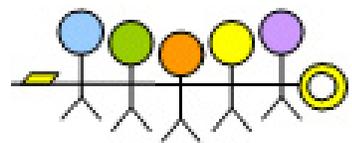


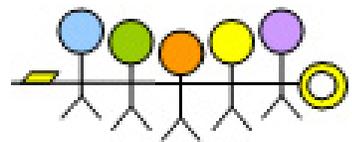
SaltStack进行Ceph部署 和运维实战

休伦科技 郑伟

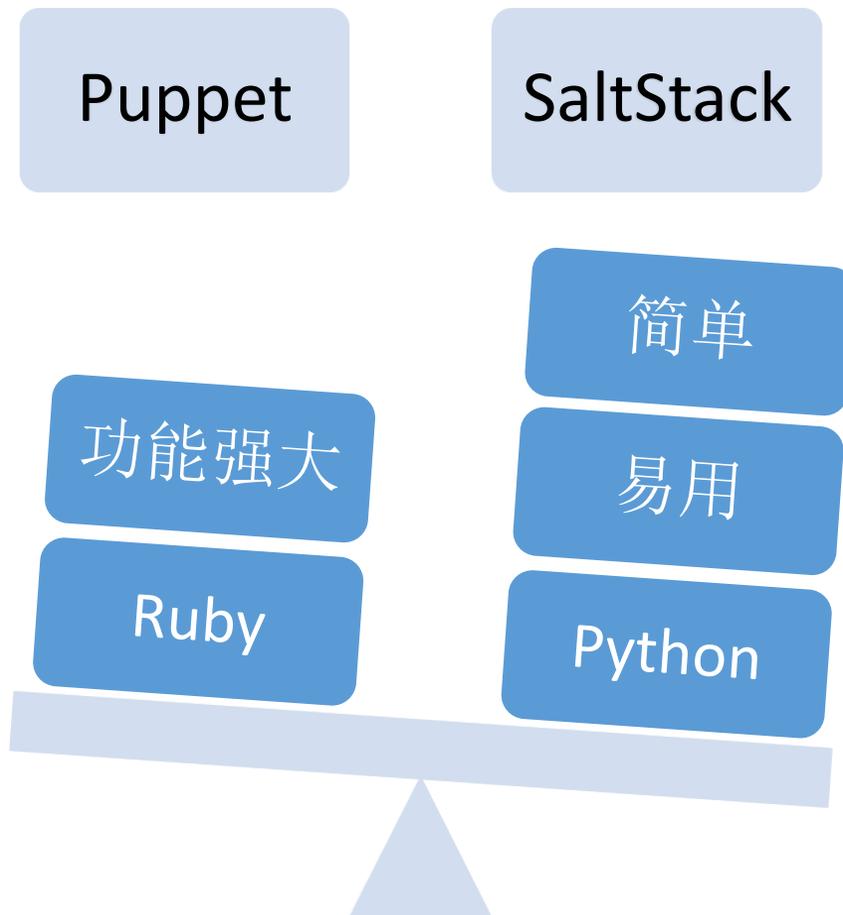


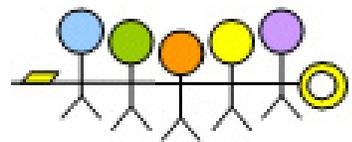
Ceph部署进化史

- 2014年3月初，开始设计ceph自动化部署
- 2014年3月中，0.1的版本release
- 2014年5月初，1.0的版本release
- 2014年8月初，平台迁移——Ubuntu到CentOS
- 2015年3月初，添加Web页面 2.0

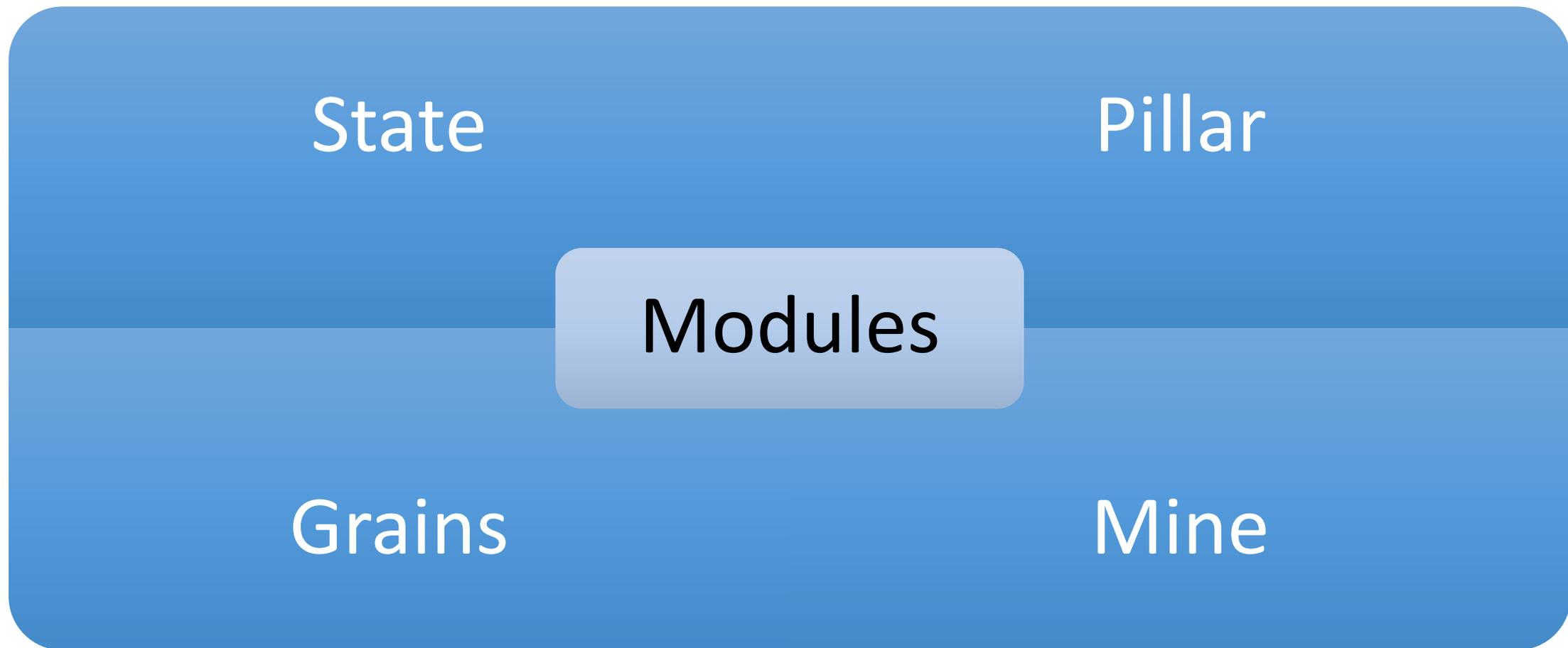


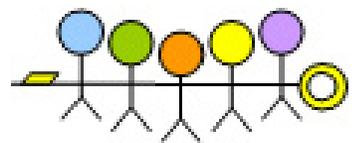
从0开始——部署工具的选择





SaltStack简介





Ceph-deploy部署流程

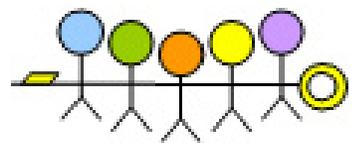
建立ssh信任

安装软件包

初始化添加mon

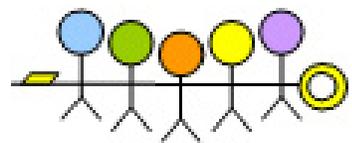
生成keyring

添加osd



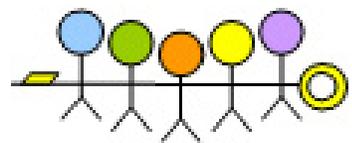
V0.1 遇到的问题

- ~~1. 启动服务，没有反应~~
- ~~2. 创建keyring超时，导致部署任务失败~~
- ~~3. Keyring同步问题~~
4. 添加的节点同时运行部署任务
5. 添加新节点必须需要修改salt的配置文件
6. OSD的添加不够灵活



V1版本重构

1. 通过在节点的Grains中定义ceph_mon的roles，代替pillar中设置mon节点
2. 将部署任务拆分为mon，osd，ceph_conf
3. 设置触发任务，当添加新的mon节点后，更新集群内所有节点的ceph.conf 文件



V1版本部署流程

定义角色

- salt-call grains.set_val roles ["ceph_mon"]

更新状态

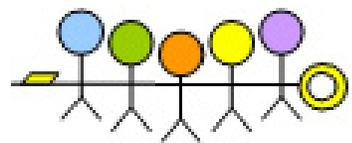
- salt-call mine.update

部署

- salt-call state.sls ceph.mon

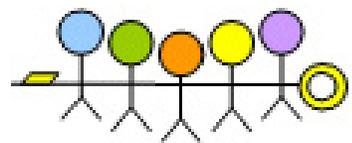
更新配置

- salt-call event.fire_master 'add_mon' 'update_ceph'



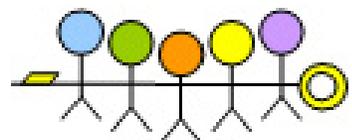
V1版本缺点

- 操作比较麻烦
- 需要salt基础
- 网络的配置
- Ceph osd 的添加还是不够灵活



通过Web管理配置和部署

- 抽象部署逻辑
- 不通过命令行修改配置参数
- 监控部署进度和结果
- 展示集群状态
- 自定义OSD的添加



Ceph部署配置

Ceph副本数量

3

Public Interface

eth4

Public Netowrk

10.80.10.0/24

Cluster是否复用Public



是



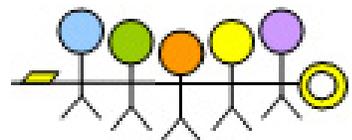
否

Cluster Interface

eth4

Cluster Netowrk

10.80.10.0/24



MON节点列表

Ceph集群信息

MON

OSD

刷新Ceph集群信息

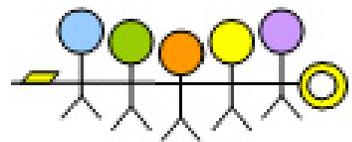
添加MON

添加OSD

Search



节点类型	ID	主机	操作
mon	compute1.econe.com	compute1.econe.com	删除
mon	compute2.econe.com	compute2.econe.com	删除
mon	seed.econe.com	seed.econe.com	删除



OSD节点列表

Ceph集群信息

MON

OSD

刷新Ceph集群信息

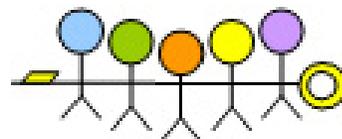
添加MON

添加OSD

Search



节点类型	ID	主机	状态	IN/OUT	Apply Latency(ms)	Commit Latency(ms)	Used/All	操作
osd	0	seed.econe.com	up	in	0	0	5/488	OUT 删除
osd	1	seed.econe.com	up	in	2	1	7/488	OUT 删除
osd	2	seed.econe.com	up	in	0	0	6/488	OUT 删除
osd	3	seed.econe.com	up	in	0	0	7/488	OUT 删除
osd	4	seed.econe.com	up	in	1	1	18/976	OUT 删除
osd	5	seed.econe.com	up	in	2	2	7/488	OUT 删除
osd	6	compute1.econe.com	up	in	2	1	16/976	OUT 删除
osd	7	compute1.econe.com	up	in	0	0	8/488	OUT 删除



添加OSD

添加OSD



物理机

compute1.econe.com

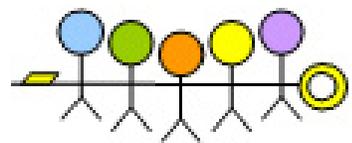
磁盘

None selected

日志磁盘

将日志存放到数据盘

创建OSD



OSD磁盘的初始化

- 软Raid信息

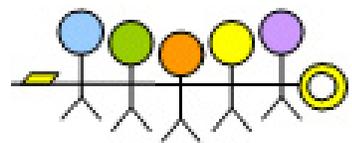
```
for i in `mdadm --detail --scan | awk '{print $2}'`; do mdadm -S $i; done
```

- 逻辑卷信息

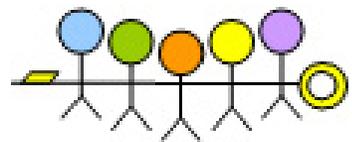
暂时未作处理

- Journal盘残留的分区信息

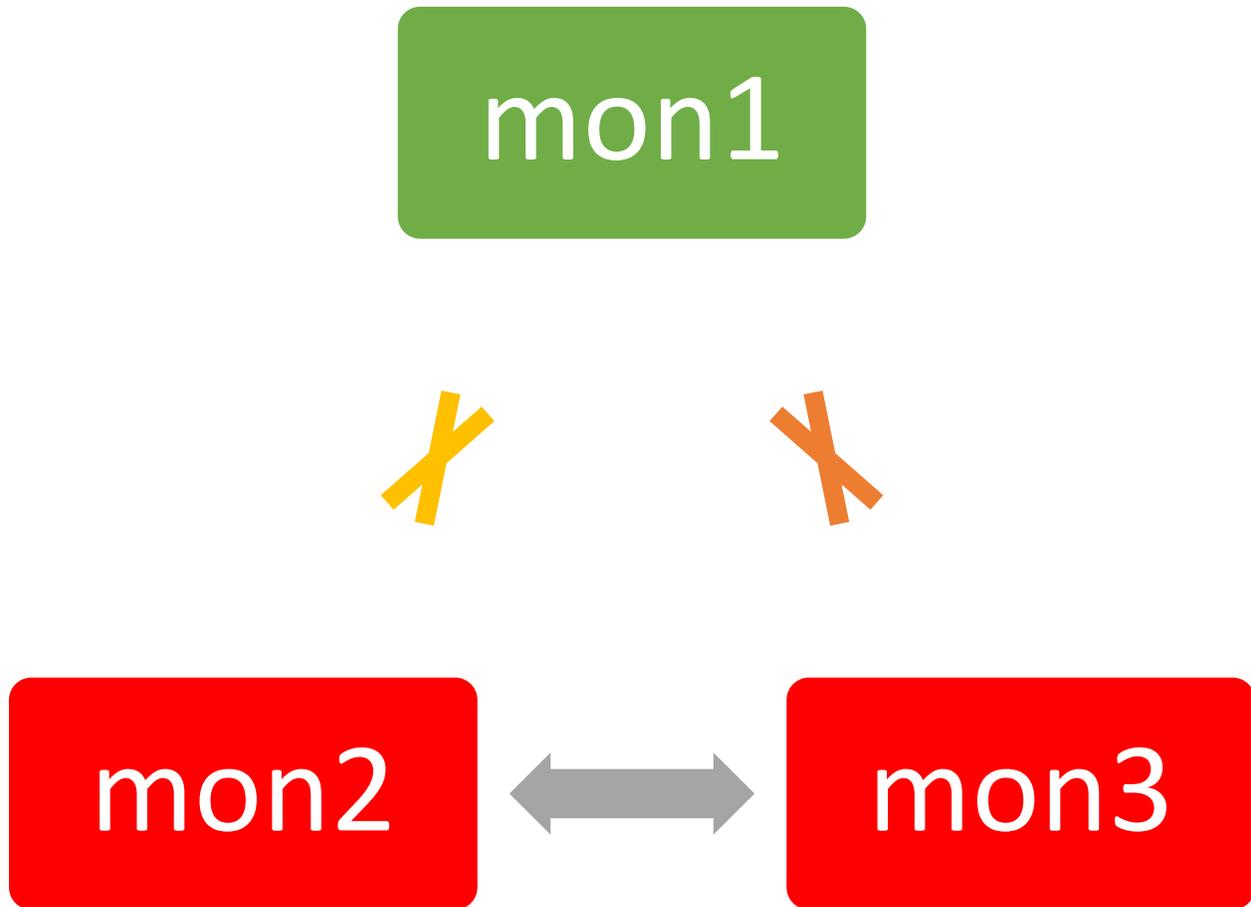
1. 设置空文件做tag，标识磁盘已经被清理过
2. `dd if=/dev/zero of=/dev/sdb bs=1M count=10 && sgdisk -zap /dev/sdb`

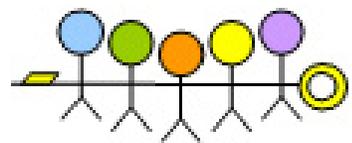


故障与运维



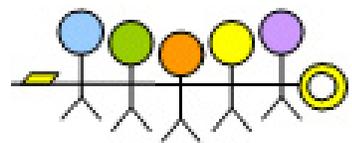
Mon节点失效过多





至少有一个节点存活

1. 使用monmaptool 导出 ceph mon map
2. 将生成的monmap 和 mon keyring 远程复制到目标机器
3. 通过mon map 创建新的mon
4. 启动服务，使集群恢复正常



monmaptool 导出 mon map

monmaptool

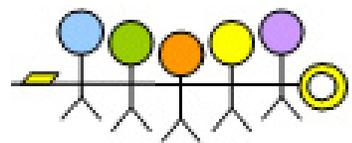
--create

--add node1.econe.com 192.168.0.1:6789

--add node2.econe.com 192.168.0.2:6789

--fsid 059f27e8-a23f-4587-9033-3e3679d03b31

--clobber /root/monmap



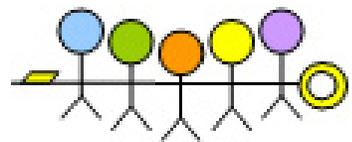
通过mon map 和 mon keyring 创建新mon

将monmap 和 mon keyring 拷贝到需要恢复的节点上。

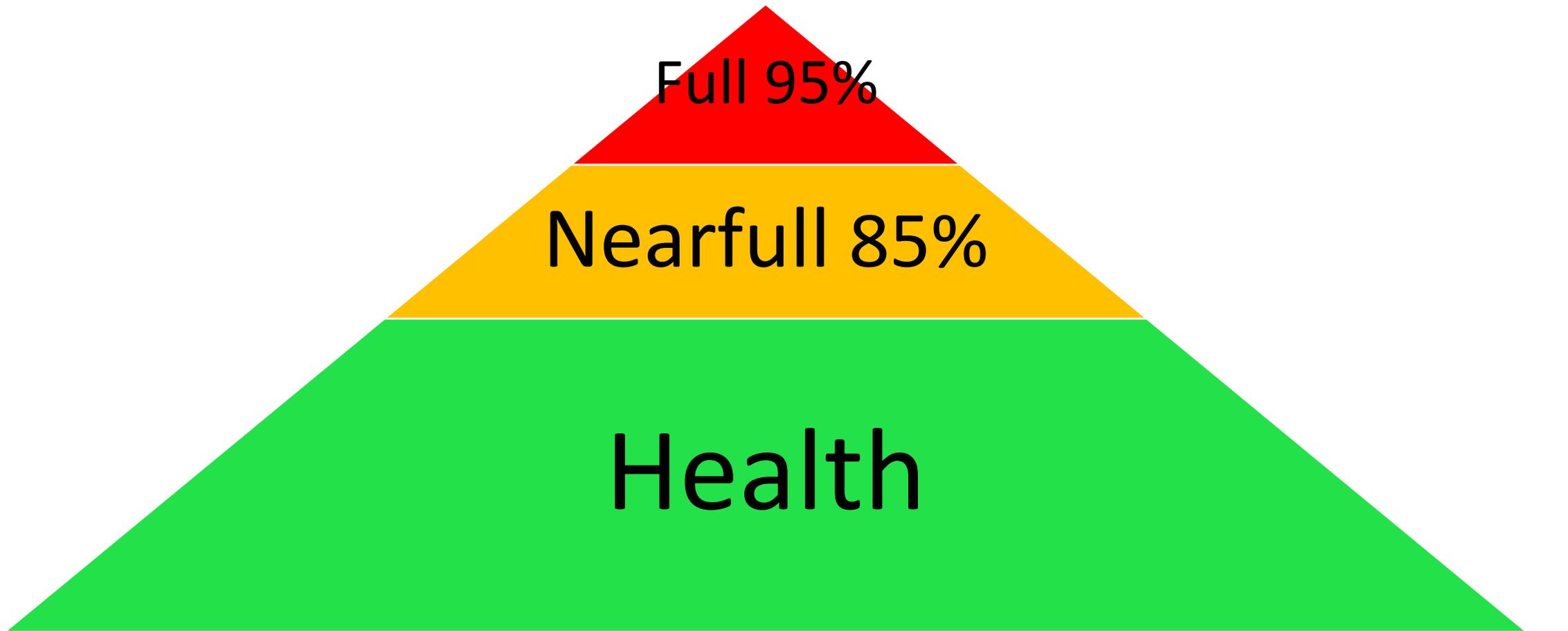
```
ceph-mon
```

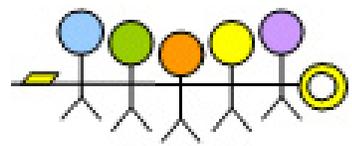
```
--cluster ceph  
-i node2.econe.com  
--mkfs --monmap /opt/monmap  
--keyring /opt/mon.keyring
```

```
touch /var/lib/ceph/mon/ceph-node2.econe.com/sysvinit  
/etc/init.d/ceph start mon #启动服务
```



集群使用量大于95%!!!





添加临时OSD应急

1. 创建一个虚拟磁盘镜像 #dd 或者qemi-img
2. 添加成为loop设备，并格式化为xfs文件系统
3. `ceph osd create` #创建一个OSD，假设返回的id是18
4. `mkdir /var/lib/ceph/osd/osd-18`
5. `mount /dev/loop0 /var/lib/ceph/osd/osd-18`
6. `ceph-osd -i 18 --mkfs --mkkey`
7. `ceph auth import -i keyring`
8. `ceph auth caps osd.18 mon 'allow profile osd' osd 'allow *'`
9. `touch /var/lib/ceph/osd/ceph-18/sysvinit`
10. `/etc/init.d/ceph start osd.18`