释放Stable Diffusion 无限可能

基于Kubernetes的大规模部署最佳实践

于昺蛟 亚马逊云科技
郑予彬 亚马逊云科技

Content

目录

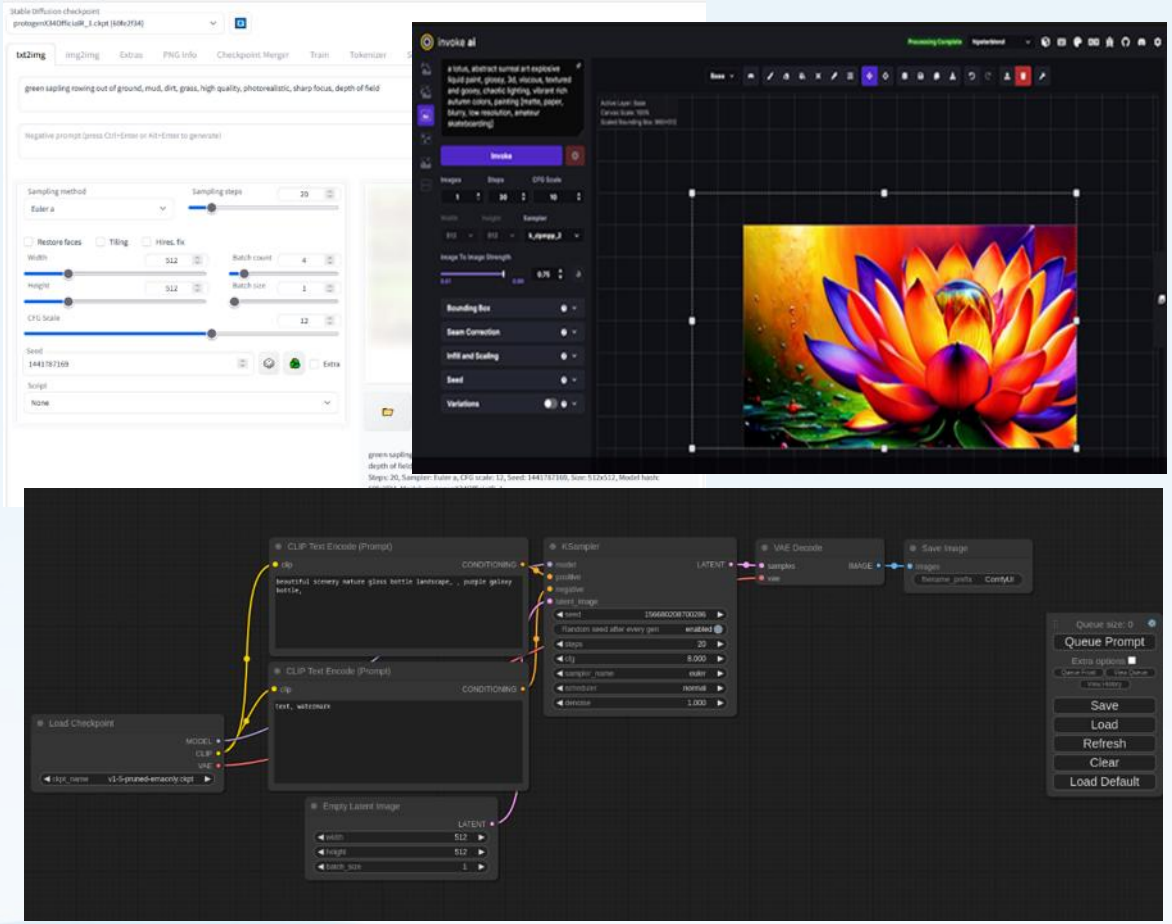
Stable Diffusion

由初创公司 Stability AI， Runway 与慕尼黑大学的 CompVis 研究团体合作开发的

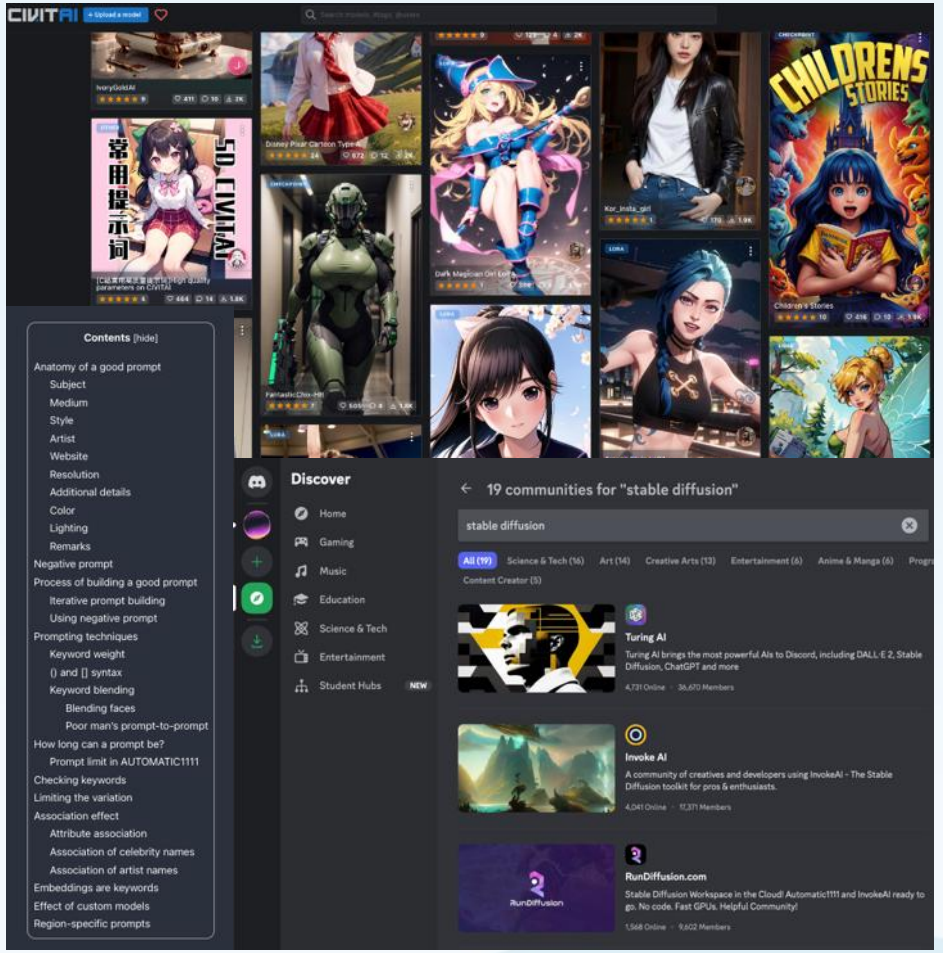
首次发布 2022/8/22

火热的 Stable Diffusion 开源社区

开发环境

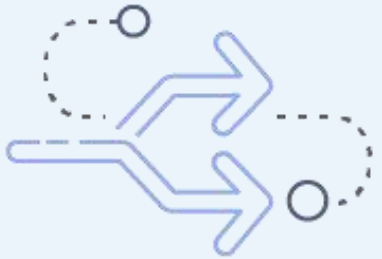


开源素材

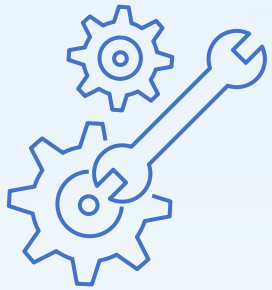


如何实现 Stable Diffusion 的大规模产品化部署?

部署 Stable Diffusion 需要考虑的因素



推理速度



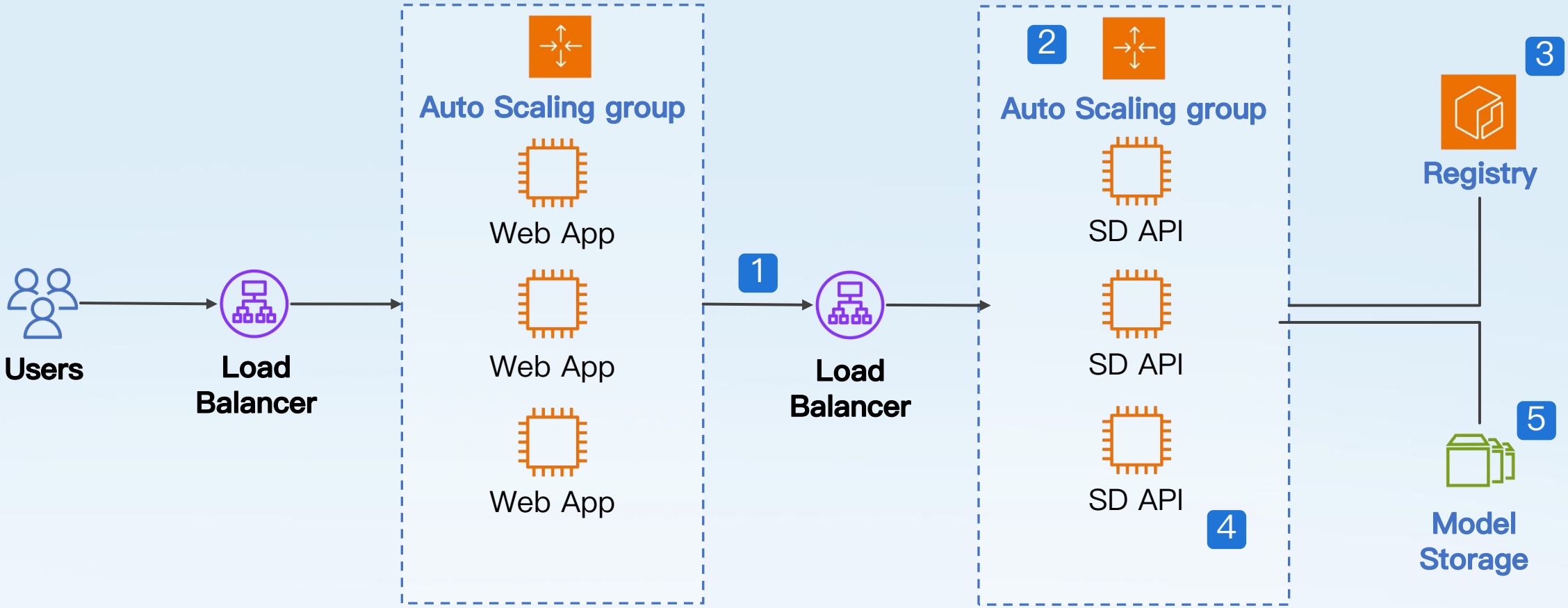
运维难度

个性化和可扩展性

弹性伸缩

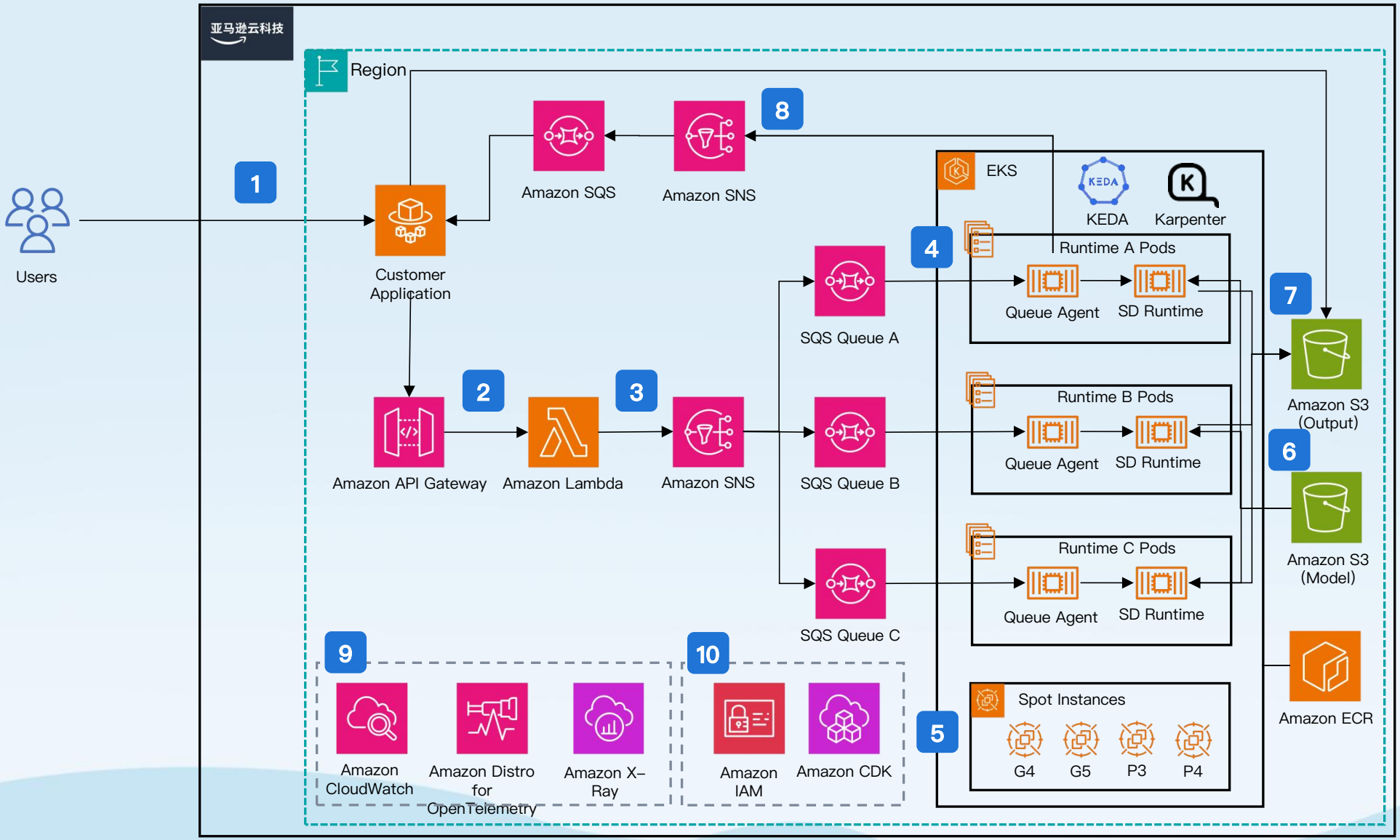
成本

常见部署方式的痛点


Auto Scaling group
Web App
Web App
Web App
Users
Load Balancer
1
Load Balancer
2
Auto Scaling group
SD API
SD API
SD API
4
3
Registry
5
Model Storage

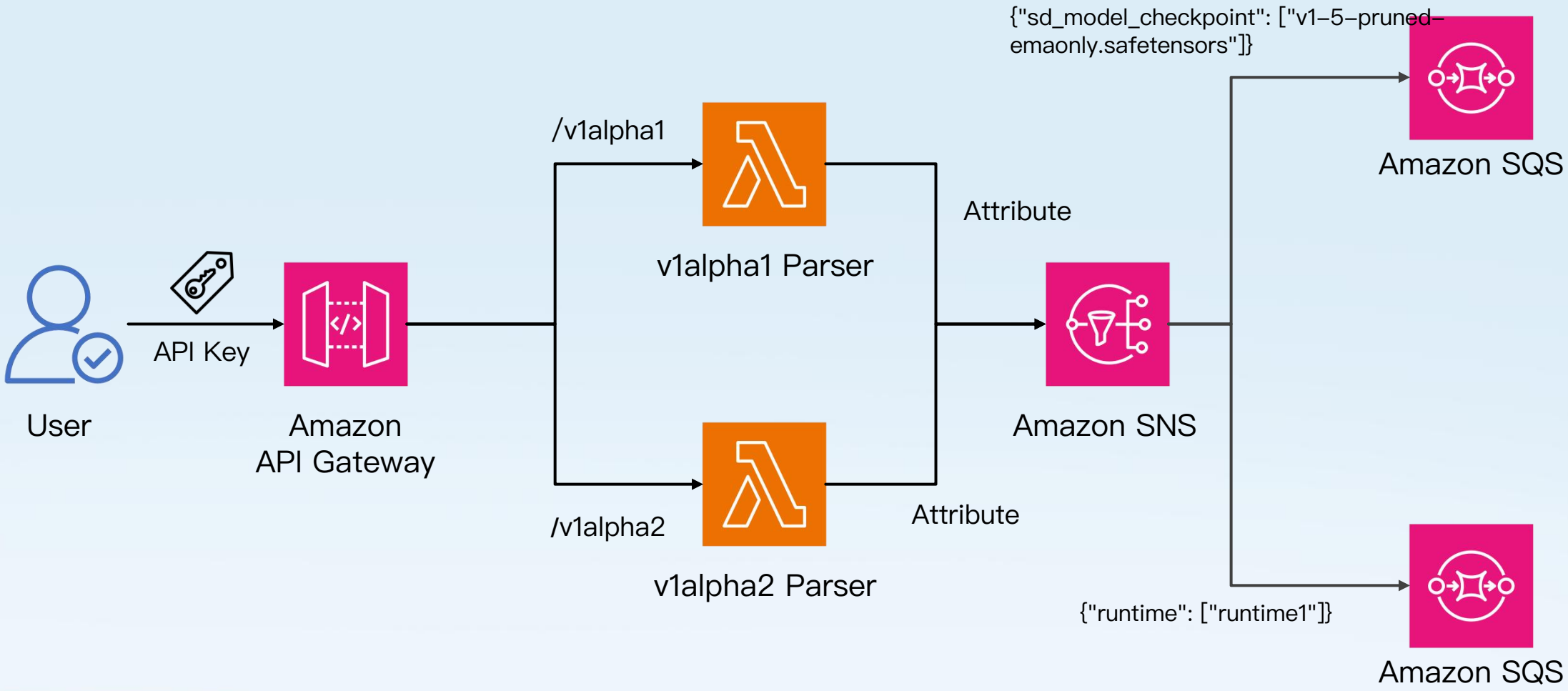
Stable Diffusion on Kubernetes

解决方案概览

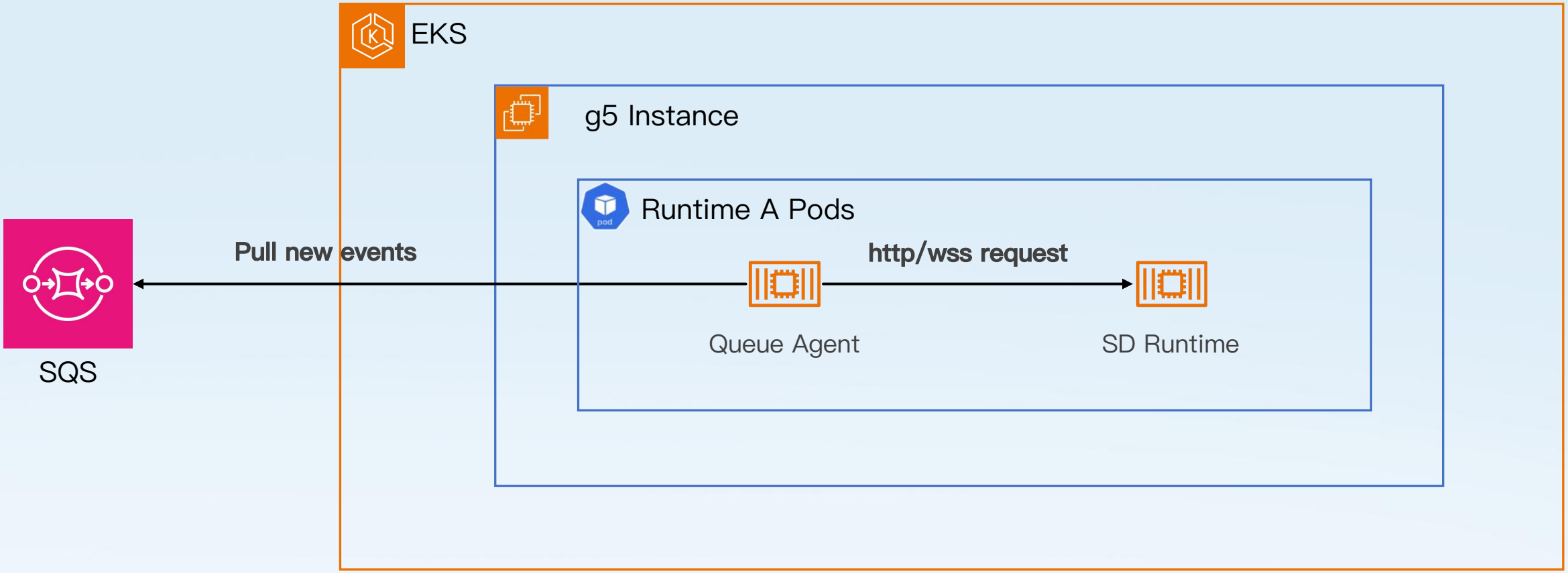


1. 用户将请求（模型，Prompt等）发送业务应用，业务应用将请求发送至 Amazon API Gateway 提供的API端点
2. 请求通过Amazon Lambda进行校验，并投送至 Amazon SNS 主题
3. Amazon SNS根据请求中的运行时名称，基于请求过滤机制，将请求投送至对应运行时的SQS队列
4. 在EKS集群中，KEDA会根据队列内消息数量扩充运行时的副本数
5. Karpenter会启动新的GPU实例以承载新的副本，这些实例运行BottleRocket操作系统，采用Spot/On-demand混合购买方式，且通过EBS快照预载Stable Diffusion运行时的容器镜像
6. Stable Diffusion 运行时启动时会通过 Mountpoint for Amazon S3 CSI Driver，直接从S3存储桶中加载模型
7. Queue Agent会从 Amazon SQS 队列里接收任务，并发送给Stable Diffusion运行时生成图像
8. 生成的图片由Queue Agent存储至 Amazon S3存储桶中，并将完成通知投送至 Amazon SNS 主题，SNS可将响应投送至SQS或其他目标中
9. 该解决方案提供完整的可观测性和管理组件，包含基于CloudWatch和ADOT的数值监控和日志，基于AWS X-Ray的全链路跟踪
10. 该解决方案通过基于AWS CDK的基础设施即代码部署方式进行部署和配置，通过IAM和API Key提供安全和访问控制

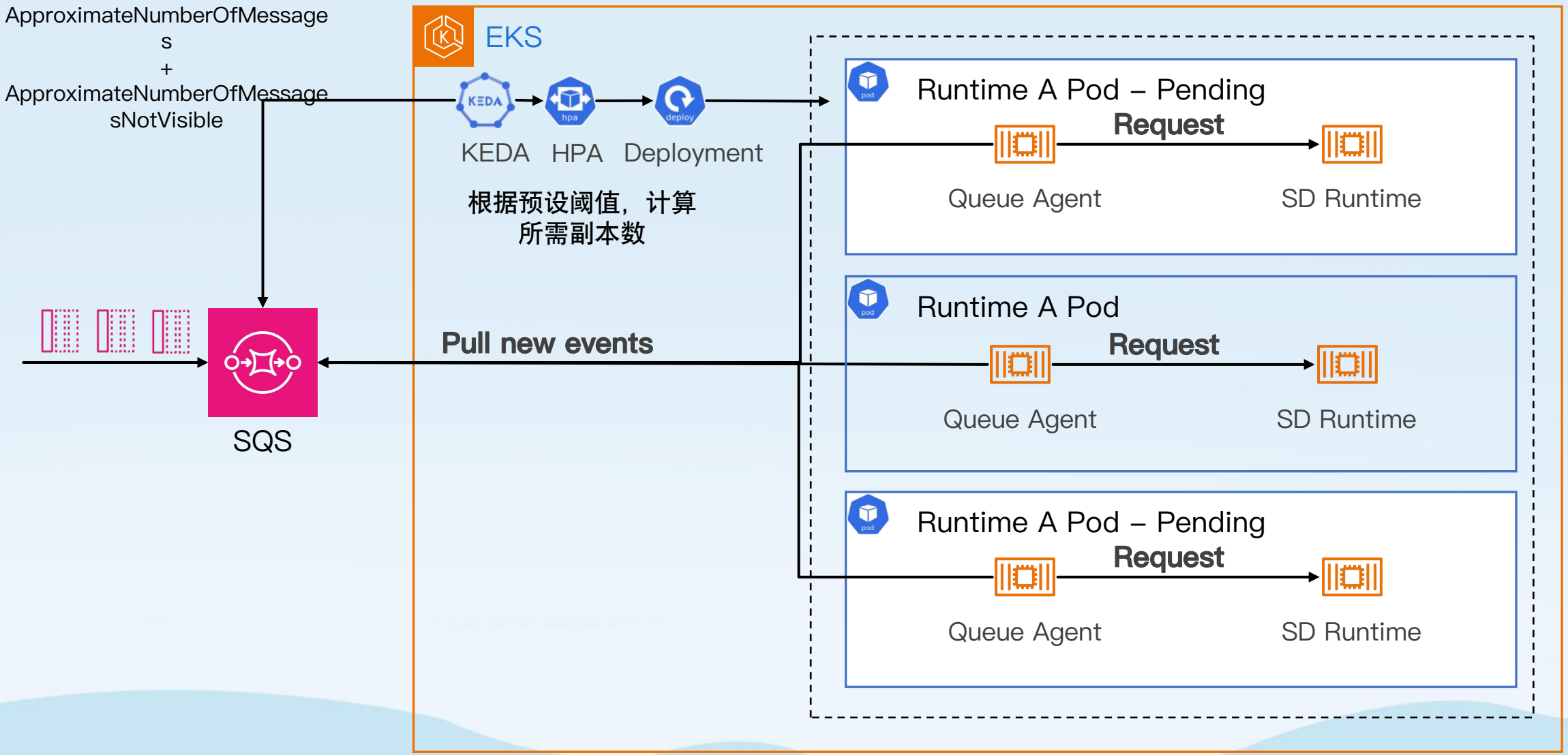
API 接入和事件传递


{"sd_model_checkpoint": ["v1-5-pruned-emaonly.safetensors"]}
/v1alpha1
v1alpha1 Parser
Attribute
Amazon SQS
API Key
User
Amazon API Gateway
Amazon SNS
/v1alpha2
v1alpha2 Parser
Attribute
{"runtime": ["runtime1"]}
Amazon SQS

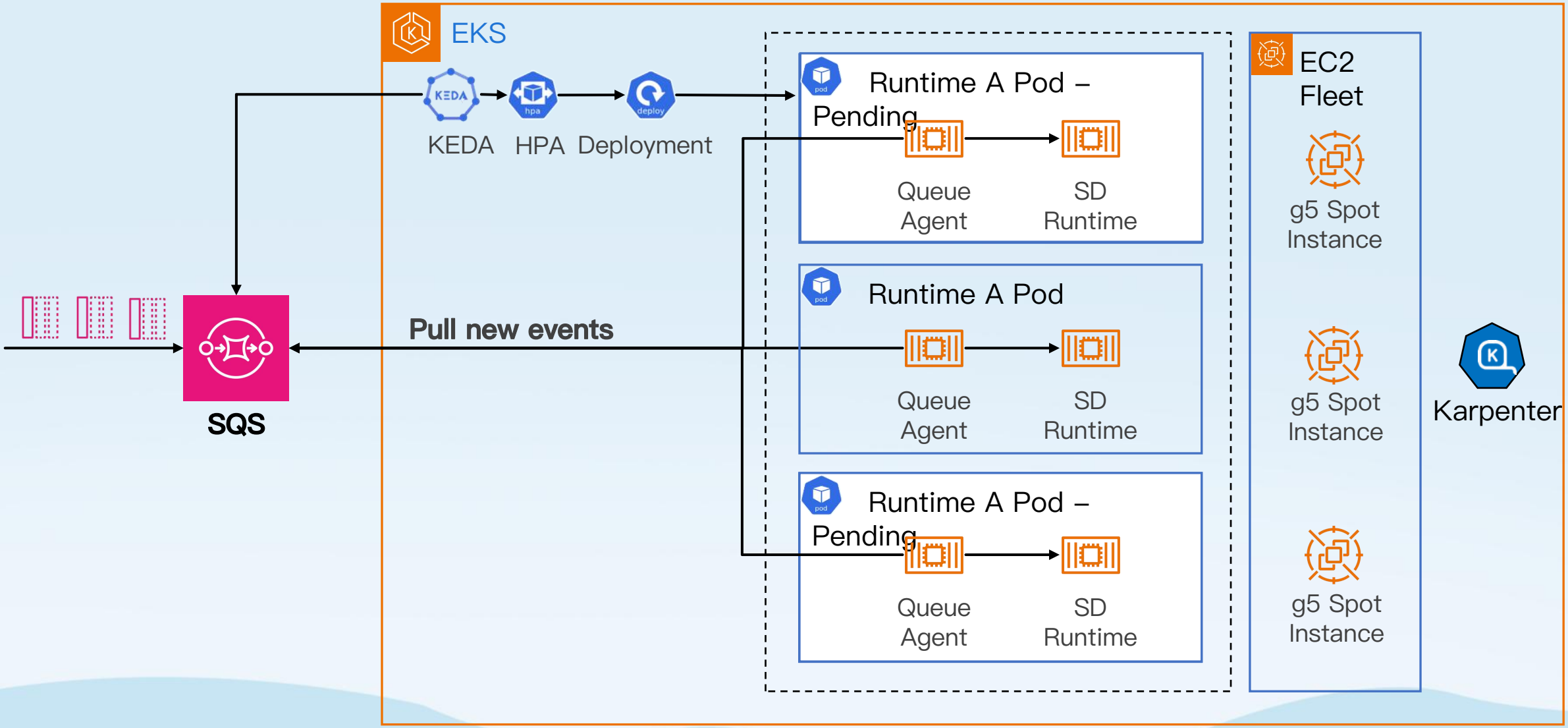
基于 GPU 实例进行 Stable Diffusion 推理



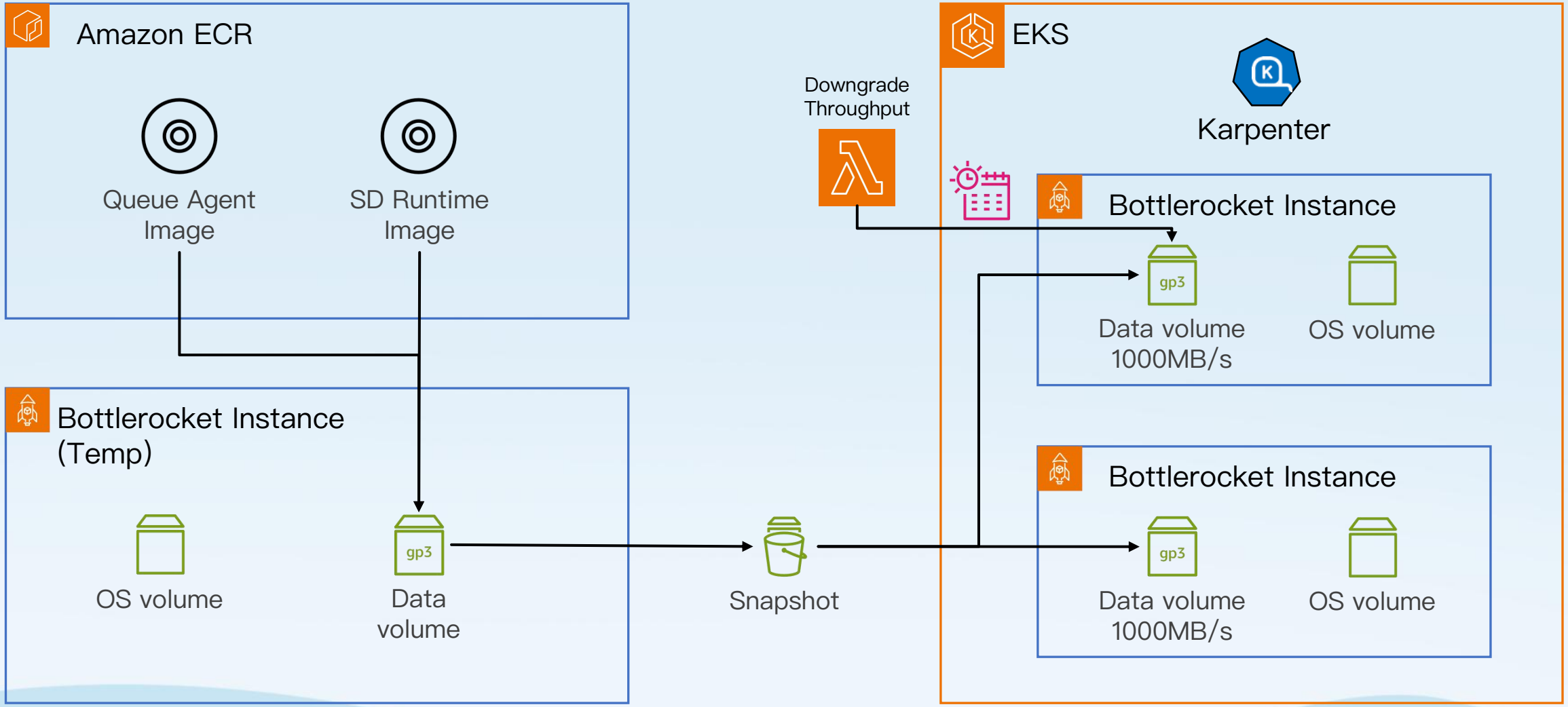
利用 KEDA 基于队列长度自动扩缩容器副本



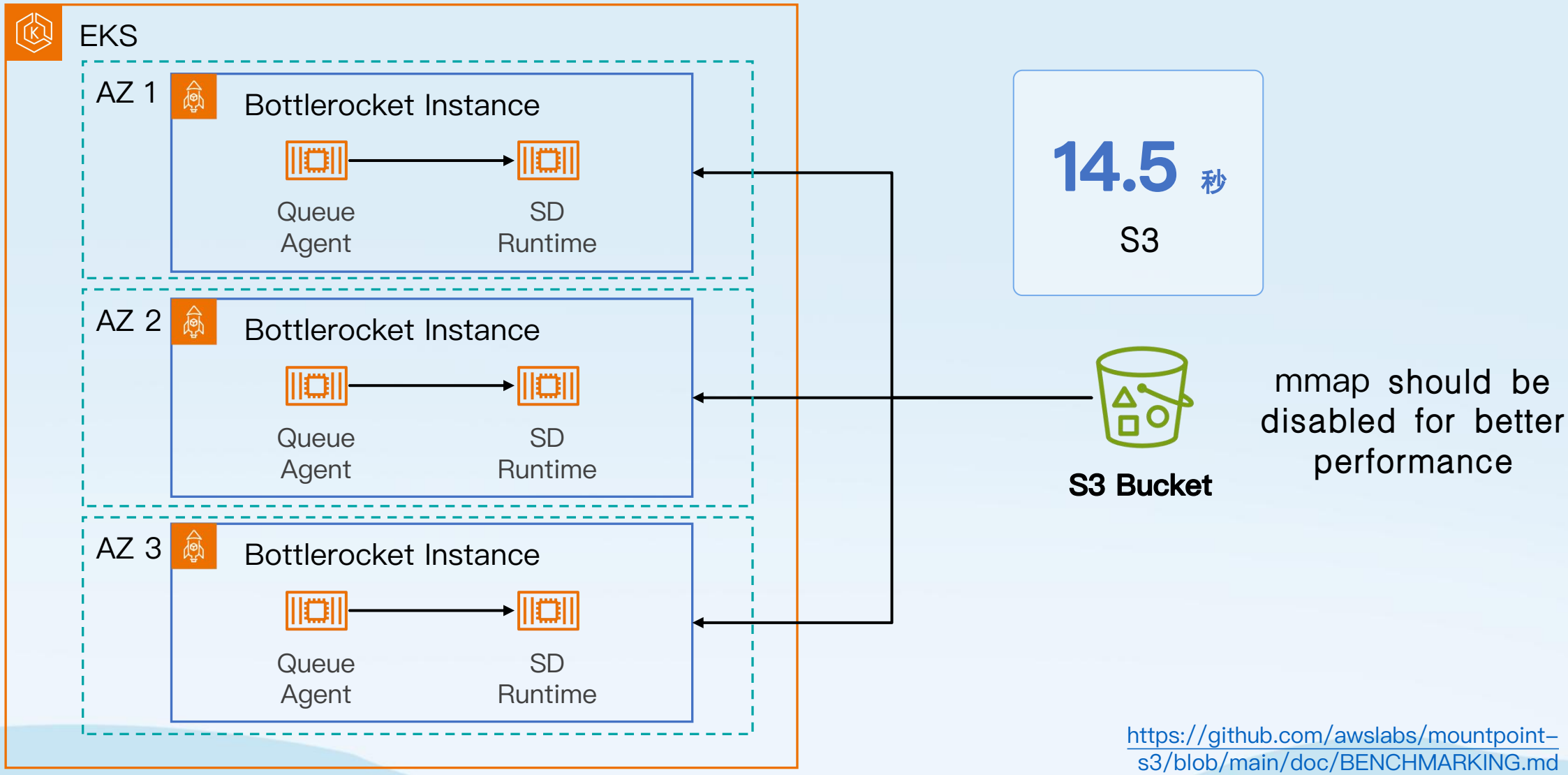
利用 Karpenter 自动扩缩实例



基于 Bottlerocket 实现容器镜像缓存

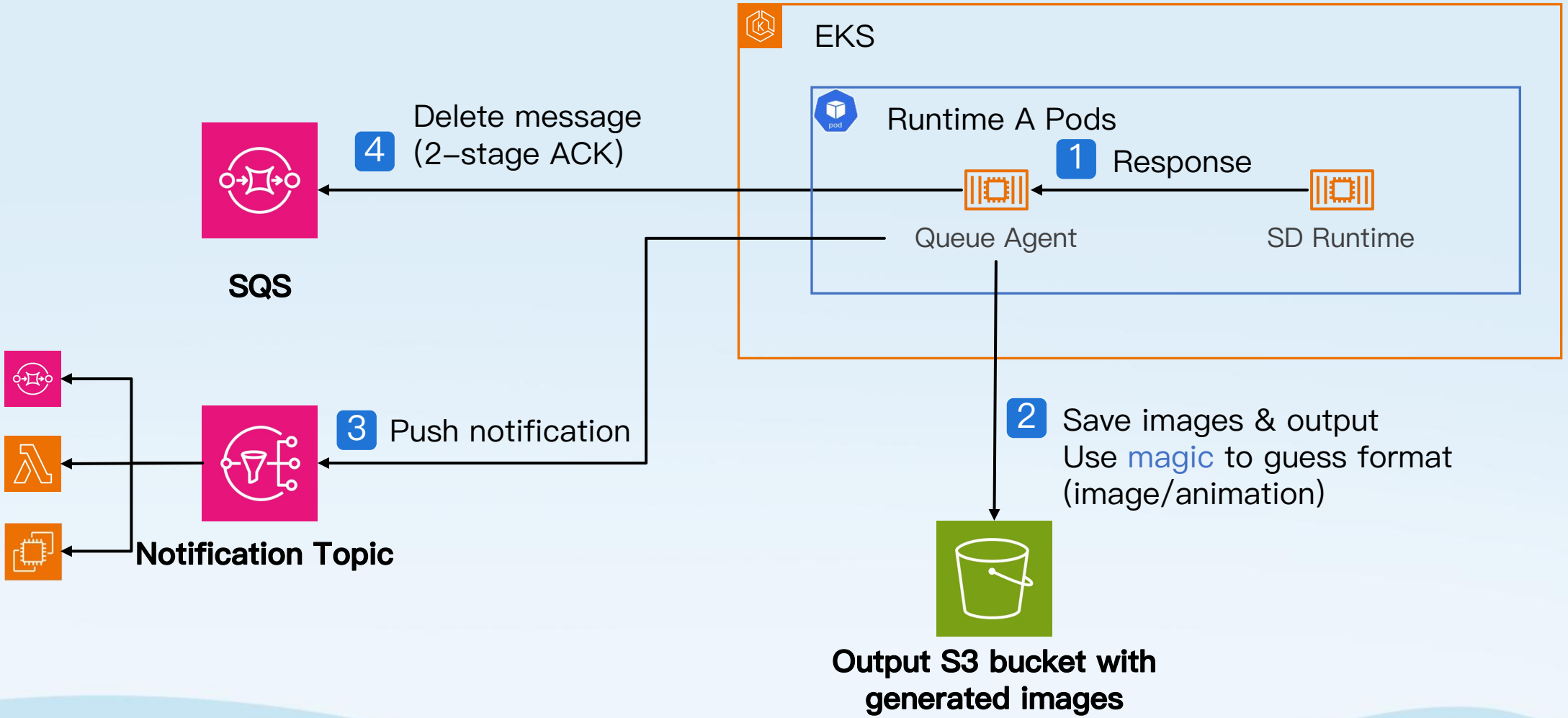


利用对象存储支撑大量模型动态读取



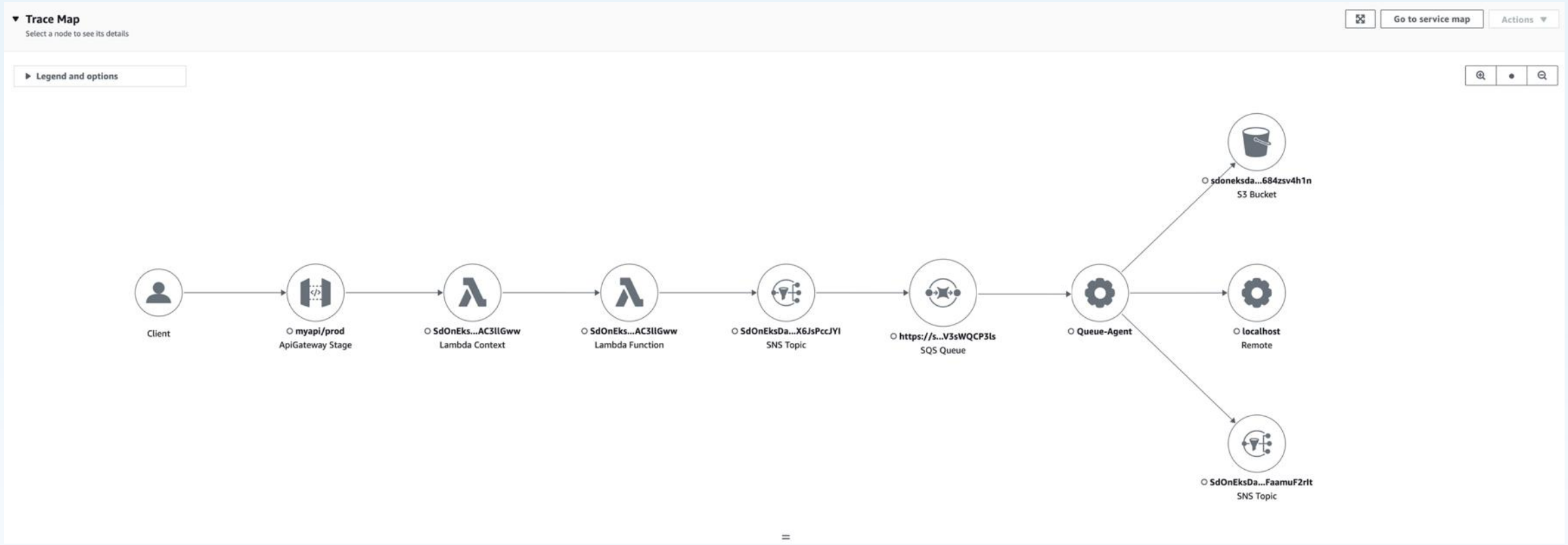
https://github.com/awslabs/mountpoint-s3/blob/main/doc/BENCHMARKING.md

推送处理结果


EKS
Runtime A Pods
1 Response
Queue Agent
SD Runtime
4 Delete message
(2-stage ACK)
SQS
3 Push notification
Notification Topic
2 Save images & output
Use magic to guess format
(image/animation)
Output S3 bucket with
generated images

X-Ray 端到端全链路追踪

可视化展示端到端请求路径



X-Ray 端到端全链路追踪

全链路追踪清晰展示调用时间线

性能与成本

扩容速度

添加新实例：20 – 40 秒

冷启动端到端：67 – 100 秒

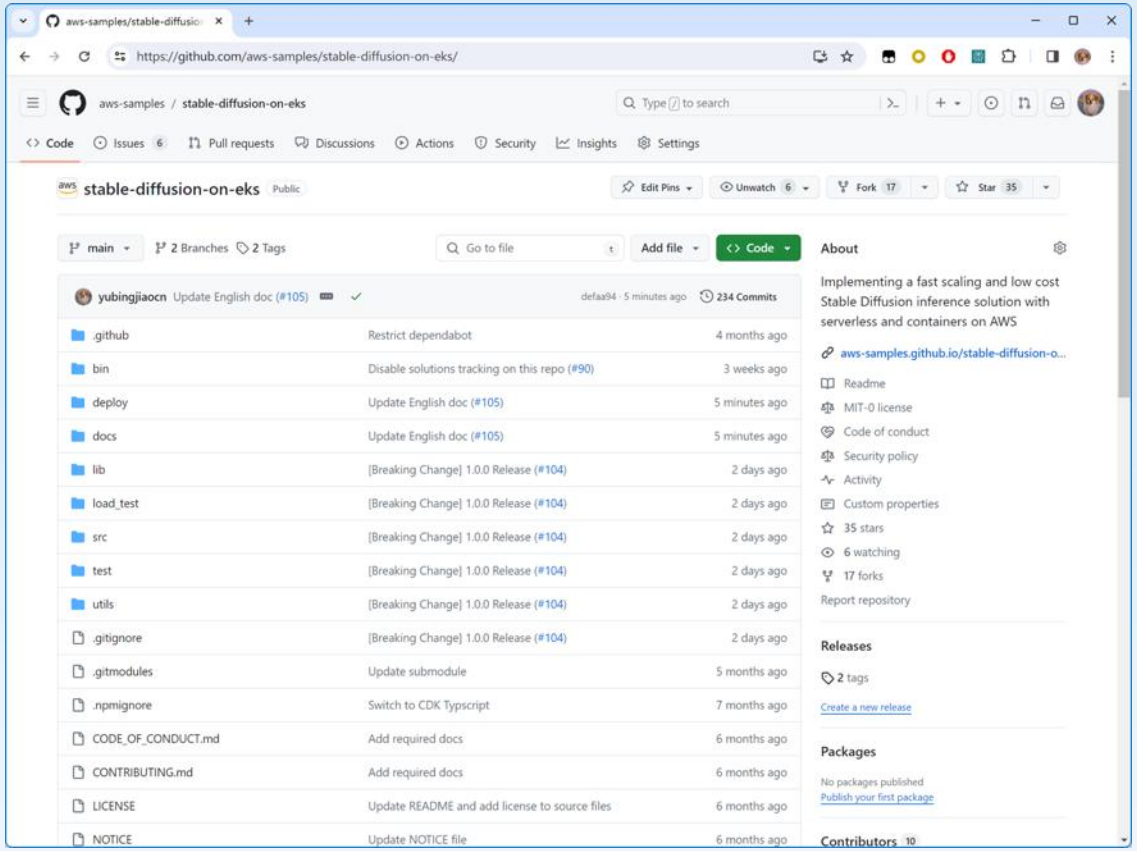
推理成本

G5.xlarge Spot 实例：节省 70%

G5.2xlarge Spot 实例：节省 64%

注：此数据只反映该解决方案在亚马逊云科技美国西部（俄勒冈）区域的性能表现和成本节省，实际性能表现和成本视您选择的区域，实例类型，当前Spot实例定价等因素而不同。此数据不作为亚马逊云科技对您作出的任何性能承诺和报价。

欢迎使用



https://github.com/aws-samples/stable-diffusion-on-eks/

Thanks.