KSManage can finely monitor the energy consumption of various parts of the infrastructure, such as data center, computer room, cabinet, server, etc., collect energy consumption data, and provide accurate basic information for energy consumption analysis. Combined with the energy consumption model, the energy consumption of computer room and device are analyzed, and comprehensive energy-saving optimization suggestions and strategies are provided for user to help enterprises reduce energy consumption costs and improve energy efficiency. It supports visual display function, and displays energy consumption through intuitive charts and dashboards, which helps managers intuitively understand the overall energy usage status and optimization direction.

Dual-channel batch configuration to shorten the launch cycle

KSManage provides functions such as batch firmware update, hardware configuration, system deployment, and software deployment, significantly increasing server launching and O&M efficiency. At the same time, KSManage integrates KSManage Boot to realize BMC+PXE dual-channel automated site deployment delivery of servers.

Improved version management efficiency

KSManage provides two methods to manage firmware and OS images: local management and automatic synchronization with the remote official website. This streamlines software and hardware version management of data center devices.

Standardized northbound interfaces for easy integration and interfacing

KSManage provides standard Redfish and SNMP interfaces, based on which other functions can be extended for easy integration and interfacing.

2 System Architecture

2.1 Software Architecture

2.1.1 System Functional Architecture

The overall functional architecture of KSManage covers multiple modules such as resourcs management, monitor management, alert management, configuration management, knowledge base management, power efficiency management, statistical analysis remote management, and system management.

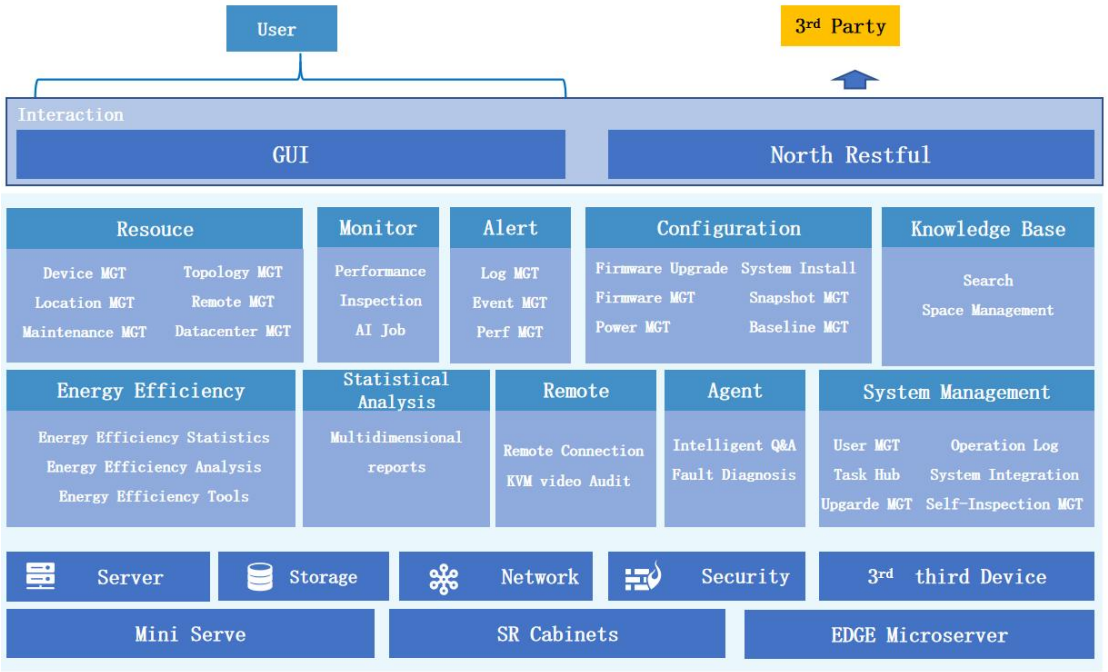

Figure 2- 1 Functional Architecture

Centralized management and scheduling center

- **Basic features:** Monitor, alert, upgrade, security, DFX, etc.
- **Functions:** resources management, monitor management, alert management, configuration management, energy efficiency management, knowledge base management, remote management, report management, agent, and system management.

Management of network-wide devices

- Supports a full range of products, including general-purpose rack servers, AI servers, blade servers, IT application and innovation industry servers, storage devices, network devices, edge servers, all-in-one servers, and other high-end server products. For details on models, see the KSManage Specification List.
- Products from different manufacturers, including servers, storage devices, network devices, and other devices, can be customized.

Highly available, highly scalable and flexible architecture

- Supports monolithic and distributed architectures. Distributed deployment supports horizontal scaling.
- Supports probe-like data collection and unified management of multiple data centers.
- Supports multiple deployment modes, flexibly adapting to management scales ranging from 100 to 10,000 servers.

Flexible Architecture

KSManage caters to a variety of customer groups, such as the CSP, finance, and telecommunications industries, with resources ranging from a few servers to tens of thousands of servers. To provide consistent services for different customer groups in different scenarios, KSManage utilizes a modular-oriented architecture (MOA). This architecture enables it to flexibly adapt to customers of different sizes and needs, and meet a variety of specific needs with modular components.

MOA provides high flexibility and scalability by distinguishing between technical modules (responsible for system functions such as calling management, caching, and queue processing) and service module. It allows the free assembly and combined deployment of modules to suit different service scenarios, and supports horizontal scaling of specific modules during high workloads, making it particularly suitable for complex, dynamic, and large enterprise application environments.

2.1.2 System Technical Architecture

The technical architecture is divided into four layers from top to bottom: the facade layer, the service layer, the collection layer, and the registration center. During runtime, it mainly consists of two services: Node and Mono. Node is responsible for the front-end business and rendering; Mono is responsible for all the back-end business and adopts a merged deployment method, integrating all back-end businesses into the same process. The overall architecture is as follows:

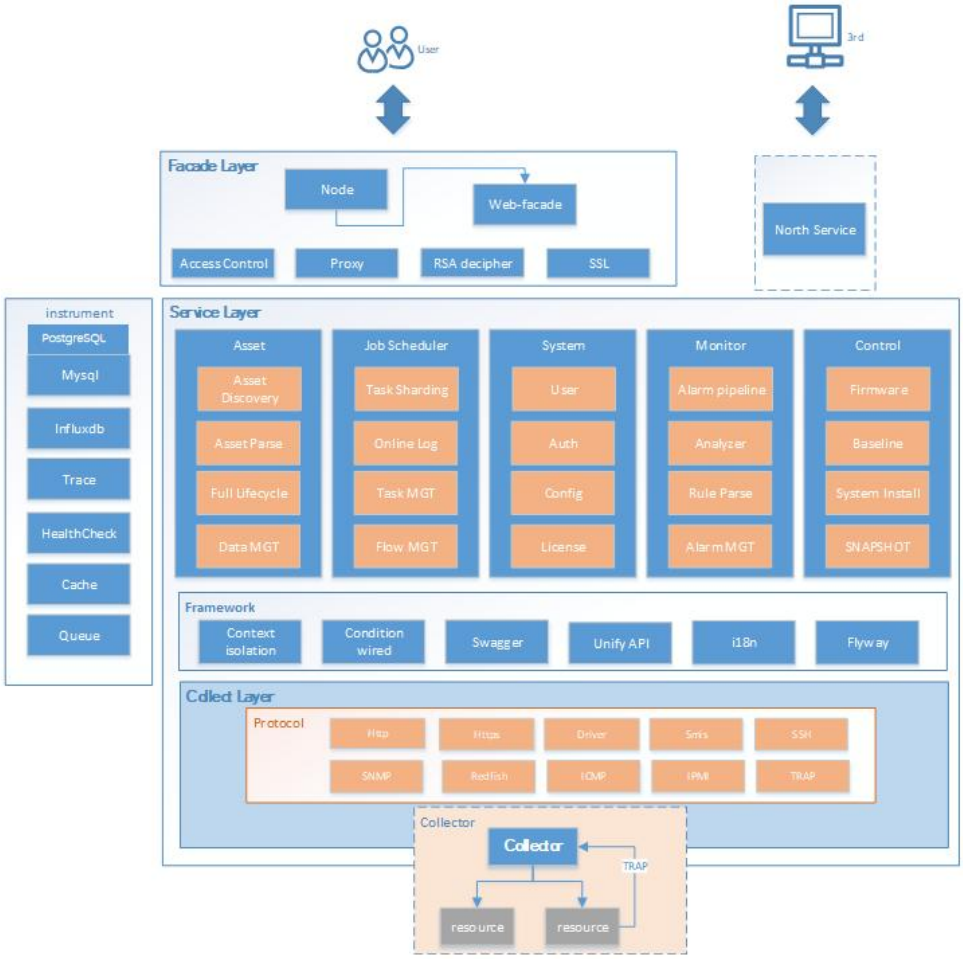

Figure 2- 2 KSManage monolithic technical architecture

Node: It is responsible for front-end services and page rendering, providing user with Web GUI. It deals with SSL uninstallation and certain security issues and is mainly responsible for interactions with user.

Mono: It includes all the functions of the facade layer, service layer, and collection layer, and combines the logic of the three layers into one process. It deals with all the back-end service logic. Its internal modules have the same functions as those in the distributed architecture.

Collection layer: The collection layer is responsible for communication with devices, triggers the collection task through the self-developed task scheduler, obtains the binding relationship from the collector gateway, and delivers it to the collector for information collection and reporting. Callbacks from devices are also handled by collectors, such as SNMP TRAP and REDFISH events.

Registry: service registration, discovery, health check, cluster master selection, responsible for service governance and health detection.

Northbound service: It provides services for northbound integration and is independently deployed to isolate the primary service resources and runtime to ensure interface security.

Relational database: mainly stores assets, monitor, system, control and other configuration and business data.

Time series database: stores time series data, mainly stores performance data and power consumption data.

2.2 KSManage Context Connection Modes

As an infrastructure manager, KSManage supports northbound and southbound connections. The southbound connection is mainly for model compatibility scenarios and system access, and the northbound connection is mainly for integrating KSManage with third-party systems.

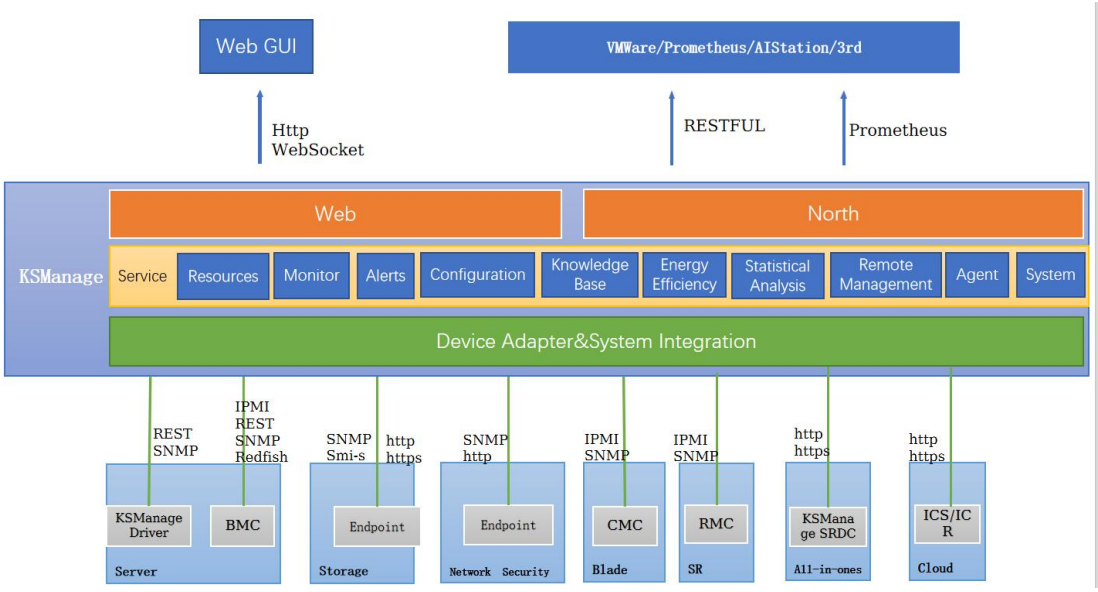

Figure 2- 3 KSManage Context Connection Modes

2.2.1 Southbound Connection

- Supports server management and connects to BMC and KSManage Driver through protocols such as IPMI, SNMP, Redfish, HTTP, and HTTPS.
- Supports storage server management and connects to and manages the controller through SNMP and SMI-S.
- Supports the management of network and security devices and connects to remote management ports through HTTP and SNMP.

- Supports chasis management and connects to CMC through IPMI and SNMP.
- Supports SR cabinet support, connects to RMC, and supports the protocols of IPMI and SNMP.
- Supports all-in-one machine devices, integrates with KSManage SRDC, and supports protocols of http and https.
- Supports management of cloud resources, integrates with ICS/ICR, and supports protocols of http and https.
- Support the integration of GB200 devices, and achieve automated operation and maintenance as well as full lifecycle management through the management platform.
- Introduce the AI training platform (AlStation) and the network platform (ICE) and provide a comprehensive description.

2.2.2 Northbound Connection

- Northbound connection supports 2 connection scenarios: Web GUI and northbound interface.
- Connection through Web GUI, which is designed for the KSManage page, mainly serves the O&M administrator. Through this connection, the administrator can effectively manage and monitor the system.
- Connection through northbound interfaces is an integrated solution for the upper-layer management software and third-party management systems. It is especially suitable for management system scenarios that require integration with VMWare, Prometheus, AIStation, and other systems. And it enhances the interoperability between systems, making management more efficient.

3 Platform Features

3.1 Home

3.1.1 Definition

The home page is a comprehensive information center, focusing on the number of data center, resource statistic and alert overview. It's not just a window for an overview of the data, but also allows you to drill down into details like alert, power statistic, and hot device. User can customize the home page layout, add new server, set alert rules, realize personalized management and efficient monitor, and ensure the smooth operation of the data center.

3.1.2 Value

As the core entry of KSManage, the home page is focused and intuitive to show the key information of the data center, including the number of data center, resource and alert statistic, so that user can quickly grasp the overall situation. At the same time, key data such as alert detail, power consumption statistic and high temperature device can be viewed to help user respond in time. In addition, the functions of customizing the home page, adding servers and alert rule setting further improve the flexibility and practicability of the platform.

3.1.3 Function

View the relevant information of the default home page of the system, and perform actions such as customizing the home page, adding servers, and adding alert rules.

On the home page, user can view global overview information such as data center, resource statistic and alert statistic in KSManage. Click on the statistics for the data center, resource statistic, or alert statistic to navigate to the corresponding administration page to view the relevant details.

KSManage home page supports to view alert information, power consumption statistic, high temperature device, device health status, overwarranty statistic, device statistic, alert type information. Hover the mouse over each trend chart, you can also view the details of the computer room power consumption, temperature, alert and other metrics at a specified time.

3.1.4 Principle

KSManage home page lies in the integrated information display and operation management. It aggregates and displays key data such as the number of data center, resource, and alert statistic, while providing alert information, power consumption statistic, and a detailed view of high temperature device. User can customize the home page layout, add server and alert rule to achieve personalized management and efficient monitor, and ensure the stable operation of the data center.

3.2 Resource Management

3.2.1 Definition

Resource management is the core module of efficient operation and maintenance of IT infrastructure, which aims to realize the whole life cycle control of global resources through a systematic digital platform. The system integrates physical device, virtual resources containers and business units from a global perspective, covers servers ,superpods,network devices storage clusters power supply facilities, thermal units and container platforms management of infrastructure such as servers, network devices, storages, clusters, and provides a complete closed-loop management from resource input, state monitor to decommissioning and recycling.

3.2.2 Value

KSManage realizes the whole life cycle management of resources, tracking, management and control from each link of procurement, shelf, maintenance, removal and scrap, optimizes the utilization of resources and cost control, and ensures the standard, efficient and compliance of resource management.

Through resource visualization, resources are presented in intuitive forms such as graphics and charts, enabling managers to quickly and accurately grasp the overall picture and status of resources. It also clearly displays the physical connection between resources and network topology. Visualized data center rack diagrams make it convenient to plan physical space and manage device racking and de-racking, preventing capacity overflow.

KSManage transforms all information about resources and their status, components, performance, firmware, etc. into structured digital data comprehensively, accurately and in real time, and uses this data for automation, analysis, insight and decision support. It can collect resource data from a variety of different sources, which can include hardware information of the device (such as the vendor, model, serial number, etc.), running state data (such as temperature, power consumption, etc.), performance index

data (such as disk read and write rate, etc.), and integrate these data from different systems and store them in the database.

3.2.3 Function

As the platform's core control module, the resource management module provides unified management, visual monitor, and refined operation capabilities for heterogeneous resources across the data center. Its functions cover diverse resource types, from traditional physical facilities to modern cloud-based and intelligent solutions, while empowering daily operations with efficient workflow tools.

Resource Panoramic View and Visual Management

This section provides a multi-dimensional resource view from global to local, helping managers gain a comprehensive overview of assets.

- Overview: A centralized dashboard for platform resources, displaying key metrics including total resources, the top 10 most common device models, and the top 10 most common vendors. It serves as the primary portal for operations management.
- Data Center: Visualizes the physical layout and spatial capacity of data centers, server rooms, floors, and micro-modules in 2D/3D formats, enabling centralized management through a unified 'one-map' system.
- Cabinet: Detailed display of cabinet U-space, power supply, load-bearing capacity, and device placement. Supports drag-and-drop simulation for device placement to optimize space utilization.
- Micro module: Centralized management of integrated facilities including containerized data centers and intelligent micro modules, with real-time monitor of their internal power, environmental, and IT device interconnection status.

The physical infrastructure management

Realizes the whole life cycle management of all kinds of entity hardware device.

- Server: Managed infrastructure encompasses all computing hardware including racks, blades, mainframes and midranges, with real-time monitor of asset data, hardware health status, performance metrics, and firmware versions.
- Storage: Centralized management of storage devices including SAN and NAS, with real-time monitor of storage pool capacity, performance throughput, and disk health status.
- Network devices: Automatically discover and manage network devices such as switches and routers, enabling visual network topology and link state monitor.
- Security devices: Connect to security hardware devices such as firewalls and intrusion detectors, and monitor their operational status and policy logs uniformly.

Other resources: Manage auxiliary infrastructure devices such as printers, projectors, and business systems that are not separately classified.

Virtualization and Cloud Resource Management

Enhance the management capabilities of pooled and abstracted resources, enabling integrated control across cloud, network, and endpoint.

- Cloud Resources: Connect to mainstream public and private cloud platforms, and display usage and cost overviews of cloud resources such as cloud servers, cloud disks, and VPCs in a synchronized and unified manner.
- Container: Deeply integrated with Kubernetes and other container platforms, it provides a comprehensive view, performance monitor, and correlation analysis of multi-layer resources including clusters, nodes (host machines), and Pods.
- AI Cluster: Specifically designed for AI training and inference scenarios, this solution manages GPU server clusters to monitor GPU utilization, memory usage, temperature, and job status.

Critical Infrastructure Management

The environment and power facilities that guarantee the operation of IT device are monitored in detail.

- Power supply: Centralized management of power racks and PDU (Power Distribution Unit), monitor electrical parameters (voltage, current, power, and energy) for input and output, with remote control capability for power circuits.
- Thermal: Monitor cooling device such as the CDU (Cooling Distribution Unit) and precision air conditioning, tracking key parameters including water temperature, flow rate, and supply/return air temperature to ensure optimal refrigeration efficiency.

Resource Operation and Process Support

Provide efficient tools for resource delivery, daily operations, and recycling management.

- IP Pool: A centralized management system for data center IP addresses, enabling automatic allocation, recycling, conflict detection, and utilization rate tracking.
- Discovery: Provides automatic discovery capabilities based on IP network segments, protocols, and other methods, with discovery task and result management functions to achieve automatic resource management.
- Workbench: Provides customized operation views for O&M personnel, including customizable lists of frequently used devices, pending alerts, and quick task shortcuts, to enhance daily O&M efficiency.

- Recycle Bin: Temporarily stores logically deleted or removed devices to prevent data loss from accidental operations. Supports one-click recovery or complete deletion.

Manage Configuration

Provide system-level configuration capability for resource management modules.

- Settings: Includes system configuration features such as resource model customization, monitor policy configuration, automatic grouping rule definition, and permission template management, to meet enterprise-specific management needs.

3.2.4 Principle

Resource management is realized by unified resource model and automatic discovery engine.

- Resource Modeling and Access: The platform predefines data models for various resources such as servers, networks, storage, and container clusters. Devices can be connected to the platform through automatic discovery (based on IP network segments and protocol scanning) or manual addition, with their attributes and status data collected and populated into the corresponding models.
- Data Aggregation and Categorization: Collected data is centrally stored in a resource database. The system automatically groups devices into corresponding management views—such as servers (including hyper nodes), power supplies (including racks and PDU units), containers (including clusters, hosts, and PODs), and thermal (including CDUs)—based on device type, predefined rules, or user-defined grouping criteria (e.g., by service, region, or organization), forming a hierarchical resource tree.
- Status Synchronization and Visualization: The platform continuously updates resource status (performance, alerts, configuration changes) through regular polling or event reception. These real-time data drive front-end interface updates, presenting information in charts, lists, and other formats across resource overview, device list, and specialized management views for user monitor and analysis.

3.2.5 Metrics

Each device added to KSManage takes up one License capacity.

Batch import device supports the import of 10000 devices at a time.

The system supports up to 30 custom attribute fields, which can be applied to servers, storage devices, network devices, security devices and other data center IT device.

3.3 Monitor Management

3.3.1 Definition

Monitor management refers to the real-time or regular monitor, evaluation and management of infrastructure through a series of technical means and management methods to ensure its normal operation, optimize performance, prevent failure, and meet the relevant regulations and standards.

3.3.2 Value

The monitor management module, through the collaborative effect of multiple functions such as health monitor, performance monitor, log monitor, operation monitor, and inspection management, comprehensively ensures the stable operation of the system. Its values include:

- Real-time monitor of the health status of various resource components such as servers, network devices, and storage devices can promptly identify potential fault hazards.
- Continuously tracking the performance metrics of hardware devices, such as CPU utilization rate, memory utilization rate, and Driver status, can quickly identify performance bottlenecks, assist operation and maintenance personnel in optimizing in advance, and ensure system stability.
- By comparing the metrics of the same type of device, the efficiency of resource utilization has been improved. In a multi-device operating environment, the horizontal comparison function accurately identifies devices with abnormal performance, ensuring the balanced and reliable operation of the system.
- The platform can significantly enhance the flexibility and efficiency of operation and maintenance work by collecting and analyzing server logs, and by leveraging a powerful search engine and flexible search methods, greatly improving the efficiency of fault detection.
- Inspection management comprehensively covers infrastructure, promptly identifying potential issues and preventing faults, and supports personalized inspection range Settings. These functions work in concert to ensure the stable, efficient and reliable operation of infrastructure in all aspects.
- Ping tool and network detect enable operation and maintenance personnel to quickly locate network layer faults and shorten the average repair time.

- The diagnostic tool, by integrating SNMP, IPMI, Redfish, HTTP and model diagnosis functions, can significantly improve the efficiency of fault detection, enhance system reliability, optimize operation and maintenance processes, and support the management of hybrid cloud and heterogeneous environments.

3.3.3 Function

Health Monitor

- KSManage's health monitor function provides user with device monitor, superpod monitor, virtualisation monitor, Kubernetes monitor, AI job monitor, power supply monitor , thermal monitor, etc. Server restart record and log download functions, which facilitates user to grasp the operating status of device at any time, including performance, wear and failure warning and other information. This helps user to take maintenance measures in time to prevent device failures and prolong the life of device, while optimizing the efficiency of device use to ensure business continuity and production safety.
- Device monitor:Supports health monitor of servers, storage, and network devices, allowing you to view the status and details of each component.
- Superpod monitor:View health status, server power status, network status, and other statistics for associated devices of the hypernode.
- Virtualisation monitor:Monitor virtualization resources to check the usage of virtual machine resources, compute pools, network pools, storage pools, and related alert information.
- Kubernetes monitor:View cluster health status, Node status, Pod status, Service Container, and performance information.
- AI Job monitor:monitor AI tasks allows you to diagnose them and view the results.
- Power supply monitor:Implement granular monitor for critical power infrastructure in data centers. The monitor targets include power racks and PDU units, with core metrics covering input/output voltage, cumulative power consumption, and power consumption.
- Thermal device monitor:The monitor metrics of CDU device include inlet/outlet temperature of main circuit, fan duty cycle, secondary supply/return liquid temperature, secondary flow rate, etc.

Performance Monitor

Performance monitoring includes performance view, metrics compare and performance predict. This feature helps user quickly understand the big picture and supports multiple filtering criteria to quickly locate the performance status of a particular device.

- **Comprehensive performance metrics display:** The performance list provides multiple performance metrics including server name, IP address, in-band IP, serial number, model, vendor, performance curve, Driver status, and so on. It provides user with a full range of device performance information to help user quickly understand the operation of device.
- **Driver Status:** The performance list can be used to check whether the device has a Driver installed and whether the Driver is online.
- **CPU and Memory utilization:** By monitor CPU utilization and memory utilization, user can identify system bottlenecks and performance hotspots, and then adjust or optimize resources accordingly.
- **Support index comparison for the same device type:** Support comparison of the same device at different time periods to monitor performance changes.
- **Multiple group comparison configuration:** The user can set multiple comparison groups, allowing the comparison of different device groups or performance metrics at the same time.

Log Monitor

Log monitor includes log search, log collector log download and index management. It can process log data automatically and intelligently, achieving efficient collection, filtering, analysis and storage of logs, thereby significantly improving the efficiency and accuracy of log processing.

- **Use a powerful search engine:** Choose a search engine that supports full-text search and complex query terms to ensure that the log data you need can be retrieved quickly.
- **Support a variety of retrieval methods:** Provide keyword search, regular expression search, time period search and other retrieval methods to meet the different retrieval needs of user.
- **Use index lifecycle management:** With the help of index lifecycle management capabilities provided by search engines such as Elasticsearch, we can automatically manage the creation, update, and deletion of indexes, reducing operational costs.
- **Support for custom templates:** Provide flexible custom template function, allowing user to define personalized collector templates according to specific log sources and business requirements.
- Support for log download:Users can customize batch device selection and create log generation tasks, which can be downloaded after successful completion.

Inspection Management

KSManage supports user to customize and add device inspection job. The system will automatically inspect device status and generate inspection record. User can preview and export inspection record from the system based on inspection record. At the same time, it supports automatically sending the inspection record to the customer, which includes the following functions:

- It supports user-defined inspection cycle and inspection device range, and the inspection type can be selected as disposable, interval or every week. Support user-defined out-of-band components for server inspection, including: CPU, memory, hard disk, PCIE, power, fan, temperature, etc.
- Support user-defined in-band metrics and desired range settings for server inspection, including: CPU usage, memory usage, hard disk usage, IO usage, swap usage, SELinux Expected value, firewall status, etc.
- User are supported to bind notifications to user in the inspection job, and the system will automatically send the inspection report to user after the inspection is completed.
- It supports user-defined health status inspection ability of various device components.

O&M Tools

- **Ping tool:** Detects the network connectivity of the target IP or domain name, assesses latency and packet loss rate, and is suitable for troubleshooting basic network issues.
- **Network Detect:** Verify the network connection status of devices such as servers and switches.
- Diagnotic tools
  - It supports user to conduct network connectivity tests on IP /host name based on PING.
  - It supports user to diagnose according to SNMP, IPMI, Redfish, HTTP protocol and Model.

3.3.4 Principle

Architecture design: Following the concept of Service Mesh, a distributed architecture of "one center and multiple grids" is designed to solve the problem of high complexity of cloud data center management.

Collection optimization: High-performance data collection technology for massive infrastructure was designed, and methods such as IPMI coroutines and session management were adopted to solve the problem of delayed detection of device failures.

Large-scale scheduling: Based on large-scale discrete scheduling at the job center, it resolves the performance bottleneck of data collection in large-scale monitor scenarios and enhances system stability.

Log gateway: Based on open-source log components, it realizes multi-type low-load log collection, improves the storage compression algorithm of open-source components, and reduces storage space.

3.4 Alert Management

3.4.1 Definition

Alert management provides alert data receiving, processing, notification and other functions, supports the whole life cycle management of alert, and helps operation and maintenance personnel to quickly troubleshoot according to alert information.

3.4.2 Value

Through alert management, user can centrally monitor the alert of the device itself and quickly locate the faults that have occurred in the system and network.

The value of Alert management includes:

- A variety of alert filtering methods are provided to help operation and maintenance personnel quickly screen the concerned alerts and achieve accurate monitor.
- It provides alert full processing function to avoid the alert that the user is concerned about being dumped.
- Flexible alert rule configuration was supported, and massive alert were correlated and compressed to reduce alert noise and improve monitor efficiency
- The remote notification function is provided, and the reported alert is sent to the ICT system maintenance personnel by E-mail or short message, so that they can understand the alert situation in time.
- The user can view the relevant traceability information of the alert and the warning information that may be triggered in the real-time alert.
- Alert correlation analysis and alert trend analysis are provided for user to help user view the internal logic and correlation between alert events.
- The alert database overflow dump function is provided to avoid the loss of alert due to the lack of database space.

3.4.3 Function

Alert management provides alert subscription, alert rule management, alert noise reduction and compression, alert correlation, alert notification and other functions.

Alert Subscription

The system automatically checks the device subscription status based on the user-defined subscription policy. User can detect or initiate alert subscription based on SNMP Trap and Redfish protocols at any time. At the same time, user can check the messages sent by the devices received by the system, and customize the ability to identify SNMP Trap Oid based on the content of the messages. It specifically includes the following content:

- Support alert subscription policy settings, support custom protocol settings and intelligent subscription, the system can automatically find idle channels for subscription, does not affect other monitor platforms.
- Support alert subscription status query of the managed device.
- Support one-click subscription and one-click detection of device alert subscription status.
- Support auto-subscribe settings for unsubscribed devices.
- Support subscription detection cycle setting and subscription failure retry times setting.
- Support device subscription status export.
- Support system each collector gateway alert subscription target IP s ettings.
- Support for each device alert subscription log, subscription status detection log view.
- Support SNMP Trap Oid customization, Redfish definition, accurate analysis and identification of alert.
- Support SNMPv3 USM add, delete, update lookup and automatic message parsing.
- Support SNMP Trap and Redfish message query.

Alert Notification

KSManage provides multi-channel alert notification methods such as email, SMS, wechat, Dingtalk, Slack, LarkSmartRobot, PagerDuty, so that user can receive notifications according to their preferences and availability. At the same time, it supports custom email and SMS notification content templates, which can notify the alert information to the operation and maintenance personnel according to the notification policy formulated by the customer. This includes the following:

- Support email, SMS, Dingtalk, wechat, Slack, LarkSmartRobot and PagerDuty notification, provide custom configuration function.
- Support custom email and SMS notification content templates. It includes alert name, alert location, alert description, alert severity, alert type, clearance method, asset name, asset sequence number, possible cause, repair suggestion, first occurrence time, last occurrence time, recovery time, resource attribution, business attribution, asset IP, asset location, asset model, asset manufacturer, component name, component sequence number, etc.
- Supports customization of notification policies. The content covers: notification time, notification method, alert type, alert severity, notification user, whether to carry device log notification, etc.
- Support the filtering and tracing of alert notification records, support the combination screening of notification content, notification method and notification results, and quickly locate the concerned notification records.
- Support one-click on/off notification function.
- Support to set notification policy according to alert severity, alert name, and can also set notification policy according to specific alert /event.
- Support alert in a certain period of time is not confirmed, not resolved, custom notification to the relevant user.

Alert Rule Management

Alert rule includes alert redefinition rules, alert suppression rules, alert acknowledge rules and so on. System maintenance personnel can flexibly set alert monitor rules according to different scenarios. User can choose to customize the threshold configuration for some devices, and support the definition of alert conditions for each device type to meet different business requirements.

- Support to set blocking rules to block the alert/event to be generated and not concerned.
- Support alert severity redefinition and alert name redefinition.
- Support to set automatic confirmation rules by level to move the current alert in the clear state to the historical alert list.
- Support user to set different alert tone function according to the alert severity.
- Support user to choose whether to enable the alert sound function.

Alert Noise Reduction Compression

In Alert management, various factors (such as temporary network fluctuations, device transient instability, etc.) may lead to a large number of repeated alert messages, many

of which may be "noise" that does not really represent the problem. Through the alert noise reduction compression, the "noise alert " can be compressed and displayed centrally, reducing the interference and the pressure of the subsequent alert processing process. In addition, the alert information frequently generated/recovered by some known problem devices will also bring trouble and extra labor cost to the server monitor. The trigger threshold can be set by alert noise reduction compression to avoid the generation of this type of alert. This includes the following:

- Support for flexible creation of noise reduction rules. The alert generation/recovery count is supported according to the threshold, and the alert that does not reach the count is filtered.
- Supports the flexible creation of alert compression rules. Alerts are compressed according to different dimensions, and default compression rules are built according to user feedback and operation and maintenance experience.
- Support compressed alert /real alert centralized display, support diversified alert query.
- Support compression alert compression range and its life cycle unified display.
- Support compression alert in accordance with the independent frequency of alert notification.

Alert Correlation

KSManage has built-in alert model and algorithm, which dynamically generate alert dependency graphs based on system alert data. It supports user to quickly query the current alert dependency and propagation path in the system, helps to quickly gain insight into the internal logic and correlation of alert events, and provides customers with alert traceability and alert prediction capabilities. This includes the following:

- Support alert correlation customization ability.
- Support system to automatically update the alert correlation graph based on real-time and historical alerts.
- Support user to specify alert types for related alert queries.
- Support user to export alert correlation data.
- Support user to query the root alert and derivative alert.
- Support alert details to jump to view the root alert and derivative alert.
- Support user to view the relevant traceability information of the alert and the warning information that may be triggered in real-time alert.

Custom Alert

The custom alert feature enables users to flexibly define alert rules, thresholds, and notification policies based on business needs. This allows precise and proactive alerts for different monitor objects, effectively filtering out irrelevant information, focusing on critical issues, and improving operational response efficiency and system reliability.

- Set custom alert rules based on actual scenario requirements.
- Supports setting alert rules for part replacement.
- Edit or delete custom alert rules.

Alert Link

- It serves as the core engine for intelligent and automated operations response. By configuring correlation rules, the system automatically executes predefined repair or mitigation actions when specific alerts are triggered. For example, upon detecting a server outage alert, it automatically attempts to restart services on the standby node. This feature transforms passive alerts into proactive handling, significantly enhancing both fault recovery speed and operational automation levels.
- Support alert linkage to automatically operate the device power.
- Support alert linkage to automatically operate the UID light.
- Supports automatic execution of alert scripts.
- View execution records of alert linkage rules.
- Edit and delete alert linkage rules.

Alert Management Functions

Table 3- 1 Alert Management Functions

Type	Description
Active and Passive Alert Collection	KSManage provides real-time monitor and fault analysis of devices with active polling and passive receipt of alerts, thus reducing potential service risks.
Event Message Parsing	You can view SNMP Trap and Redfish messages received by the system. At the same time, you can customize message parse based on Trap OID attributes to convert them into device alerts.
Alert Display	KSManage displays alerts on the alert panel and in the real-time historical alert list according to alert severity, device type, alert source, room, and cabinet, thus controlling

	the real-time running status of devices managed by KSManage.
Alert Statistics	Supports multi-dimensional statistics, such as alert severity, alert quantity in the alert source, alert type, device model, server room alert quantity, and occurrence time distribution.
Alert Auto Confirm	User can perform the **enable/disable** operation on the alert, support the delay or immediate confirmation of the alert, and support the delay or immediate clearance of the confirmed alert information.
Alert Search	You can filter and search for recent and historical alerts by combining search conditions such as alert status, alert severity, alert type, location, server room, cabinet, logic type, alert source, and component type to quickly locate and handle the alert.
Alert Redefinition	Supports the redefinition of alert names and severities, including the customization of alert types and the customization of severities of specified resources. Supports flexible conversion of various alerts and events.
Alert Suppression	You can create alert suppression rules to block certain unimportant or unconcerned alerts, thus avoiding redundant information.
Alert Sound	You can select whether to enable the alert sound based on the alert severity.
Fault Repair	KSManage provides functions such as connection setting of alert repairs, repair rules setting, and repair record tracking.

3.4.4 Principle

Alert management provides an alert processing mechanism to simplify alert, help system maintainers free from massive alert, and improve the efficiency of handling alerts.

Alert Source: The alert source attribute is used to describe the way alerts are generated, including device-side push notifications, which indicate that the alerts are initiated by the device side through subscription, and the platform receives and parses the generated alerts. Active monitor refers to the alerts generated by the platform through comparative analysis based on the collected device data in combination with the built-in or custom configured alert rules. System health monitor indicates alerts from the platform itself, such as service anomalies and abnormal user logins on the platform. And some alerts in special scenarios, such as virtualization, are used to distinguish alerts from virtualization platforms; Driver, used to distinguish alerpushed by an in-band driver.

Alert generation process: In most scenarios, the main sources of alerts on the platform are "active monitor" and "device-side push". Among them, the generation of active monitor alerts relies on the collection of device data, the setting of alert rules, as well as corresponding data analysis and comparison.

- **Data collection:** The KSManage utilizes various data collection methods, such as sensors and log collection tools, to obtain real-time operational data of the infrastructure. These data cover the performance metrics of the device (such as CPU usage rate, memory occupancy rate, disk I/O, etc.) and network status (such as bandwidth utilization rate, packet loss rate, latency, etc.).
- **Alert rules:** Based on the normal operation range of resources and business requirements, set reasonable thresholds for each monitor metrics. At the same time, define the alert rules, clearly stating that an alert will be triggered when the monitor metrics exceed the threshold or meet specific conditions. Alert rules can be based on a single metrics or a combined logic of multiple metrics.
- **Data analysis and comparison:** The platform conducts real-time analysis and processing of the collected real-time data, and compares it with the set thresholds and alert rules. Once the data exceeds the threshold or meets the conditions of the alert rules, the system immediately triggers the alert mechanism.

The generation of device-side push alerts depends on the subscription of device alerts, the reception of device alerts, and the parsing of the original alert text:

- **Alert Subscription:** The platform provides alert subscription templates to facilitate subscription management of devices. User can specify subscription templates for devices when they first scan and manage them, or uniformly manage subscription templates for all devices in the alert subscription management function list.
- **Alert reception:** After the managed device is configured and subscribed, an alert will be pushed to the platform subscription address in real time when it occurs. After receiving the message, the platform records the original alert information in the message list.

- **Alert analysis:** Generally, the original alert information pushed by devices is difficult to understand and lacks standardization. The platform has made plug-in adaptations based on different vendors and models of devices, extracting key information from the original alert message, such as alert severity and alert occurrence location, and mapping them into standardized alert name for display.

Alert notification: After an alert is triggered, the platform promptly notifies relevant management personnel of the alert information through multiple ways.

Alert Severity: According to the severity and impact range of the alert, the alert is graded, such as fault, serious, moderate, slight, and event, as shown in Table 3- 2. Different severities of alerts adopt different processing priorities.

Alert Severity Definition

Table 3- 2 Alert Severity Definition

Severity	Description
Cricial	An alert that interrupts service and requires immediate troubleshooting
Major	An alert that affects business and requires immediate troubleshooting
Medium	An alert that does not impact existing service operations but needs to be repaired to prevent deterioration
Minor	An alert that does not impact existing service operations but may have such impact if left unattended, for which countermeasures may be taken as required
Event	An event that does not affect existing business

Alert All-round Monitor

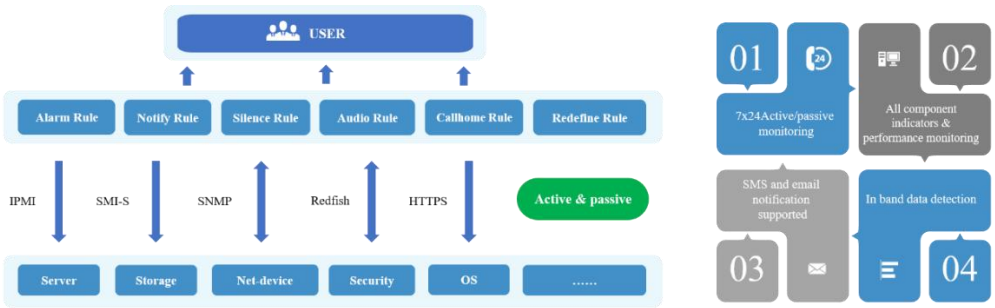

Figure 3- 2 Alert All-round monitor

3.4.5 Metrics

Table 3- 3 Key Metrics of Alert Management

Metrics Item	Metrics **Value**
Real-time alert storage capacity	One hundred thousand
History alert storage capacity	Three hundred thousand
Event alert storage capacity	Three hundred thousand
Trap packets storage capacity	One hundred thousand
Redfish message storage capacity	One hundred thousand
Alert processing capability	One hundred Note: Continuous alert processing capability :100 alerts per second. Storm alerts processing capability :1000 alerts per second. The maximum processing capability is 15 minutes. If the alert exceeds the threshold, the alert may be reported late or lost.

3.5 Configuration Management

3.5.1 Definition

Through software and standardized process, task such as deployment, upgrade, configuration of resources is automatically performed, and human intervention is reduced to improve efficiency, accuracy, reliability, and consistency. It covers a range of techniques and practices such as automation script, configuration management tool, continuous integration/continuous deployment tool, etc. to achieve efficient, reliable, repeatable operations.

3.5.2 Value

Through standardized templates and automated pipeline technology, the configuration management module upgrades large-scale device configuration management from manual operation to intelligent orchestration, which significantly improves the efficiency of large-scale device upgrade and deployment. Its value is not only reflected in efficiency improvement and cost optimization, but also provides a solid guarantee for efficient operation and business continuity of data centers through standardization, traceability and intelligent means.

- **Efficient Commissioning:** Full-process automation significantly shortens the business launch cycle. By using commissioning templates to initialize multiple servers simultaneously, only key parameters need to be confirmed manually. The traditional deployment process, which used to take much longer, is now compressed to just a few hours, greatly improving deployment efficiency.
- **Improve deployment efficiency:** The use of PXE opening technology, efficient configuration of server resources, automatic deployment of basic systems and key applications, greatly shorten the deployment time, improve work efficiency. Automatic deployment and precise parameter setting reduce the risk of errors caused by human operation and ensure the accuracy and consistency of the server environment.
- **Improve firmware management:** Firmware upgrade is an important task of the daily operation and maintenance of the server. By updating the firmware version of the server hardware, you can fix potential security vulnerabilities and improve the performance and stability of the device.
- **Personalized firmware configuration:** Personalized firmware configuration is made based on the specific requirements of the server, such as BMC log setting, BIOS configuration, NTP configuration, SMTP setting, etc.
- Integrate multiple automated management functions, simplify daily operations, enhance operation and maintenance efficiency, and ensure system stability and security.
- By parsing the hardware parameters of the server and the firmware support capabilities, a configuration model is automatically generated, which is convenient for direct invocation when new devices are delivered and avoids configuration deviations caused by differences in experience among operation and maintenance personnel.

3.5.3 Function

Commissioning

KSManage provides a complete automated process solution from server listing to normal service. It can be divided into the following processes: configure and run the commissioning dependent services, auto-manage devices, import image files, create operation and maintenance templates, select target devices, execute the configuration process, and view the configuration history. In addition, KSManage provides a unified service management interface to automatically manage the dependency environment of the start delivery. Separate device management interface to monitor and manage each device in the opening delivery state.

PXE

PXE management through efficient configuration of service resources, automatic deployment of basic systems and key applications, accurate parameter setting, generation of detailed delivery reports, to ensure that the server environment quickly ready for the stable operation of the follow-up business to build a solid foundation. User can complete the operating system installation, application deployment and system update in a short time, which significantly improves the operation and maintenance efficiency. At the same time, PXE technology reduces hardware costs, enhances system security, free task orchestration, provides multi-scenario operation and maintenance capabilities, brings higher flexibility and scalability to the data center, and is an indispensable management tool for modern data centers.

Firmware Upgrade

The platform offers upgrade configuration functions for firmware such as BMC, BIOS, PSU, main board, and rear baack panel, covering the upgrade functions of mainstream server firmware.

Firmware configuration

Firmware configuration is the process of personalizing the firmware based on the specific requirements of the server. Administrators need to understand the functions and features of the server hardware and make relevant configurations based on the actual situation, such as BMC log setting, BIOS configuration, NTP setting, SMTP setting.

Software Automation

The software automation management integrates functions such as file distribution, script execution, user management, service management, software installation, and driver management, aiming to achieve rapid deployment and efficient operation and maintenance in IT environments. By automating the toolchain, daily operations are simplified, response speed is enhanced, and system stability and security are ensured. This solution is applicable to multi-platform environments, helping enterprises achieve intelligent management of IT resources and accelerate business innovation and development.

Model

The server model is an abstract manifestation of the server's configuration capability. Through specific metric parameters, the management model of the server is defined.

- **Template:** Through the existing server device, quickly build each parameter defined for the server. The server configuration model is constructed by calling the interface to obtain the parameters of the relevant firmware and the support capabilities of this baseline for each configuration function, etc.
- **Configuration File:** Based on the server configuration model, configuration file instances can be quickly produced. These configuration file instances include SMTP, user, TRAP, NTP, DNS, etc, for parameter editing. Once the editing is completed, the edited data can be saved as server configuration files.

3.5.4 Principle

Based on standardized in-band and out-of-band management protocols, a complete automated management system for the entire lifecycle of servers has been established. It proposes an integrated automation process from “device shelving” to “application readiness,” integrating operations such as hardware initialization, system deployment, container configuration, and business deployment. This achieves a closed loop of full-lifecycle infrastructure automation for both hardware and software, breaking the traditional separation between “hardware maintenance and software delivery.”

Commissioning

Through self-research infrastructure automation standard workflow, BMC configuration, BIOS configuration, RAID configuration, system deployment and other operation and maintenance links are standardized, streamlined and automated. A unified adaptation layer is constructed, implemented by a hierarchical architecture, the device type model is abstracted, and the hardware details are hidden based on object-oriented interface. The compatibility efficiency of heterogeneous devices is greatly improved.

Software Automation

Deeply integrating out-of-band management technology and the Ansible ecosystem, it is equipped with multiple hardware and software management template components to precisely control core components such as BIOS, BMC, PSU, and RAID of heterogeneous devices from multiple vendors. Existing automated deployment files (such as Playbook, Role and Inventory) can be reused, and over 7,000 functional modules can be invoked, transforming complex operations such as deploying Kubernetes clusters and configuring Nginx load balancing into "one-click execution".

3.5.5 Metrics

Single node mode can ensure concurrent batch task execution of more than 50 devices, and complete 1000+ servers on the shelf in a single day.

3.6 Energy Efficiency Management

3.6.1 Definition

Energy efficiency management is a systematic process that focuses on the core goals of improving energy utilization efficiency and controlling carbon emissions. It involves comprehensive means of technology, policy and management to monitor, analyze, optimize and provide decision support for the entire chain of energy production, transmission and consumption. It formulates and implements refined power consumption strategies to monitor and intelligently control the energy consumption of devices in real time. It aims to enhance energy utilization efficiency, reduce costs, and support sustainable development goals.

3.6.2 Value

Reduce operating costs: Powerefficiency management monitors and intelligently controls the energy consumption of resources in real time through refined power consumption strategies. In data centers, optimizing the power consumption policy of servers can reduce energy waste and electricity expenses, significantly lowering the operating costs of enterprises.

Optimizing device operation and extending lifespan: The energy efficiency optimization function conducts in-depth analysis of device operation data, providing maintenance personnel with detailed insights into the device's operational status. For servers that have been operating at high energy consumption for a long time, the energy efficiency optimization function can analyze their workload and resource allocation, and suggest reasonable adjustments to task allocation or hardware upgrades to reduce power consumption and optimize device performance. This optimization not only helps the device operate in an energy-efficient manner, but also reduces the risk of failure caused by overheating, overload and other issues, thereby extending the service life of the device and lowering the cost of device renewal.

Supporting sustainable development and compliance: The carbon emission management function can precisely monitor and control carbon emissions resulting from energy consumption. The system accurately measures the carbon emissions generated by energy consumption, supports the generation of visual reports, and helps enterprises set scientific carbon reduction targets. This helps enterprises comply with relevant regulations and industry standards, enhance their social image and brand value, and strengthen their competitiveness in sustainable development.

Facilitating scientific decision-making: Energy efficiency management provides detailed energy consumption data and analysis reports, offering a basis for management decision-making. Analyzing energy consumption trends can predict future demand

changes, plan energy supply schemes in advance, evaluate the effectiveness of energy conservation and emission reduction measures, and select the optimal energy efficiency improvement strategies. This data-driven decision-making model helps enterprises make more scientific and forward-looking decisions in energy management and sustainable development, laying a foundation for long-term stable development.

3.6.3 Function

Power Consumption Policies

KSManage allow user to formulate corresponding power consumption limit strategies for a single device or multiple devices within the same physical space to limit the maximum power consumption of the server. The policy includes:

- **Enabled or not:** Once a policy is established, it can be individually disabled or enabled at any time.
- **Time period:** Once the policy is enabled, it will take effect within the set time period.
- **Power consumption Upper limit:** The main role of the policy is to limit the power consumption of the device by reducing the CPU frequency, etc. When the policy is enabled and implemented, the power consumption of the device is limited near the set power consumption upper limit.

Energy Efficiency Optimization

KSManage assists operation and maintenance personnel in comprehensively reducing the energy consumption of the data center through six major functions, including temperature analysis, utilization analysis, power distribution analysis, server power consumption characteristics, load distribution analysis, and advanced power consumption models.

Carbon Emission Management

KSManage provides carbon emission management, which can manage carbon assets and carbon emissions of data centers. After user input carbon assets, carbon emission management can calculate, analyze, verify and reduce the carbon emissions of the data center, and generate a carbon emission trend graph.

- **Carbon Asset:** user can input carbon asset information according to their own actual situation. After employment, they can know the usage and distribution types of carbon assets in the data center.
- **Carbon Emission Management:** data center related information needs to be maintained in carbon emission management, including PUE, quota, carbon emission coefficient information. KSManage will calculate the carbon emission trend chart,

monthly utilization rate of carbon quota, and estimated use days of carbon assets according to the maintenance information.

3.6.4 Principle

Energy Efficiency Management

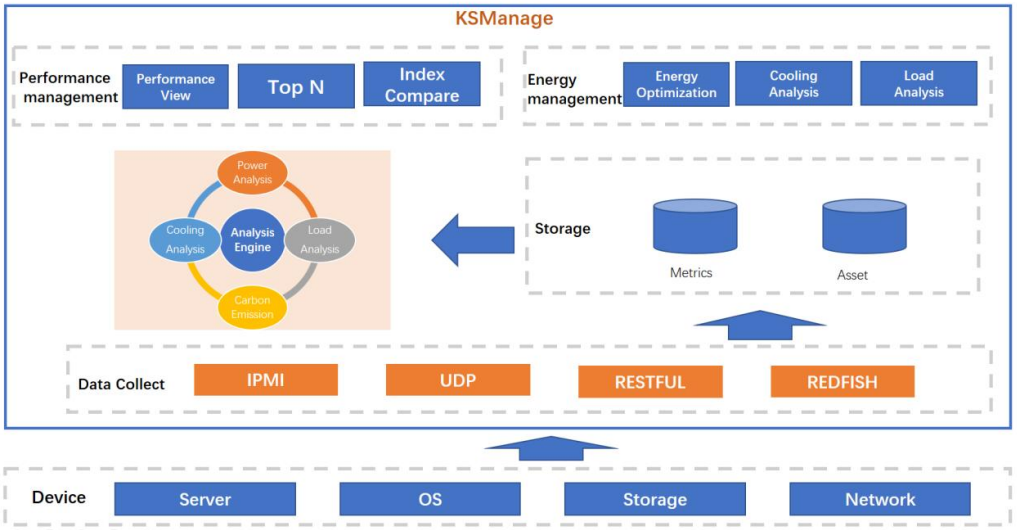

Figure 3- 3 The Principle of energy efficiency management

Power Consumption Policies

Power consumption strategies are used to dynamically limit the upper limit of power consumption for a single or multiple devices. The principle is that within a server power consumption collection cycle, based on the historical power consumption data of the server, the upper limit of the total power consumption of multiple devices is reasonably allocated to each server. Then, based on the calculated upper limit of the expected power consumption for each device, the power consumption limit command is executed for each device. Repeat the above steps until the next cycle.

Energy Efficiency Optimization

Statistical analysis is conducted using various methods based on the collected data such as power consumption, temperature, and CUP of the device.

- **Power consumption characteristics:** Statistically analyze the upper and lower limits of historical power consumption and their distribution for each model of device.
- **Power consumption prediction:** Establish a power consumption calculation model based on the historical power consumption of the device and CUP data.
- **Load distribution:** Discover the target devices available for migration within the data center based on the CUP data of the devices.

- **Temperature analysis:** Based on the temperature information of the device the temperature distribution in the computer room is statistically analyzed to obtain the assessment result.
- **Utilization analysis:** Based on the historical power consumption of the device and CUP data, the utilization of the device within the data center is statistically analyzed.
- **Power distribution analysis:** Based on the power consumption data of the device, assess the power distribution risks of the cabinets within the data center.

Carbon Emission Management

Based on the power consumption data of the device within the data center, calculate and summarize the carbon emissions of the device.

3.6.5 Metrics

Refrigeration Analysis

- ASHRAE recommended temperature 18℃-27℃.
- ASHRAE Level 1 permit temperature 15℃-32℃.
- ASHRAE secondary license temperature 10℃-35℃.

3.7 Knowledge Base Management

3.7.1 Definition

Knowledge base management is a system that stores and manages all kinds of knowledge information related to the enterprise IT infrastructure, such as user manuals, troubleshooting, frequently asked questions, etc. Through structured organization, it is easy to retrieve, share and apply, and effectively improve the efficiency and problem-solving ability of the operation and maintenance team.

3.7.2 Value

The value of knowledge space lies in centralizing the storage, organization and sharing of professional knowledge, improving the efficiency of information retrieval, accelerating problem solving, promoting team collaboration and innovation, and ensuring the continuous updating and inheritance of knowledge, which brings higher work efficiency and decision-making quality to the organization. It is a key tool to promote business development and knowledge management. Reduce repetitive learning and trial-and-error costs by sharing best practices and FAQs from our knowledge space.

Knowledge base management serves as the "intelligent hub" of KSManage. Through intelligent retrieval and spatial structuring, it transforms scattered knowledge assets such as operation and maintenance experience, manuals, fault cases, and operation norms into reusable digital resources. It not only enhances operation and maintenance efficiency and reduces costs, but also helps enterprises build core competitiveness and support digital transformation and sustainable development through knowledge accumulation and reuse. Its core value lies in:

- Operation and maintenance personnel can quickly locate solutions by querying the knowledge base, reducing reliance on experts and saving time and resources.
- Structured Spaces support cross-team knowledge sharing, break down information silos, and promote cross-departmental collaboration.
- The space management function facilitates timely updates and maintenance of knowledge entries, ensuring the accuracy and usefulness of the knowledge base content.

3.7.3 Function

Knowledge Search

Knowledge space ssearch is an efficient information query method, which can quickly locate and obtain the required knowledge or information in the knowledge space through specific retrieval systems or tools. Enter keywords or sentences, and the system uses advanced algorithms to deeply scan various knowledge sources to help user quickly find relevant document, solution, or best practices. Quick queries can also be made through the frequently asked question provided by the system.

Space Management

Knowledge space is a collection specially used for storing, managing and retrieving knowledge, which is organized in a specific structure and classification way, so that user can easily search and obtain the information they need. Through reasonable planning, integration and optimization of knowledge and information in the knowledge space, the efficiency of knowledge storage and the accuracy of knowledge retrieval are improved.

3.7.4 Principle

Taking into account multiple factors such as performance, scalability, and ease of use, structured word segmentation and inverted indexing technology are used to build an industry-specific dictionary. Combined with intelligent word segmentation algorithms, precise segmentation is achieved, the initialization of the knowledge base is completed, and functions such as knowledge updates and version control are supported to ensure the accuracy and usability of the knowledge base.

Structured Word Segmentation

By using the word segmentation technology in natural language processing, continuous text sequences are split into word units with independent semantics. In the device operation and maintenance scenarios, a large amount of unstructured text such as device log, fault description, and maintenance record is structured. The word segmentation algorithm based on rules, statistics, and deep learning is customized and optimized for content such as device professional terms, industry abbreviations, and special expressions. Ensure the accuracy and professionalism of word segmentation.

Inverted Index

The inverted index takes words as the core and establishes a mapping relationship of "word-document list". By associating each word in the document with all the documents containing that word, an efficient data retrieval network is formed. The data structure consists of two key parts: the dictionary and the inverted list. The dictionary stores all non-repetitive words, while inverted list records the position, frequency of occurrence and other information of each word in the document. By structuring and indexing disordered text data, the retrieval efficiency is increased exponentially, greatly reducing the time cost of obtaining effective information from massive data.

3.8 Remote Management

3.8.1 Definition

Remote management capabilities enable user to access, monitor, configure and maintain the IT infrastructure managed by the platform in real time.

3.8.2 Value

Remote management not only greatly improves the work efficiency of the operation and maintenance team, reduces the cost caused by site visits, but also enhances the flexibility and response speed of the organization in operation and maintenance management, and ensures the security and reliability of remote operation by integrating multiple security authentication and encryption technologies, providing solid technical support for the organization's digital transformation and business development.

3.8.3 Function

Remote Connection

Remote connection support is realized through a variety of flexible methods, not only covering KVM (including VNC connection, achieving remote access and operation through an intuitive graphical interface; Java connection, with its cross-platform feature,

ensures stable connection in different environments. H5 connection, no additional plugins required, can be conveniently accessed through a browser and Terminal (command-line interface), achieving comprehensive control over remote devices.

KVM Video Audit

The KVM video auditi function has comprehensive and detailed recording capabilities, providing complete and accurate records of all operations conducted through KVM, covering key information such as operation time, operator, specific operation instructions, and operation screens, to ensure the compliance and security of operations. At the same time, it provides user with convenient query and retrieval functions to quickly locate specific operation records. In addition, it supports the download function of operation videos to meet user's local storage and backup needs, and provides an online playback function.

3.8.4 Principle

Remote management utilizes KVM or Terminal technology to achieve remote access and control of different resource devices. User initiate connection requests through the platform interface. The platform establishes a secure remote session channel based on the request type and resource device information. Once the connection is established, administrators can remotely operate the infrastructure through the command line interface. These operation instructions are transmitted to the server-side devices or systems via the network to perform the corresponding management tasks.

3.8.5 Metrics

KSManage connects to a maximum of five devices remotely at a time.

3.9 Statistical Analysis Management

3.9.1 Definition

Statistical analysis management provides an end-to-end data analysis framework and report presentation platform, supporting administrators to view and compare data from different dimensions and generate required reports.

3.9.2 Value

By integrating, analyzing and visualizing massive data, statistical analysis management helps the management to fully understand the system operation status and find and solve potential problems in time. At the same time, statistical analysis management also supports multi-dimensional analysis and data mining, which provides strong support for

optimizing resource allocation and improving operation and maintenance efficiency. It is a key tool for enterprises to realize intelligent and systematic management.

- **Real-time monitor and management:** Provide detailed information on various resources in the infrastructure to help administrators understand the operational status of resources in real time. By monitor the usage of resources, it helps administrators optimize resource allocation and improve resource utilization.
- **Rapid response and handling:** Record current and historical alert information to help the operation and maintenance team respond and handle problems quickly, reducing downtime due to faults.
- **Problem analysis and Prevention:** By conducting correlation analysis and trend analysis on alerts, it helps the operation and maintenance team identify potential problems and abnormal situations, and take measures in advance for prevention and resolution.
- **Device maintenance and cost control:** Record information such as the maintenance status of device to help administrators understand the maintenance status of device, reasonably arrange maintenance budgets, and control maintenance costs. Remind the operation and maintenance team to carry out timely device maintenance and repair to ensure the stable operation of the device.
- **Performance optimization:** Through the analysis of performance data, it helps administrators identify performance bottlenecks, make optimization adjustments, and improve device performance.
- **Customization and flexibility:** User can customize timed reports to meet the specific needs of different user. Automatically generate and send reports to reduce the workload of manual collection and summary of reports and improve work efficiency.

3.9.3 Function

Resource Report

Resource report display detailed information of various resources in the infrastructure, including server reports, disk arrays/tape libraries report, room and cabinets reports, business reports, network device reports, security device reports and other reports. Real-time monitor of the server's operating status, including CPU, memory, network card, disk, network port, etc.

Alert Report

Record the current and historical alert information, including the alert name, severity, alert source, and occurrence time, helping the O&M team quickly respond to and handle problems. Correlation analysis and alert trend analysis can be conducted on the alerts.

Maintenance Report

Record the maintenance status, maintenance type, purchase time, and expiration time of the device to remind the operation and maintenance team to maintain the device in a timely manner to ensure the stable running of the device.

monitor Report

Track all performance metrics of the device, including CPU utilization, memory usage, and CPU temperature, enabling administrators to monitor real-time operational status at a glance.

Performance Report

Record various performance metrics of the device, such as CPU utilization, memory utilization, CPU temperature, etc. of the server, allowing administrators to have a clear understanding of the real-time operating status of the device at a glance.

Custom Report

Support user-defined scheduled reports. User can choose the type of report by themselves, and different report types can be set with different execution times.

3.9.4 Principle

The principle of the report module on KSManage is to collect all kinds of data such as asset, server, alert, storage, performance and maintenance, and use data processing technology to sort out, count and analyze these original data. You can select a preset report template or customize report content and set scheduled tasks as required. The system automatically generates reports based on the configuration to display data in an intuitive and accurate manner, helping managers to gain insight into system status in a timely manner, optimize resource allocation, and ensure efficient and stable infrastructure operation.

Each service module periodically reports report data, which is processed by the report management module and displayed in charts. Figure 3- 4 shows the implementation principle of report management.

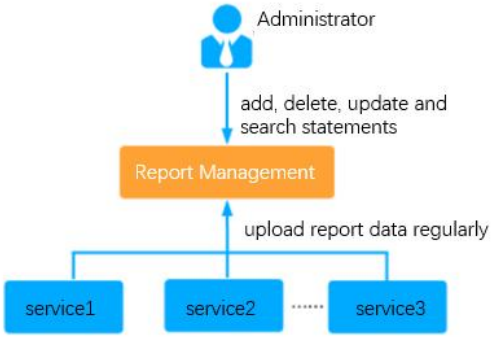

Figure 3- 4 Report Principle

3.10 Bussiness View

3.10.1 Definition

The Business View serves as a visual management interface for infrastructure platforms, focusing on operational dimensions. It establishes a hierarchical resource management system centered on business logic, utilizing system presets (e.g., regional, group, organizational, and business views) and user-defined views. This view correlates physical resources with business tiers, enabling unified monitor, analysis, and maintenance of all-domain resources through left-side group navigation and right-side multidimensional statistical analysis panels. It empowers users to swiftly assess resource distribution, status, and health from a business perspective.

3.10.2 Value

- **Business-oriented resource management:** By categorizing and aggregating resources by region, group, organization, and business, administrators can manage resources based on actual business needs, improving operational efficiency and decision accuracy.
- **Multidimensional data visualization analysis:** Integrates key metrics such as total resources, operational status, alert statistics, and aging rate to present a comprehensive resource overview in chart form, helping users quickly identify anomalies, assess capacity, and evaluate risks.
- **Flexible customization**: Users can create custom views and grouping rules to meet their management needs, adapting to data center scenarios of varying scales and architectures, and enhancing the platform's scalability and adaptability.
- **Closed-loop operation support**: By integrating the three-tier information view (Overview-Device List-Alert), it provides comprehensive analysis from macro

statistics to detailed data, enabling closed-loop management for problem discovery, localization, and resolution.Function.

3.10.3 Function

The Business View provides two types of management interfaces: system preset views and custom views. The specific functions are as follows:

1. View Type

System preset view:

- **Regional view**: Resources are divided by China and overseas regions, supporting further drilling down to specific data centers or regions.
- **Group view**: Display resources by technology or function (e.g., production clusters, test environments).
- **Organization view**: Show resource distribution by department or team.
- **Business View:** Aggregated resources by business system or application.Custom Views: Users can create personalized views based on rules such as attribute labels and resource types.

Custom View

- Users can create personalized views based on attribute labels, resource types, and other rules.

2. View interface layout

- **Left navigation bar:** Shows the hierarchical grouping structure in the current view, supporting expansion/folding and quick switching.
- **Right-side statistics panel:** Displays key metrics within this view, including total devices, active devices, current alerts, and device obsolescence rate (based on service life or maintenance status).
- **Bottom tab: Overview:**
  - Shows aggregated charts such as resource type distribution (servers, storage, networks, etc.), vendor model statistics, resource status ratios, and alert model rankings.
  - Device list: Shows detailed information for all devices in this view. You can filter, sort, and export by properties.
  - Alert: Displays real-time alert lists for connected devices, including alert level, status, occurrence time, and processing progress.

3. Interaction and Operation

- Supports selecting from the view layer to the management page of specific devices.
- You can manage and configure devices directly in the view, including status changes and alert confirmation.
- View configuration allows you to adjust grouping rules, display fields, and statistical metrics.

3.10.4 Principle

The business view is dynamically generated based on the platform's resource data model and user-defined classification rules.

- **Data Aggregation Works as Follows:** The platform automatically collects device data through its resource discovery and entry features. By analyzing device attributes (e.g., geographic location, business labels, organizational affiliation) and user-configured grouping rules, it categorizes devices into corresponding views. Real-time data such as resource status and alerts are synchronized to the view statistics panel via the monitor module.
- **View Generation Logic**
  - Preset View: Build a resource index tree based on system-built dimensions (e.g., region, business), and associate resources with views through tag matching.
  - Custom View: Users define the view scope by setting filter conditions (e.g., device type, manufacturer, custom attributes). The platform dynamically filters and presents the resource set based on the rules.
- **Visualization and Interaction**: The view frontend retrieves aggregated data from the resource management database via API, renders statistical metrics using the chart engine, and optimizes large-scale resource list display performance through pagination and lazy loading. User actions (e.g., group switching, device operations) update the view state in real time via an event-driven mechanism.
- **Integration with resource management:** The business view serves as a visual interface for resource management functions, deeply integrated with modules such as resource access, monitor, and configuration. For example, device operations triggered in the view will call the resource configuration interface, while data on the alert tab comes from the platform's unified alert center.

3.11 O&M Assistant

3.11.1 Definition

The Operations Assistant is an AI-powered intelligent maintenance tool that leverages machine learning to rapidly identify solutions from a predefined knowledge base. Through automated and intelligent processes, it provides maintenance personnel with smart decision-making recommendations, helping enterprises and organizations manage IT infrastructure, applications, and services more efficiently. This enhances system reliability and performance while reducing operational costs.

3.11.2 Value

O&M assistant reconstructs the infrastructure management system through the twin engines of intelligent Q&A and alert analysis, and its core value is lied in:

- **Enhance O&M efficiency:** The intelligent Q&A engine compresses the knowledge acquisition path from traditional document retrieval to a response within seconds. Alert analysis engine improves the efficiency of average fault location through the reasoning chain analysis of AI agent. Through multi-round dialogue and collaboration, key details are confirmed, precise repair suggestions are provided, the complexity of operation and maintenance and resource consumption are reduced, and efficiency is improved.
- **Reduce operating cost:** 7× 24-hour self-service inquiry and response cover high-frequency questions, reducing reliance on human customer service. Common functional issues are provided with solutions by product R&D engineers to the knowledge base, forming a dynamically updated "experience pool" to avoid repetitive inquiries, effectively solve problems, promote the intelligent transformation of operation and maintenance work, and help enterprises reduce costs and increase efficiency.
- **Precise problem location:** Quickly analyze the root cause of alerts in complex systems, avoid repeated trial and error in surface repair, save time for fault detection and repair, and ensure business continuity.

3.11.3 Function

Intelligent Q&A

This platform builds an intelligent Q&A system, which can transform product documents, operation manuals and other materials into a unified knowledge base and achieve natural language interaction with the help of large models. User only need to ask questions in natural language, and the system can parse the semantics in real time,

quickly match and generate precise answers. The large model will synchronously update the operation guide to ensure the timeliness of knowledge. The multi-round dialogue function can further clarify user needs and improve the accuracy of responses.

Alert Analysis and Fault Prediction Diagnosis

This platform is equipped with an AI Agent intelligent analysis engine, which can quickly process information when an alert occurs and hand it over to a large model for analysis. It can deeply analyze the root cause of the alert and provide repair suggestions. User can also conduct exploratory troubleshooting through multi-round conversations to achieve efficient and precise problem-solving.Furthermore, the platform provides real-time monitor and predictive maintenance for critical hardware components like hard drives and memory modules. By integrating advanced fault diagnosis capabilities, it enables intelligent O&M management that seamlessly transitions from early warning to actionable resolution.

3.11.4 Principle

Intelligent Q&A

The intelligent Q&A function is based on advanced natural language processing technology to accurately parse the semantics of user input and grasp the core of the question. It relies on a vectorized processing knowledge base, and combines the semantic understanding and generation capabilities of large model to create precise and practical responses. The real-time update mechanism of large model ensures the timeliness and accuracy of information. In addition, intelligent Q&A supports multi-round conversations. Through continuous interaction, the system can better understand the context and gradually clarify the specific needs of user, thereby providing more accurate answers.

Alert Analysis

The alert analysis function relies on the built-in AI Agent intelligent analysis engine. When an alert occurs in the system, operation assistant will first preprocess the alert information and then send the processed information to the large model for in-depth analysis. Large model leverages their powerful reasoning and pattern recognition capabilities to interpret alert information and identify the root cause of the problem. Through analysis, the intelligent system will infer the root cause of the alert and generate corresponding repair suggestions based on the preset rules and solutions in the knowledge base.

3.12 Intelligent Computing Center

3.12.1 Definition

The operation and maintenance management of intelligent computing center is a comprehensive intelligent control center for AI computing power infrastructure. By building a three-in-one operation and maintenance system of "resource – task - network", it realizes the unified management and multi-dimensional monitor of cluster computing, storage, and network resource, thereby enhancing the operation and maintenance efficiency of intelligent computing center.

3.12.2 Value

Panoramic resource visualization and in-depth insight: It provides a unified monitor view of the cluster, which can globally view the cluster's resource usage, including real-time monitor of core performance metrics such as computing, storage, network, and job, and quickly identifying global hotspots and anomalies.

Network link fault location: It provides real-time monitor of end-to-end network link performance, capable of conducting second-level monitor of key metrics such as network link traffic, delay, jitter, and packet loss based on network topology, accurately locating network fault points, and identifying network anomalies and bottlenecks.

Job monitor and Diagnosis: It provides job link graph display, capable of assessing the health status of job operation based on the device status and port monitor data on the job link. Combined with multi-dimensional job monitor data aggregation analysis, it supports job performance bottleneck analysis and job fault cause location.

3.12.3 Function

Cluster Management

- **Multi-dimensional resource monitor:** It supports an integrated view of node, storage, network, and task status, enabling a global grasp of the cluster's resource status and usage.
- **Active alert management:** It supports timely alerts for resource anomalies, capable of triggering multi-level alert notifications based on threshold rules, and associated display of related resources and logs to accelerate fault diagnosis and recovery.

Compute Management

- **Node performance Insights:** It supports real-time monitor of over 200 core metrics such as CPU, memory, disk, I/O, second-level monitor of key performance

data and log information of resources, promptly identifying abnormal issues of computing nodes to ensure business continuity.

- **Accelerator card real-time monitor:** It supports statistical monitor and management of heterogeneous accelerator cards, supports monitor of accelerator card usage, GPU memory usage, temperature and power consumption at the second level, and can perform comprehensive analysis based on performance trend statistics to realize active early warning of performance deterioration.

Storage Management

- **Efficient space management:** It supports refined management of storage pool, directory and storage node, real-time monitor of storage space usage, rational allocation of storage resources, avoiding space waste and chaotic data access, and ensuring the security and integrity of data.

Network Management

- **Network Anomaly and bottleneck Analysis:** It supports real-time monitor of end-to-end network link performance, conducting second-level monitor of key metrics such as traffic, delay, jitter, and packet loss, and accurately locating network fault points.
- **Network topology management:** It can generate a cluster network topology relationship graph based on the physical and logical relationships of cluster resources, support automatic perception and update of topology changes, and support automatic marking of abnormal nodes and interrupted links.

AI Job Management

- **Job path visualization:** It supports the generation of job link graph based on the compute, network, and storage associated with job operation, and can evaluate the health status of job operation according to the device status and port monitor data on the job link.
- **Root cause analysis of job failure:** Based on the fault information, job log characteristics and storage capacity status of compute, storage and network resources on the job path during the job operation period, through full-link anomaly layer-by-layer diagnosis and multi-source data correlation analysis, the cause of task failure is automatically located and a diagnostic report is output.

3.12.4 Principle

Unified cluster management: By collecting the key performance data and log information of cluster resources in real time, the second level monitor of cluster resources was realized, and the multi-level alert notification was triggered based on the threshold rule to actively warn potential risks. The integrated view of node, storage,

network and task status is supported to help user control the overall operation situation of the cluster.

Accelerator card real-time monitor: It supports unified monitor and management of heterogeneous accelerator cards, supports monitor of accelerator card usage, GPU memory usage, temperature and power consumption at the second level, and can perform comprehensive analysis based on performance trend statistics to realize abnormal alert, which helps customers identify computing resource problems in advance.

Network intelligent topology: It can generate a cluster network topology relationship graph based on the physical and logical relationships of cluster resources, support automatic perception and update of topology changes, and support automatic marking of abnormal nodes and interrupted links, also support real-time monitor of end-to-end network link performance.

Root cause analysis of job failure: The job link graph is generated based on the compute, network and storage of job running association, and the full-link anomaly layer-by-layer diagnosis and multi-source data association analysis can be performed according to the resource failure information, log information and storage capacity information on the graph link, so as to locate the cause of job failure.

3.13 System Management

3.13.1 Definition

System management through a series of functions and technical means, users, logs, notifications, inspection, jobs, security and upgrades of the platform are comprehensively and efficiently managed. It ensures the stable operation of the platform, optimizes resource utilization, and guarantees data security.

3.13.2 Value

System management reduce manual operation and improve management efficiency through automated and integrated management functions. At the same time, it also provides a solid guarantee for data security, with strict security control and audit functions to ensure that the confidentiality, integrity and availability of system data are not infringed; In addition, system management can also achieve optimal allocation of resources through in-depth data analysis and monitor, and effectively reduce operating costs.

- **Precise data access control:** As the core of platform data permission control, scope management can precisely define users' access permissions to services and data.

- **Scalability and hierarchy:** Create sub-scopes based on the existing scope to achieve more detailed permission division and meet the permission management requirements of complex organizational structures and business scenarios.

3.13.3 Function

User Management

You can create, modify, delete, and assign rights to user accounts to ensure that user can access system resources based on their roles and rights.

Scope Management

Scope is the core mechanism for achieving data permission control on the platform. By flexibly configuring scope, it is possible to precisely define which user can serve which data. Scope supports both dynamic and static types, meeting the permission control requirements of different scenarios. Child scopes can be created based on the existing scope.

- **Dynamic:** By dynamically filtering the data access scope through data attributes and setting a set of condition rules to dynamically screen eligible data, applicable scenarios include resource data classification management and automatic synchronization of data additions/deletions.
- **Static:** By precisely controlling the data access scope through specified data entries, applicable scenarios include users being able to access fixed devices or data.

Log Management

Record and store user operation logs during the running of the system, and provide log query and download functions.

Notification Management

Send system notifications and alert information to user through email, SMS, wechat, Dingding, Slack, Feishu, PagerDuty, etc.

Task Hub

The task hub includes system operations and user operations, and recordsthe work periodically executed by the system and the user-defined work.

O&M Tools

The O&M Tools module provides professional tools including network connectivity testing, diagnostic tools, and routing diagnostics, enabling users to quickly identify and analyze network anomalies or connection failures between the platform and managed devices. Through visual diagnostics and automated detection, it achieves integrated

processing from problem discovery to root cause identification, thereby enhancing network operation efficiency and system stability.

System Self-Test

Periodically check the system comprehensively, including service inspection, module inspection, and resource inspection, and generate inspection reports. User can add and manage inspection tasks, preview and download inspection reports.

Online Upgrade

Support online system upgrade and patch installation to ensure system timeliness and stability. Provides upgrade record and rollback functions.

3.13.4 Principle

The principle of system management is to realize the comprehensive monitor and maintenance of the IT environment by integrating multiple management functions into KSManage. It covers key aspects such as user rights allocation, log recording and analysis, notification push, job scheduling, system health inspection, license management, and KVM video audit to ensure efficient and secure resource operation. At the same time, through system integration and other functions, to achieve seamless docking with other systems and automated services. Security controls and online upgrade mechanisms ensure system stability and data security, while flexible setting options meet individual needs, and together constitute a comprehensive system management framework.

3.14 Service Self-monitor

3.14.1 Definition

IOPS is a background management system that processes input/output operations (IOPS) per second on server devices. It covers functional modules such as overview, server list, database operation and maintenance, data collection, backup and restore, etc., for comprehensive monitor and device performance, ensuring efficient data reading and writing, secure storage, and fast recovery. IOPS provides real-time information on device status to optimize performance and ensure business continuity.

3.14.2 Value

As a comprehensive management tool, the value of the IOPS page lies in the comprehensive and in-depth focus on the performance and operation of storage devices. With the overview function, you can quickly learn about the overall status of the storage system. The server list provides detailed storage device information, facilitating

performance comparison and troubleshooting. The database operation and maintenance function ensure the efficient operation of the database, while the data collection function helps user collect and analyze the performance data of storage devices to provide strong support for optimization decisions. In addition, the backup and restore function ensures data security and integrity. In summary, the value of IOPS pages is to improve storage device performance and O&M efficiency through comprehensive function and data support.

3.14.3 Function

The IOPS page displays component status and operating status.

The server list is displayed. On this page, you can view the server's name, ip address, online status, cpu usage, memory usage, disk usage, CPU core, total memory capacity, and total disk capacity.

On this page, you can view metrics such as the database type, ip address, port, database name, SQL statement, and execution result.

Data collection screen allows you to select collected data, including logs, SQL records, model mapping, system log, ulimit and JVM crash log.

IOPS backup and restore. On the backup and Restore page, you can perform automatic backup, manual backup, and rollback operations. During the backup and restore process, infrastructure management platform system services are restarted.

3.14.4 Principle

On KSManage, IOPS records server details and processes, database operation and maintenance, data collection (including logs, SQL records, and model mappings, system log, ulimit and JVM crash log), and supports backup and restoration operations. In addition, IOPS enables user to view and manage background information on the IOPS page. Its principle is to centrally manage resources, optimize resource allocation, improve system response speed, and ensure data integrity and business continuity.

3.15 APP

3.15.1 Definition

APP is a comprehensive management tool that integrates asset management, information monitor and quick search.

3.15.2 Value

This environment is imported by scanning the QR code on the page of the cluster version APP. The APP provides user with comprehensive asset management and information acquisition services through four modules: "Home Page", "alert", "My" and "Search". The home page displays statistics and asset entry. The alert module tracks historical and current alerts in real time to ensure that problems are handled in a timely manner. My module easily manages account information, and the search function accurately locates devices and alerts, improves management efficiency, and ensures asset safety and business continuity.

3.15.3 Function

The APP page is divided into **Home**, **Alert**, **My** and **Search** modules. The home page supports viewing statistics and entering asset module, asset module asset list, asset basic information and hardware information. alert module real-time alerts, historical alerts, and alert details. My module views account information and logins about us and logout. The search module supports device and alert search.

3.15.4 Principle

After installing the corresponding APP, keep the mobile phone network consistent with the system network you want to access, and select the corresponding Wi-Fi for connection. After the network connection is complete, open the APP and enter the "Host list" page automatically after startup. Enter the IP address of the system you want to access and click the newly configured item on the list page to access the login page. On the login page, enter the account and password of the interconnection system, and click to log in to the system to view related information, such as asset information and alert information.

4 Deployment Options

4.1 Deployment Mode

Based on the module-oriented architecture, KSManage provides a variety of deployment scenarios based on the number of nodes being managed, service scenarios, and server resource configurations provided by customers.

4.1.1 Single-Node Deployment

Single machine deployment means that all functions of KSManage are deployed on the same server.

The single-machine deployment scheme is suitable for the scenario where the network scale is not large and the reliability requirement is not high.

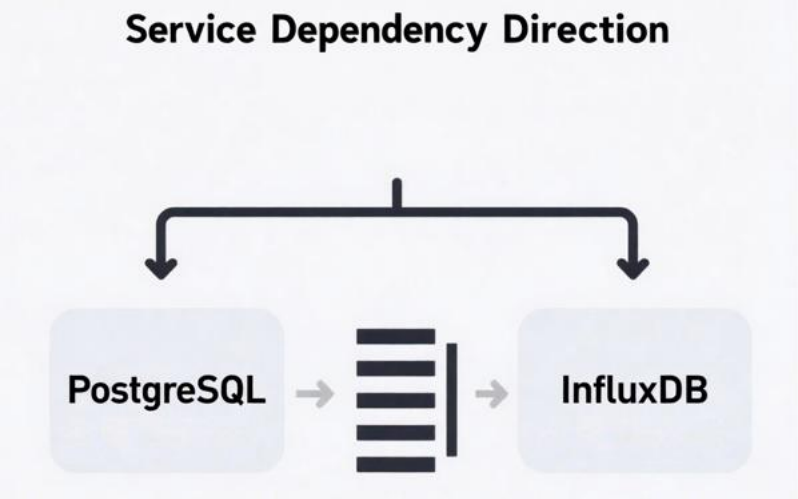

Figure 4- 1 single-node deployment diagram

4.2 Upgrade Method

- KSManage provides a version update package for one-click secure and reliable updates. After the update, you need to restart KSManage. It takes about 10 minutes
- KSManage provides an online update function that supports one-click download and one-click update, which is secure and convenient. After the update, you need to restart KSManage. It takes about 10 minutes.

5 Security

5.1 Network Constraints

KSManage has occupied ports 3306, 8086, 32314, 32315, 32316, 32317, 32318, 32319, 32320, 32321, 32322, 32323, 32324, 32325, 32326, 32327, 32229, 161, 162, and 623. During port planning, avoid using these ports on other KSManage-related devices.

Table 5- 1 Networking Constraints

Source Device	Source IP	Source Port	Destination Device	Destination IP Address	Destination (Listening) Port	Protocol	Port Description	Authentication Mode
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	9200	TCP/UDP	Port for Elasticsearch data storage service	User name and password
Elasticsearch node	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	9300	TCP/UDP	Port for communication between Elasticsearch nodes	User name and password
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	6379	TCP/UDP	Port for Redis data storage service	User name and password
Redis node	127.0.0.1	1 to 65535, randomly allocated by	Local device	127.0.0.1	26379	TCP/UDP	Port for Redis HA data synchronization	User name and password

		the socket					service	
RabbitMQ	127.0.0.1	1 to 65536, randomly allocated by the socket	Local device	127.0.0.1	5672	TCP/UDP	Port for RabbitMQ message queue service	User name and password
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	3306	TCP/UDP	PostgreSQ storage service port	User name and password
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	8086	TCP/UDP	Port for InfluxDB performance data storage service	User name and password
Local device	Any	1 to 65535, randomly allocated by the socket	Managed node	Any	161	UDP	Port used by the collector to obtain hardware data via SNMP	SNMPv1 and SNMPv2c use community strings; SNMPv3 uses USM-User/MD5 and SHA authentication

								passwords.
Node managed by KSManage	Any	1 to 65535, randomly allocated by the socket	Local device	Any	162	UDP	Port used by the collector to receive SNMP Trap alert messages	SNMPv1 and SNMPv2c use community strings; SNMPv3 uses USM-User/MD5 and SHA authentication passwords.
Local device	Any	1 to 65535, randomly allocated by the socket	Managed node	Any	623	UDP	Port used by the collector to obtain hardware data via IPMI protocol	User name and password
Node managed by KSManage	Any	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	514	UDP	Port used by the collector to receive the Syslog	

Any	Any	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	22	TCP	Port for SSH service	User name and password
Any	Any	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	9141	TCP	Port for northbound HTTPS interface service	User name and password
Any	Any	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	80	TCP	Port for accessing HTTP service on the node page (By default, it redirects to port 443)	User name and password
Any	Any	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device,	443	TCP	Port for accessing HTTPS service on the node page	User name and password

				which can be specified				
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	30100	TCP	Web-facade	User name and password
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32301	TCP		
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32310	TCP	Asset	
Node managed	127.0.0.1	1 to 65535, randomly	Local device	127.0.0.1	32311	TCP		

by KSManage		allocated by the socket						
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32326	TCP	Monitor	
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32327	TCP		
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32330	TCP	System	
Node managed by KSManag	127.0.0.1	1 to 65535, randomly allocated by	Local device	127.0.0.1	32331	TCP		

e		the socket						
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32340	TCP	Job-schedule	
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32341	TCP		
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32350	TCP	Control	
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32351	TCP		

Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32360	TCP	Collector-gateway	
Node managed by KSManage	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	32361	TCP		
Node managed by KSManage	127.0.0.1	1 to 65536, randomly allocated by the socket	Local device	127.0.0.1	32380	TCP	IOPS	
Any	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	32370	TCP	Collector-worker	

Any	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	32371	TCP		
Any	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	32372	TCP		
Any	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	32373	TCP		
Any	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device,	32374	TCP		

				which can be specified				
Any	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	32320	TCP		
Node managed by KSManage	127.0.0.1	1 to 65536, randomly allocated by the socket	Local device	127.0.0.1	32333	TCP	Logs	
Local device	Any	1 to 65535, randomly allocated by the socket	Remote device	Any	32390	TCP (HTTPS)	HingeClient	License and authentication certificate
Local device	127.0.0.1	1 to 65535, randomly	Local device	127.0.0.1	32391	TCP		

		allocated by the socket						
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	111	TCP/UDP	rpcbind	User name and password
Any	Any	1 to 65535, randomly allocated by the socket	Local device	Any IP address of the local device, which can be specified	9140	TCP	Port for northbound HTTP interface service	User name and password
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Remote device	Any	18081	TCP	CenterHub	User name and password
Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	20048	TCP/UDP	rpc.mount	User name and password

Local device	127.0.0.1	1 to 65535, randomly allocated by the socket	Local device	127.0.0.1	30100	TCP	momo	User name and password

5.2 System Security

KSManage uses secure and stable database version and other middleware versions for system security. OS security is vital to the security of KSManage. The OS where KSManage is deployed should adhere to the following principles for security reinforcement.

- Disable unused ports and services. According to the principle of least privilege, non-essential access channels, unused services, and unused TCP/UDP ports are disabled by default, such as prohibiting login via Telnet.
- Minimize file permissions and enhance security configurations of key system files and directories.
- Control accesses, such as disabling the remote access of the root account and adding the IP address blacklist and whitelist.
- Perform authentication and authorization, such as setting account and password security policies, login, and audit.
- Ensure secure access protocols, using secure access channels and securely configured SSH services.
- Ensure secure system settings by shielding the login banner information and disabling the system core dump to generate core files.
- Support vulnerability management to fix system vulnerabilities regularly and fix high-risk emergency response vulnerabilities in time.

5.3 Application Security

To protect the application security of KSManage, five types of security strategies are introduced: authentication and authorization, data protection, protocol security, session management, and log audit.

5.3.1 Authentication and Authorization

When you access KSManage via a web browser, the "user name + password" authentication method is used for local authentication. Strong password rules and password modification policies are used. A login failure lockout mechanism is provided to prevent password attacks and brute-force attacks.

- Strong password rules:
  - The password length is at least 8 characters.

- The password must contain uppercase/lowercase letters, numbers, and special characters.

- Password change strategies:

- Prompts the administrator to change the initial password.

- The old password needs to be verified before a user changes the password to a new one.

- In the interface, the password cannot be displayed in plaintext.

- The password and user name cannot be the same.

- The password is saved in ciphertext.

- Login failure lockout mechanism. Your account will be locked for 20 minutes if you enter the wrong password in 5 consecutive attempts. The administrator account can unlock locked accounts.

5.3.2 Data Protection

The most sensitive data in KSManage system, including system key parameter data, account information, and user privacy data, has been encrypted and protected using complex encryption algorithms such as SHA-256, HMAC-SHA-256, AES, and PBKDF2. Besides, functions like memset are used to cover or clear the unencrypted sensitive data generated during system operation and unencrypted sensitive data in heap, stack, and data segments.

5.3.3 Protocol Security

- Enables user to log in to KSManage via HTTPS (HTTP over SSL) protocol.

- Supports accessing devices via security protocols such as SNMPv3, HTTPS, and IPMI 2.0.

- Enables user to upload and download SFTP that supports secure encryption.

5.3.4 Session Management

- Uses token for session management.

- Supports anti-session fixation mechanism.

- Supports session timeout mechanism. If a session sits for 10 minutes, the session times out and exits, and the system clears session information.

- Provides user **Log out**/**Sign out** menus.

- The system will clear the session information after the user signs out.

5.3.5 Log Audit

- Enables auditing of security practice and operation logs.
- Log information contains the user's name, user IP address, operation time, and operation content.

5.4 Release Version Security

Code Security Scanning

Code security is the basis of system security. The release version of KSManage has been scanned by static code analyzers (Fortify and Coverity) and no high and medium vulnerabilities are found.

Security Scanning

The release version of KSManage has been scanned by vulnerability scanners (NESSUS, AppSca, and NSFOCUS) and no high and medium vulnerabilities have been found.

Version Security

The release version of KSManage comes with the hash value and digital signature. Before updating the product, version or installing a patch, you can check the product hash value or digital signature and verify the legitimacy of the software, thus avoiding unauthorized tampering or replacement of the software.

6 Reliability

6.1 Cluster Reliability

KSManage provides the 3-node cluster deployment. When cluster nodes are running normally, each node is in the active state. If one node fails, other nodes share the load of the faulty node and continue to provide services in a balanced manner.

6.1.1 Microservice Reliability

Each microservice has two or more instances deployed on three nodes. Each node processes services independently. When a single node or service instance fails, it automatically switches to another node.

6.1.2 Database Reliability

The PostgreSQL database is deployed on nodes 1, 2, and 3 of the cluster in master-slave replication mode to perform real-time data redundancy backup. If the instance of the master node or master database is faulty, it can be automatically switched to the slave node or database, while the original master node is degraded as the slave node.

6.2 Data Reliability

The backup and recovery function are an important guarantee to ensure that the system can quickly resume normal operation in case of system abnormality.

KSManage supports database backup and recovery. You can set the backup policy to automatic or manual according to system conditions. You can set the periodic backup period and backup path.

KSManage plans a high availability protection solution at the application layer to protect against unknown risks caused by hardware or software faults, ensuring the secure and stable operation of KSManage.

7 Configuration Requirements

KSManage can be installed on a virtual machine (VM) or a physical machine (PM). The configuration requirements for the server are shown in Table 7- 1.

Table 7- 1 KSManage Server Configuration Requirements

Item	Description
OS	CentOS7.9
CPU	Below 100 nodes: ≥2 cores Below 200 nodes: ≥4 cores Below 500 nodes: ≥8 cores Below 2,000 nodes: ≥16 cores
Memory	Below 100 nodes: ≥4 GB Below 200 nodes: ≥8 GB Below 500 nodes: ≥16 GB Below 2,000 nodes: ≥64 GB
Drive	≥500 GB Note: When the management scale is greater than 1,000 nodes, it is recommended to add 100 GB per 1,000 nodes.
NIC	≥1
IP	1 static IP address

NOTE

- Before deploying KSManage, deploy the CentOS 7.9 Minimal operating system first.
- After the OS is deployed, install the KSManage with the tar package.

A Getting Help

A.1 Collect Necessary Fault Information

You need to collect the necessary information before troubleshooting, including:

- Customer name and address
- Contact person and telephone number
- Time when the fault occurred
- Detailed fault description
- Device type and software version
- Measures already taken after the fault occurs and the related results
- Fault severity and expected troubleshooting deadline

A.2 How to Use Documents

KAYTUS provides comprehensive guidance documents and delivers such documents with the devices. These documents can help you address common problems encountered during routine maintenance or troubleshooting. To better solve problems, use the documents before you contact us for technical support.

A.3 Obtaining Technical Support

KAYTUS provides user with prompt and efficient technical support through local branch offices, technical guidance over the phone, remote technical support, and on-site technical support.

Our technical support system includes:

- Global service hotline: +1 800 611 8899 / +65 6611 8899
- Official website: https://www.kaytus.com/

B Terms and Abbreviations

Term	Definition
KSManage	Infrastructure Management Platform
BMC	Baseboard Management Controller
BIOS	Basic Input Output System
RAID	Redundant Arrays of Independent Drives
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
IPMI	Intelligent Platform Management Interface
SNMP	Simple Network Management Protocol