促进软件开发领域知识与创新的传播

InfoQueue

ArchSummit

大型商业银行故障处理实践
应急标准化方法论

张春林

2015年7月18日

大纲

ArchSummit

- 黑天鹅事件 —— 哲学思辩
- 应急处理的思考点 —— 理性分析
- 应急标准化方法论 —— 成果分享

- 黑天鹅事件，总是在我们自以为是的时刻发生。







- 从天灾到人祸
- 从基础设施损坏到人为操作风险

- 无论机构规模大小
- 无论IT成熟与否
- “互联网+”背后的危机

脆弱性理论

ArchSummit

- 金融体系的脆弱性理论告诉我们，金融危机的发生是由金融体系的脆弱性内生决定的。同理，信息科技风险也是由信息系统的脆弱性内生决定的。

- 硬盘的坏盘率
- 软件代码的BUG率
- 人员更替
- 以已知对抗未知

应对之策

ArchSummit

纳西姆·尼古拉斯·塔勒布在《黑天鹅》中给出的建议：

1、不要预测。

2、谨慎预防。

3、保持充足冗余。

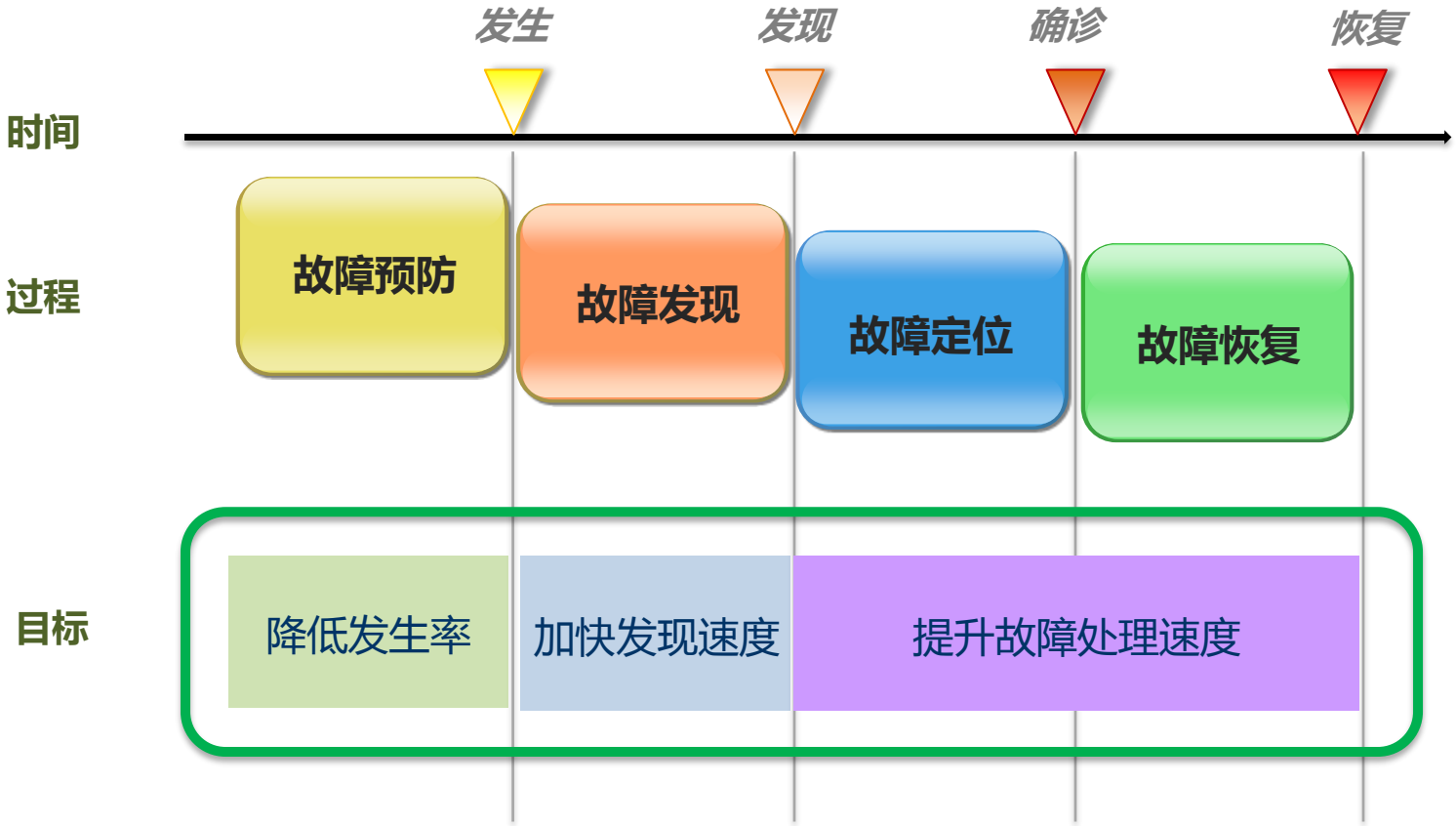
大纲

ArchSummit

- 黑天鹅事件 —— 哲学思辩
- 应急处理的思考点 —— 理性分析
- 应急标准化方法论 —— 成果分享

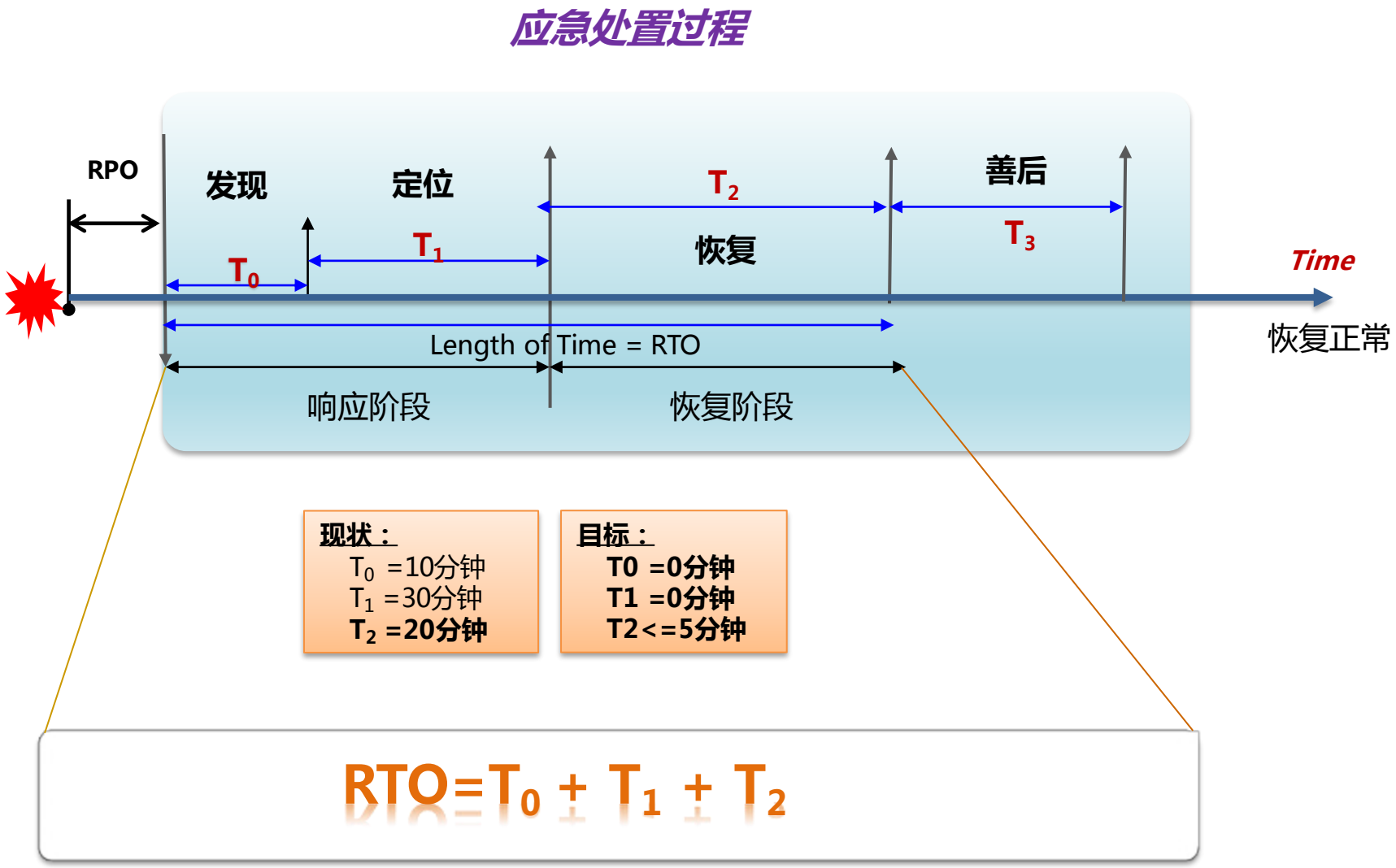
故障处理的思考点

- 故障恢复时长（RTO）是衡量一个业务连续性的关键指标，它的长短决定了业务影响。
- 从故障发生开始到故障恢复截止，整个过程包括发现、定位、恢复三阶段。



应急效率提升——RTO分析

ArchSummit



应急效率提升——现状分析

分析方法：

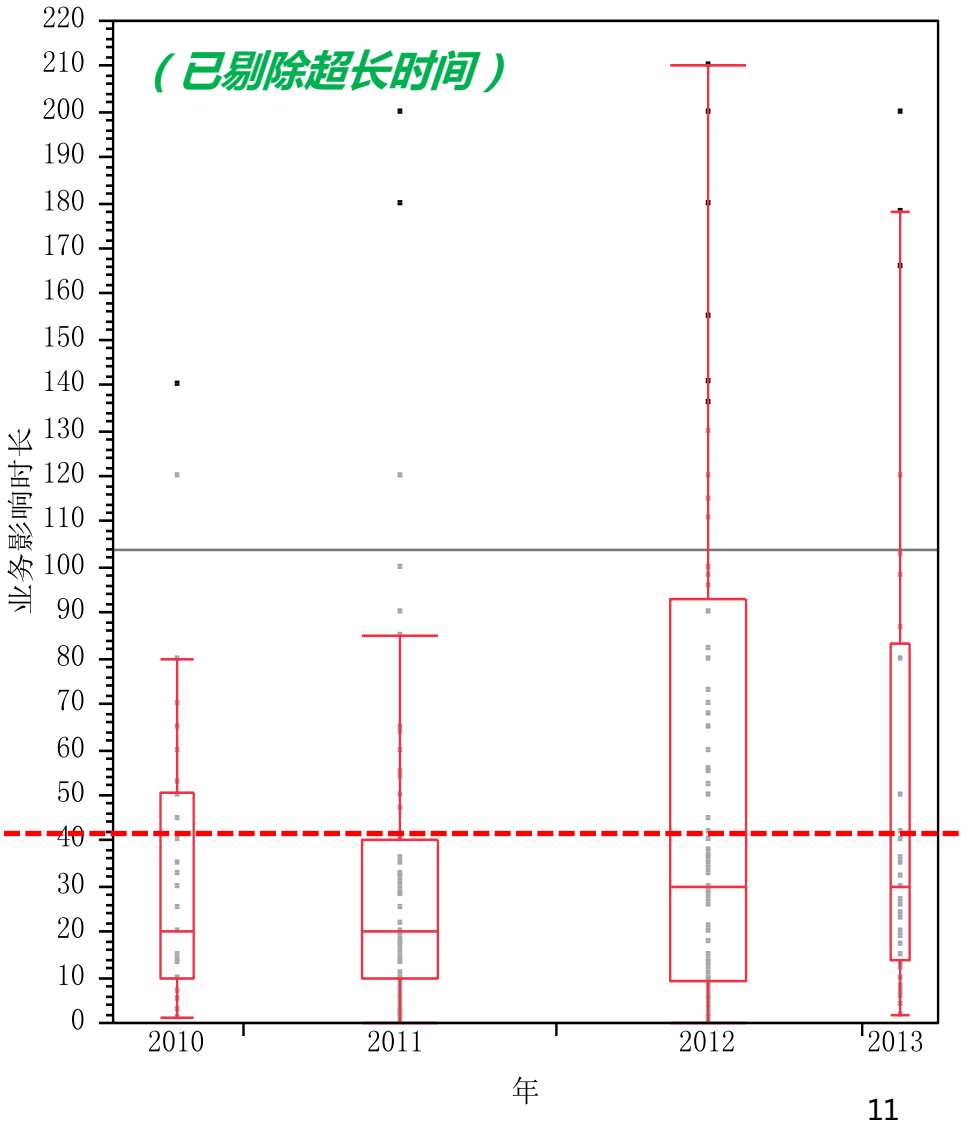
精益六西格玛

现状概览：

1、很多故障的处理时长超过30分钟；

2、很多故障无法通过切换来解决，平均每月2起。

大量的事件处理时长超过30分钟



应急效率提升——VOC/P/B

ArchSummit

VOC (客户之声)

1. 出现故障后不知道如何处理，找不到适当的应急预案
2. 不知道找谁处理
3. 现场人员混乱

VOP (流程之声)

1. 缺少通用故障处理流程，过于依赖A角；
2. 演练不到位，处理不够熟练
3. 应急环境（ECC、工具）缺少规范化管理

VOB （管理之声）

1. 应急通讯手段落后
2. 无统一指挥和标准化流程
3. 30分钟以上故障必须报银监会

- 搜集用户的声音、流程的问题、管理者的声音
- VOC-Voice Of Customer；VOP-Voice Of Procces；VOB-Voice Of Business

应急效率提升——改进评估标准

ArchSummit

Y=F（X）

Y：故障恢复时长

定义：故障恢复时长RTO，
故障开始时间至故障恢复时间。

测量：提取2010年至今的所有故障影响时长，取平均值。

X1：工具保障操作一致化

定义：**具备**电子化**工具**，包括监控工具、恢复工具和通讯工具

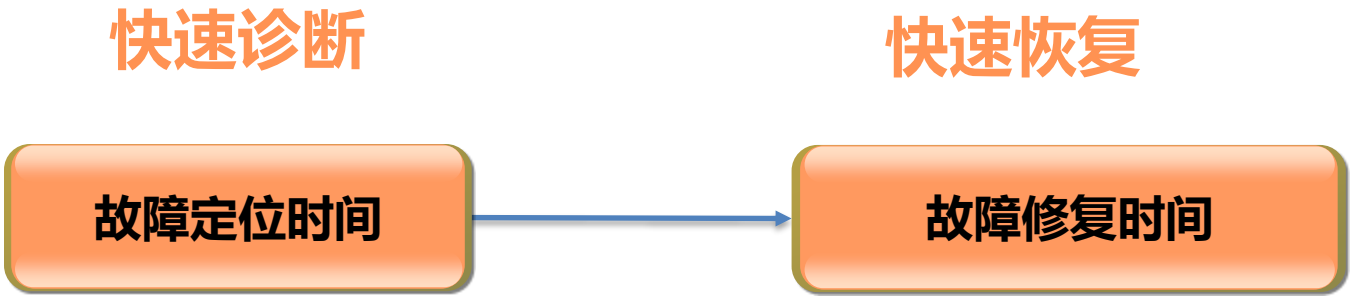
测量：通过多个项目同步建设工具并联动，实现操作一致化。

X2：流程的范围与熟练度

定义：**具备流程**，包括通用应急指挥流程和通用故障处理流程

测量：通过各种不定时演练抽查**提高熟练度**

应急效率提升——具体改进措施



- 应急预案
- 监控平台
- 应急平台

- 故障自愈进程
- 一键恢复工具
- 业务补账工具
- 流量清洗服务

大纲

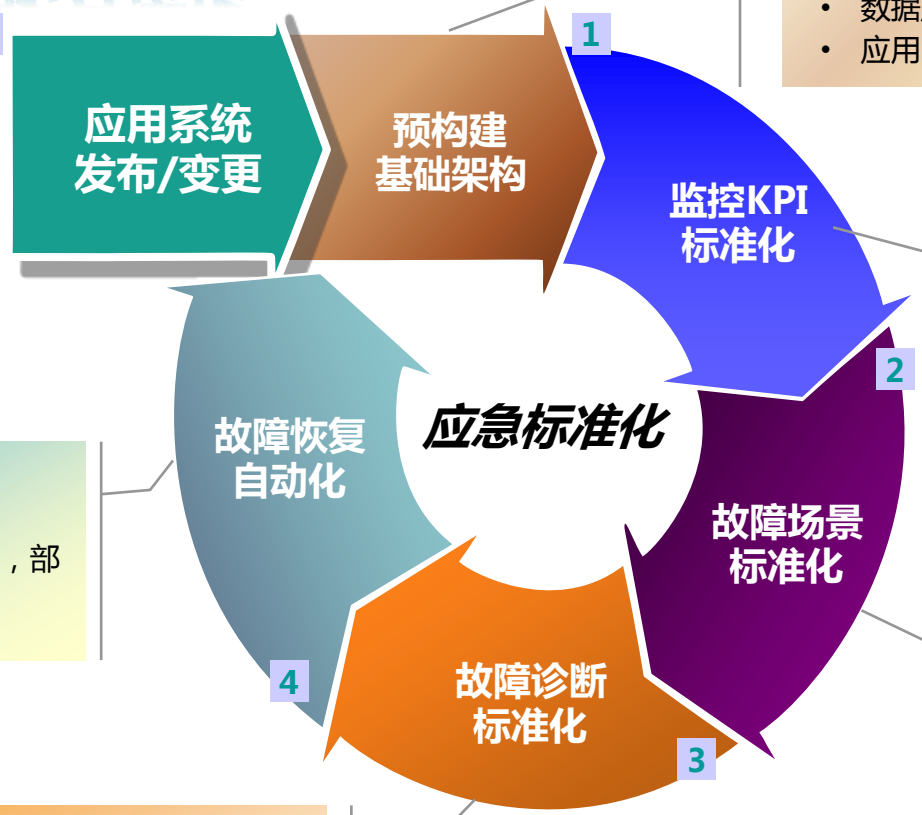
ArchSummit

- 黑天鹅事件 —— 哲学思辩
- 应急处理的思考点 —— 理性分析
- 应急标准化方法论 —— 成果分享

第三部分：成果分享

ArchSummit

应急标准化方法论



架构标准化
- 网络　—双设备+双机房+双中心
- 服务器 —负载均衡/HA
- 数据库 —HA+HADR/RAC+DataGuard
- 应用　—负载均衡/HA

KPI标准化
- 小型机监控KPI
- WIN监控KPI
- DB监控KPI
- 应用监控KPI
- 网络监控KPI

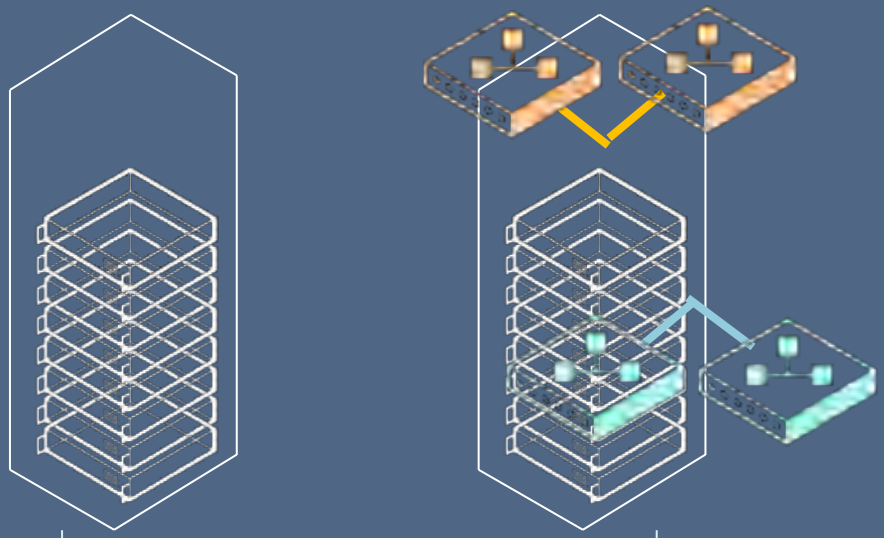
场景标准化
- **KPI组合，定义故障场景**
- **基础架构故障场景标准化**
- 应用故障场景部分标准化，特色部分单独梳理

恢复标准化
- 梳理快速处置策略，部署快速恢复脚本

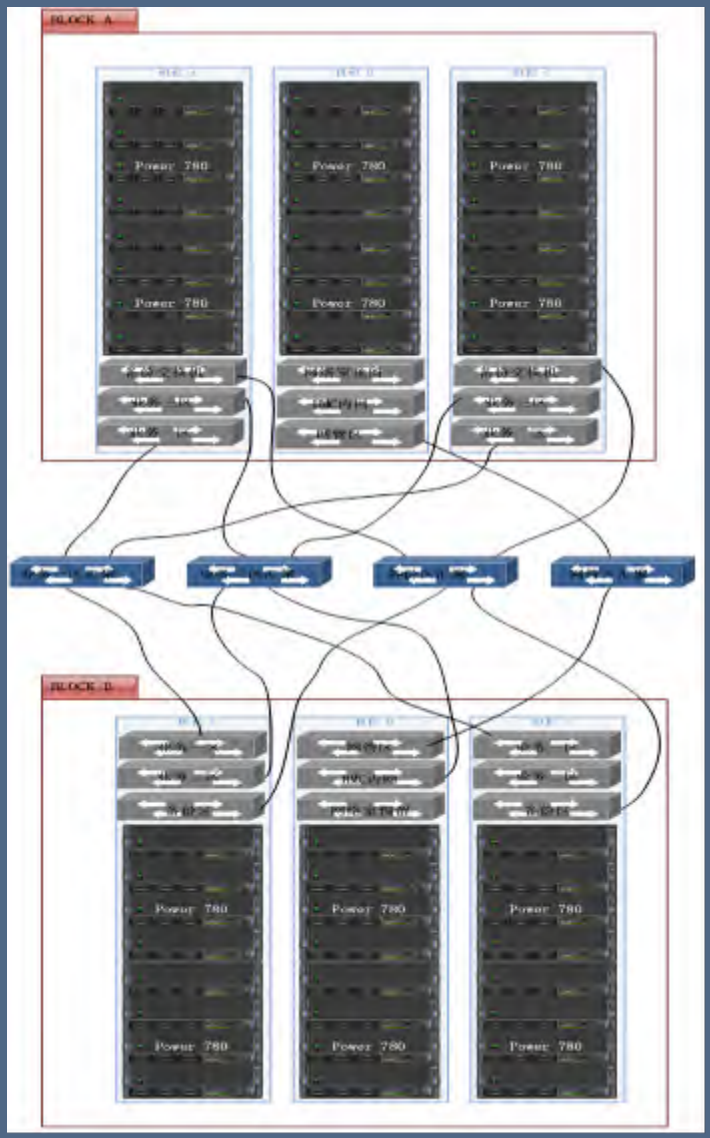
诊断标准化
- 标准化诊断流程（诊断树）
- 定时抽取监控告警，自动匹配场景
- 一键诊断脚本辅助确诊

架构标准化

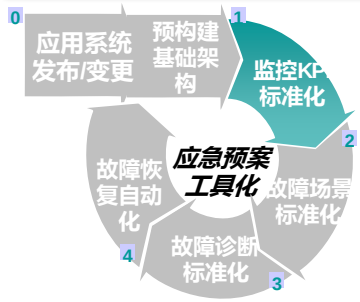
- 标准化硬件配置和部署方式
- 网络双路双区接入
- 标准化同时考虑高可用性
- 组建服务器集群



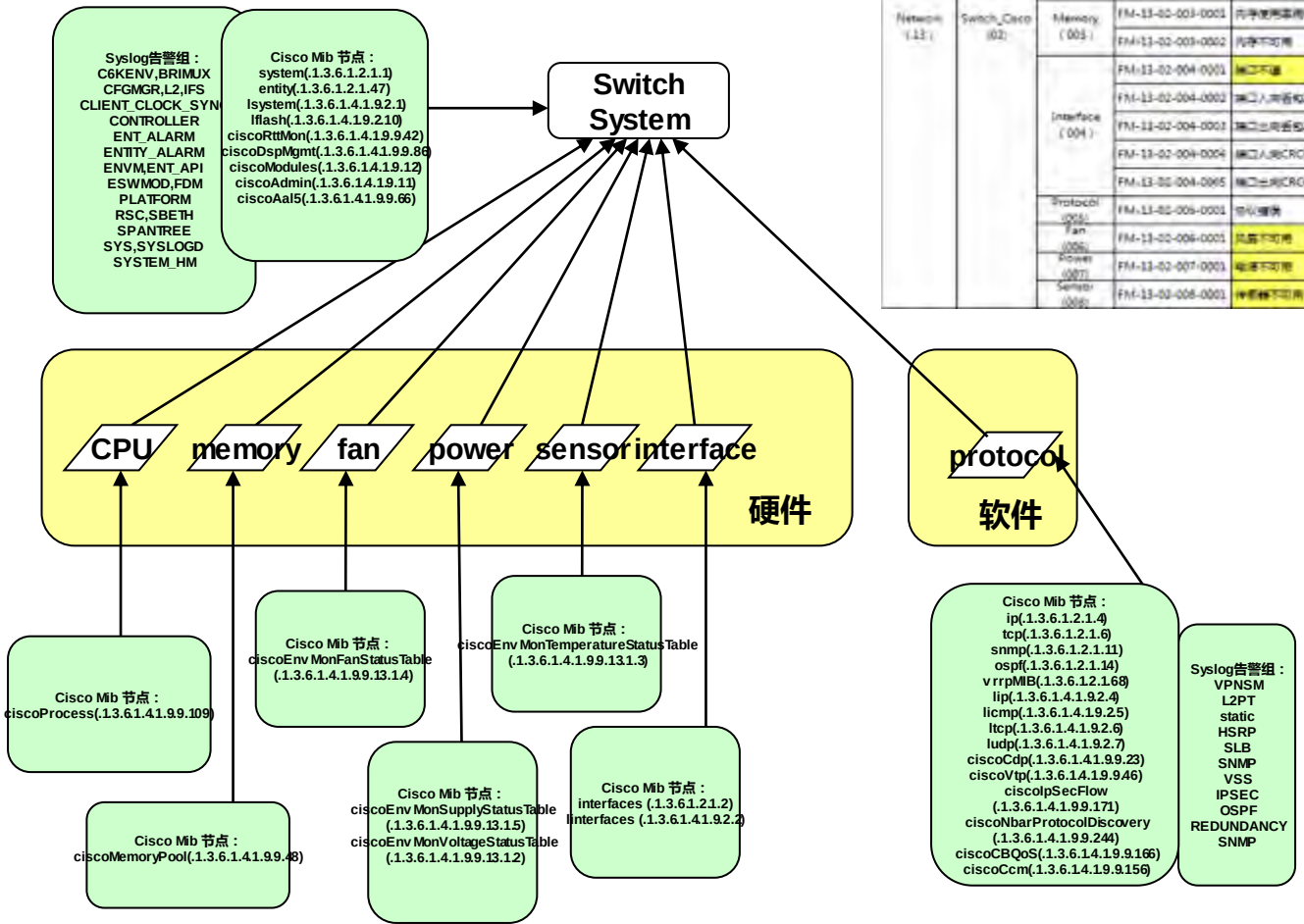
- ✓ 预构建、可伸缩的基础架构单元
- ✓ 一组计算、存储和网络资源



监控指标标准化



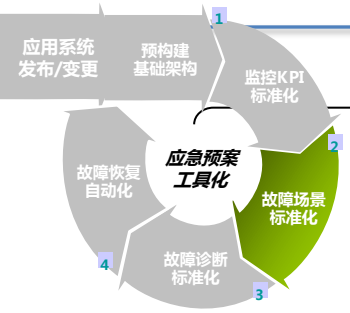
Network (11)	Switch_Cisco (02)	System (001)	FM-11-02-001-0001		RD-11-02-001-0001			
			FM-11-02-001-0002		RD-11-02-001-0002			
			FM-11-02-001-0003		RD-11-02-001-0003			
			FM-11-02-001-0004		RD-11-02-001-0004			
			FM-11-02-001-0005		RD-11-02-001-0005			
			FM-11-02-001-0006		RD-11-02-001-0006			
			FM-11-02-001-0007		RD-11-02-001-0007			
		CPU (002)	FM-11-02-002-0001	CPU使用率高告警	RD-11-02-002-0001			
			FM-11-02-002-0002	CPU不可用	RD-11-02-002-0002			
		Memory (003)	FM-11-02-003-0001	内存使用率高告警	RD-11-02-003-0001			
			FM-11-02-003-0002	内存不可用	RD-11-02-003-0002			
		Interface (004)	FM-11-02-004-0001	端口不通	RD-11-02-004-0001	端口ping不通	端口ping不通	5分钟
			FM-11-02-004-0002	端口入向丢包率高告警	RD-11-02-004-0006		>=10%	5分钟
			FM-11-02-004-0003	端口出向丢包率高告警	RD-11-02-004-0007		>=10%	5分钟
			FM-11-02-004-0004	端口入向CRC错包率高告警	RD-11-02-004-0008		>=10%	5分钟
			FM-11-02-004-0005	端口出向CRC错包率高告警	RD-11-02-004-0009		>=10%	5分钟
		Protocol (005)	FM-11-02-005-0001	协议错误	RD-11-02-005-0001	可用性Syslog		即时
		Fan (006)	FM-11-02-006-0001	风扇不可用	RD-11-02-006-0001	风扇不可用		即时
		Power (007)	FM-11-02-007-0001	电源不可用	RD-11-02-007-0001	电源不可用		即时
		Sensor (008)	FM-11-02-008-0001	传感器不可用	RD-11-02-008-0001	温度传感器不可用		即时



监控KPI标准库

监控对象组件模型

故障场景标准化——标准故障场景库



故障场景分类体系

- 应用系统
  - 小型机
    - 应用服务器
      - 服务器
        - 服务器硬件故障
        - 操作系统故障
        - 心跳中断
      - 服务器集群
        - 进程故障
        - 仲裁盘故障
        - 应用进程故障
      - 业务应用
        - 应用逻辑故障
      - 性能
      - 存储
    - • • •
  - Windows
    - 应用服务器
      - 服务器
      - 服务器集群
      - 业务数据库
        - DB连接故障
        - DB引擎故障
        - 数据库故障
        - 数据库逻辑故障
      - 性能
      - 存储
    - • • •
  - DB
    - 数据库服务器
    - • • •
  - Network
    - 外联线路
    - 防火墙
    - 负载均衡

标准场景库

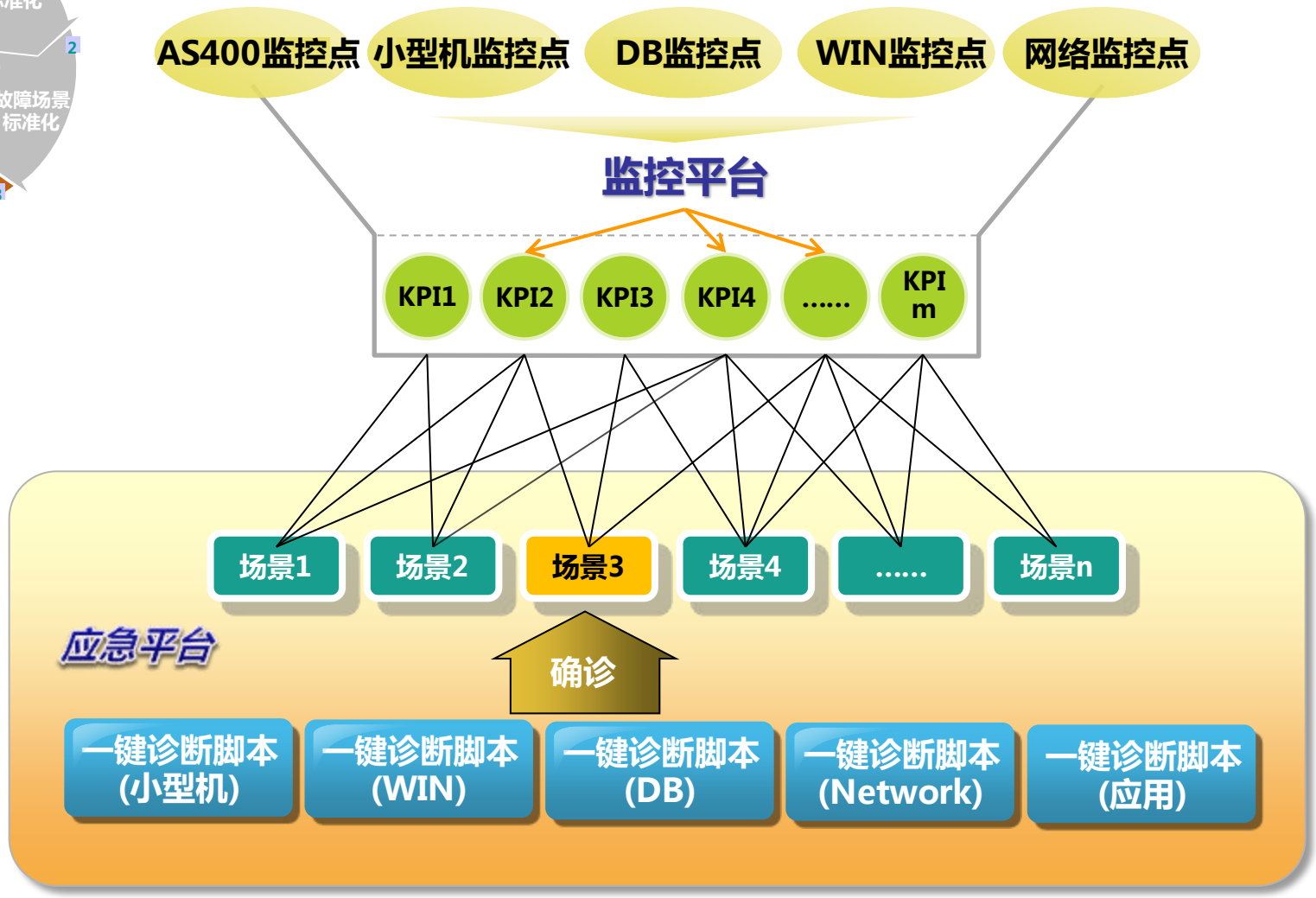
- 主机Down
- 应用进程异常
- 主机HA异常
- 服务端口异常
- • • • • • •
- 网上支付短信验证码队列堵塞
- 二代支付MQ应用进程某个通道异常
- 二代支付MQ应用进程某个队列深度异常
- o o o o o o
- 某应用系统清算异常
- o o o o o o
- 运营商线路异常或者对端异常
- 外联路由器端口异常
- 外联路由设备异常
- 防火墙策略有异常
- 防火墙HA状态异常
- 应用状态异常
- 负载均衡设备/HA/路由异常

基础架构场景标准化

应用故障场景部分标准化，特色部分单独梳理

诊断标准化——监控告警**自动定位**故障场景

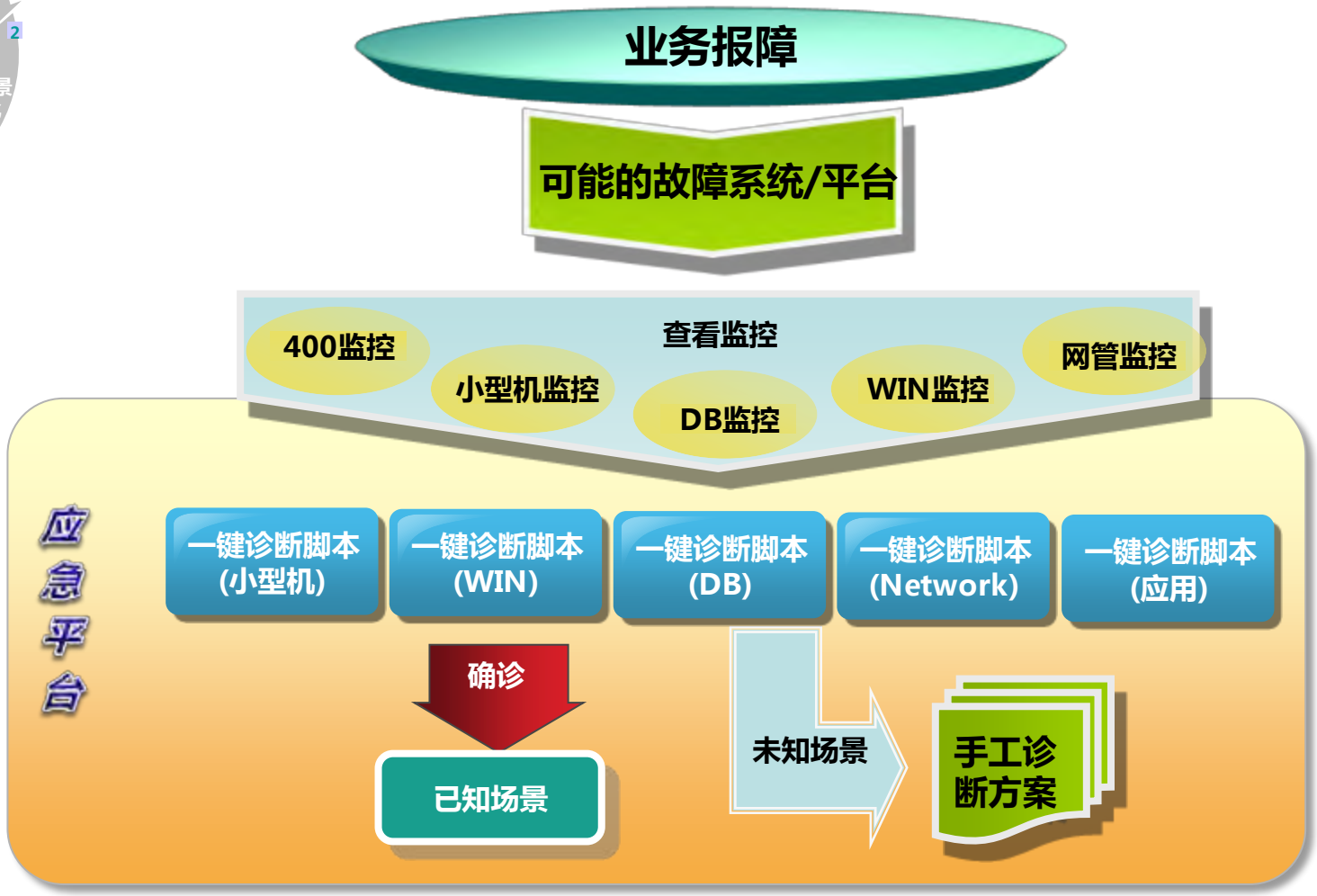
- 自动诊断
- 一键诊断
- 通用流程
- 专家诊断



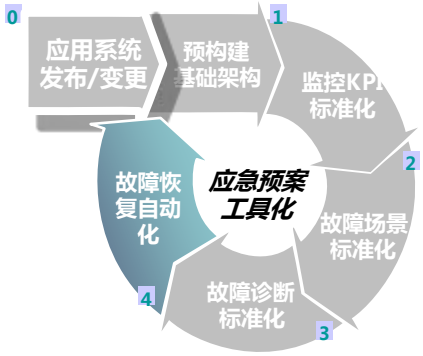
诊断标准化——一键诊断脚本及手工诊断方案

ArchSummit

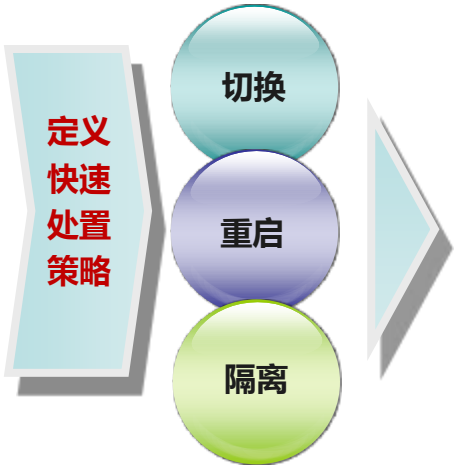
- 自动诊断
- 一键诊断
- 通用流程
- 专家诊断

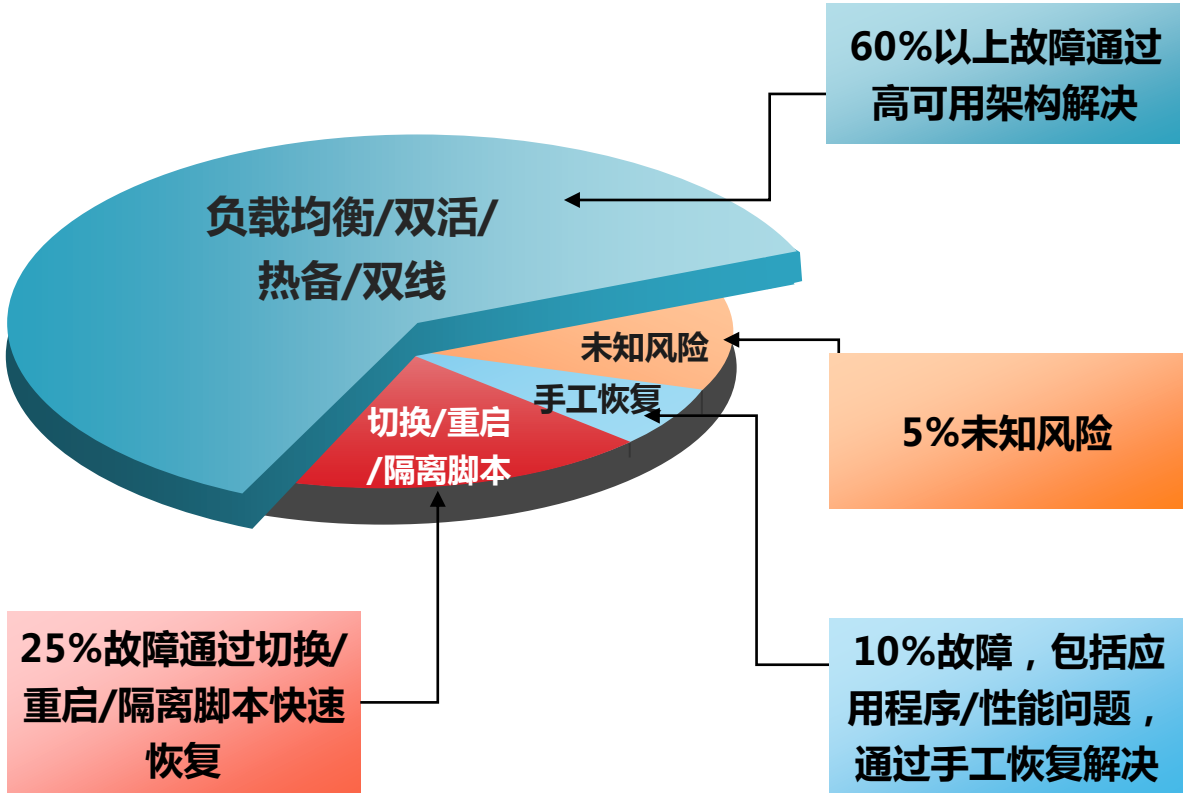


恢复标准化



• 故障自愈 • 一键恢复 • 手工恢复





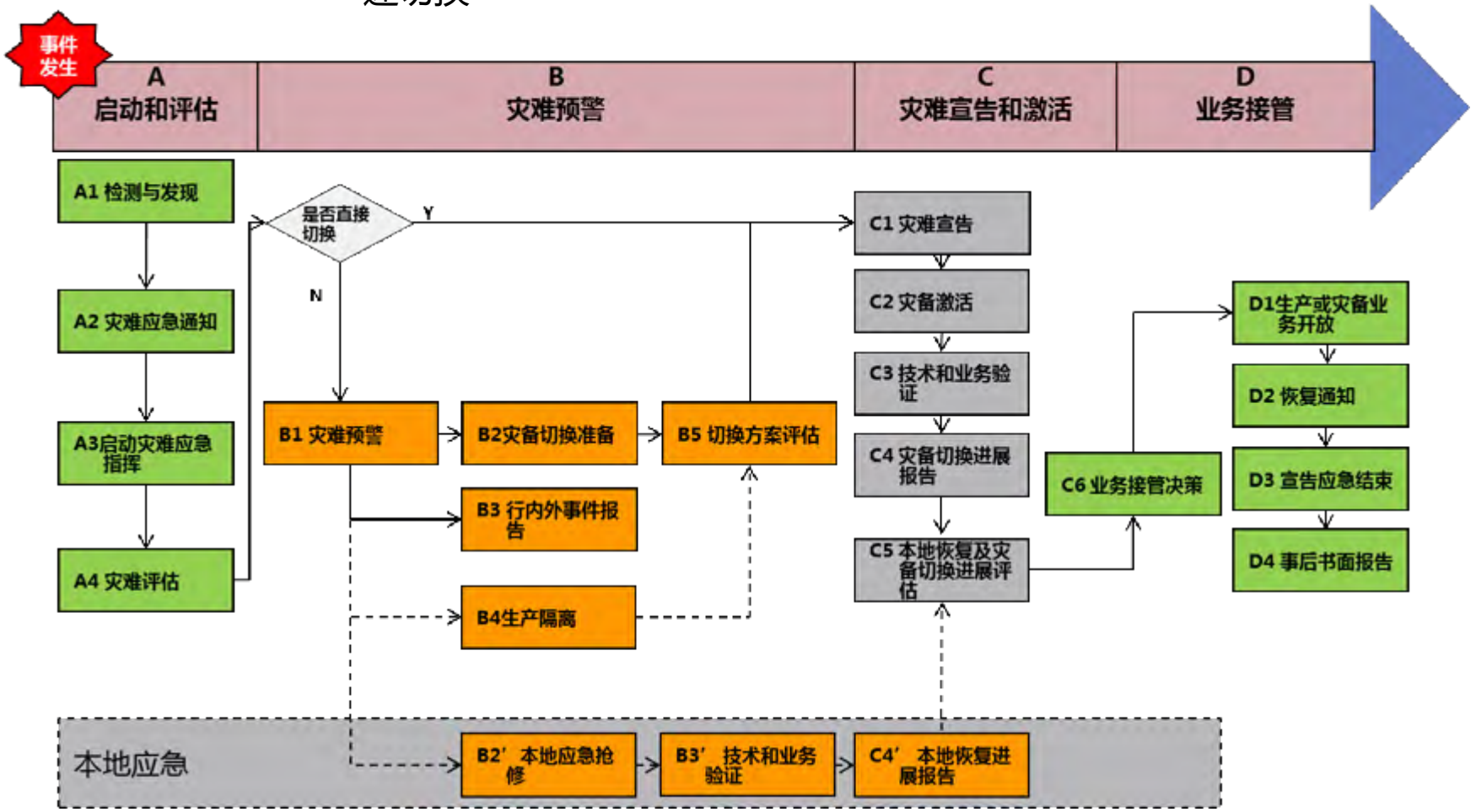
异地灾难切换

ArchSummit

低频高损事件

快速切换：
通过灾备导航工具，一键快速切换

全程监控：
执行过程全景监控



策略：基于守护进程的应用自修复

ArchSummit

高频低损事件

通过与监控系统建立接口，对该类事件发生信息进行及时推送、记录和汇总。

提醒：

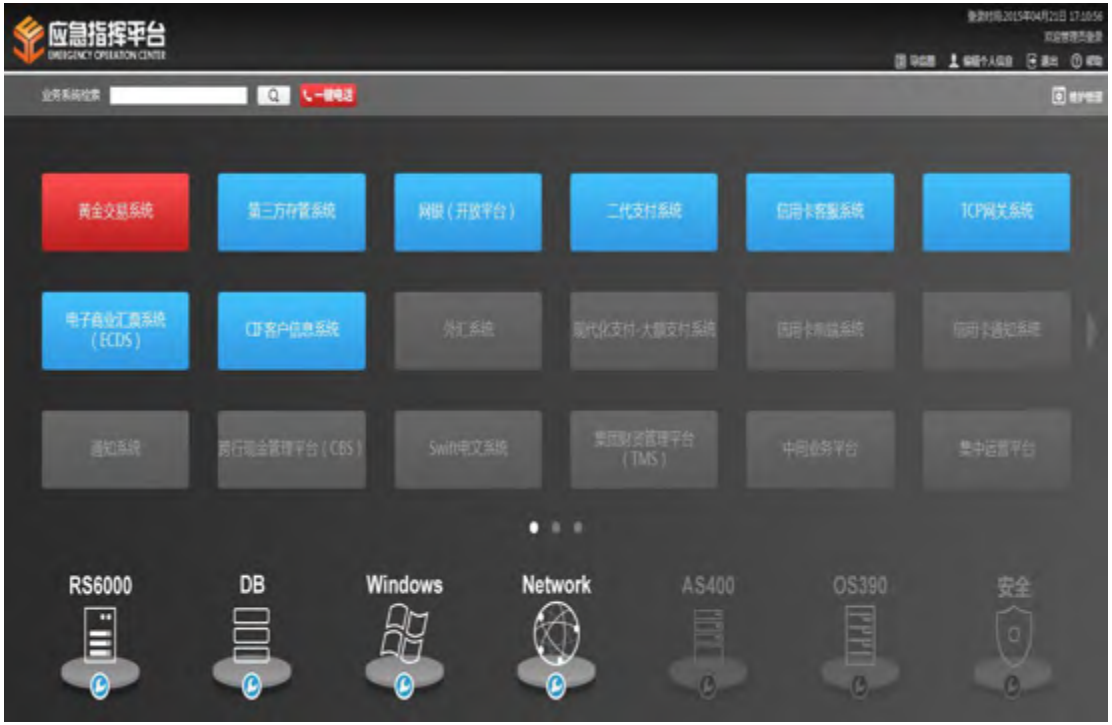
通过自动恢复可以解决的事件，在事件发现阶段、自动恢复阶段、恢复完成阶段，均有信息发送至值班经理和值班管理员公告板处进行提醒。

留存：

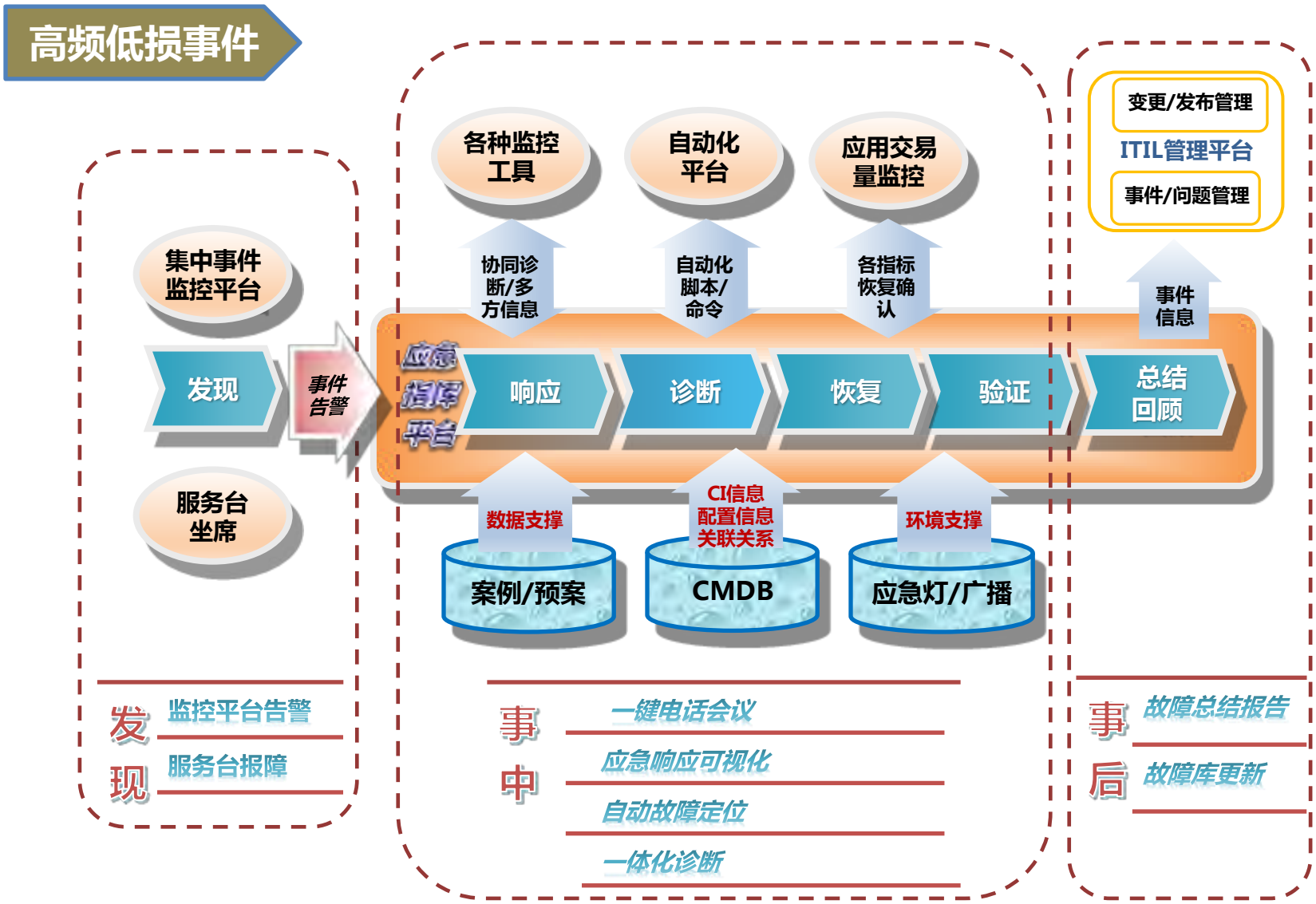
通过公告板及日志模块，可以查看事件处理的过程信息。

分析:

定期汇总生成报告，并进行统计分析。



策略：运维工具/专家协同


高频低损事件
集中事件监控平台
发现
事件告警
服务台坐席
发现
监控平台告警
服务台报障
各种监控工具
自动化平台
应用交易量监控
协同诊断/多方信息
自动化脚本/命令
各指标恢复确认
应急指挥平台
响应
诊断
恢复
验证
总结回顾
事件信息
变更/发布管理
ITIL管理平台
事件/问题管理
数据支撑
CI信息配置信息关联关系
环境支撑
案例/预案
CMDB
应急灯/广播
事中
一键电话会议
应急响应可视化
自动故障定位
一体化诊断
事后
故障总结报告
故障库更新

ArchSummit 2015

谢谢！