

基于OpenStack构建网易 云主机服务

网易杭州研究院
张晓龙

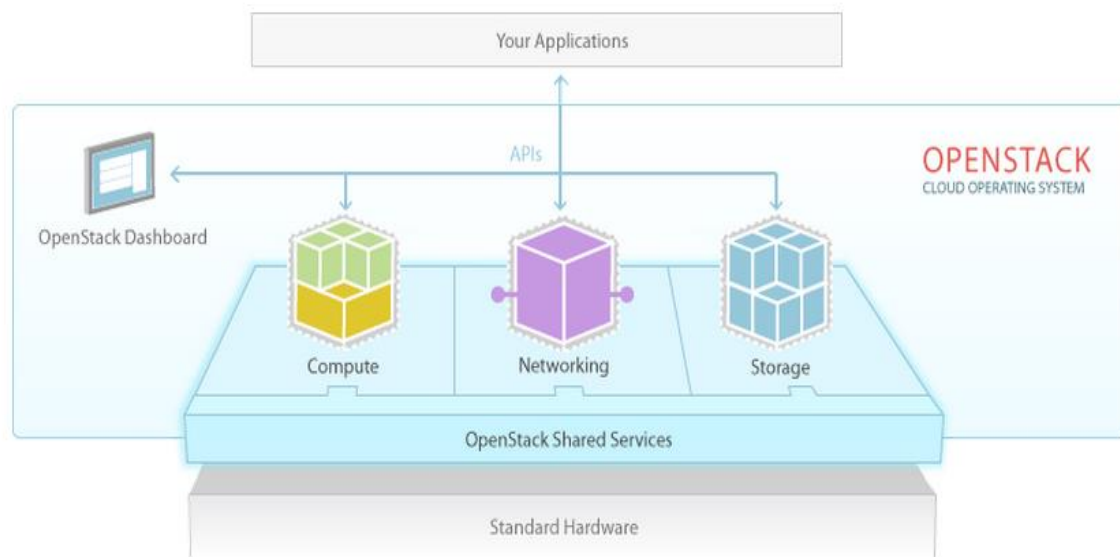


内容

- **OpenStack项目介绍**
- **网易私有云平台**
- **网易云主机服务**
- **OpenStack开发实践**
- **配置部署情况**
- **社区参与和未来工作**
- **Q&A**

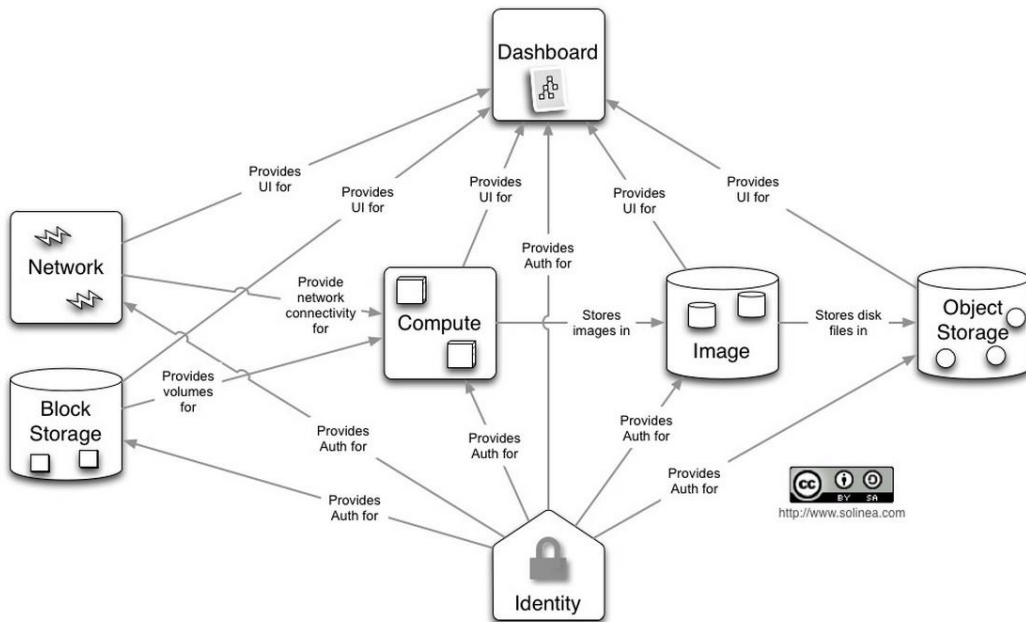
OpenStack项目

- 目标：
 - 为公共及私有云建设提供一个开放、简单易实现、高可扩展性的云计算平台
- 特征
 - 开放源代码：由NASA和Rackspace联合开源
 - 体系架构清晰、组件化，易定制及二次开发
 - 社区发展迅速，6个月发行一个大版本
 - OpenStack生态系统日趋完善



OpenStack组件

- 核心项目组件
 - Compute (Nova)
 - Networking (Quantum)
 - Block Storage (Cinder)
 - Object Storage (Swift)
 - Identity (Keystone)
 - Image Service (Glance)
 - Dashboard (Horizon)
 - Orchestration Service (Heat)



OpenStack	AWS
Nova	EC2
Quantum	VPC
Swift	S3
Cinder	EBS
Keystone	IAM
Heat	CloudFormation
Horizon	Console

OpenStack VS. Amazon AWS

网易私有云平台

- 目标：为网易公司大量的WEB类互联网产品提供统一的云计算平台
 - 提高硬件资源利用率，促进资源共享，以降低硬件成本
 - 提高资源管理与系统运维的自动化水平，以降低运维成本
 - 提高资源使用弹性，以增强业务波动的适应能力
 - 促进公共技术研发与应用，使业务获得更好的基础技术服务

网易私有云平台

- 核心IaaS服务

提供计算、存储、网络等核心IT设备的虚拟化

- 核心PaaS服务

提供海量结构化与非结构化数据存储、管理与检索功能，满足互联网典型平台需求



网易云主机服务

- 目标

- 作为网易私有云平台核心服务，提供弹性、高效、稳定可靠的虚拟机服务，满足公司产品上线、开发测试等对IT基础设施的需求

- 现状

- 基于开源云平台OpenStack开发
- 对OpenStack进行了全面测试
- 修复了OpenStack若干Bugs
- 开发并优化了OpenStack若干功能
- 完成了与私有云其他服务的整合
- 目前已提供了稳定云主机服务



主要功能

- 云主机生命周期管理
- 云主机状态详情查看
- 镜像快照管理
- 网络与访问安全管理
- 云主机计费管理
- 云主机监控报警管理

开发部署历程

- 2012.3月开始研究OpenStack
- 2012.4月部署OpenStack E版本供网易相册试用
- 2012.11月发布基于OpenStack E版本开发的云主机服务，随后网易相册和网易云课堂正式上线
- 2013.3月发布基于F版本开发的云主机服务
- 2013.4月完成网易博客的迁移上线
- 截止2013年7月
 - 已稳定运行8个多月
 - 中间经历一次云主机服务的在线平滑升级
 - 一共13个产品上线
 - 规模：几十个物理节点，几百台云主机

OpenStack优化和新功能开发

- 云主机服务质量保证
- 镜像快照存储处理优化
- 调整云主机规格优化
- 用户自助服务和运维管理平台
- 云主机监控报警功能
- 云主机实例存储配额功能
-
- ...

云主机服务质量保证

- 云平台物理资源共享带来挑战
 - 云主机性能指标变模糊，无法精确定义
 - 云主机间以及云主机与宿主机竞争资源，使云主机性能不稳定
- 目标
 - 提供性能指标明确、性能稳定可靠的云主机
- 方案
 - 明确定义云主机性能指标
 - 控制云主机资源占用，避免云主机间相互影响
 - 预留一定物理资源给宿主机，避免影响宿主机正常运行
- 当前实现的服务质量保证（QoS）
 - 计算（CPU）、网络带宽

计算资源 QoS

- 定义性能指标

- 提出以网易EUCU为基本单位来衡量云主机计算能力
- **1EUCU**定义为**1/4 Intel E5 2650单核**或**1/3 AMD 6276单核**
- 制定多种计算规格(**1VCPU X 1EUCU, 1VCPU X 2EUCU, ...**) , 以量化不同规格云主机计算能力

- 性能指标需求

- 相同规格云主机的计算能力基本相近
- 不同规格云主机的计算能力差异可量化
- 无论宿主机整体负载如何，云主机计算能力不应出现大幅波动

计算资源 QoS

- 性能指标保证策略

- 利用cgroup cpu子系统控制云主机计算能力
- 根据规格配置cgroup参数：`share/period/quota`
 - 根据ECU数目设置`cpu.shares`(ECU数 X 1024)
 - `cpu.cfs_period_us`统一设置为100ms
 - 测试确定各规格云主机在不同机型上`cpu.cfs_quota_us`值
- 将物理核分两类：宿主机保留以及云主机使用
- 保留物理核给宿主机以保障系统控制、网络I/O等正常运行，同时也提升了虚拟化性能（20%+）

计算资源 QoS (续)

- 设定云主机虚拟CPU范围绑定到宿主机物理核集合上，可减少云主机计算性能表现波动（10%内）
- **OpenStack实现**
 - 将云主机计算能力（ECU）当做新资源有效管理
 - 向云主机规格（Flavor）中加入ECU相关信息
 - 往OpenStack调度器中增加新调度器ECUFilter
 - 在计算节点上增加ECU资源信息上报流程，在控制节点上增加ECU资源信息统计流程
 - 在Libvirt Driver层支持设置ECU相关的cgroup参数
 - 增加ECU配额管理功能

网络带宽资源QoS

- 性能指标需求
 - 外网带宽、内网带宽
- 性能指标保证策略
 - 利用Linux TC控制云主机对内外网带宽资源占用
 - 设置TC rate参数，保证云主机带宽性能
 - 设置TC ceil参数，当带宽富余时提升云主机带宽
 - 外网带宽：创建云主机时由用户指定
 - 内网带宽：根据云主机计算能力(ECU)制定默认控制策略，创建云主机时无需指定

网络带宽资源QoS（续）

- 预留一定带宽资源给宿主机，保证其正常带宽需求
- **OpenStack实现**
 - 将网络带宽当做新资源有效管理
 - 往OpenStack调度器中增加新调度器NetworkFilter
 - 在计算节点增加带宽资源信息上报流程，在控制节点增加带宽资源统计流程
 - 增加网络带宽管理API，支持查看、修改带宽大小
 - 增加外网带宽配额管理功能

镜像快照存储处理优化

- 问题
 - 镜像快照太多太大，给存储系统带来巨大负担
 - 上传下载镜像快照会占用云平台大量带宽资源
- 启发
 - 不同镜像快照之间实际上会存在大量重复数据块（50%+数据块重复）
- 解决方案
 - 实现镜像快照分块处理策略：固定大小(4M)分块
 - 增加数据块缓存以加速镜像快照下载

用户自助服务



网易云 beta
cloud.163.com

下午好, **opstest**

[安全退出](#) | [帮助](#)

云主机

云硬盘

对象存储

关系型数据库

分布式数据库

云监控

云搜索

云主机首页

云主机首页

云主机管理

镜像管理

安全组管理

快照管理

网络资源管理

密钥管理

创建云主机

云主机是网易提供的云端计算服务，用户可以根据自己的需求创建云主机，系统负责对云主机进行全生命周期管理。

[创建云主机](#)

我的配额

当前云主机服务中包含的配额如下：

[刷新](#)

云主机配额（台）：30

云主机CPU配额（个）：60

ECU配额（个）：120

内网浮动IP配额（个）：30

外网浮动IP配额（个）：0

安全组配额（个）：30

实例存储容量配额（GB）：-1

内存容量配额（GB）：120

外网带宽配额（Mb/s）：100

我的资源

当前云主机服务中包含的资源如下：

[刷新](#)

云主机（个）：7

云主机CPU（个）：17

ECU（个）：52

内网浮动IP（个）：2

外网浮动IP（个）：0

安全组（个）：2

实例存储容量（GB）：90

内存容量（GB）：36

外网带宽（Mb/s）：20

相关链接

[> 特性介绍](#)

[> 用户手册](#)

[> 产品价格](#)

[> 配额申请](#)

ADC·阿里技术嘉年华

运维管理平台

首页

云主机管理

统计

管理操作

报警管理

用户管理

系统资源

可用域资源

节点资源

用户资源

用户操作

云主机数 (台)

[查看历史详情](#)

已用: 67

VCPU (个)

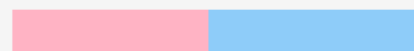
[查看历史详情](#)

已用: 136

ECU (个)

[查看历史详情](#)

总量: 720



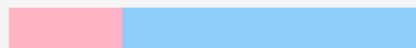
已用: 345

可用: 375

内存 (GB)

[查看历史详情](#)

总量: 757



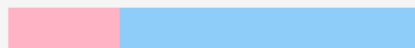
已用: 209

可用: 549

实例存储 (GB)

[查看历史详情](#)

总量: 8,244



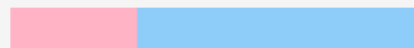
已用: 2,230

可用: 6,014

内网浮动IP (个)

[查看历史详情](#)

总量: 1,022



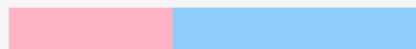
已用: 315

可用: 707

外网浮动IP (个)

[查看历史详情](#)

总量: 30



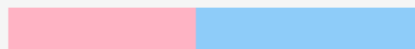
已用: 12

可用: 18

内网总带宽 (Mb/s)

[查看历史详情](#)

总量: 6,000



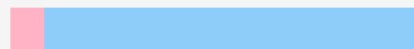
已用: 2,720

可用: 3,280

外网总带宽 (Mb/s)

[查看历史详情](#)

总量: 3,000



已用: 248

可用: 2,752

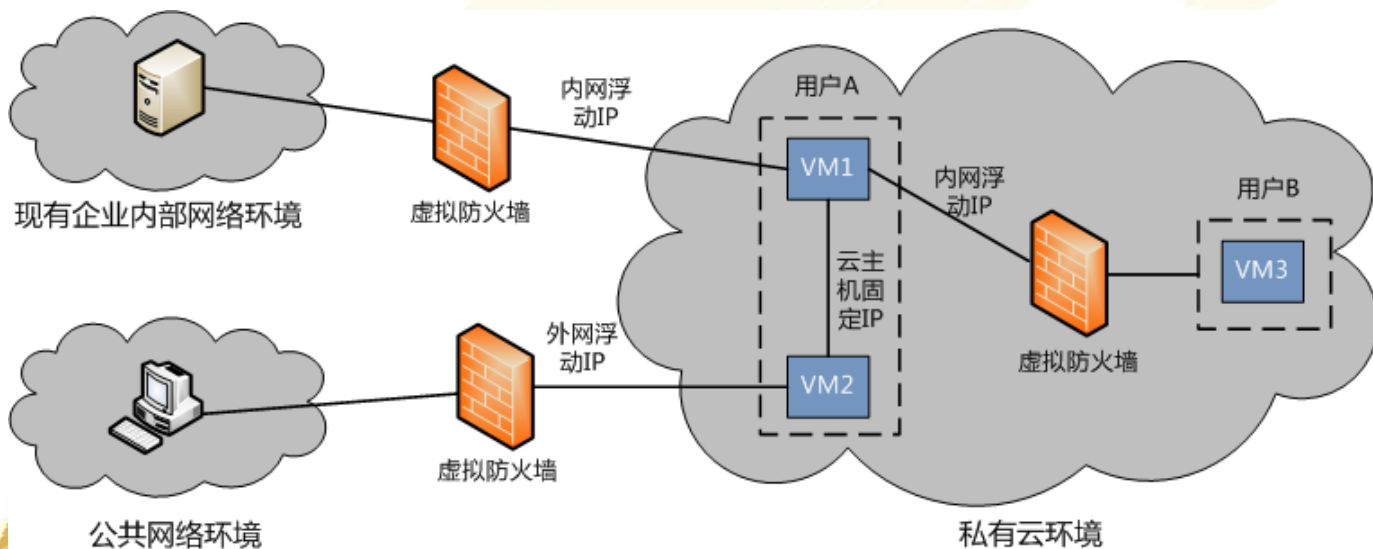
OpenStack网络开发

- 云环境租户网络隔离
 - 使用OpenStack FlatDHCP网络管理模式（简单灵活）
 - 不用VLAN模式的原因：
 - VLAN号受限（最大为4096）
 - 需机房预分配VLAN号并配置物理交换机（不灵活）
 - 实现：IPSET + IPTABLE
- 云环境与非云环境网络互通
 - 引入内网浮动(Floating)IP

OpenStack网络开发

- 私有云环境网络访问控制策略

- 云主机有三类IP：固定IP、内网浮动IP、外网浮动IP
- 利用云主机安全组（**虚拟防火墙**）控制网络访问该云主机
- 租户内云主机间网络访问不控制，以固定IP实现无条件网络互通
- 不同租户云主机间以及云主机与现有企业内网之间通过内网浮动IP访问互通，受安全组控制
- 公网用户通过外网浮动IP访问云环境云主机，受安全组控制



基本特征

- 高效管理IT物理资源
- 按需使用，弹性扩容
- 网络访问控制
- 服务质量保证
- 系统资源不超售
- 提供易用、友好的自助服务和运维平台

配置情况

- 平台相关配置
 - 宿主机系统：Debian 7.0 + Linux 3.2.x内核
 - KVM虚拟化 + Qcow2镜像格式
 - 宿主机系统盘RAID1，实例存储磁盘为RAID0
 - 使用网易对象存储服务存储镜像快照
 - 网络模式：FlatDHCP + multi_host
- 性能相关配置
 - 打开VHostNet，利用内核加速KVM网络性能
 - 打开宿主机透明大页支持，减少缺页及虚拟地址转换（性能提升10%+）

部署情况

- 物理硬件异构
 - CPU: Intel/AMD
 - 机型：机架服务器/刀片
 - 网络：万兆/千兆
- 服务高可用
 - RabbitMQ、Glance、Keystone、Nova-api
- 基于Puppet实现节点自动部署
- 错误日志监控
- 物理节点状态监控

参与社区情况

- 主要参与Nova、Glance组件的社区开发
- 向社区报告若干Bugs，并及时提交修复Commit
- 积极参与OpenStack社区的代码Review
- 在G版本开发周期中，共向Nova组件贡献16个Bugfixs
- 对H版本的贡献仍在继续




















```
Top changeset contributors by employer
Red Hat                374 (19.9%)
IBM                    366 (19.5%)
Rackspace              250 (13.3%)
HP                     127 (6.8%)
Cloudscaling           102 (5.4%)
Canonical               92 (4.9%)
Nebula                 86 (4.6%)
Intel                  42 (2.2%)
boris@pavlovic.me     30 (1.6%)
Metacloud              28 (1.5%)
Nicira                 24 (1.3%)
Cloudbase Solutions   24 (1.3%)
Citrix                 24 (1.3%)
AT&T                  20 (1.1%)
hanlind@kth.se        20 (1.1%)
VirtualTech            20 (1.1%)
NetEase                16 (0.9%)
NTT                    11 (0.6%)
ISI                    10 (0.5%)
Yahoo!                 10 (0.5%)
Covers 89.291422% of changesets
```

对Nova组件的贡献（G版本）

参与社区情况（续）

- 团队成员在社区上的经验值（贡献度）已比较高

Overall

Person	Project Karma	Total Karma
 Russell Bryant	55637	57593
 Thierry Carrez	17929	71936
 Vish Ishaya	16440	16779
 Alessandro Pilotti	3984	4569
 John Garbutt	3464	3616
 Boris Pavlovic	2810	4669
 rerngvit yanggratoke	2241	2254
 Dan Smith	2191	2215
 Christopher Yeoh	2141	2484
 Devananda van der Veen	2018	9158
 Senhua Huang	1878	1972
 Yaguang Tang	1782	2489
 dhardiputra	1649	6952
 Kiran Kumar Vaddi	1643	1881
 Chris Behrens	1610	1658
 Andrew Laski	1432	1509
 Anthony Harrington	1316	25907
 wangpan	1286	1297
 Aarti Kriplani	1284	1284

报告Bug及提交Bug Fixed情况

Importance	Status	Description	Links
Critical	Fixed Release	GET /v2.0/tokens/{token_id}/endpoints not implemented	https://bugs.launchpad.net/+bug/1006777
Critical	Fixed Release	Admin API /v2.0/tenants/{tenant_id}/users/{user_id}/roles doesn't validate token	https://bugs.launchpad.net/essex/+bug/1006815
High	Fixed Release	Heavily loaded nova-compute instances don't send reports frequently enough	https://bugs.launchpad.net/nova/+bug/1045152

Component	# of Testing API	Bugs Reported	Bug Fixed
Nova	121	21	16
Glance	14	2	2
Keystone	46	16	15
Total	181	39	29

未来工作

- 研发网络虚拟化
 - Quantum、Vxlan、Open vSwitch
- 让OpenStack更好支持容器级虚拟机LXC
- 运维管理相关新功能开发
-
- ...

Q & A

线下交流：xiaolongzhang.zju@gmail.com (Gtalk)

上传镜像

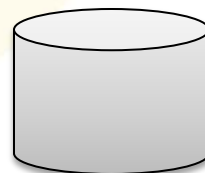
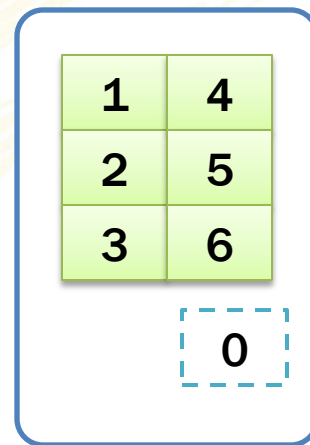
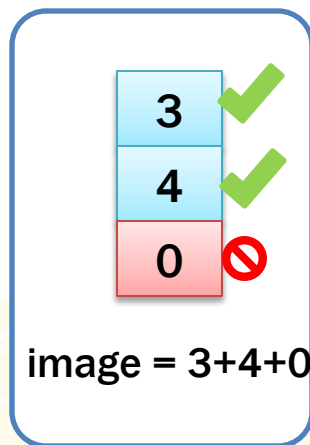


glance



backend storage

Upload:



Metadata Database

下载镜像

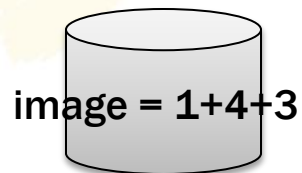
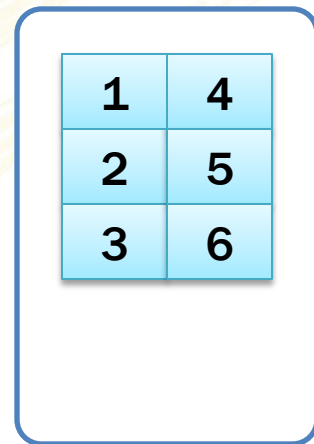
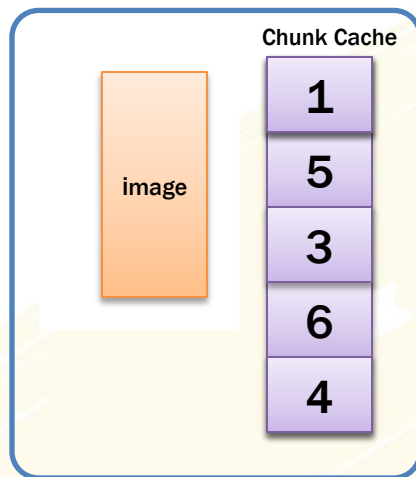


glance



backend storage

Download:



Metadata Database

调整规格优化

- 社区调整规格操作的问题
 - 调整规格 (**Resize**) 操作时间过长，代价高
 - 安全性：配置让宿主机相互免密码ssh登陆访问
- 原因分析
 - **Resize**流程：将Qcow2格式镜像Base只读部分和Cow修改部分**合并**成新镜像，再传输新镜像到目的宿主机
 - 使用rsync shell工作模式，不安全
- 优化
 - 不合并产生新镜像，只拷贝传输Cow部分
 - 确保resize中传输新镜像的流量务必通过内网
 - 使用rsync daemon模式

监控报警和实例存储配额

- 监控报警

- 监控：查看云主机运行状态，如CPU利用率、内存占用量、网络带宽流量、磁盘分区使用等13个指标
- 报警：状态维度报警，设置报警规则（报警项、聚合区间、报警阈值）和报警组（监控人员、通知方式）

- 实例存储配额

- 实例存储：系统提供云主机使用的临时块存储（系统卷和临时卷）
- 实例存储配额：控制用户实例存储使用量，避免资源耗尽

遇到问题及解决

- 宿主机高负载下创建云主机经常TimeOut
 - 原因：Nova-compute服务单进程运行，计算请求处理/定时任务/心跳三类任务通过协程处理，定时任务时间长，其他任务处理被阻塞
 - 方案：Fork Nova-compute 服务进程，将三类任务放到不同进程处理，提高处理并行性
- 云主机创建时注入文件不成功
 - 原因：挂载Windows镜像时报warning，注入文件后卸载NBD流程没执行，NBD设备被占满
 - 方案：修改文件注入流程，已被社区合并到主线

遇到问题及解决

- **Conntrack表满了，云主机网络中断**
 - 云环境下宿主机为云主机进行NAT网络路由，须使用conntrack表跟踪记录所有网络连接状态
 - 原因：宿主机网络连接数超过默认值65536，内核丢包
 - 方案：将参数net.ipv4.netfilter.ip_conntrack_max改大
- **节点重启后，Conntrack表配置参数不生效**
 - 原因：节点重启执行sysctl -p时，nf_conntrack_ipv4还没加载，导致Conntrack表配置无法生效
 - 方案：提升nf_conntrack_ipv4模块的加载顺序

遇到问题及解决

- 当云平台宿主机故障时，应用如何实现高可用？
 - 利用OpenStack中可用域（Availability Zone）功能，平台分为多个可用域
 - 云主机从属某个可用域，不同可用域云主机位于不同宿主机
 - 将应用部署在多个可用域中，避免应用单点失效
- 云主机中应用程序无法绑定浮动IP
 - 打开云主机系统内核参数ip_nonlocal_bind
- 应用无法判断云主机OS是否启动完毕
 - 新增API以查看云主机OS状态

工作模式

- 开发模式与社区类似，使用jira/git/gerrit/jenkins工具链
- 实现自动构建发行版，一天可发布多次

