基于OMAT的一平台多中心监控方案设计与实现

胡庆宝
huqb@ihep.ac.cn

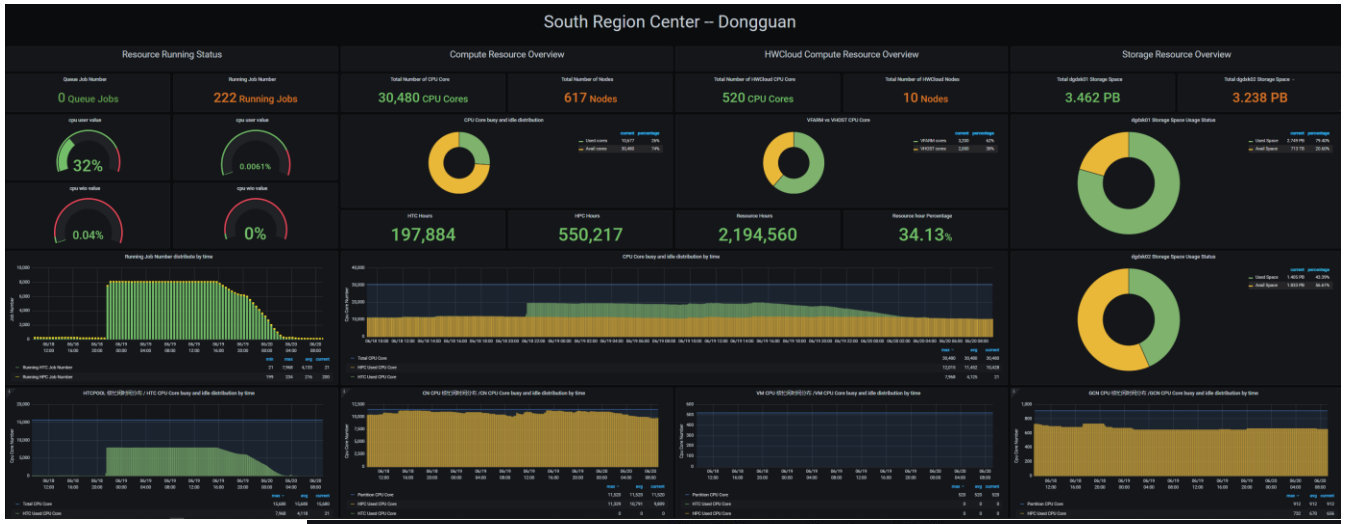
2023-07

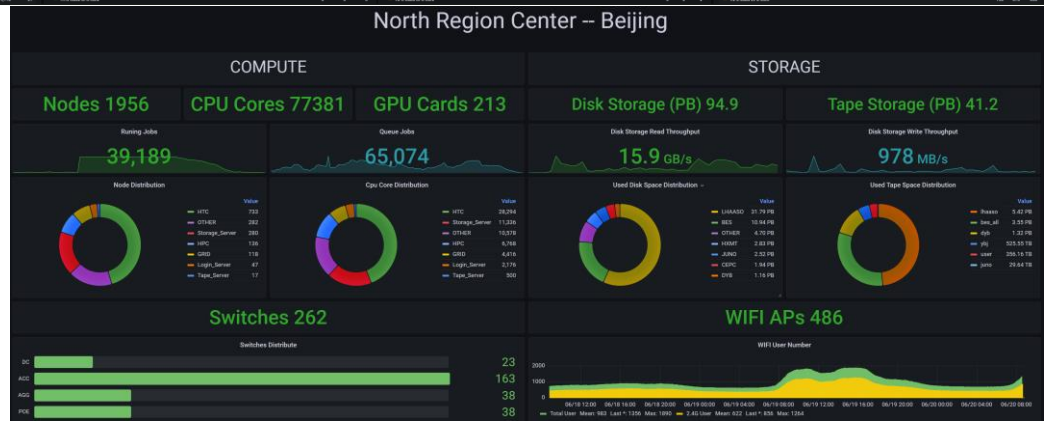
提纲

- 高能物理分布式计算平台
- 运维挑战与运维目标
- OMAT监控平台设计框架及功能
- 运维平台应用案例
- 总结和展望

高能物理分布式计算平台

- 先进高能物理分布式计算平台（一平台多中心）
  - 北方区域中心
    - 高能所计算中心
      - HTC/HPC/网格
      - Lustre/EOS/磁带库
  - 南方区域中心
    - 散裂中子源&东莞大科学智算中心
      - HPC/云计算
      - 华为OceanStor 9000
  - 众多远程站点
    - 四川稻城、西藏阿里等大型实验装置配套IT资源
    - 科大、山大、兰大、山东高研院等高校资源
    - 可弹性接入国家超算中心、商业云等外部资源
  - 面向多学科领域提供科学数据计算分析服务





高能物理分布式计算平台

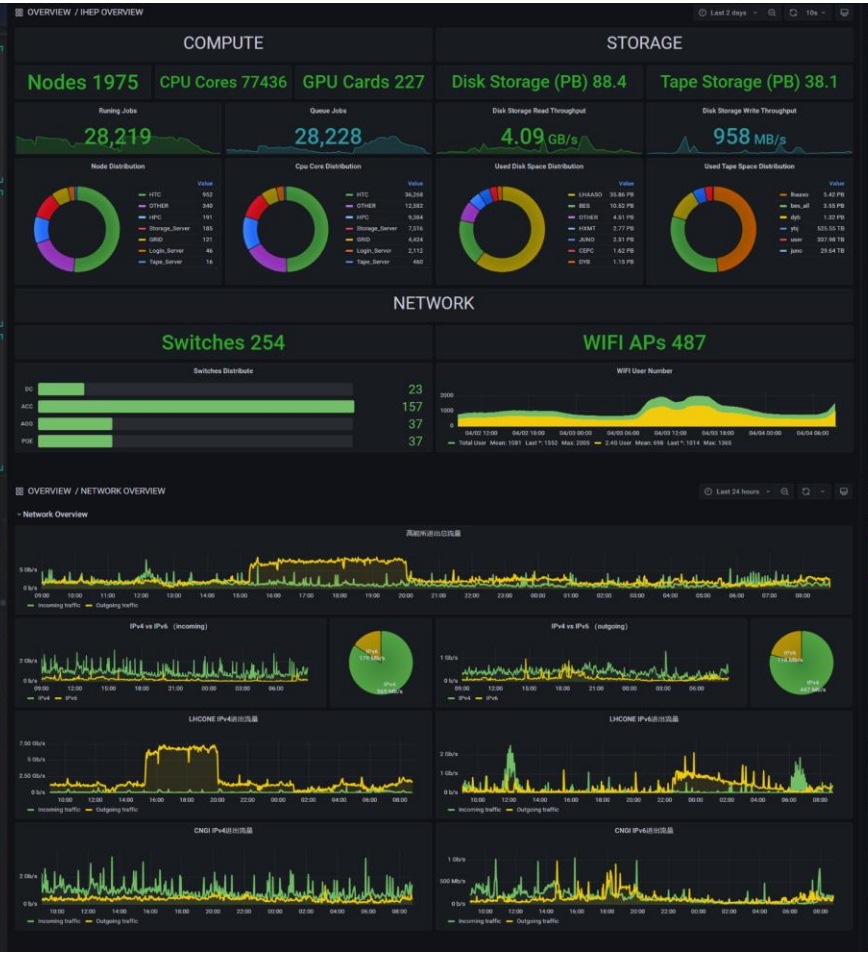
- 分布式计算平台资源状态监控

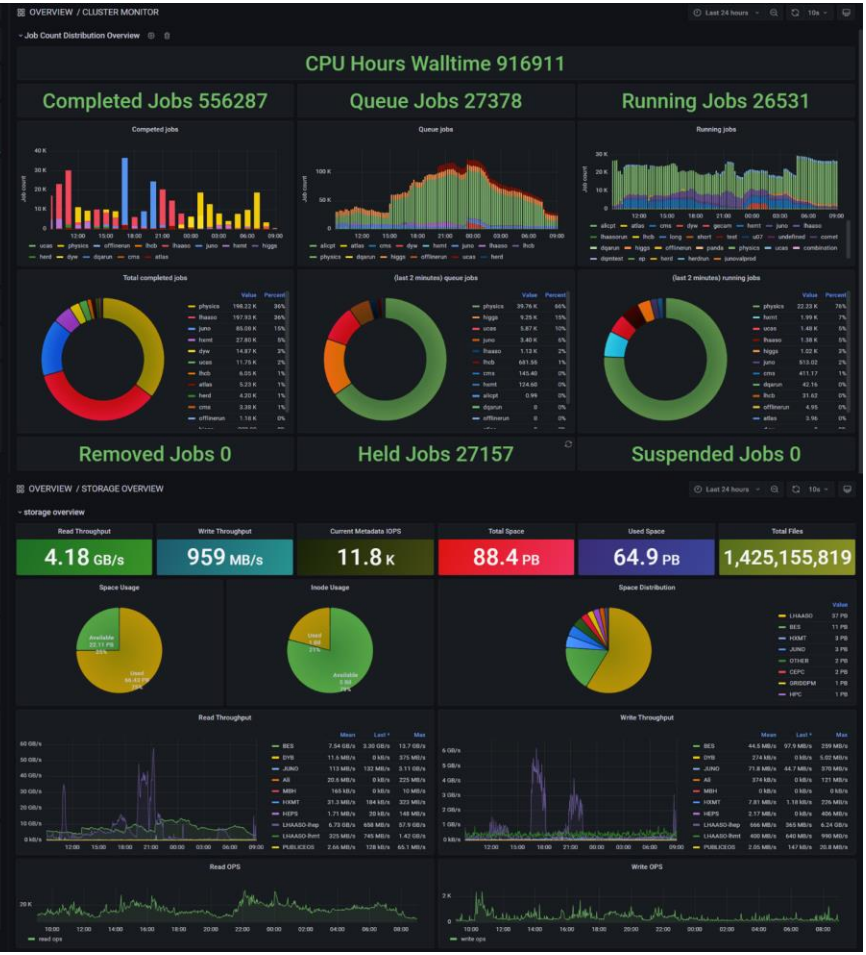
一平台多中心运行状态实时监控
2023.06.01~2023.06.07

站点	CPU资源 (CPU Cores)	CPU资源利用率 (%)	磁盘存储空间	存储数据量	完成作业数量 HTC&HPC	作业运行时间 (CPU小时)
高能所计算中心	51,788	81.82%	94.74 PB	60.81 PB	2,220,896	7,018,398
东莞大科学中心	27,400	59.15%	6.70 PB	4.13 PB	6,544	2,541,507
LHAASO稻城	3,320	18.56%	5.33 PB	4.29 PB	1,675	101,332
散裂中子源	5,572	19.02%	689.8 TB	350.0 TB	5,496	178,979
山东大学	1,504	21.39%	352.9 TB	258.3 TB	3,078	36,249
中国科技大学	3,838	34.59%	1.17 PB	980.4 TB	9,901	225,074
兰州大学	1,768	27.53%	341.8 TB	297.0 TB	1,565	52,504

高能物理分布式计算平台

- 高能所计算中心运行监控





提纲

- 高能物理分布式计算平台
- 运维挑战与运维目标
- OMAT监控平台设计框架及功能
- 运维平台应用案例
- 总结和展望

运维面临的挑战与需求

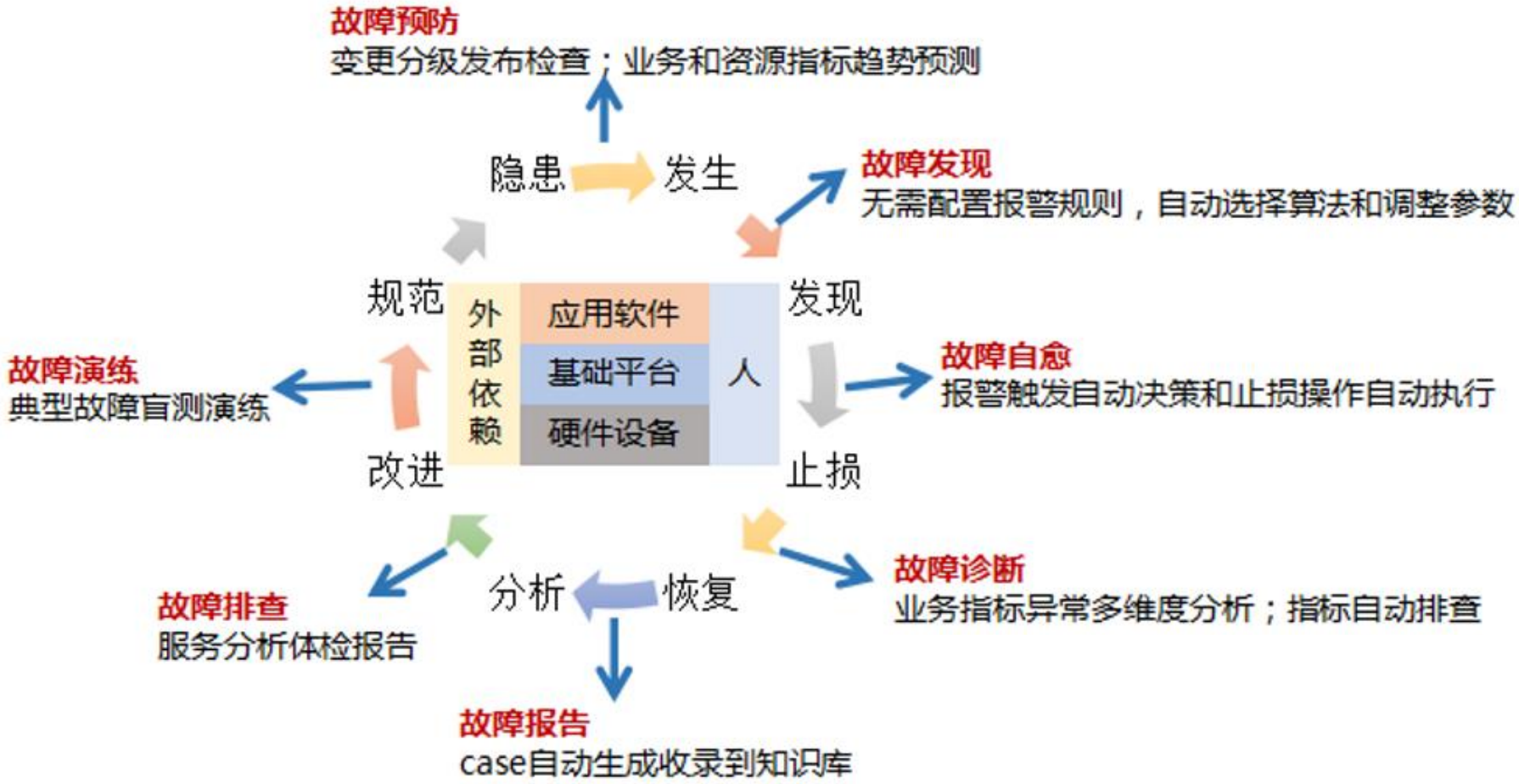
- 计算平台服务面临挑战
  - 分布式系统——解决了单机CPU、内存、存储资源受限问题，
  - 跨地域分布式系统——需求激增情形，共享异地算力资源，加速科学计算。
  - 随着分布式计算平台规模的不断扩大
    - 计算环境日益复杂、支撑服务种类多样化、平台监控数据规模激增。
    - 面向单一领域的自动化运维技术无法有效整合多站点、多种类海量运维数据，呈现多维度监控数据的综合分析结果。
- 大数据&人工智能的平台化运维解决方案
  - 丰富的探针工具、安全的信息传输、可扩展的关联分析框架、高效的数据索引仓库、多样的数据呈现方式

监控平台的设计目标(1/2)

- 企业级的基础数据管理平台（稳定提供“丰富、好用”的数据）
  - 经济性 开放性 统一性 可靠性 安全性
  - 共享性 实用性 易用性 应用广泛性
- 智能化的快速分析决策平台（专家系统和AI的价值体现）
  - 支持先存后算和边采边算处理模式
  - 支持数据的关联分析:
    - 指标-指标、多指标-事件、事件-事件、多个事件->故障传播
  - 面向AIOps的算法整合：支持回归、聚类、分类算法
    - 既定策略的快速判定和机器学习的故障推演：基于已知故障库快速识别诊断故障，基于未知故障，结合机器学习进行故障特征推演。
- 敏捷型的自动化控制平台（运维能力的最终实现手段）
  - 多样化的异常告警：微信、邮件、短信等
  - 故障止损 & 故障治愈

监控平台的设计目标(2/2)

- 规范化的智能监控管理流程（实现智能化运维管理的闭环）



提纲

- 高能物理分布式计算平台
- 运维挑战与运维目标
- OMAT监控平台设计框架及功能
- 运维平台应用案例
- 总结和展望

平台架构

资源调度层

Kubernets

Prometheus

资源分配

网络管理

服务告警

数据应用

全局视图 统计报表 日志分析 计算服务策略联动 科学数据共享统计 数据中心网络安全

数据仓库

Elasticsearch Graphite 索引查询 快照备份 数据访问

数据处理层

特征提取 关联分析 模式识别 规则应用 决策优化 实时告警

数据缓存层(汇聚&缓存) 消息发布/订阅

配置中心

权限控制

备份策略

告警配置

决策模型

专家系统

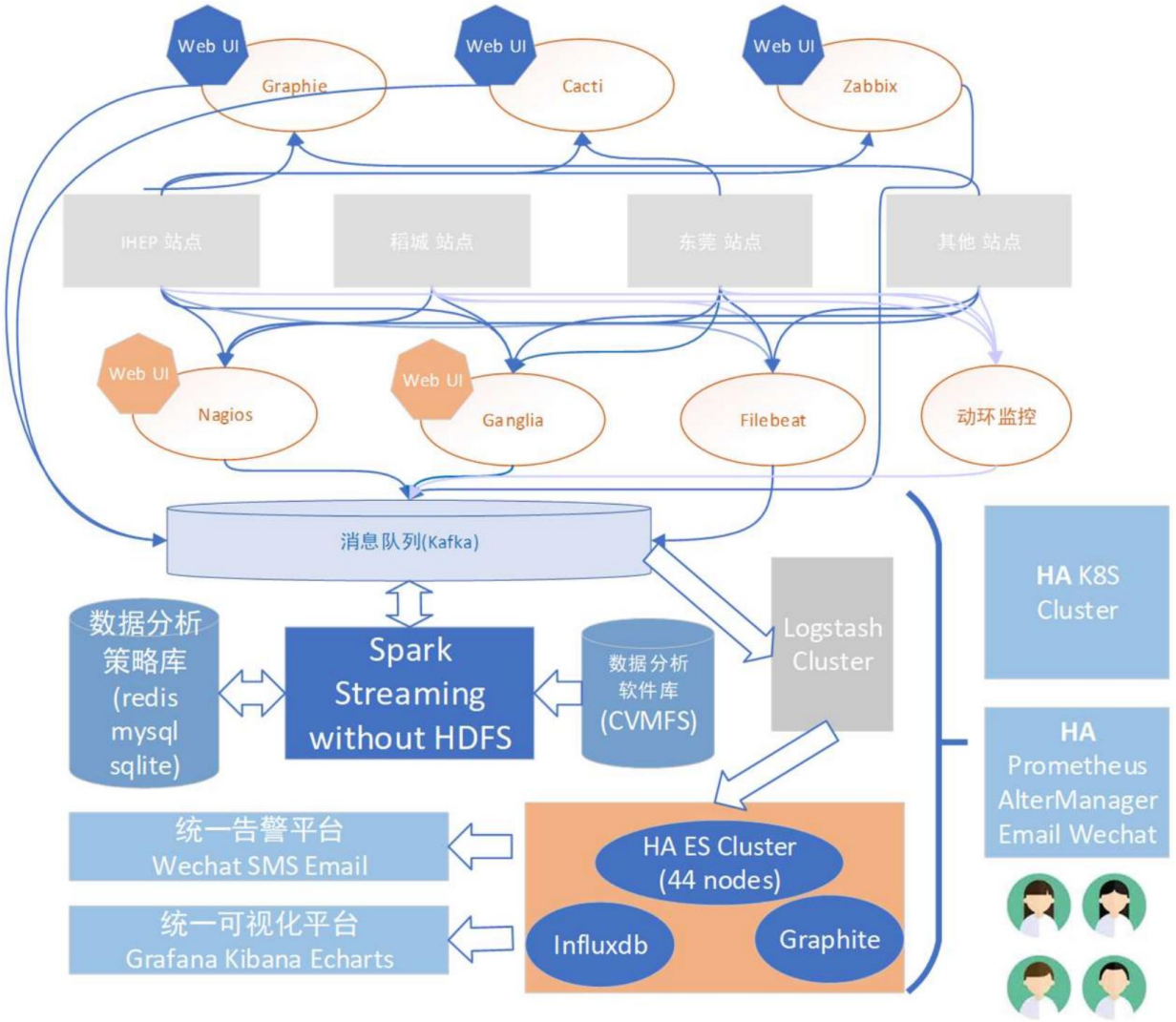
采集配置

生态集成层(采控平台)

Ganglia Nagios Zabbix Collectd Cacti Syslog BACNet WAF 流量探针

计算服务 存储服务 网络服务 操作系统 用户行为 虚拟机 服务器设备 交换设备 安全 动力环境

平台技术栈



Nagios®

ZABBIX

Ganglia

cacti

ASHRAE BACnet™

kafka

elastic stack

redis

CernVM
File system

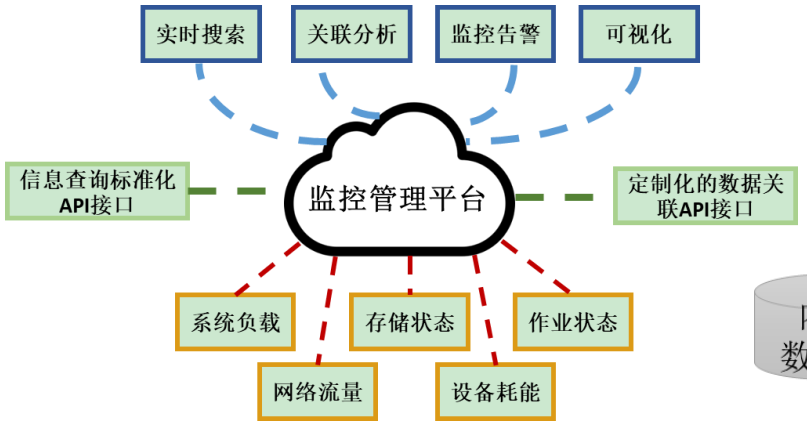
Spark
Streaming

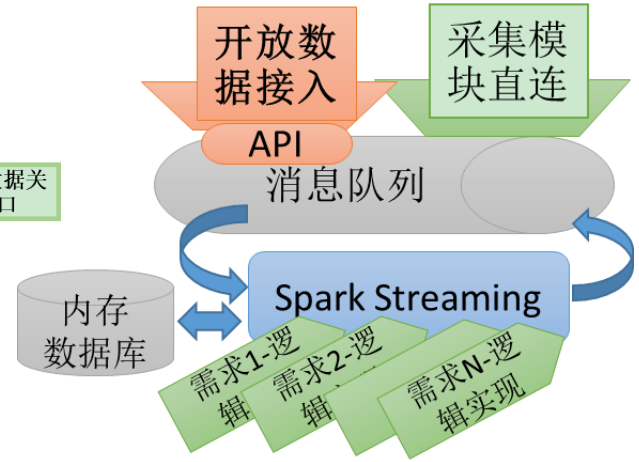
influxdb

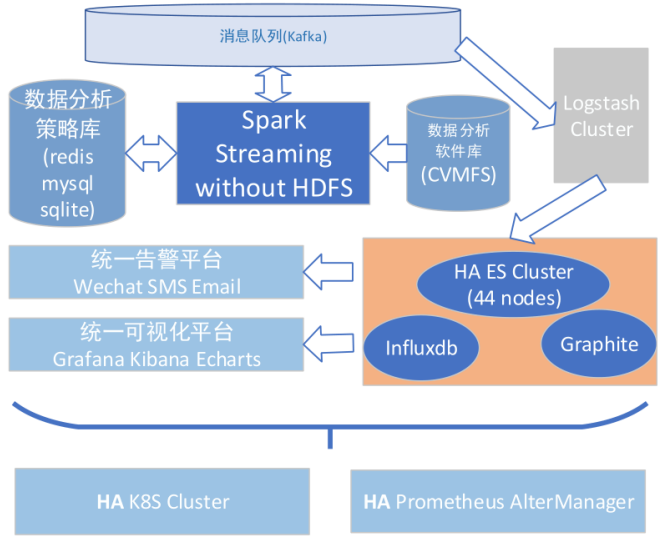
graphite

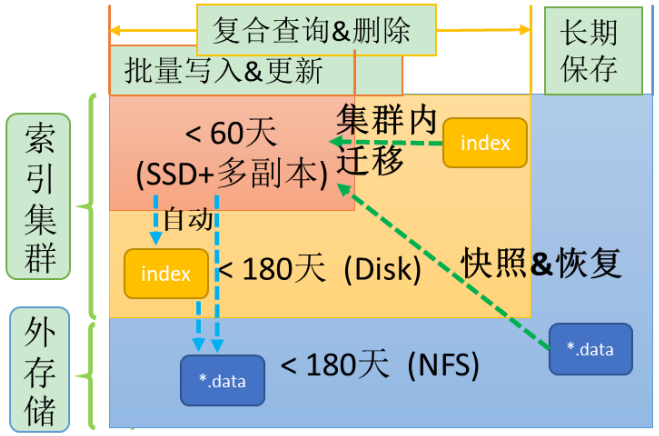
平台架构和特点

- 平台特点：
  - 统一融合的运维数据系统，统一的数据汇入和标准化的数据输出
  - 灵活高效的处理分析，可定制化的运维数据分析流程和近实时的数据处理能力
  - 运维数据和应用场景融合，支持数据业务化、分析智能化
  - 运维数据分析系统云化，容器化部署，降低运维成本
  - 数据分析和存储高可靠高可用
    - 机柜级容灾、多副本
    - 无单点瓶颈、可横向扩展



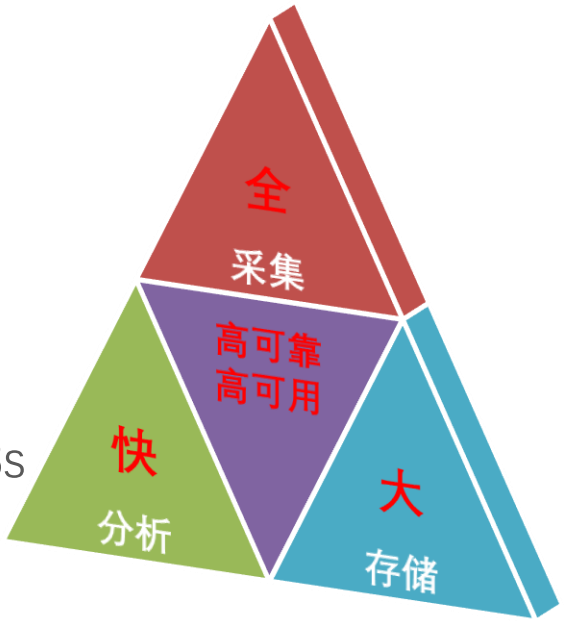


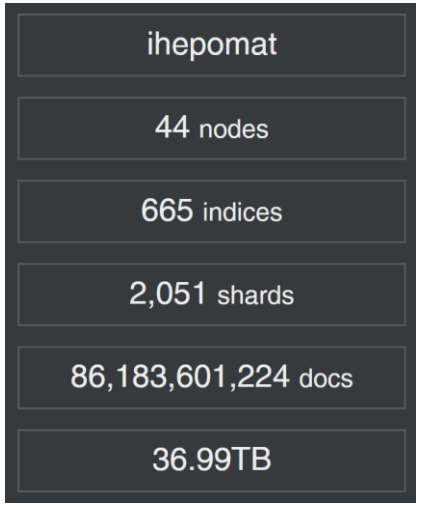




部署条件和性能(2/2)

- 平台硬件和部署
  - 9*CPU服务器 & 1*GPU服务器
    - 2台CPU服务器用于部署高可用数据缓存模块
    - 1台GPU服务器用于训练异常模型。
    - 7台CPU服务器部署高可用容器集群管理系统K8s





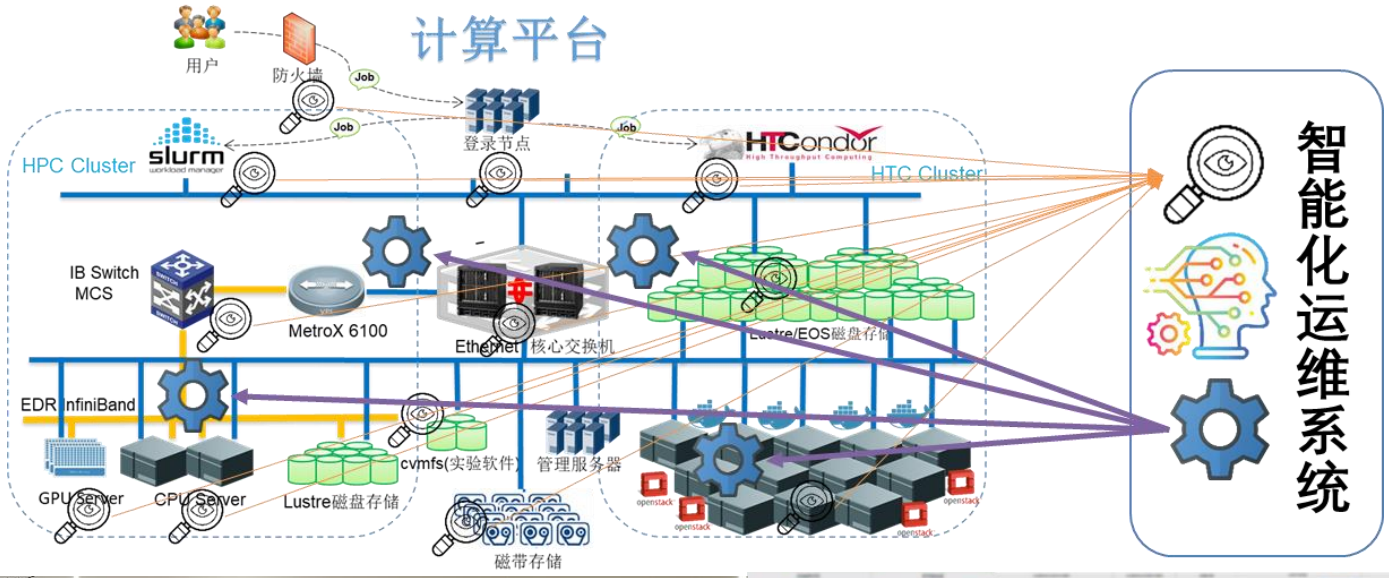
- 平台性能
  - 数据最大采集能力 ~150k doc/s （覆盖北京本部和异地资源站点 5k+节点）
  - 数据平均处理能力 ~60k doc/s 数据处理延迟 ~2s（Spark Streaming on K8s）
  - 数据最大索引能力 ~180k doc/s （44节点 ES集群 on K8s）
  - 覆盖400+运维数据指标种类
  - 累计存储原始数据850+亿条 & 拥有11k+加工汇总后关键运维指标数据库

提纲

- 高能物理分布式计算平台
- 运维挑战与运维目标
- OMAT监控平台设计框架及功能
- 运维平台应用案例
- 总结和展望

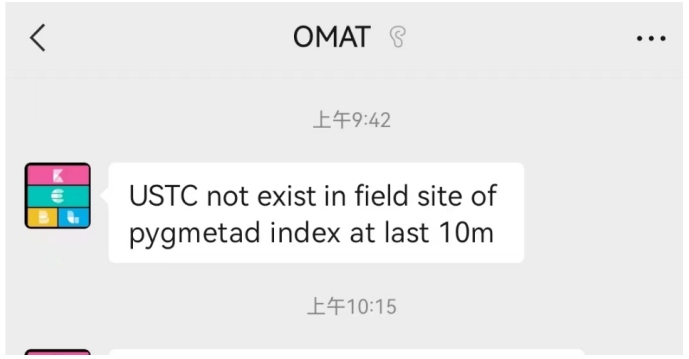
运维支撑和分布式管理应用

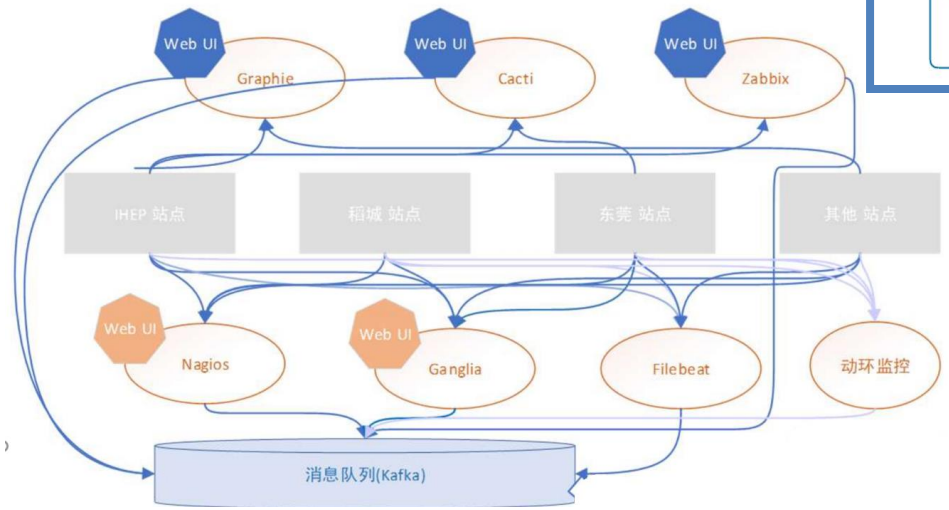
- 平台覆盖范围
  - 机房动力环境
  - 硬件设备、虚拟机、容器
  - 系统性能、存储性能、网络性能
  - 作业调度、作业数据访问行为、资源管理
  - 安全认证、用户行为、网络攻击
  - 异地站点监控等多种应用场景

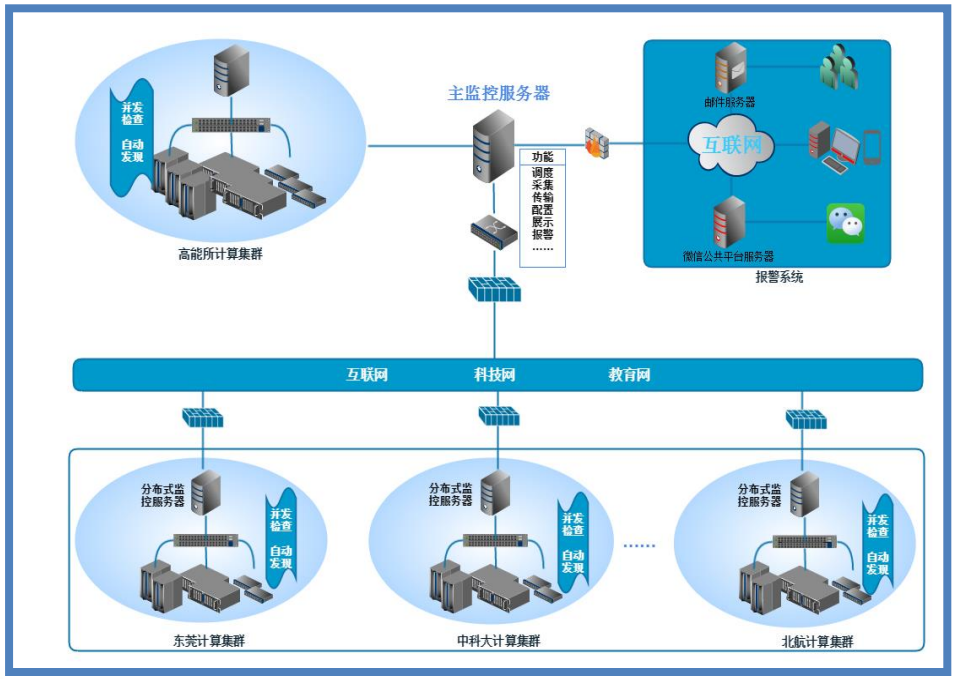


跨域运维关键技术

- 域内采集技术
  - 汇聚网关
- 域标定技术
  - 定制探针域源标签
- 断点续传技术
  - 网关缓存
- 跨域汇聚技术
  - 数据加密
  - 数据压缩
- 域漏采告警

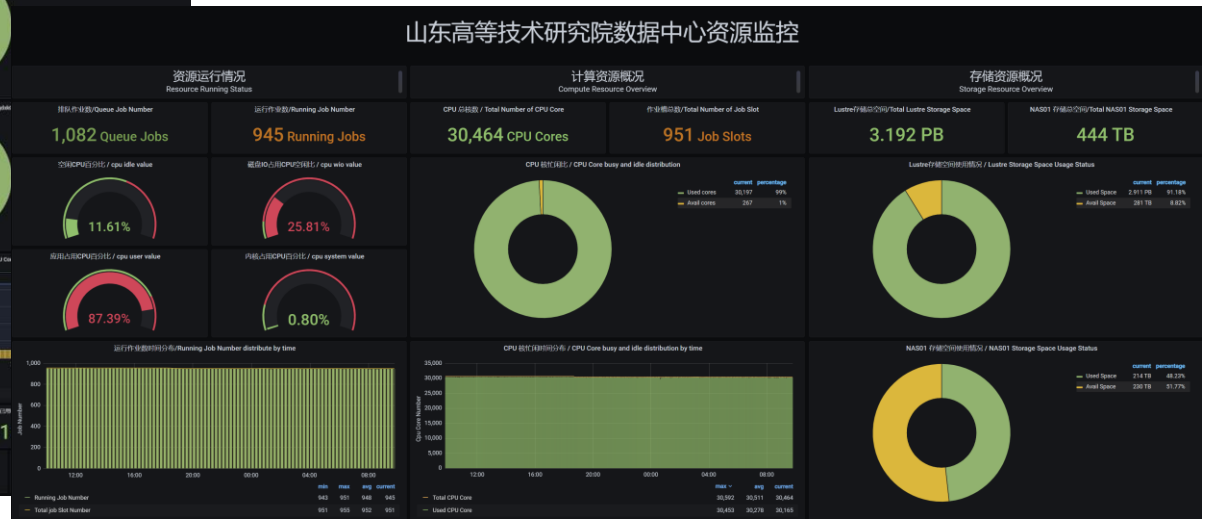






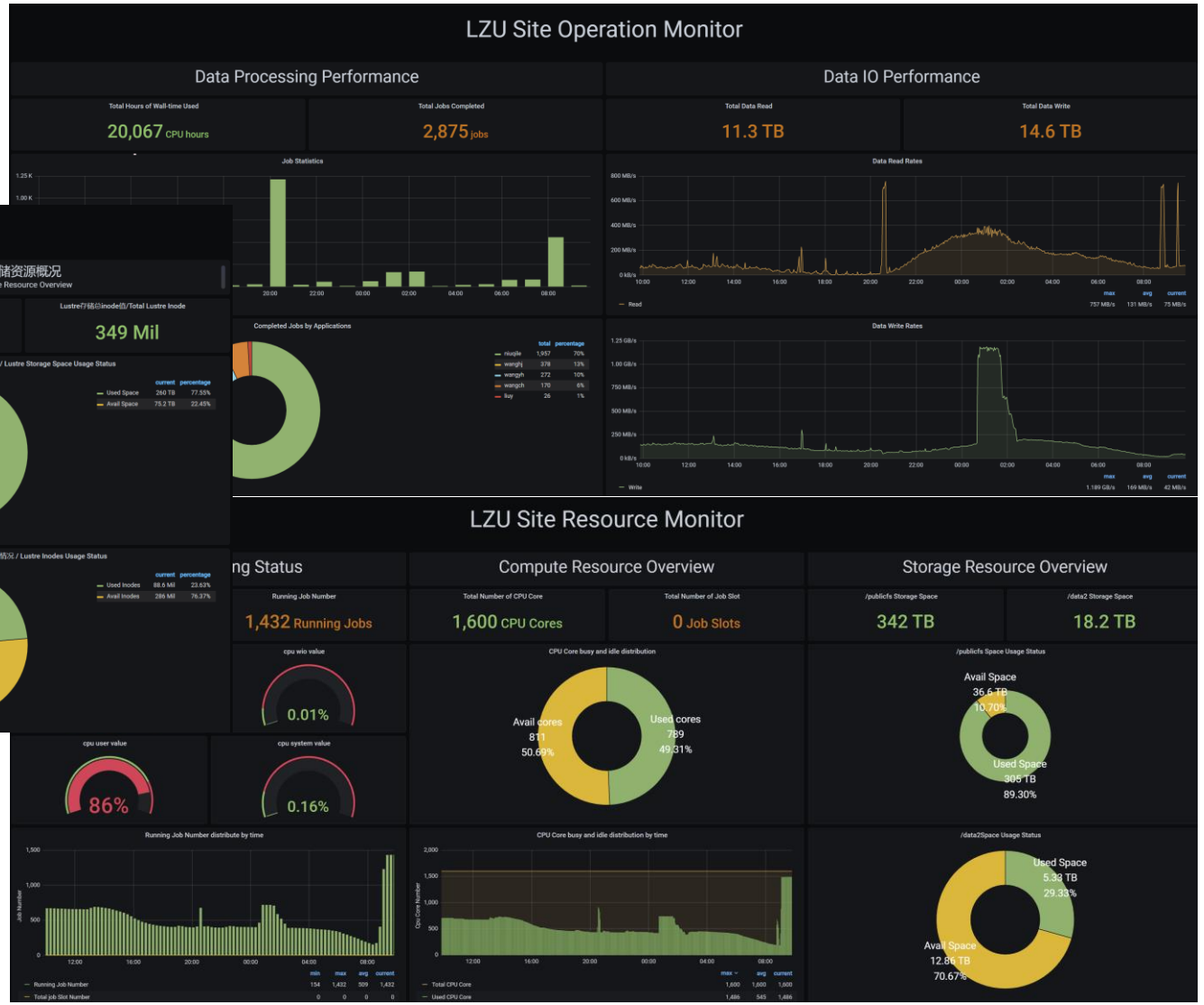
跨域运维应用

- 山东高研院监控
- 东莞大科学智算中心



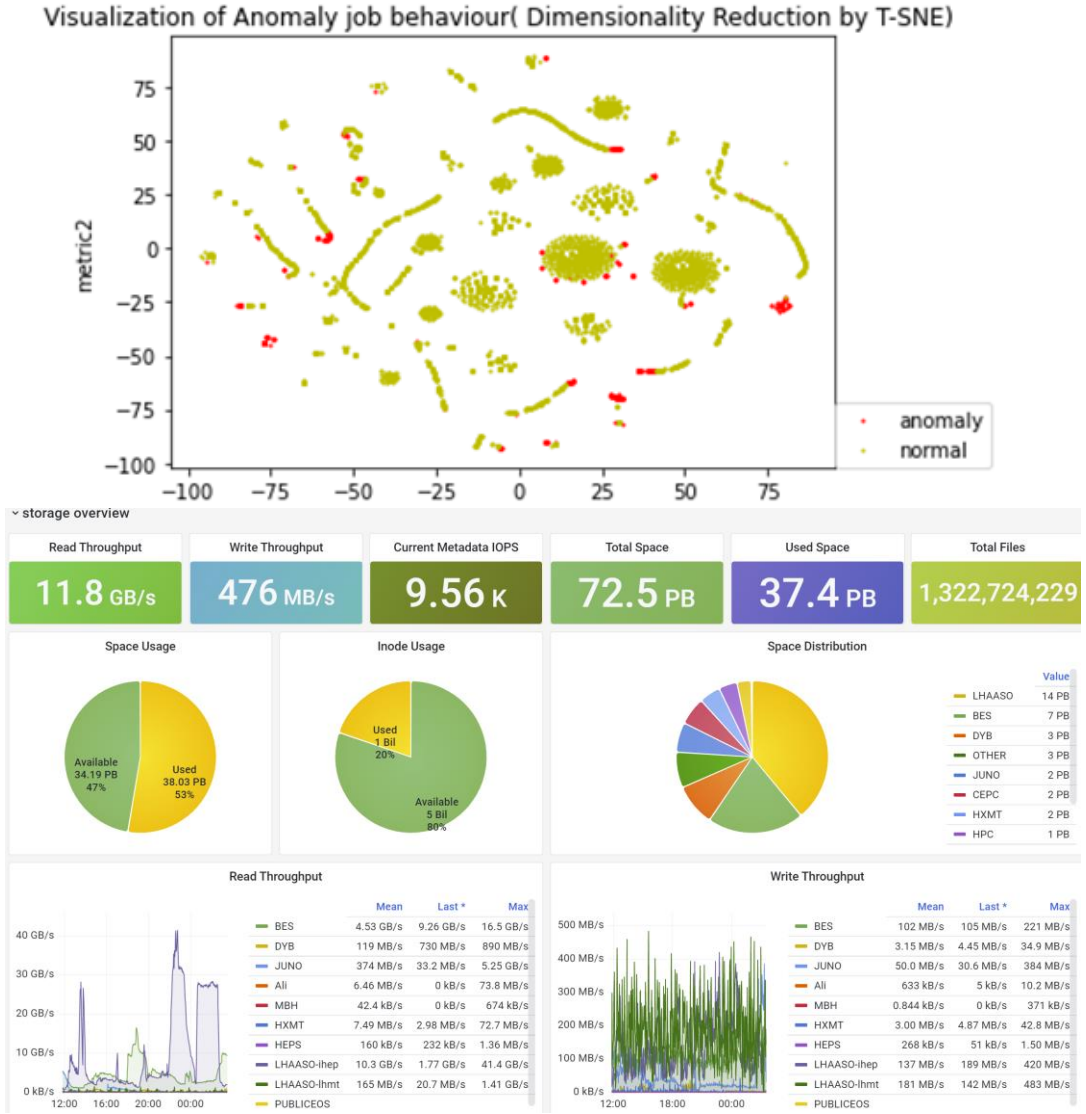
跨域运维应用

- 高校站点监控



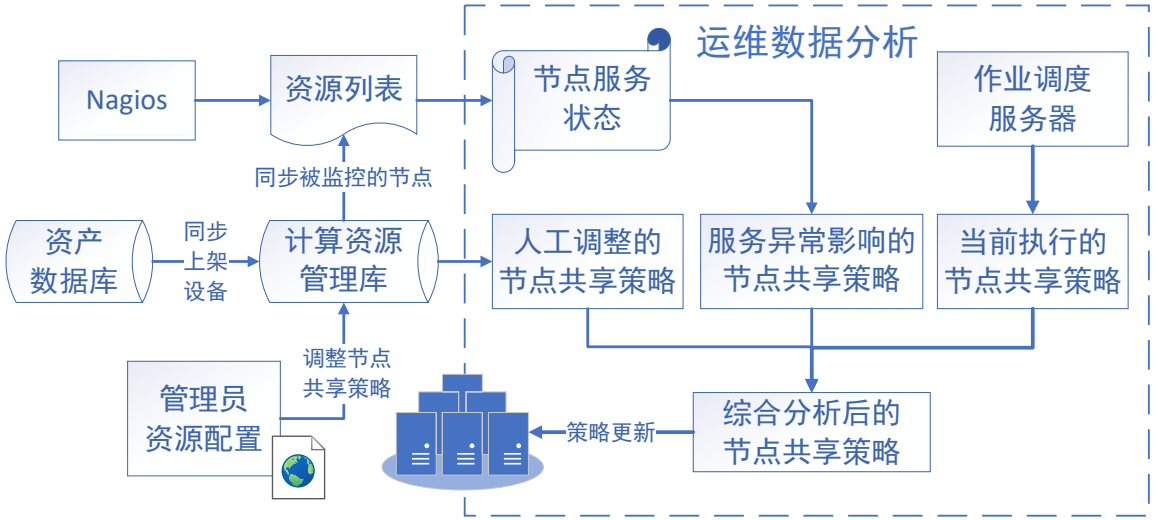
数据访问和存储服务应用

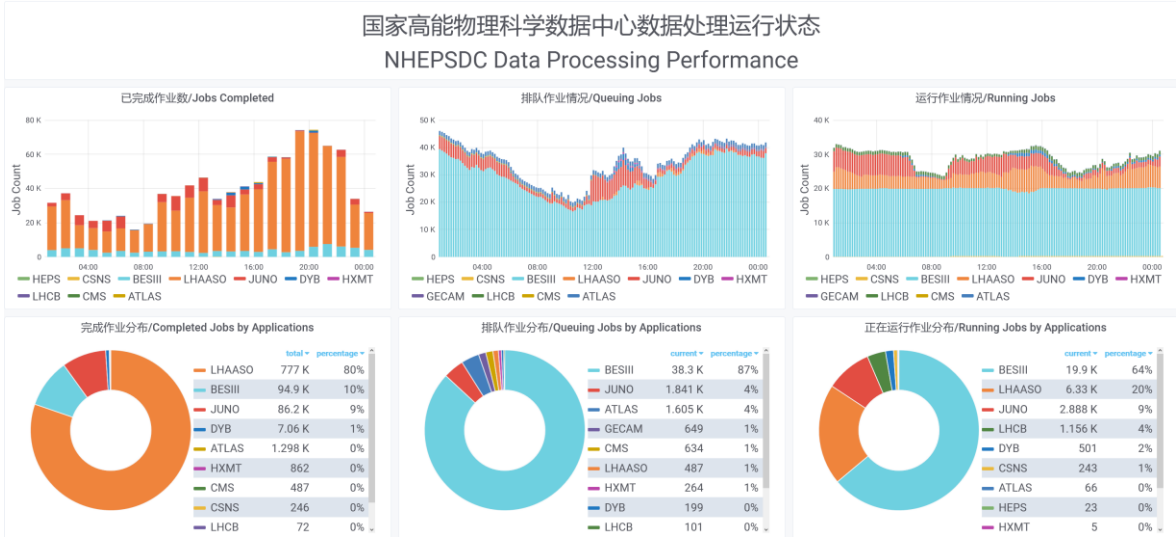
- 基于机器学习的作业异常IO访问行为探测
  - 采集分析作业的I/O模式特征
    - 30万用户作业同时运行，产生15万/秒的open、read、seek、write、close、attr等文件访问行为数据。
    - 关联每个访问行为和对应的作业ID、作业所属实验等信息。
  - 存储每个实验专有I/O数据库
  - 采用孤立森林等无监督学习算法，查找异常I/O行为作业



资源管理和计算服务应用

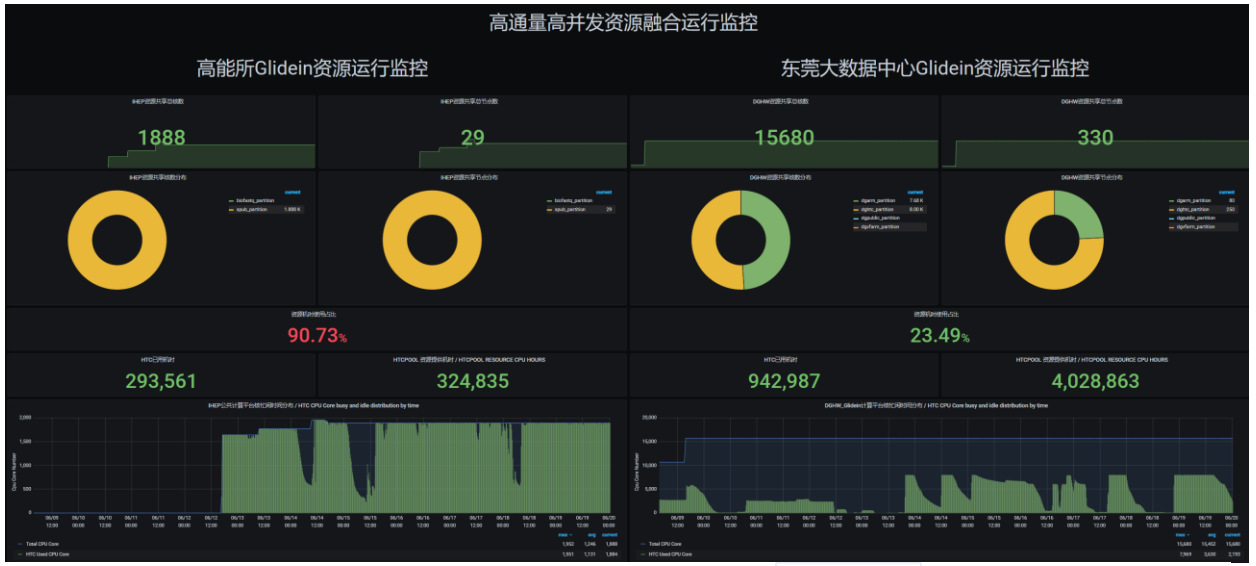
- 自动化资源池管理系统
  - 故障探测->自动化故障止损
    - 故障发现到资源隔离10s内完成
    - 极大的减少了服务故障对用户作业的影响
  - 服务恢复->自动化资源使用
    - 节点服务恢复后联动作业调度系统，将用户作业分发到节点
  - 设备资产管理-计算资源管理
    - 设备上架或修复，自动划入资源池
    - 硬件故障维修，自动摘除资源池



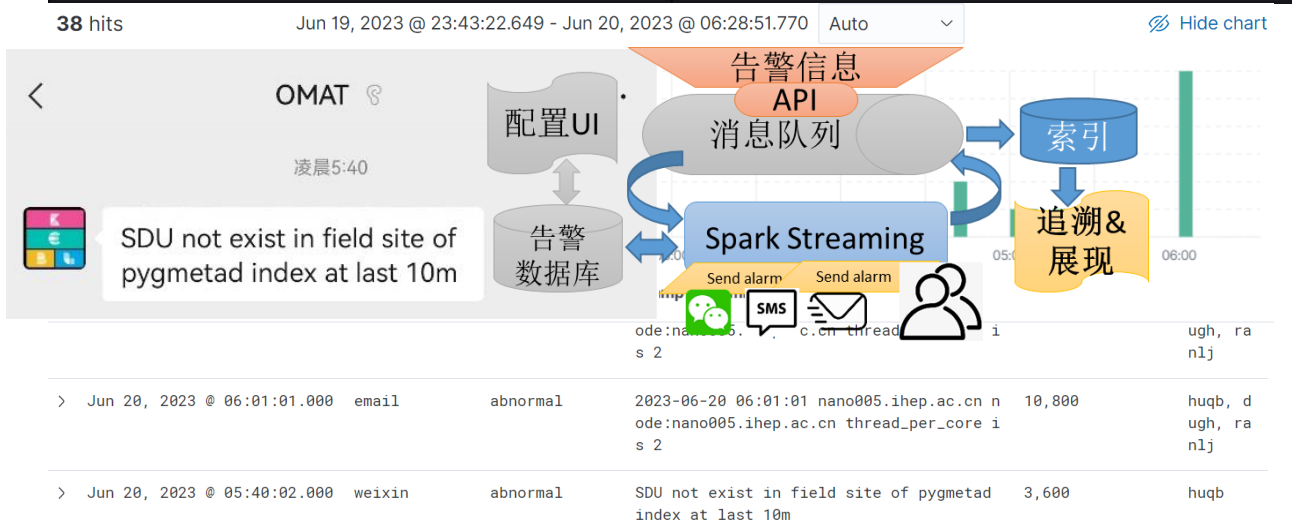


其他应用

- dHTC动态资源池管理系统
  - HTC作业槽监控和心跳监测告警
    - Slurm Glidein作业实时监控
    - HTCondor作业槽实时监控和Slurm作业关联分析



- 统一消息推送平台
  - 支持邮件、短信、微信推送方式
  - 历史消息可追溯、可查询
  - 支持消息抑制、去重合并等功能



提纲

- 高能物理分布式计算平台
- 运维挑战与运维目标
- 智能化运维平台设计框架及功能
- 运维平台应用案例
- 总结和展望

总结和展望(1/2)

- 基于大数据和机器学习领域技术，设计实现了一套可广泛应用的数据中心智能运维系统，实现了海量运维数据的价值演进，解决了大规模数据中心日益复杂的运维难题。
- 实现了专业化的DataOps平台
- 在部分运维领域结合机器学习算法，对AIOps进行了一定探索

总结和展望(2/2)

- 完善运维元数据地图建设
  - 直观展现不同层次的运维数据在整个系统的流转情况
  - 帮助用户了解不同系统的运维数据流动路径
  - 快速定位数据实体在不同系统所处的位置
  - 评估数据实体异常影响的系统范围
- 搭建丰富的运维学件库，结合智能运维系统，提供AIasService
  - 基于已积累的监控数据训练丰富的运维学件模型
  - 支持用户导入自己的监控数据，灵活选择学件模型来实施 AIOps
  - 降低实施 AIOps 的技术门槛，让广大运维人员能够快速使用 AIOps

Thanks for your attentions!

谢谢！