

校园私有云建设实践

以PKU CLab为例

孙远航

PKU CLab, Why?

想法

- 契机：23年参加开源软件论坛，听中科大的云介绍
 - 云并非高深莫测，学生也能搞
- 动机：我当时个人电脑是Mac，但需要运行windows负载
 - 无论搞出来有没有用，最终我肯定会用
- 现实需要：有很多课程需要自己配虚拟机
 - 上课很痛苦，当助教更痛苦
 - MIT的Athena上个世纪就搞了，课程使用很方便
- 到了搞私有云的时候了！

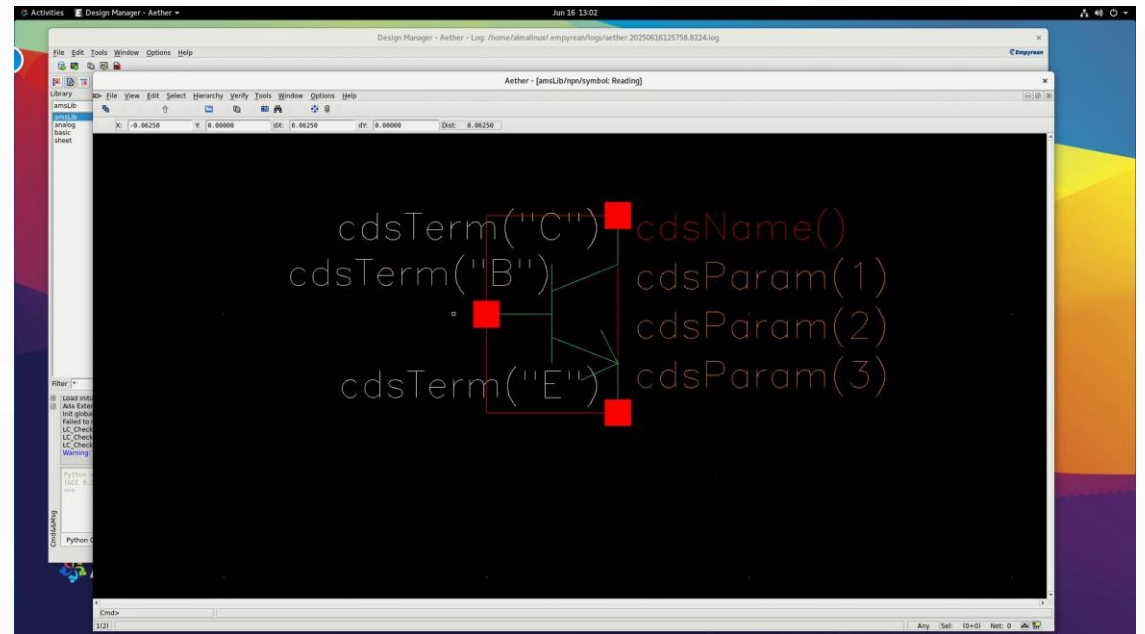
目前的情况

- 2024年8月上线，稳定运行
- 免费给全校所有师生发放4c4g资源+200G储存
- 1500+用户，来自48个院系单位，累计2K+虚拟机，30+操作系统镜像，活跃虚拟机1K左右
- 支持计算机系统导论、操作系统、模拟集成电路设计方法工具与流程、人工智能引论等20+课程的学生Lab
- 支持PKUHub、PKUInfo，音乐与数学课程教学平台等应用部署
- 支持高性能计算竞赛、信科AI Game等多种比赛

目前的情况



平台首页



支持国产EDA课程

目前的情况-硬件

- 硬件：
 - 5 * Intel Xeon Gold 5118 服务器
 - 3 * Intel Xeon Platinum 8160 服务器
 - 2 * AMD Epyc 7773X 服务器
 - 希捷2U24 SSD盘柜, 80T可用容量
 - 自组Ceph储存, 368T裸容量
 - 100Gbps RoCE 交换机
- 物理核心数: 520, 内存总量: 5280GB

目前的情况-软件

- 几乎全部基于开源实现
- 集群管理：MAAS+Ansible
- 云软件：开源版OpenStack，维护了非常非常小的修改
- 储存：开源版Ceph
- 前端：基于开源版skyline修改+重写登录页面
- 认证：接入北京大学IAAA认证，实现一键登录

一波三折的开发过程

看似顺利的开始

- 23.02.17: 做0th HPCGame技术分享的时候讨论了搞“北大云”的可能性
- 23.04.05: 做了云的初版架构设计，招募了十几人的核心小组
- 23.04.26: 找了樊春老师，拿到了四台测试用服务器

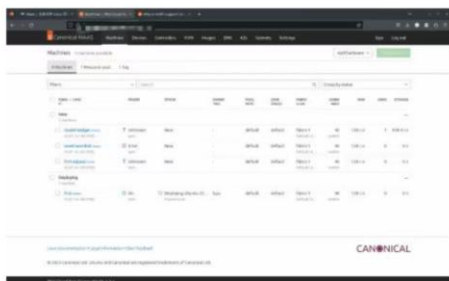
看似顺利的开始



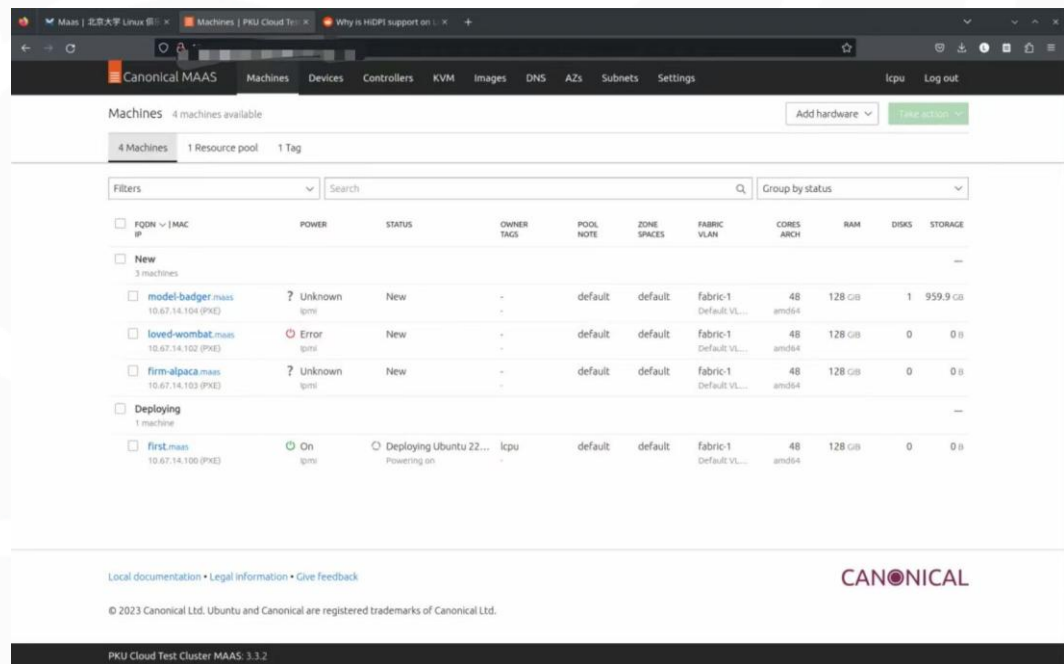
孙远航

多年以后，syh回想起北大云的时候，应该不会忘记组建起原始集群的那个夜晚（失败模仿）

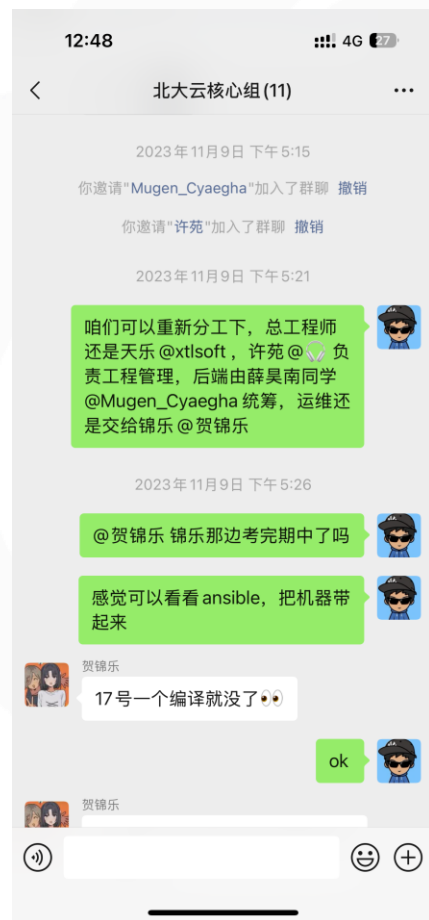
#定



2023年4月29日 凌晨 2:30



但是长期没有任何进展

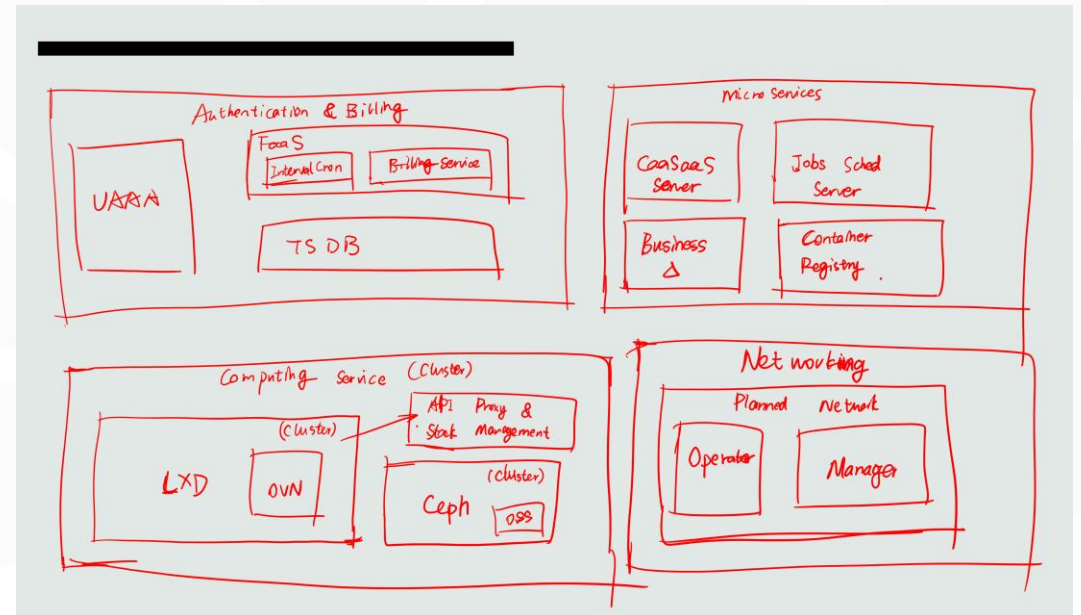
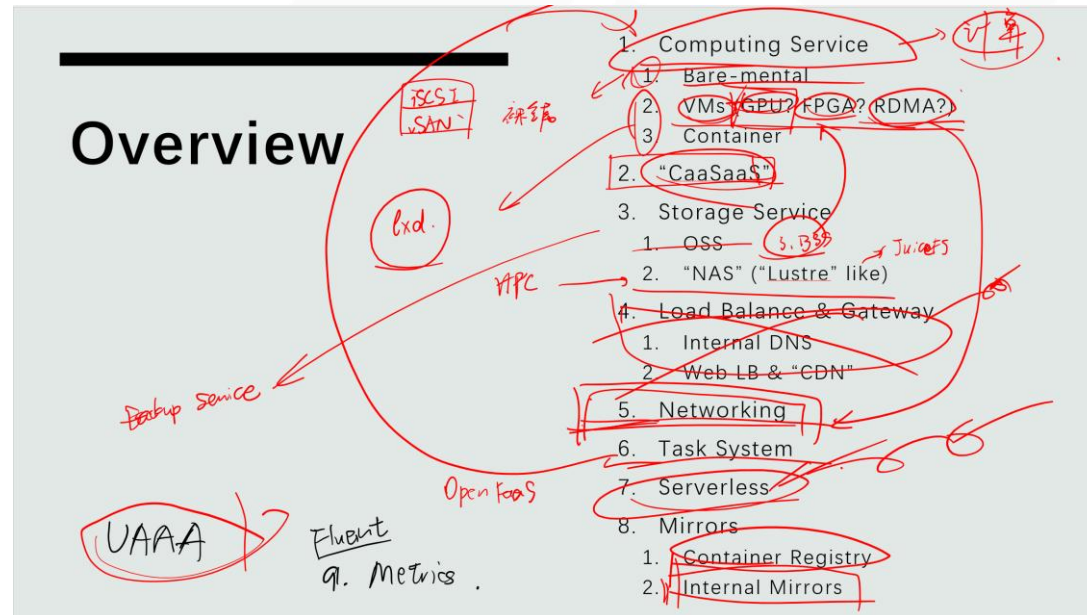


- 2023.7: 宣称完成了某个组件的设计（实际并没有）
- 2023.11: 宣称要引入工程管理和工程统筹（实际并没有左右）
- 为何呢？

过度设计、过早优化、过于自信

过度设计

- 自认为是优秀架构师的大一同学（虽然目前确实是），提出了完全重写的云架构：



过早优化

- 为了解决上面架构中需要管理特别多对象和消息的问题，抽象除了“Entityd”组件（类似于消息队列的作用，但会负责管理更多对象）
- 这个组件是性能热点，又是实现难点，而且全组只有一个人知道应该怎么写
- 从23年四月到23年七月，一行可用代码都没有



- 架构师没有动静，我尝试写了一下这个组件的文档，完全不知道如何实现

过于自信

- 没有任何同学知道云是怎么回事
- 各种组件也从来没有研究过
- 负责储存的同学花了两个月才成功部署了读写速度50MB/s的Ceph
- 宣称一周就能写的核心组件拖了六个月，目前也还没写好
- 重新造了一些没有用的轮子，选用了firecracker作为云的后端，但是功能太弱很多需要手写

一些教训

- 面对复杂系统时，派出再多小白进行研究都是不会有任何进展的，需要找到成长路径
- 一个系统首先要能用，然后再谈其他
- 尽量不要重复造轮子
- 过早优化是万恶之源
- 不要试图绕过“是什么”，先想“怎么做”，甚至是直接动手

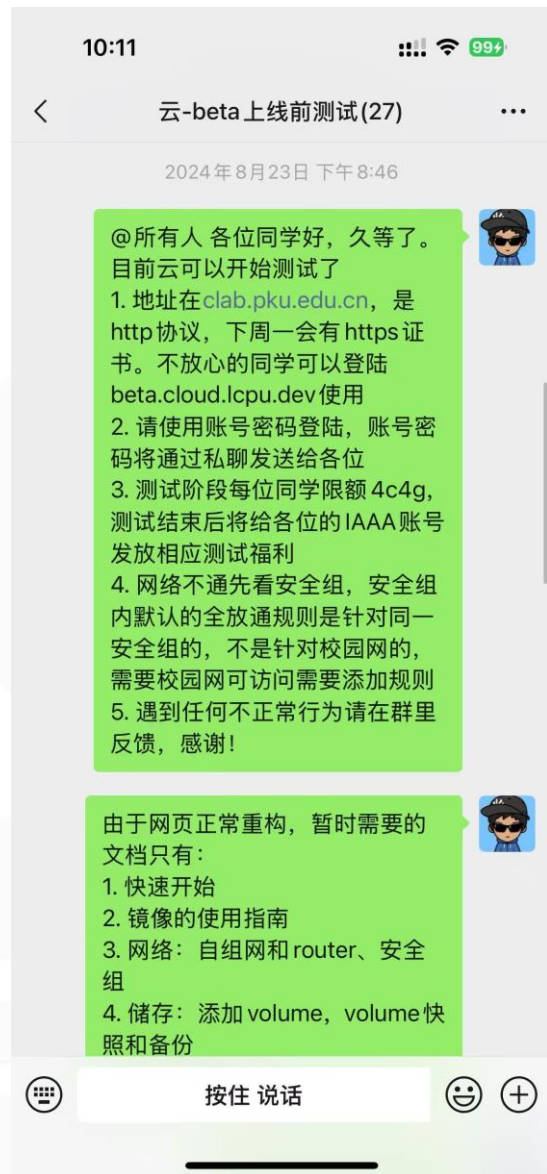
基于OpenStack部署

开工!

- 24年初，在尝试从头开发无果后，转向部署openstack方案
- 先是尝试基于k8s部署，但并不熟悉k8s->遇到各种bug，没有进展
- 接着基于kolla (openstack ansible) 部署，但该项目在rocky上bug很多，同时因为不熟悉openstack，无法理解几千行的配置是什么意思
- 最后选择使用裸机部署+手写ansible playbook的方式，踩了很多坑，终于上线

上线和测试

- 24年5月校庆假期，完成openstack部署，成功接入校园网，开始alpha测试
- 24年8月暑假，完成前端设计，接入统一认证，beta上线
- 24年秋季学期在ICS课程投入使用，安全度过压力测试
- 25年春学期增加4台服务器，新增EDA适配



一些教训

- 面对复杂系统时，试图简化流程的方法通常不会降低复杂度
- 配置文件应该使用git追踪，尽量让部署可复现
- 搞云一定要有和云适配的硬件，最好上下游都能自己掌握
 - 尝试接入校园网时，发现无法发DHCP包，最后关闭DHCP snooping就一切正常
 - 但是这个交换机我没有权限，debug了整整一周

CLab v2和未来展望

CLab v2

- 我们仍然没有放弃自研云项目的规划
- 目前上线的云让我们搞清楚了云是什么
- 在学校的广泛使用积累了很多用户和开发者
- 认识了现有openstack云的缺陷
- 25年4月，开始了CLab v2工程

CLab v2

- 更好的部署和日志收集
 - 切换到k8s, 使用kubevirt作为虚拟化引擎: PoC完成
- GPU云环境呼声较高, 高效利用有限GPU资源挑战大
 - Remote GPU: PoC完成
- 解决虚拟化RDMA问题, 助力HPC上云
 - vRDMA: PoC完成
- 高性能存储和备份系统
 - 基于DOAS优化全闪文件系统: 正在逐步推进
- 解决裸金属虚拟化问题
 - 基于DPU开发裸金属虚拟化系统
- 更好的前端、更好的用户体验

CLab v2

- 团队目前有16位同学，分前端、系统、内核、运维和自动化等方向
- Mentor制，有人带，有劳务（虽然不多），有线下实习
- 欢迎大家加入我们！
- 可以发邮件到sunyh AT stu.pku.edu.cn

感谢大家聆听！