阿里文娱技术 阿里云 开发者社区

从DevOps到AIOps
运维技术精选

基础架构 | 成本管控 | AIOps探索

—— 阿里文娱技术精选系列——
智能运维技术

关注我们

（阿里文娱技术公众号）

关注阿里技术

扫码关注「阿里技术」获取更多资讯

加入交流群

1）添加“文娱技术小助手”微信

2）注明您的手机号 / 公司 / 职位

3）小助手会拉您进群

By 阿里文娱技术品牌

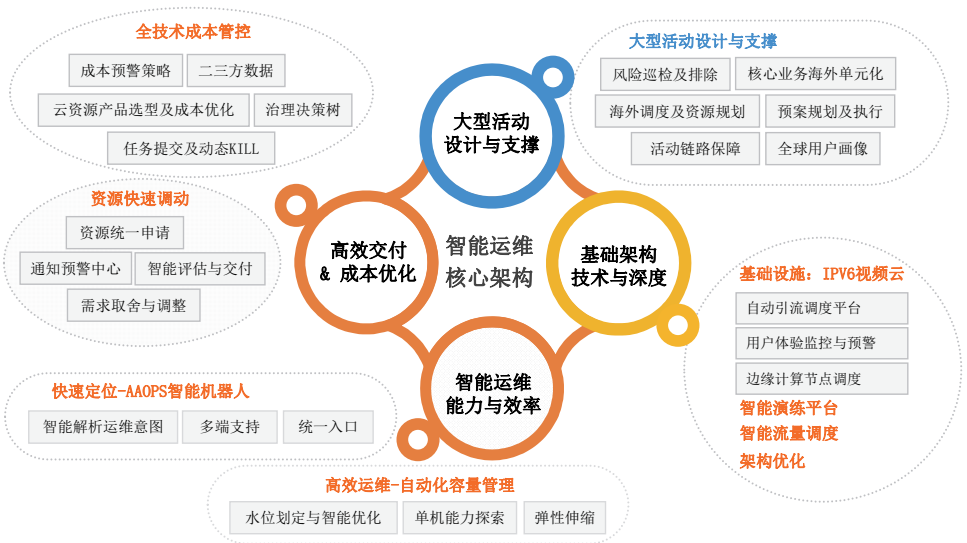
更多电子书

扫码获取更多技术电子书

目 录

序

优酷应用架构团队负责优酷的稳定性能力建设和资源精细化运营工作。我们通过稳定性能力系统的建设，重在解决由于业务快速迭代发展而可能造成的不稳定，从源头变更的管控到线上的风险自动治理，从架构体系改造完成到每月定期的容灾及故障演练；通过资源精细化管控，重在解决资源交付混乱低效，线上资源浪费的问题。从预算申请到资源申请，从资源交付到资源优化，最终资源下线，有效保障了资源在生命周期中快速交付及有效使用。我们不再关注于单点问题，而是从全局去分析，通过架构演进和智能平台相结合的方式，从源头上预测和解决可能会发生问题的点。在 AI 和数据化的加持下，平台化智能化的运维系统，真正的做到助力优酷业务更快更稳更优！



本章以阿里文娱运维团队的实践为蓝本，摘录了具有代表性的经验和总结。包括架构升级、智能运维平台的建设、整个技术资源的流转，以及在资源有限的情况下，如何让高高的业务目标快速落地；如何做整体稳定性，包括用户体验升级，降低卡顿率，而对于 AIOps，我们正在做积极的尝试，未来我们希望在智能化方向上进行更多的探索来提升整体稳定性以及资源稳定高效使用。

阿里文娱应用架构部高级技术专家 秦庸

2020.1.15

5G 的基础架构：如何让数亿用户无缝支持 IPv6?

作者| 阿里文娱技术专家 盖优

一、概述

什么是 IPv6？IPv6=互联网网络层传输协议第六版。广义的讲，IPv6 就是下一代互联网的基础设施，5G 物联网的核心基础。与现行 IPv4 相比，它更安全，更高效，更省成本，几乎用不完的 IP 地址，业务无限拓展。

本文将详解优酷 IPv6 技术改造实践，包括遇到的实际问题、技术解法，希望对大家有借鉴。

1. 背景

Internet 在形成及演化的初期，经历了一个纷繁复杂的过程，随着网络技术的发展，以 IP 协议为核心，包括：地址格式、数据封装及数据转发等 Internet 网络结构的基本框架，已经稳定运行 30 多年不曾改变。5G 来临的时候，Internet 的现有结构无法快速适应用户及上层应用需求。主要体现在以下几点：

1）IPv4 资源枯竭：海外大量购买资源，私网地址仅可支撑三年；

2）IPv6 用户增长：海外美国 50%，印度、欧洲 20%；

3）Apple 审核：2016 年 6 月起，Appstore 就开启了 IPv6 only 审核；

4）商业竞争态势：Google、Facebook 已覆盖 50% IPv6 终端，IPv6 云产品发布；

5）战略意义：由于 IPv4 报文的限制，使得域名根服务器的总数有限制。IPv6 出现以后，这个限制被打破，可写入更多的根服务器地址。目前，全球已完成 25 台 IPv6 根服务器架设，中国部署了 4 台，打破了中国过去没有根服务器的困境。

China	中国	4	一主三辅
USA	美国	3	一主二辅
Japan	日本	1	主根
India	印度	3	辅根
France	法国	3	辅根
Germany	德国	2	辅根
Russia	俄罗斯	1	辅根
Italy	意大利	1	辅根
Spain	西班牙	1	辅根
Austria	奥地利	1	辅根
Switzerland	瑞士	1	辅根
Netherlands	荷兰	1	辅根
Chile	智利	1	辅根
South Africa	南非	1	辅根
Australia	澳大利亚	1	辅根

（图　全球 IPv6 根服务器分布）

2. 目标阶段

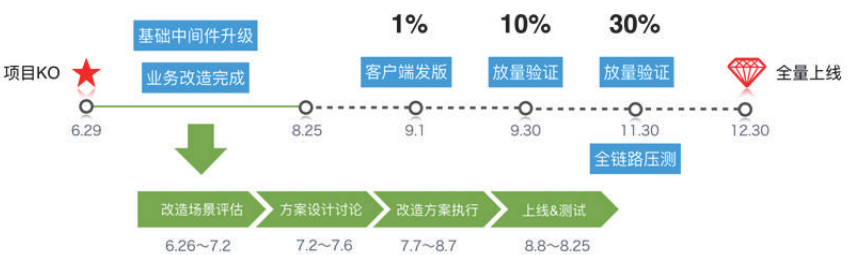
优酷 IPv6 项目将分成二步走，包括 IPv4/IPv6 双栈阶段，以及内外网 IPv6-only 阶段。

1）IPv4/IPv6 双栈改造：实现应用快速对外服务，以 Web/APP 请求服务为核心，满足 IPv6 生态发展的需求，并且以外网拉动应用的不同需求。爬虫、邮箱、DB、存储等直接交互，需要内外网服务器采用 IPv4/IPv6 双栈能力，整网双栈交付；

2）IPv6 Only：当超过 50%应用逐步迁移到 IPv6 后，新应用采用 IPv6 准入，遗留一些老旧应用继续采用 IPv4 服务，内网采用 4over6 进行封装。相对双栈而言，IPv6 only 成本更低，查表转发速度更快，只需维护一套协议栈，全面开启 IPv6 Only 时代。

3. 实施周期

优酷 2018 年 6 月开始，实施了 IPv6 的全业务改造，完成客户端及服务端改造，实现应用快速对外服务。2018 年底，实现全量部署上线，总耗时 6 个月。所以，与优酷同等用户规模和服务体量的企业，基本上 6 个月甚至更短的时间就可以完成整体 IPv6 的改造。



（图　优酷 IPv6 改造计划）

二、遇到的问题点及解法

IPv6 整个改造过程中可能会遇到的几个典型问题和解决方法，简单的总结如下：

1. 没有 IPv6 环境

众所周知，一般应用从开发到上线，至少要经过开发测试环境、预发环境、Beta 环境，最后上正式环境。一开始，基础环境还没有具备的时候，我们使用 IPv6 over IPv4 链路 VPN 的方式连入测试环境 ，需要 PC/客户端加证书改 Hosts，移动端无法改 Hosts 的，需要 root 越狱。然后，我们加强了基础网络和 IT 合作，在多个阿里园区部署多个 IPv6 的接入环境，打通 IPv6 出口，打通办公网和机房的 IPv6 链路，实现慢慢外网 IPv6，日常环境通、预发通、正式通，慢慢使业务能够测试逐步提升到 IPv4 相同的测试体验，通过域名劫持等手段，跳过了 Hosts 配置的尴尬，达到标准的测试效率。

2. OS 网络模块问题

需要让容器支持从请求头部获取 IPv6 地址，那么就需要把用户 IP 一级一级透传过来，就需要在各级的服务器上升级网络模块，扩展报文头部。例如 TOA 模块，TOA 模块是为了让后端的 Real Server 能够看到真实的 ClientIP 而不是 LVS 的 DIP。同时，Tengine/Nginx 等应用需要升级到支持 IPv6 的版本（支持新 TOA 模块等），由于历史原因存在各种老版本无法升级，导致升级受阻。我们通过推动应用接入统一接入改造，避免自行升级网络模块带来的风险。

通过老版本应用的升级，去 Nginx 的方式，统一升级安装 Tengine-proxy（安装在 ECS 测试机器或宿主机上都可以），为了能减少业务改造工作量，在接入层架构我们做了大量的改造。

3. 地址库特殊需求

首先，统一 IP 地址库，要求所有业务必须统一使用 IP 地址库。其次，协调集团地址库生产方，满足优酷使用场景需求，使统一过程中业务改造工作量减少。再次，对于广告等必须要使用行业统一地址库的场景，我们也制定了多套方案去解决。

兜底方案：将广协地址库中的地区编码，加入到集团地址库中，使集团库具备行业库的能力，在行业库没有完全产出之前，广告业务可以临时使用集团地址库进行改造和测试，保障业务不受损。

长期方案：主动出击，联系广协等行业协会，加快产出 IPv6 地址库，并且主动无偿提供阿里集团地址库数据，更加快了整个广协会员单位的改造进度。

4. MTU 问题

IPv4 时代，中间网络三层设备会进行分片，所以一般设置为 1500 的最大值，以降低网络开销。但 IPv6 协议为了减轻中间网络层设备繁杂度及成本，中间设备不再分片，由两端的协商指定。导致默认 mtu1500 的情况下，中间设备出现大量丢包，原因是 NAT 转换，TCP Option 等额外叠加，实际超过 1500。开启 SYN Proxy，通过 MSS 与端进行协商，调整 MTU 为最小值 1280。发现中间层 MTU 小于 1280 时，进行网络报障等办法。

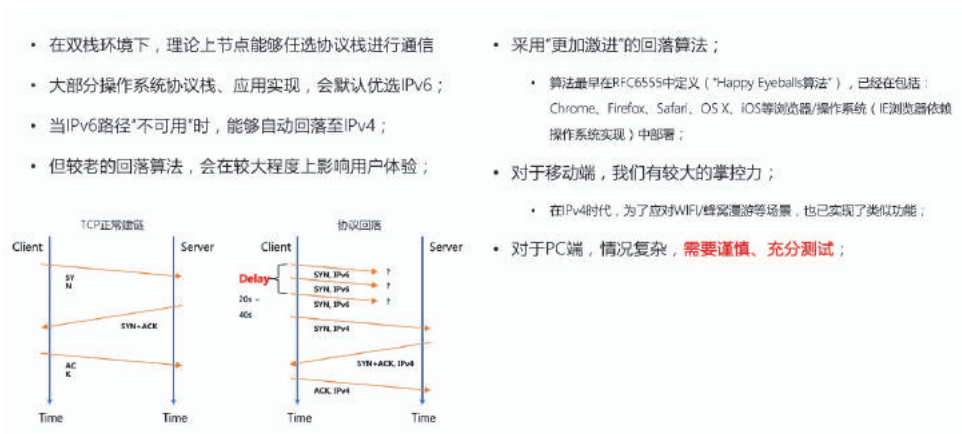
5. 客户端是否 IPv6，如何验证

这是一个很现实的问题，我的网络已经是 IPv6 了，业务也能正常运行，但怎么确认网络是运行在 IPv6 上，没有被降级呢？主要有以下两个手段：

1）抓取客户端日志：这也是最笨太最准确的手段。具体抓日志的方法有很多，就不再重复介绍了；

2）业务改造，加入网络检测能力：将优酷客户端当做网络测试的工具。

6. 协议回落问题



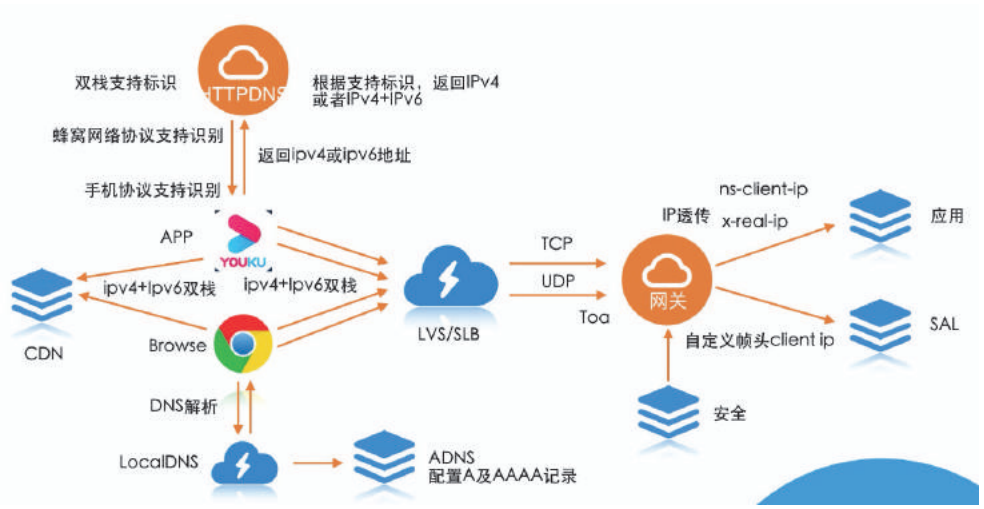
（图　协议回落）

7. CDN 灰度问题

CDN 域名由阿里云等 CDN 服务提供商进行调度控制，用户请求链路和业务服务是不一样的，导致业务服务是 IPv6，CDN 走的是 IPv4；也可能 CDN 是 IPv6，业务服务是 IPv4，无法和

业务同一灰度范围。解决方案：使用 HTTPDNS 能力，让 CDN 域名和业务域名共同管理，同步开启灰度的地域和运营商。同时，增加 IPv6 专属 CDN 域名，APP 侧通过业务侧增加业务逻辑，分别下发不同的域名来实现同一灰度节奏能力。当业务服务返回客户端的出口 IP 是 IPv6 时，调用 IPv6 的 CDN 域名；当业务服务返回客户端的出口 IP 是 IPv4 时，调用 IPv4 CDN 域名。

三、架构设计



（图　优酷 IPv6 改造架构图）

从客户端到服务端，所有涉及到的设备、网络、APP、服务器、业务等都是改造范围。

1）用户端的网络，包括移动网络和局域网：这部分移动网络依赖运营商，目前三大运营商的 4G IPv6 支持率>70%，固定宽带内部局域网等总体支持率不足 3 成，家庭路由器等也需要升级；

2）用户终端设备：依赖手机等终端设备厂家更新升级固件，小品牌的终端就听天由命了。部分安卓手机需要分配到 64 段的 IPv6 才能正常连接上 IPv6 的 Wi-Fi；

3）OS/浏览器：依赖苹果、谷歌等的更新节奏，需要客户端 OS 及浏览器都更新至最新版本，老 OS 基本不支持；

4）客户端 APP/PC 端网页：网络底层包需要支持 IPv6 以及降级能力，实施方案中详细说明；

5）HTTPDNS：基于一定的策略对支持双栈网络的客户端下发 IPv6 地址，需 HTTPDNS 端改造支持；

6）Local DNS：需要 DNS 支持 IPv6 解析，同时域名解析记录中添 AAAA 记录；

7）网络链路：运营商需要支持 IPv6，包括用户端的出口网络和服务端的机房出口，网络路由等；

8）LVS：所有服务的出口，需要支持 IPv6，将请求转发至 RS（反向代理服务）；

9）反向代理层：将请求转发至具体业务服务器，并带上客户端 IPv6 地址；

10）业务服务：请看下一节。

四、详细实施步骤

整个改造过程包括：客户端 APP 及 PC/H5 端/业务服务端的改造，安全测试及灰度保障能力。

1. 客户端 APP

1）更新网络底层包：涉及到集团二方包或者第三方网络库的，需要升级到最新版本。第三方网络库需要确认具备 IPv6 能力，否则需要重新选择其它网络库；

2）升级 IP 地址库：端上集成有 IP 地址库的，需要升级到包含 IPv6 记录的 IP 地址库；

3）升级 HTTPDNS 服务库：使用 HTTPDNS 服务的，需要确认支持 AAAA 记录的下发；使用 Local DNS 解析的，需要改造实现 DNS 服务请求参数中添加 AAAA 记录解析的标识；

4）改造支持降级能力：使用三方库已经具备 IPv6 链路质量不佳时自动降级 IPv4 能力的，可以不改造。否则，需要业务或者架构侧进行 IPv6 网络质量的判断，并实现降级功能；

5）探测埋点改造：弱网、DNS 耗时的情况下，探测能否正常，IPv6 下埋点是否正常上报；

6）测试手法：所有功能需要在 IPv4 only，IPv6/IPv4 双栈测试通过。IPv6 only 有条件时也需要测试通过。

2. 业务服务端/PC 端/H5 端

1）IP 地址库使用：是否有用到地址库，对用户 IP 进行地域来源等判断。有的话需要升级到 IPv6 地址库，并更新调用方法；

2）IP 地址格式判断：是否对用户 IP 进行验证，有的话，需要加入 IPv6 地址格式的正则表达式判断；

3）IP 地址保存：是否对 IP 有存库等保存操作，需要修改相应字段的长度，IPv6 长于 IPv4，MySQL 建议字段类型 VARBINARY(16)；

4）依赖链路上的修改：是否会将 IP 作为接口参数传递给下游依赖业务。有的话，下游依赖业务也需要改造；

5）客户端 IP 地址的取得方式：如果从客户端请求的头部获取，那么在双栈环境中，同一请求，你只能获取到 IPv4 和 IPv6 地址中的一个，不可能两个都获取。如果是通过请求正文中的某个字段，把客户端地址传上来的，那么，你需要考虑是否需要获取客户端的 v4 和 v6 的所有地址；

6）日志：当用了第三方的采集工具，如果采集工具不支持 IPv6 的话，那么采集上来的数据会和服务端的请求日志无法对齐，产生 GAP。所以第三方数据产品等都需要能够支持用户 IPv6 数据的采集；

7）监控：存在用户 IP 作为判断条件/统计条件的监控配置时，需要改造；

8）大数据统计：存在用户 IP 作为判断条件/统计条件的内容时，需要业务改造。

3. 安全，测试及灰度保障

主要包括上线前的测试保障及上线后的灰度引流能力。

1）测试保障：抓取客户端日志；客户端业务改造，加入网络检测能力；客户端增加 IPv6 链路日志，服务端日志工具支持对 IPv6 客户端地址进行分析汇总；IPv6 流量压力测试能力；模拟 IPv6 网络限速，延迟增加能力。

2）灰度引流能力包括两种方式：

HTTPDNS 方式：基于用户设备的白名单；基于地域+运营商+百分比+用户设备白名单；基于 APP 版本的全量百分比。

Local DNS(ADNS)方式：ADNS 新开发并上线了一个能力，支持一个域名下配置多 CNAME 解析功能，并且每条解释都可以配置权重，通过修改 IDNS 的 CNAME 权重配置来达到比例控制。同时加上自有的线路和运营商的选择能力，满足地域级的灰度需求。

3）自动化能力：我们开发了自动化的灰度系统，根据起始参数和灰度目标，自动规划灰度比和时间节奏，实现完全自动化的灰度引流。监控预警+自动回滚能力，边喝咖啡边看灰度量，就是这么简单。

大促下的运维挑战：如何抗住双 11 猫晚

作者| 阿里文娱技术专家 子霖

一、背景

2019 双 11 猫晚在全球近 190 个国家和地区播出，海外重保是首要任务，如何提升海外用户观看猫晚的体验？本文将详解双 11 猫晚国际化的技术挑战和技术策略。

二、播前成功率改进

1. 海外主站链路优化

计算猫晚海外 CDN 带宽用量，明确海外直播 CDN 资源分布及其调度回源链路，结合现有直播赛事，分析海外各地域国家直播卡顿率，重点分析卡顿率高的地域。对比访问主站各地域

国家 TCP 建连时间，调整主站 TCP 建连时间。

2. 直播服务单元化

1）为什么要做单元化

用户体验及资源瓶颈：随着业务体量和服务用户群体的增长，用户需要更优的访问速度，单机房无法支持长期的服务的持续扩容；

服务异地容灾：异地容灾已经成为核心服务的标配，有些服务虽然进行了多地多机房部署，但数据还是只在中心机房。要实现真正意义上的异地多活，就需要对服务进行单元化改造；

全球化战略：全球化战略带来的不只是用户增长，同时数据也会快速增长，全球数据都集中部署在少数几个机房显然不太现实，基于地理区域划分、数据维度驱动的单元化架构是未来全球化战略的一个技术方案储备。

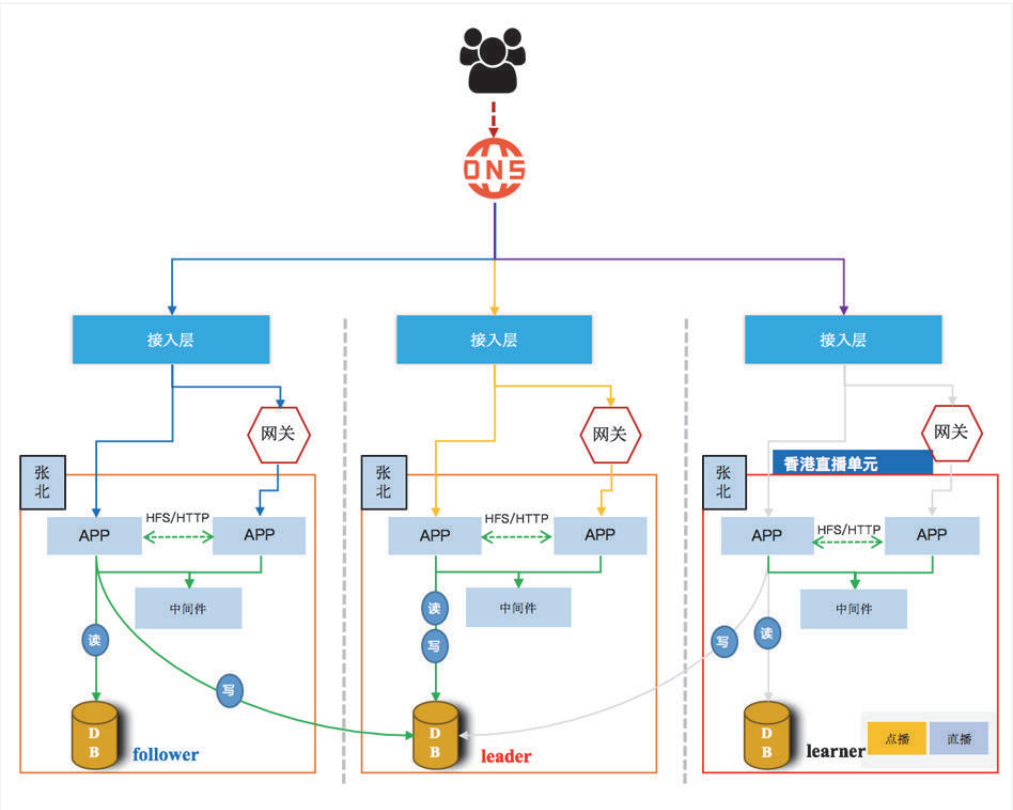
2）单元化收益

一是容灾：多地域容灾。任何一个城市异常，核心交易下单行为能够在秒级内全部恢复，非核心行为在 2 小时内恢复；中心有全量的数据，单元的数据写入后会同步到中心；若单元故障，单元提供的交易流量会切换到中心；如果是中心故障，中心会切换到中心备份环境；

二是扩展性：单元化后，系统扩展不受机房的部署和资源限制，可在不同地域选址横向扩展来满足系统日益增长的需求；在多地部署应用后，按照就近接入原则：用户请求落到离它最近的站点，提升用户体验；

三是稳定性：单元快速部署和验证，可在一个空的单元引流来验证功能，降低系统风险；可以对一个空站点做全链路压测，能快速得到站点容量；

四是成本：减少系统对机房部署的强依赖，提供更大的灵活性；能够将单元的规模，所需的资源信息确定下来，用环境的标准化快速部署来节省成本。



3. 部署压测兜底演练

在直播单元压测完成，扩容完成后，进入直播链路专项压测阶段。在这个阶段，对存在性能风险的场景和链路进行放大流量的集群层面的压力探测，对集群层面的指标 CPU(avg/max)、Load(avg/max)、RT(avg/max)等进行监控，并对直播入口首页播放页兜底压测，力求发现系统潜在风险。2019 年优酷 Java 应用测试团队梳理出多个存在性能风险的链路和场景，进行了专项压测保证。单链路专项压测的压测方式不尽相同，这一部分是最灵活和最有效的性能与稳定性筛查。

4. 成功率验证

多场次直播拉取数据对比验证各场次播前成功率是否符合预期，分析不符合预期的国家进行相应的回源调度调整。双 11 当天整体表现稳定，接入层和各上云应用 OPS、成功率、RT 均符合预期。

三、卡顿率改进

1. 开启智能档

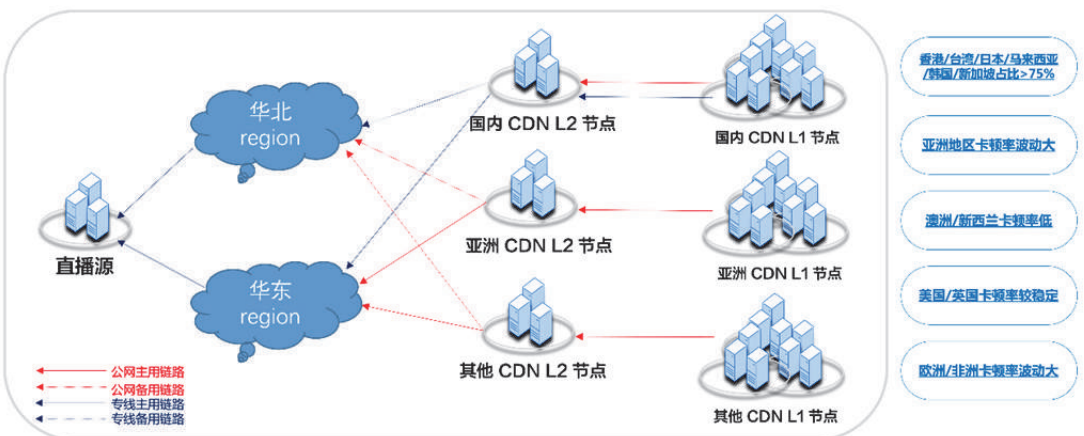
今年的双 11 天猫晚会，优酷 APP 上清晰度列表中有了智能清晰度。智能清晰度是什么，简单来讲，就是自动调整实际播放的清晰度，而调整的依据就是用户当前的 IP，网络状态等信息，结合用户所处网络环境为用户选择合适的清晰度，减少卡顿，并提升播放体验。智能档，也就是码率自适应，优酷在点播场景下已经用了起来且相对较成熟，而今年的双 11，我们将其应用在直播的场景下，进一步提升用户的播放体验。猫晚海外各端默认全开智能档。目前优酷的码率自适应实现，基于主流的 HLS 协议，进行分片级别的切档。

2. 直播链路优化

优化了协议栈，客户端播放使用 HLS 协议，开启 TS 预热。CDN 边缘节点调度优化，边缘节点回源按地域国家统一调度到指定的二级回源节点。

3. L2 节点优化

L2 节点回源站，优先走高速通道。高速通道故障后，切到备用链路。其中东南亚国家 L2 回源链路由公网回源切到了香港 L2 高速通道回源。保障了直播回源链路的稳定。



4. 卡顿率验证

通过开启智能档，调整边缘节点的调度以及 L2 回直播源站的链路。通过多场次直播验证，整个播放卡顿率比预期得到极大的改善。

四、资源成本效能

通过建站平台完成海外直播单元一键建站，单机压测平台输出单机能力，从而评估出应用扩容所需资源。将所有扩容应用输入平台，一键完成应用扩容。切流平台保障单元间容灾时平滑切换。这么复杂的直播链路，通过智能业务链路，对上千个接口进行自动监测、梳理以及容量的自动评估。其节点部署及交付是小时级别，从成本、用户体验上领跑行业，让海外播放页面及互动游戏体验和国内一致。

有了上述对关键点的梳理，那么一套服务于大促资源全生命周期保障的平台系统 就应运而生。确定目标如下：

1）资源需求：需求收集-单机能力评估-历史活动数据对比-工单生成-批量扩容执行-压测调整-结束资源回收，全链路 100%平台化流程化；

2）单机压测：单机压测能力覆盖所有大促应用，实时提供最精准的能力数据；

3）业务巡检：资源健康度巡检，低水位低利用率低 OPS 应用数据输出，自动资源回收，提升资源利用率。

1. 一键扩缩容

非核心应用资源动态支援能力，10 分钟 1000 台回收交付能力；

整体能力分为两个大类：资源需求线和资源保障线。

2. 总体流程

明确需求收集范围->改进需求收集方法->实现单机能力自动获取->实现历史容量数据自动获取->应用上下游依赖链路自动获取-> [业务目标->技术目标转换]->完成资源需求评估。实现需求上报渠道能力，单机实时压测能力，目标转化能力，上下游链路及流量平衡自动评估能力。

3. 资源保障总体流程

建设整体资源容量盘点能力->建设应用级别线上水位巡检能力->优化资源快速交付/调整能力->建设非核心应用应急容量挪用能力 ->快速回收资源 ->完成闭环的生命周期，实现高效交付，强化资源盘点及 buffer 保障的能力。

五、重保预案及措施

1. 主站重报预案及措施

接入层水位高：水位达到预定阀值执行自动扩容。不需要人工干涉；

用户超预期：降低码率，执行限流；

单机房出口故障：平台自动执行切流，切到正常的机房；

单元公网出口故障：切流平台执行切流，切到正常单元。

2. CDN 重保预案及措施

预案执行演练、兜底策略执行演练、大盘实时监控、快速故障处理应急小组。

六、项目总结

每次参加双 11 战役，都收获颇丰。精准的计算资源用量、技术方案评审到落地，都要经过多次反复验证，通过每年的双 11 技术沉淀到平台，大大减轻了人力成本。单机压测平台提供单机的能力从而计算出集群的水位，建站平台提供主站及单元部署的一键建站；扩容平台提供大促一键扩容及一键缩容能力，切流平台提供机房容灾切换能力，运维机器人提供各监控项的快速查看；资源成本平台提供大促成本能力。后面平台优化的地方还很多，从主站到 CDN 实现全链路无缝对接，借助双 11 充分发挥平台优势。

和阿里文娱学“技术资源成本管控”

作者| 阿里文娱技术专家 杨琦

一、背景

2017 年，阿里文娱对业务机器资源做了管理控制，业务方按量按需申请使用，完成了机器的分配、上线、下线回收及预算管控的自动化。但随着业务的发展，其他资源用量的需求在逐渐增加，机器资源的成本占比逐渐减小，因此急需对业务使用的全量资源进行管控。我们对用到的技术产品、业务方进行了综合的调研分析，发现几个突出问题：

1）技术产品分散，多个技术体系都有各自的计量、账单模块；

2）产品收费规则查询费时费力，难以快速预测一次技术方案的成本；

3）业务方对技术成本关注度较低；

4）资源产品生命周期未闭环，部分产品资源没有涵盖在统一运维管控系统。

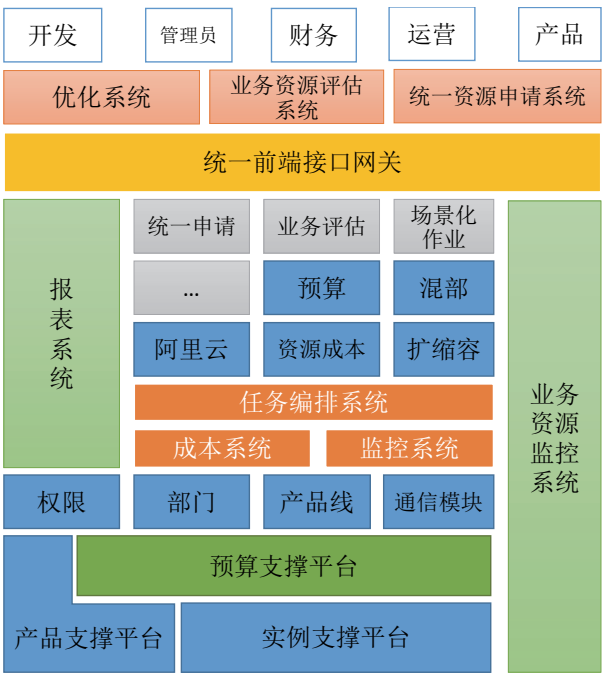
二、文娱成本管理解决办法

（图：资源全生命周期管控平台）

基于上述问题，优酷应用架构团队计划在资源使用的前、中、后三个阶段分别制定策略，构建从资源申请到释放的全生命周期监控体系。

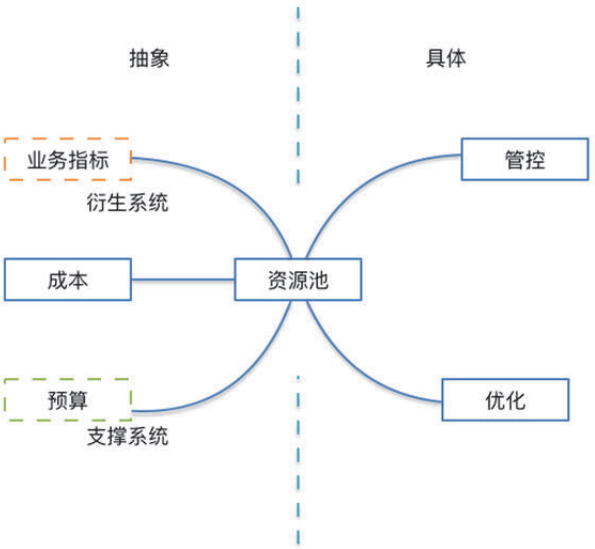
1. 资源池构建及资源运营

资源分散在多个技术体系，将各个技术体系资源计量及成本收口到一个入口，方便财务、业务查询用量，减少平台间的切换成本。



（图：统一平台基础架构）

技术资源实例、归属 OWNER、归属部门、计量单位、账单单位等通过产品 ID 与实例 ID 关联，搭建“统一资源池”，涵盖产品、实例、业务团队支撑，具备资源用量数据的统一查询输出能力，基于资源池的构建，可以输出实例、产品、团队的资源明细/占比，以及分析用量趋势等数据。



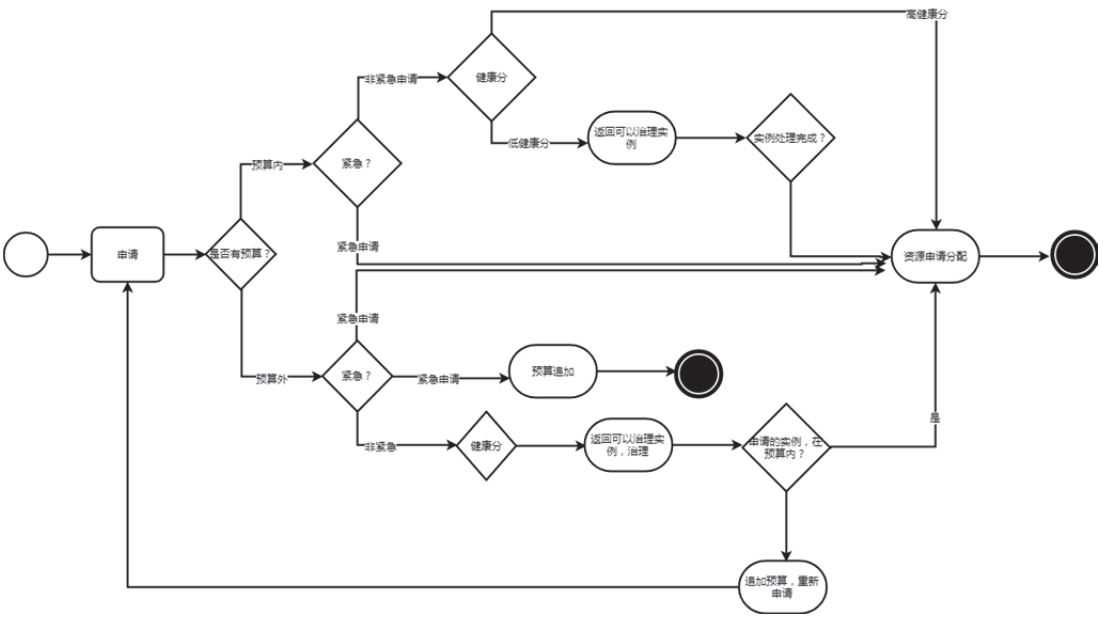
（图：资源池）

2. 资源申请管控

存量及用量数据虽然完成了收录，但还无法具备资源管控的能力，因为真正的“口子”还是开放的，业务方资源需求是直接提交到资源方系统的。在资源方有资源，而业务方有“足够业务诉求”的场景下，资源方会尽量满足业务需求，那么这个入口就必须管控起来，基于业务提交的财年预算数据和资源池采集存量数据，将资源申请审批下放给业务 PE、主管进行有效把控，在入口处减少资源不必要的申请浪费。

3. 资源下线、治理

对于有状态（数据）的资源产品实例，业务下线后，对业务关联的实例根据实例监控指标，确认实例是否还在使用，根据周期的检测反馈，对存量实例进行相应的操作，同时根据资源属性、监控指标等设置资源“健康分”的百分制体系，根据实例得分趋势确认该实例是否还在使用，各个产品根据属性设置报警阈值，进行自动化预警，尽可能做到 T+1 内通知异常资源负责人进行处置，完善资源规范使用，避免资源低效、浪费使用。



（图：资源申请流程概览）

在存量、申请都有管控后，完善运维平台业务与资源关联，类似容器，和业务是强相关，如果业务下线那么容器也将回收。

通过以上的功能设计，我们构建了资源全生命周期流程，规范资源的使用，通过资源用量透明化、资源申请管控等措施，加强了业务对资源成本、预算的认识，加深了大家对资源成本的理解。

三、文娱成本管理收益

1. 资源用量、成本透明化

基于技术产品资源池，我们搭建了资源运营体系，可以满足日常所需的计量明细、团队成本、历史用量趋势等多维度的数据查询需求，同时每月初按照业务分布及团队分布，输出上月的资源用量、资源成本占比及团队趋势排名等，让技术团队一目了然地掌握技术成本用量。

2. 大型活动成本支持

每年优酷会有很多大型活动，双 11 猫晚、跨年晚会、春晚等常规直播外，还有如世界杯等

超大体量的活动。成本系统上线前，每次活动前的资源评估耗时较久；资源成本系统上线后，可以参考历史活动的 DAU 等指标，根据运营提供的预估人数，输出活动的技术资源、成本的预估，为活动进行提供强有力的支撑；且在活动结束后的 T+1 内输出活动的人均成本、及资源成本明细数据，为技术复盘提供数据支持。

详解 Ops 智能运维机器人，故障处理又快又准!

作者| 阿里文娱高级开发工程师 见乔

一、背景

对优酷来说，核心业务全年需要有很高的业务可用率。对于故障处理则有 1-5-10 的目标，即 1 分钟发现、5 分钟定位、10 分钟恢复。当前我们的技术架构越来越复杂，线上的一次请求，可能会经过非常复杂的调用链路，当业务出现问题时，如何快速发现和止血，是当前系统运维体系的核心点之一。

在稳定性建设这条路上，我们已经沉淀非常多的经验：监控预警、业务链路、变更查询、日志查询分析……每一个故障排查手段都对应了可能不止一个运维平台。所以 PE 在故障处理时所面临的问题，不是没有平台或工具，而是平台太多，想要在 5 分钟快速定位线上问题，非常考验每个 PE 的能力。

随着人工智能在全球领域越来越热，ChatBot 作为应用场景之一，它的功能也日益强大。为了让 PE 的运维方式更加智能，优酷应用架构团队，也借助钉钉机器人，在 ChatBot 领域展开了运维领域的实践，为 PE 量身打造了一款智能运维机器人。

二、常用场景介绍

运维机器人通过聊天的方式，智能处理用户的输入，将运维结果快速反馈给用户。用户不用关心众多运维平台和具体术语概念，只需要聚焦于运维对象和运维操作，机器人会帮助用户去处理一些脏活、累活和重复性工作。另外，机器人天然具备秒级响应、一触即达的能力，以及在移动端的优势，更是让每个 PE 都能随时随地通过手机进行运维，大大提高了故障响应能力。

运维机器人在优酷的具体使用场景：

1. 变更操作

1）实例重启：线上实例异常时，想快速重启应用服务，只需要在钉钉里告诉机器人想重启哪个实例即可；

2）实例替换：线上实例异常时，如想直接替换新的实例（如替换容器），只需要在钉钉里告诉机器人想替换哪个实例即可；

3）订阅应用发布：开发、测试同学，经常因为各种原因需要关心上下游某些应用的发布情况，可以提前在钉钉群里告诉机器人想订阅哪个应用的发布情况，在该应用开始发布时，机器人就会在钉群里通知，相关同学便能在第一时间判断此次发布的影响范围。

2. 应用查询

1）应用信息查询：想最快知道一个应用的相关信息？直接将应用名发送给机器人即可；

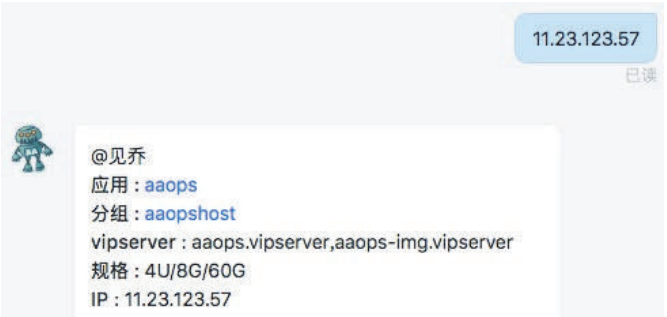
2）Java Dump。

排查 Java 应用的线上问题时，经常需要 Dump 堆栈或堆内存信息来进行分析。直接告诉机器人想 Dump 哪个实例上的应用即可快速 Dump。

3. 系统网络

1）IP 查询

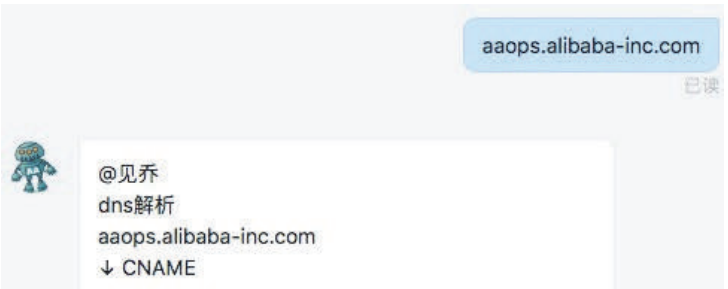
无论是外网还是内网 IP，都可以发送给机器人进行查询。



（图：IP 查询）

2）域名查询

无论是内网还是公网域名，都可以发送给机器人快速查询相关信息。



（图：域名查询）

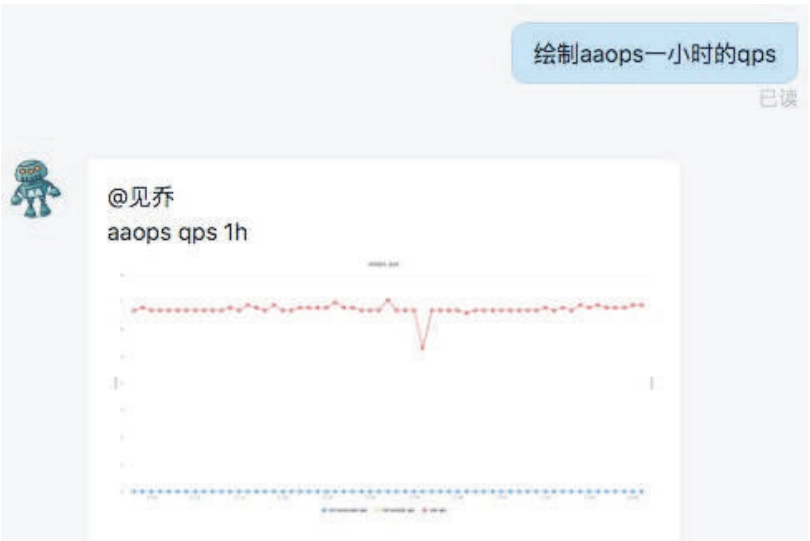
3）VIP 查询

通过机器人，快速查询 VIP 信息，及 VIP 下 RS 的信息。

1. 监控预警

1）系统监控图绘制

可以通过告诉机器人相关指标及时间，快速查看对应的监控指标图。



（图：监控图绘制）

2）异常诊断

底层对接了集团的日志分析平台，可以快速诊断应用是否存在异常日志。

三、技术实现方式

1. 消息收发

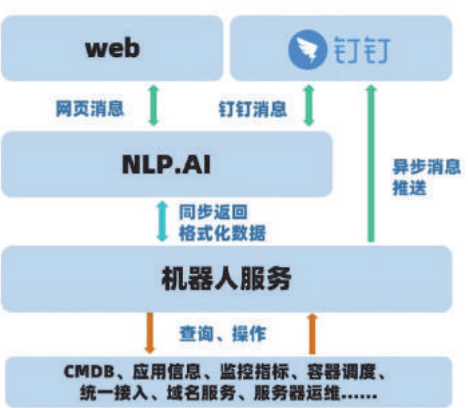
在实现上，机器人主要是依赖了钉钉提供的群机器人功能来进行钉钉消息的收发。

钉钉支持 Incoming 机器人和 Outgoing 机器人。在企业内部往往有很多自研后台系统，例如 CRM 系统、交易系统、监控报警系统等等，有时候大家想把这些自研系统的事件同步到钉钉的聊天群，通过钉钉的 Incoming 机器人，就可以快速实现这个功能。只需要在钉群里创建群机器人，通过向群机器人的 webhook 来发送特定消息格式的请求即可实现。

而对于复杂的运维场景，仅仅通过 Incoming 机器人做消息推送还不够，所以我们用到了 Outgoing 机器人，根据具体的运维场景，来定制用户与机器人的交互。当用户@机器人时，钉钉将用户发送的消息内容实时地发送到机器人服务上。

2. 请求处理流程

机器人服务会针对用户的输入，首先进行意图判断，如果用户是以自然语言的方式进行输入，首先会通过 NLP 模块解析用户的意图，最终拆分成命令+参数的方式，识别出具体的可处理组件，然后交给组件去处理业务逻辑（例如判断此次输入是进行服务器信息相关查询还是应用信息相关查询），最后由组件根据自身逻辑调用底层的具体服务。

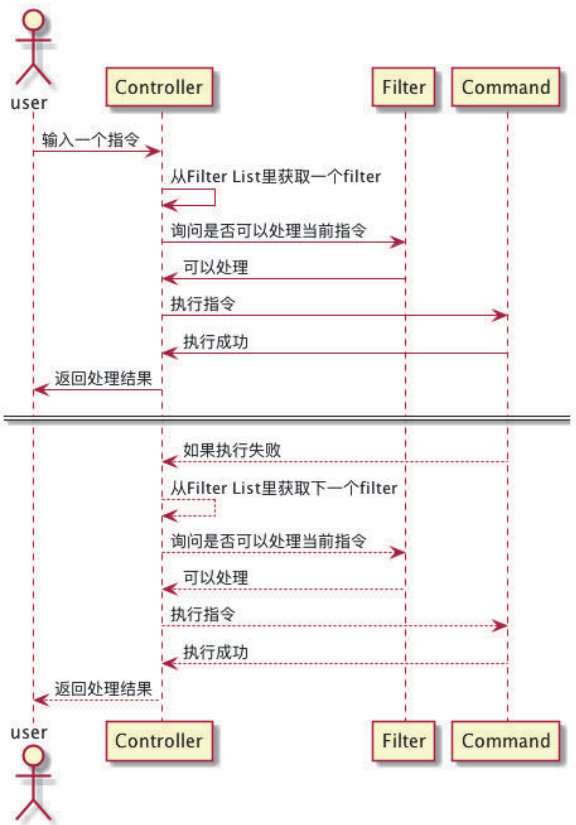


（图：技术架构）

请求处理的整体框架设计比较简单：通过统一的 Controller 进行消息接收，然后遍历每个已加载的组件，首先询问该组件是否能处理该请求，如果可以处理，则交给该组件进行处理；

对于处理失败的组件，会再继续询问下一个组件。

大致流程如下图：

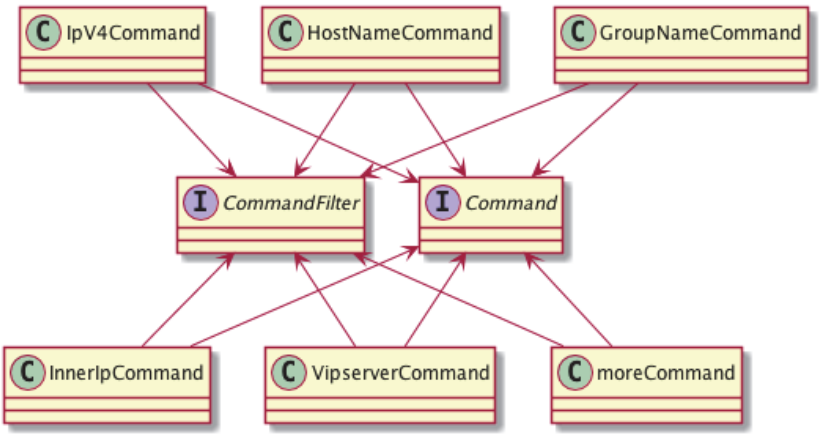


（图：请求处理过程）

对于用户而言，用户不需要了解具体的运维相关概念，例如输入一个 IP 时，用户并不需要关心这是一个公网 IP、内网 IP 又或者是一个 VIP，甚至用户不需要关心自己输入的字符串到底代表什么，因为相应的判断都由各个组件自己决定。

3. 组件接口设计

接口设计上，每个组件都会实现 CommandFilter 和 Command 这两个接口，CommandFilter 主要是回答能否处理某个用户输入，Command 则是具体地处理某个输入：



（图：组件接口）

这样一来，机器人服务也非常利于拓展。对于新增的业务场景，只需要新增一个组件即可。

四、总结

机器人服务向下打通了十几个底层基础设施平台，同时提供了统一的交互方式给到真实用户，解决了长期以来运维难、应急响应慢的痛点。当前在优酷，运维机器人已经成为日常工作、故障应急必不可少的助手。

机器人在智能这块，后续还会持续加强能力。因为对用户屏蔽了一系列复杂的运维平台概念，所以机器人提供的运维服务是否周到，最终还是取决于底层对接的运维平台是否足够全；而用户与机器人交互的体验是否贴心，意图识别是否准确，最终还是要看在 NLP 这块，机器人服务对于用户自然语言输入的预处理能否更加精准。

智能运维的关键：自动化应用容量管理实践

作者| 阿里文娱高级开发工程师 金呈

一、概述

1. 背景

随着业务形态发展，更多的生产力集中到业务创新，这背后要求研发能力的不断升级。我们持续倾向用更加高效、稳定、低成本的方式支持快速软件交付，保障高可用。于是运维力从“HaaS”（Handwork as a Service）到“PaaS”再到“SaaS”（System as a Service），运维生产力从 Ops 到 DevOps 再到 NoOps。

在传统应用容量管理模式下，应用、集群容量评估缺乏有效数据依据与支撑，往往牺牲效率或成本来平衡经验决策风险，另一方面，人肉决策和执行难以满足业务对稳定性和效能的追求。因此急需一个能够把优酷所有应用的容量管理起来的能力。

2. 目标

整体目标分成 2 个阶段，一是摸清各应用容量水平，二是为所有应用赋予弹性伸缩的能力，最终直观看到各应用及总体资源使用率的明显提升。

二、技术挑战与解法

1. 单机性能

既然谈到容量问题，已知的压测方案有链路压测方案、模拟流量压测方案等。为什么还要自研一套基于单机引流的压测方案来评估应用容量水平？

1）更接近日常真实水平；

2）无人工决策，纯机器决策单机性能瓶颈；

3）全自动，比如配置成发布结束后进行单机性能压测。

2. 弹性

弹性指标选择：仅靠集群 CPU 水位弹性确实可以解决绝大多数类型应用，但若基于集群 QPS 水位则可更精准的进行弹性伸缩。

1）多维度弹性指标，同时也需要支持自定义指标；

2）多方位弹性方案，使用条件编排策略来达到多个弹性指标之间的协同。

三、技术方案

1. 全自动单机性能探索

与各接入层对接，自动配置权重完成单机引流，配合性能拐点算法支撑，完成自动识别性能拐点并终止压测，最后一并记录单机能力值。并允许每天定时压测、发布结束压测来感知每天、或每次发布给单机性能带来的变化，使用户更亲近自身应用的容量水平。

2. 弹性伸缩

与底层交付系统联动，打造从规划，交付，计费，弹性水平扩展、回收、资源排布调度全生命态面向业务需求的自驱动统筹调度资源管理系统。一方面，业务资源统一平台构建与维护，挖掘空闲资源，共享弹性计算力，整体提高部署密度，降低业务单位成本；另一方面，面向不同应用场景，自动化容量管理，按需分配，保证应用服务可用性，提高容量运维效能。

3. 方案结合

全自动单机性能探索与弹性伸缩的结合框架，如下图：

1）应用接入：无缝接入；

2）单机性能压测自定义配置：无需再定义配置过多配置，默认“智能探索”可以支持大部分场景；

3）单机引流：从集群中其他机器的流量引流至被压测机器；

4）智能拐点识别：使用时间序列数据趋势转折点提取算法，进行拐点识别；

5）单机性能预测：详见技术细节；

6）基础弹性配置：详见技术细节。

四、技术细节

1. 单机引流权重优化

1）调整权重就是调整单机流量，且权重越高，单机流量越高；

2）增加自动化调整权重策略方法。

权重优化：用于已经识别出拐点，保证下一次压测接近 MAX 权重保持平缓；

权重递增：用于未触发拐点，保证下一次压测能引更多的流量。

2. 响应时间拐点识别

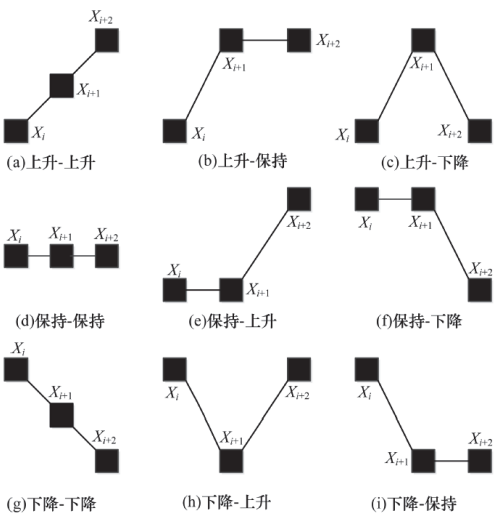
使用算法：箱线图。基于 IQR 定制多组 k 的箱线图上限的异常提取，上限=Q3+k*IQR 实现。而效果能够定位到多数拐点，并且一般拐点前的一个时间点的值为单机能力值。

3. 成功率拐点识别

0 错误代表 100%成功率

成功率拐点识别相比响应时间拐点识别更加严格。虽然同样是“基于 IQR 定制多组 k 的箱线图”实现，但此时 k 必须收紧，因为成功率指标较为敏感，稍有波动就应该终止压测。

4. 拐点提取参考了“时间序列数据趋势转折点提取算法”文章



比如在 3 个连续点 Xi、Xi+1、Xi+2 的判定上，它们的发展趋势共有 9 种情况：当 Xi+2-Xi>0，即图 a，b，e 这 3 种情况属于总体趋势上升，当 Xi+2-Xi<0，即图 f，g，i 这 3 种情况属于总体趋势下降，当 Xi+2-Xi=0，即图 c，d，h 这 3 种情况属于总体保持不变。

而我们通过“基于 IQR 定制多组 k 的箱线图”可以识别出上升和下降 2 种拐点，分别对应不同的场景，如响应时间拐点识别（上升拐点识别），成功率拐点识别（下降拐点识别），而 k 的定义方式也参考近期数据。

比如某个应用日常响应时间稳定在 100-200ms 和某个应用日常响应时间稳定在 2-3ms 的 k 值是不一样的，不合适的 k 用在 2-3ms 的这种数据上会导致异常识别较为频繁及不准确。

5. 单机性能预测方案

单机性能与什么有关，系统指标？如果是 JAVA 应用还和 JVM 相关指标有关？而应用本身有会有各种池的限制，如 JVM 相关池、TOMCAT 相关池、DB 相关池、Redis 相关池、队列相关池等，这些都可以作为预测单机性能的特征。先基于 PCA 抽象出 N 个特征，也称降维，可将两两线性相关的因素进行整合或排出，降维后建立线性回归模型，而拟合度较高的模型将予以采纳并进行预测。同时预测参数也需要实事求是，比如日常 CPU 区间为 2-60%，那预测参数可以为 80%，此时若超过 100%那将毫无意义。

6. 流量驱动弹性方案

基于 CPU 指标的弹性伸缩：比如 CPU 超过 60%则执行弹性扩容，CPU 低于 20%则执行弹性缩容。扩容与缩容允许按机器数比例进行伸缩：如按 5%的机器数进行弹性扩容。定义弹性区间：如 10-20，机器数会在 10-20 区间变动。

一般低峰期会处在最低机器数区域，高峰则会处在最高机器数区域，基于外挂单机能力模型。允许基于 QPS 水位指标进行弹性，可随 QPS 增加而增加机器数，反之则减少机器数。

五、总结

自动化容量管理与弹性伸缩的深度结合解决了当前容量预估的问题，使得资源能够被合理的使用。一方面，用户专注业务层，做基于业务需求的容量规划、交付和维护，革命性改变生产关系，提高研发迭代效率；另一方面，更加细粒度的弹性伸缩，比如小时、分钟的资源的快速流转，资源粒度分解到具体硬件计算垂直伸缩，也是一种更优的解决方案，使得弹性更加迅速能做到秒级能力，进一步压缩集群密度，降低单位成本。

关注我们

（阿里文娱技术公众号）

关注阿里技术

扫码关注「阿里技术」获取更多资讯

加入交流群

1）添加“文娱技术小助手”微信

2）注明您的手机号 / 公司 / 职位

3）小助手会拉您进群

By 阿里文娱技术品牌

更多电子书

扫码获取更多技术电子书