智能语音V2.X部署架构与资源标准

服务清单

部署架构

- 基于ClientSide高可用方案
- 基于VIP高可用方案
- 基于阿里自研的vipserver高可用方案
- 双机高可用方案

资源标准

- 测试开发环境标准
- 生产环境配置标准
  - 基础服务
  - 录音文件识别
  - 一句话识别
  - 实时语音转写
  - 语音合成
  - 语音识别定制化(语言模型自学习)
  - 语音合成定制化
  - MRCP

名词解释：

单实例	为完成指定功能特性(例如录音文件识别)的最小原子化全链路（或理解成单机版，一体机），下文以Instance为代称，语音功能详情请参考官网 https://ai.aliyun.com/nls
高可用	为规避单点故障而采取多实例(N>1)提供服务的方案，本推荐方案里使用全主模式，不考虑主备模式
中间件	语音服务依赖的底层存储服务，比如业界常用的mysql，redis，mongodb，下文以Mideleware 代称

服务清单

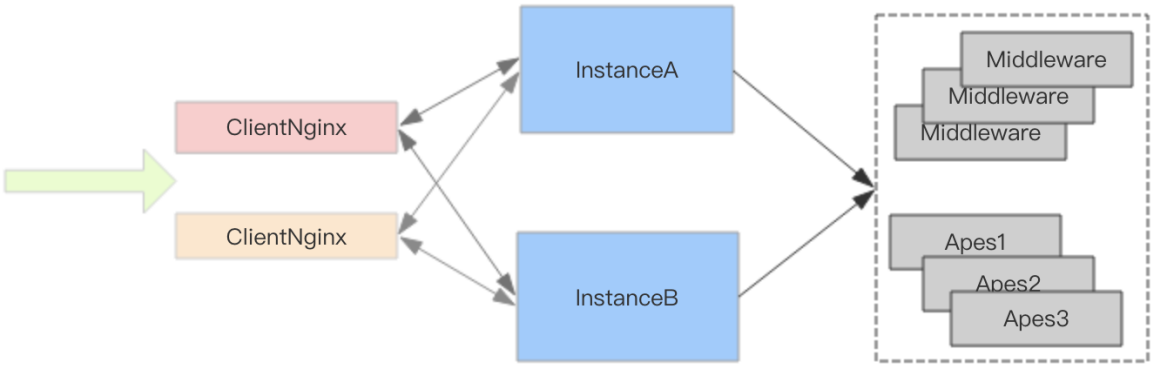
编号	软件名	描述	其他代称
1	apes	授权&网格服务	
2	animus	语音管控服务	
3	nls-cloud-gateway	语音网关服务	gateway
4	nls-cloud-asr	语音识别基础服务	asr，speech-alisr
5	nls-cloud-realtime	语音识别实时转写服务	realtime
6	nls-cloud-unify-post	语音识别统一后处理服务	unifypost，统一后处理
7	nls-cloud-filetrans	录音文件识别前置服务	filetrans
8	nls-cloud-slp	语音识别定制化服务	slp，自学习
9	nls-cloud-ai-container	语音识别模型训练服务	aicontainer
10	nls-cloud-tts	语音合成基础服务	tts，nls-tts
11	nls-cloud-tts-evolution	语音合成定制化服务	ttsevo，evolution，nls-tts-veolution
12	nls-cloud-sdm	语音媒体资源接口前置服务	alimrcp-server、mrcpserver、sdm

部署架构

语音2.x淘汰了原有1.x繁琐的内部部署实施的细节，采用单实例横向扩容为集群的模式，部署实施者不再需要过多关注语音服务内部调用关系，部署和扩容效率相对原有方案有极大提升。
下文是推荐的高可用方案：

基于ClientSide高可用方案

- 适合场景： 调用方和语音服务均部署在生产环境，服务对服务的调用。如在同一个网络域，推荐使用基于阿里自研的vipserver方案
- 如下是在调用测自主做轮训调用，实际是通过nginx，或者调用侧代码实现均可。



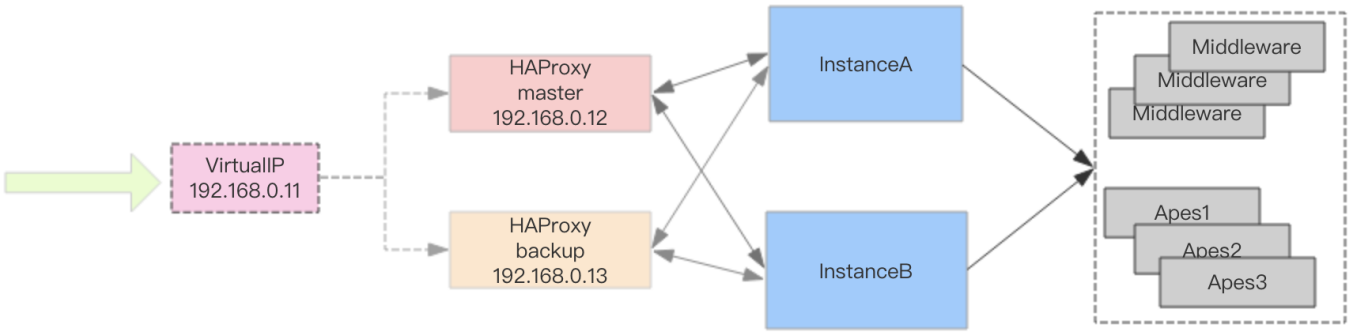
基于VIP高可用方案

1：基于HAPorxy+Keepalived 等类似开源方案 http://www.keepalived.org/

2：基于F5 等类似方案 https://f5.com/zh

3：阿里云SLB服务 https://www.aliyun.com/product/slb

- 适合场景： 客户应用为CS架构，C端量比较大且直接调用语音服务，修改调用端配置相对困难的环境。



基于阿里自研的vipserver高可用方案

在v2.5版本及以上，基于java，linux c++ sdk提供自研的vipserver的负载均衡高可用方案

技术介绍：https://blog.csdn.net/heyc861221/article/details/80126013

1：需要调用方和语音服务在同一个网络域中

2：使用v2.5版本及以后提供的sdk，详情请参考sdk文档

双机高可用方案

- 郑重声明我们不提供主备的方案，只提供多活，对于小规模，也就是常规的只有双机的情况下，其高可用方案如下：

1：理解成2个单实例(一体机)提供服务，调用侧自行进行轮训负载均衡以及异常重试等机制

2：在出现异常时将损失一半的处理能力，所以对调用量需合理评估，以及做好监控和及时恢复

资源标准

客户须知(重要，请务必认真了解)：

1. 生产环境请使用高可用配置（实例数N>1，全主模式)，如果不使用高可用(即使用单实例版N=1)，本协议则视作客户知晓并承担相应单点故障导致的损失；
2. 性能为N * T（单实例最高性能），为了保证整体服务稳定性，平稳并发或吞吐量应保持水位80%以下；实例每离线1，则减少对应T，本协议视作客户知晓并承担相应实例减少导致的损失；
3. 推荐配置为实际生产实践的最佳配置，如果客户主动选择低于该推荐配置，则视作客户知晓并承担相应风险以及后续售后服务的困难性；
4. 中间件服务比如mysql，redis，mongodb可以复用，但请保证负载合理分配，基础中间件所有权归原作者所有，基础中间件由用户自行维护，语音服务提供商不提供相关运维和可靠性，易用性等指标承诺，自带中间件镜像仅供测试使用；
5. 所有资源按独占规划，如果存在共享(超卖),则视作客户知晓并承担相应风险以及后续售后服务的困难性；
6. 目前仅支持物理机，阿里云ECS，基于物理机或ECS的容器化方案，对于其他虚拟化设备，其稳定性，性能无法承诺相关指标；
7. 资源配置合规时不同应用可以混部在同一设备上，由于默认未做资源强隔离，负载过高时会互相影响，本协议视作户客户知晓该风险点；

注意：

- N=1　单实例模式
- N>1　多实例高可用模式
- Apes是必选服务，下面所有模组都需要Apes进行中心授权和管理

测试开发环境标准

测试开发环境可以所有应用单机部署，最低配置如下：

应用	CPU	核数	内存	硬盘	总台数
语音识别全系服务	主频2.5GHz+	32C	128G+	500G+	1
语音合成全系服务	主频2.5GHz+	32C	128G+	500G+	1

生产环境配置标准

基础服务

应用	CPU	核数	内存	硬盘	总台数
animus	主频2.5GHz+	4C	8G	20G	N
apes	主频2.5GHz+	4C	2G	10G	1 单实例版本 3 集群版本 注意： 该服务不支持 双实例模式
mysql	主频2.5GHz+	4C	8G	50G	1 单实例模式 2 主备模式 >=3 集群模式 UTF-8 编码

录音文件识别

应用	CPU	核数	内存	硬盘	总台数
nls-cloud-asr	主频2.5GHz+	32C	128G（具体取决于模型，最小64G）	500G	N
nls-coud-filetrans	主频2.5GHz+	8C	16G	50G	N
nls-cloud-gateway	主频2.5GHz+	4C	8G	50G	N
redis	主频2.5GHz+	2C	8G	50G	1 单实例模式 2 主备模式 >=3 集群模式

一句话识别

应用	CPU	核数	内存	硬盘	总台数
nls-cloud-asr	主频2.5GHz+	32C	128G（具体取决于模型，最小64G）	500G	N
nls-cloud-gateway	主频2.5GHz+	4C	8G	50G	N
nls-cloud-unify-post	主频2.5GHz+	4C	8G	50G	N
redis	主频2.5GHz+	2C	8G	50G	1 单实例模式 2　　主备模式 >=3 集群模式

实时语音转写

应用	CPU	核数	内存	硬盘	总台数
nls-cloud-asr	主频2.5GHz+	32C	128G（具体取决于模型，最小64G）	500G	N
nls-cloud-gateway	主频2.5GHz+	4C	8G	50G	N
nls-cloud-realtime	主频2.5GHz+	4C	8G	50G	N
nls-cloud-unify-post	主频2.5GHz+	4C	8G	50G	N
redis	主频2.5GHz+	2C	8G	50G	1 单实例模式 2　　主备模式 >=3 集群模式

语音合成

应用	CPU	核数	内存	硬盘	总台数
nls-cloud-tts	主频2.5GHz+	32C	96G（男女2个声音模型）	300G	N
nls-cloud-gateway	主频2.5GHz+	4C	8G	50G	N
mysql	主频2.5GHz+	4C	8G	50G	1 单实例模式 2 主备模式 >=3 集群模式 UTF-8 编码

语音识别定制化(语言模型自学习)

- 定制化服务属于旁置链路，除Redis外，其他由用户自行选择是否做容灾
- 不含声学模型训练，声学训练需要GPU支持

应用	CPU型号	核数	内存	硬盘	总台数
nls-cloud-ai-container	主频2.5GHz+	32C	64G	500G	N
nls-cloud-slp	主频2.5GHz+	4C	8G	100G	N
mysql		4C	8G	100G	1 单实例模式 2 主备模式 >=3 集群模式
mongodb		4C	8G	100G	1 单实例模式 2 主备模式 >=3 集群模式

redis	主频2.5GHz+	2C	8G	100G	1 单实例模式 2　　主备模式 >=3 集群模式 和 实时转写，一句话 所用的redis必须是同一套

语音合成定制化

- 定制化服务属于旁置链路，除mysql外，其他由用户自行选择是否做容灾

应用	CPU	核数	内存	硬盘	总台数
nls-cloud-tts-evolution	主频2.5GHz+	4C	8G	50G	N
mysql	主频2.5GHz+	4C	8G	50G	1 单实例模式 2　　主备模式 >=3 集群模式 UTF-8 编码

MRCP

媒体资源控制协议（Media Resource Control Protocol, MRCP）是一种通讯协议，用于语音服务器向客户端提供各种语音服务(如语音识别和语音合成)。

应用	CPU	核数	内存	硬盘	总台数
nls-cloud-sdm	主频2.5GHz+	4C	8G	500G+ （如需保存录音资源）	N