# EDA场景下E-HPC混合云搭建手册

阿里云E-HPC弹性高性能计算

## 前言

### 为何混合云是EDA场景下hpc集群交付为主流

**从用户的业务角度来讲**，EDA的业务具有潮汐性和突发性，因此选择云上资源满足突发需求可优化成本；**从EDA行业的工具栈和软件栈的角度来看**，EDA行业使用的调度器、行业应用多为商业版本，客户调度器管控节点和许可证服务器通常只能部署于线下，因此混合云的方式能兼顾线下资源和线上弹性资源；**从EDA行业对保密性的需求来看**，半导体公司多年的设计验证结果是集群计算的结果，选择混合云可以最大化的保证安全性。

因此，需要为EDA场景设计兼容性高、稳定性高、弹性能力强的混合云架构。

### E-HPC混合云交付的优势

阿里云E-HPC深耕超算多年，在多个超算行业中积累了丰富的经验，基于对于EDA行业应用支持的积累以及对超大规模集群架构的理解，设计完成针对EDA行业的E-HPC混合云产品。充分的展示了高兼容性、高稳定性以及高弹性的优势。具体体现在：

1. 高兼容性。E-HPC混合云集群构建时支持对多种客户自持调度器的接入，包括客户自持lsf、slurm等，客户可以基于调度器插件的方式对接E-HPC的管控能力和弹性伸缩能力，做到对多种调度器的灵活兼容；E-HPC支持接入多种域控管理工具，包括windows AD server、Openldap、nis等；支持灵活的挂载线上或线下共享存储；这些能力可以使客户自有集群可快捷的获得线上资源弹性，同时，无需额外部署配置调度器，无需额外开启新一套的域账号体系，无需额外安装一套EDA工具。

2. 高稳定性。E-HPC混合云在线上节点扩容时，支持多种扩容策略；其中池内弹性的扩容策略，用户可以将线上集群的ip地址规划与其自有集群的规划统筹考虑，无需担心弹性资源的引入导致其已有集群的解析混乱；同时，基于池内弹性的扩容策略，E-HPC混合云通过调度器预置节点的方式，可以在不对调度器进行重启或reconfig的操作，即可完成弹性资源并入计算集群，做到混合云线上的弹性扩容不影响线下正在进行的业务。

3. 高弹性。E-HPC借助阿里云已有的丰富资源池和云企业网能力，可以将多个地域的资源供给给线下调度器统一管理，自动伸缩服务借由混合云的代理节点感知集群当前任务负载，做到任务级的云上资源自动伸缩，同时伸缩方案可以定制多款实例类型。

## 一、局点环境确认

局点环境是指客户在进行混合云创建前，对其自身环境的确认。由于线下环境的差异，E-HPC混合云的尝试以最大的兼容性来满足不同的线下集群构建混合云需求，若下文中的技术细节依然有不清晰，请联系阿里云同学进行支撑。

混合云构建的局点环境确认，硬件层面需要确认打通线上线下网络，确认共享存储在云上云下可同时挂载；集群服务层面，需要确认DNS服务同时可以对云上云下节点进行域名解析，需要确认域控组件可对云上云下节点进行注册，客户需要根据其网段规划在其dns中录入预设节点的正/反向域名解析，同时需要客户将预置节点信息(hostname)写入调度器基础配置文档。具体环节如下：

### 云上云下互通测试

假设云上节点为ecs01，云下节点为offline01：

1）确认节点间可以相互ping通；

2）确认两节点间ssh可以相互访问，即ssh使用的端口（通常tcp协议的22端口）可以相互访问，可以使用nc或telnet进行确认；调度器多基于ssh服务来完成集群维护。

3）确认DNS服务可以使用，通常为53端口（注意DNS需要UDP协议），可以使用nc或telnet进行确认；

4）其他服务可互通确认，如windows AD server，Openldap，nis等需要的端口。

### 客户侧挂载点确认

假设待使用挂载点为：03cd949fd9-pis41.cn-hangzhou.nas.aliyuncs.com

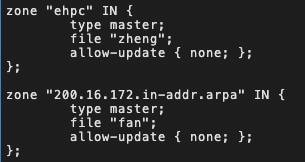
在挂载点上建立文件夹/test1, 使用mount命令，在云上云下节点建立挂载点，测试挂载后能读到挂载点上内容。

### 客户侧dns状态确认

我们以centos上bind组件实现的dns server为例（windows平台dns server搭建后，测试方式一致）：

1）云上DNS配置：

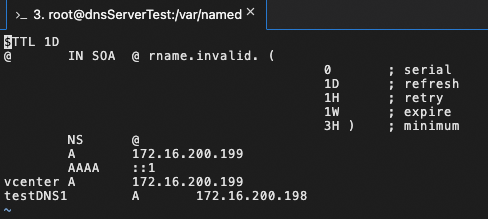
a）集群所在zone确定，常规配置在/etc/named.rfc1912.zones，也可在 /etc/named.conf中，因其引用named.rfc1912.zones。具体配置如下图，其中文件zheng负责正向解析，即hostname到ip解析，文件fan为反向解析，负责ip到hostname解析。



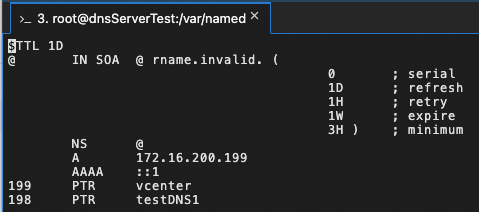
b）dns服务端口及访问限制，如下图所示，设定为53端口，任何主机都可以使用 该dns server。



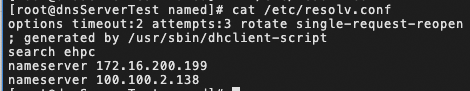
c）正反向解析中录入集群节点信息，我们以vcenter和testDNS1两个节点为例， 配置文件位置位于/var/named/目录下，具体配置如下，这里我们将vcenter.ehpc的域名解析映射到172.16.200.199，将testDNS1.ehpc域名映射到172.16.200.198。

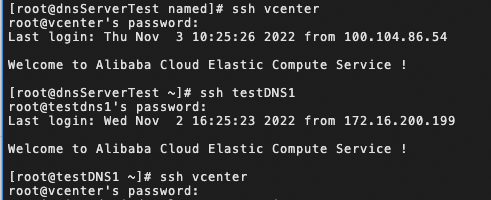


d）反向解析配置



e) 重启named服务，测试server端可以提供dns服务，之后测试vcenter和testDNS1可以相互解析，注意，为了配合之后调度器中对节点的命名，我们在resolv.conf文件中加入对ehpc的搜索，并将dns解析服务器都设置成之前配置的dns服务器位置。

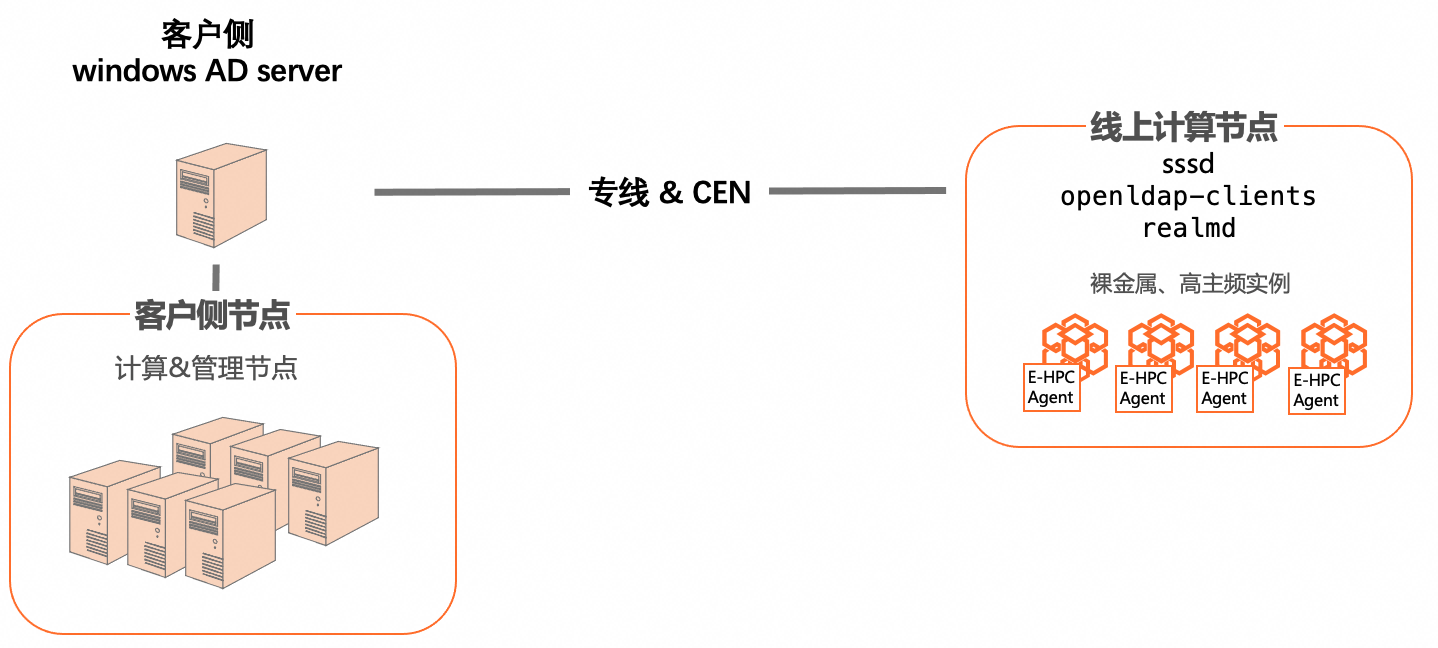




### 客户侧域管理状态确认

1）windows AD server状态确认

E-HPC混合云方案推荐使用云上资源直连线下AD server的方案，如下图：



如上图案例中，线下AD server构建的域为ldap-win002.lework.com，那么云上ecs01加入该节点通过如下命令可以完成，该命令需要输入AD server端进行认证的用户的密码（如Administrator，也可以是其他）。

realm join --user=Administrator -v ldap-win002.lework.com

之后根据客户使用习惯对sssd文件进行修改并重启sssd服务，即可在centos主机上使用AD server中已有的账户，如wutest。

systemctl restart sssd

su wutest

2）Openldap状态确认

同AD server的接入方案的架构，我们推荐云上节点直连线下Openldap服务。具体命令如下，其中ldapserver的值为线下ip，basedn也是现在使用的域：

authconfig --enableldap --ldapserver='ldap://192.168.0.232/,ldap:///' --ldapbasedn='dc=ehpc-sh-MDiQgx5qaQ,dc=com' --update

systemctl restart nslcd

### 客户侧调度器状态确认

1. 确认已挂载LSF安装目录

由于LSF客户端的安装需要获取LSF安装目录，所以需要确定挂载的共享存在是否包含LSF安装目录，例如安装目录为/opt/lsf（以下内容将以该目录为LSF安装目录叙述）

1. 修改lsf.cluster.ehpc加入节点并启动LSF客户端

修改/opt/lsf/conf/lsf.cluster.ehpc 文件，将客户端节点以对应格式加入到配置文件节点列表中并保存

执行 /opt/lsf/10.1/install/hostsetup --top="/opt/lsf" --boot="y"

1. 配置lsf.sudoers文件

echo 'LSF\_STARTUP\_PATH=/opt/lsf/10.1/linux3.10-glibc2.17-x86\_64/etc/' >> /etc/lsf.sudoers

echo 'LSF\_STARTUP\_USERS="lsfadmin"' >> /etc/lsf.sudoers

chmod 600 /etc/lsf.sudoers

1. 配置LSF调度器环境变量并启动客户端服务

执行 echo 'source /opt/lsf/conf/profile.lsf' >> /etc/profile

执行 echo 'source /opt/lsf/conf/cshrc.lsf' >> /etc/csh.cshrc

执行 source /etc/profile之后，执行 service lsf status进行服务状态确认

5）检查客户端节点是否加入

在线下调度器节点上执行 yes | lsadmin reconfig 重新配置lsf集群，reconfig之后检查 qhost，查看是否存在该客户端节点

客户端节点执行 yes | lsadmin limstartup ， yes | lsadmin resstartup，yes | badmin hstartup来启动LSF守护进程服务

## 二、混合云创建API

由于线下集群的差异性，建议使用E-HPC的api完成集群创建，具体使用参数如下：

const Core = require('@alicloud/pop-core');

var client = new Core({

accessKeyId: 'xxxxxxxxxx', //客户keyId

accessKeySecret: 'xxxxxxxxxx', //客户key秘钥

endpoint: 'https://ehpc.cn-hangzhou.aliyuncs.com', //混合云云上集群创建服务点

apiVersion: '2018-04-12'

});

var params = {

"RegionId": "cn-hangzhou",//混合云云上集群创建区域

"Name": "armOnline6-zailong",//混合云云上集群名称

"OsTag": "CentOS\_7.8\_64",//集群所需操作系统标签

"ImageId": "m-bp136x83f14rf27u3z9c",

"VpcId": "vpc-bp1q4vw1dug6bojdg61cd",//集群所需VPC

"VSwitchId": "vsw-bp1fjyhlhiqbuuyxhznid",//集群所需交换机

"EcsOrder.Compute.InstanceType": "ecs.c7.large",//集群创建所需实例类型

"ZoneId": "cn-hangzhou-k",//实例所在区域

"Nodes.1.HostName": "manager",//线下集群管理节点命名

"Nodes.1.IpAddress": "172.16.100.78",//线下集群ip地址

"Nodes.1.Role": "ResourceManager",//线下集群该节点角色描述

"Nodes.1.AccountType": "custom",//线下节点使用的域控管理，如客户自持则为custom

"Nodes.1.SchedulerType": "custom",//线下调度器，如果是客户自持都选择custom

"Nodes.2.HostName": "proxymgr",//云上代理节点名称

"Nodes.2.IpAddress": "172.16.200.252",//云上代理节点IP

"Nodes.2.Role": "AccountManager",//不用修改

"Nodes.2.AccountType": "custom",//不用修改

"Nodes.2.SchedulerType": "lsf",//云下调度器名称

"Location": "ProxyOnline",//集群位置，eda混合云都位ProxyOnline

"VolumeType":"local",

"VolumeMountpoint": "03cd949fd9-pis41.cn-hangzhou.nas.aliyuncs.com",//挂载点位置

"Password": "",//集群密码

"WinAdPar.AdUser": "Administrator",//使用windows AD时，AD侧的用户名

"WinAdPar.AdUserPasswd": "",// AD侧的用户的密码

"WinAdPar.AdDc": "ad-hybrid001.ehpcad.com",//AD侧给云上主机使用规划的用户域

"WinAdPar.AdIp": "172.16.200.33",//线下windows AD server的IP地址

// "OpenldapPar.LdapServerIp": "172.16.200.23",//线下Openldap的IP地址

// "OpenldapPar.BaseDn": "ehpctest",//线下Openldap管控账户所使用的Dn

"ClientVersion": "1.0.1"//集群版本

"Plugin": '{"pluginMod":"image", "pluginLocalPath":"/opt/plugin" }' //插件模式及路径

}

var requestOption = {

method: 'GET',

};

client.request('CreateHybridCluster', params, requestOption).then((result) => {

console.log(JSON.stringify(result));//混合云创建接口

}, (ex) => {

console.log(ex);

})

### 线下挂载点接入

当前，混合云接入的方式有两种，可以灵活的支持线上或线下的挂载点。

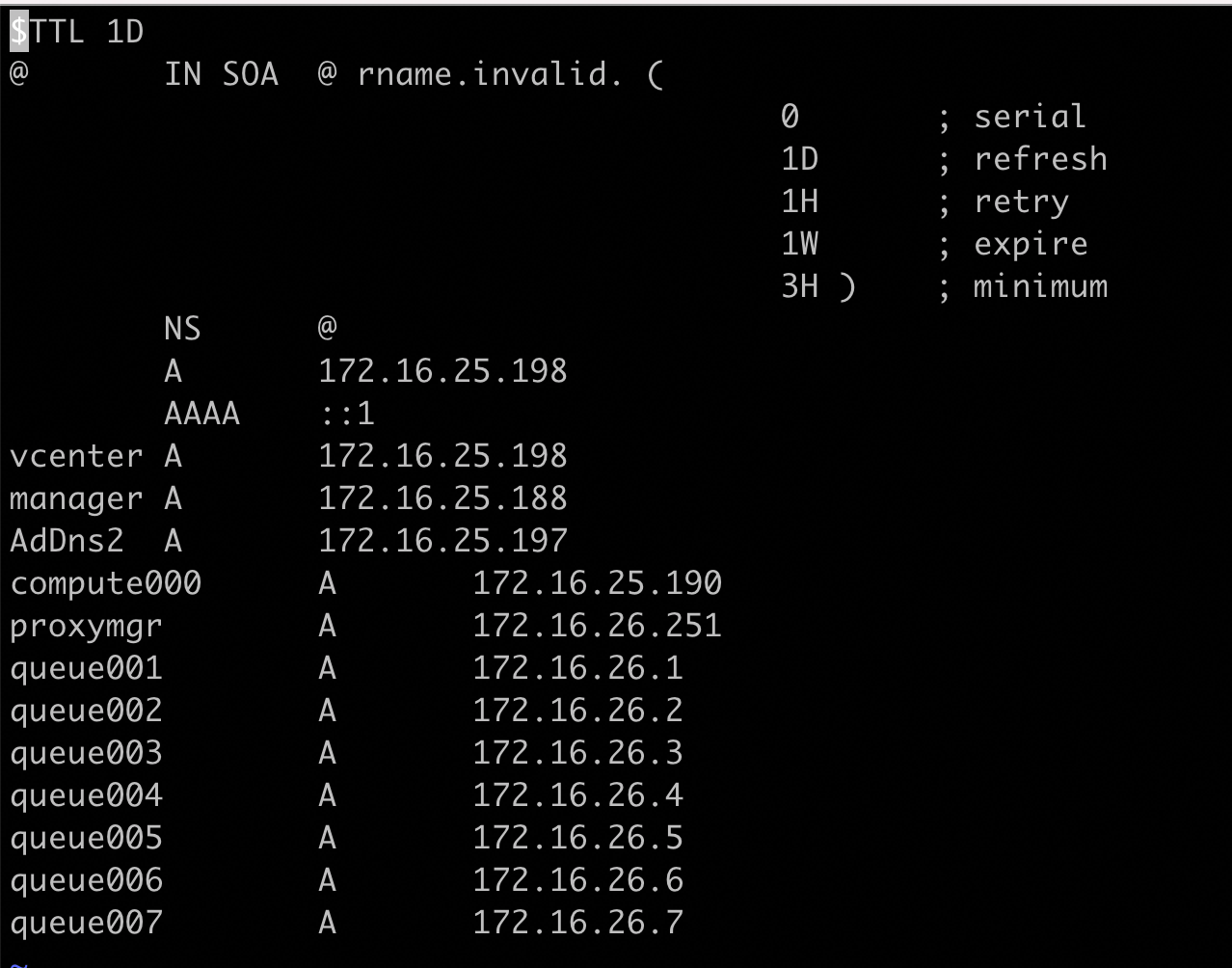
若"VolumeType":"local",则ehpc混合云线上集群会按照镜像（下章详述）中已挂载的位置，这样方面线上节点直接对接完成客户线下的应用（如/opt下的应用），同时直接对接完成/home目录下的用户文件。推荐使用此种方式。

还可以在线上线下共享存储上创建ehpc所需要的文件路径，并进行挂载。如："VolumeType"赋值为空，同时对"VolumeMountpoint"赋值（如上节api示例），E-HPC会默认在该挂载点下创建集群文件夹，例如ehpc.xxxxx，该文件夹内包含/opt, /home, /ehpcdata三个文件。如需挂载其他文件夹，可在镜像中进行约定。线下集群挂载线上节点的场景下，可使用此种方式。

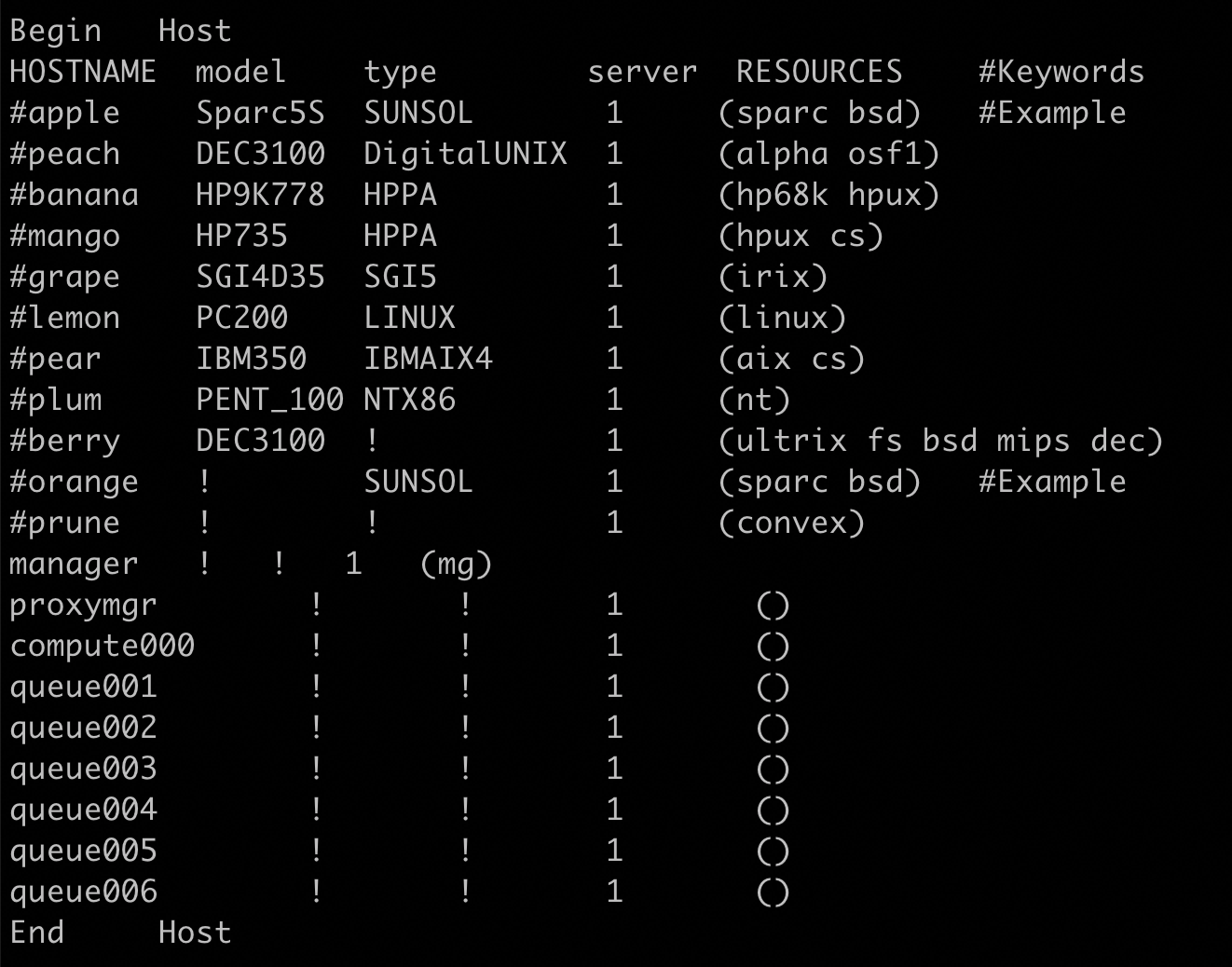
### 云上弹性节点在线下的预设

在客户侧dns中，将预设节点的hostname和ip节点的正向解析进行录入。我们约定云上节点的命名规则为，计算节点名称前缀+数字索引，单个虚拟交换机下支持一种计算节点名前缀，例如：计算节点名为queue，则进行池内弹性伸缩的节点列表为: queue001，queue002，最大为queue250。索引值为该节点的ip地址最后一位，例如: 云上节点规划使用的交换机网段为：172.16.26.0/24, 则queue001，对应的ip地址为172.16.26.1，queue002对应172.16.26.2，queue250对应172.16.26.250。这些预设节点都需要写入到dns的正向解析中（反向选填）。同时将这些预设节点名称，写入到lsf调度器对应配置文件内。具体如下图。

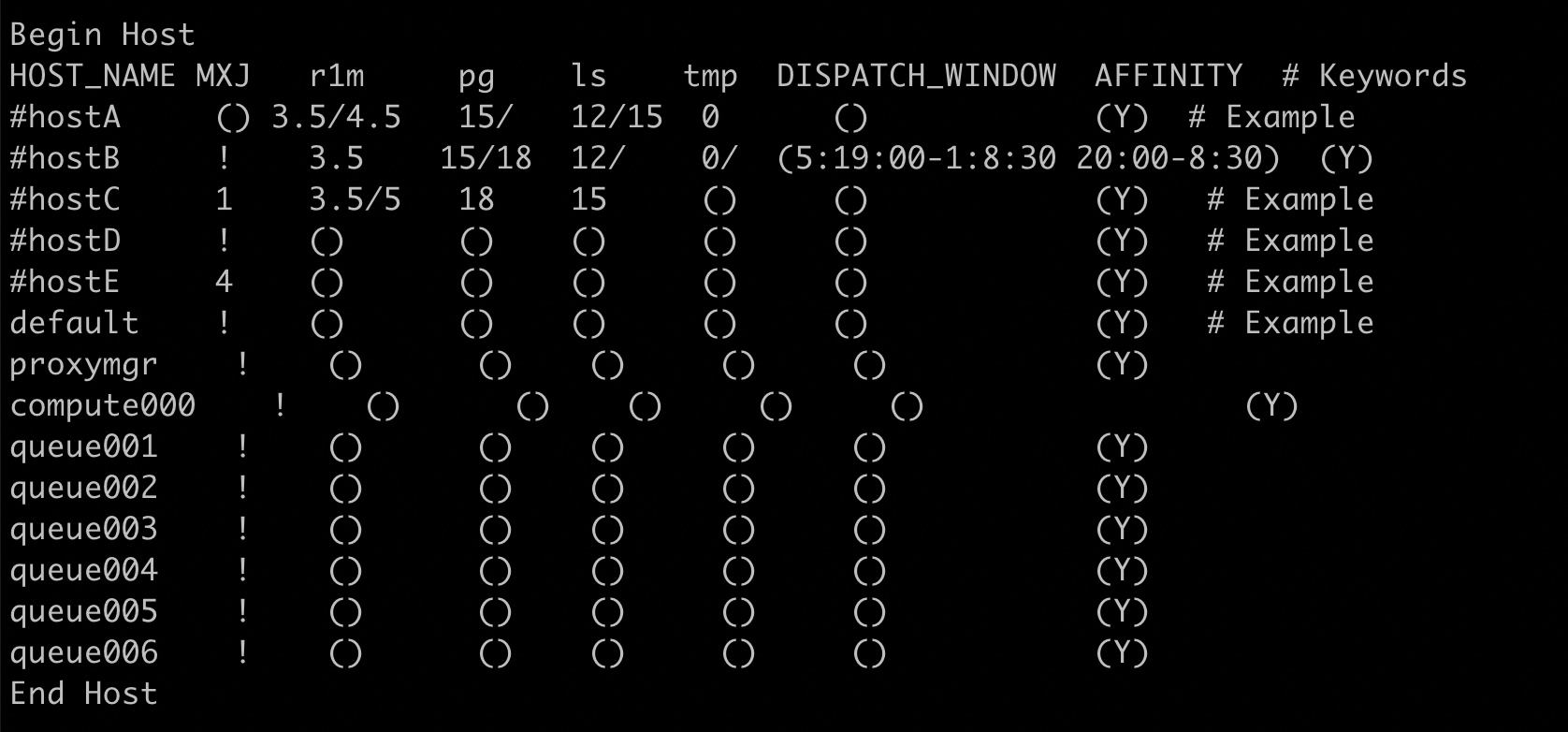
1. DNS服务器上正向解析的预设节点配置



1. lsf.cluster文件中的预设节点配置



1. lsb.hosts文件中预设节点的配置



1. 将云上计算节点的前缀名写入到proxymgr节点上的 /usr/local/cloudnode\_prefix 文件中，以便能够正确识别云上计算节点负载

### 客户侧windows AD server接入

"WinAdPar.AdUser": "Administrator",//使用windows AD时，AD侧的用户名

"WinAdPar.AdUserPasswd": "Alihpc123",// AD侧的用户的密码

"WinAdPar.AdDc.1": "ad-hybrid001.ehpcad.com",//AD侧给云上主机使用规划的用户域

"WinAdPar.AdIp": "172.16.200.33",//线下windows AD server的IP地址

如上所示，为线下AD server接入时需要输入的配置参数。AD server中需要负责对用户管理进行配置。推荐在AD server中创建至少两个组（group）对用户进行管理，以支持区分sudo权限用户和普通用户。

### 客户侧Openldap接入

"OpenldapPar.LdapServerIp": "172.16.200.23",//线下Openldap的IP地址

"OpenldapPar.BaseDn": "ehpctest",//线下Openldap管控账户所使用的Dn

如上所示，线上扩容节点会以线下openldap为server，并获取其上对用户、用户组和权限的配置。

### 客户侧lsf调度器接入--调度器插件部分

1. 将插件解压到manager共享路径（联系团队获取插件）

mkdir /opt/plugin

mv /root/lsf\_hybrid\_dns\_plugin\_v2.zip /opt/plugin/

cd /opt/plugin/

unzip lsf\_hybrid\_dns\_plugin\_v2.zip

1. 根据实际LSF软件环境调整插件内容
2. 在制作LDAP自定义镜像的同时来配置计算节点LSF客户端启动服务（需要根据具体LSF安装位置调整）

3.1 创建script目录：mkdir -p /usr/local/lsf/script

3.2 创建log目录：mkdir -p /usr/local/lsf/log

3.3 创建启动进程脚本：/usr/local/lsf/script/lsfinit.sh

#!/usr/bin/env bash

#

# EHPC lsf, (c)Copyright 2022

#

# Install the LSF master, compute in ALIYUN EHPC platform

LSF\_DIR=/opt/lsf

LOG\_FILE=/usr/local/lsf/log/lsfinit.log

EHPC\_CONF=/root/ehpc.conf

LSF\_ADMIN=lsfadmin

IS\_MANAGER=

IS\_LOGIN=

IS\_COMPUTE=

LSF\_CLUSTER\_CONF=${LSF\_DIR}/conf/lsf.cluster.ehpc

LSF\_HOST\_CONF=${LSF\_DIR}/conf/lsbatch/ehpc/configdir/lsb.hosts

LSF\_QUEUE\_CONF=${LSF\_DIR}/conf/lsbatch/ehpc/configdir/lsb.queues

LSF\_PARAM\_CONF=${LSF\_DIR}/conf/lsbatch/ehpc/configdir/lsb.params

LSF\_PROFILE\_ENV="source ${LSF\_DIR}/conf/profile.lsf"

LSF\_CSHRC\_ENV="source ${LSF\_DIR}/conf/cshrc.lsf"

echo -e "" | tee -a $LOG\_FILE

echo -e "-------------------------------------------------------" | tee -a $LOG\_FILE

mkdir -p /usr/local/lsf/log

touch ${LOG\_FILE}

# 0. Check lsf service

source /etc/profile && service lsf status > /dev/null 2>&1

if [ "$?" -eq "0" ]; then

echo -e "INFO: lsf service is startup, skip follow steps" | tee -a $LOG\_FILE

exit 0

else

source /etc/profile && service lsf start > /dev/null 2>&1

source /etc/profile && service lsf status > /dev/null 2>&1

if [ "$?" -eq "0" ]; then

echo -e "INFO: lsf daemons is restartup, skip follow steps" | tee -a $LOG\_FILE

exit 0

else

echo -e "INFO: lsf daemons is , skip follow steps" | tee -a $LOG\_FILE

fi

fi

# 1. Check EHPC config

if [[ ! -f /root/ehpc.conf ]] ; then

echo -e "ERROR: E-HPC service is not installed." | tee -a $LOG\_FILE

exit 1

fi

# 2. Retry check user lsfadmin

count=0

while [ 0 -eq 0 ]

do

id -u $LSF\_ADMIN > /dev/null 2>&1

if [ "$?" -eq "0" ]; then

echo -e "INFO: check lsf\_admin success." | tee -a $LOG\_FILE

break;

else

let count=${count}+1

if [ ${count} -eq 60 ]; then

echo -e "ERROR: check lsf\_admin exceed 5 min, exit" | tee -a $LOG\_FILE

exit;

fi

echo -e "WARN: check lsf\_admin ${count} times failed." | tee -a $LOG\_FILE

sleep 5

fi

done

# 3. Get current node type by ehpc.conf Role

NODE\_ROLE=`cat ${EHPC\_CONF} | grep Role | awk -F "=" '{print $2}'`

# 4. Loop the Roles

for ROLE in `echo $NODE\_ROLE | tr "," " "`

do

# Scheduler

if [[ "$ROLE" == "ResourceManager" ]] ; then

IS\_MANAGER=1

fi

# Login

if [[ "$ROLE" == "LoginNode" ]] ; then

IS\_LOGIN=1

fi

# Compute

if [[ "$ROLE" == "ComputeNode" ]] ; then

IS\_COMPUTE=1

fi

# ProxyMgr

if [[ "$ROLE" == "ProxyManager" ]] ; then

IS\_PROXYMGR=1

fi

done

# 5. Install LSF client

if [[ "X$IS\_COMPUTE" == "X1" ]] || [[ "X$IS\_PROXYMGR" == "X1" ]] ; then

COMPUTE\_HOST=`hostname -s`

echo -e "INFO: Start to install LSF ${COMPUTE\_HOST} ..." | tee -a $LOG\_FILE

${LSF\_DIR}/10.1/install/hostsetup --top="${LSF\_DIR}" --boot="y" | tee -a $LOG\_FILE

echo "${LSF\_PROFILE\_ENV}" >> /etc/profile

echo "${LSF\_CSHRC\_ENV}" >> /etc/csh.cshrc

source /etc/profile && service lsf start | tee -a $LOG\_FILE

source /etc/profile && service lsf status | tee -a $LOG\_FILE

echo -e "INFO: Finish to install LSF ${COMPUTE\_HOST} ..." | tee -a $LOG\_FILE

fi

echo -e "INFO: Init successfully." | tee -a $LOG\_FILE

exit 0

3.4 创建启动服务： lsf-init.service

[Unit]

Description=Install AIP by E-HPC node roles

After=local-fs.target network-online.target cloud-final.service ehpc-config.service

Requires=network-online.target cloud-final.service ehpc-config.service

Wants=local-fs.target

[Service]

Type=oneshot

ExecStart=/bin/bash /usr/local/lsf/script/lsfinit.sh

RemainAfterExit=yes

Restart=no

[Install]

WantedBy=multi-user.target

3.6 拷贝启动服务到系统目录：cp lsf-init.service /lib/systemd/system

3.7 load启动服务目录：systemctl daemon-reload

3.8 enable启动服务目录：systemctl enable lsf-init

1. 使用上一步的自定义镜像通过API创建带有插件的混合云集群

## 镜像制作

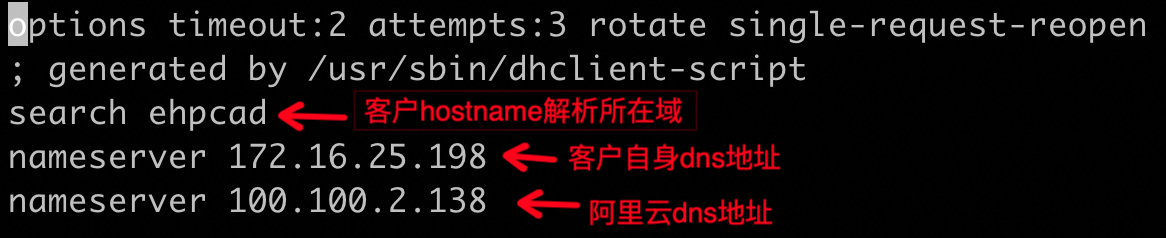
上一章节中，我们讲到了E-HPC混合云场景下在调度器及域控组件对接上的必要输入及基本原理。本章节将围绕如何制作适合该场景下的镜像展开，镜像是云计算提供的能力，按需制作镜像并在集群扩容时选择所需镜像，可以保证计算节点配置的一致性，并大大节约运维与操作成本。

### DNS配置的录入

若客户对阿里云云助手或挂载阿里云上共享存储有需求，则线上扩容节点的操作系统中需要在dns的寻址配置文件/etc/resolv.conf文件中保留阿里云默认地址中的一个，另一个地址可配置用户自己的dns。

阿里云ecs上的dns配置修改，请参考如下链接，注意/etc/cloud/cloud.cfg的配置。

<https://help.aliyun.com/knowledge_detail/188420.html?spm=a2c4g.11186623.0.0.5c001c217pUglM>



### 挂载点配置

挂载点的使用配合local模式下的VolumeType。注意制作带挂载点的镜像时，除了需要进行mount操作外，还需要在/etc/fstab内进行修改。

### AD server场景下sudo组对接

若客户需要接入AD中的sudo权限组和普通权限组，则需要在镜像的sudo配置文件中进行修改，例如ad server中我们规定adminSudo组内的用户为sudo权限用户，则需要visudo后，在最优一行加入授权操作，如下。

%adminSudo ALL=(ALL) ALL

[root@queue001 ~]# id lsfadmin

uid=272801108(lsfadmin) gid=272800513(domain users) 组=272800513(domain users),272801114(adminsudo)

## 集群测试

1. 集群创建测试



1. 线下集群中用户同步至线上节点

[root@queue001 ~]# id lsfadmin

uid=272801108(lsfadmin) gid=272800513(domain users) 组=272800513(domain users),272801114(adminsudo)

1. 线上节点挂载线下存储

使用df -h 命令查看。

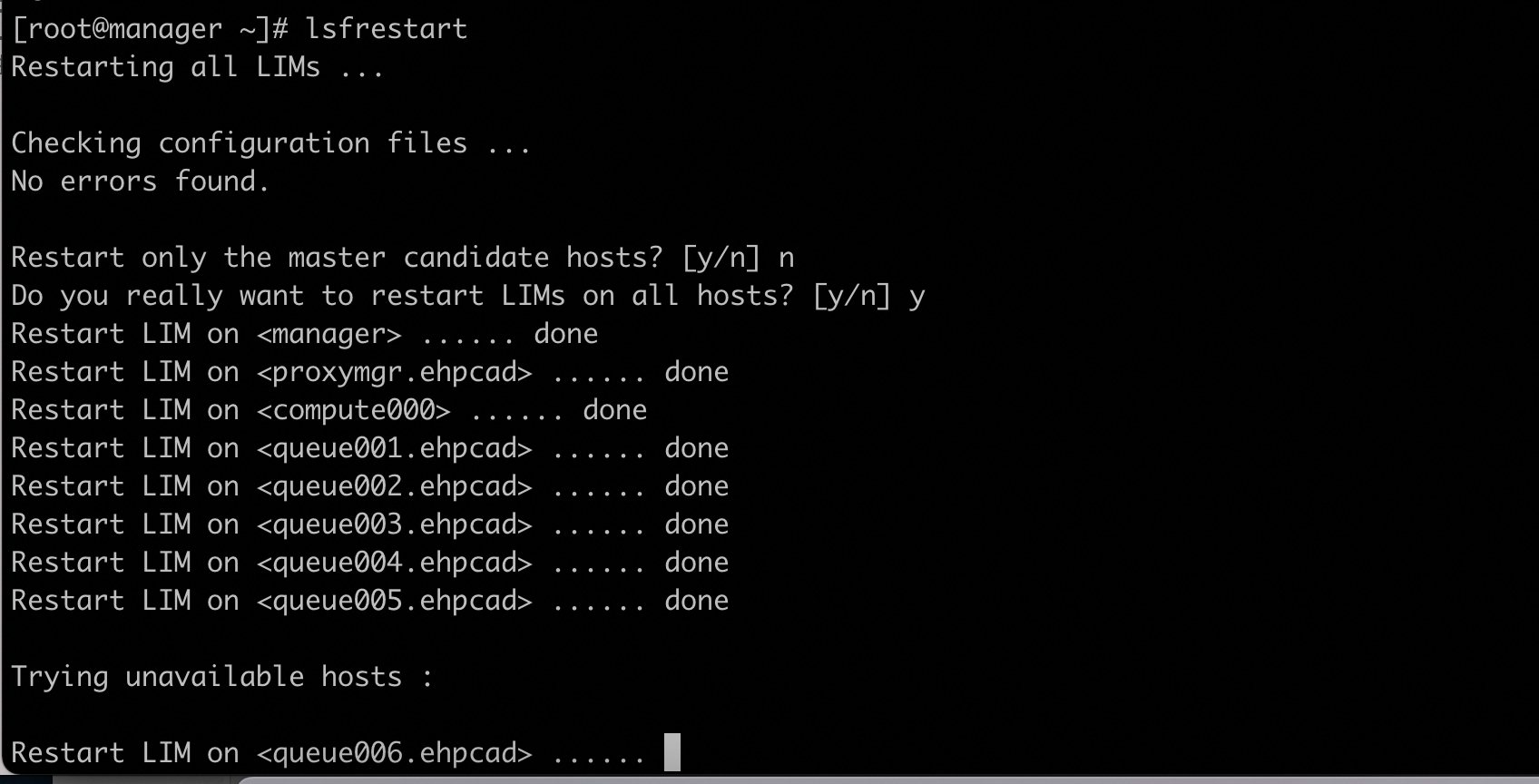
1. 线上线下节点间互ping及互ssh

此处目的为测试dns的能力。

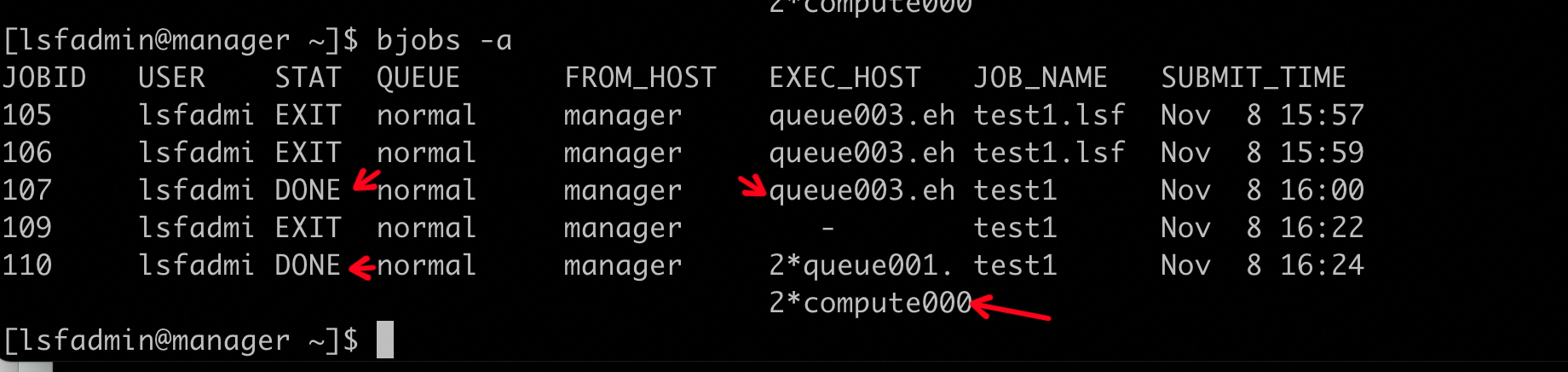
1. 用户在线上线下节点间的免密环境

使用root以外的用户，在节点间ssh，可以免密跳转。

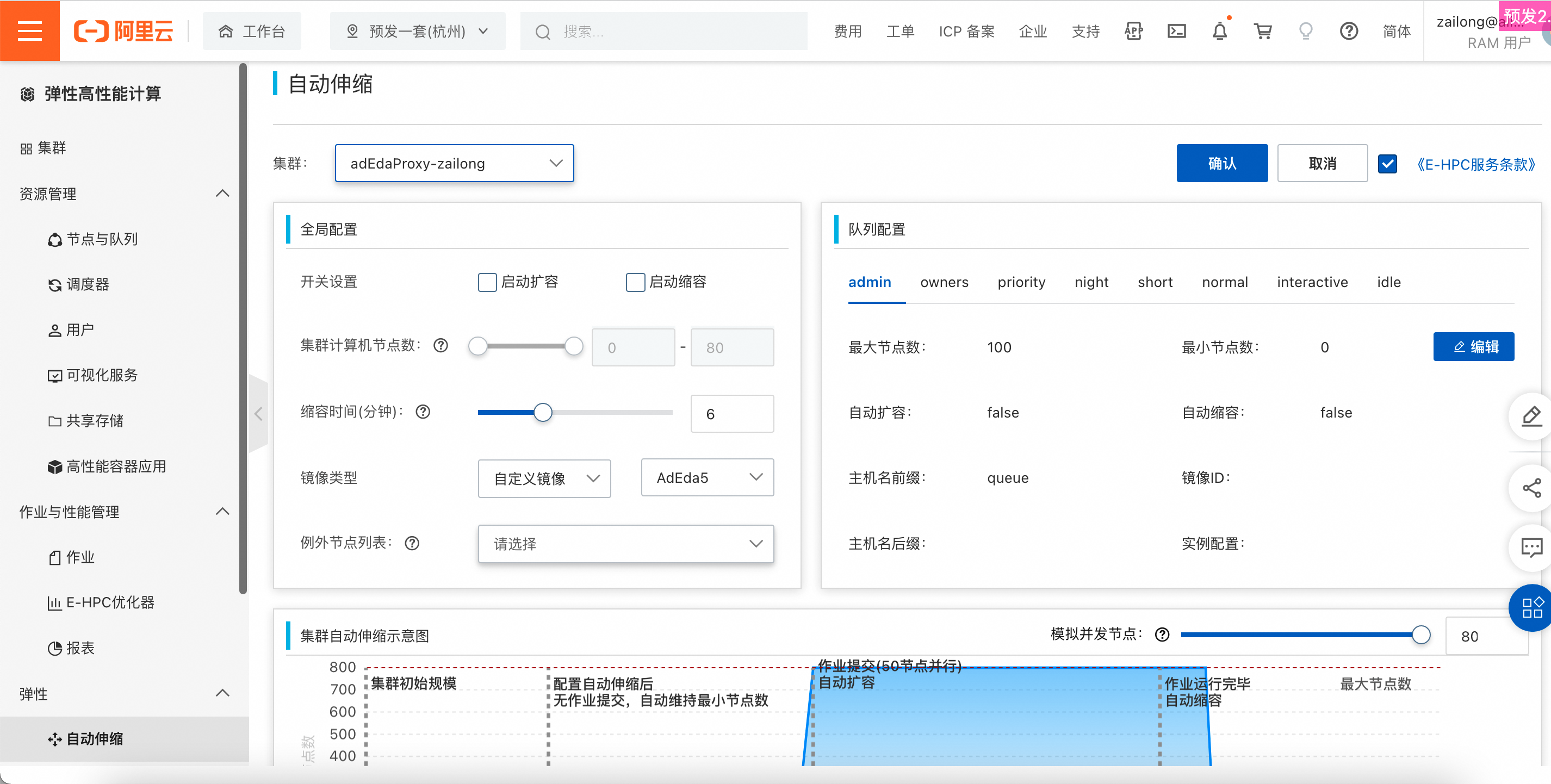
1. 预设节点配置完成后调度器重启后对预设节点的发现



1. 作业提交后，可以在线上线下一起运行



1. 作业提交和完成后，集群的自动伸缩



eda工具的输入和输出分清