
Architect
SACC
2022 中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2022
激发架构性能 点亮业务活力
云上会议 网络直播 | 2022年10月27-29日
ITPUB

58同城深度学习推理平台 基于Istio的云原生网关实践

58同城 AI Lab 魏竹斌

个人及部门简介

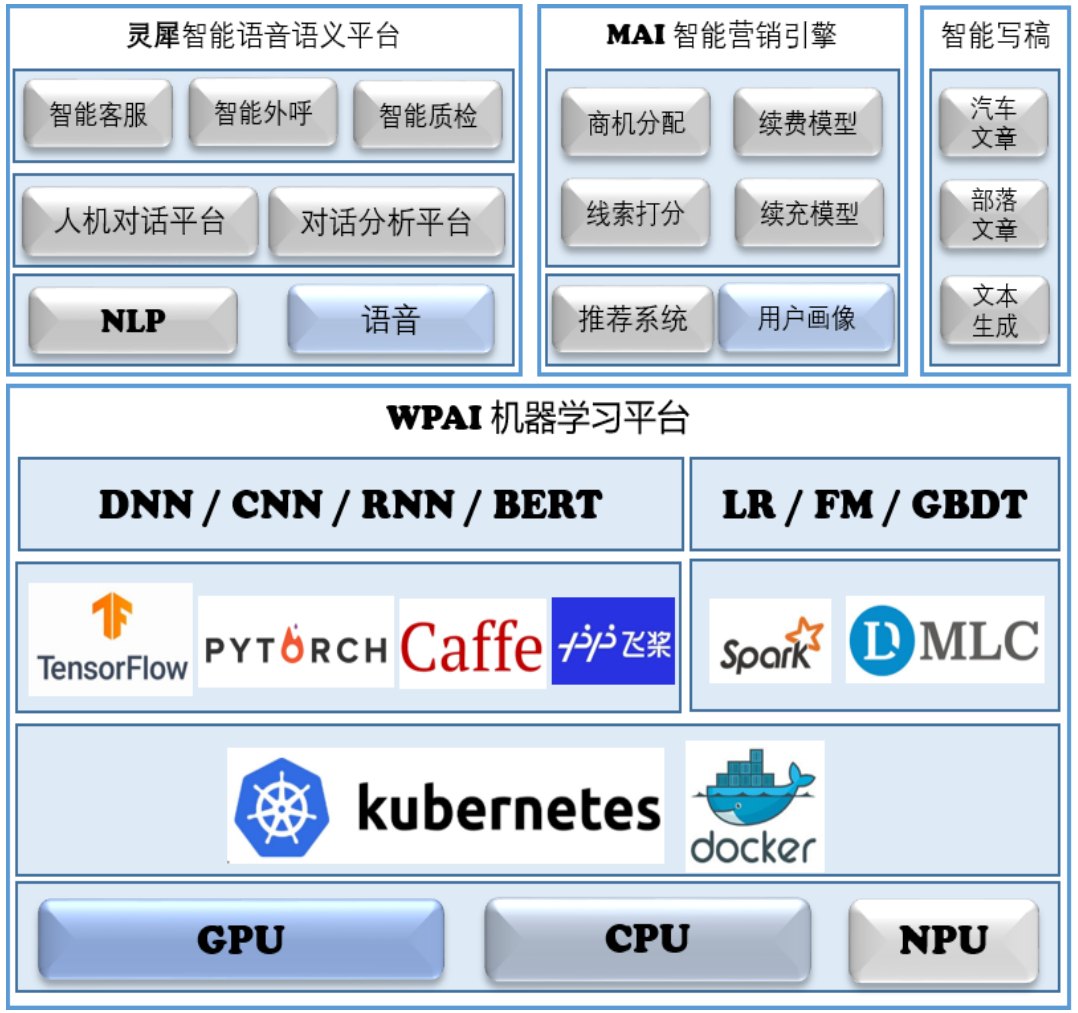
魏竹斌

- 58同城深度学习推理平台负责人，2020年加入58TEG-AI Lab，先后负责推理加速、向量检索平台和深度学习推理平台
- 硕士毕业于中国矿业大学（北京），曾就职于北斗航天集团从事后端开发工作

AI Lab

- 2018年5月21日成立，隶属于58同城TEG技术工程平台群
- 旨在推动AI技术在58同城的落地，打造AI中台能力，以提高前台业务人效、收入和用户体验

AI Lab公众号

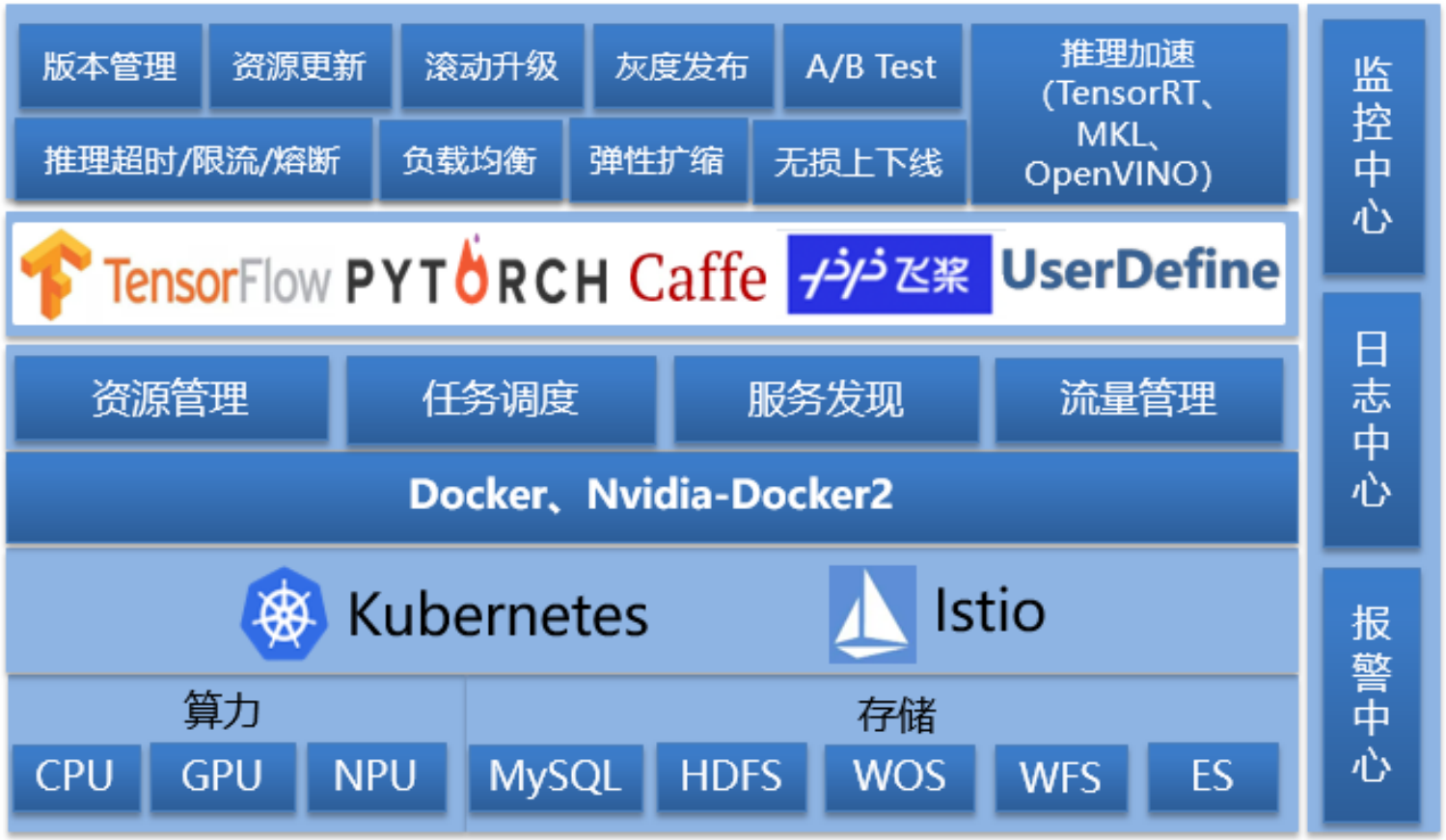


AI Lab产品技术架构

推理平台Istio云原生网关应用实践

- 推理平台1.0架构实现及不足
- 推理平台2.0架构设计及效果
- 2.0架构下的流量治理能力建设
- 2.0架构下的可观测能力建设

深度学习推理平台整体架构



- 平台定位

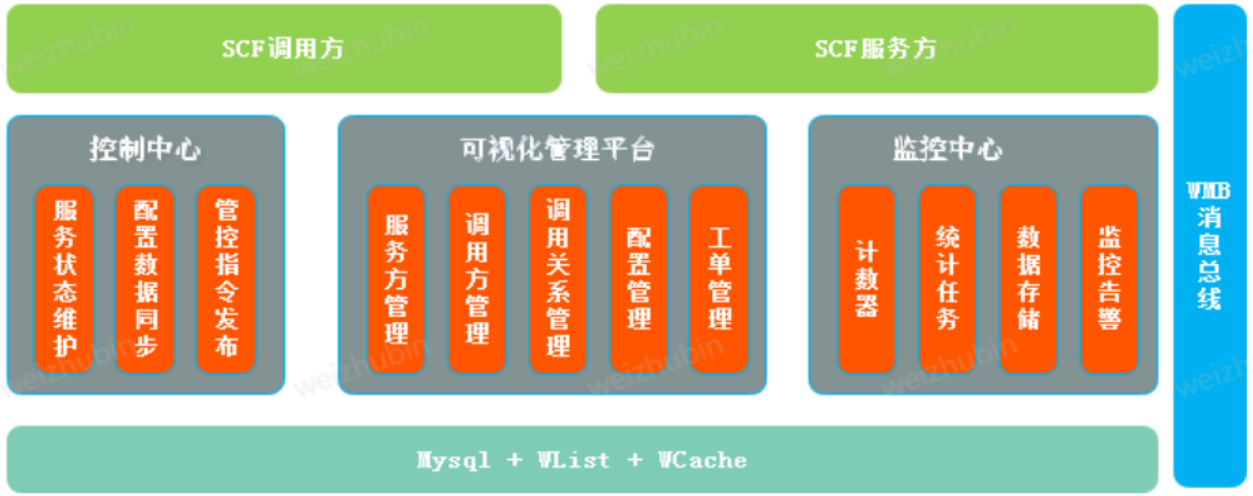
  将算法人员使用深度学习框架训练出来的模型部署到生产环境，提供高性能、高可用的在线推理服务

- 应用情况
  - 上线模型数1000+
  - 运行节点数4000+
  - 日均请求量30亿+
  - 峰值QPS6.6万

推理架构1.0实现背景

- 集团对AI平台化能力的迫切需求
  - 各业务部门为实现AI应用落地目标各自为战，但因为缺乏平台化能力，导致研发、运维效率低下
  - 算法人员深陷工程泥潭，模型迭代效率低

- SCF具备成熟的服务治理能力
  - 58自研java系RPC框架
  - 服务节点自动注册与发现
  - 负载均衡、服务鉴权
  - 全方位监控、完善的告警配置

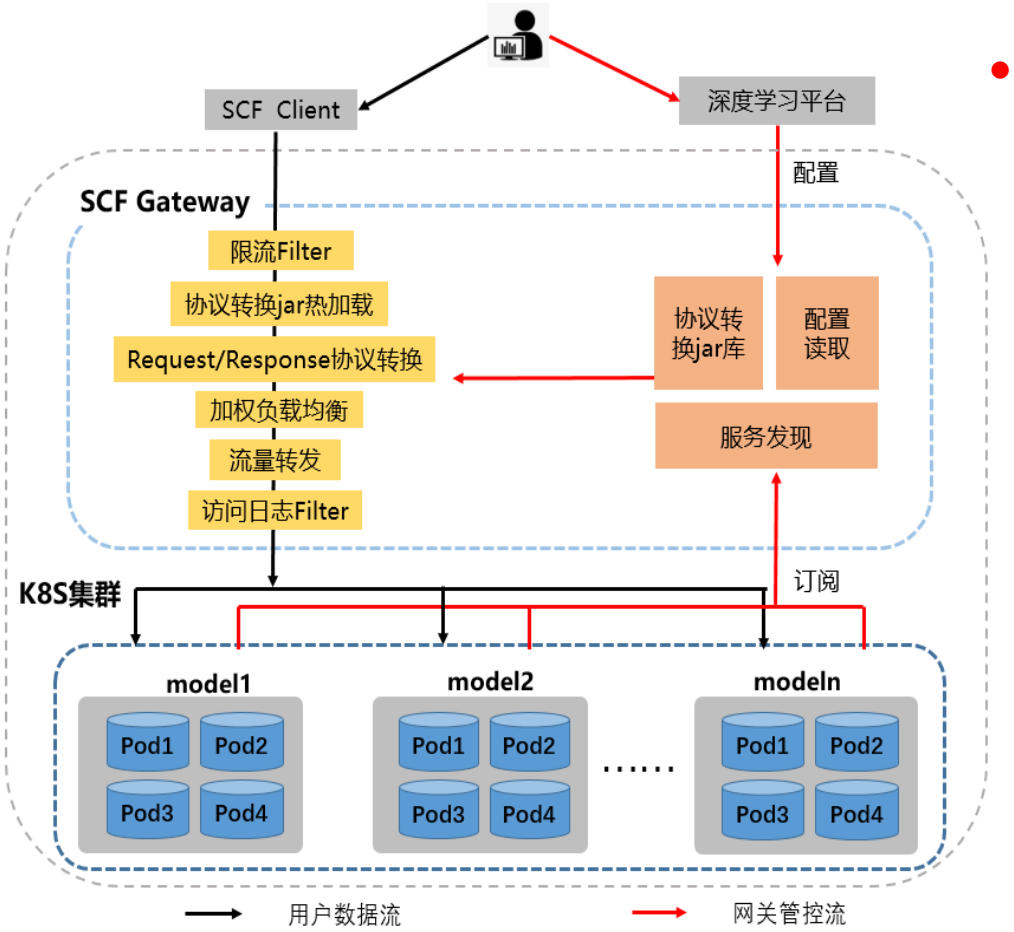


SCF架构图

推理架构1.0实现

- 数据面逻辑
  - 基于preFilter实现任务鉴权、秒级限流功能
  - 通过定义协议转换接口 + 类加载器机制支持不同协议转换策略热加载
  - 通过加权负载均衡算法实现服务熔断容错功能
  - 基于postFilter统一日志输出与异常处理方式

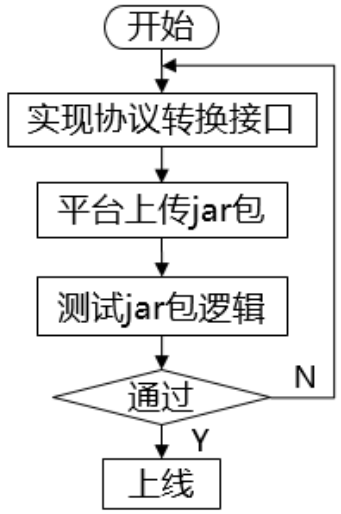
```
/**
 * 在线推理request/response协议转换接口
 */
public interface IPredictOnlineIntercepter {
    /**
     * 将SCF输入数据转换成PredictRequest
     * @param requestData SCF接口请求数据
     * @return tensorflow-serving推理请求数据
     */
    PredictRequest predictOnlineBefore(List<Object> requestData);

    /**
     * 将推理返回结果转换成SCF数据结构
     * @param response tensorflow-serving推理返回数据
     * @return SCF返回数据结构
     */
    Object predictOnlineAfter(PredictResponse response);
}
```



- 控制面逻辑
  - 基于K8S List/Watch机制实现服务发现功能，构建upstream连接池
  - 通过WConfig(58自研配置中心)及时同步任务参数的变更
  - 通过WOS（58自研对象存储）打造协议转换jar插件中心

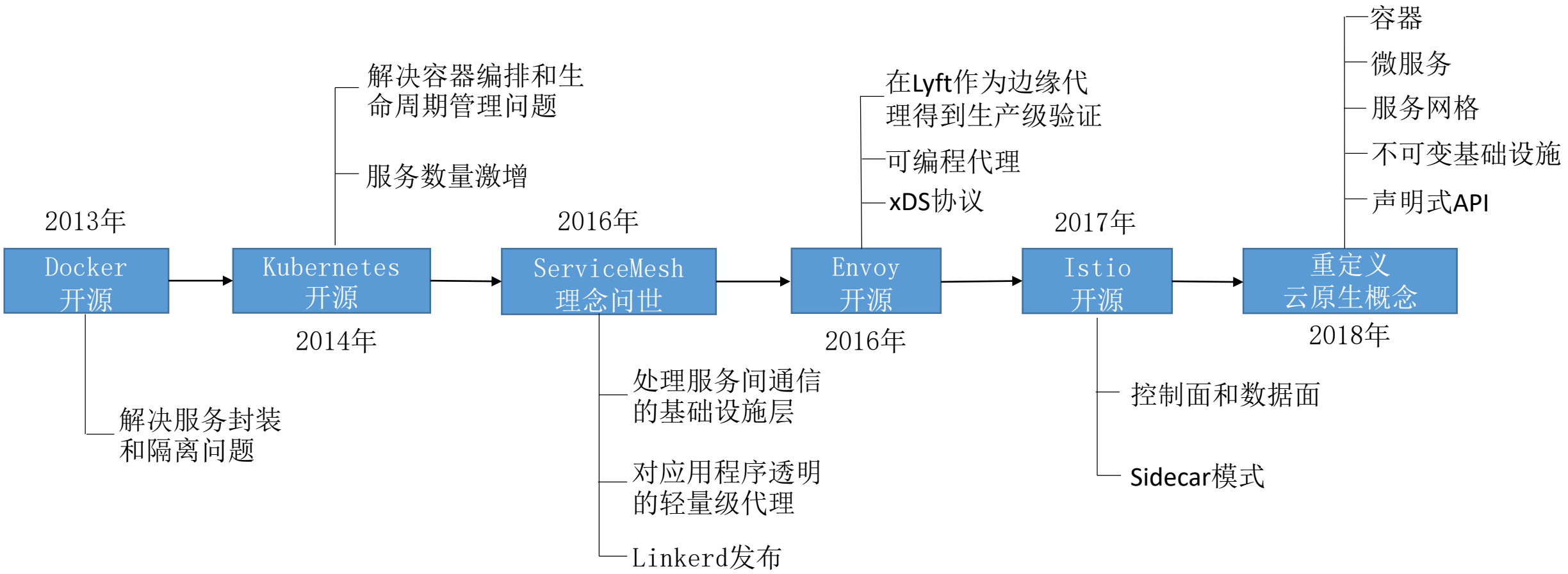
推理架构1.0不足

- 业务接入
  - 接入流程复杂（如右图），增加了算法人员调试成本
  - 接入方式单一，不支持HTTP方式接入



- 服务性能
  - SCF与gRPC请求协议互转延时损耗大
  - 底层Netty SocketChannel自适应缓冲区内存配置策略对size较大请求不友好，SCF客户端连接数直接决定服务端老年代内存占用，随着接入规模增加会因为gc问题导致性能抖动
- 开发运维成本
  - 与第三方库紧密耦合，集成新功能或第三方库升级都需要对网关进行整体升级，成本较高

Istio的诞生


2013年
Docker
开源
解决服务封装
和隔离问题
2014年
Kubernetes
开源
解决容器编排和生
命周期管理问题
服务数量激增
2016年
ServiceMesh
理念问世
处理服务间通信
的基础设施层
对应用程序透明
的轻量级代理
Linkerd发布
2016年
Envoy
开源
在Lyft作为边缘代
理得到生产级验证
可编程代理
xDS协议
2017年
Istio
开源
控制面和数据面
Sidecar模式
2018年
重定义
云原生概念
容器
微服务
服务网格
不可变基础设施
声明式API

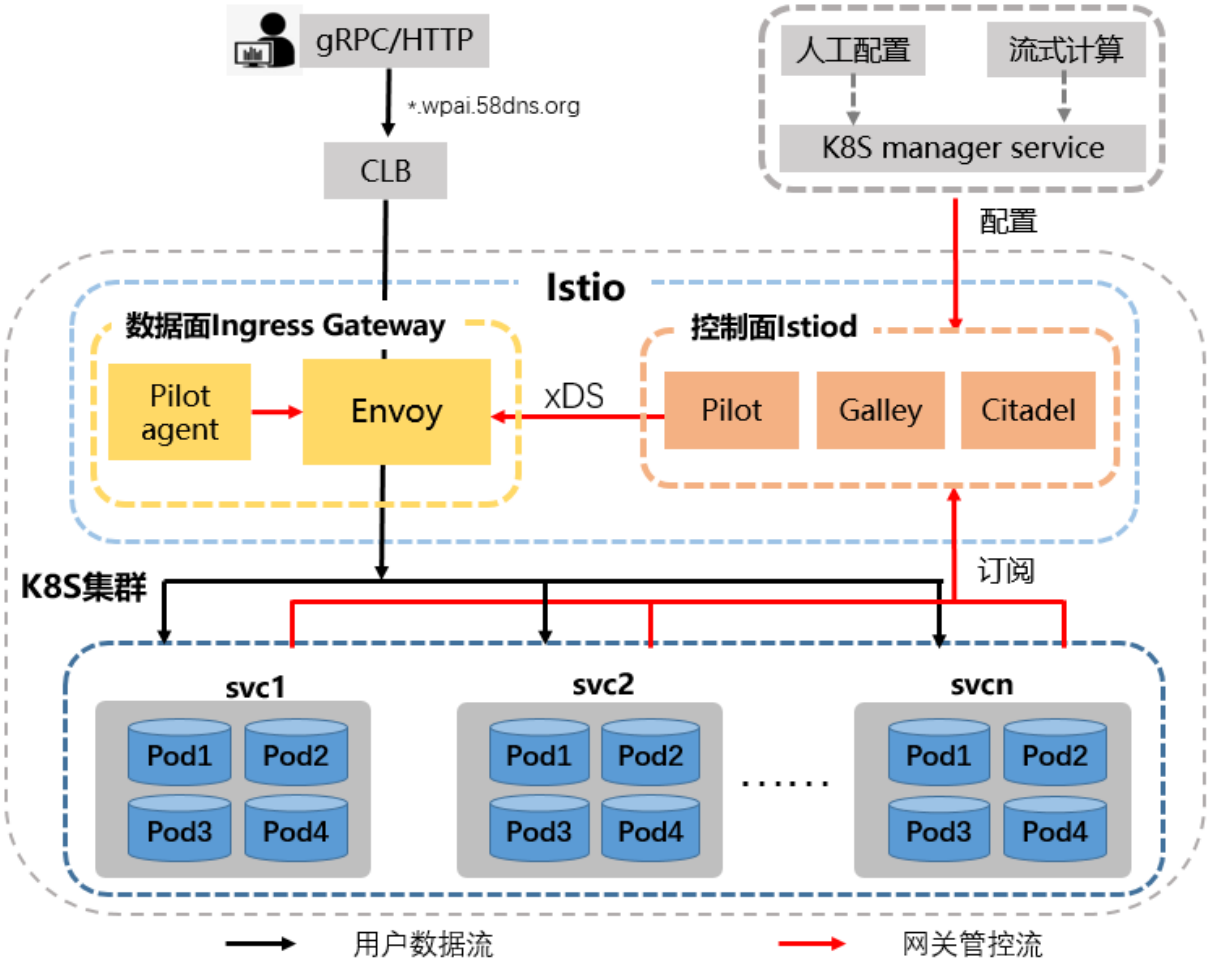
网关新解法：Istio云原生网关

- 优质的基因
  - Envoy作为边缘代理在Lyft公司中得到生产验证，随后成为云原生计算基金会（后简称CNCF）第三个毕业的项目
  - CNCF正式将服务网格（Service Mesh）写入云原生第二版定义 ，Istio也于近期成为 CNCF 孵化项目
  - 控制面和数据面隔离架构，搭配xDS(x Discovery Service)动态配置同步方案
- 全面的能力
  - 强劲的代理性能，基于c++11标准实现，全异步事件机制驱动
  - 丰富的流量治理能力，如请求路由、负载均衡、超时、限流、熔断等，开箱即用
  - 强大的可观测性支持，具有详细的监控指标，完整的访问日志
  - 灵活的可扩展性，可以基于Filter、Lua和WASM方式增强功能
- 强劲的势头
  - 2020年CNCF中国云原生调查显示：去年排名第四的Envoy近1年内使用量明显上升，从15%的份额增长到29%，超过F5和HAProxy跃居第二
  - Istio/Envoy在谷歌、微软、阿里、腾讯等等国内外头部公司大规模落地应用，已然成为服务网格/数据面代理的事实标准

推理架构2.0设计实现

- 以业务部门为粒度做多租户隔离，通过域名 + CLB组合实现网关侧负载均衡和高可用

- 端到端推理场景，没有东西向流量需求，为不影响推理性能及降低运维复杂度，所以服务节点未注入Sidecar代理



- 封装K8S manager service，做为业务操作K8S+Istio资源统一入口,标准化操作行为

- Istio控制面专注服务信息与策略配置的分发，数据面依据配置高效执行流量管理

推理架构升级后效果

- 通过升级实现了性能、稳定性、易用性的全面提升
  - 更强劲的请求转发性能，推理耗时相对于原始架构减少了50%以上

序列号	数据大小	请求方式	平均耗时(ms)	耗时相对减少
1	0.6M	新架构	6.08	58.86%
2		原始架构	14.78	
3	7.12M	新架构	74.89	60.08%
4		原始架构	187.59	

  - 数据面和控制面从部署层面实现资源隔离，功能更加内聚、服务更加稳定
  - 提供丰富、开箱即用的流量治理功能，极大地方便后续开发、运维工作

推理架构升级后丞待解决的问题

- 原架构下通过SCF分组能力实现的多业务方资源隔离功能，新架构下如何实现？
- 原架构下通过SCF服务逻辑代码实现的灰度发布、A/B Test功能，新架构下如何实现？
- 原架构下通过SCF服务逻辑代码实现的推理超时、秒级限流、节点动态加权负载均衡功能，新架构下如何实现？
- 离在线混部、推理服务自动扩缩容功能的应用使得服务节点上、下线操作变得频繁，如何保证上、下线期间请求不受影响？
- 原架构下通过SCF框架+服务管理平台提供的可观测性支持，新架构下如何实现？

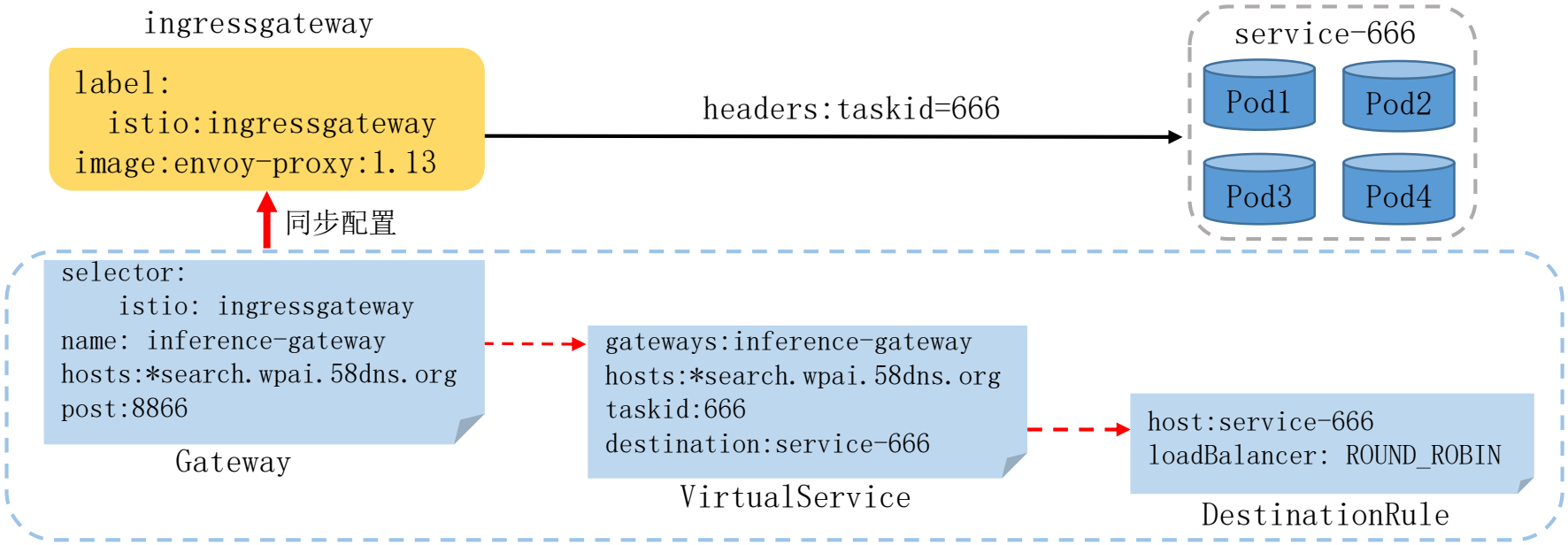
……

推理平台Istio云原生网关应用实践

- 推理平台1.0架构实现及不足
- 推理平台2.0架构设计及效果
- 2.0架构下的流量治理能力建设
- 2.0架构下的可观测能力建设

Istio流量治理基础-声明式API

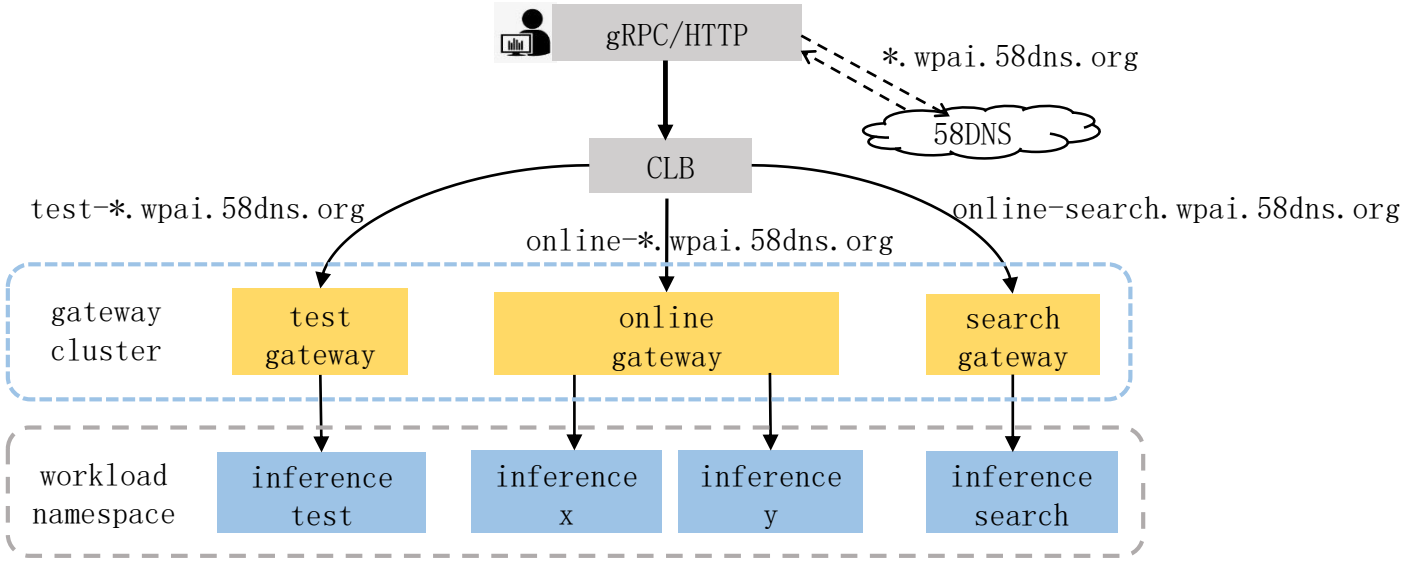
- Gateway 抽象网关在L4-L6层负载均衡属性，例如暴露端口、协议等
- VirtualService 配置L7层路由策略，基于请求自身特征路由到特定Service
- DestinationRule 定义路由发生后更精细流量控制策略，例如Endpoint级别负载均衡
- EnvoyFilter Istio插件机制，定制Envoy请求处理逻辑，例如服务Metrics统计、限流等



声明式API架构

网关多租户实现-Gateway拆分

- 依据流量特征拆分网关，减少网关故障爆炸半径
- 网关集群与工作负载命名空间之间是1:1或1:N关系，平衡推理质量与网关资源使用率



网关集群拆分示意图

```
spec:
  selector:
    istio: ingressgateway-search
  servers:
  - port:
      name: grpc-inference
      number: 8866
      protocol: grpc
    hosts:
    - inference-search/online-search.wpai.58dns.org
```

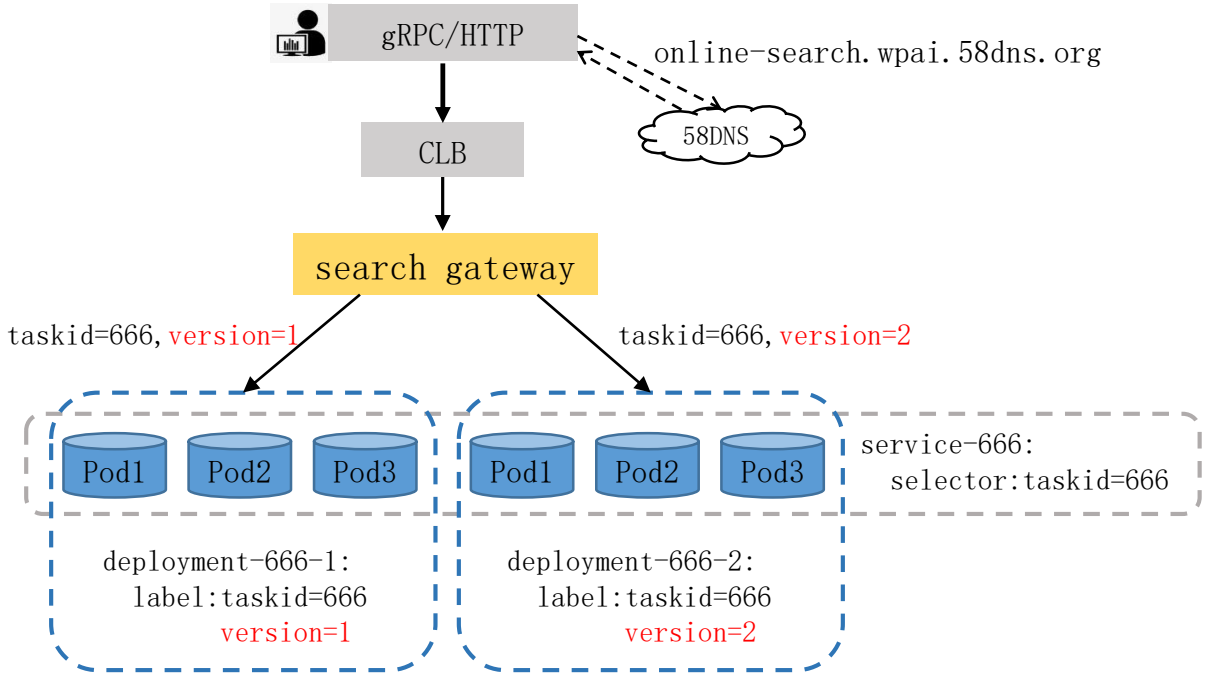
search gateway配置示例

```
spec:
  selector:
    istio: ingressgateway-test
  servers:
  - port:
      name: grpc-inference
      number: 8866
      protocol: grpc
    hosts:
    - inference-test/test-ai-lab.wpai.58dns.org
    - inference-test/test-anjuke-bi.wpai.58dns.org
    - inference-test/test-search.wpai.58dns.org
    ...
```

test gateway配置示例

A/B Test实现

- VirtualService + DestinationRule实现流量精准路由



A/B Test流量路由示意图

```
spec:
  host: service-666
  subsets:
  - labels:
      version: "2"
    name: v2
  - labels:
      version: "1"
    name: v1
```

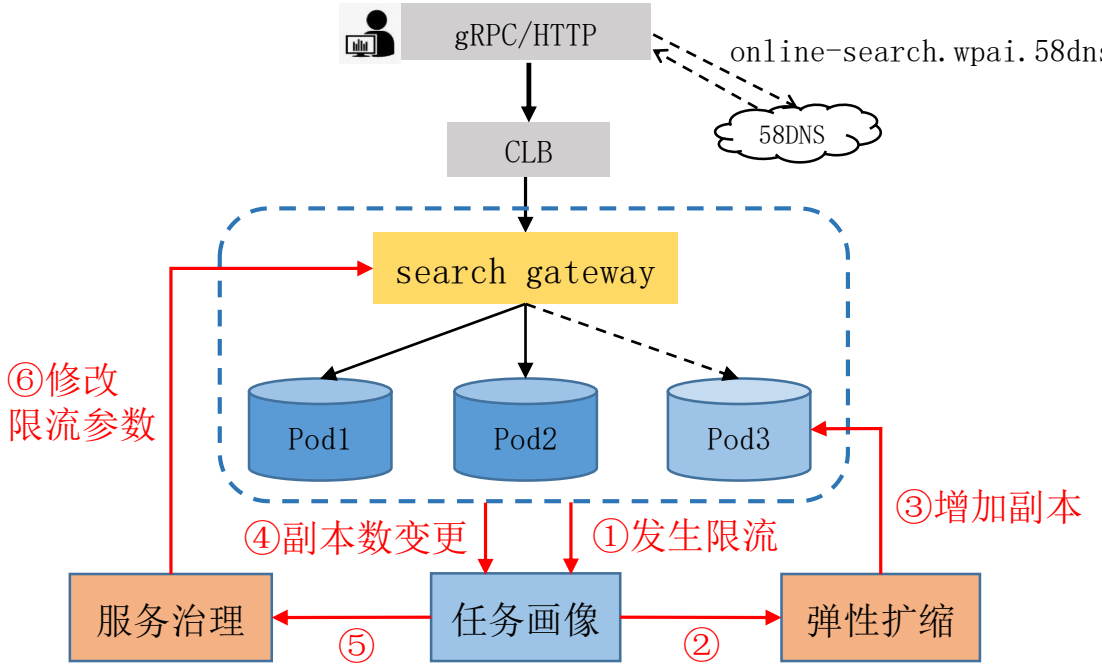
DestinationRule配置示例

```
spec:
  gateways:
  - ingressgateway-search
  hosts:
  - online-search.wpai.58dns.org
  http:
  - match:
    - headers:
        taskid:
          exact: "666"
        version:
          exact: "2"
      port: 8866
    route:
    - destination:
        host: service-666
        port:
          number: 8866
        subset: v2
  - match:
    - headers:
        taskid:
          exact: "666"
        version:
          exact: "2"
      port: 8866
    route:
    - destination:
        host: service-666
        port:
          number: 8866
        subset: v1
```

VirtualService配置示例

秒级限流实现

- 基于EnvoyFilter实现：令牌桶算法 + 本地限流
- 基于任务副本数及TCP链接分布自动调整本地限流参数



本地限流参数自动调整流程

```
spec:
  workloadSelector:
    labels:
      istio: ingressgateway
  configPatches:
  - applyTo: HTTP_ROUTE
    match:
      context: GATEWAY
      routeConfiguration:
        vhost:
          route:
            action: ANY
            name: inference-hdp-teu-dia
    patch:
      operation: MERGE
      value:
        route:
          rate_limits:
          - actions:
            - request_headers:
                descriptor_key: TASKID
                header_name: taskid
        typed_per_filter_config:
          envoy.filters.http.local_ratelimit:
            '@type': type.googleapis.com/udpa.type.v1.TypedStruct
            type_url: type.googleapis.com/envoy.extensions.filters.
            value:
              descriptors:
              - entries:
                - key: TASKID
                  value: "666"
                token_bucket:
                  fill_interval: 1s
                  max_tokens: 888
                  tokens_per_fill: 888
```

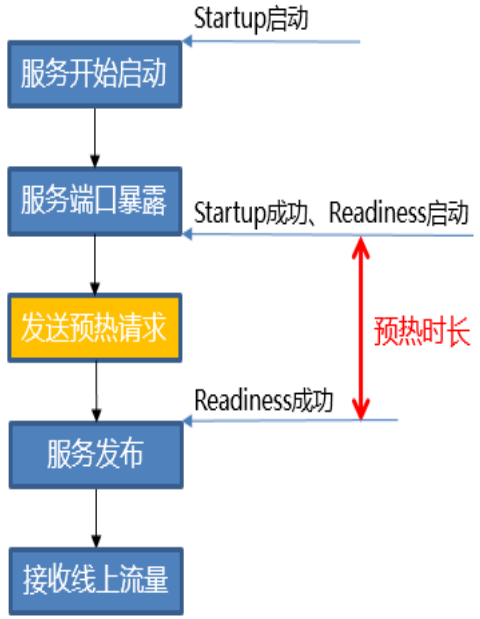
EnvoyFilter本地限流配置

无损上线实现-模型预热

- 服务端口暴露后内部可以发送预热请求
  - 抽象推理请求，形成预热配置规约
  - 不同protobuf打造不同预热客户端

```
{
    "signature_name": "prediction",
    "model_name": "tensorflow-666",
    "batch_size": 1,
    "input": [
        {
            "name": "x",
            "data_type": "float",
            "dims": [1,6,8,3]
        }
    ],
    "output": [
        {
            "name": "y",
            "data_type": "float",
            "dims": [1,1]
        }
    ]
}
```

- 服务端口暴露到发布间有可控时间间隔
  - Readiness探针，决定服务发布及下线时机
  - Startup探针，决定Readiness探针启动时机

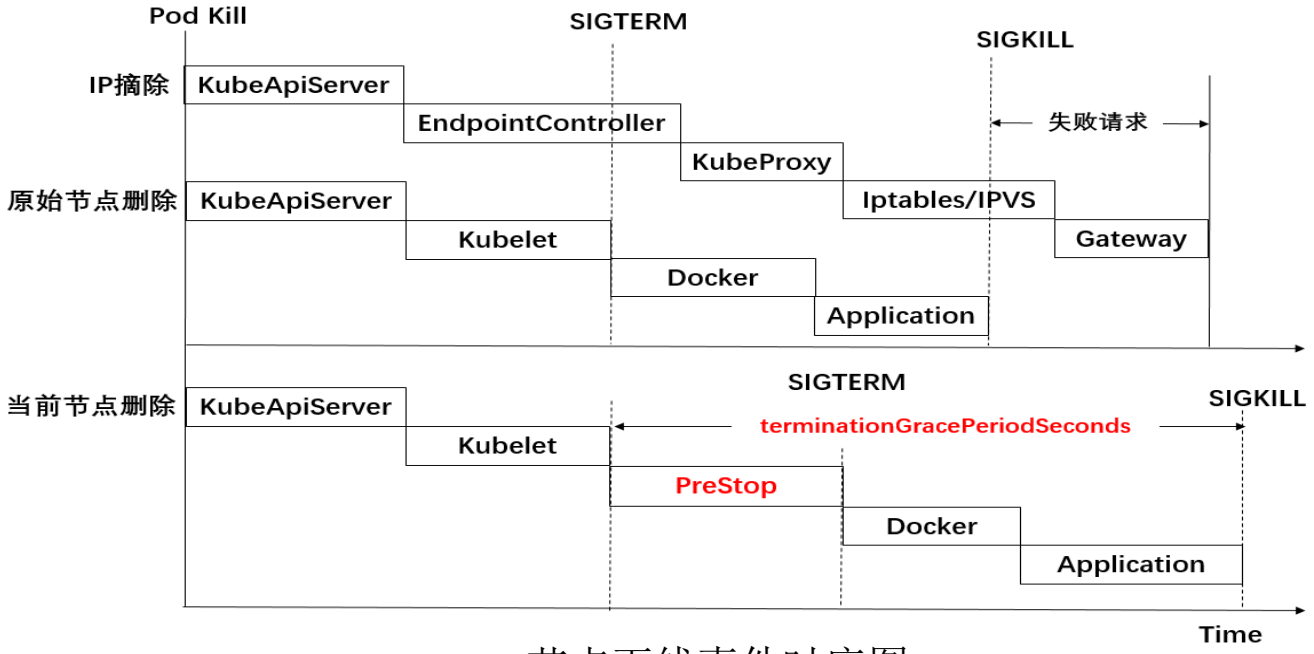



```
spec:
  host: warmup-test-service
  trafficPolicy:
    loadBalancer:
      simple: LEAST_REQUEST
      warmupDurationSecs: 10s
```

DestinationRule预热配置（istio 1.14）

无损下线实现-优雅停服

- 以可执行脚本形式配置preStop钩子，例如休眠等待
- 配置terminationGracePeriodSeconds强杀时间，预防脚本卡死



节点下线事件时序图

```
terminationGracePeriodSeconds: 30
containers:
- lifecycle:
    preStop:
      exec:
        command:
        - /bin/bash
        - -c
        - sleep 10s
```

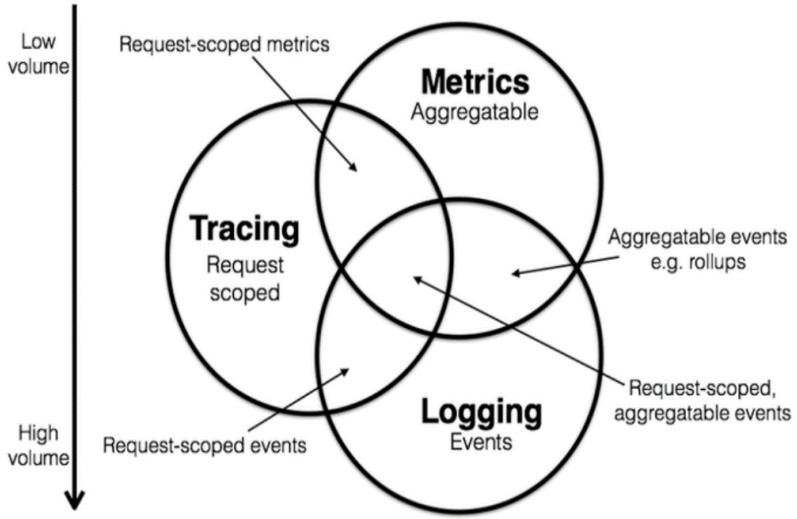
Deployment优雅停服配置

推理平台Istio云原生网关应用实践

- 推理平台1.0架构实现及不足
- 推理平台2.0架构设计及效果
- 2.0架构下的流量治理能力建设
- 2.0架构下的可观测能力建设

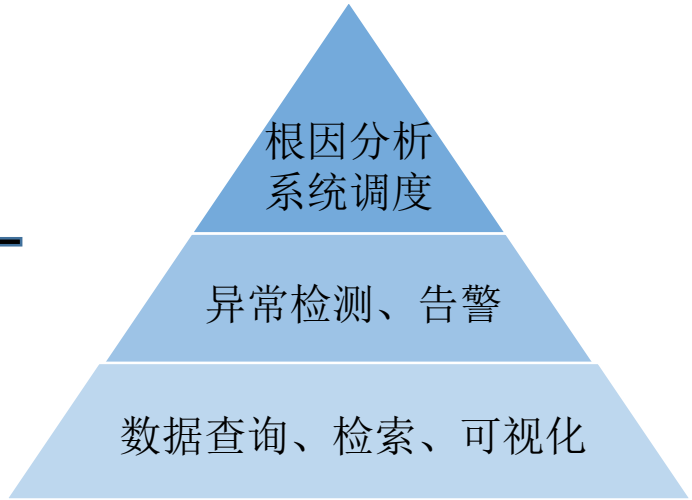
可观测性概述

- 控制理论领域：指系统可以由其外部输出推断其其内部状态的程度
- IT领域：由工具逐渐演变为完整的理论体系，成为管理复杂分布式系统的重要组成部分



数据模型

产品工具



能力金字塔

Istio可观测性支持

- 访问日志（Logging）
  - 基于EnvoyFilter插件access-log实现
  - 可定制日志输出路径、格式和字段内容
  - 通过ELK组件实现采集、传输、存储及可视化

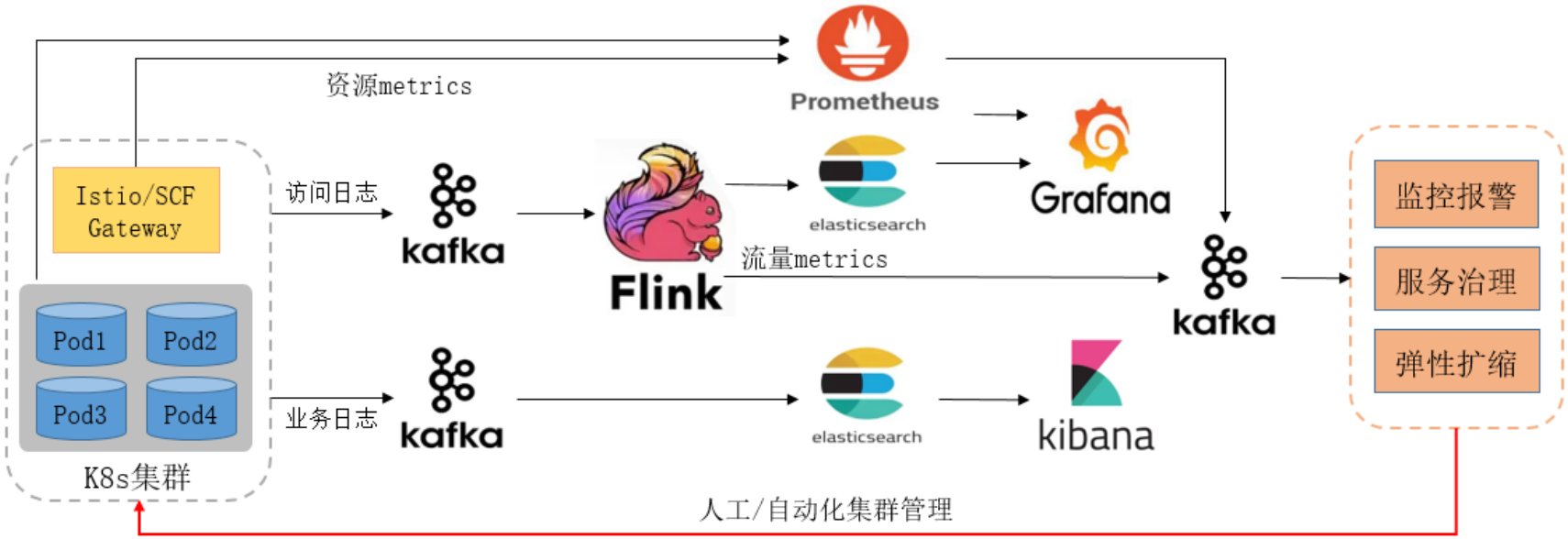
```
access_log:
- name: envoy.access_loggers.file
  typed_config:
    '@type': type.googleapis.com/envoy.extensions.access_loggers.file.v3.FileAccessLog
    path: /opt/isito-gateway-proxy/access.log
    json_format:
      authority: '%REQ(:AUTHORITY)%'
      bytes_received: '%BYTES_RECEIVED%'
      bytes_sent: '%BYTES_SENT%'
      downstream_local_address: '%DOWNSTREAM_LOCAL_ADDRESS%'
      downstream_remote_address: '%DOWNSTREAM_REMOTE_ADDRESS%'
      duration: '%DURATION%'
      grpc_status: '%GRPC_STATUS%'
      hostname: '%HOSTNAME%'
      method: '%REQ(:METHOD)%'
      path: '%REQ(X-ENVOY-ORIGINAL-PATH?:PATH)%'
      protocol: '%PROTOCOL%'
      response_code: '%RESPONSE_CODE%'
```

access-log插件配置概览

- 监控指标（Metrics）
  - Sidecar + EnvoyFilter插件stats-filter-1.xx实现网格内工作负载服务指标计算
  - 以标准Prometheus数据格式提供服务级和工作负载级的请求响应指标，提供Prometheus和Grafana插件负载采集和展示，开箱即用

2.0架构可观测性建设

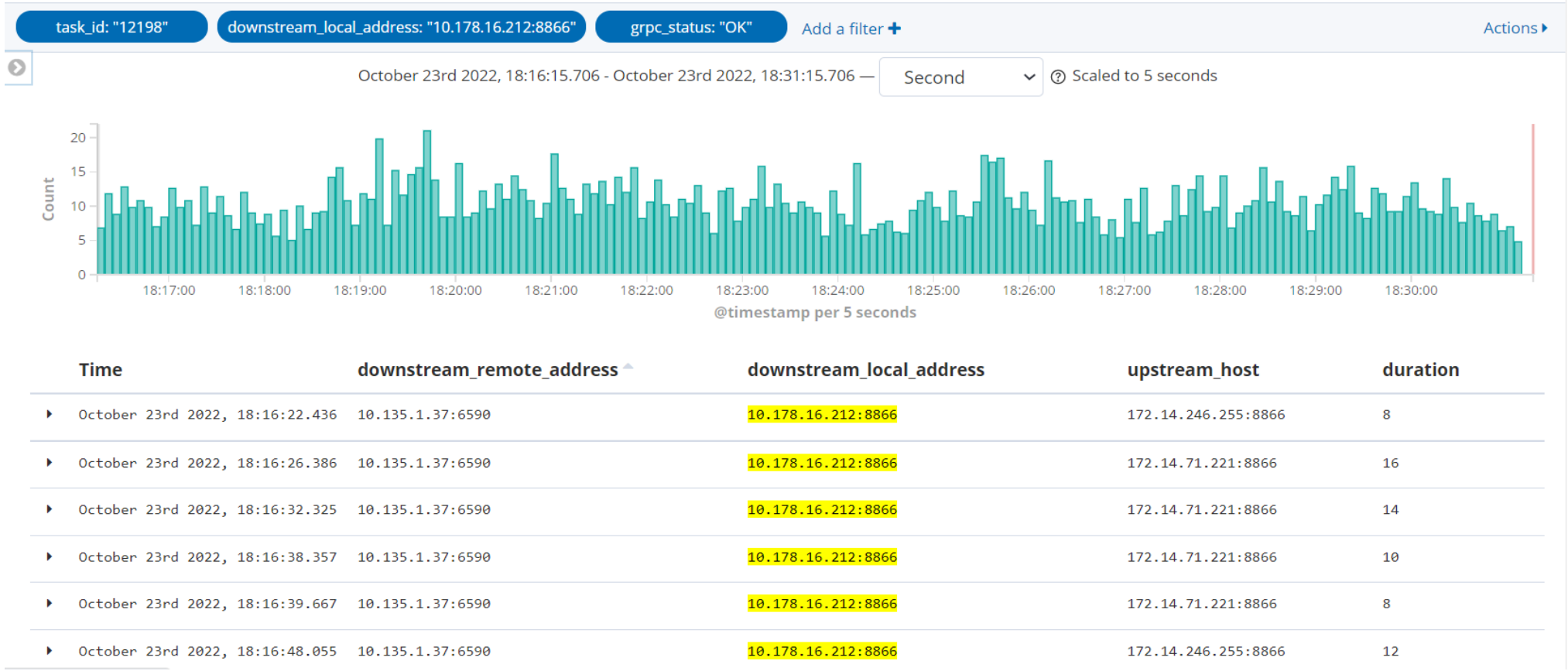
- 网关结构化访问日志(json)及业务非结构化日志皆从磁盘统一采集、存储至ES，通过Kibana检索
- 流量监控指标基于网关结构化访问日志通过流式计算引擎Flink计算所得并存储至ES；资源监控指标由cAdvisor采集计算，Prometheus负责传输和存储
- 监控指标通过Grafana大盘实现查询、可视化工作；并为监控报警、服务治理和弹性扩缩等功能提供数据支撑



2.0推理架构下可观测体系

2.0架构可观测效果展示

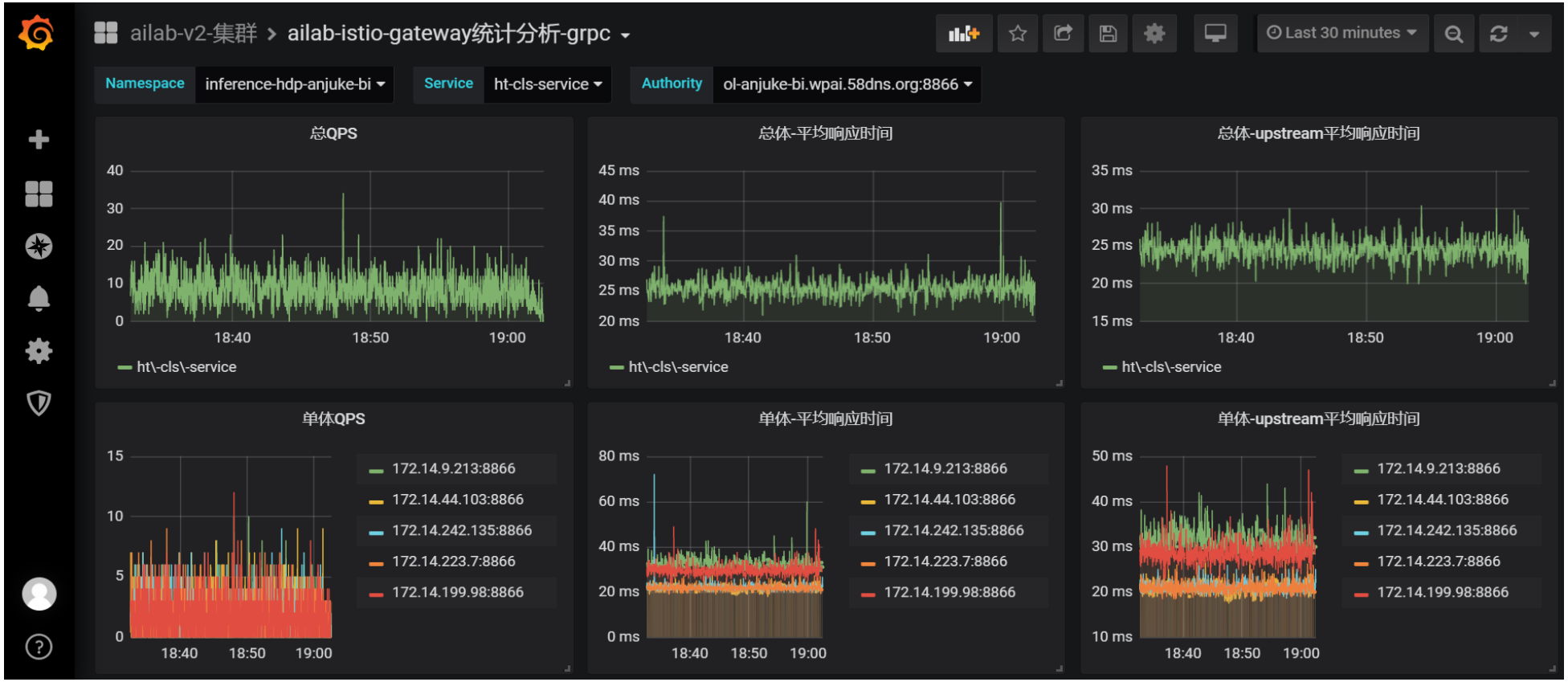
- 访问日志：异常请求第一案发现场，通常用于排查问题发生的根本原因



Time	downstream_remote_address	downstream_local_address	upstream_host	duration
October 23rd 2022, 18:16:22.436	10.135.1.37:6590	10.178.16.212:8866	172.14.246.255:8866	8
October 23rd 2022, 18:16:26.386	10.135.1.37:6590	10.178.16.212:8866	172.14.71.221:8866	16
October 23rd 2022, 18:16:32.325	10.135.1.37:6590	10.178.16.212:8866	172.14.71.221:8866	14
October 23rd 2022, 18:16:38.357	10.135.1.37:6590	10.178.16.212:8866	172.14.71.221:8866	10
October 23rd 2022, 18:16:39.667	10.135.1.37:6590	10.178.16.212:8866	172.14.71.221:8866	8
October 23rd 2022, 18:16:48.055	10.135.1.37:6590	10.178.16.212:8866	172.14.246.255:8866	12

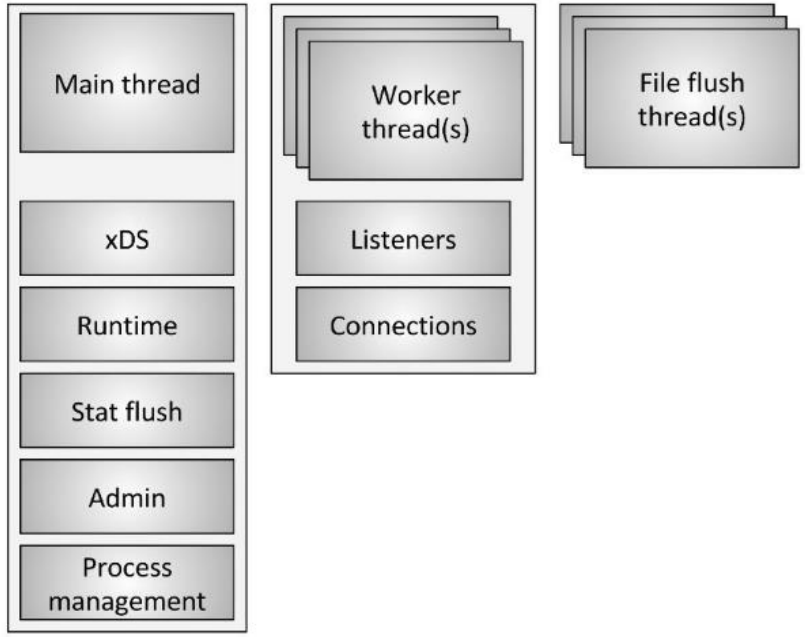
2.0架构可观测效果展示

- 指标监控：展示流量和资源在多个维度的统计信息，用来观察集群工作负载的运行状态和趋势



从Envoy线程模型看网关性能优化

- 线程模型：
  - 主线程：负责所有xDS API处理，运行时，统计刷新，管理控制等功能；单线程执行，各功能间竞争cpu资源
  - 工作线程：负责监听每个侦听器，接受新连接，为每个连接实例化过滤器栈，处理所有连接生命周期内IO事件；提供服务统计数据时需访问“stat store”锁
  - 文件刷新线程：工作线程写文件时，数据实际上被移入内存缓冲区，最终通过文件刷新线程刷新至磁盘。需访问进程范围内锁



Envoy线程模型

- 性能优化措施
  - 简化访问日志格式，减少工作线程锁保持时间
  - 禁用服务Metrics统计功能，使工作线程专注IO事件处理
  - 关闭服务Metrics采集接口，使主线程尽可能专注xDS配置的同步

平台下一步工作方向

- 持续跟进K8S、Istio等基础设施提供的新特性，丰富平台功能，提升推理性能
  - 推理工作负载绑核、绑NUMA部署
  - 测试环境节点自动化压测能力，辅助部署资源申请
- 持续完善平台可观测体系建设，让运维更智能化
  - 实现K8S Event事件、流量Metrics和资源Metrics等数据的联动展示、分析能力
  - 基于可观测数据不断提升弹性扩缩、离在线混部等功能的准确性、及时性

欢迎关注

AI Lab公众号

58技术公众号

开源项目dl_inference

地址：*https://github.com/wuba/dl_inference*

简介：dl_inference是58同城开源的通用深度学习推理工具，可在生产环境中快速上线由TensorFlow、PyTorch、Caffe框架训练的深度学习模型，集成TensorRT、MKL(Math Kernel Library)加速模型推理。

欢迎使用，并 Star、Issue、PR ！

《58同城机器学习平台资源使用率优化实践》《端到端语音识别技术在58同城的探索实践》

《基于微服务架构的智能对话分析平台》《多模态推荐算法在CRM商机推荐系统中的应用》

…

THANKS