阿里云

ZStack for Alibaba Cloud

共享文件系统高可用 使用教程

产品版本：V3.0.1

文档版本：20180930

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的”现状“、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含”阿里云”、Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

表 -1: 格式约定

格式	说明	样例
	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	**禁止**： 重置操作将丢失用户配置数据。
	该类警示信息可能导致系统重大变更甚至故障，或者导致人身伤害等结果。	**警告**： 重启操作将导致业务中断，恢复业务所需时间约10分钟。
	用于警示信息、补充说明等，是用户必须了解的内容。	**说明**： 导出的数据中包含敏感信息，请妥善保存。
	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	**说明**： 您也可以通过按**Ctrl** + **A**选中全部文件。
>	多级菜单递进。	**设置** > **网络** > **设置网络类型**
粗体	表示按键、菜单、页面名称等UI元素。	单击 **确定**。
courier字体	命令。	执行 **cd /d C:/windows** 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	**bae log list --instanceid** *Instance_ID*
[]或者[a|b]	表示可选项，至多选择一个。	**ipconfig** *[-all|-t]*
{}或者{a|b}	表示必选项，至多选择一个。	**swich** *{stand | slave}*

目录

法律声明 I
通用约定 I
1 安装与部署 1
1.1 概述 1
1.1.1 共享文件系统高可用 1
1.1.2 高可用技术 2
1.1.3 网络拓扑规划 2
1.2 安装部署 5
1.2.1 安装操作系统 5
1.2.2 存储网络 9
1.2.3 管理网络 10
1.2.4 云主机数据网络 11
1.2.5 网络共享存储（NFS/SMP） 12
1.2.6 高可用套件 13
1.2.6.1 功能介绍 13
1.2.6.2 部署过程 13
1.2.7 集群升级 18
1.2.7.1 内嵌服务升级 18
1.2.7.2 高可用升级 19
1.2.8 管理节点迁移 19
1.2.9 配置更新 20
1.3 其他操作 21
1.3.1 卸载操作 21
1.3.2 日志输出 21
2 高可用测试与恢复 22
2.1 计划运维 22
2.1.1 单节点需要维护 22
2.1.1.1 单节点关闭 22
2.1.1.1.1 该节点运行非管理节点主机 22
2.1.1.1.2 该节点运行管理节点主机 23
2.1.1.2 单节点启动 24
2.1.2 三节点需要维护 25
2.1.2.1 三节点关闭 25
2.1.2.2 三节点启动 26
2.2 异常处理 28
2.2.1 单节点异常处理 28
2.2.1.1 单节点异常掉电 28

2.2.1.1.1 该节点运行非管理节点主机 28
2.2.1.1.2 该节点运行管理节点主机 29
2.2.1.2 单节点网络异常 31
2.2.1.2.1 该节点运行非管理节点主机 31
2.2.1.2.2 该节点运行管理节点主机 32
2.2.2 两节点异常处理 33
2.3 节点修复 34
2.3.1 单节点故障修复 34
2.3.2 两节点故障无法修复 36
3 命令行使用手册 37
3.1 简介 37
3.2 -h 帮助内容 37
3.3 -v 版本信息 38
3.4 check-config 配置检查 38
3.5 install -p -c 安装命令 38
3.6 uninstall 卸载命令 40
3.7 config-sample 样本配置生成 40
3.8 export-config 当前配置生成 41
3.9 migrate 迁移命令 41
3.10 import-config 配置升级 42
3.11 status 状态信息 42
3.12 stop 集群关闭 44
3.13 start 集群启动 45
3.14 reset-MNVM-password 管理节点主机重置root密码 45
专有云术语表 46
混合云术语表 49

1 安装与部署

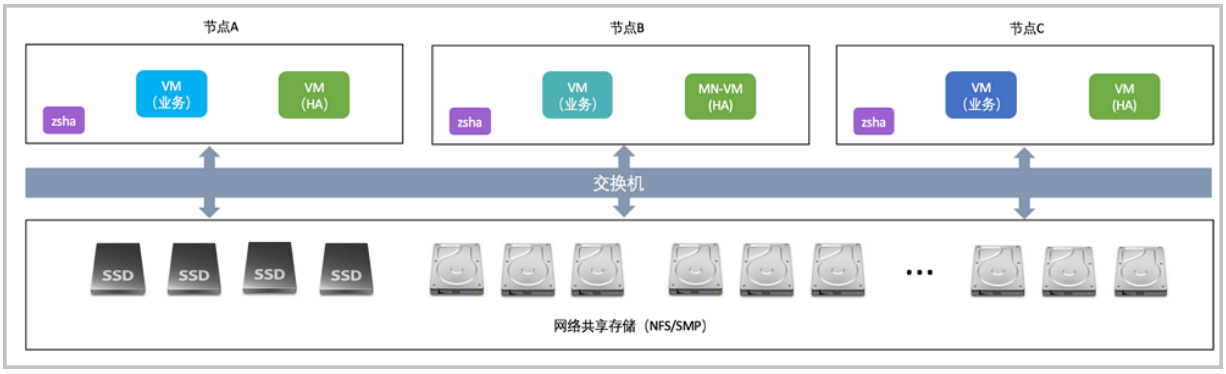
1.1 概述

1.1.1 共享文件系统高可用

ZStack for Alibaba Cloud提供基于网络共享存储的管理节点高可用方案，本文档主要介绍NFS/SMP共享文件系统场景。

以下是图 *1-1: 基于NFS/SMP共享文件系统高可用 三节点经典模型*：

图 1-1: 基于NFS/SMP共享文件系统高可用 三节点经典模型



共享文件系统高可用方案有以下特点：

- 通过NFS/SMP共享文件系统，由各个节点的固态硬盘（SSD）和机械硬盘（HDD）提供统一的存储系统
- 通过虚拟化技术提供云主机服务，并将其数据存放在NFS/SMP共享文件系统中
- ZStack for Alibaba Cloud管理节点主机以云主机形式运行，管理整个系统的计算、存储和网络资源分配与调度

ZStack for Alibaba Cloud共享文件系统高可用方案，最重要的技术亮点是高可用服务，其中：

- NFS/SMP共享文件系统提供存储级别高可用能力
- zsha守护服务提供ZStack for Alibaba Cloud管理节点主机高可用
- ZStack for Alibaba Cloud管理节点主机提供业务云主机的高可用

三重高可用技术的保护能有效支撑云主机正常运行，下节将分别介绍其功能和作用。

1.1.2 高可用技术

NFS/SMP共享文件系统高可用

NFS/SMP共享文件系统可提供存储级别高可用能力，具体请参考相应存储厂商提供的产品资料。

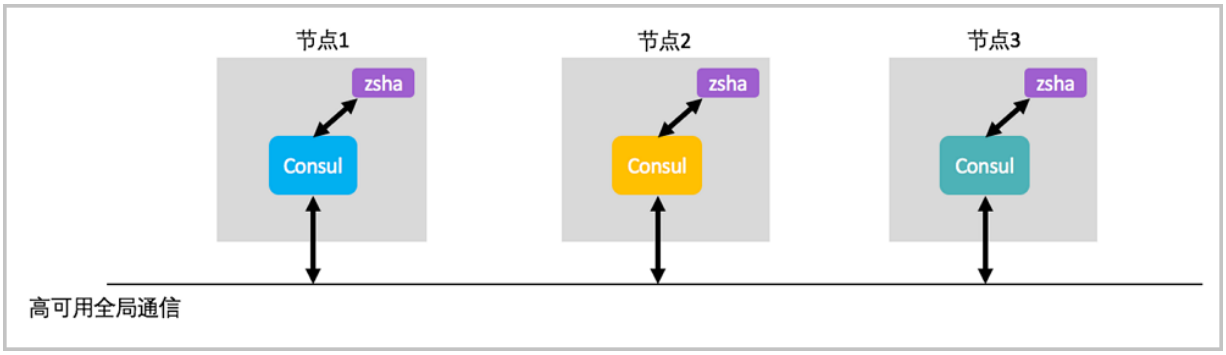
管理节点高可用

通过**zsha守护服务**提供强一致选举机制，保障ZStack for Alibaba Cloud管理节点主机的运行状态。

在三节点经典模型下，每个节点均运行**zsha和consul守护服务**，通过两两间建立**通信服务**（**TCP 8300**）维护强一致共享数据结构。

其运行通信原理如图 *1-2: zsha与consul服务消息通信原理*所示：

图 1-2: zsha与consul服务消息通信原理



业务云主机高可用

共享文件系统高可用方案，通过管理节点主机，对HA集群内其它节点进行**心跳检测**，在网络或节点失效情况下，将业务云主机**自动切换**到其他健康的节点运行，能快速恢复业务系统，提高业务系统的可用性。

1.1.3 网络拓扑规划

NFS/SMP共享文件系统高可用方案，支持以下网络流量模型：

- 云主机数据网络
- 管理网络
- 存储网络

云主机数据网络

云主机数据网络，即云主机向外提供应用服务的网络，或云主机之间相互沟通的网络。

该网络也承载云路由三层流量。根据业务负载类型，建议采用**双链路1GbE或10GbE以太网**。

说明：

- 若计划部署的云主机采用单一扁平网络，建议此网络的交换机端口设定Access模式，以下行文以此为主体描述；
- 若计划部署的云主机考虑多个扁平或云路由网络，建议此网络的交换机端口设定Trunk模式；以下行文并未提及该场景操作过程，如需了解，请联系官方技术支持团队。

管理网络

管理网络主要承载**管理节点主机**与**物理主机**的消息通信，包括任务下发和云主机迁移。

考虑到管理网络的业务负载需求，建议采用**双链路1GbE或10GbE以太网**。

存储网络

存储网络主要承载**NFS/SMP共享文件系统存储流量**和**zsha守护服务通信**的网络流量。该网络是保障存储集群、管理节点和业务云主机高可用的关键网络。

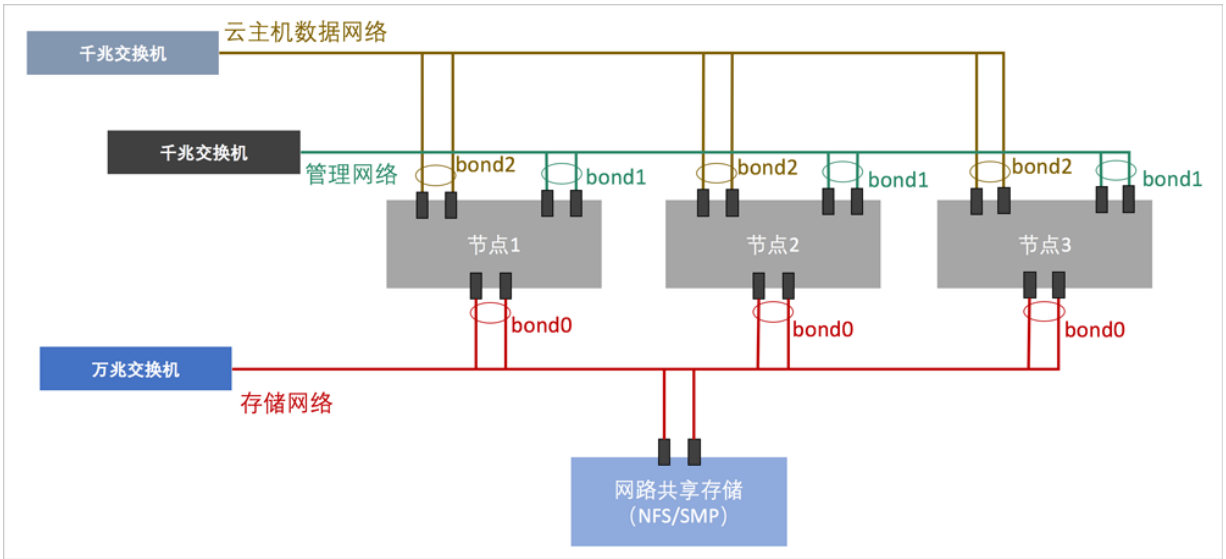
考虑到存储网络的业务负载需求，建议采用**双链路10GbE以太网**。

说明：

- 建议此网络的交换机端口设定Access模式（Trunk模式亦支持，但本文未描述）
- 查看交换机是否支持巨型帧（Jumbo Frame），若支持则建议开启，全链路MTU设定值为9000

综上所述，在三节点经典网络中，图 *1-3:* 基于*NFS/SMP*共享文件系统高可用的三节点网络拓扑如下：

图 1-3: 基于NFS/SMP共享文件系统高可用的三节点网络拓扑



管理员根据上述的网络架构图，对网络设备和服务器进行上架并连线。

表 *1-4:* 服务器节点的配置如下，管理员可根据业务性能需求，合理调配CPU、内存和硬盘的容量配比，也达到合适的平衡状态。

表 1-1: 服务器节点的配置

	配件	型号	数量	总数
服务器节点	CPU	Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz	2	3个
	内存	DDR4 16GB	8	
	主板	双路服务器标准主板	1	
	阵列卡	阵列卡支持SAS/SATA RAID 0/1/10 支持直通模式	1	
	固态硬盘	Intel SSD DC S3610 480GB	2	
	机械硬盘1	SAS HDD 300GB 3.5"，15k rpm	2	
	机械硬盘2	NL SAS HDD 2TB 3.5"，7.2k rpm	6	
	千兆网口	以太网1GbE，RJ45	4	
	万兆网口	以太网10GbE，SFP+	2	
	光电模块	-		

光纤HBA卡	-	
远程管理	DELL iDRAC企业版	1
电源	标准电源1100W	2

1.2 安装部署

本章节描述ZStack for Alibaba Cloud基于NFS/SMP共享文件系统高可用的安装部署过程。请管理员准备以下必要的软件包，以便安装部署过程顺利执行：

- **操作系统**
- **管理节点镜像**
- **高可用服务套件**
- **内嵌服务套件**

以上安装包最新版本请见这里

1.2.1 安装操作系统

操作步骤

1. 准备

   管理员对上架的网络设备和服务器加载电源，手动启动服务器进入BIOS，检查以下内容：

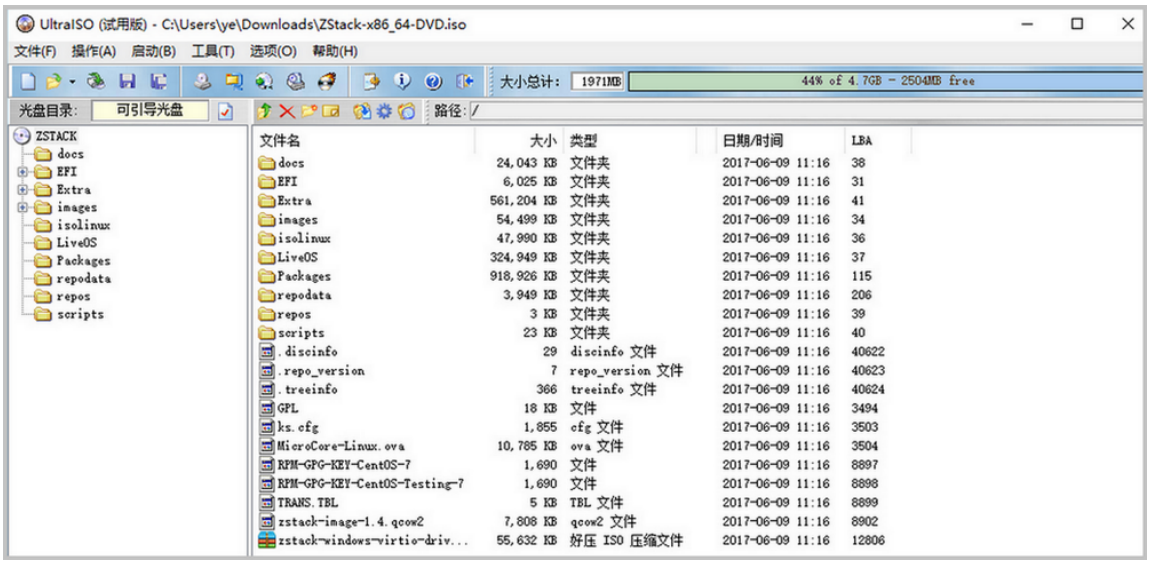
   - 激活所有CPU核心和启用超线程功能，设定系统性能为最高性能状态；
   - 打开硬件虚拟化VT功能，支持硬件虚拟化技术加速优化功能；
   - 进入阵列卡设定，对两块系统硬盘配置RAID1（Mirror），其余硬盘设定直通模式。

2. 在UltraISO打开ZStack for Alibaba Cloud DVD镜像

   - ZStack for Alibaba Cloud操作系统ISO镜像可通过DVD-RW设备刻录成安装光盘，也可通过UltraISO工具将把ISO文件刻录到U盘。
   - 打开UltraISO，点击**文件**按钮，选择打开已下载好的ISO文件。

     如图 *1-4: 在UltraISO打开DVD镜像*所示：

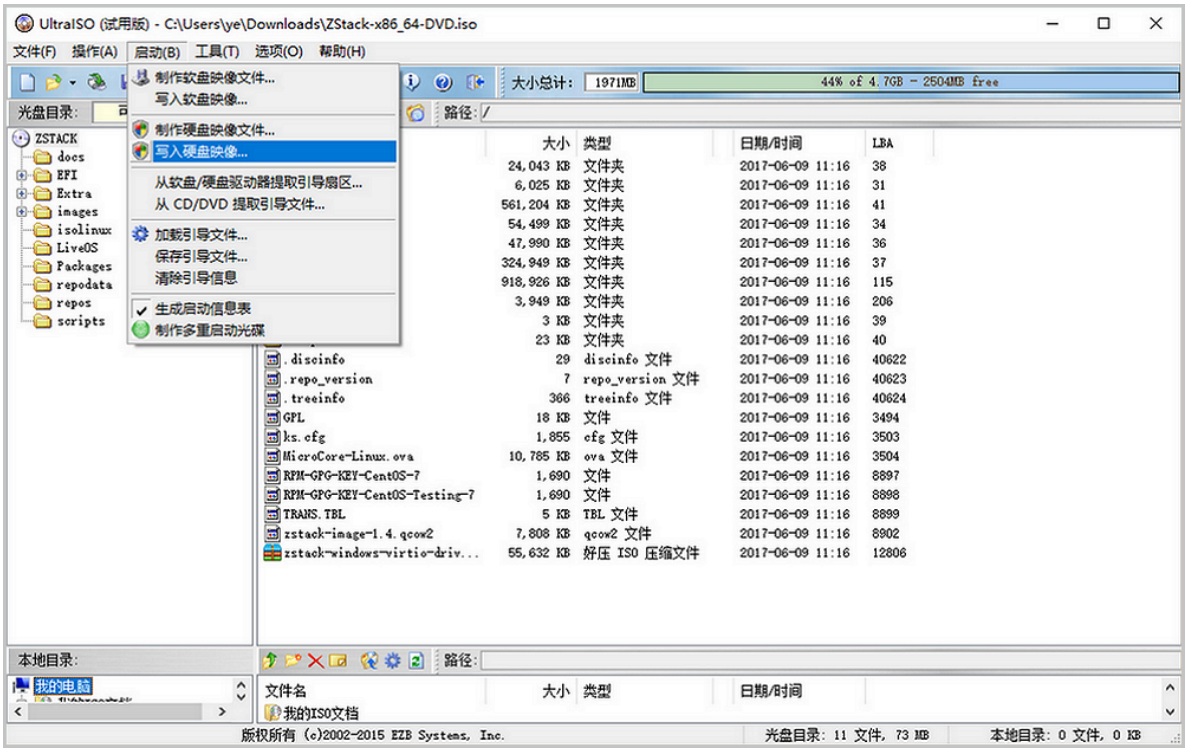
图 1-4: 在UltraISO打开DVD镜像



3. 写入硬盘镜像

在UltraISO点击**启动**按钮，选择**写入硬盘镜像**，如图 *1-5: 在UltraISO写入DVD镜像*所示：

图 1-5: 在UltraISO写入DVD镜像



4. 在UltraISO确认写入ZStack for Alibaba CloudDVD镜像

- 如果系统只插了一个U盘，则默认以此U盘进行刻录和写入，在刻录前，**注意备份U盘之前的内容。**
- 其他选项，按照默认设置，无须额外配置，点击**写入。**

如图 *1-6: 在UltraISO确认写入ISO镜像*所示：

图 1-6: 在UltraISO确认写入ISO镜像

5. 进入安装导航

ISO镜像已经刻录到U盘。此时U盘可用来作为启动盘，支持Legacy模式和UEFI模式引导。管理员通过安装介质，引导节点启动，并进入安装导航，如图 *1-7: U盘引导界面*所示：

图 1-7: U盘引导界面

6. 安装操作系统

   默认选择**Install ZStack for Alibaba Cloud**开始安装操作系统。

   在进入安装界面后，已经预先配置默认选项：

   - **时区**：亚洲东八区
   - **语言**：English(United States)
   - **键盘**：English(US)

   一般情况下管理员无需更改配置。管理员需自行执行硬盘的分区，推荐分区如下（UEFI 模式）：

   - */boot/efi*：创建分区500MB
   - */boot*：创建分区1GB
   - *swap*（交换分区）：创建分区32GB
   - */*（根分区）：配置剩下容量

   分区配置完后，选择**Software Selection**进入服务器安装角色候选，选择**ZStack for Alibaba Cloud Compute Node**计算节点模式，确定后回到主界面。

   点击**Begin Installation**进行安装。安装过程将会自动进行，管理员需要设定root账户密码。

   安装结束后，重新引导服务器并拔掉U盘。如安装成功，则服务器重启后进入操作系统登录提示符，使用root和设置的密码登录到操作系统。

说明：

管理员可根据自身需要更改密码。

1.2.2 存储网络

管理员对三节点安装ZStack for Alibaba Cloud操作系统后，可对**存储网络**进行如下设定：

节点	网卡 1	网卡 2	聚合接口	网桥	地址
节点1	p5p1	p5p2	bond0	br_bond0	*192.168.93.3/24*
节点2	p5p1	p5p2	bond0	br_bond0	*192.168.93.4/24*
节点3	p5p1	p5p2	bond0	br_bond0	*192.168.93.5/24*

以上是示例数据，管理员请根据具体部署环境的网卡和网络地址进行配置。以**节点 1**为例，其他节点配置相似。执行配置命令：

```
# 创建聚合网卡bond0
[root@localhost ~]# zs-bond-lacp -c bond0
# 将网卡p5p1与p5p2均添加到bond0
[root@localhost ~]# zs-nic-to-bond -a bond0 p5p1
[root@localhost ~]# zs-nic-to-bond -a bond0 p5p2
# 配置上述链路聚合后,请管理员在对应的交换机网口配置LACP聚合
# 创建网桥br_bond0,指定网络IP、掩码和网关
[root@localhost ~]# zs-network-setting -b bond0 192.168.93.3 255.255.255.0 192.168.93.1
# 查看聚合端口bond0是否创建成功
[root@localhost ~]# zs-show-network
...
---------------------------------------------------------------------------
| Bond Name  | SLAVE(s)       | BONDING_OPTS                              |
---------------------------------------------------------------------------
| bond0      | p5p1           | miimon=100 mode=4 xmit_hash_policy=layer2+3       |
|            | p5p2           |                                           |
---------------------------------------------------------------------------
```

说明：

- p5p1和p5p2加载到bond0后，对应交换机的端口需要配置LACP聚合，否则网络通信将异常；如果交换机不支持LACP聚合，请联系网络设备厂商更换设备；
- 通过bond0创建网桥后，网桥命名为**br_bond0**，将提供Ceph存储集群通信和zsha守护服务通信；
- 关于网桥的IP地址、子网掩码和网关参数，用户需按照实际情况填写；

- 存储网络配置完成后，可通过**ping**命令进行检测；若配置正确，则三节点的存储网络对应的IP地址可两两互**ping**。

存储网络配置完成后，随之可配置管理网络。

1.2.3 管理网络

管理员对三节点安装ZStack for Alibaba Cloud操作系统后，可对**管理网络**进行如下设定：

节点	网卡 1	网卡 2	聚合接口	网桥	地址
节点1	eth0	eth1	bond1	br_bond1	*172.20.198.3/24*
节点2	eth0	eth1	bond1	br_bond1	*172.20.198.4/24*
节点3	eth0	eth1	bond1	br_bond1	*172.20.198.5/24*

以上是示例数据，管理员请根据具体部署环境的网卡和网络地址进行配置。以**节点 1**为例，其他节点配置相似。执行配置命令：

```
# 创建聚合网卡bond1
[root@localhost ~]# zs-bond-lacp -c bond1
# 将网卡eth0与eth1均添加到bond1
[root@localhost ~]# zs-nic-to-bond -a bond1 eth0
[root@localhost ~]# zs-nic-to-bond -a bond1 eth1
# 配置上述链路聚合后,请管理员在对应的交换机网口配置LACP聚合
# 创建网桥br_bond1,指定网络IP、掩码和网关
[root@localhost ~]# zs-network-setting -b bond1 172.20.198.3 255.255.255.0 172.20.198.1
# 查看聚合端口bond1是否创建成功
[root@localhost ~]# zs-show-network
...
----------------------------------------------------------------------------
| Bond Name  | SLAVE(s)       | BONDING_OPTS                                |
----------------------------------------------------------------------------
| bond1      | eth0           | miimon=100 mode=4 xmit_hash_policy=layer2+3       |
|            | eth1           |                                             |
----------------------------------------------------------------------------
```

说明：

- eth0和eth1加载到bond1后，对应交换机的端口需要配置LACP聚合，否则网络通信将异常：如果交换机不支持 LACP 聚合，请联系网络设备厂商更换设备；
- 通过bond1创建网桥后，网桥命名为br_bond1，将提供管理节点主机与物理主机的消息通信；
- 关于网桥的IP地址、子网掩码和网关参数，用户需按照实际情况填写；

- 管理网络配置完成后，可通过**ping**命令进行检测；若配置正确，则三节点的管理网络对应的IP地址可两两互**ping**。

存储网络和管理网络配置完成后，随之可配置云主机数据网络。

1.2.4 云主机数据网络

管理员对三节点安装ZStack for Alibaba Cloud操作系统后，可对**云主机数据网络**进行如下设定：

节点	网卡 1	网卡 2	聚合接口	网桥	地址
节点1	em1	em2	bond2	-	-
节点2	em1	em2	bond2	-	-
节点3	em1	em2	bond2	-	-

以上是示例数据，管理员请根据具体部署环境的网卡和网络地址进行配置。以**节点1**为例，其他节点配置相似。执行配置命令：

```
# 创建聚合网卡bond2
[root@localhost ~]# zs-bond-lacp -c bond2
# 将网卡em1与em2均添加到bond2
[root@localhost ~]# zs-nic-to-bond -a bond2 em1
[root@localhost ~]# zs-nic-to-bond -a bond2 em2
# 配置上述链路聚合后,请管理员在对应的交换机网口配置LACP聚合
# 云主机数据网络,无需创建网桥
# 查看聚合端口bond2是否创建成功
[root@localhost ~]# zs-show-network
...
--------------------------------------------------------------------------------
| Bond Name  | SLAVE(s)       | BONDING_OPTS                                  |
--------------------------------------------------------------------------------
| bond2      | em1            | miimon=100 mode=4 xmit_hash_policy=layer2+3   |
|            | em2            |                                               |
--------------------------------------------------------------------------------
```

说明：

em1和em2加载到bond2后，对应交换机的端口需要配置LACP聚合，否则网络通信将异常；如果交换机不支持LACP聚合，请联系网络设备厂商更换设备。

云主机数据网络配置完成后，随之可安装NFS/SMP网络共享存储。

1.2.5 网络共享存储（NFS/SMP）

建议用户在部署NFS/SMP共享存储服务之前，先查看当前存储服务器系统的内核版本，同一个NFS/SMP软件在不同版本，内核之间是有差异的，部署方法也随之不同。具体请参考相应存储厂商提供的产品资料。

ZStack for Alibaba Cloud要求NFS/SMP服务器端提供以下三个共享目录：

- *your_NFS/SMP_server_IP:/share/mngt*

  用于存放管理节点镜像

- *your_NFS/SMP_server_IP:/share/bs*

  用于存放镜像服务器模板

- *your_NFS/SMP_server_IP:/share/ps*

  主存储，用于存放云主机数据

三节点客户端均需挂载上述三个共享目录：

```
# 分别登录三节点，执行以下命令，将共享目录your_NFS/SMP_server_IP:/share/mngt分别挂载到三节点本地目录/opt/smp/mngt/
[root@localhost ~]# mkdir -p /opt/smp/mngt
[root@localhost ~]# mount -t nfs your_NFS/SMP_server_IP:/share/mngt /opt/smp/mngt/

# 分别登录三节点，执行以下命令，将共享目录your_NFS/SMP_server_IP:/share/bs分别挂载到三节点本地目录/opt/smp/bs/
[root@localhost ~]# mkdir -p /opt/smp/bs
[root@localhost ~]# mount -t nfs your_NFS/SMP_server_IP:/share/bs /opt/smp/bs/
```

说明：

- 对于NFS，主存储共享目录*your_NFS_server_IP:/share/ps*会自动挂载到三节点本地目录（该本地目录由ZStack for Alibaba Cloud生成）
- 对于SMP，主存储共享目录*your_SMP_server_IP:/share/ps*需要手动挂载到三节点本地目录*/opt/smp/ps/*

```
# 分别登录三节点，执行以下命令，将共享目录your_SMP_server_IP:/share/ps分别挂载到三节点本地目录/opt/smp/ps/
[root@localhost ~]# mkdir -p /opt/smp/ps
[root@localhost ~]# mount -t nfs your_SMP_server_IP:/share/bs /opt/smp/ps/
```

设置开机自动挂载：

```
# 将mount命令写在rc.local文件中
[root@localhost ~]# chmod +x /etc/rc.local
[root@localhost ~]# vim /etc/rc.local
```

```
# 以NFS为例：

...

touch /var/lock/subsys/local

mount -t nfs 172.20.14.112:/share/mngt /opt/smp/mngt/
mount -t nfs 172.20.14.112:/share/bs /opt/smp/bs/

...
```

NFS/SMP共享文件系统挂载到三节点后，随之可通过共享存储快速部署高可用。

1.2.6 高可用套件

1.2.6.1 功能介绍

ZStack for Alibaba Cloud基于网络共享存储场景提供了高可用套件，保障ZStack for Alibaba Cloud管理服务长期处于在线运行，与业务云主机高可用机制相结合，最大程度降低物理设备异常情况下的云主机失效带来的业务中断风险。

高可用套件，是基于Consul开发的强一致性消息通信组件，部署在高可用节点集群中，部署数量满足2N+1条件。在三节点经典模型中，能容忍1个节点的离线或失效，保证ZStack for Alibaba Cloud管理服务正常运行。

在部署高可用之前，确保每个节点均安装 libvirt 和 qemu-kvm-ev 软件包：

```
# 每个节点均安装libvirt和qemu-kvm-ev软件包
[root@localhost ~]# yum -y --disablerepo=* --enablerepo=zstack-local,qemu-kvm-ev \
install libvirt qemu-kvm-ev qemu-img-ev
```

1.2.6.2 部署过程

操作步骤

1. 导入镜像

   管理员已经获得ZStack for Alibaba Cloud管理节点镜像(QCOW2 格式)，通过共享目录，将镜像快速传输到三节点：

   ```
   # 登录节点1，将镜像拷贝到本地目录/opt/smp/mngt/，并命名为mnvm.img
   [root@localhost ~]# cp ZStack for Alibaba Cloud-Management-Server-2.x.qcow2 /opt/smp/
   mngt/mnvm.img
   # 拷贝完成后，由于共享目录机制，可在三节点本地目录/opt/smp/mngt/均查看到该镜
   像mnvm.img
   ```

2. 编写配置

管理员完成ZStack for Alibaba Cloud管理服务的镜像导入后，接下来编辑ZStack for Alibaba Cloud高可用套件的初始化配置文件。

执行以下命令，生成并编写配置文件：

```
# 生成的样本配置文件临时保存在/tmp/sample.FileConf.config.json,用户可自行拷贝到其它目录
[root@localhost ~]# chmod +x ZStack for Alibaba Cloud-HA-Suite-2.x.bin
[root@localhost ~]# ./ZStack for Alibaba Cloud-HA-Suite-2.x.bin config-sample FileConf
# 编辑样本配置文件
[root@localhost ~]# cp /tmp/sample.FileConf.config.json /tmp/config.json
[root@localhost ~]# vim /tmp/config.json
...
{
  "Node": [
    "192.168.255.205 nfs-1",
    "192.168.255.221 nfs-2",
    "192.168.255.183 nfs-3"
  ], //请修改三节点的IP地址以及主机名称
  "MemorySizeInGB": 8, //管理节点主机的内存配置,最小值为8GB
  "CPU": 4, //管理节点主机的CPU数目,最小值为4核
  "Type": "FileConf", //配置文件的类型
  "MonAddrs": null, //无需设置监控节点
  "ChronyServers": [
    "192.168.255.205"
  ], //请根据实际情况设置chrony时间服务器IP地址
  "PoolName": "", //管理节点主机所在的存储池名
  "ImageFolder": "/opt/smp/mngt/", //管理节点镜像存放的本地目录
  "DNS": [
    "172.20.198.1", //管理节点主机的私网DNS信息
    "223.5.5.5" //管理节点主机的公网DNS信息
  ],
  "Network": [
    {
      "Bridge": "br_zsn0", //管理节点主机的网卡所挂接的网桥名称
      "MacAddress": "02:98:54:4b:a5:c0", //管理节点主机的网卡MAC地址,单次随机生成
      "Ipaddr": "172.20.198.3@nfs-1, 172.20.200.3@nfs-2, 172.20.210.3@nfs-3",  //管理节点主机的IP地址
      "Netmask": "255.255.255.0@nfs-1, 255.255.255.0@nfs-2, 255.255.255.0@nfs-3", //管理节点主机的掩码
      "Gateway": "172.20.198.1@nfs-1, 172.20.200.1@nfs-2, 172.20.210.1@nfs-3", //管理节点主机的网关
      //管理节点主机的IP地址以及对应掩码和网关支持per-node格式,可实现管理节点主机跨网段启动
      "IsMgmt": true, //管理网络设置，本示例设为true
      "IsDefRoute": false //默认路由配置，本示例设为false
    },
    {
      "Bridge": "br_zsn1", //管理节点主机的网卡所挂接的网桥名称
      "MacAddress": "f6:75:d8:5d:73:73", //管理节点主机的网卡MAC地址,单次随机生成
      "Ipaddr": "203.114.54.10", //管理节点主机的IP地址
      "Netmask": "255.255.255.0", //管理节点主机的掩码
      "Gateway": "203.114.54.1", //管理节点主机的网关
      "IsMgmt": false, //管理网络设置，本示例设为false
      "IsDefRoute": true //默认路由配置，本示例设为true
    }
```

```
  ],
  "ConsoleProxyOverriddenIP": "" //ZStack管理服务的控制台服务地址,支持per-node格式
  "AllowResetRootPassword": false //默认为false;设为true时,管理节点主机才允许重置root密
码
}
...
```

管理员需要按照具体部署场景，修改上述参数。

说明：

- CPU和MemorySizeInGB设置值推荐如下：

场景	配置	备注
中小规模	CPU核4个/内存8GB	管理小于100个物理主机与1000个云主机
特大规模	CPU核8个/内存16GB	• 管理大于100个物理主机与1000 个云主机内 • 管理小于1000个物理主机与10000个云主机
超大规模	CPU核12个/内存32GB	管理大于1000个物理主机与10000个云主机

- 关于管理节点主机的网卡MAC地址，需确保在二层网络里是唯一的。
- MonAddrs设置为默认null，无需设置监控节点。
- 关于ChronyServers的设置：
  - ChronyServers可使用三存储节点作为时间源，示例如下：

```
"ChronyServers": [
    "192.168.255.205"
  ],
```

  - 管理员也可自行设置chrony时间源，示例如下：

```
"ChronyServers": [
    "time1-7.aliyun.com "
  ],
```

  - ChronyServers中指定的时间服务器不可对自身做时间同步。

说明：

补充说明，对于未部署管理节点高可用服务的场景，默认情况下管理节点作为NTP时间源；若希望指定时间服务器，方法如下：

- 在管理节点的**zstack.properties**文件中添加：

```
chrony.serverIp.0 = xx.xx.xx.xx
# xx.xx.xx.xx为时间服务器IP地址,可为管理节点IP地址,也可自行指定其它chrony时间源
```

- 重连物理主机生效。
- chrony.serverIp.0中指定的时间服务器不可对自身做时间同步。

- 管理节点主机跨网段启动

  在**IsMgmt**为**true**情况下，管理节点的IP地址、以及对应的掩码、网关均设置为**per-node**格式，可实现管理节点主机跨网段启动，示例如下：

```
"Network": [
  {
    ...
    "Ipaddr": "172.20.198.3@nfs-1, 172.20.200.3@nfs-2, 172.20.210.3@nfs-3",
    "Netmask": "255.255.255.0@nfs-1, 255.255.255.0@nfs-2, 255.255.255.0@nfs-3",
    "Gateway": "172.20.198.1@nfs-1, 172.20.200.1@nfs-2, 172.20.210.1@nfs-3",
    "IsMgmt": true,
    ...
  },
```

  ConsoleProxyOverriddenIP也支持**per-node**格式，当管理节点主机跨网段启动后，可打开相应管理服务控制台。

- 管理节点主机允许重置root密码为系统默认值（即**password**）

  在**AllowResetRootPassword**为**true**情况下，管理节点主机允许重置root密码为系统默认值（即**password**），详情请参考配置更新章节。

3. HA初始化

在上一节中,管理员已完成对config.json配置文件的编写，接下来可执行高可用套件的初始化。执行以下命令：

```
# 为三个节点安装高可用套件
[root@localhost ~]# ./ZStack for Alibaba Cloud-HA-Suite-2.x.bin install \
-p password -c /tmp/config.json
```

说明：

- 安装命令中，**-p** 传递三个节点账号root密码。建议首次初始化时，对三个节点设置一致密码，安装完成后，基于安全考虑可按需修改或关闭密码登陆。
- */tmp/config.json*为填写完成的配置文件路径，需与上述步骤中真实保存的文件绝对路径保持一致。

高可用套件初始化完成后，可执行以下命令查看集群状态:

```
# 查看 HA 集群状态
[root@nfs-1 ~]# zsha status
>>>MN-VM Running On
192.168.255.205 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
192.168.255.205 nfs-1
192.168.255.221 nfs-2
192.168.255.183 nfs-3

>>>Consul Members
Node           Address              Status  Type    Build  Protocol  DC
192.168.255.205  192.168.255.205:8301  alive   server  0.7.2  2         dc1
192.168.255.221  192.168.255.221:8301  alive   server  0.7.2  2         dc1
192.168.255.183  192.168.255.183:8301  alive   server  0.7.2  2         dc1


>>>Management Node IP
172.20.198.3

>>>Public Network IP
203.114.54.10


>>>Checking Host I/O
done.
```

至此，ZStack for Alibaba Cloud基于NFS/SMP共享存储的高可用云管理平台安装完成。

当所有节点正常加入并开始运行高可用服务程序后，将选择其中一个节点运行管理节点主机，高可用保护机制开始正常运转，ZStack for Alibaba Cloud管理节点也开始正常提供服务。

4. 云平台初始化

管理员可通过浏览器访问 ZStack for Alibaba Cloud管理服务界面（http://manage-server-ip:5000），并完成云平台初始化操作。如登陆界面所示：

图 1-8: 登录界面

初始化的操作说明，具体参照*ZStack*官网文档《用户手册》。

- 初始化过程中，添加镜像服务器时，URL请输入：*your_NFS/SMP_server_IP:/share/bs*
- 初始化过程中，添加主存储时，URL请输入：*your_NFS/SMP_server_IP:/share/ps*

1.2.7 集群升级

1.2.7.1 内嵌服务升级

管理员获得新的内嵌服务套件后，可用于升级当前高可用环境的ZStack for Alibaba Cloud管理节点主机。

请管理员将新的内嵌服务套件上传到ZStack for Alibaba Cloud管理节点主机的目录*/root/*。上传完成后，登陆ZStack for Alibaba Cloud管理节点主机，执行命令：

```
# 登陆ZStack for Alibaba Cloud管理节点主机,执行内嵌服务升级操作
[root@managementnode ~]# chmod +x ZStack for Alibaba Cloud-HA-Guest-Agent-2.x.bin
[root@managementnode ~]# ./ZStack for Alibaba Cloud-HA-Guest-Agent-2.x.bin -i
# 升级完成后,检查版本信息
[root@managementnode ~]# zin -v
App Version: 1.5.6.0
Build Date: Fri Jul 28 10:37:49 GMT-8 2017
Branch Name: (detached
```

```
Commit ID: c7dc57a012f5bae01eafaea7455bf5784007c172
```

1.2.7.2 高可用升级

管理员获得新的高可用管理套件后，可用于升级当前的高可用集群环境。请管理员登陆其中一个节点执行命令：

```
# 导出当前的高可用集群配置信息
[root@localhost ~]# zsha export-config
output config json file to /tmp/current.config.json

# 若新版本升级时，需要增加/删除相关配置信息，请依照新版本的编写配置说明进行修改，再执行zsha export-config进行导出

# 关闭集群内管理节点主机以及所有zsha服务
[root@localhost ~]# zsha stop

# 使用新的高可用套件升级管理三个节点
[root@localhost ~]# chmod +x ZStack for Alibaba Cloud-HA-Suite-2.x.bin
[root@localhost ~]# ./ZStack for Alibaba Cloud-HA-Suite-2.x.bin install \
-p password -c /tmp/current.config.json
```

1.2.8 管理节点迁移

ZStack for Alibaba Cloud高可用集群在正常运行环境下，若管理员需要临时关闭一个节点进行维护，需在ZStack for Alibaba Cloud管理界面设定物理主机为维护模式。

然后，根据该节点是否运行ZStack for Alibaba Cloud管理节点主机，进行对应操作。

- 若该节点并非运行ZStack for Alibaba Cloud管理节点主机，则可以执行关机等维护操作。
- 若该节点运行ZStack for Alibaba Cloud管理节点主机，则需要执行手动热迁移操作。

```
# 登陆到其中一个节点,执行对ZStack for Alibaba Cloud管理节点主机迁移至目标节点
[root@localhost ~]# zsha migrate target ip

# 用户也可直接填写目标节点的主机名称执行迁移操作
[root@localhost ~]# zsha migrate target hostname
```

说明：

- 其中*target ip*为高可用集群中一个正常节点的IP地址。该地址必须是高可用集群内的IP地址，否则无法通过检查，将会报错并且不能执行迁移操作。
- 同样的，其中**target hostname**必须为高可用集群中一个正常节点的主机名称。

- 如果目标节点指定为正在运行管理节点主机的HA节点，将会报错，报错格式如下：

  ```
  错误:Requested operation is not valid: domain 'ZStack Management Node VM' is already active
  ```

- 迁移过程需消耗一定时间，迁移过程将会显示进度状态。如果网络为万兆网络且状况良好，一般可在数秒内完成。

1.2.9 配置更新

管理节点高可用机制将全局配置信息存储在Consul KV数据库中，如果需要修改诸如管理节点主机配置、管理节点主机网络配置时，执行以下命令即可：

```
# 导出当前的高可用集群配置信息
[root@localhost ~]# zsha export-config
# 修改导出文件的配置
# 更新三个节点的配置信息
[root@localhost ~]# zsha import-config /tmp/current.config.json
# 配置更新后,请登录管理节点主机执行关机操作,高可用服务将会触发启动
[root@managementnode ~]# halt -p
```

说明：

- 修改后的*/tmp/current.config.json*格式，应与原配置文件格式相同。
- 如需修改节点的IP地址，请按照高可用套件的安装过程执行。
- MemorySizeinGB和CPU管理节点主机性能配置可以按需修改，但内存不可低于8G，CPU不可低于4核心。
- Bridge为三个节点提供给管理节点主机使用的网桥名称。只有当该网桥名称发生变化时，才应修改此参数。
- NetworkInterface为管理节点主机内部的网络配置信息，可以按需修改。
- 由于管理节点镜像封装了qemu-guest-agent服务，管理节点主机支持修改root密码。若管理员忘记当前root密码，可通过执行reset-MNVM-password命令重置root密码为系统默认值（即**password**），但前提是**AllowResetRootPassword**需设置为**true**。

  - 假定管理员使用原root密码登录管理节点主机，希望修改root密码：

    ```
    [root@managementnode ~]# passwd root
    Changing password for user root.
    New password:
    Retype new password:
    ```

```
passwd: all authentication tokens updated successfully.
```

- 假定管理员忘记当前root密码，希望重置root密码为系统默认值（即**password**）：

```
# 登录到其中一个节点，导出当前的高可用集群配置信息
[root@localhost ~]# zsha export-config
output config json file to /tmp/current.config.json

# 修改导出文件的配置,将AllowResetRootPassword设置为true
[root@localhost ~]# vim /tmp/current.config.json
...
{
  ...
  "AllowResetRootPassword": true //默认为false;设为true时,管理节点主机才允许重置root密码
}
...

# 更新三个节点的配置信息,按照高可用套件的安装过程执行
[root@localhost ~]# zsha install -p password -c /tmp/current.config.json

# 执行以下命令,重置root密码为系统默认值（即password）
[root@localhost ~]# zsha reset-MNVM-password
Reset MNVM root password on nodex IP //nodex IP为MNVM所在节点的IP
Password set successfully for root in 173de36b-4a5d-458e-a5e3-9a99584551ef
```

1.3 其他操作

1.3.1 卸载操作

管理员若需卸载ZStack for Alibaba Cloud高可用套件，可执行以下命令：

```
# 登陆到其中一个节点,执行关闭高可用套件服务并停止ZStack for Alibaba Cloud管理节点主机
[root@localhost ~]# zsha stop
# 确认管理节点主机已关闭,避免导致数据损坏或丢失
[root@localhost ~]# zsha status
# 登陆到三个节点后,执行卸载高可用相关服务
[root@localhost ~]# zsha uninstall
```

执行卸载后，ZStack for Alibaba Cloud管理节点主机数据仍然保留在NFS/SMP共享文件系统里。管理员可保留数据，亦可删除该镜像。

若需删除镜像数据和存储，可参考导入镜像章节关于删除数据的内容。

1.3.2 日志输出

ZStack for Alibaba Cloud高可用管理套件部署后，其运行状态日志信息存放在目录*/var/log/z/*。

高可用套件守护服务在运行状态时，会产生详细的日志内容,并进行按天和容量进行切割。管理员需关注ZStack for Alibaba Cloud管理节点主机的系统日志分区（*/var/log/*）空间状况，保证预留一定量的存储空间。根据运行估算，每年产生日志10+GB。

2 高可用测试与恢复

2.1 计划运维

2.1.1 单节点需要维护

2.1.1.1 单节点关闭

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。在正常运行环境下，若管理员需要临时关闭某个节点进行维护。

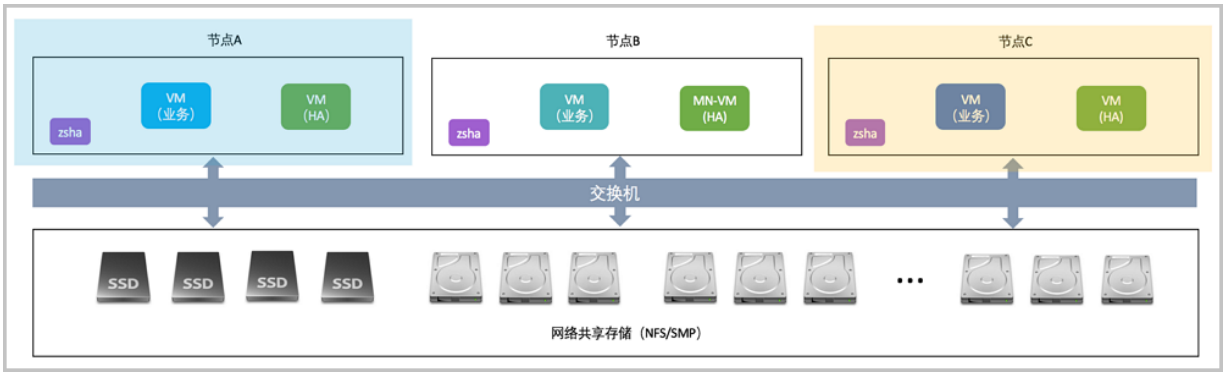
- 该节点运行非管理节点主机。
- 该节点运行管理节点主机。

2.1.1.1.1 该节点运行非管理节点主机

背景信息

需要关闭节点A或C进行维护，如图 *2-1: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图*所示：

图 2-1: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图



以关闭节点A为例。

操作步骤

1. 进入物理主机A的维护模式。

   通过账户admin登录到ZStack for Alibaba Cloud云管理平台，将节点A的物理主机进入维护模式。

   如图 *2-2: 维护物理主机A*所示：

图 2-2: 维护物理主机A

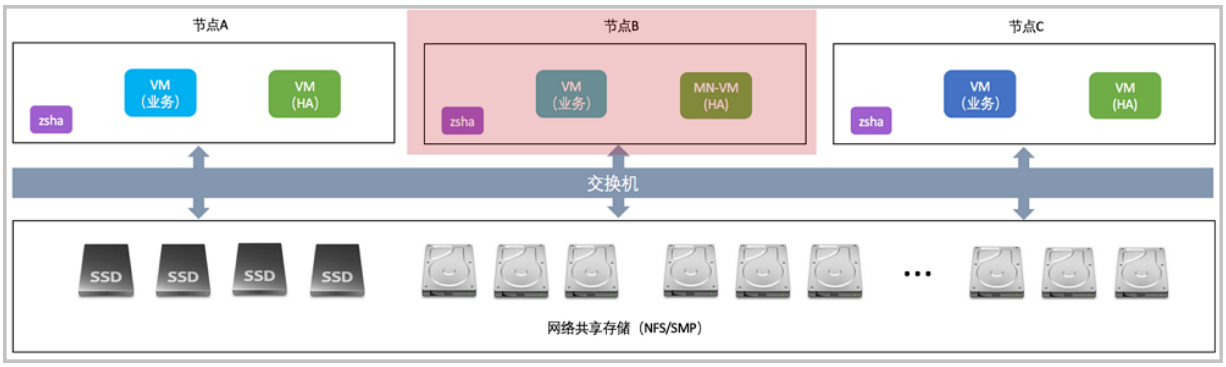
2. 对节点A进行**shutdown**关机操作。
3. 对节点A下电后进行维护。

2.1.1.1.2 该节点运行管理节点主机

背景信息

需要关闭节点B进行维护，如图 *2-3: 基于NFS/SMP*共享文件系统的高可用集群 经典三节点示例图所示：

图 2-3: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图



操作步骤

1. 管理员进行手动操作，将管理节点迁移至其他健康节点A或C。

   将ZStack for Alibaba Cloud管理节点主机迁移至目标节点。

```
# 登陆管理节点主机,执行对管理节点主机迁移至目标节点
[root@managementnode ~]# zsha migrate target ip

# 用户也可直接填写目标节点的主机名称执行迁移操作
[root@managementnode ~]# zsha migrate target hostname
```

2. 进入物理主机B的维护模式。

通过账户admin登录到ZStack for Alibaba Cloud云管理平台，将节点B物理主机进入维护模式。

如图 *2-4: 维护物理主机B*所示：

图 2-4: 维护物理主机B

3. 对接点B进行**shutdown**关机操作。
4. 对节点B下电后进行维护。

2.1.1.2 单节点启动

背景信息

ZStack for Alibaba Cloud集群由三个节点A、B、C构建。若某个节点临时关闭维护完成后，需要启动该节点。

操作步骤

1. 未启动的该节点通电后，通过手动或IPMI启动服务器。
2. 等待节点启动，成功引导操作系统。
3. 设置物理节点为启用状态。

   通过账户admin登录到ZStack for Alibaba Cloud云管理平台，设置该节点为启用状态，如图 *2-5: 启用物理主机*所示：

   图 2-5: 启用物理主机

2.1.2 三节点需要维护

2.1.2.1 三节点关闭

背景信息

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。若管理员需要临时关闭整个集群进行维护。

操作步骤

1. 关闭云主机高可用。

   通过账户admin登录到ZStack for Alibaba Cloud云管理平台，进入**设置** > **全局设置**，将**云主机高可用全局开关**的值设置为**false**，停止高可用心跳检测，如图 *2-6:* 关闭云主机高可用所示：

   图 2-6: 关闭云主机高可用

2. 将三台物理主机设置为维护模式。

   将三台物理主机均设置为维护模式，批量停止云主机，如图 *2-7:* 维护物理主机集群所示：

图 2-7: 维护物理主机集群

3. 通过ssh工具登陆到其中一个节点，执行**zsha stop**命令：

   登录到其中一个节点，执行关闭高可用套件服务并停止管理节点主机。

4. 对三个节点进行shutdown关机操作。

5. 三个节点下电后，维护整个集群。

2.1.2.2 三节点启动

背景信息

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。若整个集群临时关闭维护完成后，需要启动该集群。

操作步骤

1. 三个节点通电后，通过手动或IPMI启动服务器。

2. 操作系统启动引导完成，即可通过ssh工具登陆到三个节点。

3. 执行**zsha start**命令启动高可用服务套件。

   登录到其中一个节点，执行启动高可用套件服务并启动管理节点主机。

4. 检查运行状态。

   通过**zsha status**检查状态是否健康 并关注管理节点主机是否正常运行。

```
[root@localhost ~]# zsha status [root@localhost ~]# zsha status
>>>MN-VM Running On
172.20.14.215 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
172.20.13.239 : nfs-1
172.20.14.215 : nfs-2
```

```
172.20.14.251 : nfs-3
```

5. 设置节点为启用状态。

   通过账户admin登录到ZStack for Alibaba Cloud云管理平台，依次设置三个节点为启用状态，如图 *2-8:* 启用物理主机所示：

   图 2-8: 启用物理主机

6. 开启云主机高可用。

   通过账户admin登录到ZStack for Alibaba Cloud云管理平台，进入**设置** > **全局设置**，将**云主机高可用全局开关**的值设置为**true**，停止高可用心跳检测，如图 *2-9:* 开启云主机高可用所示：

   图 2-9: 开启云主机高可用

说明：

- 高可用级别为**NeverStop**的云主机将会自动恢复。

- 高可用级别为**None**的云主机需管理员手动启动，或通知云主机所有者执行启动。

2.2 异常处理

如果节点出现异常掉电/断网，请参考本章异常处理进行故障恢复。如果节点出现损坏（系统崩溃、磁盘损坏、数据丢失），请参考下章节点修复进行故障修复，如果依然无法解决问题，请联系官方技术支持团队。

2.2.1 单节点异常处理

2.2.1.1 单节点异常掉电

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。在正常运行环境下，若其中某个节点突然断电。

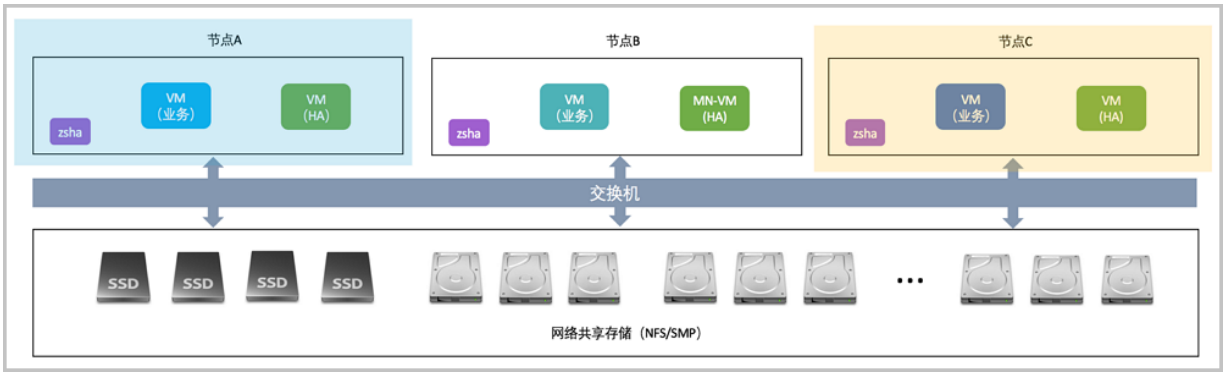
- 该节点运行非管理节点主机。
- 该节点运行管理节点主机。

2.2.1.1.1 该节点运行非管理节点主机

背景信息

节点A或C异常掉电需恢复，如图 *2-10: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图*所示：

图 2-10: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图



操作步骤

1. 该异常节点A上电后，通过手动或IPMI方式启动服务器。
2. 自动引导加载操作系统，确认是否启动成功。
3. 检查异常节点高可用状态。

执行**zsha status**命令，检查该异常节点的高可用服务是否正常。

```
[root@localhost ~]# zsha status
>>>MN-VM Running On
172.20.14.215 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
172.20.13.239 : nfs-1
172.20.14.215 : nfs-2
172.20.14.251 : nfs-3
```

4. 检查物理主机连接状态。

通过账号admin登陆ZStack for Alibaba Cloud云管理平台，检查该异常物理主机连接状态是否正常，如图 *2-11:* 物理主机连接状态界面所示：

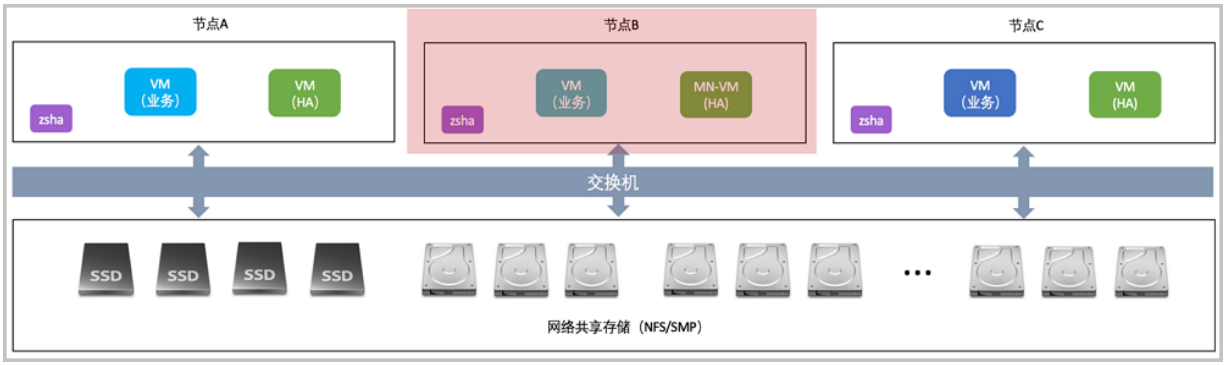
图 2-11: 物理主机连接状态界面

2.2.1.1.2 该节点运行管理节点主机

背景信息

节点B异常掉电需恢复，如图 *2-12:* 基于*NFS/SMP*共享文件系统的高可用集群 经典三节点示例图所示：

图 2-12: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图



操作步骤

1. 检查管理节点主机高可用是否已触发。

   通过ssh工具登陆到正在运行的节点A或C，执行**zsha status**命令检查管理节点主机是否正常运行，若正常运行，代表高可用套件已触发高可用切换。

```
[root@localhost ~]# zsha status
>>>MN-VM Running On
172.20.14.251 : running

>>>Last MN-VM Start Record
```

2. 该异常节点B上电后，通过手动或IPMI方式启动服务器。
3. 自动引导加载操作系统，确认是否启动成功。
4. 检查异常节点高可用状态。

   执行**zsha status**命令，检查该异常节点的高可用服务是否正常。

```
[root@localhost ~]# zsha status
>>>MN-VM Running On
172.20.14.251 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
172.20.13.239 : nfs-1
172.20.14.215 : nfs-2
172.20.14.251 : nfs-3
```

5. 检查物理主机连接状态。

   通过账号admin登陆ZStack for Alibaba Cloud云管理平台，检查该异常物理主机连接状态是否正常，如图 *2-13: 物理主机连接状态界面*所示：

图 2-13: 物理主机连接状态界面

2.2.1.2 单节点网络异常

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。在正常运行环境下，若其中某个节点网络异常。

- 该节点运行非管理节点主机。
- 该节点运行管理节点主机。

说明：

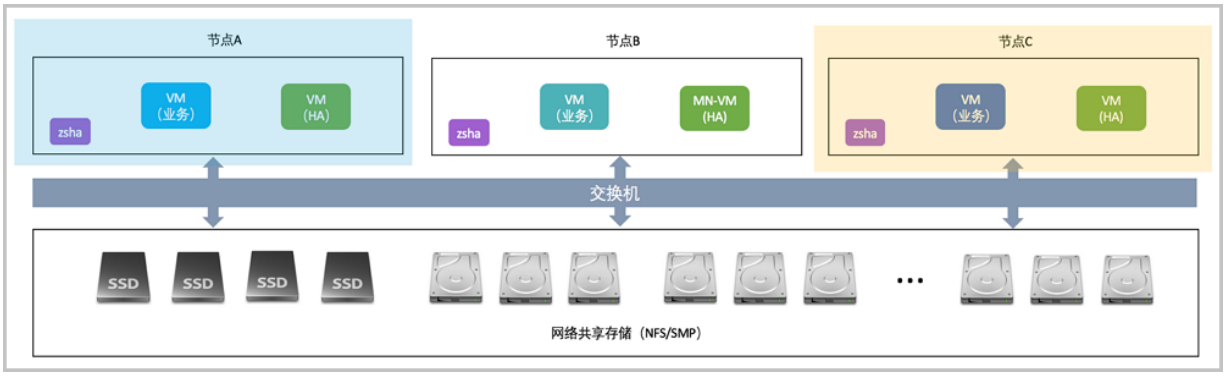
- 本手册中“网络故障”，特指管理与存储网络的故障，云主机数据网络故障，不会触发高可用切换。
- 本手册中“网络故障”，特指两个链路均故障的情况，单链路故障，不会影响节点之间的通信。

2.2.1.2.1 该节点运行非管理节点主机

背景信息

节点A或C网络异常需恢复，如图 *2-14: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图*所示：

图 2-14: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图



以节点A网络异常为例。

操作步骤

1. 检查交换机运行状态，确认网络异常节点A的端口指示灯常亮或闪烁。
2. 通过ssh登录其他正在运行的节点B或C上，检查网络异常节点A高可用状态。

执行**zsha status**，检查网络异常节点A的高可用服务是否正常：

```
[root@localhost ~]# zsha status
>>>MN-VM Running On
172.20.14.215 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
172.20.13.239 : nfs-1
172.20.14.215 : nfs-2
172.20.14.251 : nfs-3
```

3. 查看物理主机连接状态。

通过账户admin登录ZStack for Alibaba Cloud云管理平台检查网络异常物理主机连接状态，如果连接正常，则网络异常节点A已恢复，如图 *2-15:* 物理主机连接状态界面所示：

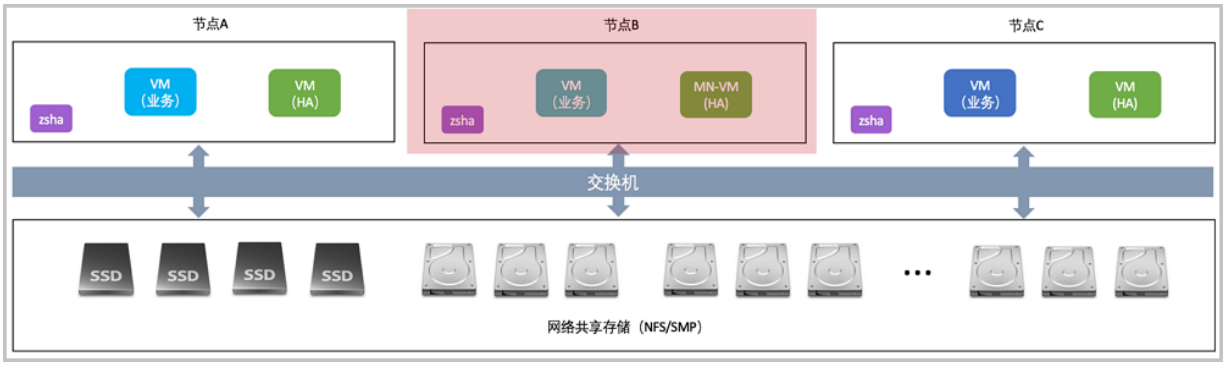
图 2-15: 物理主机连接状态界面

2.2.1.2.2 该节点运行管理节点主机

背景信息

节点B网路异常需恢复，如图 *2-16:* 基于*NFS/SMP*共享文件系统的高可用集群 经典三节点示例图所示：

图 2-16: 基于NFS/SMP共享文件系统的高可用集群 经典三节点示例图



操作步骤

1. 检查管理节点主机高可用是否已触发。

   通过ssh工具登陆到正在运行的节点A或C上， 执行**zsha status**检查管理节点主机是否正常运行，若正常运行，代表高可用套件已触发高可用切换。

   ```
   [root@localhost ~]# zsha status
   >>>MN-VM Running On
   172.20.14.251 : running

   >>>Last MN-VM Start Record
   ```

2. 检查交换机运行状态，确认网络异常节点B的端口指示灯常亮或闪烁。

3. 检查网络异常节点高可用状态。

   执行**zsha status**检查网络异常节点的高可用服务是否正常。

   ```
   [root@localhost ~]# zsha status
   >>>MN-VM Running On
   172.20.14.251 : running

   >>>Last MN-VM Start Record


   >>>ZStack HA Services Running On
   172.20.13.239 : nfs-1
   172.20.14.215 : nfs-2
   172.20.14.251 : nfs-3
   ```

4. 检查物理主机连接状态。

   通过账户admin登录ZStack for Alibaba Cloud云管理平台检查网络异常物理主机连接状态，如果连接正常，则网络异常节点B已恢复，如图 *2-17:* 物理主机连接状态界面所示：

   图 2-17: 物理主机连接状态界面

2.2.2 两节点异常处理

背景信息

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。在正常运行环境下，若其中两个节点均断电或网络异常。

操作步骤

1. 两异常节点上电后通过手动或IPMI方式启动。
2. 两异常节点加载操作系统，确认是否启动成功。
3. 检查两异常节点高可用状态。

   通过**zsha status**检查两异常节点的高可用服务是否正常，检查管理节点主机是否正常运行。

```
[root@localhost ~]# zsha status
>>>MN-VM Running On
172.20.14.215 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
172.20.13.239 : nfs-1
172.20.14.215 : nfs-2
172.20.14.251 : nfs-3
```

4. 查看物理主机连接状态。

   通过admin账户登陆ZStack for Alibaba Cloud云管理平台，检查两异常物理主机连接状态，如果连接正常，则两网络异常节点已恢复，如图 *2-18: 物理主机连接状态界面*所示：

图 2-18: 物理主机连接状态界面

2.3 节点修复

2.3.1 单节点故障修复

背景信息

ZStack for Alibaba Cloud基于NFS/SMP共享文件系统的高可用集群由三个节点A、B、C构建。若其中某个节点损坏后需要执行修复。

操作步骤

1. 调配备用服务器，使得硬件规格与故障节点相近。
2. 参考本文档安装部署章节，安装基础操作系统。安装完成后，配置root的密码和网络信息与故障节点一致。
3. 参考本文档安装部署章节，安装高可用套件。

   在健康节点，导出当前的高可用集群配置信息。

   ```
   [root@localhost ~]# zsha export-config
   output config json file to /tmp/current.config.json
   ```

   说明：

   若新版本升级时，需要增加/删除相关配置信息，请依照新版本发布说明。
4. 查看管理节点主机状态。

   通过**zsha status**检查高可用服务套件运行状态，以及管理节点主机状态。

   ```
   [root@localhost ~]# zsha status
   >>>MN-VM Running On
   172.20.14.215 : running

   >>>Last MN-VM Start Record


   >>>ZStack HA Services Running On
   172.20.13.239 : nfs-1
   172.20.14.215 : nfs-2
   172.20.14.251 : nfs-3
   ```
5. 对置换节点进行重连操作。

   通过admin账户登录ZStack for Alibaba Cloud云管理平台，对置换节点执行重连操作，如图 *2-19:* 重连物理主机界面所示：

   图 2-19: 重连物理主机界面

2.3.2 两节点故障无法修复

由于ZStack for Alibaba Cloud高可用三节点模型，只支持容忍一个节点发生故障，对于两节点损坏情况无法修复。若需满足两个节点可修复的场景，则需构建至少5个节点的集群。

3 命令行使用手册

3.1 简介

zsha是ZStack for Alibaba Cloud针对管理节点高可用场景设计的命令，帮助用户快速完成该场景下的多种操作。

zsha下有多条子命令，本文档将基于共享文件系统（NFS/SMP）的管理节点高可用场景对**zsha**每条子命令的作用和使用方法进行说明。

3.2 -h 帮助内容

描述

显示帮助，可查看**zsha**全部子命令。

```
[root@localhost ~]# zsha -h

zsha is a proprietary, high-availability suite that protects ZStack management services from
running online for long periods of time.

usage:
  zsha [-h] [-v]

All sub-commands
  install [-p] ROOT-PASSWORD [-c] CONFIGURATION-FILE-PATH
                Install Mode, Need a root password for all HA nodes, Need a configuration file path
  uninstall --yes-i-really-really-mean-it
                uninstall from all nodes
  check-config CONFIGURATION-FILE-PATH
                check configuration file
  config-sample [CephConf|FileConf]
                output sample config json file to /tmp/sample.CephConf.config.json
        or /tmp/sample.FileConf.config.json
  export-config     output current config json file to /tmp/current.config.json
  import-config CONFIGURATION-FILE-PATH
                update configuration to HA cluster
  migrate xxx.xxx.xxx.xxx
                migrate to target HA host, ip or hostname should be provided
  status            show status
  status-conf       show status with current configuration
  stop              close MN-VM and all zs-ha services in cluster, consul will continue to operate
  start             start all zs-ha services in cluster, if consul service is stopped, it will also be started

optional arguments:
  -h                show this help message
```

```
-v              show version, print execution details
```

3.3 -v 版本信息

描述

查看版本信息，包括版本号、编译日期、Branch Name和Commit ID。

```
[root@localhost ~]# zsha -v
App Version: 1.5.6.0
Build Date: Fri Jul 28 10:37:49 GMT-8 2017
Branch Name: (detached
Commit ID: c7dc57a012f5bae01eafaea7455bf5784007c172
```

3.4 check-config 配置检查

描述

检查配置文件格式和内容，该检查在安装和更新时也会被执行。

其中**ConsoleProxyOverriddenIP**用户可按需填写，也可不填，仅检查是否为有效的IPv4格式。其他各个字段必须填写，相应检查格式和内容。

管理节点主机的CPU数目和内存必须满足最低要求，可按需增加。

使用方法

参数	介绍	示例
CONFIGURATION-FILE-PATH	用户需填写完整的配置文件路径	zsha check-config /tmp/current.config.json

```
[root@localhost ~]# zsha check-config /tmp/current.config.json
config file check pass
```

3.5 install -p -c 安装命令

描述

安装命令，对HA集群内所有节点执行安装操作，要求HA集群内所有节点使用同一个password，用户需填写该password，以及配置文件路径。

使用方法

参数	介绍	示例
-p ROOT-PASSWORD	要求HA集群内所有节点使用同一个password，用户需填写该password	./ZStack for Alibaba Cloud-Enterprise-HA-Suite-2.x.bin install -p password
-c CONFIGURATION-FILE-PATH	用户需填写完整的配置文件路径	./ZStack for Alibaba Cloud-Enterprise-HA-Suite-2.x.bin install -c /tmp/config.json

```
[root@localhost ~]# ./ZStack-Enterprise-HA-Suite-2.x.bin install -p password -c /tmp/config.json
config file check pass
All nodes connect success.
check environment for 172.20.13.239
File /opt/smp/mngt/mnvm.img exists
Settings for br_eth0:
 Link detected: yes

check environment for 172.20.14.215
File /opt/smp/mngt/mnvm.img exists
Settings for br_eth0:
 Link detected: yes

check environment for 172.20.14.251
File /opt/smp/mngt/mnvm.img exists
Settings for br_eth0:
 Link detected: yes

Stop Service consul&z for: 172.20.13.239
Stop Service consul&z for: 172.20.14.215
Stop Service consul&z for: 172.20.14.251
Send /tmp/zstack-ha-installer/zsha to 172.20.13.239
Send /tmp/zstack-ha-installer/config.json to 172.20.13.239
Send /tmp/zstack-ha-installer/zsha to 172.20.14.215
Send /tmp/zstack-ha-installer/config.json to 172.20.14.215
Send /tmp/zstack-ha-installer/zsha to 172.20.14.251
Send /tmp/zstack-ha-installer/config.json to 172.20.14.251
Not Ceph Type, skip generate ceph auth.
Install for: 172.20.13.239
config file check pass
File /opt/smp/mngt/ exists
Settings for br_eth0:
 Link detected: yes

Install bash auto-completion
Set iptables finished.
Install to Current Node Finished.

Install for: 172.20.14.215
config file check pass
File /opt/smp/mngt/ exists
Settings for br_eth0:
 Link detected: yes

Install bash auto-completion
Set iptables finished.
```

```
Install to Current Node Finished.

Install for: 172.20.14.251
config file check pass
File /opt/smp/mngt/ exists
Settings for br_eth0:
 Link detected: yes

Install bash auto-completion
Set iptables finished.
Install to Current Node Finished.

Upsert configuration to consul kv finished.

Install to All Nodes Finished.
Now,You can monitor zsha's status by using command "zsha status".
More commands are shown in "zsha -h".
```

3.6 uninstall 卸载命令

描述

在HA集群中，对当前节点执行**zsha**卸载操作。建议用户自行备份当前节点内数据库再执行此操作。

```
[root@localhost ~]# zsha uninstall
WARNING: this will *DESTROY* ZStack Management Node VM.
If you are *ABSOLUTELY CERTAIN* that is what you want, followed by --yes-i-really-really-mean-it.
[root@localhost ~]# zsha uninstall --yes-i-really-really-mean-it
start clean
start shutdown cluster
Do for 172.20.13.239
Do for 172.20.14.215
Do for 172.20.14.251
shutdown cluster finished
Do for 172.20.13.239
Do for 172.20.14.215
Do for 172.20.14.251
clean finished
```

3.7 config-sample 样本配置生成

描述

生成样本配置文件并存放在临时路径*/tmp/sample.CephConf.config.json*或*/tmp/sample.FileConf.config.json*中。

使用方法

参数	介绍	示例
[CephConf|FileConf]	用户根据实际情况选择配置文件类型	zsha config-sample CephConf 或zsha config-sample FileConf

```
[root@localhost ~]# zsha config-sample CephConf
output config json file to /tmp/sample.CephConf.config.json

[root@localhost ~]# zsha config-sample FileConf
output config json file to /tmp/sample.FileConf.config.json
```

3.8 export-config 当前配置生成

描述

生成当前配置文件并存放在临时路径*/tmp/current.config.json*。

```
[root@localhost ~]# zsha export-config
output config json file to /tmp/current.config.json
```

3.9 migrate 迁移命令

描述

将管理节点主机从当前运行主机迁移到目标HA主机。

使用方法

参数	介绍	示例
xxx.xxx.xxx.xxx	目标HA主机为HA集群中一个正常的HA节点，xxx.xxx.xxx.xxx为该目标主机的IP地址	zsha migrate xxx.xxx.xxx.xxx
hostname	目标HA主机为HA集群中一个正常的HA节点，hostname为该目标主机的名称	zsha migrate hostname

```
[root@localhost ~]# zsha migrate 172.20.198.176
Migrate from 172.20.197.242 to 172.20.198.176
Migration: [100 %]

[root@localhost ~]# zsha migrate ceph-3
Migrate from 172.20.198.176 to 172.20.197.202
```

```
Migration: [100 %]
```

3.10 import-config 配置升级

描述

对整个HA集群更新配置文件，需填写完整的配置文件路径。

使用方法

参数	介绍	示例
CONFIGURATION-FILE-PATH	用户需填写完整的配置文件路径	zsha import-config /tmp/current.config.json

```
[root@localhost ~]# zsha import-config /tmp/current.config.json
```

3.11 status 状态信息

描述

显示当前整个HA集群的状态，包括管理服务主机的当前运行状态和上次启动记录、集群中所有HA节点状态、以及管理服务主机的IP地址。

```
[root@localhost ~]# zsha status
>>>MN-VM Running On
192.168.255.205 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
192.168.255.205 nfs-1
192.168.255.221 nfs-2
192.168.255.183 nfs-3

>>>Consul Members
Node            Address              Status  Type    Build  Protocol  DC
192.168.255.205  192.168.255.205:8301  alive   server  0.7.2  2         dc1
192.168.255.221  192.168.255.221:8301  alive   server  0.7.2  2         dc1
192.168.255.183  192.168.255.183:8301  alive   server  0.7.2  2         dc1


>>>Management Node IP
172.20.198.3

>>>Public Network IP
```

```
203.114.54.10
```

使用方法

参数	介绍	示例
-conf	显示整个HA集群状态的同时，显示当前配置情况	zsha status-conf

```
[root@localhost ~]# zsha status-conf
>>>MN-VM Running On
192.168.255.205 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
192.168.255.205 nfs-1
192.168.255.221 nfs-2
192.168.255.183 nfs-3

>>>Consul Members
Node           Address              Status  Type    Build  Protocol  DC
192.168.255.205  192.168.255.205:8301  alive   server  0.7.2  2         dc1
192.168.255.221  192.168.255.221:8301  alive   server  0.7.2  2         dc1
192.168.255.183  192.168.255.183:8301  alive   server  0.7.2  2         dc1


>>>Management Node IP
172.20.198.3

>>>Public Network IP
203.114.54.10

>>>Configuration
{
  "Node": [
    "192.168.0.201 nfs-1",
    "192.168.0.202 nfs-2",
    "192.168.0.203 nfs-3"
  ],
  "MemorySizeInGB": 8,
  "CPU": 4,
  "Type": "FileConf",
  "MonAddrs": null,
  "ChronyServers": [
    "192.168.0.201",
    "192.168.0.202",
    "192.168.0.203"
  ],
  "PoolName": "",
  "ImageFolder": "/storage/",
  "DNS": [
    "192.168.0.1"
  ],
  "Network": [
    {
      "Bridge": "br_eth0",
      "MacAddress": "b2:e4:c2:1e:06:44",
```

```
      "Ipaddr": "192.168.0.204",
      "Netmask": "255.255.0.0",
      "Gateway": "192.168.0.1",
      "IsMgmt": true,
      "IsDefRoute": false
    },
    {
      "Bridge": "br_eth1",
      "MacAddress": "f6:75:d8:5d:73:73",
      "Ipaddr": "172.20.0.214",
      "Netmask": "255.255.0.0",
      "Gateway": "172.20.0.1",
      "IsMgmt": false,
      "IsDefRoute": true
    }
  ],
  "ConsoleProxyOverriddenIP": ""
}


>>>Checking Host I/O
done.
```

3.12 stop 集群关闭

描述

关闭集群内管理节点主机以及所有**zsha**服务，consul服务依然继续运行。

```
[root@localhost ~]# zsha stop
start shutdown cluster
Do for 192.168.255.205
Do for 192.168.255.221
Do for 192.168.255.183
shutdown cluster finished
[root@localhost ~]# zsha status
>>>MN-VM Running On
192.168.255.205 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
192.168.255.205 nfs-1
192.168.255.221 nfs-2
192.168.255.183 nfs-3

>>>Consul Members
Node           Address              Status  Type    Build  Protocol  DC
192.168.255.205  192.168.255.205:8301  alive   server  0.7.2  2         dc1
192.168.255.221  192.168.255.221:8301  alive   server  0.7.2  2         dc1
192.168.255.183  192.168.255.183:8301  alive   server  0.7.2  2         dc1


>>>Management Node IP
172.20.198.3

>>>Public Network IP
203.114.54.10
```

```
>>>Checking Host I/O
done.
```

3.13 start 集群启动

描述

启动集群内所有**zsha**服务，对于处于停止状态的consul服务，也将被启动。

```
[root@localhost ~]# zsha start
Do for 192.168.255.205
Do for 192.168.255.221
Do for 192.168.255.183
[root@localhost ~]# zsha status
>>>MN-VM Running On
192.168.255.205 : running

>>>Last MN-VM Start Record


>>>ZStack HA Services Running On
192.168.255.205 nfs-1
192.168.255.221 nfs-2
192.168.255.183 nfs-3

>>>Consul Members
Node           Address              Status  Type    Build  Protocol  DC
192.168.255.205  192.168.255.205:8301  alive   server  0.7.2  2         dc1
192.168.255.221  192.168.255.221:8301  alive   server  0.7.2  2         dc1
192.168.255.183  192.168.255.183:8301  alive   server  0.7.2  2         dc1


>>>Management Node IP
172.20.198.3

>>>Public Network IP
203.114.54.10


>>>Checking Host I/O
done.
```

3.14 reset-MNVM-password 管理节点主机重置root密码

描述

管理节点主机重置root密码为系统默认值（即**password**），但前提是配置文件中的**AllowReset RootPassword**需设置为**true**。

该命令并未显示在*-h* 帮助内容中，仅供特殊情况使用。

```
[root@localhost ~]# zsha reset-MNVM-password
Reset MNVM root password on nodex IP //nodex IP为MNVM所在节点的IP
Password set successfully for root in 173de36b-4a5d-458e-a5e3-9a99584551ef
```

专有云术语表

区域（Zone）

ZStack中最大的一个资源定义，包括集群、二层网络、主存储等资源。

集群（Cluster）

一个集群是类似物理主机（Host）组成的逻辑组。在同一个集群中的物理主机必须安装相同的操作系统（虚拟机管理程序，Hypervisor），拥有相同的二层网络连接，可以访问相同的主存储。在实际的数据中心，一个集群通常对应一个机架（Rack）。

管理节点（Management Node）

安装系统的物理主机，提供UI管理、云平台部署功能。

计算节点（Compute Node）

也称之为物理主机（或物理机），为云主机实例提供计算、网络、存储等资源的物理主机。

主存储（Primary Storage）

用于存储云主机磁盘文件的存储服务器。支持本地存储、NFS、 Ceph、Shared Mount Point、Shared Block等类型。

镜像服务器（Backup Storage）

也称之为备份存储服务器，主要用于保存镜像模板文件。建议单独部署镜像服务器。

镜像仓库（Image Store）

镜像服务器的一种类型，可以为正在运行的云主机快速创建镜像，高效管理云主机镜像的版本变迁以及发布，实现快速上传、下载镜像，镜像快照，以及导出镜像的操作。

云主机（VM Instance）

运行在物理机上的虚拟机实例，具有独立的IP地址，可以访问公共网络，运行应用服务。

镜像（Image）

云主机或云盘使用的镜像模板文件，镜像模板包括系统云盘镜像和数据云盘镜像。

云盘（Volume）

云主机的数据盘，给云主机提供额外的存储空间，共享云盘可挂载到一个或多个云主机共同使用。

计算规格（Instance Offering）

启动云主机涉及到的CPU数量、内存、网络设置等规格定义。

云盘规格（Disk Offering）

创建云盘容量大小的规格定义。

二层网络（L2 Network）

二层网络对应于一个二层广播域，进行二层相关的隔离。一般用物理网络的设备名称标识。

三层网络（L3 Network）

云主机使用的网络配置，包括IP地址范围、网关、DNS等。

公有网络（Public Network）

由因特网信息中心分配的公有IP地址或者可以连接到外部互联网的IP地址。

私有网络（Private Network）

云主机连接和使用的内部网络。

L2NoVlanNetwork

物理主机的网络连接不采用Vlan设置。

L2VlanNetwork

物理主机节点的网络连接采用Vlan设置，Vlan需要在交换机端提前进行设置。

VXLAN网络池（VXLAN Network Pool）

VXLAN网络中的 Underlay 网络，一个 VXLAN 网络池可以创建多个 VXLAN Overlay 网络（即 VXLAN 网络），这些 Overlay 网络运行在同一组 Underlay 网络设施上。

VXLAN网络（VXLAN）

使用 VXLAN 协议封装的二层网络，单个 VXLAN 网络需从属于一个大的 VXLAN 网络池，不同 VXLAN 网络间相互二层隔离。

云路由（vRouter）

云路由通过定制的Linux云主机来实现的多种网络服务。

安全组（Security Group）

针对云主机进行第三层网络的防火墙控制，对IP地址、网络包类型或网络包流向等可以设置不同的安全规则。

弹性IP（EIP）

公有网络接入到私有网络的IP地址。

快照（Snapshot）

某一个时间点上某一个磁盘的数据备份。包括自动快照和手动快照两种类型。

混合云术语表

访问密钥（AccessKey）

用于调用阿里云API或大河云联API的唯一凭证，AccessKey包括AccessKeyID（用于标识用户）和AccessKeySecret（用于验证用户密钥）。

数据中心（Data Center）

包含阿里云的地域和可用区等地域资源，用于匹配阿里云资源的地域属性。

地域（Region）

物理的数据中心，划分地区的基本单位，ZStack混合云的地域对应了阿里云端的地域。

可用区（Identity Zone）

在同一地域内，电力和网络互相独立的物理区域，ZStack混合云的可用区对应了阿里云端的可用区（Zone）。

存储空间（Bucket）

用于存储对象（Object）的容器，ZStack使用对象存储（OSS）里的Bucket来上传镜像文件。

ECS云主机（Elastic Compute Service）

阿里云端创建的ECS实例，可在ZStack混合云界面进行ECS云主机生命周期的管理。

专有网络VPC（Virtual Private Cloud）

用户基于阿里云构建的一个隔离的网络环境，不同的专有网络之间逻辑上彻底隔离。

虚拟交换机（VSwitch）

组成专有网络VPC的基础网络设备，可以连接不同的云产品实例。ZStack混合云的虚拟交换机对应了阿里云VPC下的虚拟交换机。

虚拟路由器（VRouter）

专有网络VPC的枢纽，可以连接专有网络的各个虚拟交换机，同时也是连接专有网络与其它网络的网关设备。ZStack支持查看VPC下的虚拟路由器。

路由表（Route Table）

虚拟路由器上管理路由条目的列表。

路由条目（Route Entry）

路由表中的每一项是一条路由条目。路由条目定义了通向指定目标网段的网络流量的下一跳地址。路由条目包括系统路由和自定义路由两种类型。ZStack支持自定义类型的路由条目。

安全组（Security Group）

针对云主机进行第三层网络的防火墙控制。ZStack混合云的安全组对应了阿里云端ECS云主机三层隔离的防火墙约束。

镜像（Image）

云主机使用的镜像模板文件，一般包括操作系统和预装的软件。ZStack支持上传本地镜像到阿里云，以及使用阿里云端镜像。

弹性公网IP（EIP）

阿里云端公有网络池中的IP地址，绑定弹性公网IP的ECS实例可以直接使用该IP进行公网通信。

VPN连接（VPN Connection）

通过建立点对点的IPsec VPN通道，实现企业本地数据中心的私有网络与阿里云端VPN网络进行通信。

VPN网关（VPN Gateway）

一款基于Internet，通过加密通道将本地数据中心和阿里云专有网络VPC安全可靠连接起来的服务。用户在阿里云VPC创建的IPsec VPN网关，与本地数据中心的用户网关配合使用。

VPN用户网关（Customer Gateway）

本地数据中心的VPN服务网关。可通过ZStack混合云创建VPN用户网关，并将VPN用户网关与VPN网关连接起来。

高速通道（Express Connect）

通过物理专线（即租用运营商的专线：电缆或光纤），连通本地数据中心到阿里云专线接入点，与阿里云VPC环境打通，实现云上云下不同网络间高速，稳定，安全的私网通信。

边界路由器（VBR）

用户申请的物理专线接入交换机的产品映射。用户在物理专线上可以创建边界路由器，边界路由器负责专线上的数据在阿里云上进行转发。通过边界路由器，用户数据可以直达阿里云VPC网络。

路由器接口（Router Interface）

一种虚拟的网络设备，可以挂载在路由器并与其他路由器接口进行高速通道互联，实现不同网络间的内网互通。