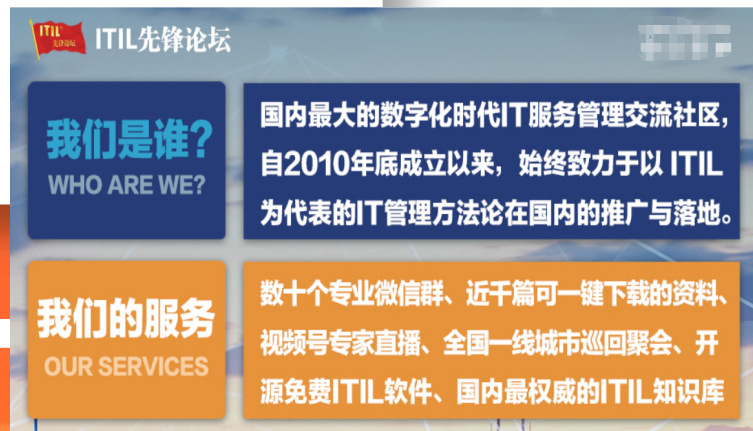


# 超大规模IT稳定性保障实践

运维负责人



ITIL 先锋论坛

**我们是谁?**  
WHO ARE WE?

国内最大的数字化时代IT服务管理交流社区，自2010年底成立以来，始终致力于以 ITIL 为代表的IT管理方法论在国内的推广与落地。

**我们的服务**  
OUR SERVICES

数十个专业微信群、近千篇可一键下载的资料、视频号专家直播、全国一线城市巡回聚会、开源免费ITIL软件、国内最权威的ITIL知识库

# 目录

## contents

- 1 困难与挑战
- 2 稳定性保障体系
- 3 工具篇
- 4 运营保障篇

# Part 01

## 困难与挑战



# 系统的变革

- 中国联通**cBSS系统**，全球最大的电信运营商支撑系统，集约31省BSS系统，支撑全国营业受理、计费、账务、客服、手厅、触点营销等业务，用户规模超**4亿**





# 面临的困难和挑战



## 超大规模

- 用户规模：4.1亿+
- 计费数量：450亿+/天
- 服务调用：30亿+/天
- 服务实例：4万+
- 容器数量：20w+



## 混合架构

- 传统架构+分布式架构
- 割接、迁转
- 长链路复杂调用
- 新技术栈不灵



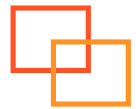
## 故障频发

- 故障调度乱
- 故障定位慢
- 铁路警察，各管一段
- 故障处置慢
- 故障定责难



## 工具不全

- 工具分散（60+）
- 数据不拉通
- 缺乏全流程的可观测
- 工具关键时候发挥不了作用



寻求解决之道

体系 (做什么) + 措施 (怎么做) + 工具 (高效做)

## Part 02

# 稳定性保障 体系



# 中国联通大IT稳定保障体系



□ 大IT生产运营体系：一个目标，依托四大保障，聚焦研运流程中Ops部分的三个阶段，对应十四项核心工作

一个目标

做实安全生产，提升中国联通大IT稳定性

四大保障

制度规范保障

组织架构保障

平台工具保障

运营机制保障

五个阶段

架构设计

研发测试

生产变更

运营保障

运营响应

十七项  
核心工作

高可用架构

需求设计

研发测试

变更管理

版本管理

链路识别

监控管理

隐患管理

应急管理

故障调度

故障定位

故障闭环

值班管理

问题运营

知识运营

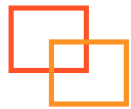
服务质量

运营分析

研运流程

Dev

Ops



# 制度规范保障

- 制度规范保障方面制定了《大IT生产运营规范》、《红线底线纪律条例》以及《大IT生产运营奖惩方案》，落实集中系统维护职责，筑牢安全风险防线，激发活力、时刻警醒。



## 大IT生产运营规范

- **总册+十三个分册**：是B域、M域、D域集中系统生产运营工作平稳推进的基础性规范和依据。明确团队组织、分工、流程和工作要求，落实集中系统维护职责



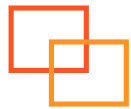
## 红线底线纪律条例

- **七大类、二十九条**：明确生产运营工作中的红线底线纪律条例，时刻保持高度警醒，筑牢安全风险防线



## 大IT生产运营奖惩方案

- **九项奖励，四项处罚**：通过表扬、通报、奖励及小额工资扣罚等方式，达到明确导向、激发活力、时刻警醒、防范风险的目的



# 一体化SRE运营团队

集团：数字化部运营管理处

运营管理

联通软件研究院 - 一体化SRE运营团队

四级安全生产运营保障团队

团队职责

应用运维

平台运维

基础设施运维

团队结构

一级负责总体

二级负责全站

三级负责核心业务链路

四级负责中心/模块

稳定性保障工具研发团队

团队职责

工具体系规划

工具建设

工具运营

团队结构

运维研发核心团队

应用运维研发团队

云平台运维研发团队

基础设施运维研发团队

一体化三级运营响应团队

团队职责

问题响应

问题分析

质量管理

数据修复

团队结构

三级修复团队

二级支撑团队

一级响应团队

生产调度

双向协同

问题响应

分子公司 - 运营团队

生产调度

双向协同

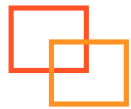
问题响应

分子公司安全生产运营保障团队

分子公司稳定性保障工具研发团队

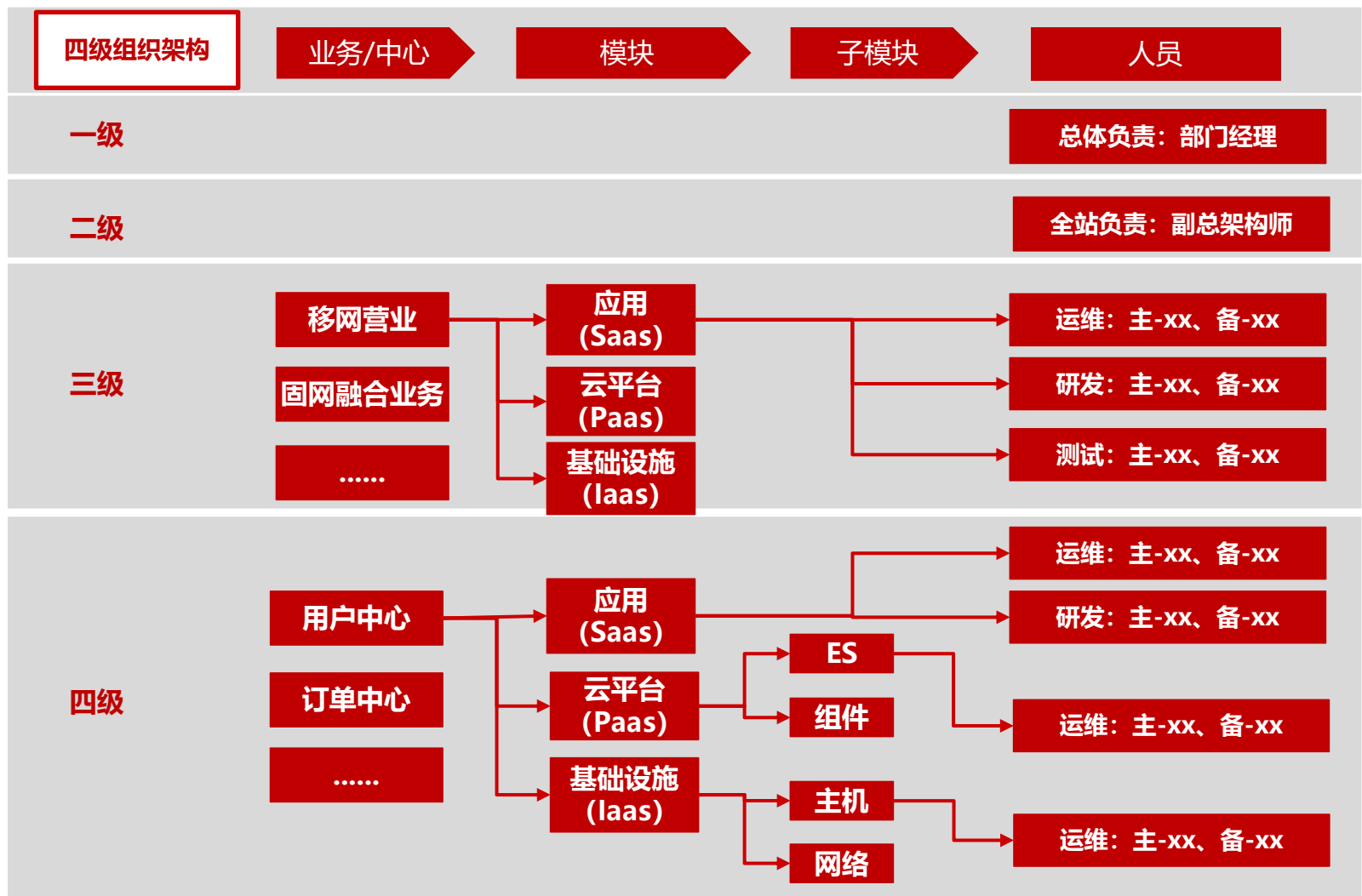
分子公司运营响应团队

一线业务



# 四级安全生产运营保障团队

## 创建以核心业务链路为重点保障的四级安全生产团队



### 职责划分

#### 四级安全生产团队

- 一级总体负责安全生产责任团队；
- 二级全站负责安全生产责任团队；
- 三级负责核心业务安全生产，涉及1个或多个四级团队；
- 四级负责各中心安全生产，可归属于多个三级核心业务；

#### 纵向职责

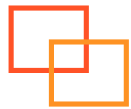
- **应用**：核心应用稳定性保障工作
- **云平台**：保障云平台的稳定性、可观测性，配合应用
- **基础设施**：保障基础设施稳定性、可观测性，配合应用

#### 角色职责

- **运维**：链路识别、监控告警、应急流程、故障定位调度...
- **研发**：业务架构、业务功能
- **测试**：功能验证、性能验证

# Part 03

## 工具篇



# 平台工具保障

数字化监控平台为运营保障2.0体系中十四项核心运营工作提供端到端、全层级、全流程工具支撑保障

## 中国联通数字化监控平台

| 监控管理                                              | 配置管理             | 自动化运维         | 变更管理           | 稳定性测试                                             | 业务连续性                | 用户体验             | 问题管理                         | 知识管理                 |
|---------------------------------------------------|------------------|---------------|----------------|---------------------------------------------------|----------------------|------------------|------------------------------|----------------------|
| 浏览器监控<br>APP监控<br>一键故障诊断<br>亚健康检查<br>业务监控<br>日志中心 | 传统CMDB<br>云化CMDB | 自动化作业<br>故障自愈 | 任务调度平台<br>变更追踪 | 接收测试<br>自动化巡检<br>压力测试<br>版本发布调度<br>APP测试<br>全链路压测 | 故障管理<br>隐患管理<br>值班管理 | 感知运营<br>cBSS感知运营 | 一体化运营<br>沃运营<br>补天平台<br>订单医生 | 沃运营<br>一体化运营<br>知识中心 |



端到端



全层级



全流程



## OPS-十四项核心工作

| 变更管理           | 版本管理           | 链路识别                     | 监控管理                                                  | 隐患管理          | 应急管理                  | 故障调度                             | 故障定位            | 故障闭环         | 日常值班                     | 问题运营                         | 知识运营                 | 服务质量         | 运营分析         |
|----------------|----------------|--------------------------|-------------------------------------------------------|---------------|-----------------------|----------------------------------|-----------------|--------------|--------------------------|------------------------------|----------------------|--------------|--------------|
| 任务调度平台<br>变更管理 | 接收测试<br>版本发布调度 | 性能指标标准<br>性能测试<br>应用性能优化 | 浏览器监控<br>APP监控<br>调用链监控<br>智能监控告警平台<br>自动化巡检<br>APP巡检 | 隐患管理<br>亚健康检查 | 故障管理<br>自动化运维<br>故障自愈 | 一键拉会<br>故障管理<br>业务监控<br>智能监控告警平台 | 一键故障诊断<br>调用链监控 | 故障管理<br>业务监控 | 值班管理<br>故障管理<br>智能监控告警平台 | 沃运营<br>一体化运营<br>订单医生<br>补天平台 | 沃运营<br>一体化运营<br>知识中心 | 沃运营<br>一体化运营 | 沃运营<br>一体化运营 |

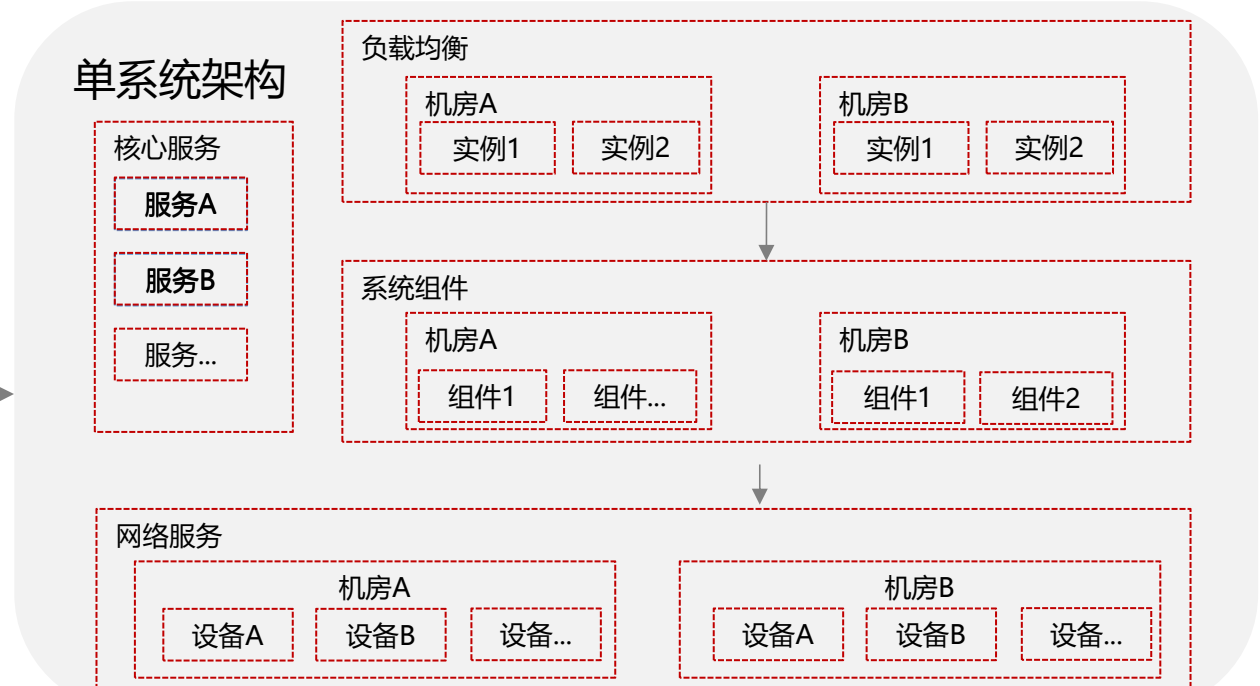
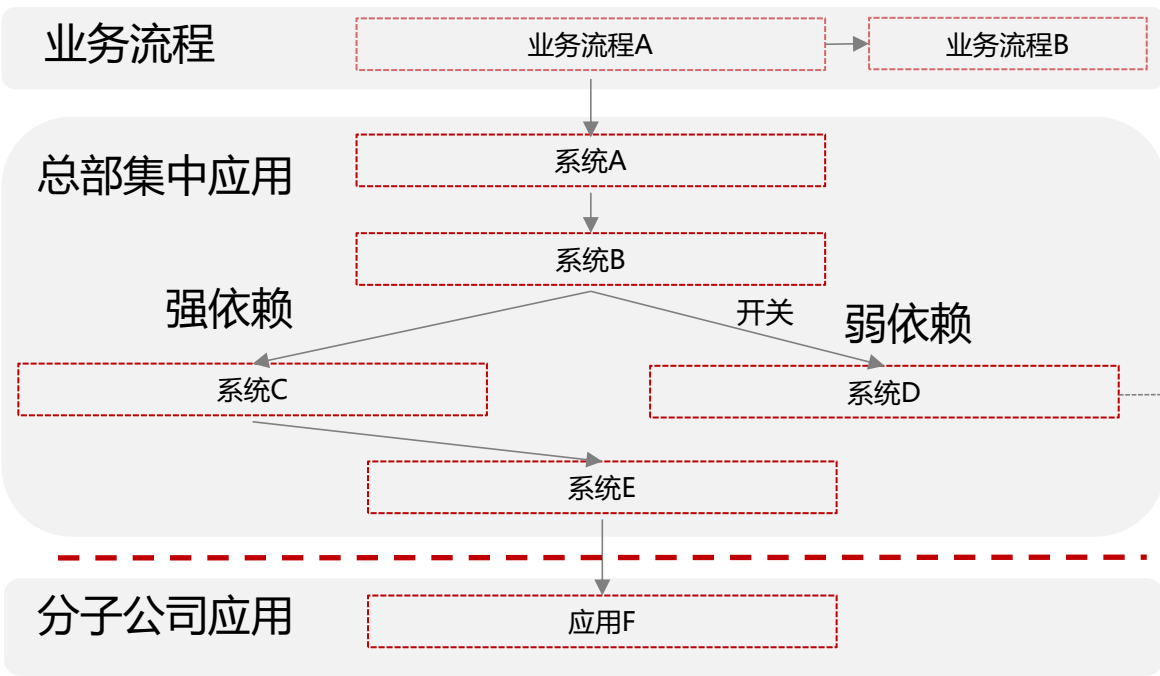
# Part 04

## 运营保障篇



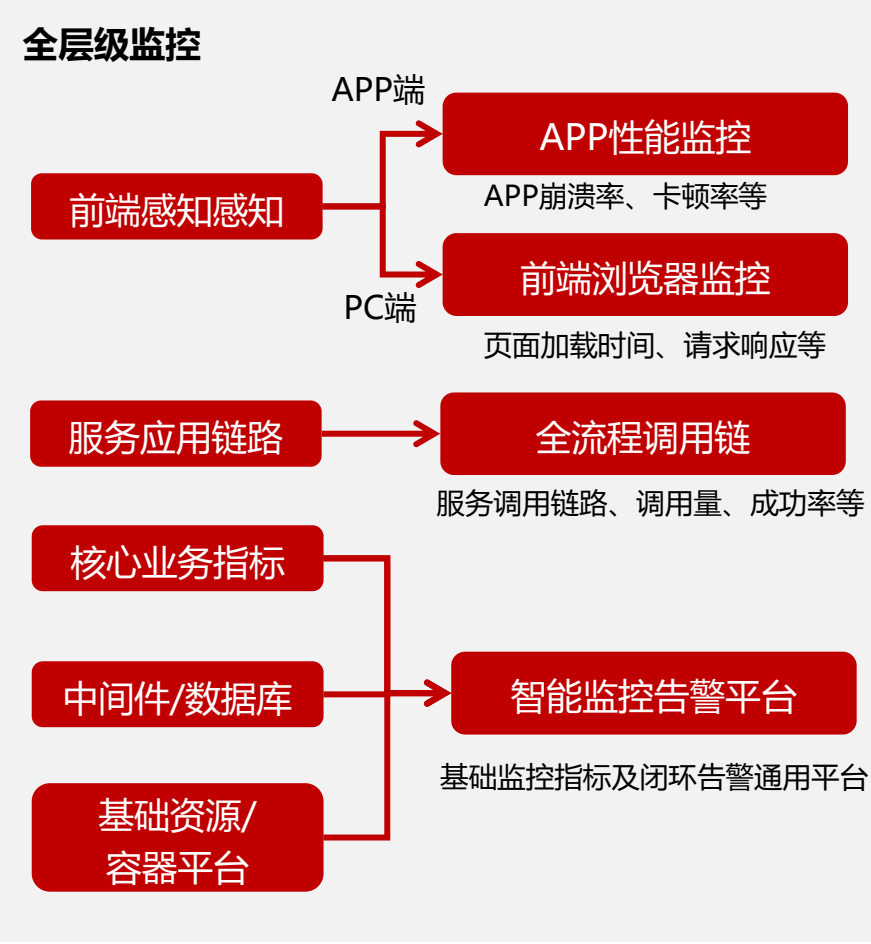
# 运营保障-链路识别

- ❑ 链路识别是指，在**二级、三级**安全生产保障团队的带领下，进行全集团的**核心业务链路的识别、治理**，贯穿**总部+分子公司**应用，从而提升运维人员对业务流程、系统架构及系统调用关系了解，进一步提升运维效率，减少故障分析时间，提前识别系统隐患，进而提升系统稳定性。
- **跨系统核心业务链路识别**：以核心业务服务入口、横向梳理业务流程、数据流程，梳理系统的上下游系统，包括不限于接口调用、异步消息、文件传输等，识别强依赖、弱依赖。
- **单系统架构部署识别**：核心业务链路中涉及到的系统、纵向梳理系统内部的网关、服务、中间件、数据库，主机、网络、机房、数据中心，统架构部署图，精确到基础设施，包括机房、网络设备等。
- **稳定性风险识别**：针对核心业务链路的关键点，识别关键点是否有监控、应急预案（业务开关、主备集群切换、多活、扩容等），架构是否合理等。



# 监控管理-全层级监控

系统需要依托全层级监控标准规范，接入数字化监控平台，实现从业务到触点、应用、中间件、云平台、基础资源的全层级指标数据监控覆盖。

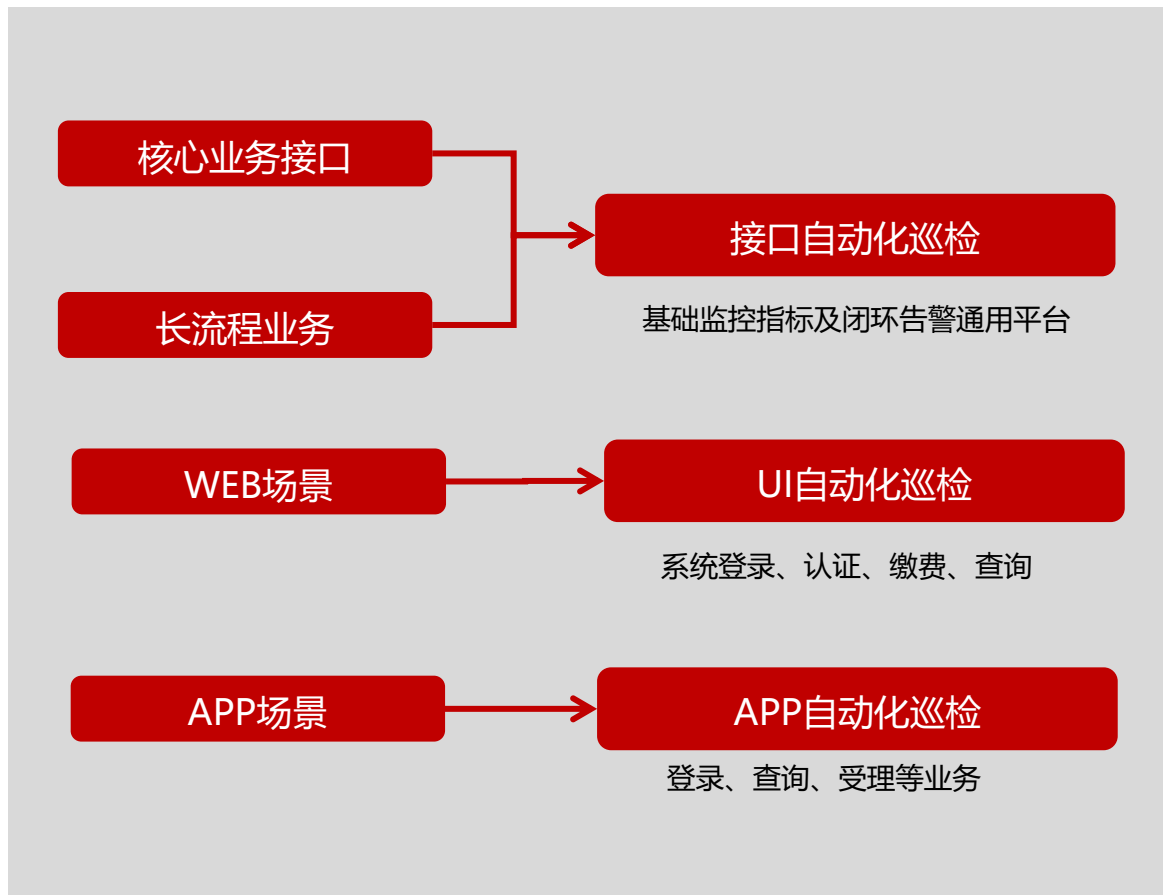


|                                                                                                 |               |
|-------------------------------------------------------------------------------------------------|---------------|
|  <b>业务</b>   | 用户发展量、工单积压量   |
|  <b>服务</b>   | 调用量、响应时间、TPS  |
|  <b>中间件</b>  | KONG、LB、MQ等   |
|  <b>数据库</b>  | RDS、ES、REIDS等 |
|  <b>容器</b>  | CPU、内存        |
|  <b>主机</b> | CPU、内存、磁盘、IO  |
|  <b>网络</b> | 域名流量、机房流量     |



# 监控管理-自动化巡检

- 模拟人类工作方式，将**核心业务流程实现自动化**，由机器或者软件替代人工完成**7x24不间断业务巡检**，发现问题出具报告。不同于监控的被动式发现，自动化巡检为**主动侦测防御**，对营业前的巡检或调用频次较低的接口有较为显著的作用、



## ➤ 制度规范

- 《中国联通软研院运营工作红线底线》;

## ➤ 注意要点:

- 营业开始前、变更操作后、业务高峰前开展自动化巡检

## ➤ 接口自动化巡检

- 覆盖范围：内外部核心、重要接口，长接口串联

## ➤ UI自动化巡检

- 覆盖范围：核心、重要场景电脑端WEB自动化巡检;

## ➤ APP自动化巡检

- 覆盖范围：APP登录、查询等核心场景

## ➤ 自动化巡检巡检频率

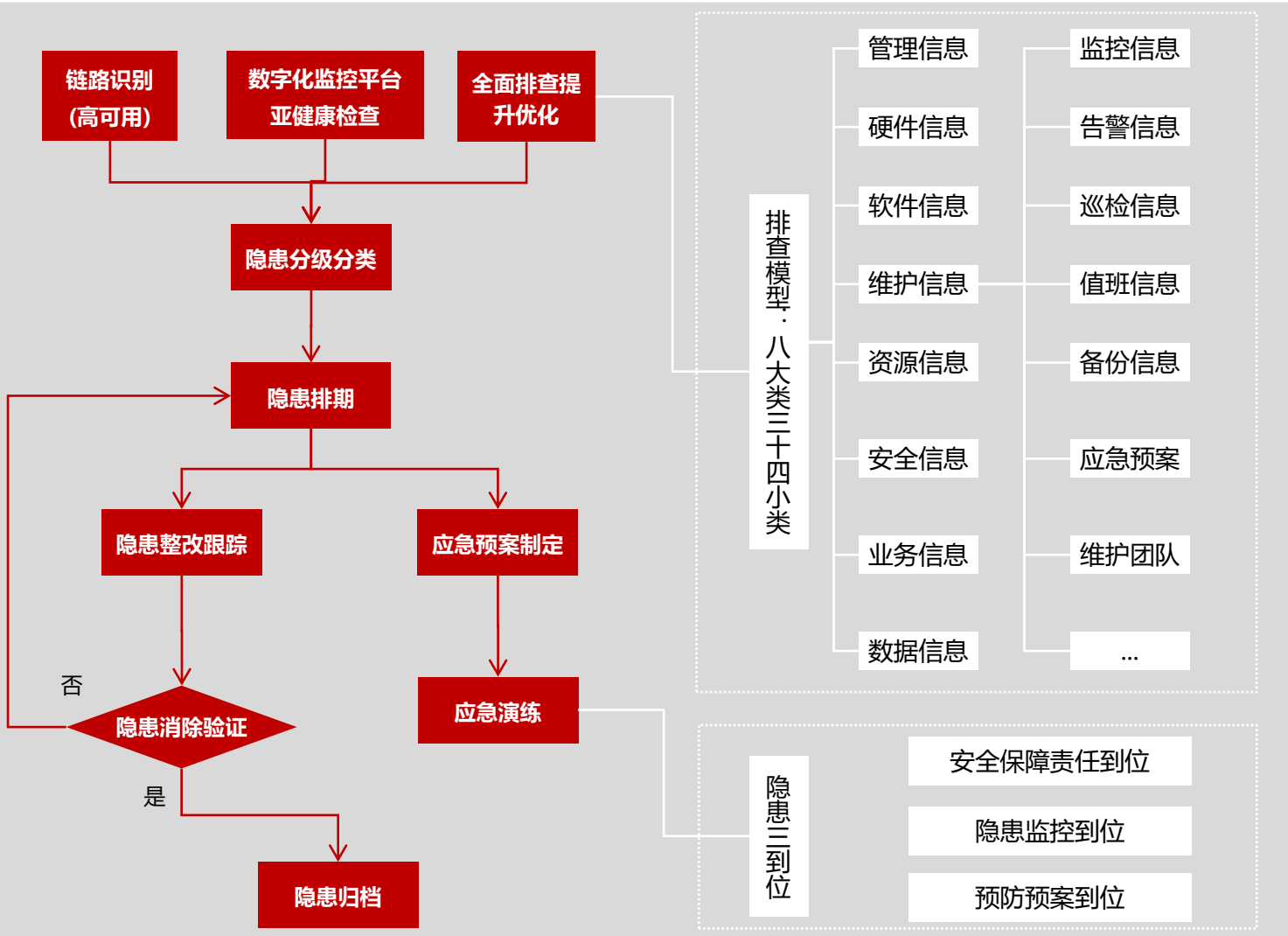
- 巡检频率：接口与UI巡检**5分/次**；APP巡检15分钟/次
- 巡检时段：最低标准为**早6点晚10点**，其他时段以实际需要为准。

## ➤ 自动化巡检巡检频率平台支撑

- 数字化监控平台-自动化巡检/APP巡检

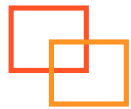
# 运营保障-隐患管理(1/2)

隐患管理是利用**运维大数据**与**专家排查**双重手段对故障主动出击进行全面**故障预防**的手段。



## 落地方法

- **制度规范**
  - ✓ 《红线底线条例》
- **注意要点**
  - ✓ **隐患发现来源：**
    - 链路识别高可用隐患 (人工)
    - 数字化监控平台-亚健康检测 (自动)
    - 全面排查提升优化 (人工)
  - ✓ **全面排查模型：**
    - 根据经验共制定**八大类，三十四小类**排查模型，涉及管理、硬件、软件、维护、资源、安全、业务、数据信息
  - ✓ **亚健康监测：**
    - 服务侧：对接口耗时长、慢SQL、系统报错重点优化
    - 组件、资源层：对高资源使用率重点优化
  - ✓ **隐患“三到位”要求：**
    - 安全保障责任到位
    - 隐患监控到位
    - 预防预案到位
- **平台工具**
  - ✓ 数字化监控平台 - 健康检查 - 故障预防：
  - ✓ 数字化监控平台 - 故障管理 - 隐患管理；



# 运营保障-隐患管理(2/2)

## 亚健康管理实践



### 周期性观测

故障预防统计以日、周、月维度统计问题项情况，观测系统阶段性运行情况



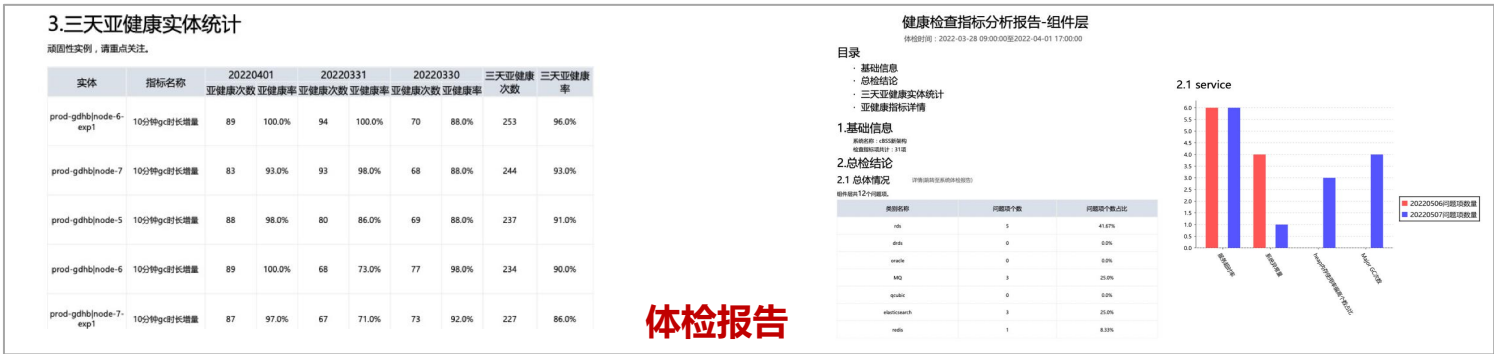
### 实时健康体检

系统实时体检实时计算全层级指标，根据阈值判断指标异常及风险程度



### 性能对比

系统性能对比页面可选取发版前后时间进行各指标性能对比，观测系统性能变化趋势



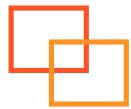
### 体检报告

系统体检与性能对比报告，找出系统异常指标标注指标含义、可能引起故障、整改举措，助力系统整改，夯实稳定性



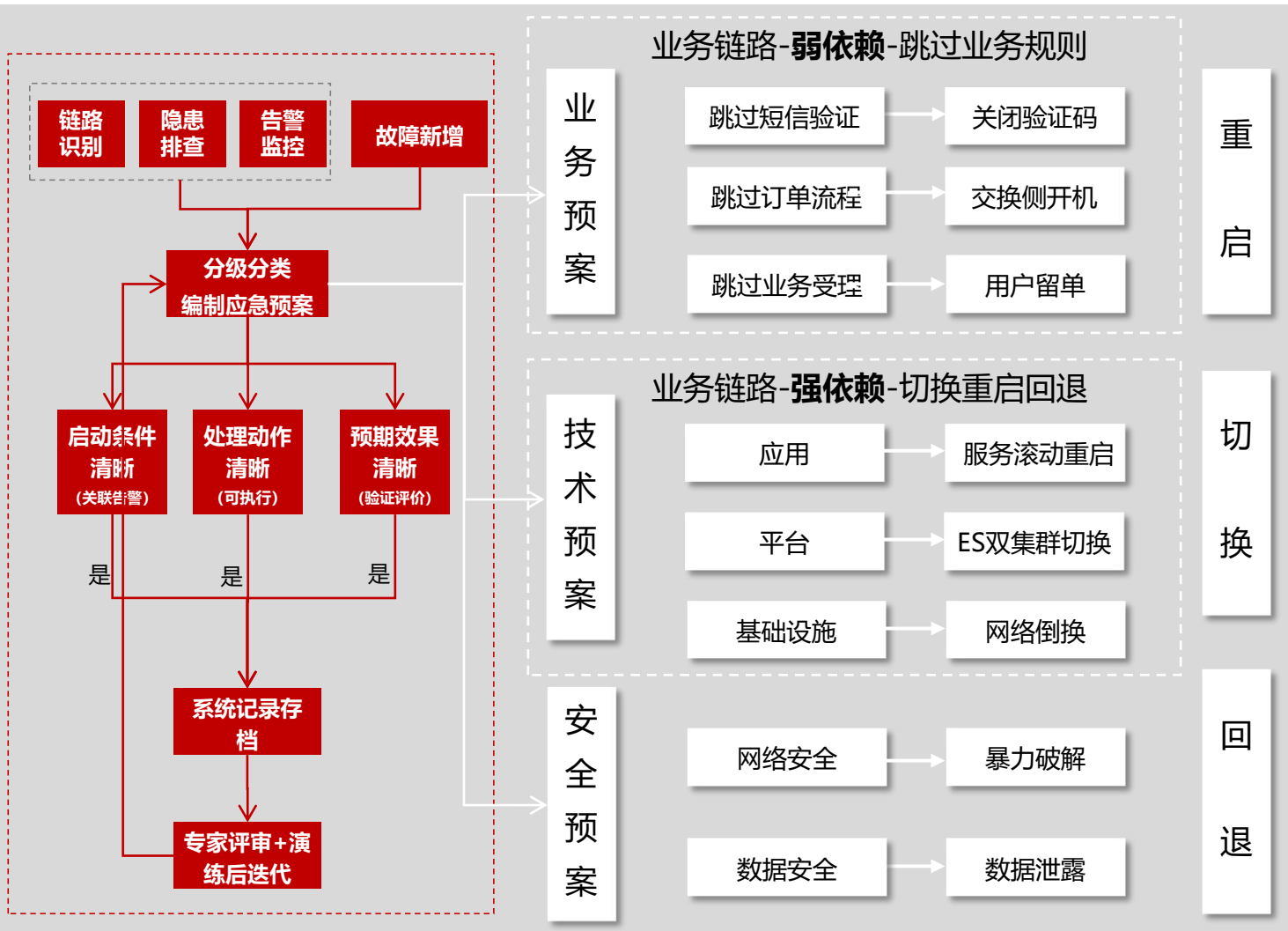
### 黄金核心指标选择

根据故障知识库与专家建议，选取页面、服务、组件、资源层核心黄金指标



# 运营保障-应急预案

□ 应急预案方面，重点关注**预案覆盖来源、应急预案分类、应急预案书写要点**三部分内容。



## 落地方法

### 制度规范

- ✓ 《中国联通大IT生产运营规范-07故障及应急管理分册 (V2.0) 》
- ✓ 《中国联通大IT生产运营奖惩办法》
- ✓ 《红线底线条例》 - 系统上线前需完成应急预案的制定工作

### 注意要点

#### 应急预案来源:

- ① **正向:** 链路识别强弱依赖 + 尚未解决隐患 + 关键监控告警
- ② **反向:** 发生的故障均需有应急预案

#### 应急预案分类 (三板斧: 重启、切换、回退):

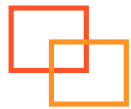
- ① **业务预案:** 跳过当前业务规则
- ② **技术预案:** 通过技术手段
- ③ **安全预案:** 解决安全类问题

#### 应急预案书写要点:

- ① 通俗易懂, **可在生产环境执行**
- ② 启动条件清晰, **关联监控告警**
- ③ 效果清晰明确
- ④ 预案负责人对**全生命周期的质量进行负责**

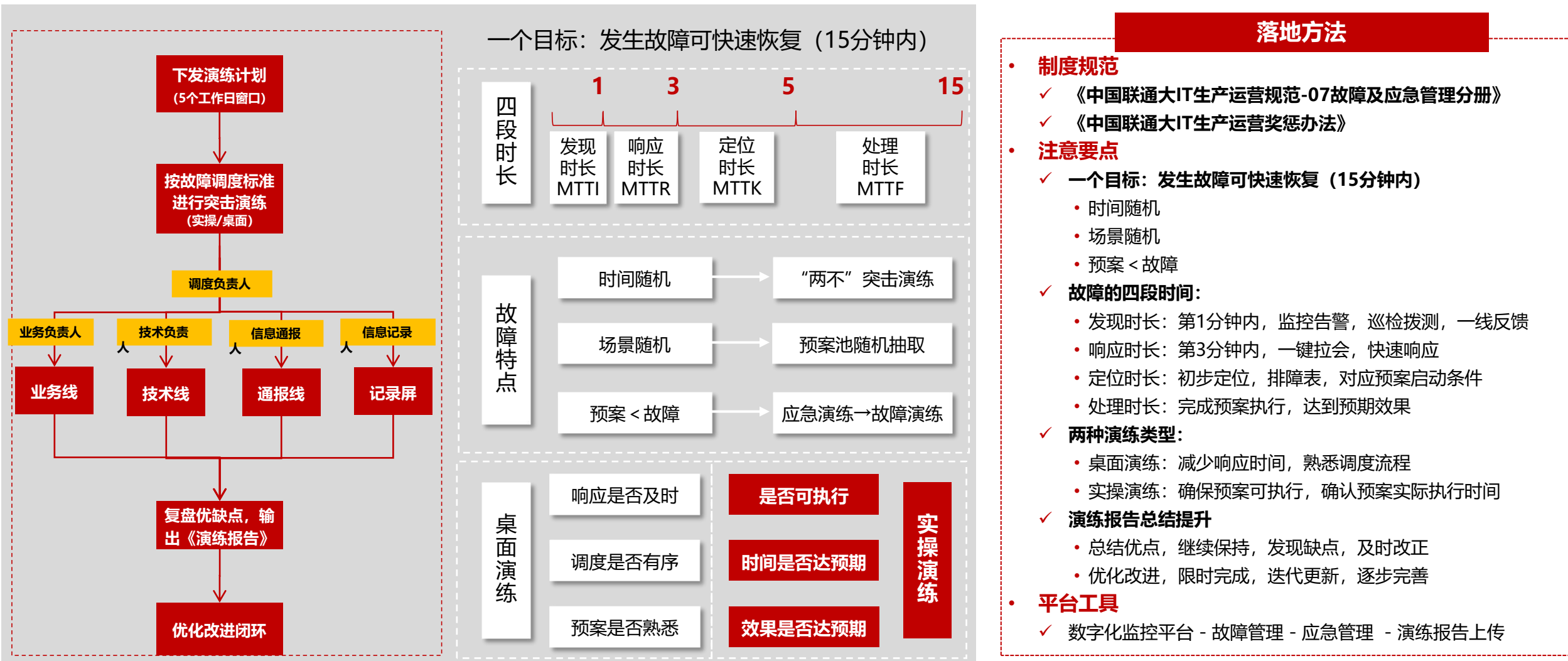
### 平台工具

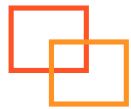
- ✓ 数字化监控平台 - 故障管理 - 应急管理 - 应急预案看板;



# 运营保障-应急演练

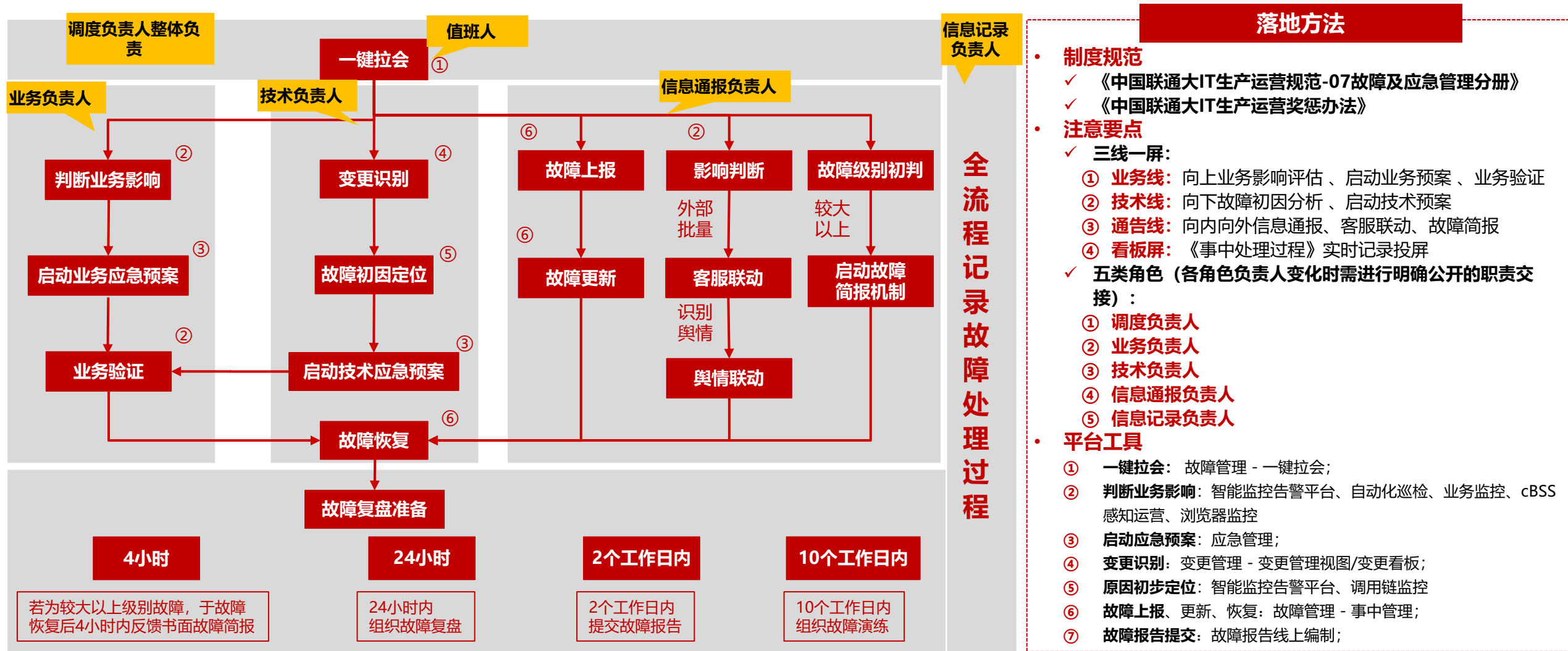
应急演练方面，围绕故障快速恢复一个目标，重点关注故障历时的四段时长，通过两类演练进行压降





# 运营保障-故障调度(1/2)

数字化系统故障发生后遵循“先抢通、后抢修”原则，重点关注“三线一屏”分析法及起对应的五类角色职责





# 运营保障-故障调度(2/2)

## 信息上报、过程看板



**【故障上报】**

信息系统名称: cBSS

信息系统级别: 核心

故障发现时间: 2022-10-01 10:00

故障发生时间: 2022-10-01 10:00

影响业务类型: 部分营业终端办理业务时出现异常弹窗, 不影响核心业务

影响地域范围: 全国

原因初步判断: 通过告警发现告警弹窗增多

故障处理进展: 启动应急预案恢复。

现场负责人: 岩



**【故障恢复】**

信息系统名称: cBSS

信息系统级别: 核心

故障发现时间: 2022-10-01 10:00

故障发生时间: 2022-10-01 10:00

故障恢复时间: 2022-10-01 10:00

影响业务类型: 部分营业终端办理业务时出现异常弹窗, 不影响核心业务

影响地域范围: 全国

原因初步判断: 通过告警发现告警弹窗增多

故障处理进展: 启动应急预案下发全国应急脚本恢复。

### 事中过程看板

**一、过程看板**

调度负责人: XXX; 00:00转交XXX; “三线一屏”分析法

业务负责人: XXX; 业务线向上进行业务影响评估

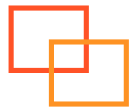
技术负责人: XXX; 技术线向下进行故障初因分析

信息通报负责人: XXX; 通告线向内向外进行信息通报

信息记录负责人: XXX; 一屏进行《事中过程看板》的实时记录投屏

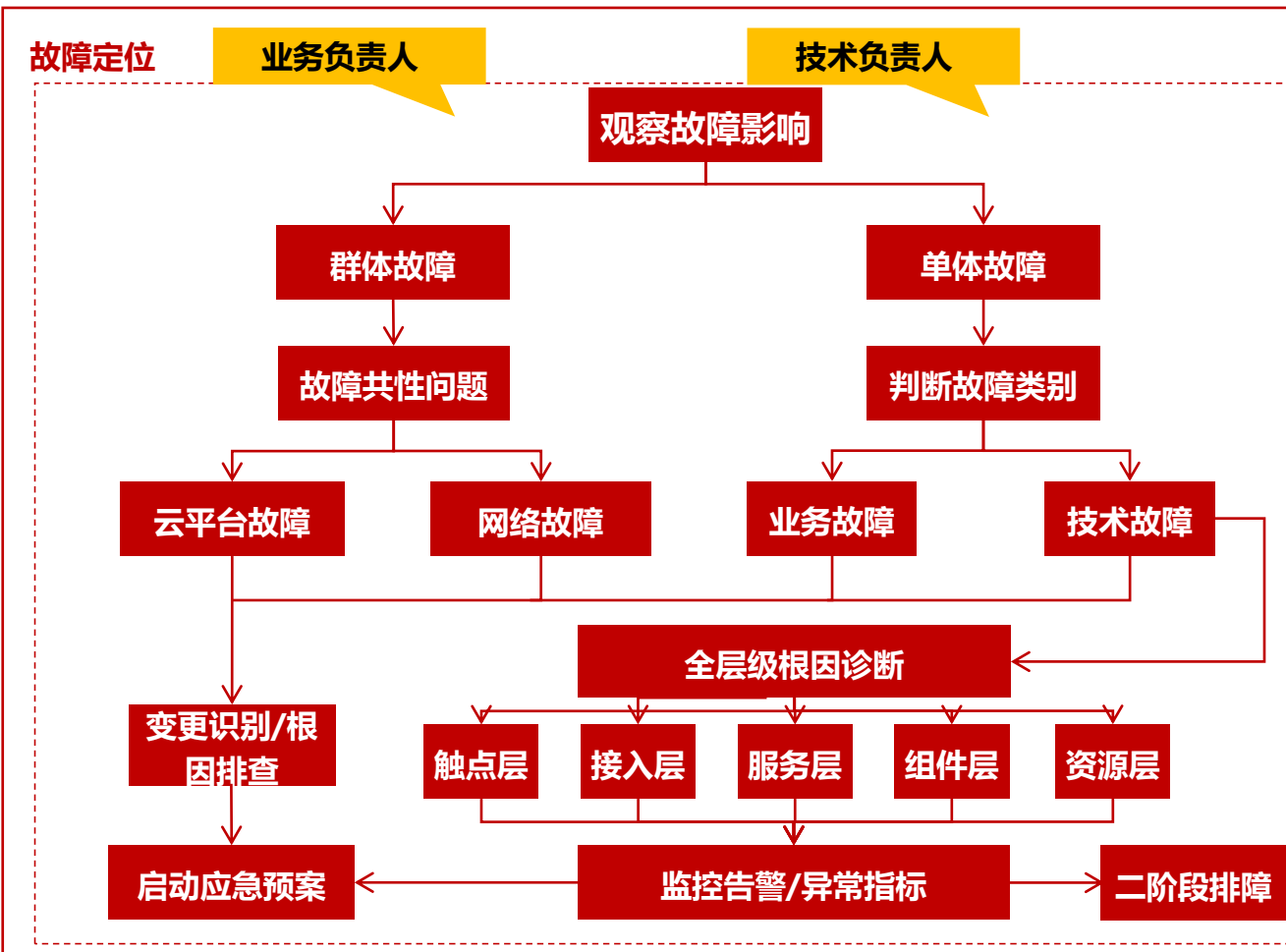
业务影响: XXX

| 时间    | 处理动作         | 处理方/人 | 预计时间 |
|-------|--------------|-------|------|
| 00:00 | 发现 (监控告警)    | XXX   | X分钟  |
|       | 近期是否存在变更     |       |      |
|       | 初因定位 (转交调度人) |       |      |
|       | 启动执行《排障检查清单》 |       |      |
|       | 处理 (启动应急预案)  |       |      |
|       | 业务验证         |       |      |
|       | 更新/恢复        |       |      |



# 运营保障-故障定位 (1/3)

❑ 五类角色第四、五类：**业务负责人、技术负责人**，建议由三/四级团队担任，依据故障定位流程**观影响、定分类、查变更、有序层级排查快速定位**



## 落地方法

### ① 观影响，判断故障影响范围：

- ✓ 群体故障：以**数据中心维度**观测，寻找共性切入点，多为**云平台**或**网络**故障影响导致
- ✓ 单体故障：单系统故障，变更识别与全层级根因诊断。

### ② 定分类，判定故障类别：

- ✓ 业务故障：通过自动化巡检、业务监控、浏览器弹窗等判断是否为系统个别业务问题，重点关注变更，多为**应用软件逻辑**变更或**数据库参数表**变更导致。
- ✓ 技术故障：全层级监控告警判断，全层级**变更识别**与**根因诊断**

### ③ 查变更，识别变更进行回滚：

- ✓ 变更识别：根据关联性排查，若为变更引发，则进行**回滚**

### ④ 有序层级排查，各层级协同处理：

- ✓ 全层级根因诊断：触点层、接入层、服务层、组件层、资源层的**全层级有序排查**
- ✓ 三四级团队协同：各层级**协同排查处理**，根据情况进行**双向协同调度**。

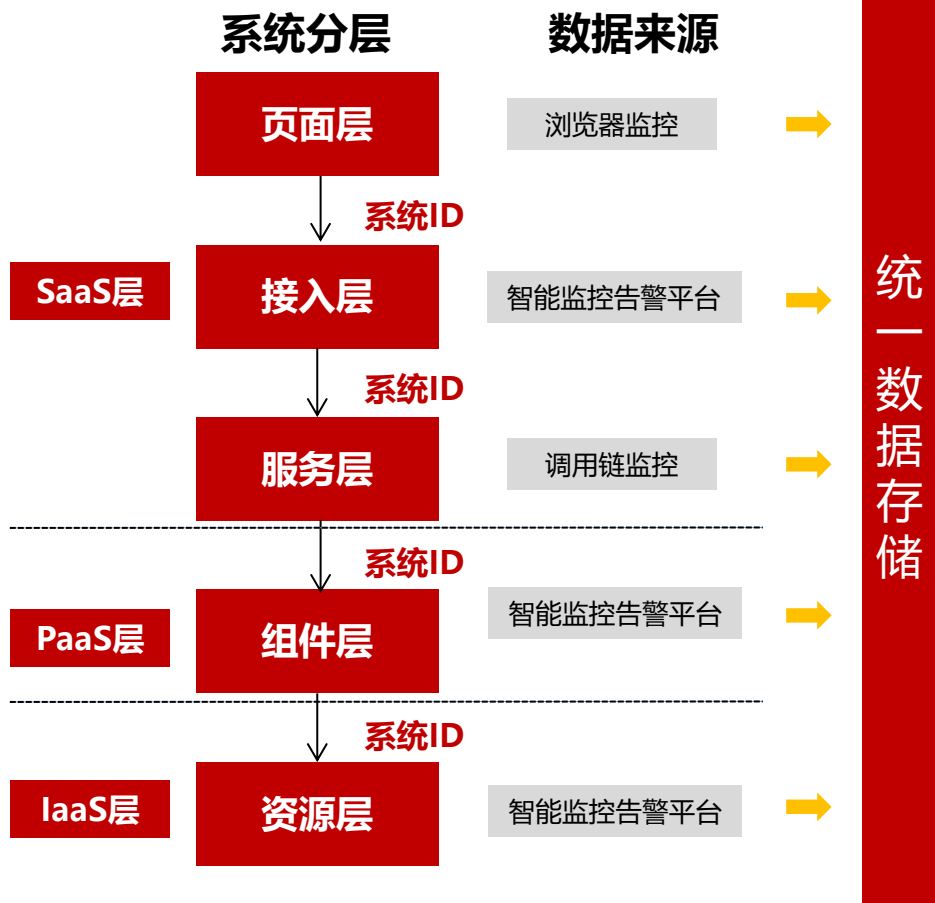


# 运营保障-故障定位 (2/3)

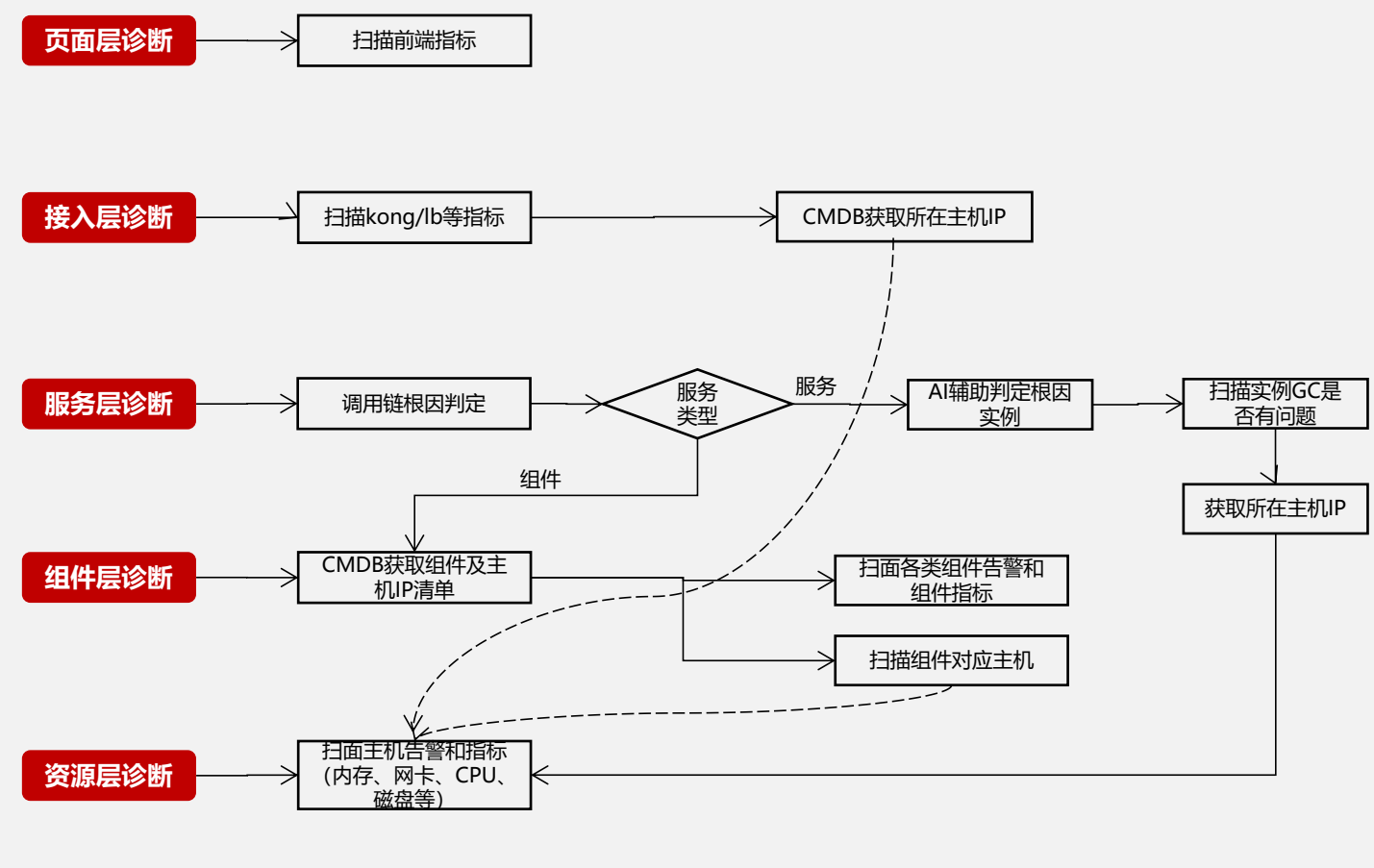
□ 依托全层级指标数据，全层级链路调用，自研云原生CMDB，以级丰富的故障知识库，以服务层为故障起点进行纵向串联，配以规则+AI的能力实现全层级一键智能故障诊断，当前准确率可以达到55%。

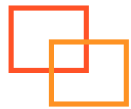
□ 关键点：耐着性子一个组件一个组件弄，一个故障场景一个故障场景弄，AI能力是辅助，故障库和专家是关键财富

## 全层级指标拉通



## 一键故障诊断



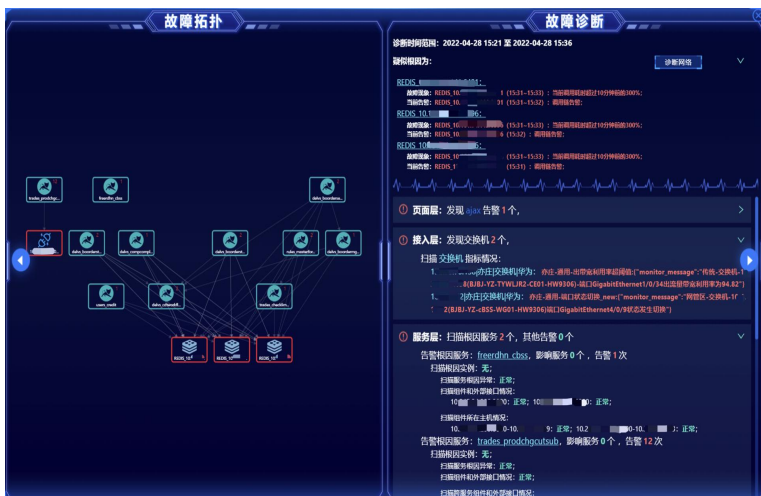


# 运营保障-故障定位 (3/3)

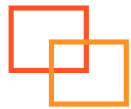
**服务实例事件异常:**  
2022-04-23 11:XX~11:XX  
根因服务实例耗时突增  
判断实例GC是引发故障根因



**ES事件异常:**  
2022-5-22 10:XX~10:XX  
es进程负载率突增导致  
上游服务连接超时是故障根因

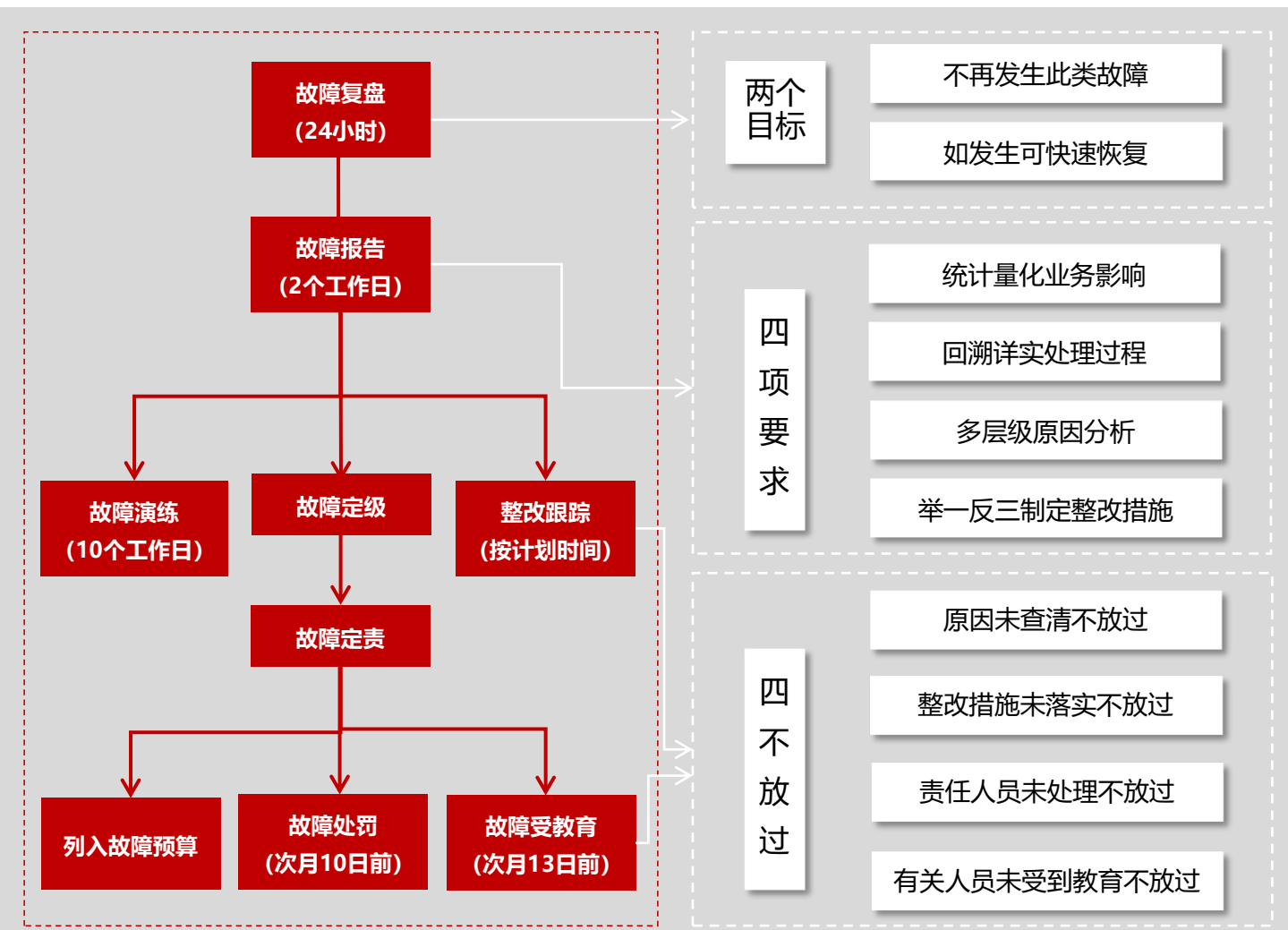


**网络事件异常:**  
2022-03-02 16:XX~17:XX  
交换机带宽使用率指标  
打满引起访问受限是故障根因



# 运营保障-故障闭环

故障复盘方面，重点关注故障复盘的**两个目标**、故障报告的**四项要求**、故障闭环的**四不放过**三部分内容



## 落地方法

### 制度规范

- ✓ 《中国联通大IT生产运营规范-07故障及应急管理分册》
- ✓ 《中国联通大IT生产运营奖惩办法》

### 注意要点

- ✓ **时间要求:** 24小时复盘, 2个工作日内提交报告, 10个工作日故障演练;
- ✓ **复盘质量要求:**
  - **两个目标:** 1、不再发生此类故障; 2、如发生可快速恢复;
  - **统计量化业务影响:** 影响时间、范围、渠道、业务量等
  - **回溯详实处理过程:** “五个时间点”(发现时间, 响应时间, 上报时间, 定位时间, 处理时间);
  - **多层次原因分析:** 多层次原因分析, 列名导致故障发生或未快速恢复的多个原因
  - **举一反三制定整改措施:** “五定原则”(定目标(可量化)、定标准(可检查)、定负责人(谁牵头、谁配合)、定计划完成时间、定里程碑)
- ✓ **整改质量要求:** 注重举措成效, 对结果负责
- ✓ **四不放过:** 原因未查清不放过、责任人员未处理不放过、整改措施未落实不放过、有关人员未受到教育不放过

### 平台工具

- ✓ 数字化监控平台 - 故障管理 - 故障报告编写审核、整改工单发起跟踪