

智算集群基础设施运维通用技术要求

General Technical Requirements for Operation and Maintenance of
Intelligent Computing Cluster Infrastructure

V 1.0

目录

编者	3
1. 引言	4
2. 缩略语与术语	5
3. 运维参考架构	7
4. 关键运维组件功能要求	8
4.1. 硬件管理	8
4.1.1. 设备纳管	8
4.1.2. 配置管理	11
4.1.3. 性能管理	13
4.1.4. 告警管理	15
4.1.5. 日志管理	16
4.2. GPU 运维	17
4.2.1. GPU 配置管理	17
4.2.2. GPU 驱动管理	17
4.2.3. GPU 指标监控	18
4.2.4. GPU 健康检查	18
4.2.5. GPU 故障自愈	19
4.2.6. GPU 故障诊断	19
4.2.7. 性能压测	19
4.3. 智算集群运维	20
4.3.1. 平台资源监控	20
4.3.2. 性能管理	21
4.3.3. 告警管理	21
4.3.4. 日志管理	22
4.4. 网络运维	22
4.4.1. 拓扑可视化	22
4.4.2. 网络检测	23
4.4.3. 网络故障诊断	24
4.4.4. 网络调优	25
4.5. 作业运维	25
4.5.1. 作业监控	25
4.5.2. 作业可观测	25
4.5.3. 作业劣化分析	25
4.5.4. 作业故障诊断	26
4.5.5. 性能调优	26
5. 修订记录	27

编者

文档参编单位：

中兴通讯股份有限公司、中国信息通信研究院、阿里云计算有限公司、上海壁仞科技股份有限公司、上海天数智芯半导体股份有限公司、浪潮电子信息产业股份有限公司、上海寒武纪信息科技有限公司、昆仑芯（北京）科技股份有限公司、瀚博半导体（上海）股份有限公司（排名不分先后）

文档参编组织：

龙蜥社区智算联盟

1. 引言

随着“全国一体化算力网建设”、“算力互联互通行动计划”的稳步推进，2026 年“十五五”规划纲要对算力基础设施建设提出了更高的要求：“加快国家枢纽算力设施集群建设，论证建设超大规模智算集群，推动绿色电力与算力协同布局，降低中小企业用算成本。”

智算基础设施建设的飞速发展，集群规模的不断扩大，使得集群基础设施运维的痛点日益突显。软硬件协同断层、跨产品运维壁垒、智能运维技术应用碎片化、面向 AI 智能体的算力支撑功能缺失等问题，将会制约产业的高质量发展。

龙蜥社区智算联盟依托社区平台，联合智算硬件设备商、集群软件厂商、软件集成商，深入研究智算集群基础设施运维体系，分析具体应用场景，结合业界运维痛点，提出了针对智算集群软硬件基础设施的运维通用技术要求。该技术要求的编写旨在打通智算软硬件、跨主体运维断点，补齐 AI 智能体底层算力协同短板；为支撑智算集群基础设施稳定运行，提供运维体系架构、关键运维功能的标准化建议。

本文件提出了智算集群基础设施端到端运维的参考架构，定义了参考架构的设计原则。从智算项目运维视角，对硬件基础设施管理、GPU 运维、智算集群软件基础设施运维、网络运维、智算作业运维等关键组件的功能进行了定义。为智算集群软硬件产品生产商、系统集成商在系统设计、产品制造、运维集成方面提供指导，为第三方测评机构实施智算集群运维体系和功能的测试和评价提供参考。

2. 缩略语与术语

1. 缩略语

缩略语	英文	中文
BMC	Baseboard Management Controller	基板管理控制器
IPMI	Intelligent Platform Management Interface	智能平台管理接口
SNMP	Simple Network Management Protocol	简单网络管理协议
SMI-S	Storage Management Initiative - Specification	存储管理倡议规范
BIOS	Basic Input/Output System	基本输入输出系统
CDU	Coolant Distribution Unit	冷却液分配单元
IOPS	Input/Output Operations Per Second	每秒输入/输出操作次数
UUID	Universally Unique Identifier	通用唯一识别码
SN	Serial Number	序列号
VM	Virtual Machine	虚拟机
BM	Bare Metal	裸金属服务器
vcjob	Volcano Job	Volcano 调度作业
ECC	Error Correcting Code	错误纠正码
BAR1	Base Address Register 1	基地址寄存器 1
SM	Streaming Multiprocessor	流式多处理器
FP	Floating Point	浮点数
AER	Advanced Error Reporting	高级错误报告
RDMA	Remote Direct Memory Access	远程直接内存访问
RNIC	RDMA Network Interface Card	RDMA 网卡
OOM	Out Of Memory	内存溢出
CUDA	Compute Unified Device Architecture	统一计算设备架构
MTU	Maximum Transmission Unit	最大传输单元
ACL	Access Control List	访问控制列表
CCL	Collective Communication Library	集合通信库
RTT	Round-Trip Time	往返延迟
DP	Data Parallelism	数据并行
PP	Pipeline Parallelism	流水线并行
EP	Expert Parallelism	专家并行

2. 术语

Redfish: 由 DMTF 推出的一种现代、安全的 IT 基础设施管理标准，定义了如何使用 RESTful API 接口、JSON 数据格式来管理服务器等硬件设备。

Telemetry: 一种流式遥测技术，支持从远程设备自动、持续地收集、传输和监控数据。

Syslog: 一种通用的日志记录标准，在 RFC5424/3164 标准文档中定义，提供了独立于应用程序、与操作系统无关的日志记录方式。

Node: Kubernetes 承载容器运行时的实体对象，可以是虚拟机或者物理服务器。

POD: Kubernetes 管理的最小计算单元，由一组容器构成。

StatefulSet: Kubernetes 管理有状态应用工作负载的内置 API。

Deployment: Kubernetes 管理应用工作负载的 Pod 资源，适用于无需保持状态的负载。

Host: Openstack

Leaf 交换机: Leaf-Spine 网络架构的接入层交换机，负责设备网络接入，以及 Spine 交换机的网络互联。

Spine 交换机: Leaf-Spine 网络架构的汇聚/核心层交换机，负责接入 Leaf 交换机，实现 Leaf 交换机之间的网络互联。

NVLink: NVIDIA 专有的高速互联协议，用于解决多处理器（GPU 与 GPU、CPU 与 GPU）之间的高带宽、低延迟通信问题。

XID: GPU 驱动内部定义的错误事件 ID，当 GPU 发生底层硬件故障、显存错误、驱动崩溃或固件异常时，内核驱动会在系统日志中打印 XID 错误代码。

Tensor: 张量是一个多维数组，人工智能领域数据处理的最基础架构。Tensor Core 是专门用于加速矩阵乘法和深度学习计算的硬件架构。

Profiler: 性能剖析器，在 AI 模型训练或系统开发中是一种用于测量和可视化代码执行过程的工具。能精确记录每个函数、每个 GPU 算子的执行时间、内存分配情况、GPU 利用率以及 PCIe/NVLink 的数据传输耗时。

Pipeline: AI 和高性能计算的流水线。

Fullmesh: 全互联拓扑，网络中的每一个节点都与其他所有节点直接相连，高性能 AI 服务器的标准设计。

All-Reduce: CCL 集合通信库原语“全局归约”，分布式训练的核心通信操作，对归属不同 GPU 卡的梯度数据进行指定数据运算，最终生成全局聚合最新梯度数据。

All-Gather: CCL 集合通信库原语“全局搜集”，提供数据分片的拼接功能。

3. 运维参考架构

1. 面向用户、端到端、可异构、可扩展的智算集群基础设施运维体系，需满足如下架构设计要求：

- (1) 分层、解耦

- 1) 运维能力按对象分层定义，横向聚焦基础数据的采集和存储。
- 2) 运维组件以用户视角设计，功能内聚，组件间松耦合，可分可合。每个运维组件可作为独立工具使用，也可组合为统一运维平台。

- (2) 可扩展

运维组件可横向扩展，覆盖智算项目建设全场景，智算业务全流程。

- (3) 云原生、微服务化

运维组件支持微服务化、云原生部署。

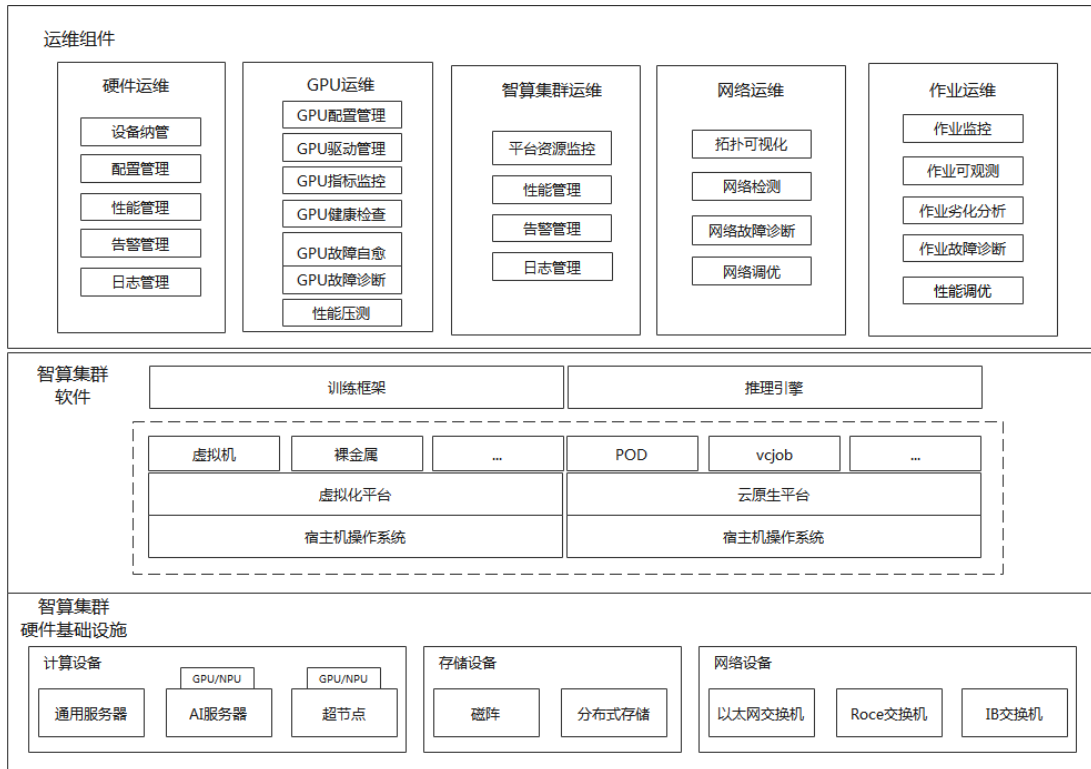
- (4) 设计目标

- 1) 构建面向用户的、端到端的运维体系标准。
- 2) 为用户按需提供场景化、可组装的运维解决方案。

2. 智算集群基础设施运维的建议参考架构如图 1 所示，具体功能包括：

- (1) 硬件管理：负责智算硬件基础设施统一管理与运维，支持管理通用服务器、AI 服务器、超节点等计算资源；支持管理磁阵、分布式（块/文件/对象）存储等存储资源；支持管理以太网交换机、Roce 交换机、IB 交换机等网络资源。
- (2) GPU 运维：负责 GPU 配置规格查询，温度、功耗指标采集等基础运维功能；支持健康检查，故障诊断、性能压测等增加运维能力。
- (3) 智算集群运维：负责虚拟化、云原生智算集群软件平台运维管理，支持对虚拟化、云原生资源对象的资源、性能、告警等基础运维功能。
- (4) 网络运维：负责智算软硬件基础设施多层次、多平面网络的运维，支持网络拓扑可视化、网络故障诊断、网络检测与调优等运维功能。
- (5) 作业运维：对训推任务关联模型、算子、网络以及云原生工作负载监控数据跨域整合，从业务视角跨层联动软硬件基础设施，实现端到端运维。

图 1. 智算集群基础设施运维参考架构



4. 关键运维组件功能要求

4.1. 硬件管理

4.1.1. 设备纳管

应支持对通用服务器、AI 服务器、超节点计算设备，磁阵、分布式存储设备，以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备的统一设备接入和硬件纳管功能。

1. 服务器纳管

- (1) 应支持通过服务器产品提供的带外管理接口接入、管理服务器，支持对服务器的添加、删除、设备信息更新等硬件生命周期管理操作。
- (2) 应支持对接通用的 Redfish 或 IPMI、SNMP 服务器带外管理接口。
- (3) 应支持批量添加、删除服务器硬件，批量更新服务器设备信息。

- (4) 应支持通过服务器名称对已纳管服务器设备进行搜索。
- (5) 应支持导出服务器的接入配置信息，内容包括：设备名称、设备型号、管理 IP 地址、设备物理位置信息；通过 Redfish 接口接入时，需包括服务器 Redfish 接口用户名及密码；通过 IPMI、SNMP 接口接入时，需包括服务器 BMC 用户及密码，SNMP 协议类型、用户名、认证等信息。
- (6) 宜支持服务器设备发现功能，通过配置发现规则，选择 Redfish 或者 SNMP 发现协议，指定 IP 地址范围，执行设备发现操作；支持对已发现服务器进行设备纳管。
- (7) 宜支持针对服务器设备的自动发现功能，通过配置发现规则，设置扫描周期，自动执行设备发现操作。

2. 超节点纳管

- (1) 应支持通过超节点产品提供的带外管理接口接入、管理超节点及超节点内的计算节点和交换节点，支持对超节点及其内的计算节点、交换节点的添加、删除、设备信息更新等硬件生命周期管理操作。
- (2) 应支持对接通用的 Redfish 或 IPMI、SNMP 超节点带外管理接口。
- (3) 应支持批量添加、删除超节点及其内部的计算节点、交换节点硬件，批量更新超节点及其内部的计算节点、交换节点信息。
- (4) 应支持导出超节点的接入配置信息，内容包括：设备名称、设备型号、管理 IP 地址、设备物理位置信息；通过 Redfish 接口接入时，需包括超节点 Redfish 接口用户名及密码；通过 IPMI、SNMP 接口接入时，需包括超节点 BMC 用户及密码，SNMP 协议类型、用户名、认证等信息。

3. 磁阵纳管

- (1) 应支持通过磁阵产品提供的带外管理接口接入、管理磁阵，支持对磁阵的添加、删除、设备信息更新等硬件生命周期管理操作。
- (2) 应支持对接通用的 SMI-S 或 SNMP 磁阵管理接口。
- (3) 应支持批量添加、删除磁阵，批量更新磁阵设备信息。
- (4) 应支持导出磁阵的接入配置信息，内容包括：设备名称、设备型号、管理 IP 地址、设备物理位置信息；通过 SMI-S 接口接入时，需包括磁阵的 SMI-S 用户名及密码；

通过 SNMP 接口接入时，需包括 SNMP 协议类型、用户名、认证等信息。

- (5) 宜支持磁阵设备发现功能，通过配置发现规则，通过 SNMP 发现协议，指定 IP 地址范围，执行设备发现操作；支持对已发现磁阵进行设备纳管。
- (6) 宜支持针对磁阵的自动发现功能，通过配置发现规则，设置扫描周期，自动执行设备发现操作。

4. 分布式存储纳管

分布式存储产品由存储软件系统和服务器硬件组成，存储服务器硬件纳管要求参考“服务器纳管”方式，针对分布式存储软件系统的纳管功能要求如下：

- (1) 应支持通过分布式存储产品提供管理接口接入、管理分布式存储软件，支持添加、删除、存储软件信息更新等生命周期管理操作。
- (2) 应支持导出分布式存储软件接入配置信息。

5. 网络设备纳管

- (1) 应支持通过 SNMP 接口接入、管理以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备，支持对添加、删除、设备信息更新等针对网络设备的硬件生命周期管理操作。
- (2) 应支持批量添加、删除网络设备，批量更新网络硬件设备信息。
- (3) 应支持通过网络设备名称对已纳管设备进行搜索。
- (4) 应支持导出网络设备的接入配置信息，内容包括：设备名称、设备型号、管理 IP 地址、设备物理位置信息、SNMP 协议类型、用户名、认证等信息。
- (5) 宜支持网络设备发现功能，通过配置发现规则，选择 SNMP 发现协议，指定 IP 地址范围，执行设备发现操作；支持对已发现网络设备进行设备纳管。
- (6) 宜支持针对网络设备的自动发现功能，通过配置发现规则，设置扫描周期，自动执行设备发现操作。
- (7) 宜支持基于 Telemetry 技术栈的订阅与推送机制对接网络设备，构建小粒度指标数据采集通道。

4.1.2. 配置管理

应支持对通用服务器、AI 服务器、超节点计算设备，磁阵、分布式存储设备，以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备的硬件配置信息采集，统一配置管理。

1. 服务器配置管理

- (1) 应支持通过服务器带外管理接口采集服务器关键部件配置信息，包括：CPU、内存、硬盘、网卡、RAID 卡、电源、风扇、GPU 等。
- (2) 应支持采集 CPU 的厂商、型号、序列号、架构、核数、线程数、频率等配置信息。
- (3) 应支持采集内存的厂商、型号、序列号、类型、容量、频率等配置信息。
- (4) 应支持采集硬盘的厂商、型号、序列号、接口类型、介质类型、容量、速率等配置信息。
- (5) 应支持采集网卡的厂商、型号、序列号、接口类型、网口数量、网口 MAC 地址、网口速率等配置信息。
- (6) 应支持采集 RAID 卡的厂商、型号、序列号、接口类型、接口速率、内存容量等配置信息。
- (7) 应支持采集电源的厂商、型号、序列号、额定功率等配置信息。
- (8) 应支持采集风扇的厂商、型号、序列号、转速等配置信息。
- (9) 应支持采集 GPU 的厂商、型号、序列号、显存类型、显存容量、功率等配置信息。
- (10) 应支持服务器关键部件配置信息查询与搜索。
- (11) 应支持服务器关键部件的配置信息导出。
- (12) 应支持对服务器的 BIOS 配置操作。
- (13) 应支持对服务器的上电、下电、重启操作。

2. 超节点配置管理

- (1) 应支持通过超节点带外管理接口采集超节点、及其内部的计算节点、交换节点关键部件配置信息。
- (2) 应支持采集超节点的厂商、型号、序列号、固件版本、电源信息，以及 CDU 的厂商、型号、序列号等信息。

- (3) 应支持采集计算节点的 CPU、内存、硬盘、网卡、RAID 卡、GPU 等关键部件信息，具体包括：

CPU 的厂商、型号、序列号、架构、核数、线程数、频率等配置信息。

内存的厂商、型号、序列号、类型、容量、频率等配置信息。

硬盘的厂商、型号、序列号、接口类型、介质类型、容量、固件版本等配置信息。

网卡的厂商、型号、序列号、接口类型、网口数量、网口 MAC 地址、网口速率等配置信息。

RAID 卡的厂商、型号、固件版本等配置信息。

GPU 的厂商、型号、序列号、固件版本、主频、显存容量、额定功率等配置信息。

- (4) 应支持采集交换节点的序列号、端口名称、速率、MAC、光模块的厂商、型号、序列号信息。

- (5) 应支持对超节点的上电、下电、重启操作。

3. 磁阵配置管理

- (1) 应支持通过磁阵管理接口采集磁阵关键部件配置信息，包括：控制器、硬盘框、网口、硬盘、电源等。

- (2) 应支持采集控制器的厂商、型号、序列号、磁盘容量、Cache 容量等配置信息。

- (3) 应支持采集硬盘框的序列号、类型、磁盘容量等配置信息。

- (4) 应支持采集磁阵网口的序列号、接口类型、网口 MAC 地址、网口速率等配置信息。

- (5) 应支持采集磁阵硬盘的厂商、型号、序列号、介质类型、容量等配置信息。

- (6) 应支持采集磁阵电源的额定功率等配置信息。

4. 分布式存储配置管理

应支持通过分布式存储管理接口采集关键部件配置信息，包括：分布式存储软件、存储服务器硬件。

- (1) 应支持采集分布式存储软件的厂商、型号、版本号、集群流控、数据重构等配置信息。

- (2) 应支持采集存储服务器硬件的关键部件配置信息，具体参考“服务器配置管理”要求。

5. 网络设备配置管理

应支持通过 SNMP 接口采集以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备配置信息。

- (1) 应支持采集以太网交换机的厂商、型号、序列号、网络吞吐量、包转发量、端口带宽、额定功率等配置信息。
- (2) 应支持采集 Roce 交换机的厂商、型号、序列号、网络吞吐量、包转发量、端口带宽、额定功率等配置信息。
- (3) 应支持采集 IB 交换机的厂商、型号、序列号、网络吞吐量、包转发量、端口带宽、额定功率等配置信息。
- (4) 应支持采集路由器的厂商、型号、序列号、网络吞吐量、包转发量、端口带宽、额定功率等配置信息。

4.1.3. 性能管理

应支持对通用服务器、AI 服务器、超节点计算设备，磁阵、分布式存储设备，以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备的硬件性能指标采集，性能数据的统一管理。

1. 服务器性能管理

- (1) 应支持通过服务器带外管理接口，周期性采集服务器关键部件性能指标。
- (2) 应支持采集服务器的温度指标，包括：GPU 温度、CPU 温度、环境温度等。
- (3) 应支持采集服务器的功耗指标，包括：GPU 功耗、服务器当前功率等。
- (4) 应支持采集服务器的风扇指标，包括风扇转速百分比等。
- (5) 应支持通过统一性能管理界面查询服务器历史性能指标数据，支持性能数据的图形化展示。
- (6) 应支持服务器性能指标数据导出操作。

2. 超节点性能管理

- (1) 应支持通过超节点管理接口，周期性采集超节点、及其内部的计算节点、交换节点的关键部件性能指标。

- (2) 应支持采集超节点性能指标包括：当前功率、电源输出电压、电流，CDU 的二次侧供液温度、回液温度、流量、供液压力、回液压力等。
- (3) 应支持采集计算节点性能指标包括：GPU 温度、CPU 温度、环境温度，GPU 功耗、服务器实际功率等。
- (4) 应支持采集交换节点的设备、端口、光模块的性能指标数据。
设备指标包括：CPU 使用率、内存使用率、温度、实际功率等。
端口指标包括：端口的流入速率、流出速率、流入丢包数、流出丢包数、流入错包数、流出错包数、流入字节数、流出字节数等。
光模块指标包括：光模块的温度、电压、接收功率、发射功率、偏置电流等。
- (5) 应支持通过统一性能管理界面查询超节点的历史性能指标数据，支持性能数据的图形化展示。
- (6) 应支持超节点性能指标数据导出操作。

3. 磁阵性能管理

- (1) 应支持通过磁阵管理接口，周期性采集磁阵关键部件性能指标。
- (2) 应支持采集磁阵的温度指标。
- (3) 应支持采集磁阵的容量指标，包括：磁阵总容量、磁阵已使用容量、存储池总容量、存储池剩余容量等。
- (4) 应支持采集磁阵的 IO 指标，包括存储池的读 IOPS、写 IOPS、读取速率、写入速率，卷的读 IOPS、写 IOPS、读取速率、写入速率等。
- (5) 应支持通过统一性能管理界面查询磁阵历史性能指标数据，支持性能数据的图形化展示。
- (6) 应支持磁阵性能指标数据导出操作。

4. 分布式存储性能管理

- (1) 应支持通过分布式存储管理接口，周期性采集分布式存储软件对象的性能指标，对象包括：分布式存储集群、分布式存储池、分布式存储节点。
- (2) 应支持采集分布式存储集群的性能指标，包括：集群已用容量、集群容量利用率、集群已分配容量、集群容量分配率等。

- (3) 应支持采集分布式存储池的性能指标，包括存储池已用容量、存储池容量利用率、存储池已分配容量、存储池容量分配率，存储池已使用缓存、存储池缓存利用率、存储池缓存命中率，存储池读速率、存储池写速率、存储池平均 IO 时延等。
- (4) 应支持采集分布式存储节点的性能指标，包括：节点 CPU 利用率、内存利用率，网口收发速率，节点文件系统已使用容量、节点文件系统使用率等。
- (5) 应支持通过统一性能管理界面查询分布式存储历史性能指标数据，支持性能数据的图形化展示。
- (6) 应支持分布式存储性能指标数据导出操作。

5. 网络设备性能管理

- (1) 应支持通过 SNMP 接口，周期性采集以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备的性能指标。
- (2) 应支持采集设备的性能指标，包括：CPU 使用率、内存使用率、温度、实际功率等指标。
- (3) 应支持采集端口的性能指标，包括：端口的流入速率、流出速率、流入丢包数、流出丢包数、流入错包数、流出错包数、流入字节数、流出字节数等。
- (4) 应支持采集光模块的性能指标，包括：光模块的温度、电压、接收功率、发射功率、偏置电流等。
- (5) 应支持通过统一性能管理界面查询网络设备历史性能指标数据，支持性能数据的图形化展示。
- (6) 应支持网络设备性能指标数据导出操作。

4.1.4. 告警管理

应支持对通用服务器、AI 服务器、超节点计算设备，磁阵、分布式存储设备，以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备的硬件告警信息实时采集，告警数据的统一管理。

- 1. 应支持通过服务器提供的 Redfish 或者 SNMP 告警接口，实时接收服务器上报的硬件告警。

2. 应支持通过超节点提供的 Redfish 或者 SNMP 告警接口，实时接收超节点上报的硬件告警。
3. 应支持通过磁阵提供的 SNMP 等告警接口，实时接收磁阵上报的设备告警。
4. 应支持通过分布式存储软件提供的 SNMP 等告警接口，实时接收分布式存储软件上报的软件告警；支持通过分布式存储节点提供的 Redfish 或者 SNMP 告警接口，实时接收存储节点上报的服务器硬件告警。
5. 应支持通过以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备提供的 SNMP 告警接口，实时接收网络设备上报的设备告警。
6. 应支持集中存储上述设备的实时告警和历史告警；提供统一告警查询页面，支持浏览或者按指定过滤条件查看告警，过滤条件包括：告警级别、告警类型、对象类型、确认状态、影响类型、发生时间等。支持根据告警描述的模糊匹配。
7. 应支持实时告警、历史告警的全量、或者部分导出功能，导出文件格式包括：EXCEL、CSV 等。
8. 宜支持通过设备提供的告警查询接口，按需对设备发起实时告警补采操作。补采操作可自定义时间段、补采对象。
9. 宜支持设备新增告警、变更告警快速适配功能，通过告警管理界面导入设备新增告警、变更告警完成适配操作。变更的告警信息包括：告警标题、告警级别、处理建议等。

4.1.5. 日志管理

应支持对通用服务器、AI 服务器、超节点计算设备，磁阵、分布式存储设备，以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备的硬件日志采集，日志数据的统一管理。

1. 应支持通过服务器提供的 Syslog 日志接口，实时接收服务器上报的日志数据。
2. 应支持通过超节点提供的 Syslog 日志接口，实时接收超节点上报的日志数据。
3. 应支持通过超节点提供的 Syslog 日志接口，实时接收超节点上报的日志数据。
4. 应支持通过磁阵提供的日志接口，采集磁阵的日志数据。
5. 应支持通过分布式存储提供的日志接口，采集分布式存储的日志数据。
6. 应支持通过以太网交换机、Roce 交换机、IB 交换机，路由器等网络设备提供的 Syslog

日志接口，实时接收网络设备上报的日志数据。

7. 应支持对上述设备日志数据的分类管理和集中存储，支持通过统一的日志管理页面筛选与搜索日志。

支持按照设备类型、日志类型筛选日志。

支持选择不同匹配规则搜索原始日志，匹配规则包括：日志文本全字符匹配、通配符匹配、多分词匹配等，支持对搜索结果按时间排序。

8. 应支持将查询日志结果或全量日志内容导出到本地。
9. 应支持提供日志采集配置开关，开启或者关闭对设备日志的采集和接收功能。

4.2. GPU 运维

4.2.1. GPU 配置管理

1. 应支持获取 GPU 的基本配置信息，包括 ID、UUID、SN、VendorName、ModelName、显存大小、驱动版本、主频、健康状态等。
2. 应支持设置 GPU 的 ECC Mode，最大功率等配置。
3. 应支持设置 GPU 的温度告警门限、功能告警门限等配置。

4.2.2. GPU 驱动管理

1. 支持查看当前系统内每个节点的驱动安装信息，版本信息
2. 支持单节点的驱动安装、升级、回退、卸载，支持不同节点安装不同驱动版本
3. 支持多节点批量安装、升级、回退和卸载
4. 支持驱动的容器化管理
5. 支持检测驱动健康状态，异常后自动恢复
6. 支持驱动安装升级时自动驱逐业务

4.2.3. GPU 指标监控

1. 应支持对 GPU 设备的性能指标进行采集，总体覆盖：
 - (1) 温度监控：采集 GPU 核心温度、显存温度及温度保护阈值，实现过温预警。
 - (2) 内存管理：覆盖显存容量/用量、BAR1 空间、ECC 退役页等，掌握显存使用与硬件退化。
 - (3) 计算利用率：从宏观 GPU 利用率到 SM/Tensor/FP 管线活跃率，定位算力瓶颈。
 - (4) 功耗：实时功耗、功耗上限、累计能耗监控，防止超出供电散热能力。
 - (5) 时钟频率：SM 核心时钟频率及降频原因，排查节流导致的性能下降。
 - (6) PCIe 接口：PCIe 链路代际、带宽、收发吞吐量及重传，发现总线瓶颈。
 - (7) P2P 网络互连：GPU 间 P2P 通道（xLink）的吞吐量和错误计数，多卡通信核心指标。
 - (8) 错误与健康：健康状态及 ECC、PCIe、NVLink、AER、XID 等硬件错误，可靠性运维核心。
 - (9) 违规/节流统计：因功率、温度、板级限制被节流的累积时长，量化性能损失来源。
 - (10) 容器/Pod 级指标：部分指标提供容器维度，实现多租户精细化资源观测。
2. 应支持对接上层系统将 GPU 指标进行上报和持久化存储。

4.2.4. GPU 健康检查

1. 应支持对特定的 GPU 设备执行健康检查（GPU 健康检查范围包含：GPU 设备 XID 故障、GPU 设备不在位、驱动无法访问 GPU 设备以及 GPU 设备发生未知原因故障）。
2. 应支持对多个 GPU 设备并行执行健康检查。
3. 应支持按照配置参数对 GPU 设备执行健康检查。
4. 应支持对 GPU 设备健康检查配置参数的修改。
5. 应支持对接上层系统将 GPU 设备健康检查结果进行上报和持久化存储。

4.2.5. GPU 故障自愈

1. 应支持对故障的 GPU 设备执行自愈动作（自愈动作至少包含：GPU 复位、GPU 所在节点操作系统复位）。
2. 应支持对多个故障的 GPU 设备并行执行自愈动作。
3. 应支持对根据 GPU 故障原因分别配置自愈动作（包括关闭自愈）。
4. 应支持对 GPU 故障自愈动作在线修改配置。
5. 应支持按照配置的自愈动作对故障的 GPU 进行自愈。
6. 应支持对接上层系统将故障 GPU 设备执行自愈的结果上报。

4.2.6. GPU 故障诊断

1. 应支持对特定的 GPU 设备执行诊断。
2. 应支持对多个 GPU 设备创建逻辑组执行批量诊断。
3. 应支持按照预置的诊断级别对 GPU 进行诊断，每个诊断级别包含特定的诊断项使用默认的参数进行诊断。
4. 应支持按照指定的诊断项对 GPU 进行诊断，支持自定义诊断参数。
5. 应支持对正在执行中的诊断操作进行取消。
6. 应支持对 GPU 的存在性检测、基本信息收集（型号、驱动、ECC 状态）、快速硬件连通性检测等诊断。
7. 应支持 PCIe 带宽检测、ECC 错误检查、内存分配与拷贝测试、温度、功耗、时钟检测。
8. 应支持内存带宽压测、计算密集任务压测、高温运行压测诊断。
9. 应支持 GPU 目标功率测试、诊断性内存测试、计算核心压力测试、链路压测试等。

4.2.7. 性能压测

智算系统的性能检测功能应至少具备以下功能：

4.2.7.1. 计算性能压测

1. 应具备对 GPU 的算力、显存带宽、P2P 带宽、PCIe 带宽等进行专项压测的能力。
2. 应支持不同强度（如快速、标准、深度）的压测策略。
3. 宜支持超节点场景下的 GPU 算力与带宽压测。

4.2.7.2. 网络性能测试

1. 应具备对参数面网络进行节点内、节点间及集群级带宽测试的能力。
2. 应具备对集合通信带宽进行测试的能力。
3. 应支持集合通信 AllReduce、Broadcast 等多种模式的性能测试。
4. 应支持 RDMA 时延测试。
5. 宜支持超节点场景下的网络带宽与集合通信测试。
6. 宜支持超节点场景下的 RDMA 时延测试。
7. 宜支持超节点 P2P 打流压测。

4.2.7.3. 训练与推理模型测试

1. 应具备调度至指定节点执行典型训练模型性能测试的能力。
2. 应具备调度至指定节点执行典型推理模型性能测试的能力。
3. 宜支持超节点场景下的训练与推理模型性能测试。

4.3. 智算集群运维

4.3.1. 平台资源监控

1. 应支持云原生资源对象监控，资源对象包括但不限于 Node、POD、vcjob、StatefulSet/deployment 等。
2. 应支持 Openstack 虚拟化资源对象监控，资源对象包括 Host、VM、BM 等。
3. 应支持云原生资源对象、虚拟化资源对象的资源拓扑管理、关联关系展示。

4. 应支持统计云原生资源对象、虚拟化资源对象的资源分配和使用情况。

4.3.2. 性能管理

1. 应支持对上报到性能管理的指标的持久化功能。
2. 应支持对上报到性能管理的指标查询的功能。
3. 应支持对上报到性能管理的指标数据导出的功能。
4. 应支持对上报到性能管理的指标按照空间或时间的汇聚功能。
5. 应支持对上报到性能管理的指标按照空间和时间规则的清理功能。

4.3.3. 告警管理

应支持智算集群平台虚拟资源对象告警管理功能。

1. 应支持浏览或者按指定条件查看当前告警，查询过滤条件支持告警级别、告警类型、对象类型、确认状态、影响类型、发生时间。并支持告警描述的模糊匹配。
2. 应支持浏览或者按指定条件查看已经恢复的告警，并设置历史告警存储日期；查询过滤条件支持告警级别、告警类型、对象类型、影响类型、告警时间。
3. 应支持监控本系统告警并在界面呈现。
4. 应支持告警确认功能，通过是否已确认的告警状态，很容易区分已处理告警和未处理告警。
5. 应支持告警反确认功能，对于已被确认的告警，提供反确认操作，可以通过反确认操作将告警返回到未确认状态。
6. 应支持告警清除功能，被清除的告警转入历史告警库，在当前告警中不再体现。
7. 应支持告警导出功能，提供将选择的当前告警、历史告警以部分或者全部方式以 EXCEL 或 CSV 文件方式导出。
8. 应支持告警屏蔽功能，设定此告警屏蔽的策略，通过对日期、时段、具体告警描述进行屏蔽规则设定，达到对符合条件的告警进行屏蔽的目的。
9. 宜支持告警补采：支持按自定义时间段以及按对象进行告警补采。
10. 宜支持告警北向上报：可以上报当前告警到上级网管，通过 SNMP 北向接口。
11. 宜支持告警通知功能，支持通过邮件、短信方式通知到指定人员。

12. 宜支持 QoS 静态阈值告警，当指标超出用户预定义的门限范围时，系统会产生告警。

4.3.4. 日志管理

1. 应支持智算集群软件平台的操作日志、安全日志、系统日志的采集和集中管理。
2. 应支持设置日志清理策略，系统按照清理策略，自动清除历史日志。
3. 应支持日志搜索功能，提供原始日志文本全字符匹配搜索、通配符搜索、多分词搜索、指定搜索字段及按时间和相关性排序。
4. 应支持日志导出功能，将查询到的日志或全部日志导出到本地。
5. 应支持日志备份功能，支持设置日志备份启用时刻、备份周期、备份文件大小等配置。
6. 宜支持日志北向接口，对接第三方日志管理系统。

4.4. 网络运维

4.4.1. 拓扑可视化

1. 全网拓扑呈现
 - (1) 应支持以智算集群基础设施全局视角展示网络拓扑，包括：管理面网络、业务面网络、普通存储面网络。
 - (2) 应支持呈现智算集群特有的 RDMA 参数面网络、存储面网络拓扑，拓扑需覆盖：服务器网口、Leaf 交换机端口、Spine 交换机端口、路由器端口之间的连接关系。
 - (3) 应支持在全局网络拓扑中，展示故障设备或设备端口，支持查看故障设备关联告警。
 - (4) 应支持在全局网络拓扑中，实现端口级流量可视，支持标识拥塞、异常的网络路径。
 - (5) 应支持展示 GPU 服务器节点机内 GPU/CPU/RNIC 硬件拓扑，拓扑需覆盖 GPU 与 RNIC 之间的 PCIe bus/switch 连接，GPU 与 GPU 的 PCIe 总线连接、或者 GPU 自有高速互联通道连接等。
2. 作业全路径拓扑可视

- (1) 应支持以训推作业视角，呈现整个作业网络拓扑，拓扑需覆盖：Pod、容器节点、Leaf 交换机、Spine 交换机等之间的连接关系。
- (2) 应支持在作业网络拓扑上展示流路径，支持查看流路径上的设备状态及告警。
- (3) 应支持展示训练作业网络“流”(通信五元组)级流量，并通过曲线图呈现跨机 DP/PP/EP 等流量/RTT 时延数据。
- (4) 应支持 GPU 服务器节点内作业流量数据可视化，支持监控作业 Pod 使用的 GPU 与 RNIC 之间，GPU 与 GPU 之间 PCIe bus/NVlink 的实时流量数据。支持以曲线图形式展示流量数据，查看异常机内流量。

4.4.2. 网络检测

1. 网络配置检查

- (1) 应支持网络设备端口、服务器网口等 MTU 一致性检查。
- (2) 应支持服务器网口 IP 地址与子网配置一致性检查。
- (3) 应支持路由策略配置一致性检查。
- (4) 应支持 RoCE 无损网络 PFC/DSCP 一致性检查。
- (5) 应支持网卡流控算法配置一致性检查。

2. 网络性能压测

- (1) 应支持指定服务器网口进行 RoCE 带宽/时延压测，支持选择网络大包、或者网络小包进行压测。
- (2) 应支持指定多台服务器进行 Fullmesh RoCE 带宽/时延压测，支持选择网络大包、或者网络小包进行压测。
- (3) 应支持支持构造多对一场景拥塞控制压测。
- (4) 应支持 CCL 集合通讯 allreduce、allgather 等通讯原语压测。

3. 网络连通性检测

- (1) 应支持指定服务器网口进行 RoCE 连通性检测。
- (2) 应支持指定多台服务器进行 Fullmesh 连通性检测。

- (3) 应支持智算集群运行时实时连通性检测，连通检测异常支持上报告警。

4. 实时网络亚健康检测

- (1) 应支持训推作业运行时带内遥测功能，可采集网络流数据每一跳的网络运维指标，进行实时网络健康分析。遥测功能支持配置开关，可按需开启或者关闭。
- (2) 应支持按需采集训推作业运行时的毫秒级、亚毫秒级网络流量/时延等指标，支持网络流量特征在线分析，识别网络异常。

4.4.3. 网络故障诊断

1. 网络连通性诊断

- (1) 应支持指定 GPU 服务器网口作为源目的端口，进行端到端 RoCE 网络连通性诊断、网络故障定界和定位。
- (2) 宜支持定位、识别 GPU 服务器网口 MTU 配置错误、交换机 ACL 配置错误、交换机端口严重丢包等故障场景。
- (3) 应支持抓包分析功能，快速定位故障点。

2. 网络亚健康诊断

- (1) 应支持基于毫秒/亚毫秒级流量采集指标，在线分析检测慢网络问题。
- (2) 宜支持定位、识别网络拥塞导致的网络性能下降问题，服务器网卡/交换机转发异常/丢包，光模块污损等导致网络性能下降问题。
- (3) 宜支持定位、识别服务器 GPU/CPU 处理异常，导致的网络性能下降问题，服务器机内 PCIe 总线降 Lane 导致的网络性能下降问题。

4.4.4. 网络调优

1. 网络流量路径规划

- (1) 宜支持全局流量路径规划，避免网络拥塞，最大化提高网络路径带宽利用率。
- (2) 宜支持针对作业断点续训，快速调整流量路径，调度新计算节点加入训练任务。
- (3) 宜支持作业运行时 switch port 故障等不可预知异常，实时调整流量路径，保证训练业务稳定运行。
- (4) 宜支持通过服务器网卡设备，处理流量路径调整引入的报文乱序问题。

4.5. 作业运维

4.5.1. 作业监控

4.5.2. 作业可观测

1. 应支持采用轻量级、低侵入式的监控方案，对训推任务的关键时间进行打点和指标采集，确保对作业性能影响最小化。
2. 应支持监控指标数据格式的标准化，支持以结构化日志的形式输出，确保与监控系统的无缝集成。
3. 应支持日志与指标数据携带统一标签，标签需包含：作业名、节点信息、Pod/容器等信息。
4. 应支持提供完整的作业分析视图，可视化展示作业的生命周期状态、性能趋势与关键事件的热力图。

4.5.3. 作业劣化分析

1. 应支持标准化的性能剖析器功能，支持对作业的全栈性能分析。
2. 应支持基于性能指标变化趋势感知作业劣化问题，如：GPU 利用率持续偏低、迭代速度异常下降等。

3. 应支持用户指定时间启动性能剖析操作，支持用户配置策略自动触发生性能剖析任务
4. 应支持基于性能指标数据自动化生成瓶颈分析报告。指标数据包括：算子耗时、CUDA 内核执行时间、CPU/GPU 内存拷贝、通信通信耗时等。
5. 应具备统一的性能瓶颈分类，支持在分析报告中明确标识性能瓶颈的源头类别，如：数据加载瓶颈、模型计算瓶颈、内存带宽瓶颈、跨节点通信瓶颈等，并指出具体的耗时算子或操作。

4.5.4. 作业故障诊断

1. 应支持建立可动态更新的训推任务故障模式库，记录常见的故障场景（如 OOM、节点故障、依赖缺失等）、表象特征（如日志错误内容、指标异常模式等）、以及对应的根因分析和解决方案。
2. 应支持通过关联故障模式库自动识别故障，支持结合作业日志、事件、资源监控及 Profiler 数据等信息，进行智能根因分析。
3. 应支持自动识别作业的典型故障状态，包括：作业中断、作业卡死、性能劣化等。
4. 应支持基于轻量级运行时性能指标、日志和全量的 Profiler 数据进行故障诊断，诊断结果以报告形式输出，给出疑似根因、推荐修复建议。

4.5.5. 性能调优

1. 应支持基于 Profiler 等工具采集数据的性能一调优。
2. 宜支持对性能调优过程设置系统的分析环节，包括：
计算侧：支持优化模型算子、混合精度策略、激活检查点等。
数据侧：支持优化数据 Pipeline、数据加载方式、存储 I/O 性能等。
通信侧：支持优化分布式训练的通信拓扑、梯度同步策略、网络参数等。

5. 修订记录

版本	修订日期	修订内容
V1.0	2026-06	完成V1.0版本内容编写。