爱奇艺AI推理平台演进和实践

张俊钦

爱奇艺深度学习平台研究员

大纲

1. 背景介绍
2. AI 推理平台架构演进
3. AI 推理平台落地优化实践
4. 总结与展望

1. 背景介绍

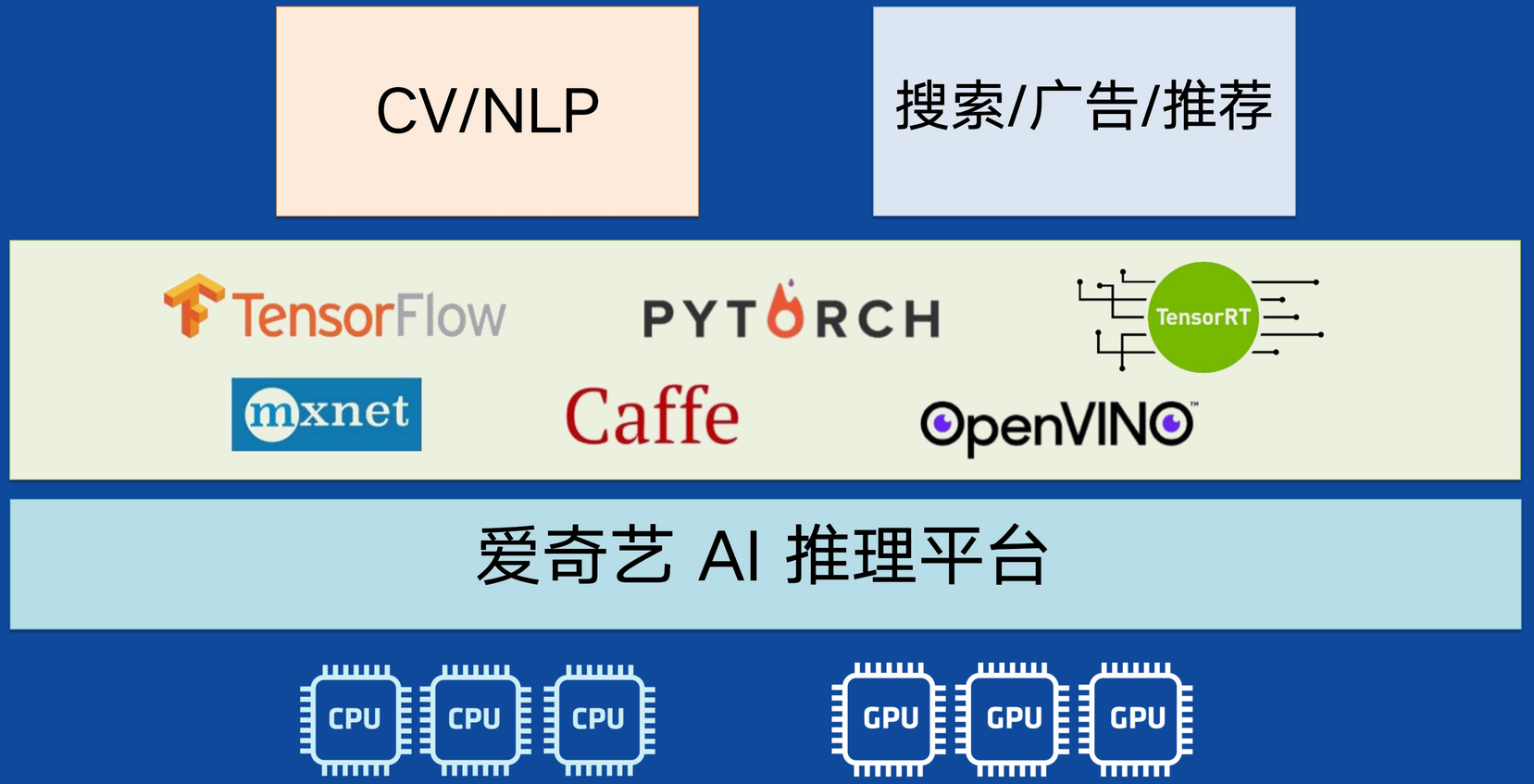
背景介绍

- AI 在爱奇艺的应用场景

AI创作	AI生产	AI标注	AI播放	AI交互	AI分发	AI变现
IP价值评估	热点内容发现	智能标签	热点预测	小艺机器人	智能搜索	个性化广告投放
流量预测	视频指纹	行为	HCDN	智能在线客服	个性化推荐 (短视频、长视频、图文)	创可贴广告
“爱创”智能小媒资	智能审核	场景	奇速播	智能呼叫中心	泡泡社区宣发	TV识商品
智能拆条	自适应编码	物体	自适应码流	Home AI	用户理解	转场点识别
Starworks	智能剪辑	音频	(AI、ABS)	奇巴布AI助手	片段打分	Video In
标题质量评价	智能生成描述关键词	姿态	绿镜（快放）	AR扫一扫	视频质量评估	Video Out
标题辅助生成	动态封面图	自动分类	只看他3.0	手语主播	封面图质量评估	
体育集锦	AI辅助后期制作	智能明星库	Zoom AI	直播劲舞		
	跨媒体内容提取	iCartoon	蒙版弹幕			
	智能插帧	微表情	奇观			
	国剧、老电影修复		智能快进			
			省流模式增强			

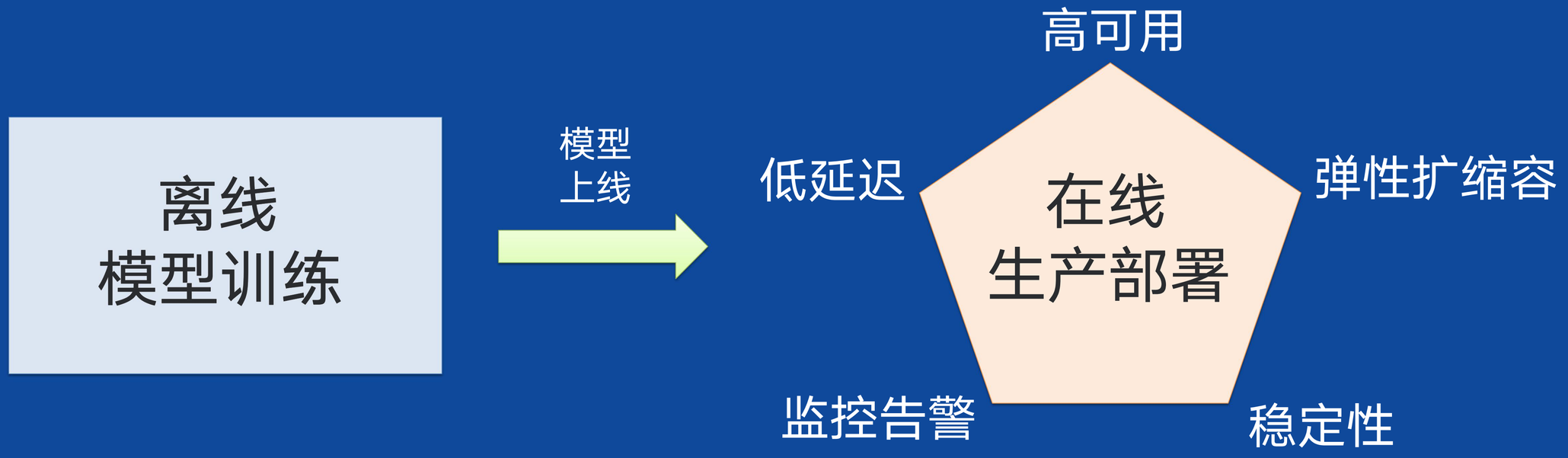
背景介绍

- AI 业务的归类



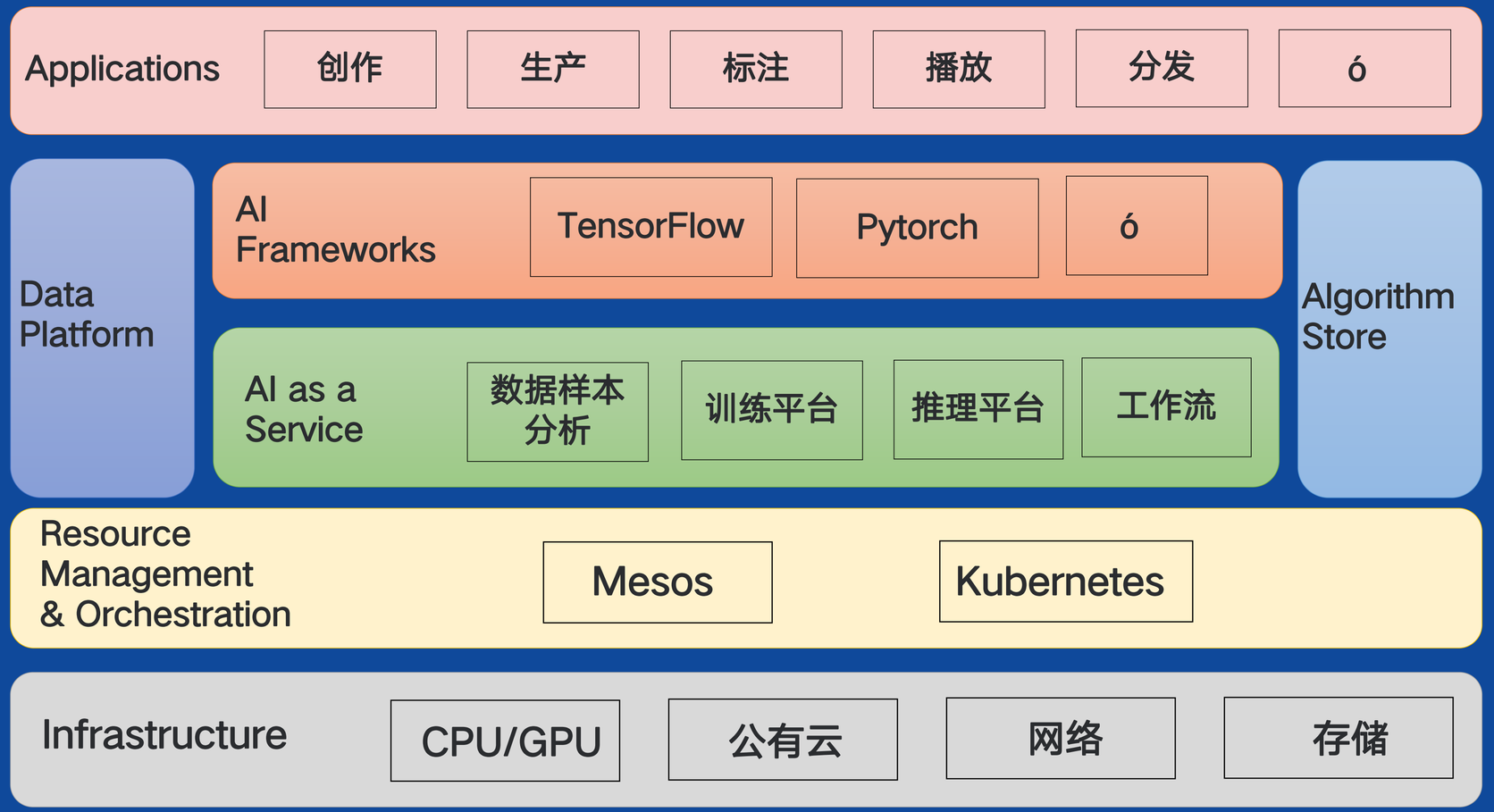
背景介绍

- 从离线模型训练到模型部署



背景介绍

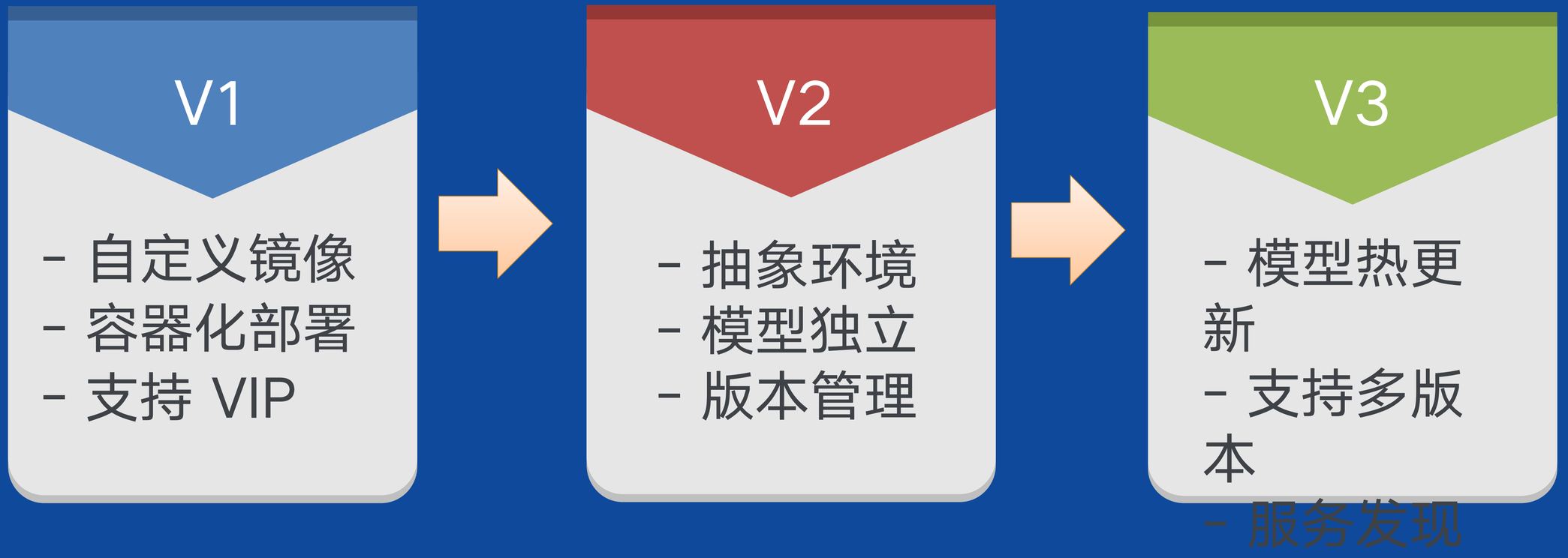
- 爱奇艺深度学习平台总体架构



2. AI 推理平台架构演进

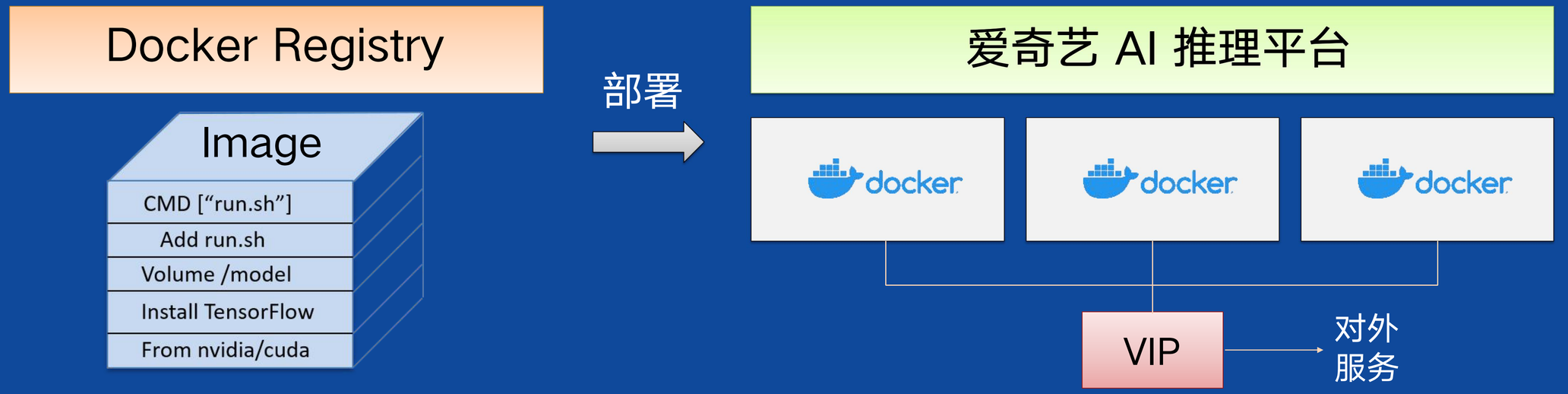
AI 推理平台架构演进

- 架构演进



架构演进 V1

- 自定义推理镜像
  - 将模型封装到镜像，推送到 Docker Registry
- Docker 容器化部署
  - 绑定 VIP，对外提供 HTTP 服务
- 模型版本升级
  - 通过更新镜像和重启容器



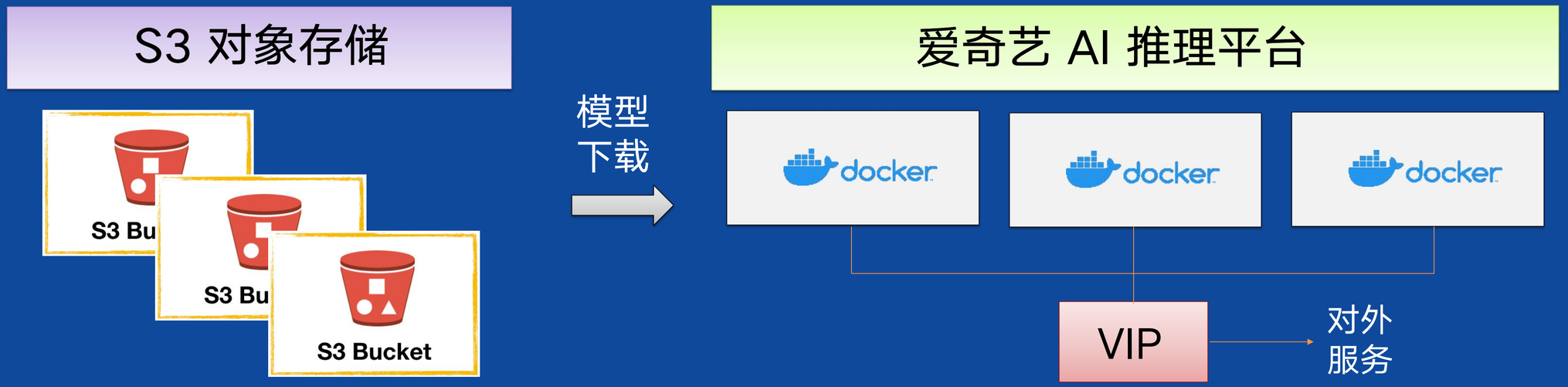
架构演进 V1

- 优点
  - 可以自定义复杂的前后处理逻辑
  - 可以封装多个相互关联 AI 模型的推理

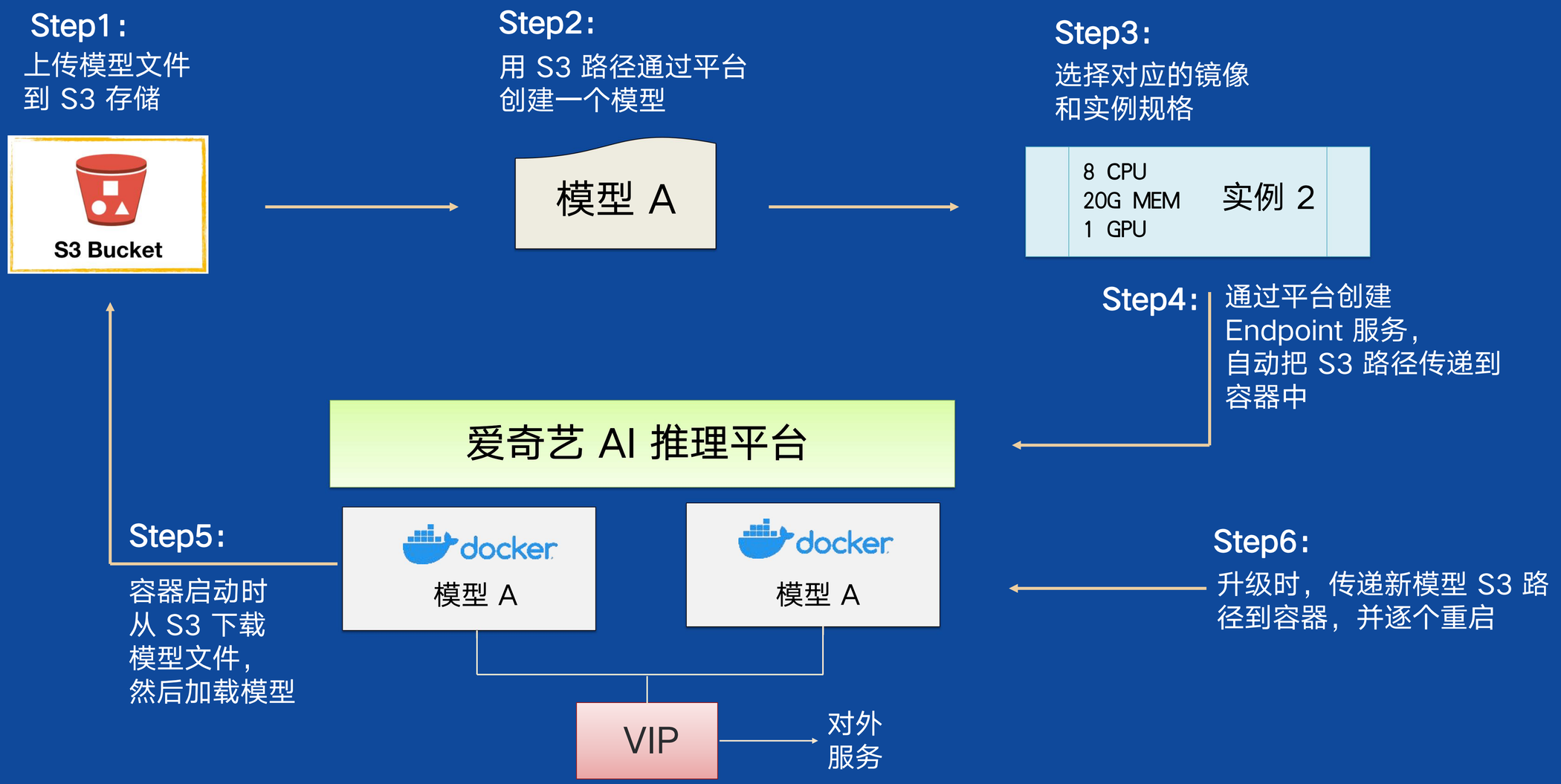
- 缺点
  - 推理镜像环境依赖和训练环境难对齐
  - 大部分是 Python 代码封装，推理效率比较低

架构演进 V2

- 抽象深度学习框架环境
  - 平台提供各个框架版本的容器镜像
- 模型文件独立管理
  - 平台提供 S3 对象存储来上传/下载模型
- 模型加载和版本升级
  - 传递模型 S3 路径到容器，启动时下载，版本更新通过重启容器



架构演进 V2


Step1：
上传模型文件
到 S3 存储
S3 Bucket
Step2：
用 S3 路径通过平台
创建一个模型
模型 A
Step3：
选择对应的镜像
和实例规格
8 CPU
20G MEM
1 GPU
实例 2
Step4：
通过平台创建
Endpoint 服务，
自动把 S3 路径传递到
容器中
爱奇艺 AI 推理平台
Step5：
容器启动时
从 S3 下载
模型文件，
然后加载模型
docker
模型 A
docker
模型 A
Step6：
升级时，传递新模型 S3 路
径到容器，并逐个重启
VIP
对外
服务

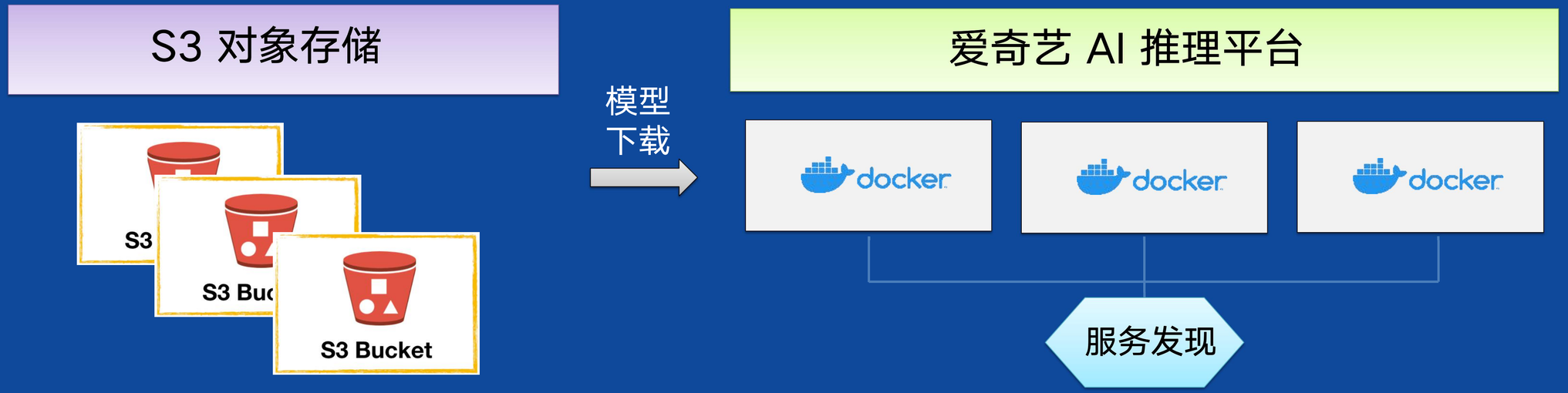
架构演进 V2

- 优点
  - 推理服务只依赖于模型文件
  - 支持模型优化，提高推理效率

- 缺点
  - 模型更新时版本不统一
  - 不能适应搜索/广告/推荐类业务对 TCP 长连接的需求

架构演进 V3

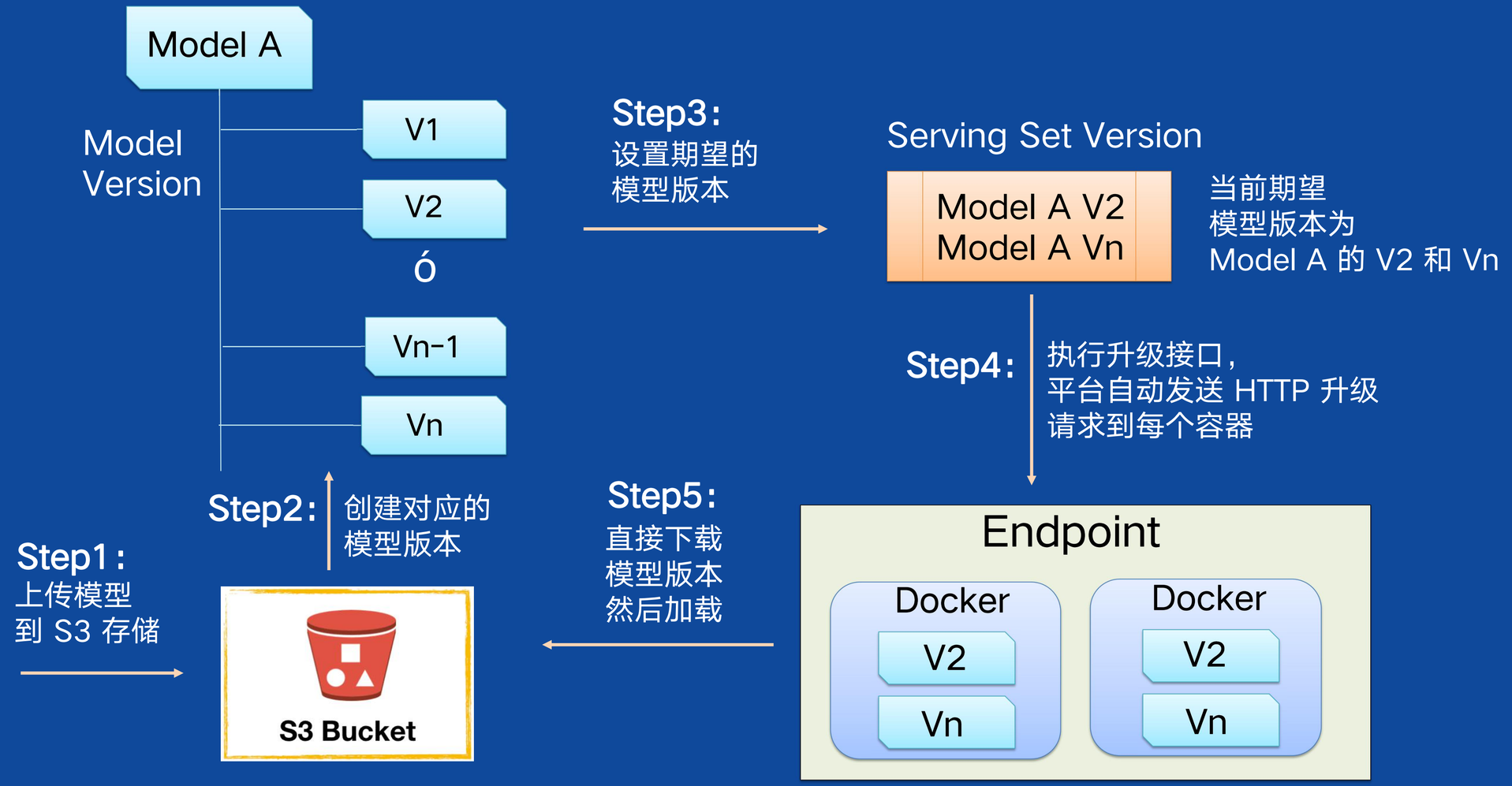
- 模型版本动态热更新
  - 通过 S3 对象存储下载模型，实时热更新，不需要重启容器
- 模型多版本管理
  - 可以同时服务模型的多个版本，支持回滚
- 服务发现
  - 支持 Consul 服务发现



架构演进 V3 概念定义

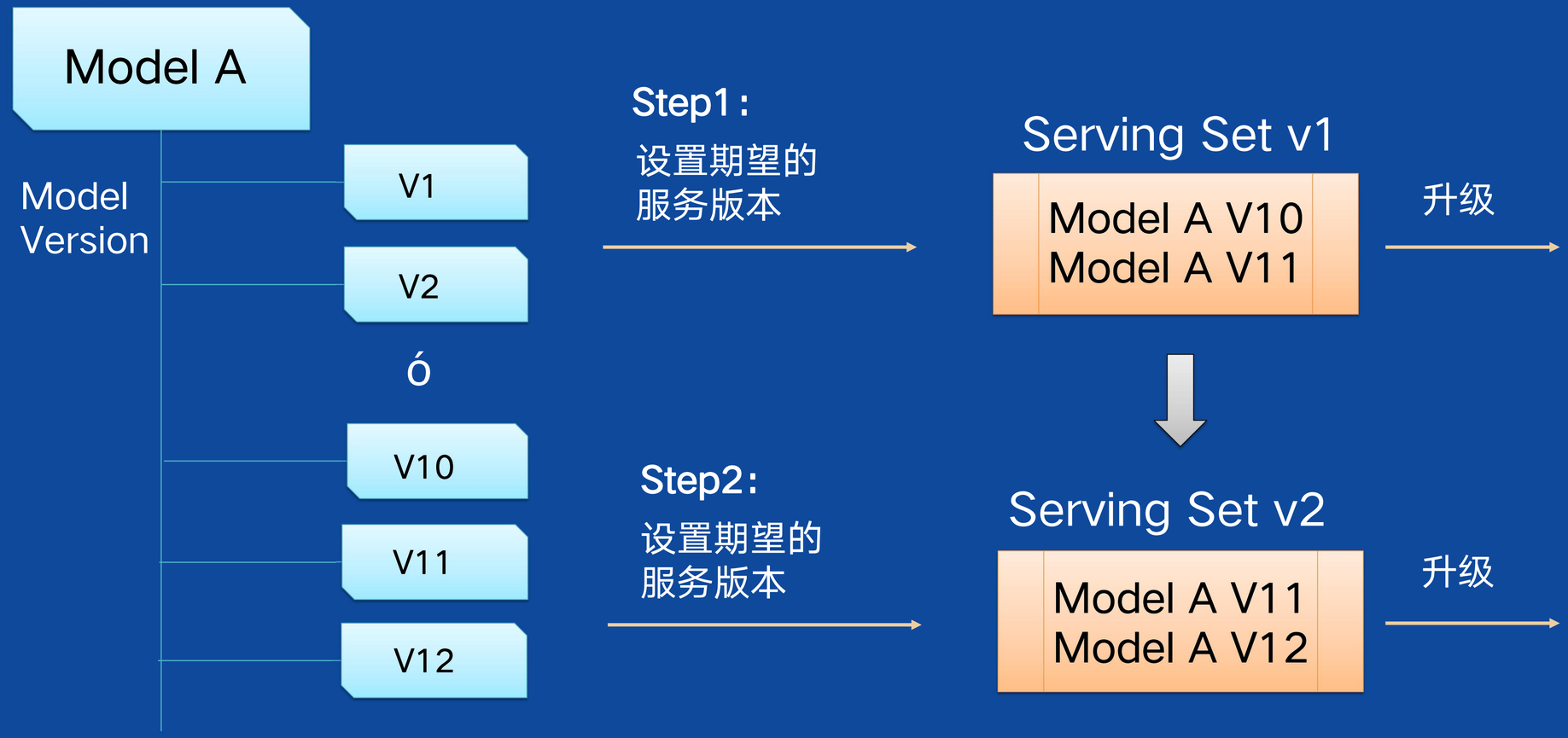
- Model
  - 表示某一个 AI 模型，一个逻辑的模型名称，并不包括任何模型实体文件
- Model Version
  - 表示某个 Model 的一个具体版本，指向一个具体的模型文件
  - 一个 Model 下可以包含多个模型版本，版本号必须是自增的
- Serving Set Version
  - 当前服务的模型版本集合，包含一个或多个 Model Version
- Endpoint
  - 一个推理服务，包括一个或多个容器实例，每个实例服务相同的模型版本

架构演进 V3


Model A
Model Version
V1
V2
ó
Vn-1
Vn
Step1：
上传模型
到 S3 存储
Step2：
创建对应的
模型版本
S3 Bucket
Step3：
设置期望的
模型版本
Serving Set Version
Model A V2
Model A Vn
当前期望
模型版本为
Model A 的 V2 和 Vn
Step4：
执行升级接口，
平台自动发送 HTTP 升级
请求到每个容器
Step5：
直接下载
模型版本
然后加载
Endpoint
Docker
V2
Vn
Docker
V2
Vn

架构演进 V3

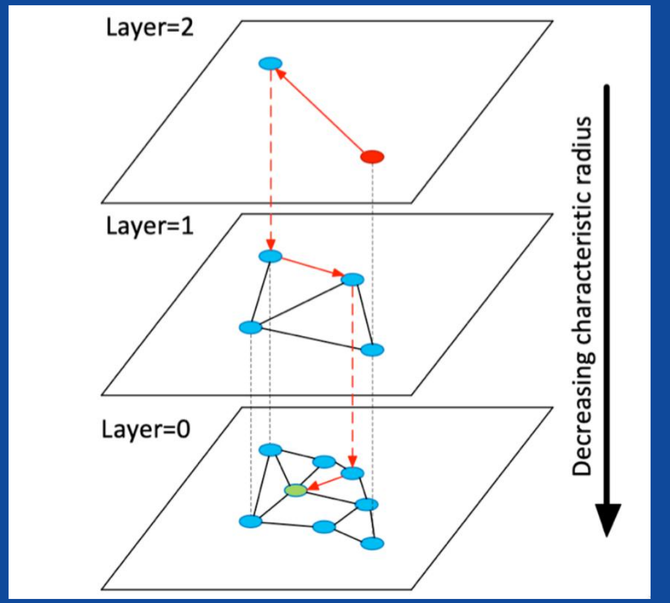
- 模型版本滚动升级



适配更多业务场景

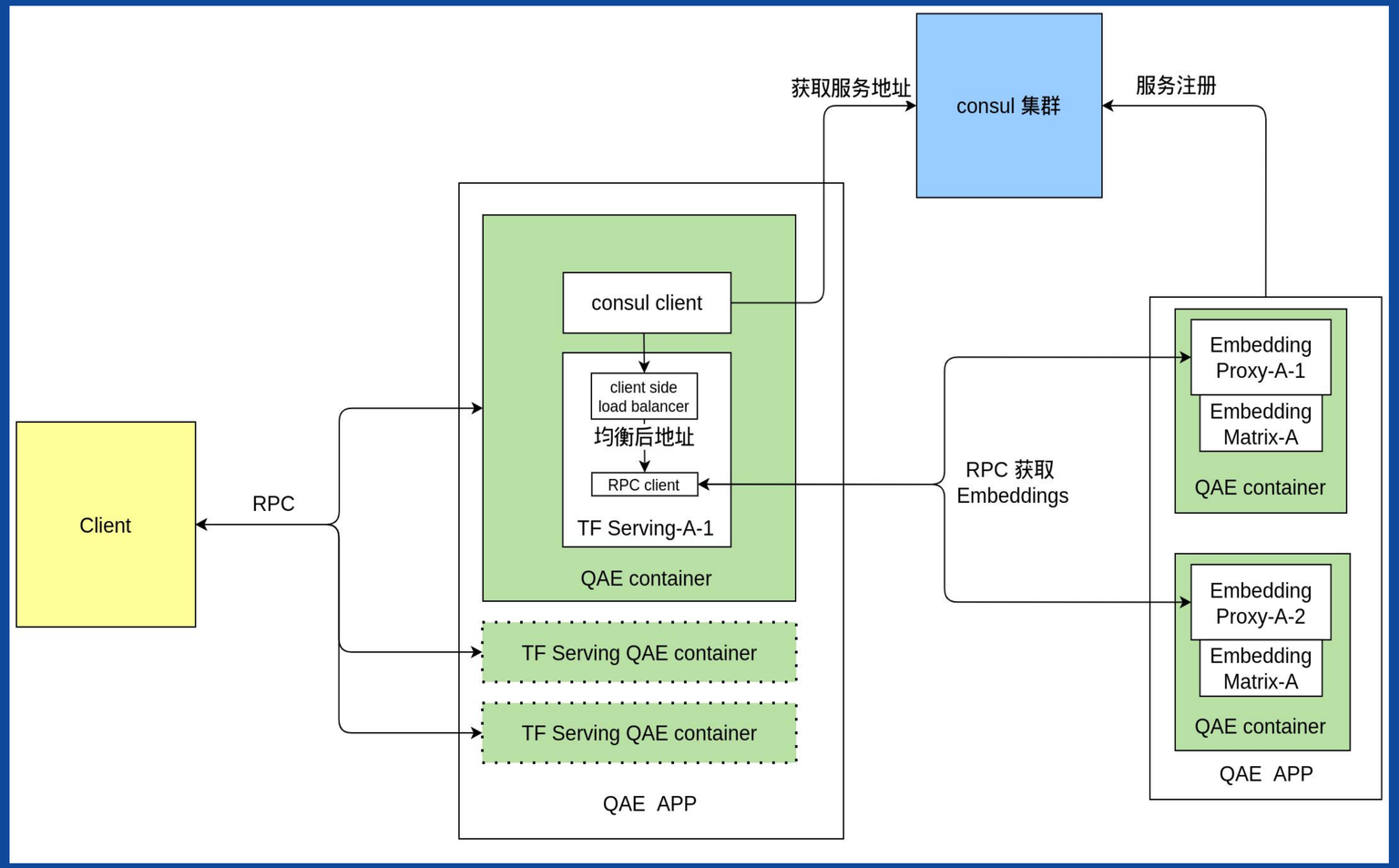
- XGBoost/FM 机器学习模型
  - 在 TensorFlow Serving 的基础上增加了 XGBoost 和 FM 模型的推理支持，并对外开源
  - https://github.com/iqiyi/xgboost-serving

- ANN 向量检索
  - 将 HNSW 算法封装成 TensorFlow OP 加入到训练和推理服务
  - 在 TensorFlow 中训练索引导出模型，部署推理服务



适配更多业务场景

- CTR 大模型分布式推理
- 将 TF 模型里的 Embedding 矩阵拆分出来独立部署
- 增加远程 Embedding lookup OP



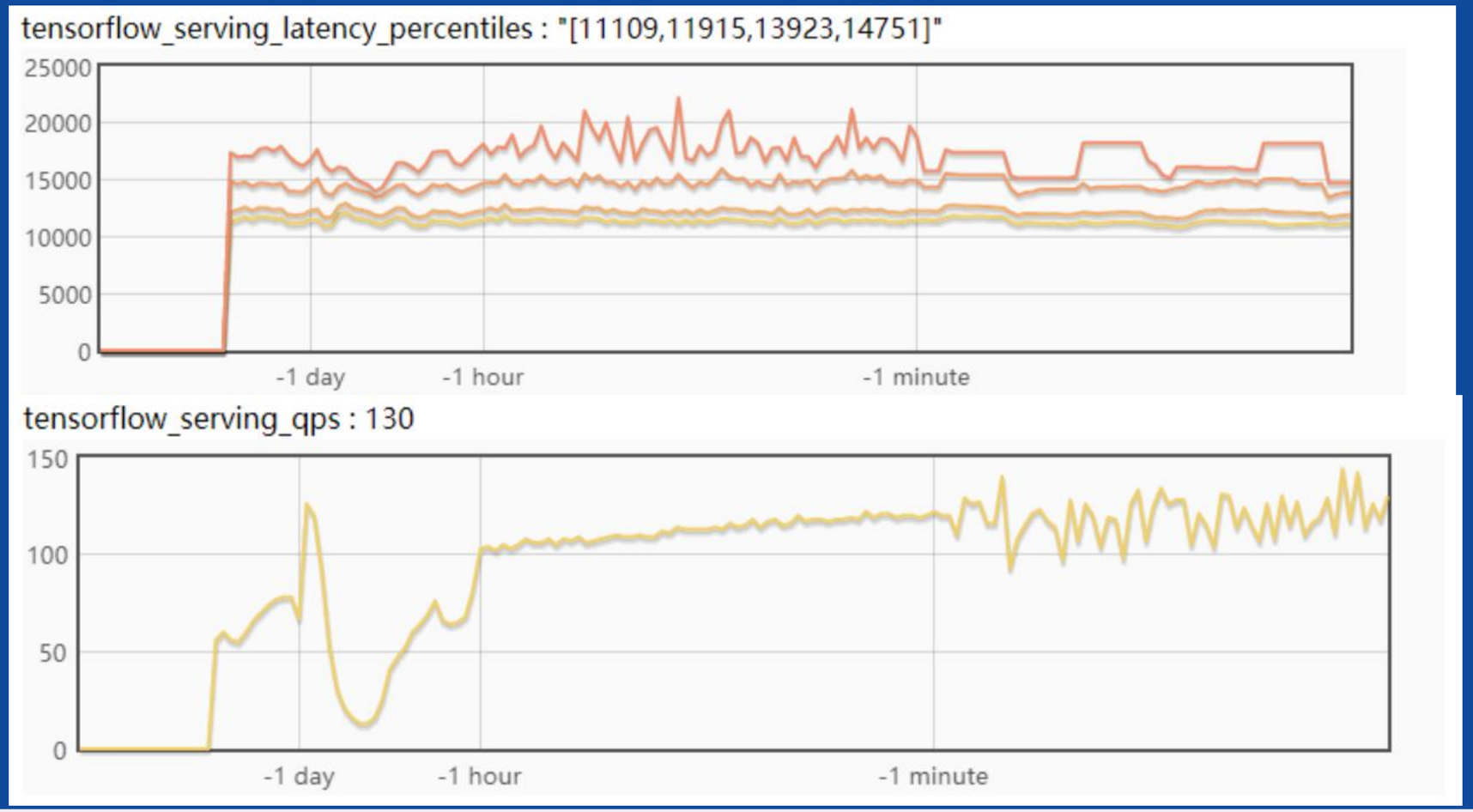
3. AI 推理平台落地优化实践

落地优化实践

- 指标监控
  - 服务内部的延迟，QPS 等监控指标
- 自动扩缩容
  - 根据监控指标变化进行动态扩缩容
- 跨地区模型下载
  - 优化其他地区的推理服务从北京地区下载 AI 模型
- 请求限流
  - 让推理服务在高负载下仍然可以降级服务
- 模型热更新请求毛刺优化
  - 模型热更新期间，客户端请求超时毛刺优化

指标监控

- 使用 Brpc bvar 增加服务内部的延迟，QPS 等监控指标
- bvar 是多线程环境下的计数器类库，几乎不会给程序增加性能开销



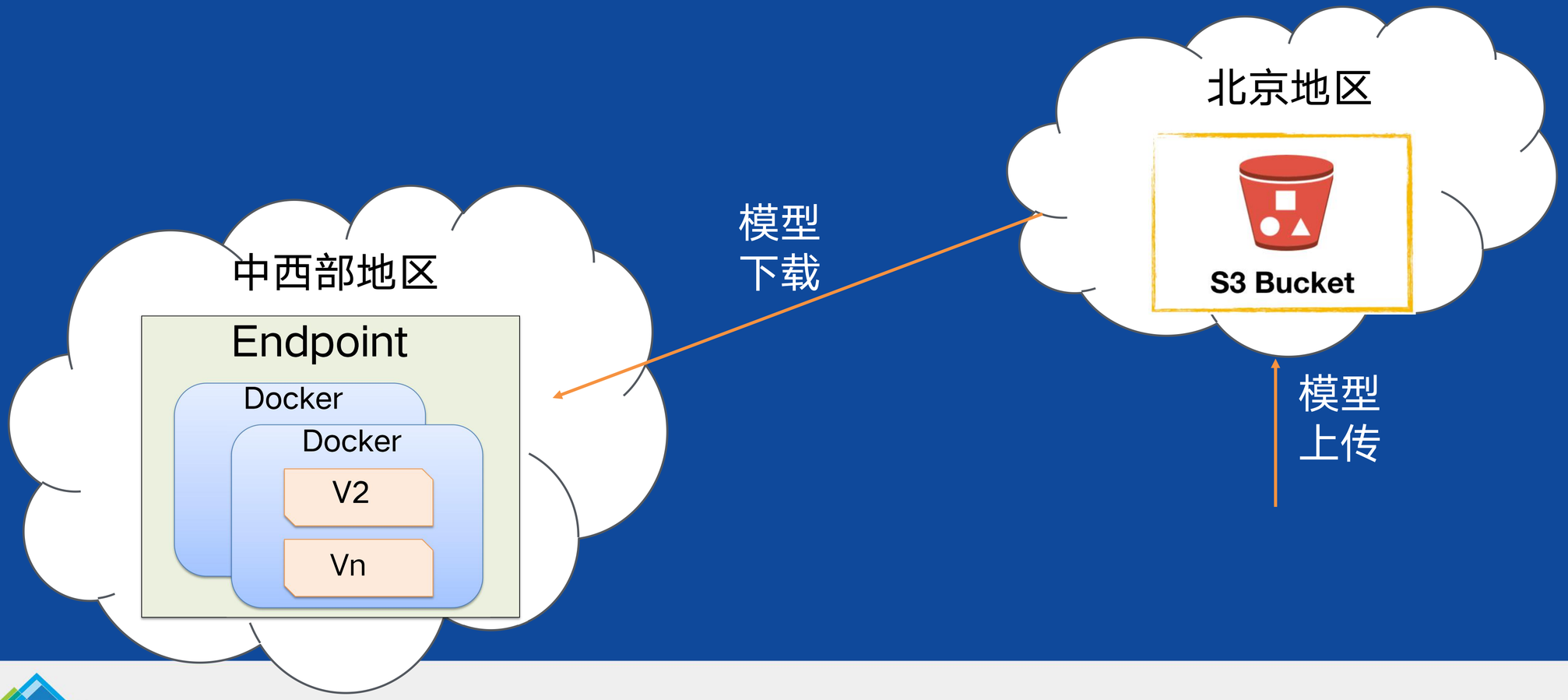
自动扩缩容

- 支持 4 种类型的自动扩缩容策略

Scheduled Rule	Range Track Rule	Target Track Rule	Predict Rule
• 支持定时的扩缩容	• 对特定指标的变化进行动态扩缩容 • 使指标保持在指定范围（range）内	• 对特定指标的变化进行动态扩缩容 • 使指标尽可能接近并小于指定目标（target）	• 对指标的变化进行提前预测，并根据预测结果提前进行扩缩容

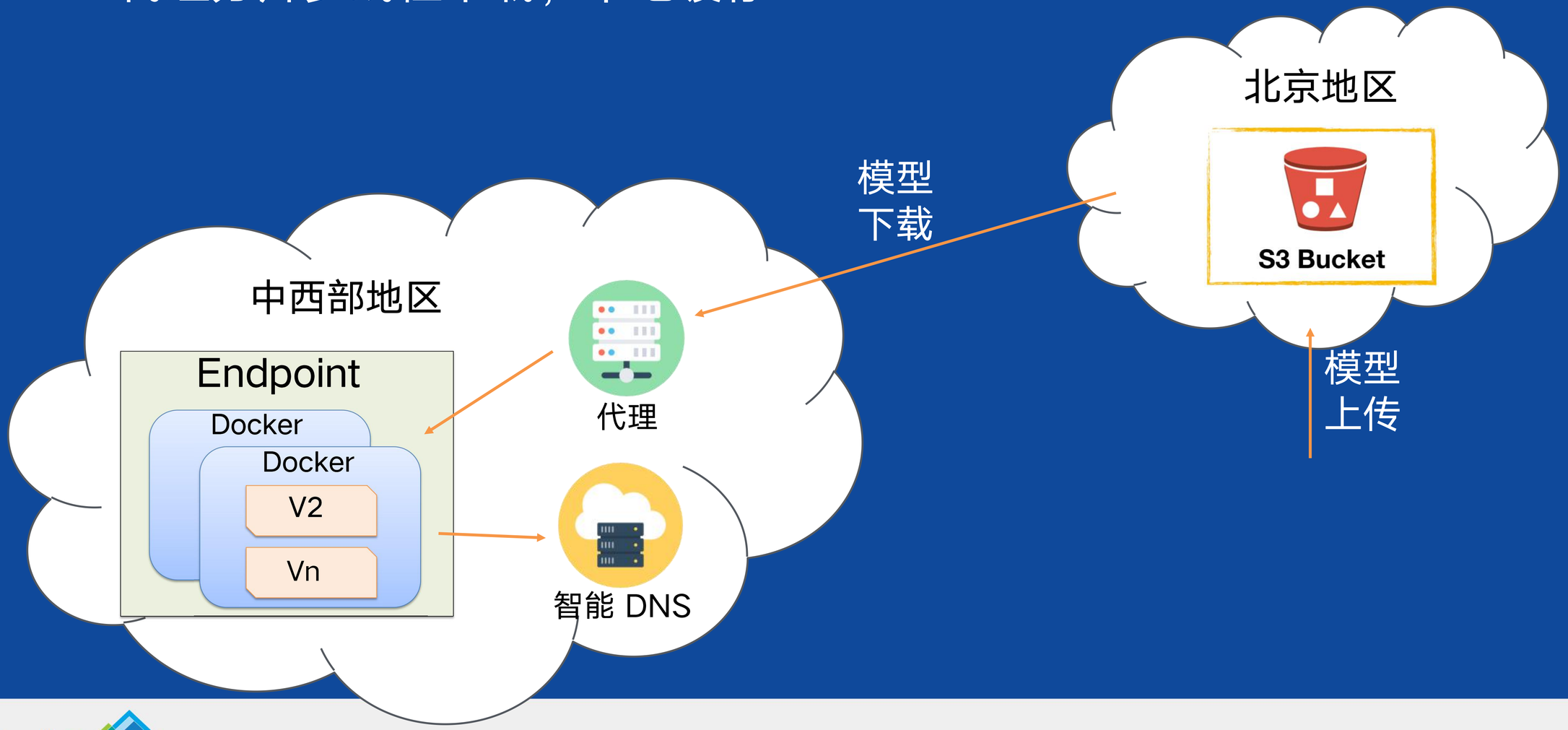
跨地区模型下载

- S3 对象存储集群部署在北京地区
- 在中西部地区的推理服务需要从北京地区下载模型
- 当实例数量过多，模型较大时对带宽压力比较大，时间过长



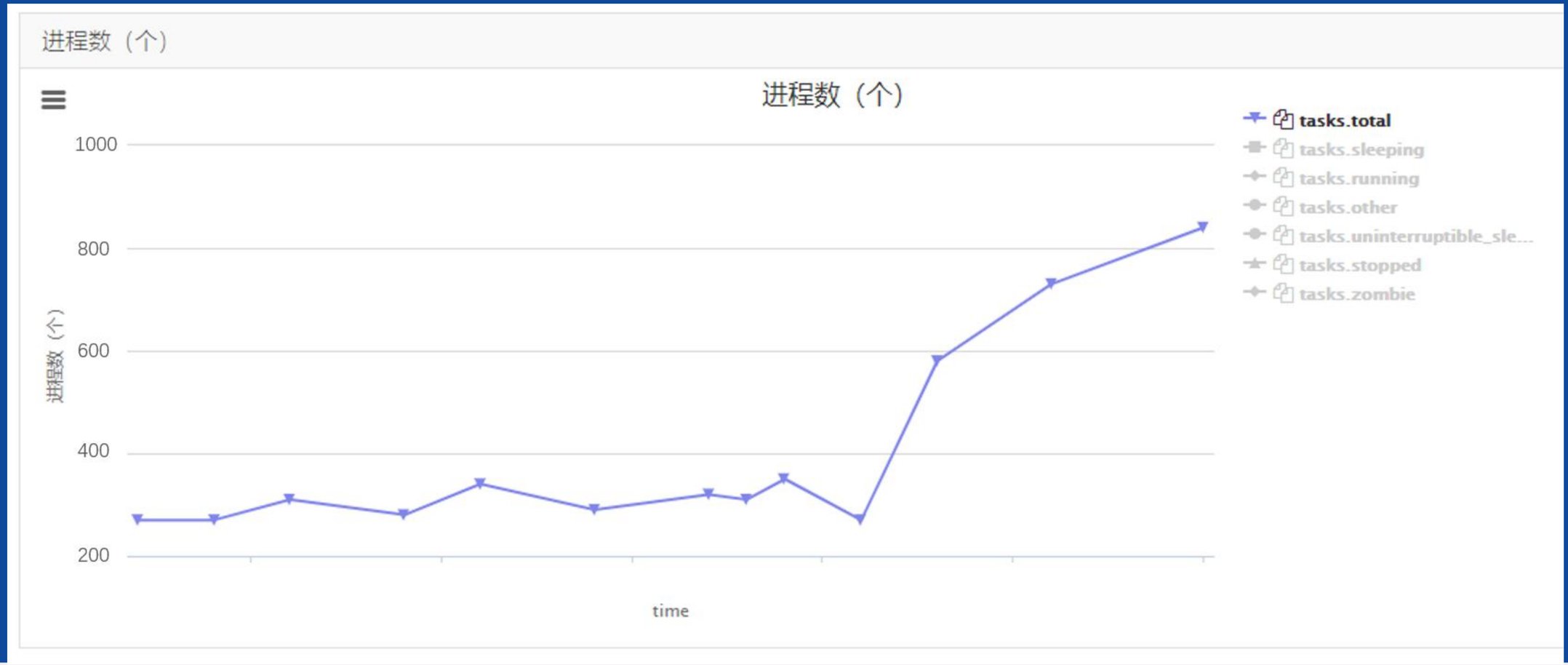
跨地区模型下载

- 增加下载代理，通过智能 DNS 重定向
- 代理分片多线程下载，本地缓存



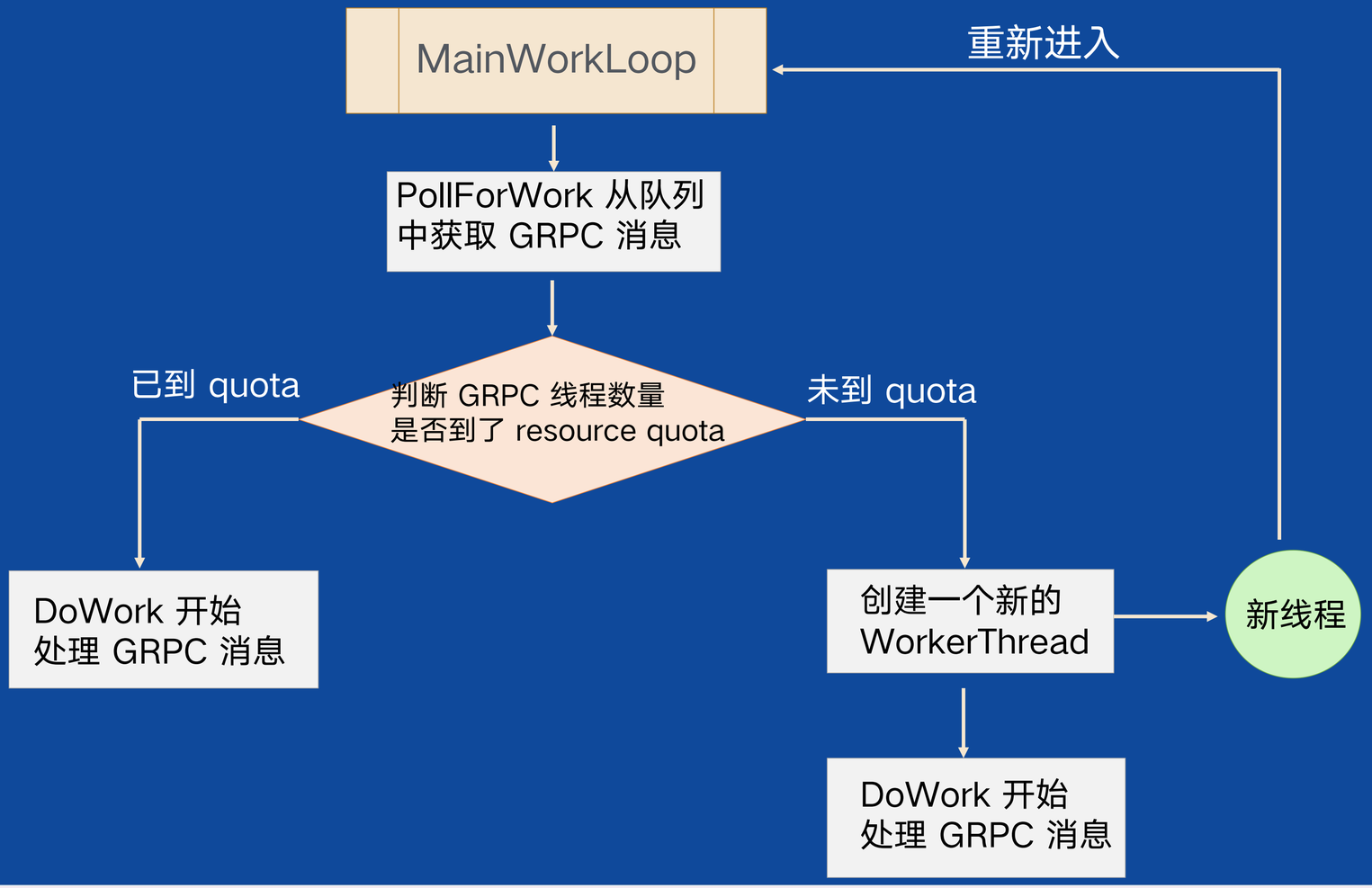
请求限流

- 线上业务在高峰期出现服务过载不断重启的情况
- 原因是 GRPC Server 处理线程不断增加导致 OOM



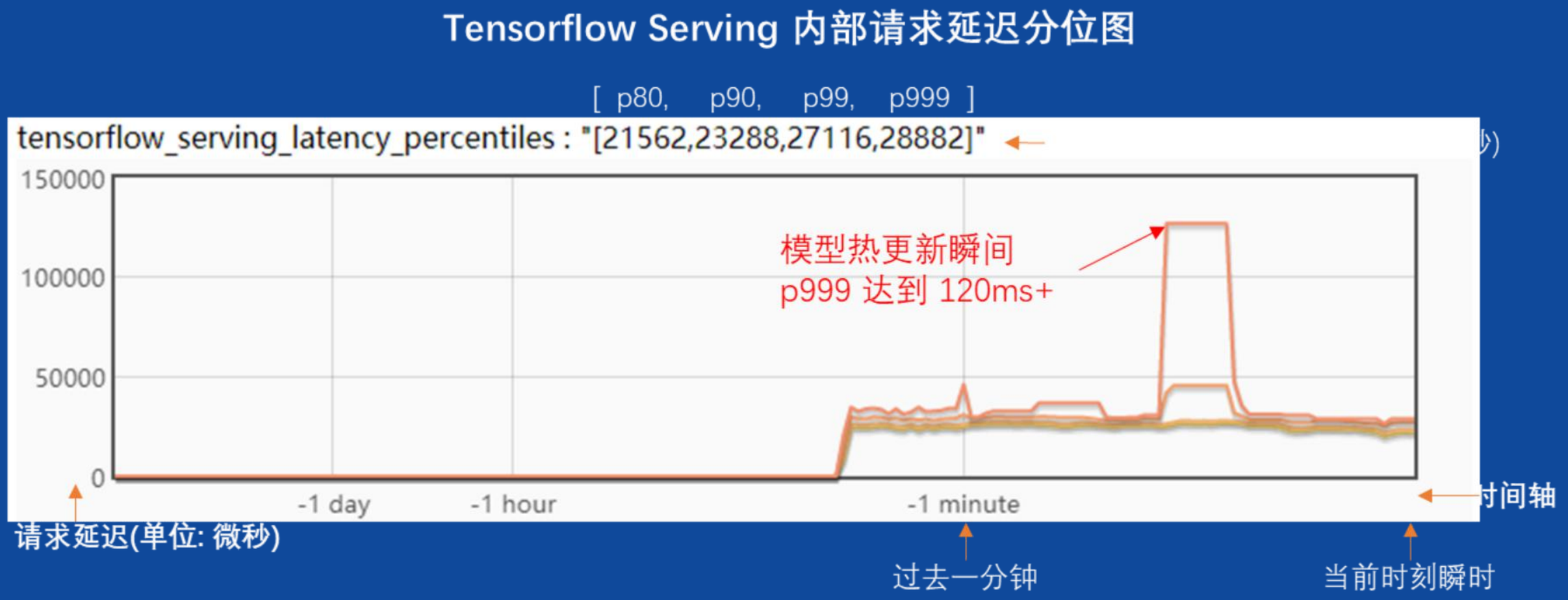
请求限流

- 限制 GRPC Server 的最大线程数量



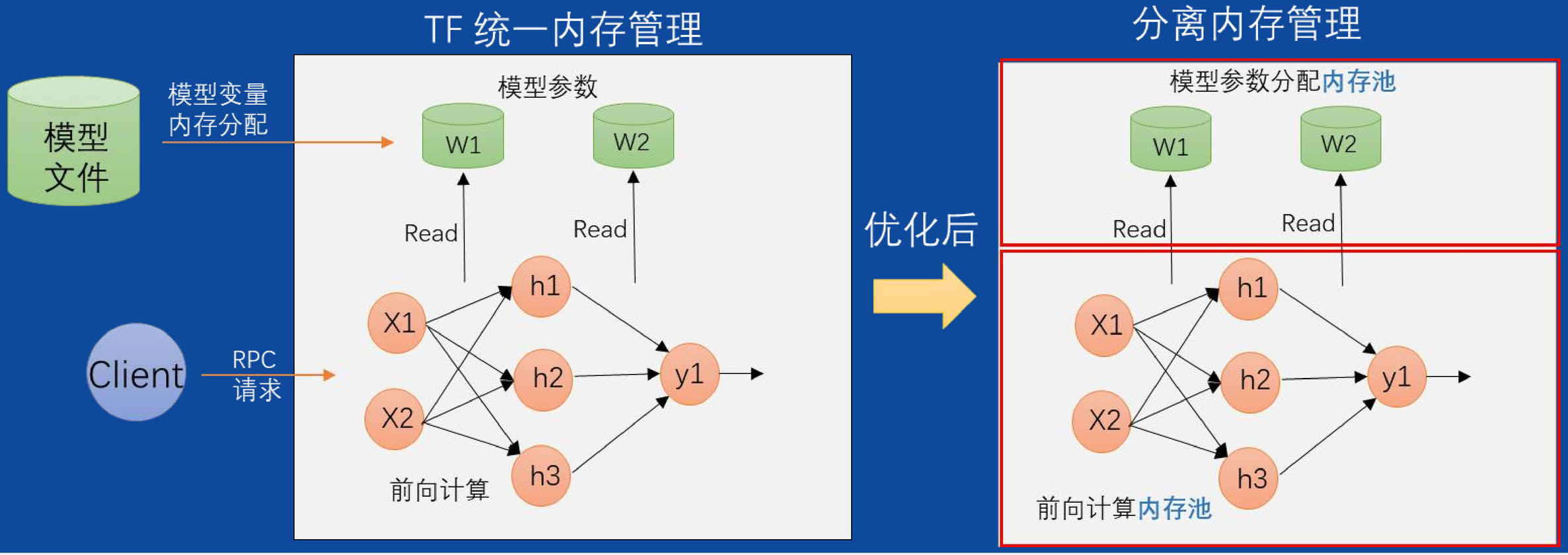
模型热更新请求毛刺优化

- 模型热更新会出现短暂的客户端请求超时现象（称之为毛刺现象）



模型热更新请求毛刺优化

- 通过 Warmup 来预热模型
- 使用 Jemalloc 做内存分配优化
- 模型参数分配和 RPC 请求内存分配分离



4. 总结与展望

总结与展望

- 目前线上以 V1 和 V3 两种为主
  - V1 的框架比较多样，主要使用是 CV/NLP 类业务
  - V3 的框架以 TensorFlow 为主，主要使用业务是搜索/广告/推荐

- 后续的方向
  - 利用 Triton 统一 CV/NLP 类模型部署
  - 推理模型自动优化
    - TVM 自动编译优化
  - 超大推荐类模型在线训练和推理支持

InfoQ

精彩继续！
更多一线大厂前沿技术案例

北京站

AiCon

全球人工智能与机器学习技术大会

时间：2021年11月25-26日

地点：北京 · 国际会议中心

扫码查看大会详情>>

北京站

PCon

全球产品创新大会

时间：2021年11月26-27日

地点：北京 · 国际会议中心

扫码查看大会详情>>

北京站

ArchSummit

全球架构师峰会

时间：2021年12月03-04日

地点：北京 · 国际会议中心

扫码查看大会详情>>

THANKS

Global

Architect Summit