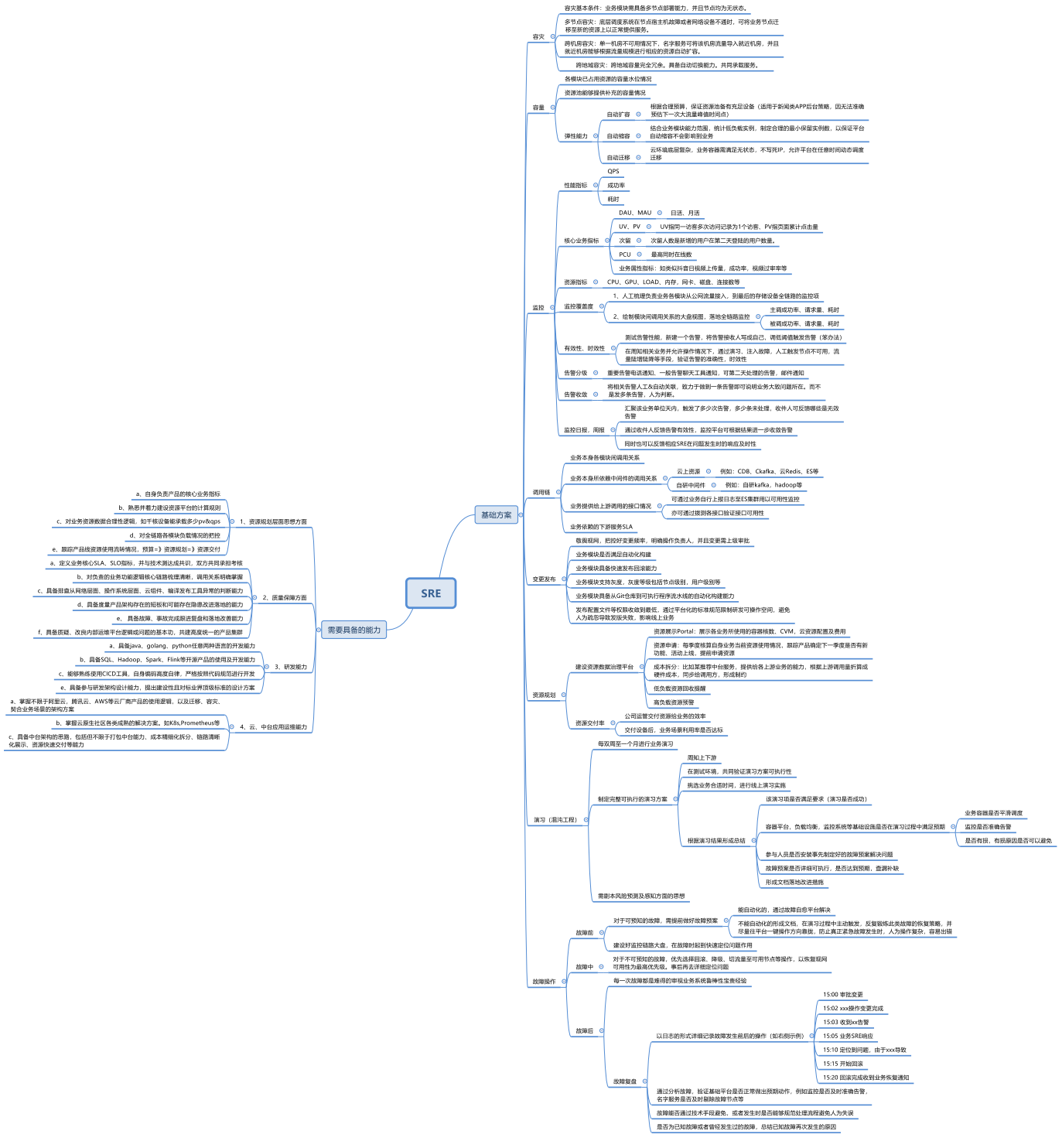
SRE
需要具备的能力
1、资源规划层面思想方面
a、自身负责产品的核心业务指标
b、熟悉并着力建设资源平台的计算规则
c、对业务资源数据合理性逻辑，如干核设备能承载多少pv&qps
d、对全链路各模块负载情况的把控
e、跟踪产品线资源使用流转情况，预算=》资源规划=》资源交付
2、质量保障方面
a、定义业务核心SLA、SLO指标，并与技术测达成共识，双方共同承担考核
b、对负责的业务功能逻辑核心链路梳理清晰，调用关系明确掌握
c、具备排查从网络层面、操作系统层面、云组件、编译发布工具异常的判断能力
d、具备度量产品架构存在的短板和可能存在隐患改进落地的能力
e、具备故障、事故完成原进复盘和落地改善能力
f、具备质疑、改良内部运维平台逻辑或问题的基本功，共建高度统一的产品集群
3、研发能力
a、具备java、golang、python任意两种语言的开发能力
b、具备SQL、Hadoop、Spark、Flink等开源产品的使用及开发能力
c、能够熟练使用CICD工具，自身编码高度自律，严格按照代码规范进行开发
e、具备参与研发架构设计能力，提出建设性且对标业界顶级标准的设计方案
4、云、中台应用运维能力
a、掌握不限于阿里云，腾讯云、AWS等云厂商产品的使用逻辑，以及迁移、容灾、
灰合业务地量的架构方案
b、掌握云原生社区各类成熟的解决方案。如K8s,Prometheus等
c、具备中台架构的思路，包括但不限于打包中台能力、成本精细化拆分、链路清晰
化展示、资源快速交付等能力
基础方案
容灾
容灾基本条件：业务模块需具备多节点部署能力，并且节点均为无状态。
多节点容灾：底层调度系统在节点宿主机故障或者网络设备不通时，可将业务节点迁移至新的资源上以正常提供服务。
跨机房容灾：单一机房不可用情况下，名字服务可将该机房流量导入就近机房，并且就近机房能够根据流量规模进行相应的资源自动扩容。
跨地域容灾：跨地域容量完全冗余。具备自动切换能力。共同承载服务。
容量
各模块已占用资源的容量水位情况
资源池能够提供补充的容量情况
弹性能力
自动扩容
根据合理预算，保证资源池备有充足设备（适用于新闻类APP后台策略，因无法准确预估下一次大流量峰值时间点）
自动缩容
结合业务模块能力范围，统计低负载实例，制定合理的最小保留实例数，以保证平台自动缩容不会影响到业务
自动迁移
云环境底层复杂，业务容器需满足无状态，不写死IP，允许平台在任意时间动态调度迁移
监控
性能指标
QPS
成功率
耗时
核心业务指标
DAU、MAU
日活、月活
UV、PV
UV指同一访客多次访问记录为1个访客，PV指页面累计点击量
次留
次留人数是新增的用户在第二天登陆的用户数量。
PCU
最高同时在线数
业务属性指标：如类似抖音日视频上传量，成功率，视频过审率等
资源指标
CPU、GPU、LOAD、内存、网卡、磁盘、连接数等
监控覆盖度
1、人工梳理负责业务各模块从公网流量接入，到最后的存储设备全链路的监控项
2、绘制模块间调用关系的大盘视图，落地全链路监控
主调成功率、请求量、耗时
被调成功率、请求量、耗时
有效性、时效性
测试告警性能，新建一个告警，将告警接收人写成自己，调低阈值触发告警（笨办法）
在周知相关业务并允许操作情况下，通过演习、注入故障，人工触发节点不可用，流量陡增陡降等手段，验证告警的准确性，时效性
告警分级
重要告警电话通知、一般告警聊天工具通知，可第二天处理的告警，邮件通知
告警收敛
将相关告警人工&自动关联，致力于做到一条告警即可说明业务大致问题所在。而不是发多条告警，人为判断。
监控日报，周报
汇聚该业务单位天内，触发了多少次告警，多少条未处理，收件人可反馈哪些是无效告警
通过收件人反馈告警有效性，监控平台可根据结果进一步收敛告警
同时也可以反馈相应SRE在问题发生时的响应及时性
调用链
业务本身各模块间调用关系
业务本身所依赖中间件的调用关系
云上资源
例如：CDB、Ckafka、云Redis、ES等
自研中间件
例如：自研kafka，hadoop等
业务提供给上游调用的接口情况
可通过业务自行上报日志至ES集群用以可用性监控
亦可通过拨测各接口验证接口可用性
业务依赖的下游服务SLA
变更发布
敏捷观网，把控好变更频率，明确操作负责人，并且变更需上级审批
业务模块是否满足自动化构建
业务模块具备快速发布回滚能力
业务模块支持灰度，灰度等吸包括节点级别，用户级别等
业务模块具备从Git仓库到可执行程序流水线的自动化构建能力
发布配置文件等权限收敛到最低。通过平台化的标准规范限制研发可操作空间，避免人为疏忽导致发版失败，影响线上业务
资源规划
建设资源数据治理平台
资源展示Portal：展示各业务所使用的容器核数、CVM、云资源配置及费用
资源申请：每季度核算自身业务当前资源使用情况，跟踪产品确定下一季度是否有新功能、活动上线，提前申请资源
成本拆分：比如某推荐中台服务，提供给各上游业务的能力，根据上游调用量折算成硬件成本，同步给调用方，形成制约
低负载资源回收报警
高负载资源预警
资源交付率
公司监管交付资源给业务的效率
交付设备后，业务场景利用率是否达标
演习（混沌工程）
每双周至一个月进行业务演习
制定完整可执行的演习方案
周知上下游
在测试环境，共同验证演习方案可执行性
挑选业务合适时间，进行线上演习实施
根据演习结果形成总结
该演习项是否满足要求（演习是否成功）
容器平台，负载均衡，监控系统等基础设施是否在演习过程中满足预期
业务容器是否平滑调度
监控是否准确告警
是否有损，有损原因是否可以避免
参与人员是否安装事先制定好的故障预案解决问题
故障预案是否详细可执行，是否达到预期，查漏补缺
形成文档落地改进措施
需剧本风险预测及感知方面的思想
故障操作
故障前
对于可预知的故障，需提前做好故障预案
能自动化的，通过故障自愈平台解决
不能自动化的形成文档，在演习过程中主动触发，反复锻炼此类故障的恢复策略，并尽量往平台一键操作方向靠拢，防止真正紧急故障发生时，人为操作复杂，容易出错
建设好监控链路大盘，在故障时能到快速定位问题作用
故障中
对于不可预知的故障，优先选择回滚、降级、切流量至可用节点等操作，以恢复观网可用性为最高优先级。事后再去详细定位问题
故障后
每一次故障都是难得的审视业务系统鲁棒性宝贵经验
故障复盘
以日志的形式详细记录故障发生前后的操作（如右侧示例）
15:00 审批变更
15:02 xxx操作变更完成
15:03 收到xx告警
15:05 业务SRE响应
15:10 定位到问题，由于xxx导致
15:15 开始回滚
15:20 回滚完成收到业务恢复通知
通过分析故障，验证基础平台是否正常做出预期动作，例如监控是否及时准确告警，名字服务是否及时最除故障节点等
故障能否通过技术手段避免，或者发生时是否能够规范处理流程避免人为失误
是否为已知故障或者曾经发生过的故障。总结已知故障再次发生的原因