CAICT 中国信通院　中国通信标准化协会 China Communications Standards Association

混沌工程在中国移动磐基PaaS平台的实践

严俊

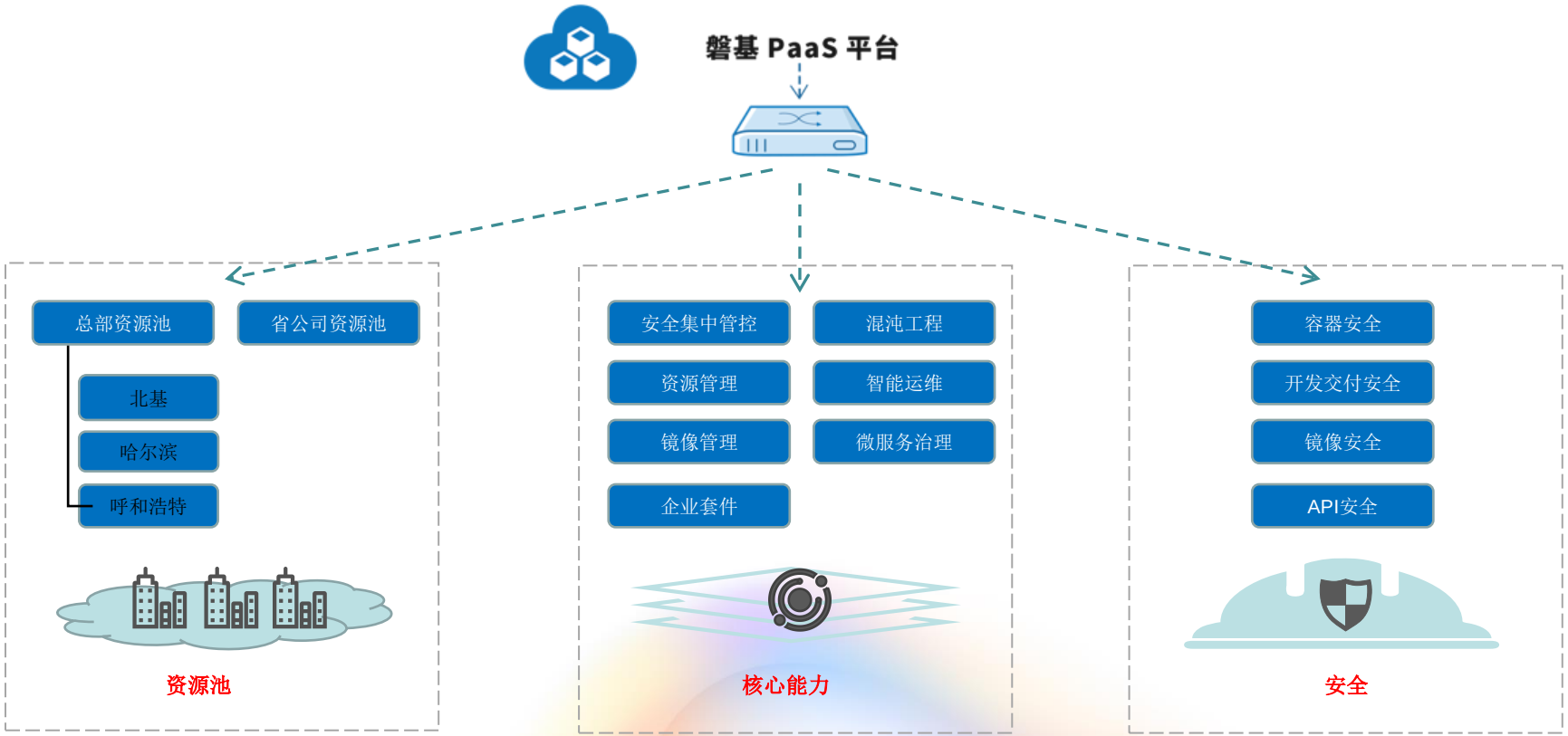
2021 可信云大会
2021 TRUSTED CLOUD SUMMIT
数字裂变　可信发展

1 从0到1构建混沌工程

2 案例分享

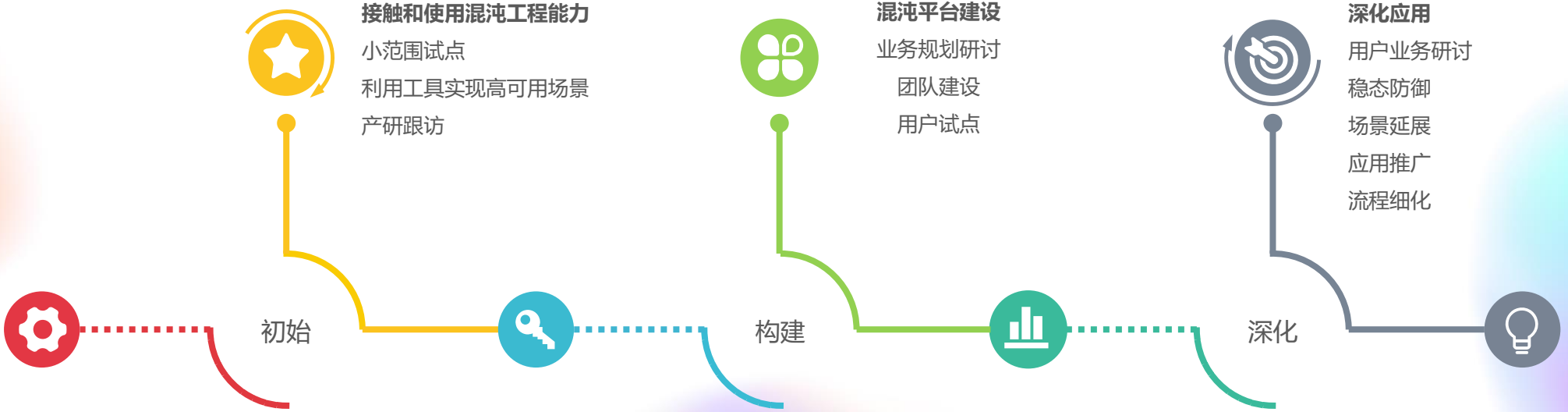
3 如何使用混沌工程能力

中国移动磐基PaaS平台



1 从0到1构建混沌工程

磐基混沌能力发展路线


接触和使用混沌工程能力
小范围试点
利用工具实现高可用场景
产研跟访
初始
混沌平台建设
业务规划研讨
团队建设
用户试点
构建
深化应用
用户业务研讨
稳态防御
场景延展
应用推广
流程细化
深化

从磐基混沌建设的初衷说起...

架构转变

- 传统IT架构云化转型
- 大型功能拆分成数个小功能
- 保持高速迭代

关平台啥事？

- 面向系统级稳定性的软件工程
- 精细化服务的SRE运维体系

如何保障？

架构评审？

平台高可用测试方案？

7*24小时高效运维服务？

监控or应急预案？

。。。。。。

用逻辑推理代替主观推理去证明系统的正确性

过往经验的正确性远没有实际操作的逻辑推理过程要高

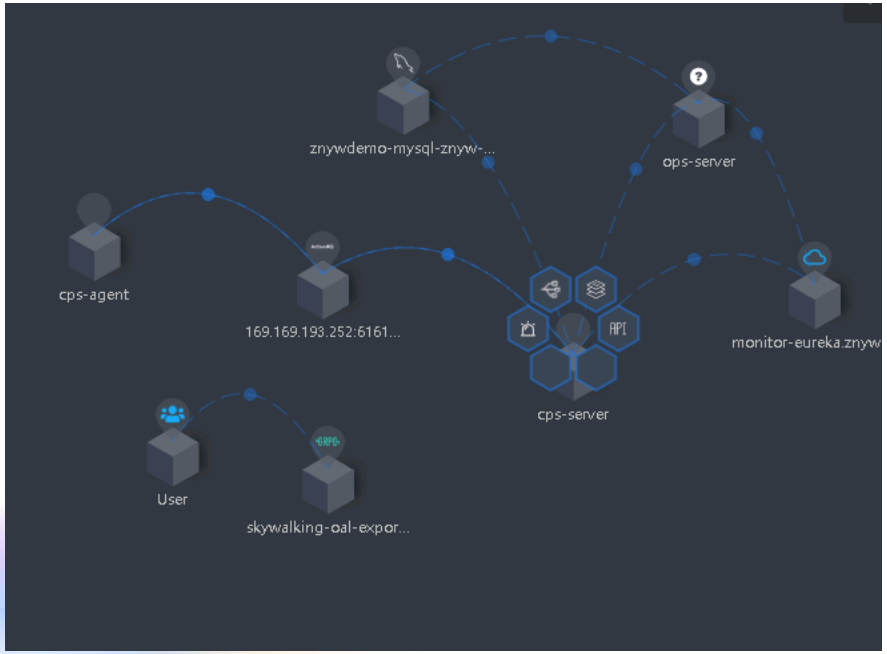
磐基混沌建设的第一步

磐基构建混沌工程体系迈出的第一步是什么？

理清原则

为什么需要 ⟶ 它是什么 ⟶ 它不是什么

混沌工程目的：建立分布式系统抵御生产环境中失控条件的能力与信心



混沌工程原则

TRUCS

不是什么？

- 分析系统内部相互作用的模型
- 已知属性的系统测试
- 故障注入
- 反脆弱

是什么？

混沌工程： 促进发现弱点的实验

通过制造问题，更好地理解系统在生产环境的能力范围，以便能够提前准备相应的应急能力。

混沌工程原则

混沌工程为什么不是测试？

思考的角度：混沌工程和测试最本质的区别是什么？

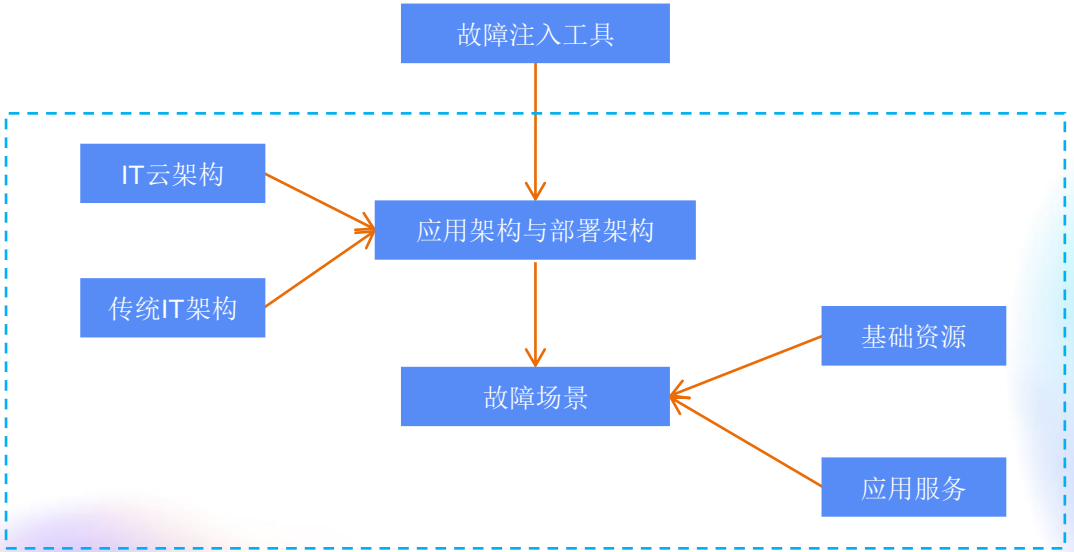
故障注入工具

Chaos Blade

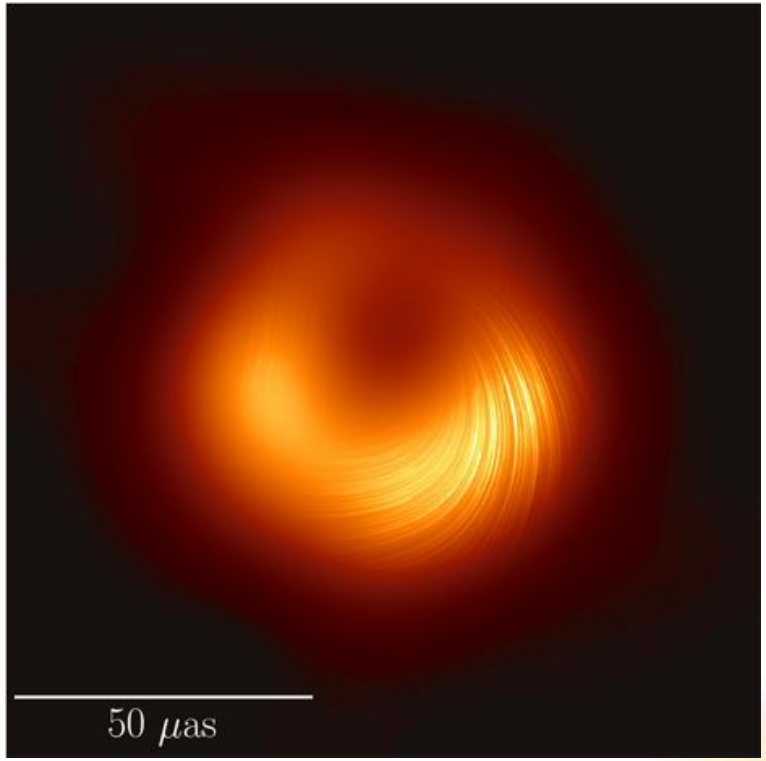
Chaos Monkey

Chaos Mesh

环境适配高，场景丰富



混沌工程可观测性



M87黑洞

混沌工程的黄金标准之一：建立关于稳态行为的假说

这意味着要关注系统预期的运行方式，并通过度量的方式表现出来。

磐基混沌平台功能迭代路线

- 故障中心
  - 场景库
  - 事件库
  - 经验库
- 视图中心
  - 场景排名
  - 记录视图
  - 状态视图
- 演练中心
  - 新建演练
    - 演练描述
    - 监控配置
    - 任务模式
    - 场景选择
    - 防护配置
    - 演练编排
  - 演练管理
    - 演练启停
    - 状态查看
    - 过程详情
    - 演练修改
    - 演练标记
    - 结果报告
  - 演练监控
    - 基础监控
    - 指标监控
    - 日志监控
  - 演练防护
    - 防护措施
    - 防护策略
- 外部接口
- 系统管理
  - 资源管理
    - 主机管理
    - 集群管理
  - 用户管理
    - 资源权限
    - 场景权限
  - 场景管理
    - 场景配置
    - 监控配置
  - 审计日志
    - 平台操作日志
    - 任务演练日志
    - 任务调度日志

2 案例分享

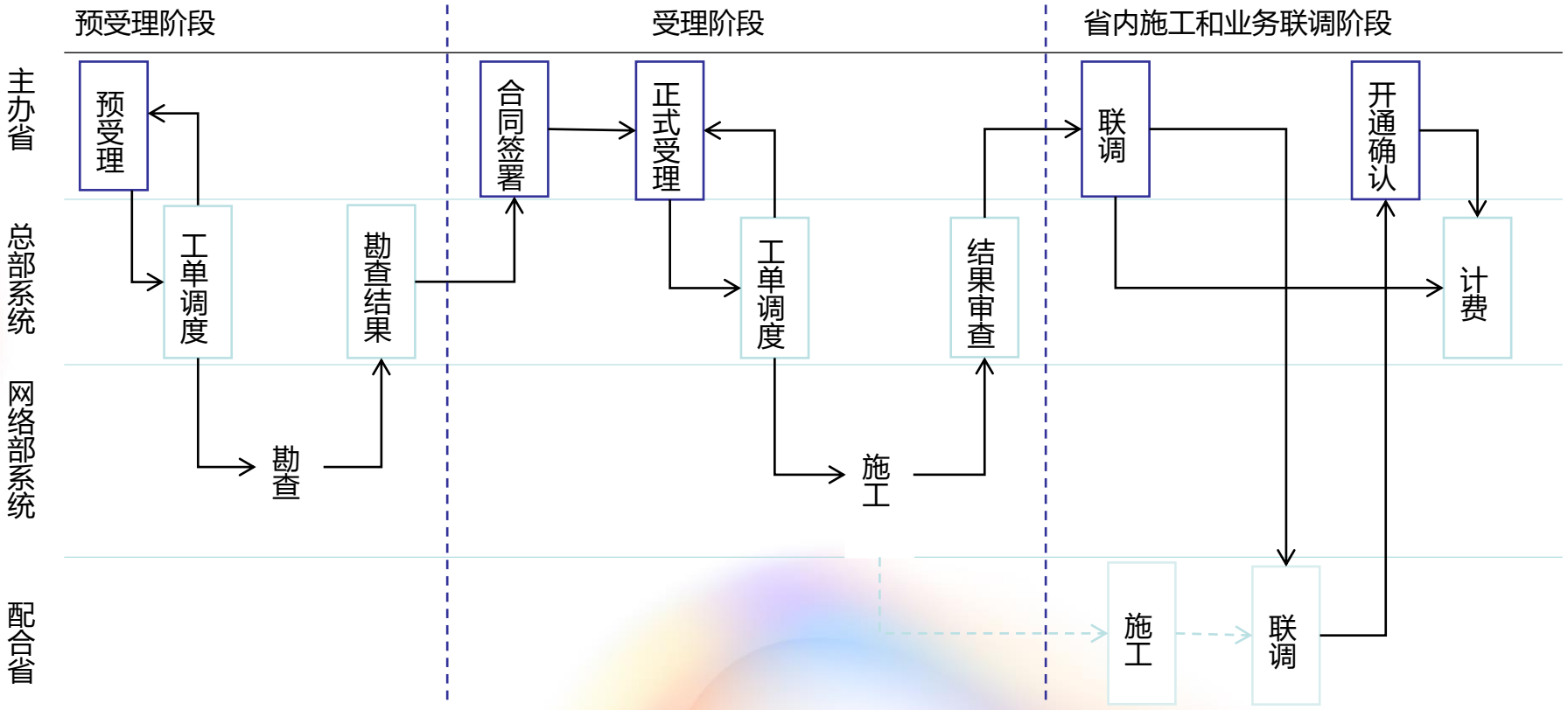
磐基PaaS平台混沌工程故障演练流程

混沌工程黄金标准

- 建立关于稳态行为的假说
- 多样化引入真实故障事件
- 在生产环境中进行实验
- 持续运行自动化实验
- 最小化爆炸半径

演练流程

案例：计费专线业务


预受理阶段
受理阶段
省内施工和业务联调阶段
主办省
总部系统
网络部系统
配合省
预受理
工单调度
勘查
勘查结果
合同签署
正式受理
工单调度
施工
结果审查
联调
施工
联调
开通确认
计费

案例：计费专线业务

穷举出故障实现方案

避免陷入简单原则，而对系统考量不全面

避免工程师的主观意识选择，而非用户视角

选定假设

简单原则

终止实例

CPU使用率100%

内存使用率100%

磁盘使用100%

关闭实例网络

实例端口占用

常见故障类别

硬件故障

软件缺陷

实例状态异常

节点网络延迟

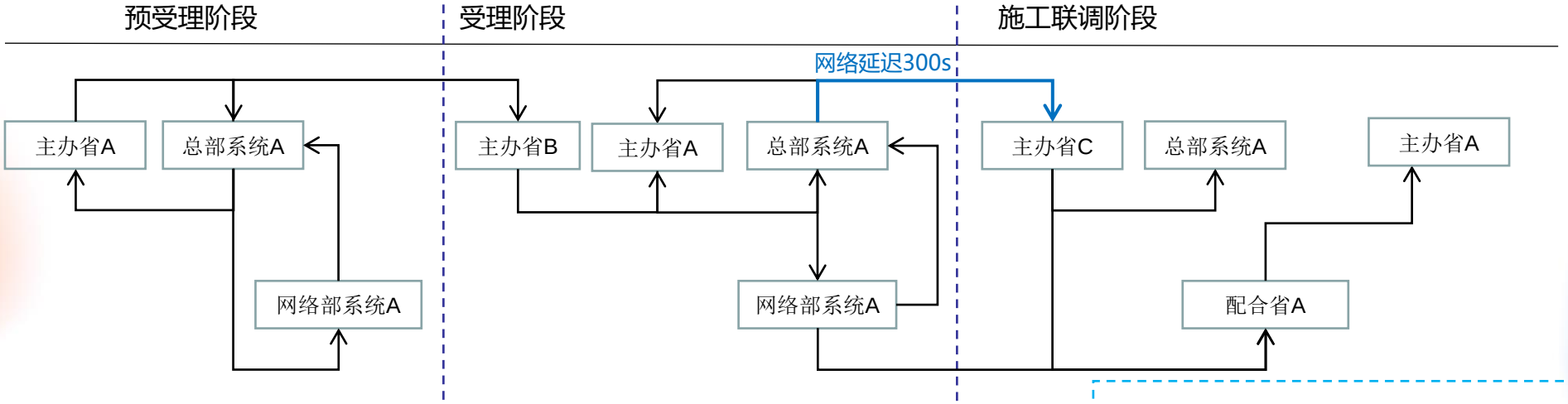
资源耗尽

服务调用异常

依赖故障

案例：计费专线业务

选定假设：业务受理大规模延迟，验证系统协同调度能力，强弱依赖能力。



设定实验范围：测试环境

识别影响范围：下游业务流

确定稳态指标：API响应时长（httpRequestDuration）、业务计费流程执行（前台报表）

实验结果：网络延迟300s内模拟下单请求1000次，当请求响应重发超时后从消息中间件继续消费，网络恢复后未超时请求恢复正常。

实验成果：验证总部系统服务A与主办省服务C间的强弱依赖关系，与系统协同调度能力

案例：监控告警系统

选定假设：业务容器资源异常，PaaS平台具备容错能力，容器检测与自愈

设定实验范围：准生产环境

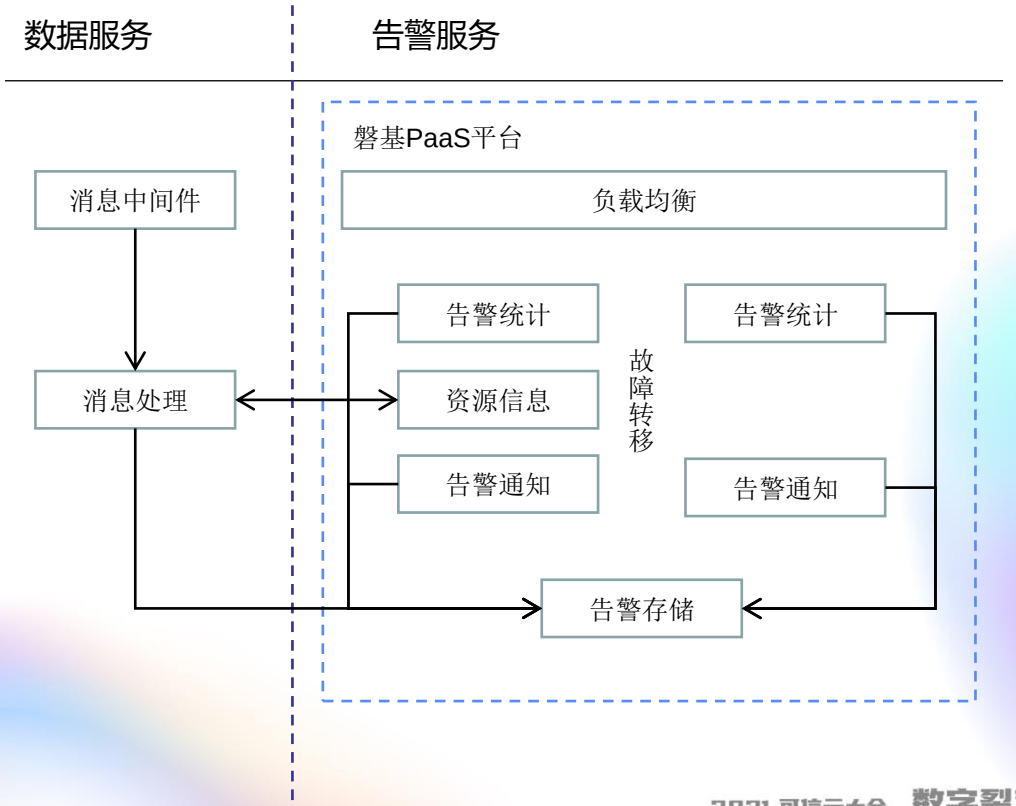
识别影响范围：告警展示与通知

确定稳态指标：磁盘使用率、告警及时率

实验目的：通过模拟服务资源异常，验证磐基PaaS平台动态扩缩容能力，平台故障转移能力、服务在资源满载后的可用性是否符合预期。

实验手段：容器进程停止、节点资源满载、k8s节点掉线

实验成果：微服务因外部因素超时后（故障恢复阶段可用节点网络异常导致容器恢复时间较长），频繁刷新redis缓存，故障阶段基础信息丢失，无有效预案。



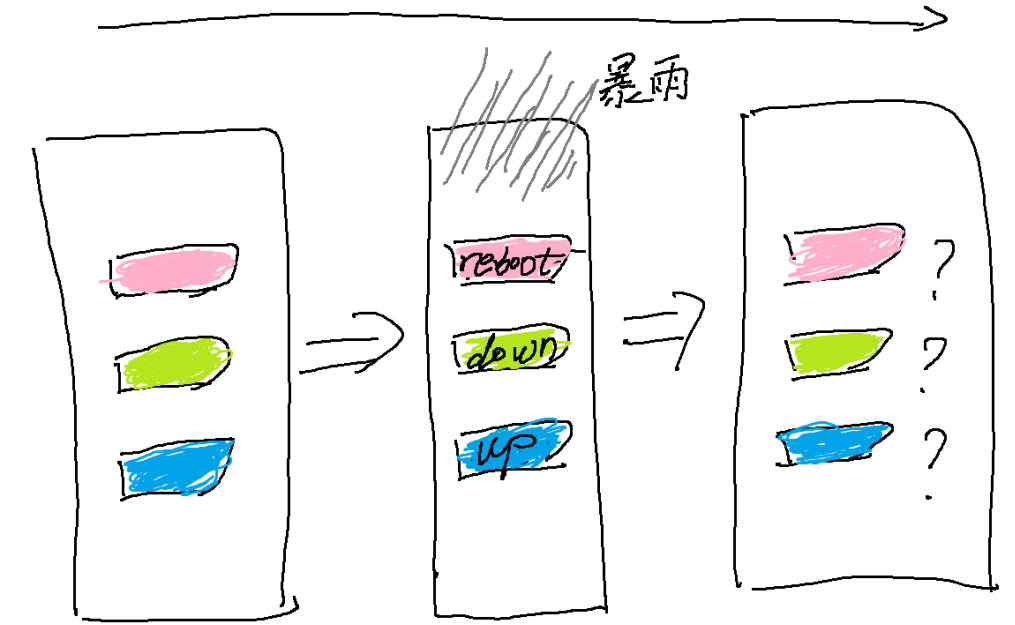
3 如何使用混沌工程能力

混沌工程实施的意义？

•业务量可能有上限，但变更是无止境的，而且会越来越频繁；

•要预见未来是不可能的，但故障是绝对存在的；

•不过，减少将来的意外是可以期望的。



机房停电、操作系统升级、漏洞修复等在生产环境时有发生

混沌工程实施的意义？

混沌工程的目标不是破坏系统，我们的目标也不是利用工具来阻止问题的发生或发现漏洞。

混沌工程是建立一种文化，在不确定的系统结果出现时保持韧性。

意外事件为我们提供了一个机会，看看不同的人对系统如何运作的思维模式有何差异。**—Nora Jones**

混沌工程可以帮助你更好地理解人和机器之间的社会技术界限。

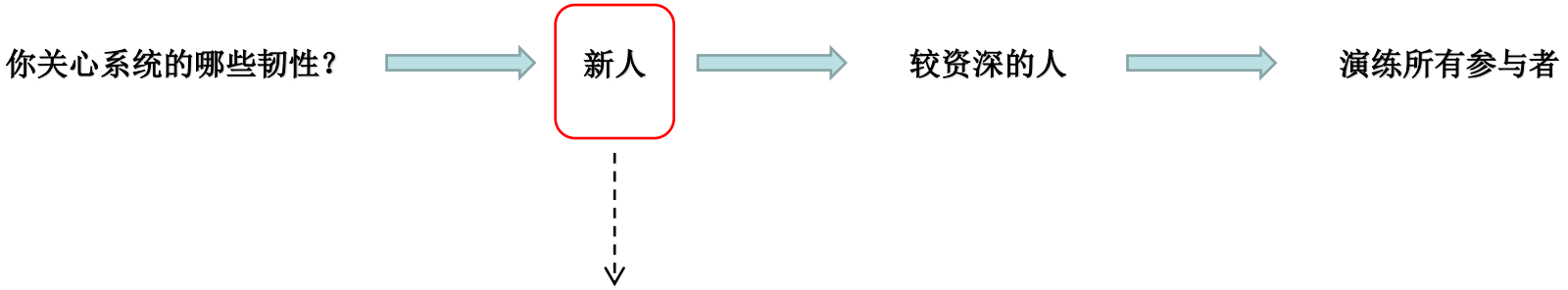
—来自《混沌工程复杂系统韧性实现之道》

通过实验前的团队讨论与实验后的总结分析，提高团队的能力

“习惯将注意力放在如何实施故障注入，而忽略准备阶段和成果分享带来的价值。”

实验前的团队学习

实验前了解团队的思维模式：



新员工受到教条的干扰较少，对不易察觉的环节具有敏锐性。

常见问题如下：

- 是否担心下游服务？
- 若你的服务出问题，上游该如何处理？
- 如果你的服务出现问题怎么办？是否有应急预案？这些方案有什么作用？如何改变服务行为？如何影响用户体验？
- 系统是否会有新缺陷产生的可能？谁应该最清楚这些变化？
- 服务是否会进入糟糕的状态？由什么导致的？
- 对系统参数设置是否有信心？
- 系统日常操作时，最担心什么？

—来自Nora Jones的分享

实验后总结分析

- 团队在实验前是如何获得信心？
- 团队之间是如何就实验的进展进行沟通的？
- 从中你学到了什么？
- 是否因故障问题借助了计划外的人员（因为他们具备某种特定技能）
- 实验过程中是否发生了意外？
- 实验哪些部分与预期不同？
- 下一次我们应该进行哪些实验？
- 如何分享这次实验带来的新知识？

—来自Nora Jones的分享

引入混沌工程至其他领域，非局限于软件工程，例如SRE

运维团队的混沌工程

这个系统中所有参与者都是人，这意味着他们的工作流程除大方向外（保障系统稳定，解决系统缺陷），无章可循。

选定假设：事件响应是团队运动，如果有人延迟响应，整个运维流程将如何工作？例如通讯工具延迟、孤点人员暂时离开。

设定实验范围：最熟悉的系统。

确定稳态指标：故障处理所花费时间。

稳态范围：应在正常响应下运行几场游戏，获得在完整团队下的故障处理效率。

结果：发现组织内沟通渠道的风险，考虑如何扩大网络。

通过混沌工程学科获取到最大的价值！

THANKS!

2021
TRUSTED CLOUD
SUMMIT