



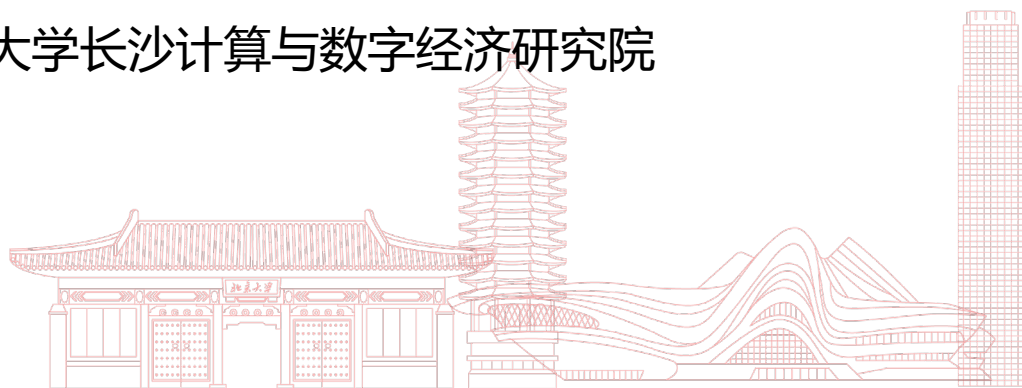
服务科学计算的系统平台软件 研发进展

樊春

北京大学计算中心

北京大学高性能计算校级公共平台

北京大学长沙计算与数字经济研究院



content

目录

- 一 | 北大超算平台建设与服务**
- 二 | 算力中心门户平台**
- 三 | 算力资源调度**
- 四 | 未名小蒜**



content

目录

一 | 北大超算平台建设与服务

二 | 算力中心门户平台

三 | 算力资源调度

四 | 未名小蒜



1963

1963年，当北京大学拥有第一台计算机时，就在全国高校率先成立计算中心，组建专门技术团队服务全校师生。

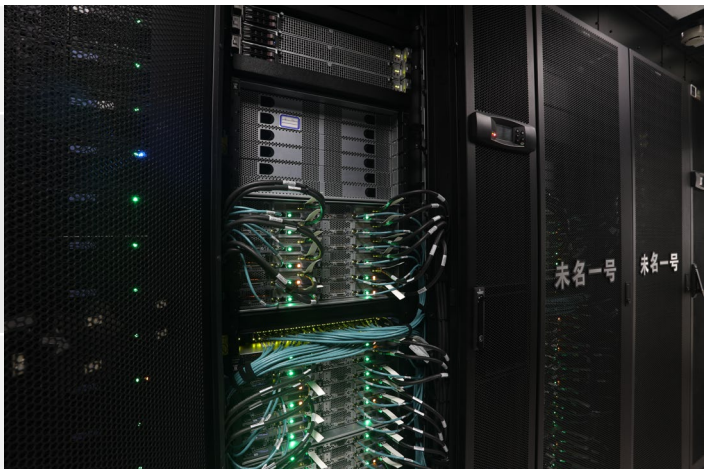
2001

于2001年和2005年购置两台计算能力分别为960亿次和1.638万亿次的高性能计算机，服务全校教学科研，取得良好效益。

2016

2016年成立北京大学高性能计算校级公共平台。

年代	1960s	1970s	1980s	2000s
计算机	103电子管计算机	6912机	Dps8/52 Elxsi6400	RS6000Sp HP dl360g4
速度	30-2000 次/秒	13万-15万次/秒	120万次/秒 600万次/秒	960亿次/秒 1.6万亿次/秒



2018年1月3日北京大学高性能计算校级公共平台正式揭牌启用。

“未名一号”、“未名教学一号”和“未名生科一号”等多套集群陆续投入运行，在数学、深度学习、大气海洋环境、新能源新材料、天文地球物理、生物医药健康等领域提供高性能科学与工程计算服务。

集群

5套

节点

755

核心

33584

计算峰值

4.39PFLOPS

存储容量

15PB

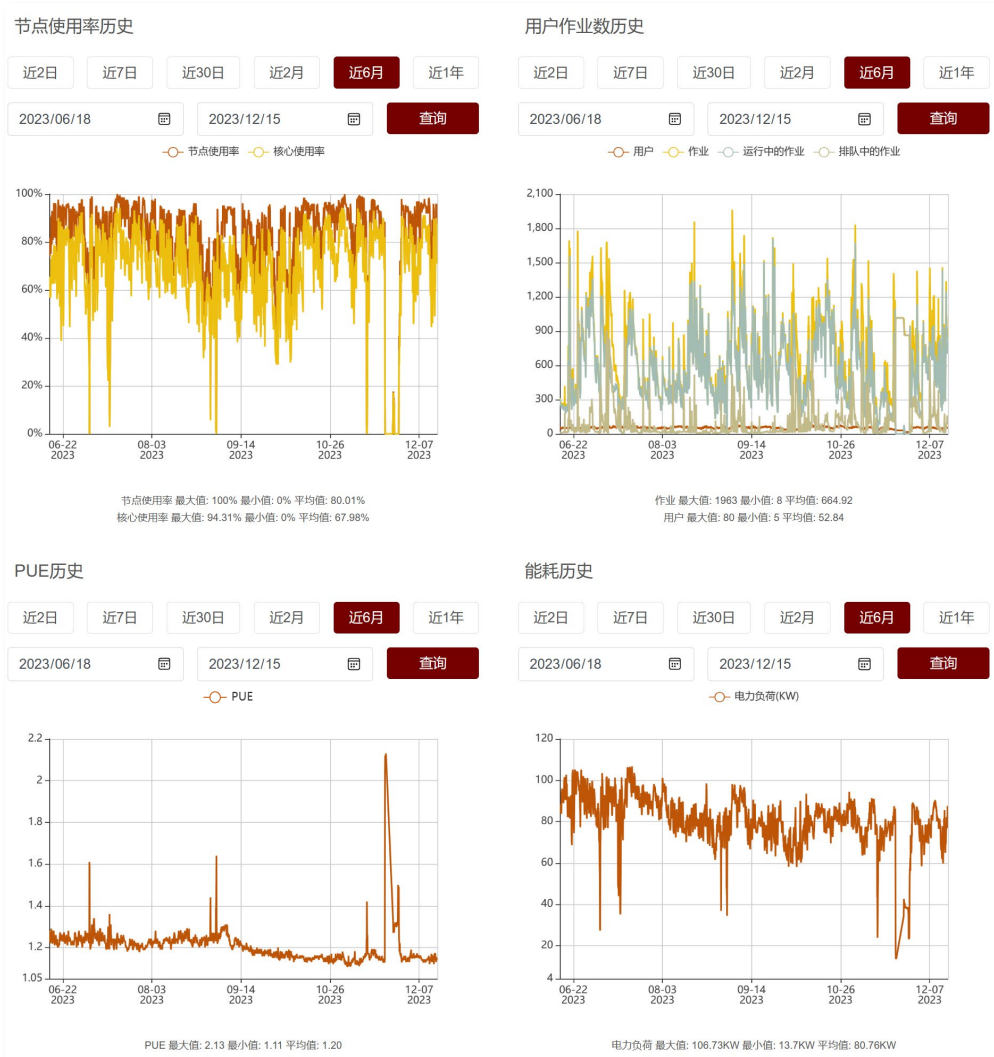
PUE值

1.1

<https://hpc.pku.edu.cn>



未名一号——国内首个大规模温水水冷集群



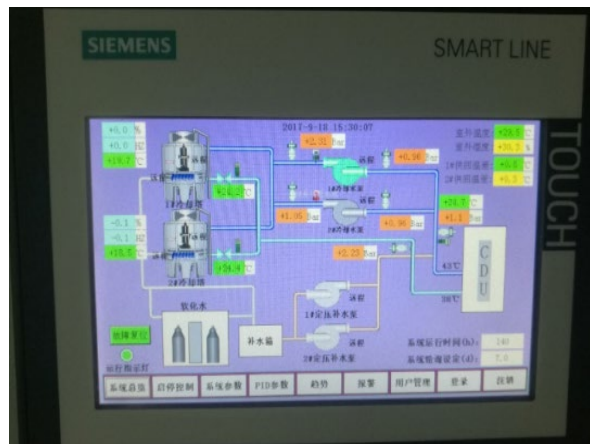
集群利用率和能耗历史数据



室外冷却塔



水冷设备间

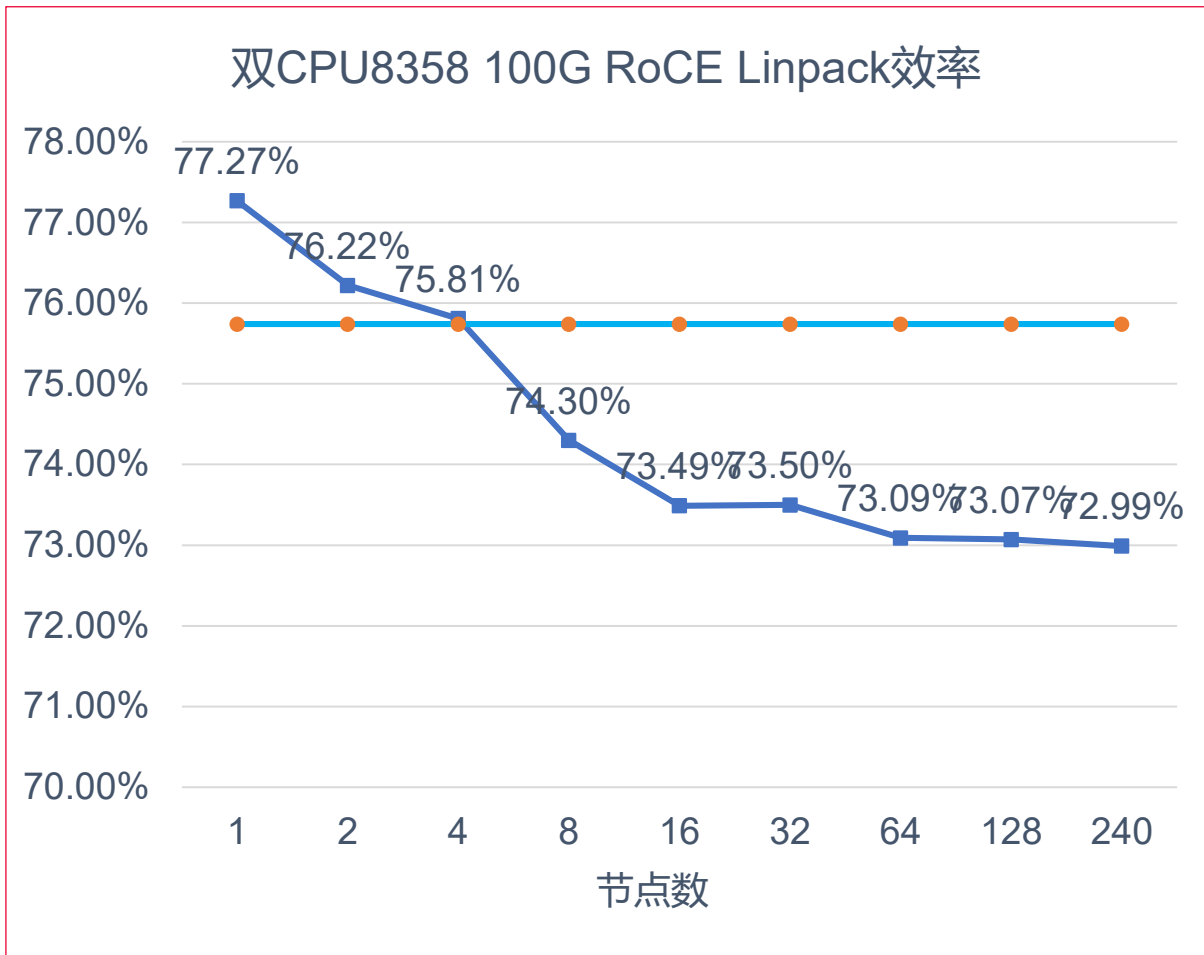
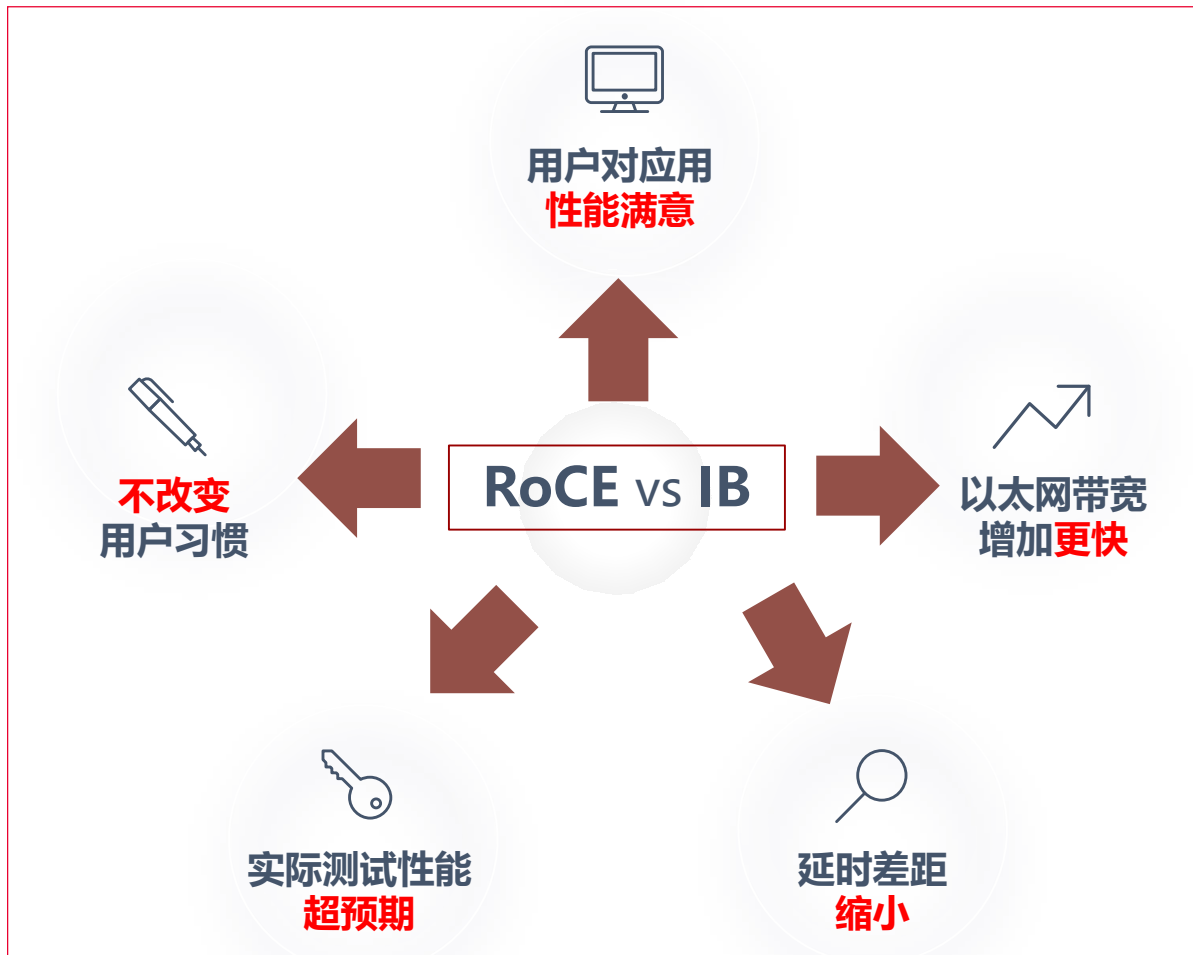


西门子自控系统

未名二号——国内高校首个国产100G无损高速网

240节点HPL测试，线性度超预期

基于超融合以太RoCE

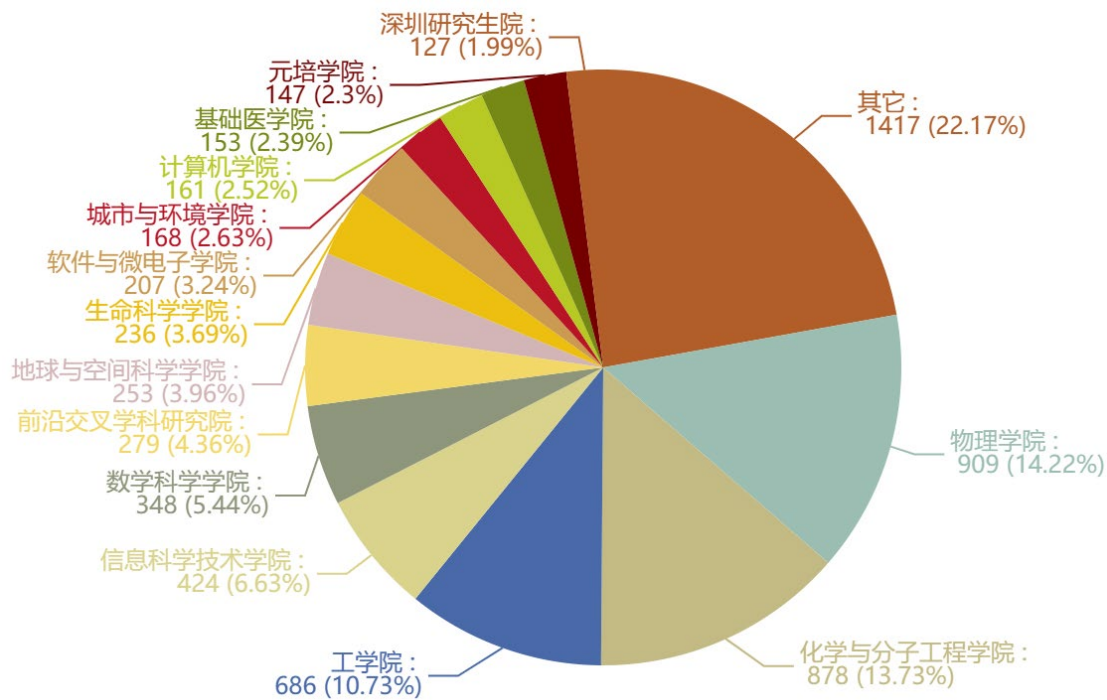


6393人
师生用户

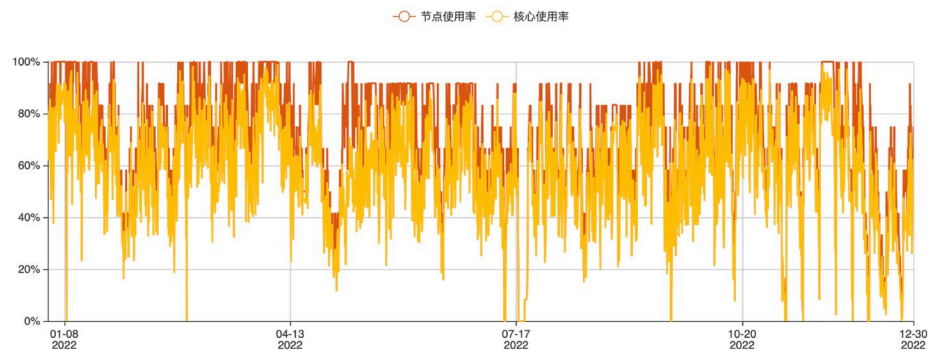
105个
院系单位

20位
两院院士

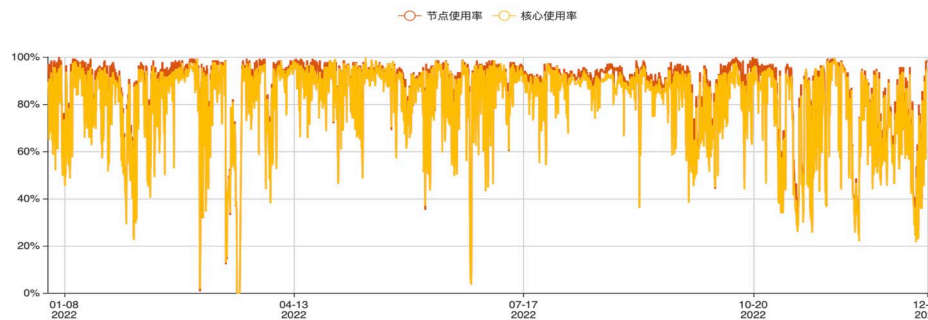
6位
高校校长



用户来源分布图



未名一号
年均使用率
87.58%



未名生科一号
年均使用率
81.30%

平台资源使用状况

用户代表性成果

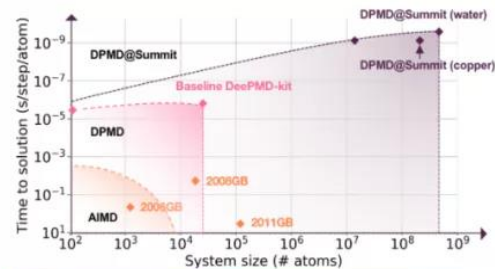
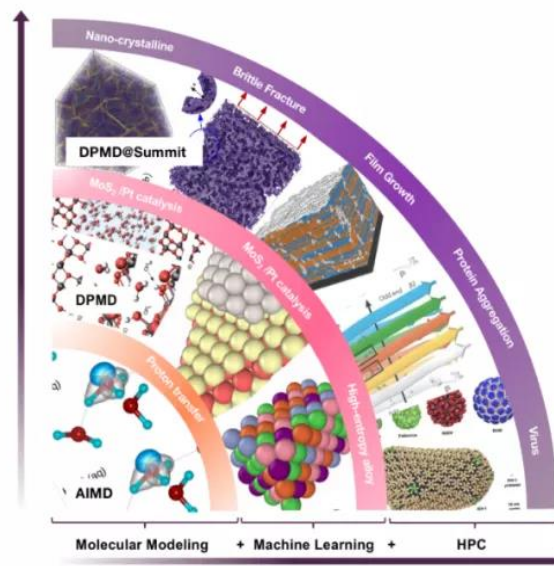
2020年度国家技术发明**一等奖**
超高清视频多态基元编解码关键技术

ACM GORDON BELL PRIZE - WINNER
Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning

2020年**戈登贝尔奖**
基于深度学习的分子动力学模拟



分子建模+人工智能+高性能计算



Time and size scales required by important Problems

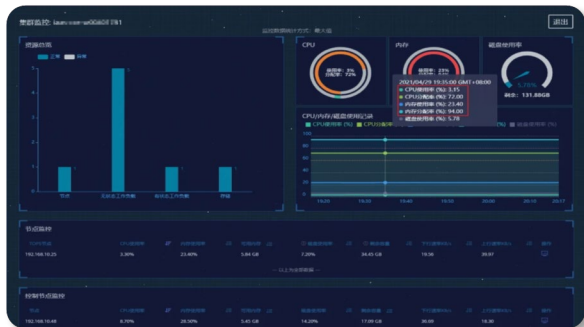
Problem	Time span [ns]	System size [#atom]
Droplet coalescence	~10	~1e+8
Dynamic fracture	~0.1	~1e+8
Strength of nanocrystalline metal	~0.01	~1e+6
Heterogeneous aqueous interfaces	~100	~1e+6

content

目录

- 一 | 北大超算平台建设与服务**
- 二 | 算力中心门户平台**
- 三 | 算力资源调度**
- 四 | 未名小蒜**





运营管理难

- **管理困难**，缺乏统一标准的管理模式，各算力中心需要各自制定管理政策
- **部署困难**，超算集群部署需要各类软硬件配置，缺乏开箱即用的管理平台
- **运营困难**，运维服务人员短缺，需要自动化工具提高运营效率

01



用户使用难

- 算力终端用户多元化，无法适应基于命令行的传统超算集群使用模式，更习惯图形化、鼠标化的操作
- 可视化交互式应用配置复杂，用户使用门槛高

02



算力融合难

- 各个算力中心计算**资源独立**，无法互联
- 各算力中心管理模式不同，使用模式不同，**难以融合**
- 各算力中心使用情况不同，算力资源**闲忙有别**，造成算力浪费

03

建立**面向算力网络的算力中心门户和管理平台SCOW**，通过简化集群软件部署流程、统一平台管理模式、降低用户使用门槛，实现算力中心资源易管理、易使用的目标，提高算力资源使用效率，满足算力中心的管理和维护需求。面向算力网络，构建标准化的平台接口，支撑算力网络平台建设。

SCOW = 平底船 -> 未名湖上的石舫



SCOW = Super COW



SCOW = Super Computing On Web



SCOW——开源中立，开放自主

- 项目由北京大学高性能计算平台、北京大学长沙计算与数字经济研究院共同研发

自主完成系统核心逻辑功能开发，具有自主知识产权；

- 项目采用木兰宽松协议开源

源码地址：<https://github.com/PKUHPC/SCOW>

文档地址：[https:// icode.pku.edu.cn/SCOW](https://icode.pku.edu.cn/SCOW)

试用地址：<https://hpc.pku.edu.cn/demo/scow>



标准化管理模式

SCOW管理平台提供了标准化的模式、模型，能够帮助新建的算力中心快速建立管理和运营制度。管理系统提供了各项管理功能，管理员和运营人员可以很容易在浏览器上实现管理团队的人员和资源，灵活分配人员权限和机时份额，封锁解封团队用户等精细化操作。

通用管理功能

标准化模式、模型



多集群统一管理



用户账户模型
(租户-账户-用户三级模型)



计费收费模型

管理系统多项功能



账户管理



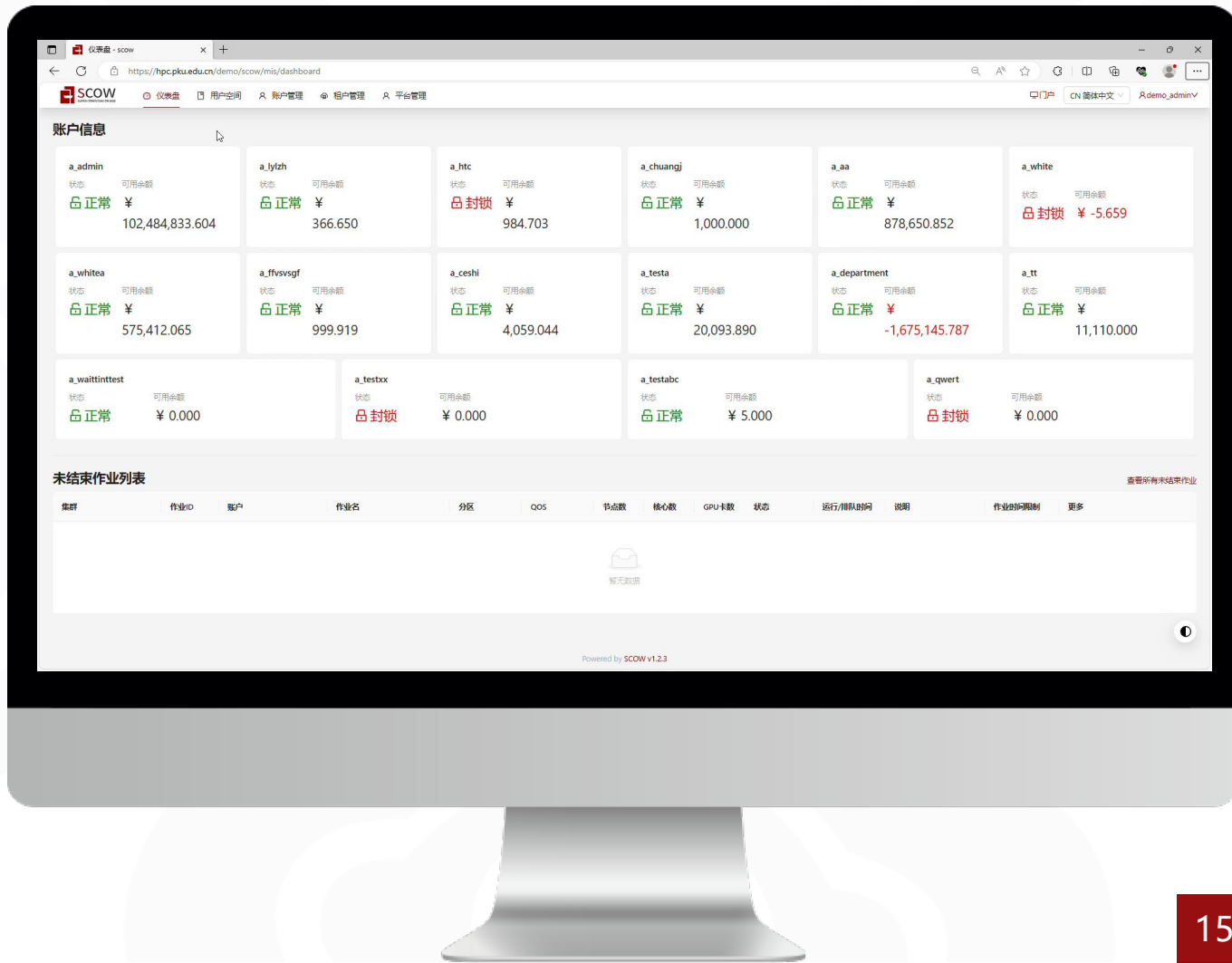
租户管理



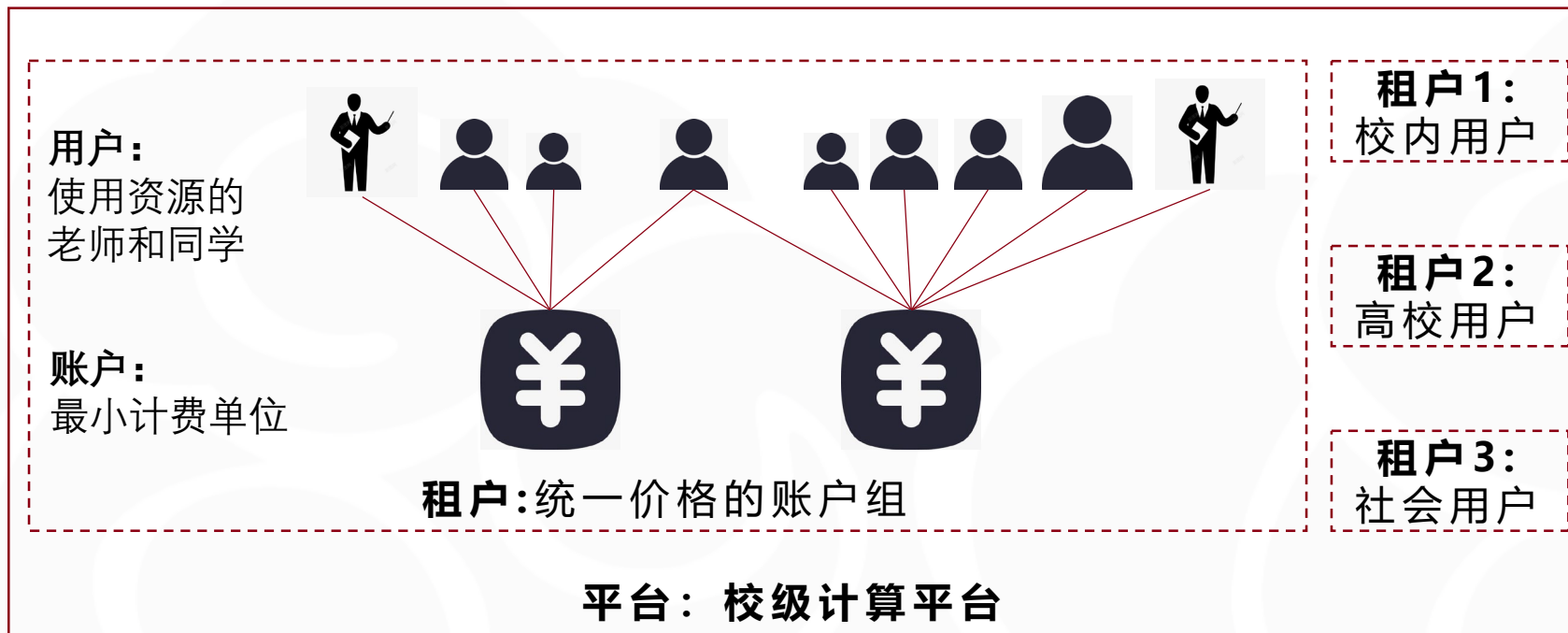
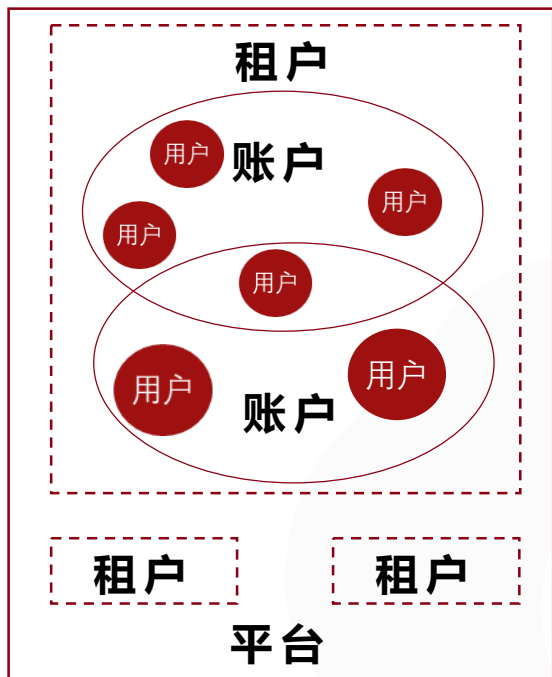
平台管理



财务管理



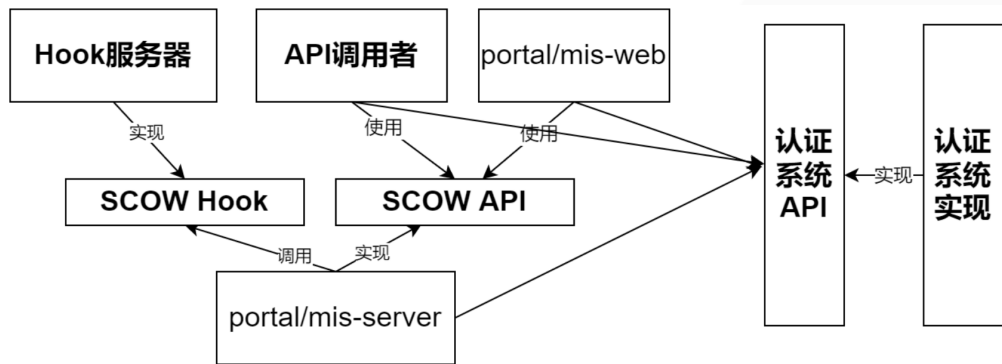
SCOW——人员管理逻辑



角色\权限	用户	账户	租户	平台
用户	登录集群使用资源			
拥有者		账户和拥有者绑定		
管理员		管理账户	管理租户	管理平台
财务人员			账户财务管理	租户财务管理

SCOW——全面的管理能力





建立SCOW API和Hook机制+自定义认证系统

API		Hook	
查询平台租户列表	查询租户扣费记录	封锁账户	解封账户
查询账户列表	查询账户当前余额	创建账户	创建用户
账户扣费	账户充值	在账户中封锁用户	在账户中解封用户
将用户加入账户	将用户移出账户	添加用户	账户扣费
查询租户信息	查询作业信息	租户扣费	作业导入
查询充值记录	作业扣费

后端提供101个API+11个Hook



自定义导航栏



微前端架构

降低使用门槛

在SCOW门户平台，超算用户无需了解和配置SSH、VNC、命令行等技术，直接在浏览器上就可以使用超算集群。不同于传统超算基于命令行的使用模式，平台门户系统提供基于web页面的各项功能，极大降低了用户使用门槛，让Linux小白用户也能顺利提交作业。

基于Web的多项功能



Web SSH连接



Web 远程桌面



Web 文件管理



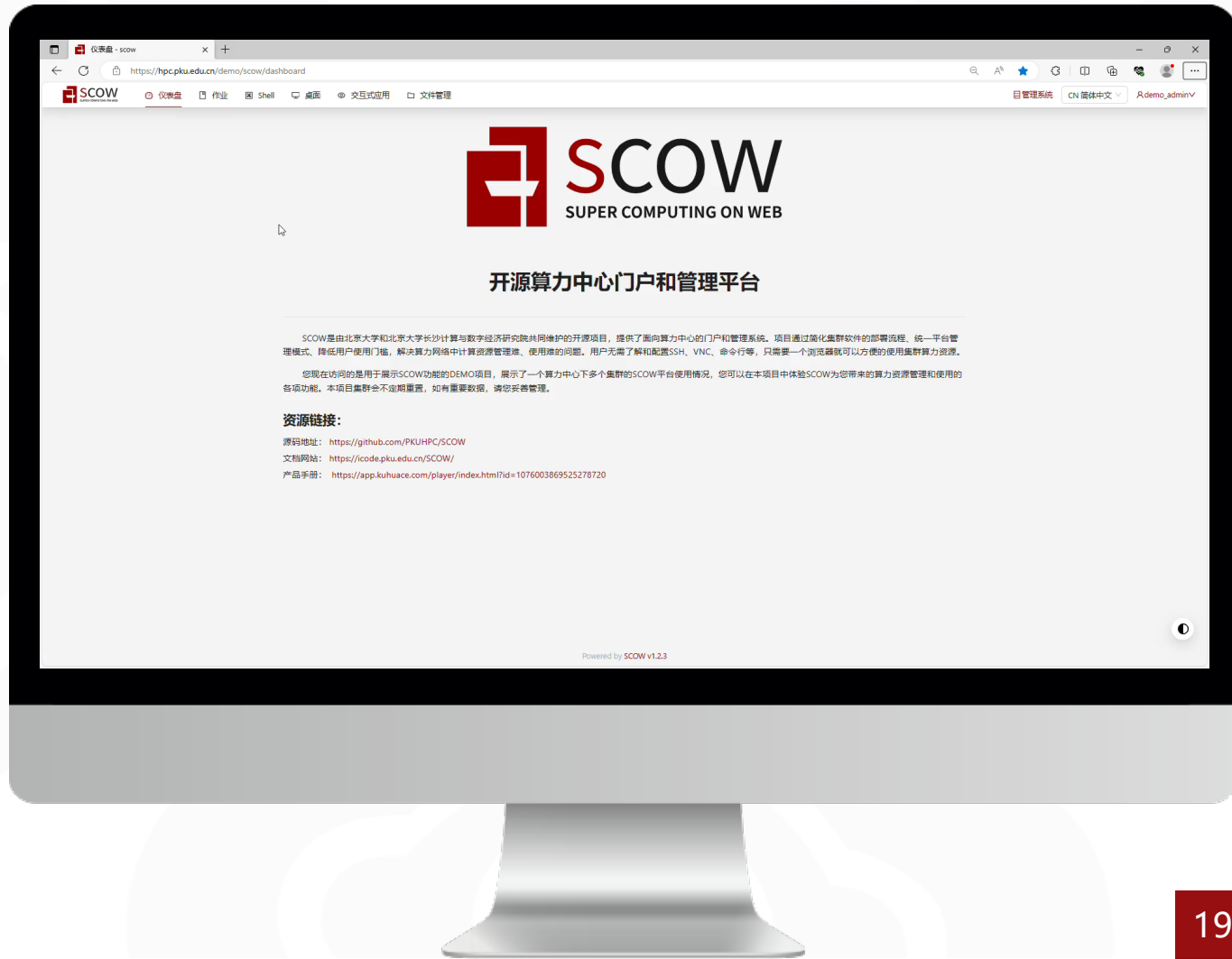
Web 跨集群数据传输



Web 交互式应用



Web 作业管理



提交作业

1、门户系统作业模块提交

SCOW 仪表盘 作业 Shell 桌面 交互式应用 文件管理 管理系统 CN 简体中文 demo_admin

未完成的作业
所有作业
提交作业
作业模板

提交作业

* 集群: 演示集群一 * 作业名: job-20231211-212754

* 命令: 1 此处参数设置的优先级高于页面其它地方, 两者冲突时以此处为准

* 账户: a_aa * 分区: compute * QOS: normal

* 节点数: 1 * 单节点核心数: 1 * 最长运行时间: 30 分钟

* 工作目录: /data/home/demo_admin/scow/jobs/job-20231211-212754 选择 * 标准输出文件: job.%j.out * 错误输出文件: job.%j.err

总节点数: 1 总CPU核心数: 1 总内存容量: 3.66 GB

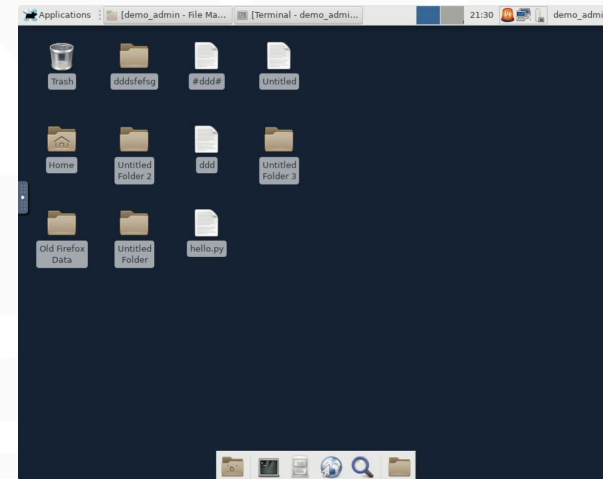
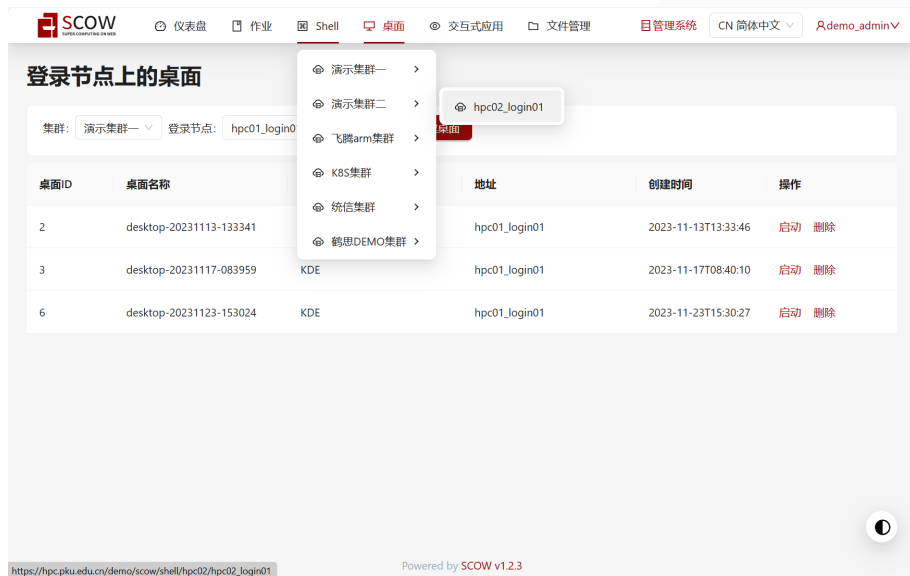
备注:

保存为模板

提交

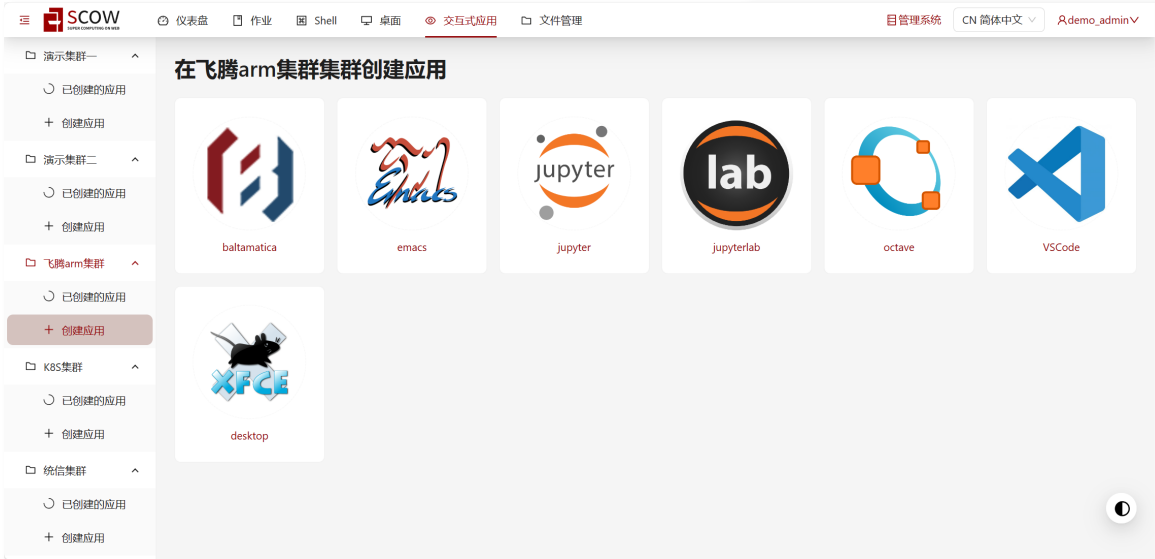
提交作业

2、通过shell和远程桌面连接到登录节点提交



提交作业

3、交互式应用中提交作业



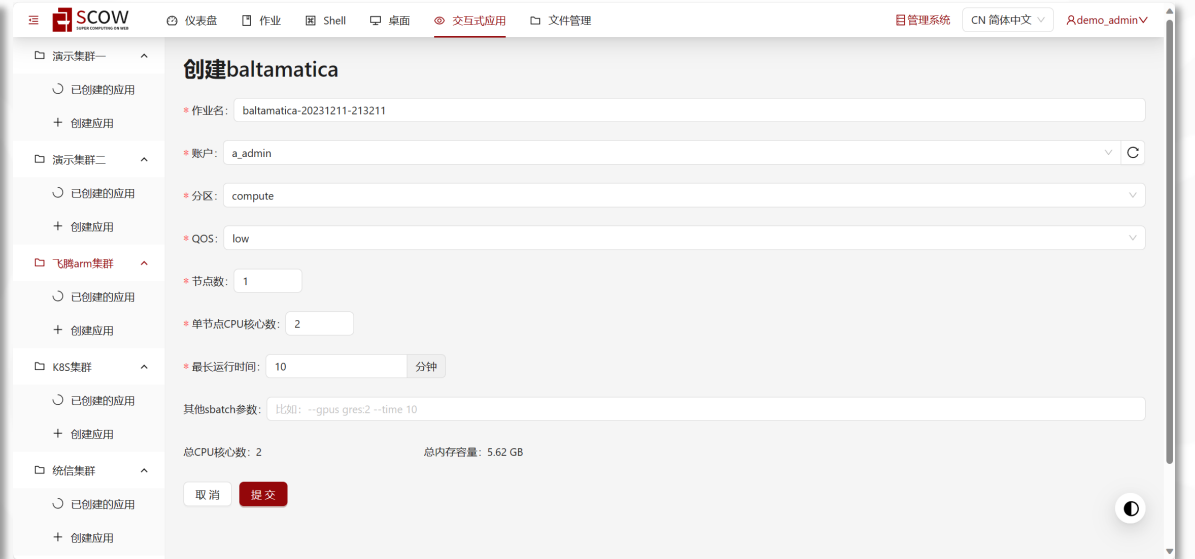
在飞腾arm集群创建应用

已创建的应用

- baltamatica
- emacs
- jupyter
- jupyterlab
- octave
- VSCode

飞腾arm集群

- desktop



创建baltamatica

作业名: baltamatica-20231211-213211

账户: a_admin

分区: compute

QOS: low

节点数: 1

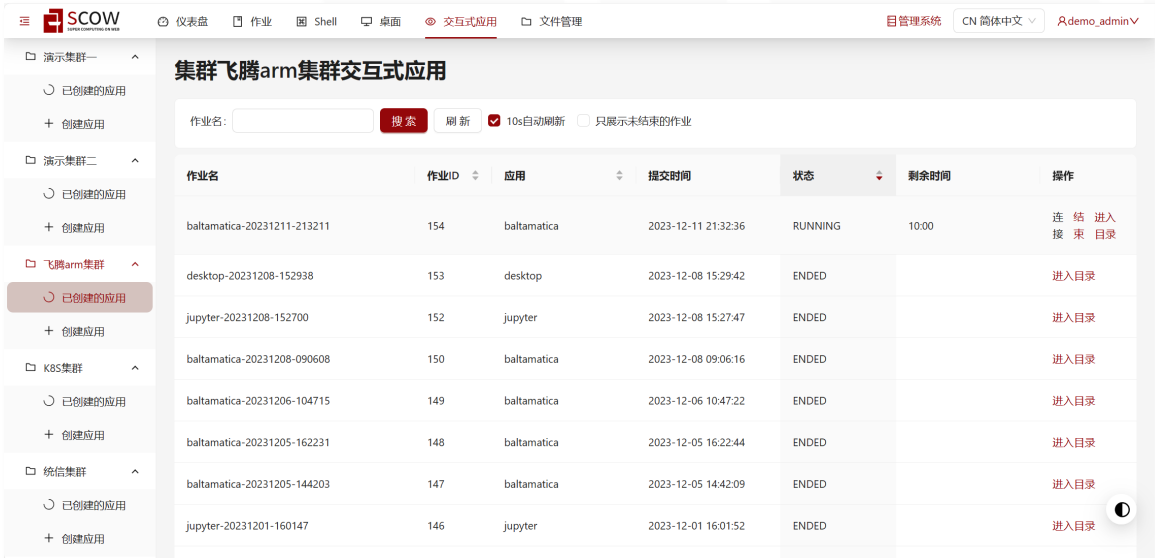
单节点CPU核心数: 2

最长运行时间: 10 分钟

其他batch参数: 比如: --gpus gres2 --time 10

总CPU核心数: 2 总内存容量: 5.62 GB

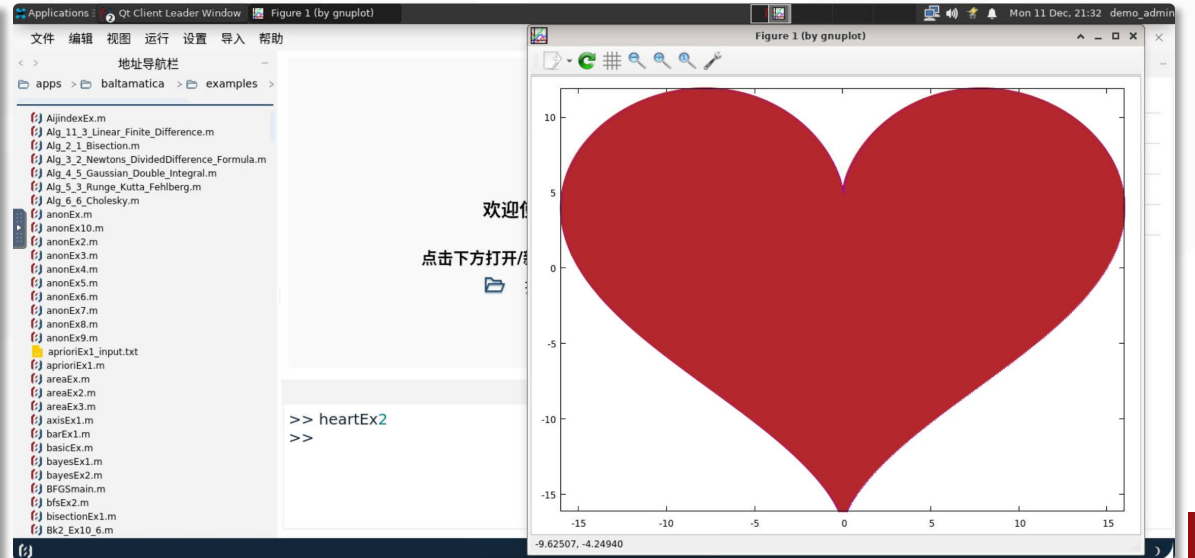
取消 提交



集群飞腾arm集群交互式应用

作业名: 搜索 刷新 10s自动刷新 只展示未结束的作业

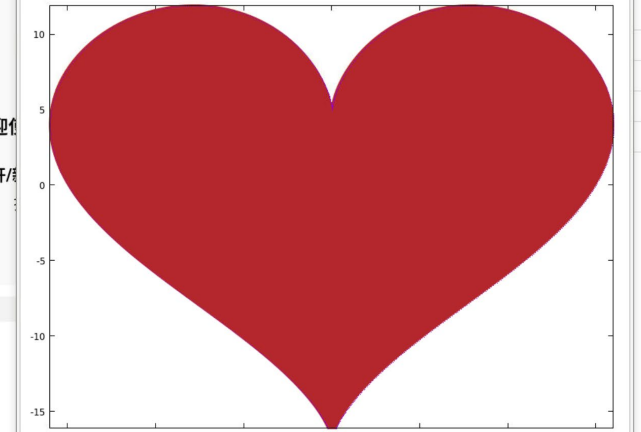
作业名	作业ID	应用	提交时间	状态	剩余时间	操作
baltamatica-20231211-213211	154	baltamatica	2023-12-11 21:32:36	RUNNING	10:00	连续进入 结束目录
desktop-20231208-152938	153	desktop	2023-12-08 15:29:42	ENDED		进入目录
jupyter-20231208-152700	152	jupyter	2023-12-08 15:27:47	ENDED		进入目录
baltamatica-20231208-090608	150	baltamatica	2023-12-08 09:06:16	ENDED		进入目录
baltamatica-20231206-104715	149	baltamatica	2023-12-06 10:47:22	ENDED		进入目录
baltamatica-20231205-162231	148	baltamatica	2023-12-05 16:22:44	ENDED		进入目录
baltamatica-20231205-144203	147	baltamatica	2023-12-05 14:42:09	ENDED		进入目录
jupyter-20231201-160147	146	jupyter	2023-12-01 16:01:52	ENDED		进入目录



Qt Client Leader Window Figure 1 (by gnuplot)

```
>> heartEx2
>>
```

Figure 1 (by gnuplot)



欢迎...
点击下方打开/

SCOW SUPER COMPUTING ON WEB

仪表盘 作业 Shell 桌面

管理系统 CN 简体中文 demo_admin

在K8S集群创建应用

- 演示集群一
 - 已创建的应用
 - 创建应用
- 演示集群二
 - 已创建的应用
 - 创建应用
- 飞腾arm集群
 - 已创建的应用
 - 创建应用
- 鹤思ARM集群
 - 已创建的应用
 - 创建应用
- K8S集群**
 - 已创建的应用
 - 创建应用**
- 统信集群
 - 已创建的应用
 - 创建应用
- 鹤思集群
 - 已创建的应用
 - 创建应用



ChatGLM



emacs



jupyter



jupyterlab



Qwen



VSCode

创建jupyterlab

Powered by SCOW v1.4.3

https://hpc.pku.edu.cn/demo/scow/apps/k8s/create/jupyter...

通义千问 Qwen-7B-Chat Bot

This WebUI is based on Qwen-7B-Chat, developed by Alibaba Cloud. (本WebUI基于Qwen-7B-Chat打造, 实现聊天机器人功能。)

Qwen-7B | Qwen-7B-Chat | [Github](#)

Qwen-7B-Chat

你是谁

我是阿里巴巴达摩院开发的一款超大规模语言

Input

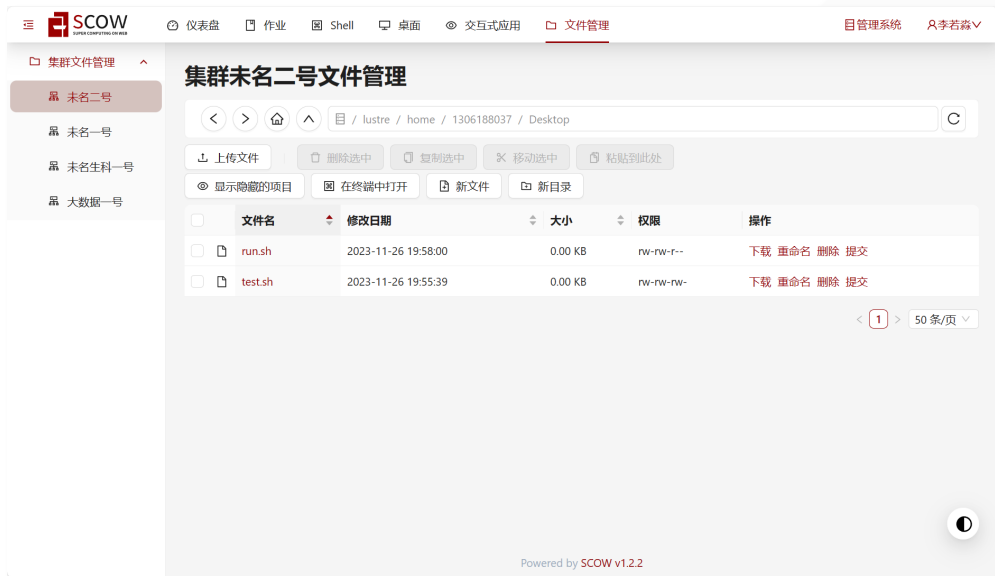
Clear History (清除历史) Submit (发送) Regenerate (重试)

Note: This demo is governed by the original license of Qwen-7B. We strongly advise users not to knowingly generate or allow others to knowingly generate harmful content, including hate speech, violence, pornography, deception, etc. (注: 本演示受Qwen-7B的许可协议限制。我们强烈建议, 用户不应传播及不应允许他人传播以下内容, 包括但不限于仇恨言论、暴力、色情、欺诈相关的有害信息。)

Use via API · 使用Gradio构建

提交作业

4、文件管理中提交作业



SCOW 仪表盘 作业 Shell 桌面 交互式应用 文件管理 目管理系统 李若森

集群文件管理

未名二号

未名一号

未名生科一号

大数据一号

集群未名二号文件管理

目 / lustre / home / 1306188037 / Desktop

上传文件 删除选中 复制选中 移动选中 粘贴到此处

显示隐藏的项目 在终端中打开 新文件 新目录

文件名	修改日期	大小	权限	操作
run.sh	2023-11-26 19:58:00	0.00 KB	rw-rw-r--	下载 重命名 删除 提交
test.sh	2023-11-26 19:55:39	0.00 KB	rw-rw-rw-	下载 重命名 删除 提交

1 / 50 条/页

Powered by SCOW v1.2.2



SCOW 仪表盘 作业 Shell 桌面 交互式应用 文件管理 目管理系统 李若森

集群文件管理

未名二号

未名一号

未名生科一号

大数据一号

集群未名二号文件管理

目 / lustre / home / 1306188037 / Desktop

上传文件 删除选中 复制选中 移动选中 粘贴到此处

显示隐藏的项目 在终端中打开 新文件 新目录

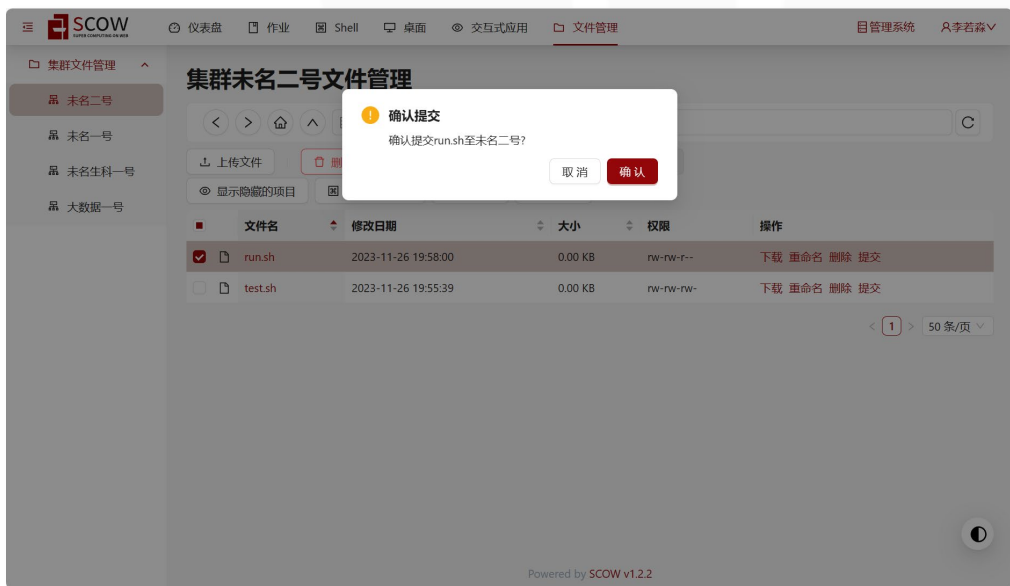
文件预览

```
run.sh
1 #!/bin/bash
2 #SBATCH -o job.%j.out
3 #SBATCH --partition=C064M0256G
4 #SBATCH -J vasp_job
5 #SBATCH -n 1
6 #SBATCH --ntasks-per-node=1
7 sleep 100
8
```

编辑

50 条/页

Powered by SCOW v1.2.2



SCOW 仪表盘 作业 Shell 桌面 交互式应用 文件管理 目管理系统 李若森

集群文件管理

未名二号

未名一号

未名生科一号

大数据一号

集群未名二号文件管理

目 / lustre / home / 1306188037 / Desktop

上传文件 删除选中 复制选中 移动选中 粘贴到此处

显示隐藏的项目 在终端中打开 新文件 新目录

确认提交

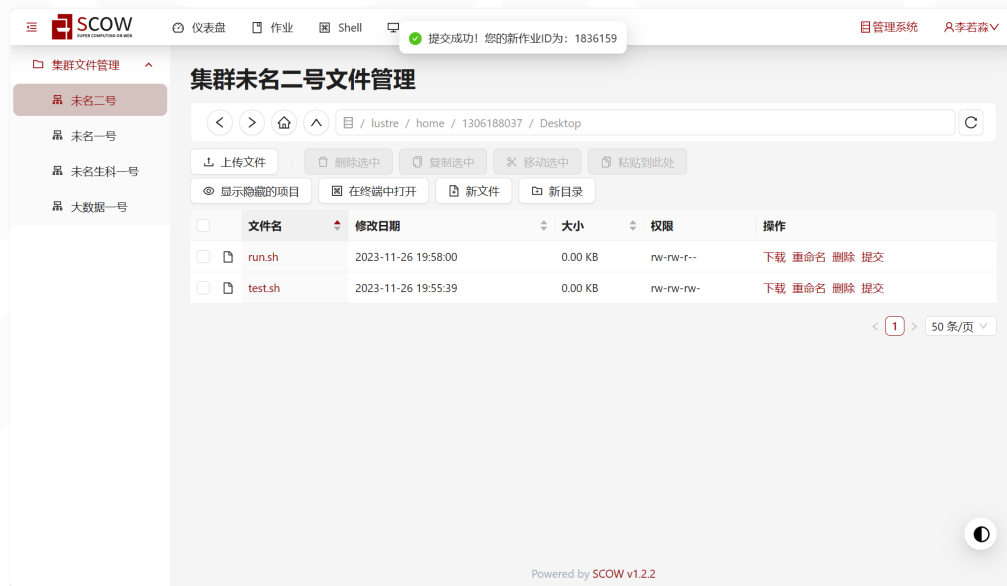
确认提交run.sh至未名二号?

取消 确认

文件名	修改日期	大小	权限	操作
<input checked="" type="checkbox"/> run.sh	2023-11-26 19:58:00	0.00 KB	rw-rw-r--	下载 重命名 删除 提交
<input type="checkbox"/> test.sh	2023-11-26 19:55:39	0.00 KB	rw-rw-rw-	下载 重命名 删除 提交

1 / 50 条/页

Powered by SCOW v1.2.2



SCOW 仪表盘 作业 Shell 桌面 交互式应用 文件管理 目管理系统 李若森

提交成功! 您的新作业ID为: 1836159

集群文件管理

未名二号

未名一号

未名生科一号

大数据一号

集群未名二号文件管理

目 / lustre / home / 1306188037 / Desktop

上传文件 删除选中 复制选中 移动选中 粘贴到此处

显示隐藏的项目 在终端中打开 新文件 新目录

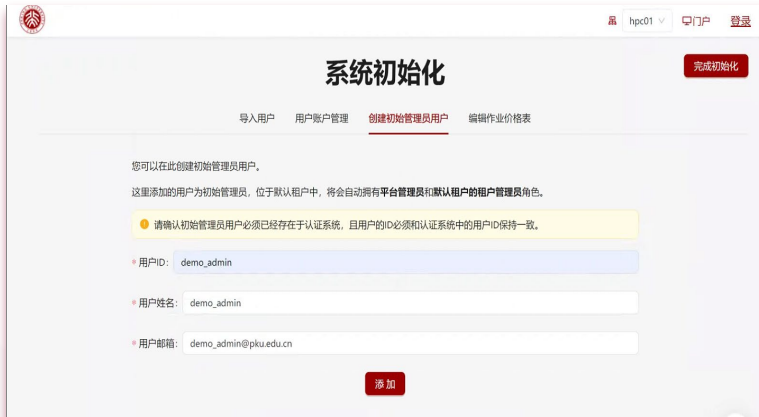
文件名	修改日期	大小	权限	操作
<input type="checkbox"/> run.sh	2023-11-26 19:58:00	0.00 KB	rw-rw-r--	下载 重命名 删除 提交
<input type="checkbox"/> test.sh	2023-11-26 19:55:39	0.00 KB	rw-rw-rw-	下载 重命名 删除 提交

1 / 50 条/页

Powered by SCOW v1.2.2


```
port: 8080
basePath: /demo/scow
image: ghcr.io/pkuhpc/scow/scow
imageTag: master
gateway:
  uploadFileSizeLimit: 20M
portal:
  novncClientImage: mirrors.pku.edu.cn/pkuhpc/novnc-client-docker:master
portMappings: {}
mis:
  dbPassword: mustchang3this
portMappings: {}
log:
  level: debug
pretty: false
fluentd:
  logDir: /var/log/fluentd/scow_logs
auth:
  portMappings: {}
```

```
[root@service scow-deployment]# ./cli compose ps
NAME                STATUS              IMAGE                                COMMAND                SERVICE      CREATED
scow-deployment-auth-1  Up 13 hours        ghcr.io/pkuhpc/scow/scow:master    "/.entrypoint.sh"    auth         13 hours
scow-deployment-db-1    Up 13 hours        mysql:8                              "docker-entrypoint.s..." db           13 hours
scow-deployment-gateway-1  Up 13 hours        ghcr.io/pkuhpc/scow/scow:master    "/.entrypoint.sh"    gateway      13 hours
scow-deployment-log-1    Up 13 hours        fluentd:vi.14.0-1.0                "tini -- /bin/entryp..." log          13 hours
scow-deployment-mis-server-1  Up 13 hours        ghcr.io/pkuhpc/scow/scow:master    "/.entrypoint.sh"    mis-server   13 hours
scow-deployment-mis-web-1  Up 13 hours        ghcr.io/pkuhpc/scow/scow:master    "/.entrypoint.sh"    mis-web      13 hours
scow-deployment-novnc-1    Up 13 hours        mirrors.pku.edu.cn/pkuhpc/novnc-client-docker:master "/docker-entrypoint..." novnc        13 hours
scow-deployment-portal-server-1  Up 13 hours        ghcr.io/pkuhpc/scow/scow:master    "/.entrypoint.sh"    portal-server 13 hours
scow-deployment-portal-web-1  Up 13 hours        ghcr.io/pkuhpc/scow/scow:master    "/.entrypoint.sh"    portal-web    13 hours
scow-deployment-redis-1    Up 13 hours        redis:alpine                        "docker-entrypoint.s..." redis        13 hours
```



01 编辑配置文件

配置部署的模块、集群基础信息、认证系统、交互式应用等信息。

02 编排与启动服务

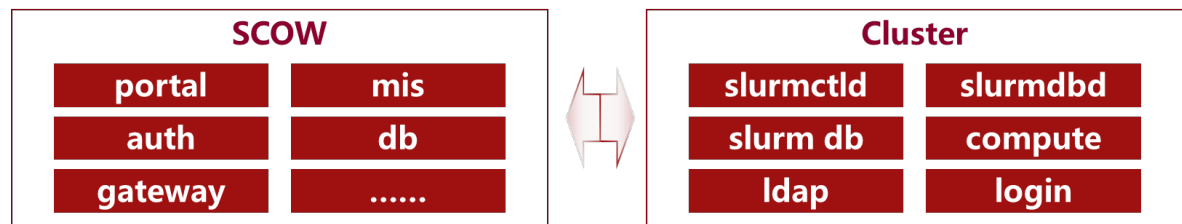
基于Docker Compose，使用自研SCOW-cli工具编排服务，一键拉取镜像，启动服务。

03 初始化管理系统

导入用户、用户账户管理、创建初始管理员、编辑作业价格表，完成初始化。



Vagrant 一键部署



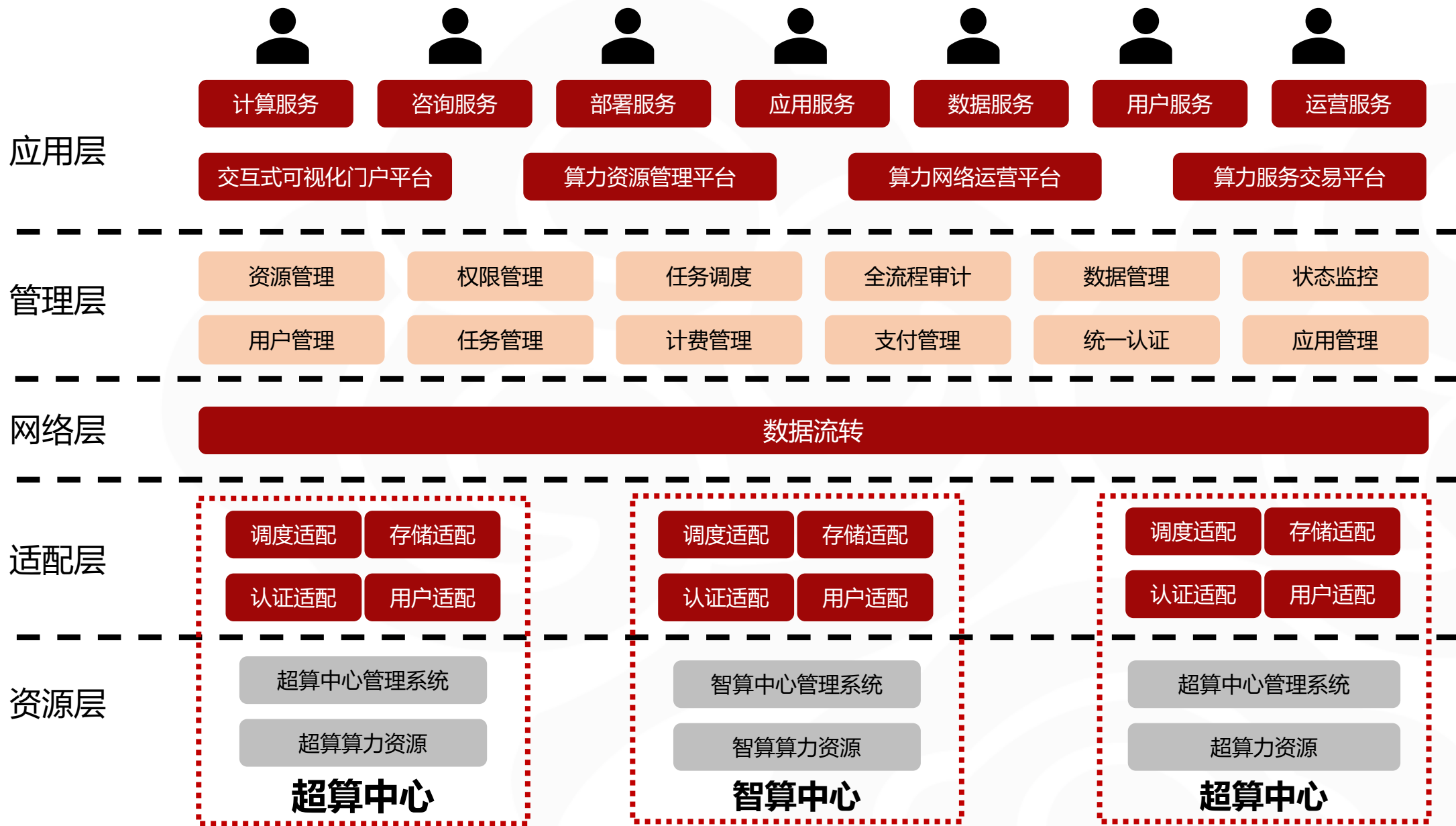
Docker 快速部署

SCOW——标准化平台，支持算力融合

SCOW提供了一套标准的平台接口，可以实现的异构平台的统一抽象，在此基础上可以进一步研制跨中心的算力融合平台，打通算力网络中各高性能计算中心的管理、使用、结算通道，连通算力孤岛，实现算力灵活接入、统一调度。



XSCOW——算力网络融合与交易平台



SCOW: 最新版本



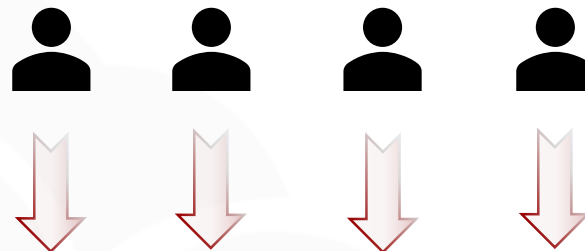
2023世界计算大会上发布SCOW 1.0版本

- **v1.0: 2023年9月14日发布**
 - 审计功能
- **v1.1: 2023年11月6日发布**
 - 1. 跨集群文件传输功能
 - 2. 定时同步账户状态功能
 - 3. 国际化功能, 提供英语版本
- **v1.2: 2023年11月19日发布**
 - 1. 增加在线编辑文件功能
 - 2. 增加系统基于域名的多风格支持
- **v1.3: 2023年12月23日发布**
 - 仪表盘首页更新集群状态
- **v1.4: 2024年1月31日发布**
 - 增加集群运营状态统计
 - 发布AI门户Beta版
- **v1.5: 2024年4月3日发布**
 - 增加后付费功能支持

SCOW——部署案例

北京大学

多套集群，分期建设，异地部署



SCOW

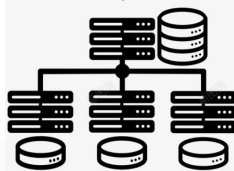
SCOW-PKU

统一认证
成果管理
消息通知



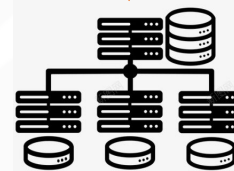
未名一号

- 2017年建
- 联想
- 理科一号楼



未名生科一号

- 2018年建
- 联想
- 理科一号楼



未名大数据一号

- 2019年建
- 联想
- 第二教学楼



未名二号

- 2023年建
- 超聚变
- 昌平新校区

SCOW——部署案例

华南理工大学

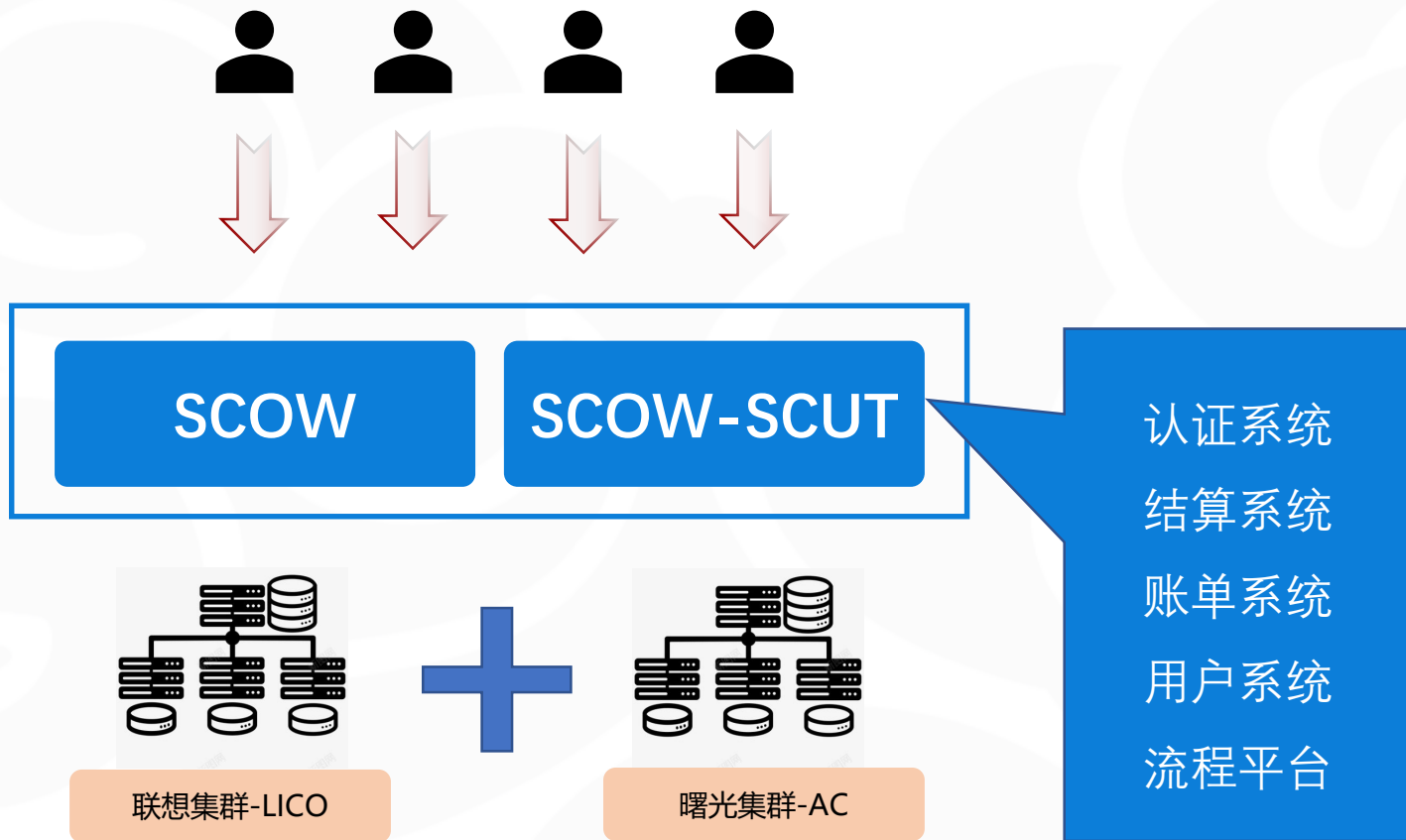
规模大，多家供应商建设，个性功能

集群1(联想集群):

- 存储9.1P,
- 3个计算分区:
 - CPU分区:320个节点
 - GPU分区:32个节点, 1机4卡
 - fat分区:3个节点

集群2 (曙光集群):

- 存储15P,
- 5个计算分区:
 - CPU分区1:16个节点
 - CPU分区2:195个节点
 - GPU分区1:3个节点, 1机8卡
 - GPU分区2:5个节点, 1机8卡
 - GPU分区3:19个节点, 1机8卡



SCOW: 推广和部署

平台部署已点亮16个省市46个算力中心

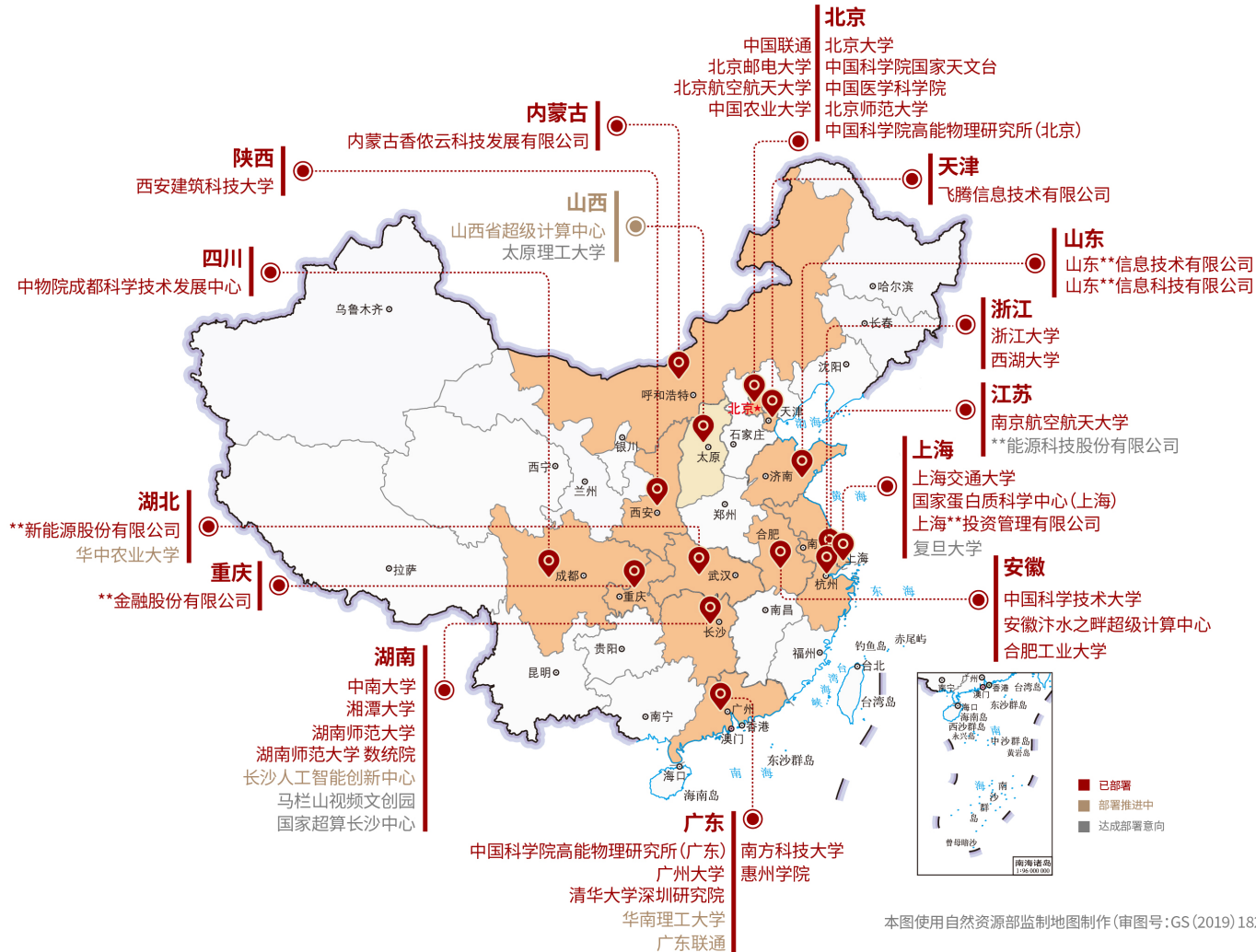
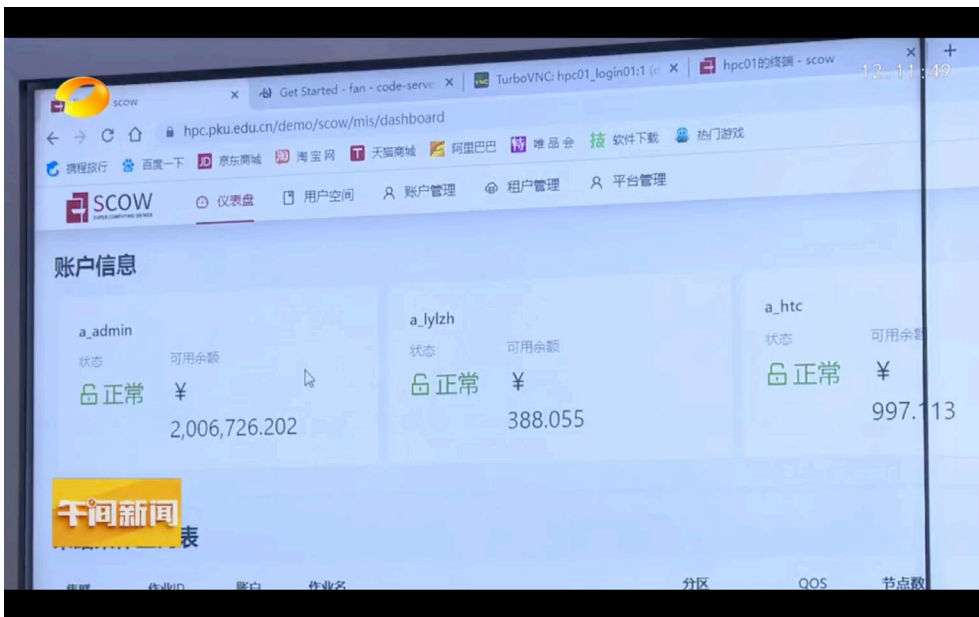
政府算力中心
4个

高校算力中心
25个

企业算力中心
11个

科研机构算力中心
6个

影响广泛，SCOW亮相湖南卫视



本图使用自然资源部监制地图制作(审图号:GS(2019)1825号)

SCOW——推广

支撑教育部高校智算融合共享平台

2023年11月13日，教育部教育管理信息中心与北京大学、中国电信正式签约，成立高校智算融合共享平台，旨在推进整合各地各校超算中心，搭建供需平台，推动以云服务方式为高校提供普惠性智能算力资源服务，促进高校AIGC应用发展。



XSCOW

高校智算融合共享平台



SCOW



XSCOW



北京大学
PEKING UNIVERSITY



浙江大学



南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



北京邮电大学
Beijing University of Posts and Telecommunications



北京航空航天大学
BEIHANG UNIVERSITY



中国农业大学
China Agricultural University



中南大学
CENTRAL SOUTH UNIVERSITY

支撑高校算力专员培训

教育部高等学校科学研究发展中心将在12月14日起举办高校算力专员培训，旨在基于SCOW平台培养算力中心建设与运维的急需人才，培训第一节课就是介绍基于SCOW的算力平台。

关于举办高校算力专员培训班(第一期)的通知 (第一轮)

各高等学校、有关单位及个人：

党的二十大首次将“教育数字化”写入报告，明确了教育数字化未来发展的行动纲领。算力作为“数字新基建”，在高校数字化转型过程中发挥着越来越大的作用，高校对算力的需求迎来突破性增长，这对高校算力中心的工作提出了更高的要求。为进一步推动高校算力中心建设、部署与实施，培养高校算力中心建设与运维方面的急需人才，在教育部高等学校科学研究发展中心指导下，《中国高校科技》杂志社、上海交通大学将开展高校算力专员培训班(第一期)。本次培训也是2023中国高校算力年会系列活动的重要内容之一。现将有关事项通知如下。

一、组织机构

指导单位：教育部高等学校科学研究发展中心
主办单位：《中国高校科技》杂志社
承办单位：上海交通大学
《中国教育网络》杂志编辑部
“中国教育数字化”新媒体平台
中国学术会议在线

二、培训对象

高校算力工作人员、学生。

五、日程安排

第一天 超算系统建设篇

- 09:00-10:30 课程一 使用SCOW构建算力聚合平台
李若淼 北京大学
- 10:30-12:00 课程二 构建异地高可用算力平台
张天阳 上海交通大学
- 14:00-15:30 课程三 超算机房规划与建设
李会民 中国科学技术大学
- 15:30-17:00 课程四 超算认证系统对接与安全防护
姚 舸 南京大学

第二天 超算系统使用篇

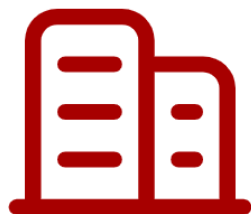
- 09:00-10:30 课程五 使用SLURM管理超算作业
高亦沁 上海交通大学
- 10:30-12:00 课程六 构建易用高效的超算软件环境
徐冬阳 上海交通大学
- 14:00-15:30 课程七 面向超算的高性能容器运行环境Singularity
张 洋 兰州大学
- 15:30-17:00 课程八 超算平台上的科学数据管理
韦建文 上海交通大学

什么是 SCOW 企业版

相比SCOW社区版本，SCOW企业版提供面向企业级应用场景的scowd性能增强组件、高等级的原厂企业级支持服务、更多可选功能等，同时提供SCOW最佳运营实践，有效助力算力中心快速构建并运营自己的算力资源。



scowd性能增强组件



原厂企业级支持服务

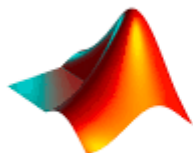


更多可选功能

为什么算力平台需要远程桌面?

- 更便捷的作业提交
- 更便捷的程序使用
- 数据结果的可视分析

RELION



MATLAB®



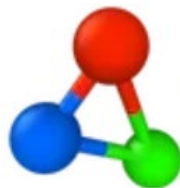
UCSF
Chimera



北天天元



ParaView



OVITO

远程桌面的传输技术主要分为以下三种：

- 帧缓冲传输：适合简单、跨平台需求，配置灵活，但在高动态内容下表现较差。

- 各种VNC



- RDP传输：适合Windows环境，高效的带宽利用率和用户体验，适用于一般办公和应用程序使用。

- 微软RDP



Remote Desktop Services

- 视频流传输：适合高性能高分辨率场景，如设计和多媒体应用，在硬件加速场景下效果更好。

- NICE DCV



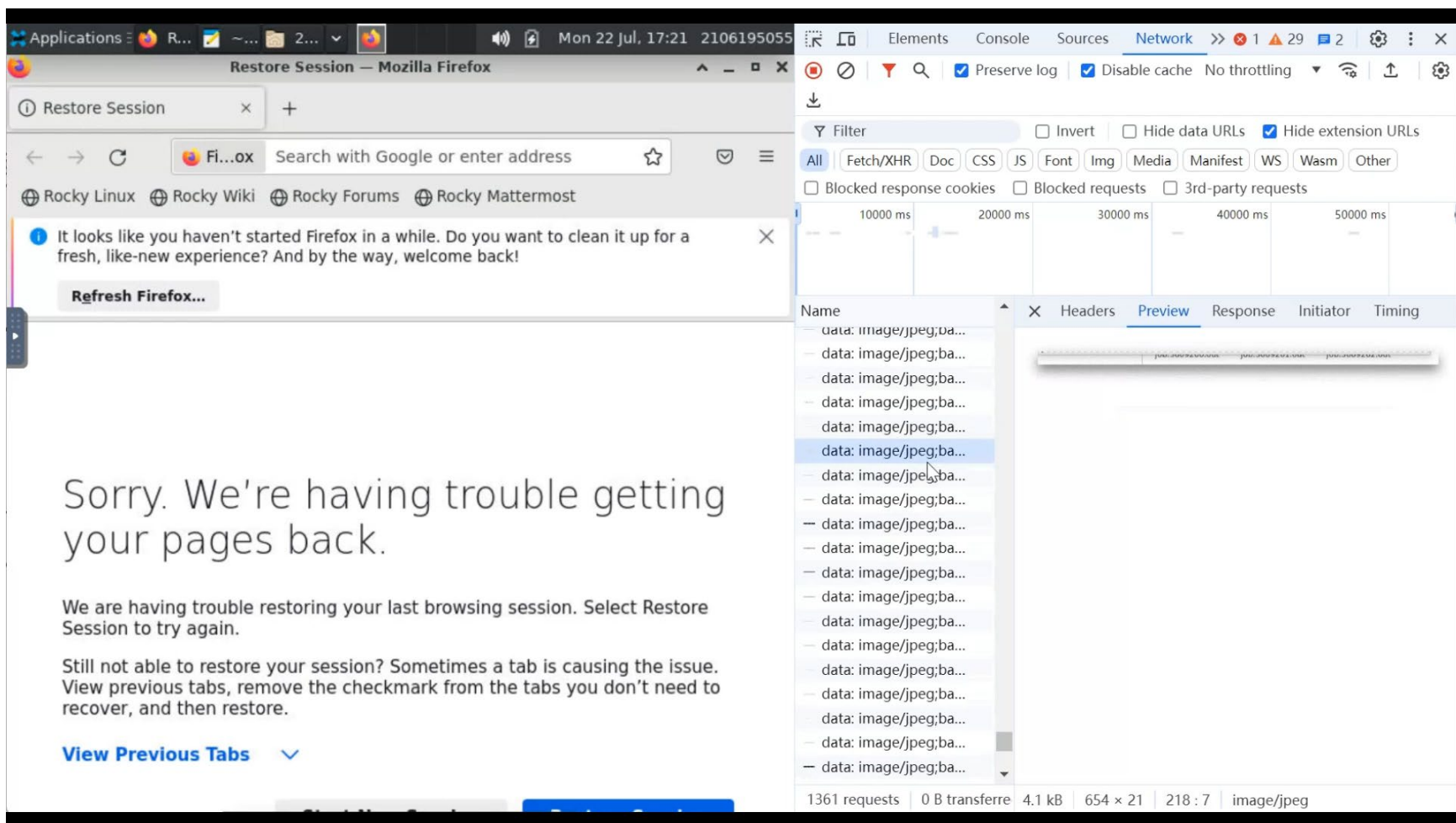
- RayLink

- Citrix Virtual Desktops



- Shadowdesk

VNC协议Protocol (RFB) 是一个简单的协议，传送服务端的原始图像到客户端（一个X,Y位置上的正方形的点阵数据），客户端传送事件消息到服务端。在被控端桌面变化较大的场景下带宽占用极高。



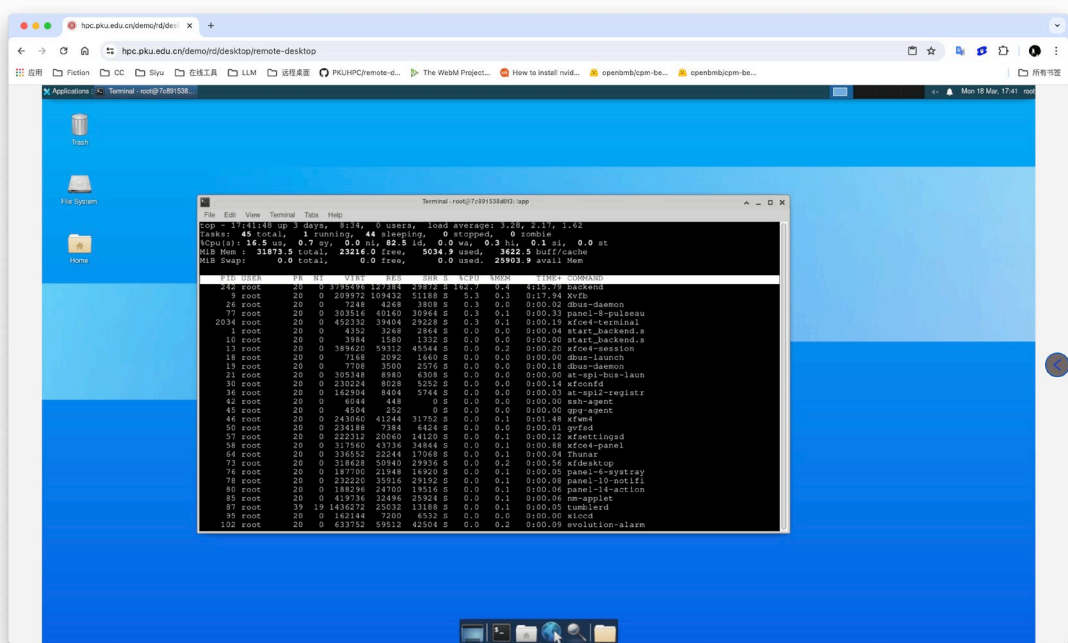
ShadowDesk远程桌面

面向更高质量要求的远程桌面访问，构建基于流式传输的远程桌面解决方案。通过硬件编码提升复杂场景处理能力。控制端使用浏览器作为载体，被控端支持Linux与Windows，提升远程桌面使用的便捷性。支持丰富的附加功能，提升远程桌面的功能完备性。

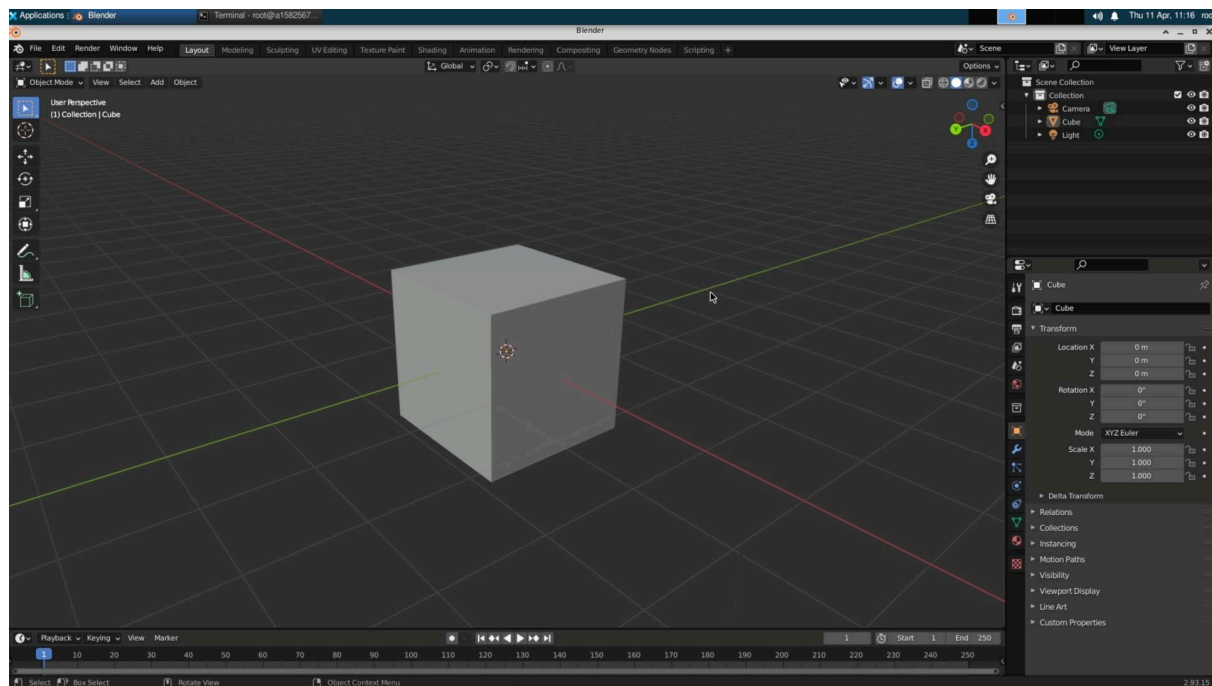


ShadowDesk远程桌面

面向更高质量要求的远程桌面访问，构建基于流式传输的远程桌面解决方案。通过硬件编码提升复杂场景处理能力。控制端使用浏览器作为载体，被控端支持Linux与Windows，提升远程桌面使用的便捷性。支持丰富的附加功能，提升远程桌面的功能完备性。



支持日常办公与开发场景



完美支持对3D建模等精细操作（Blender视频）

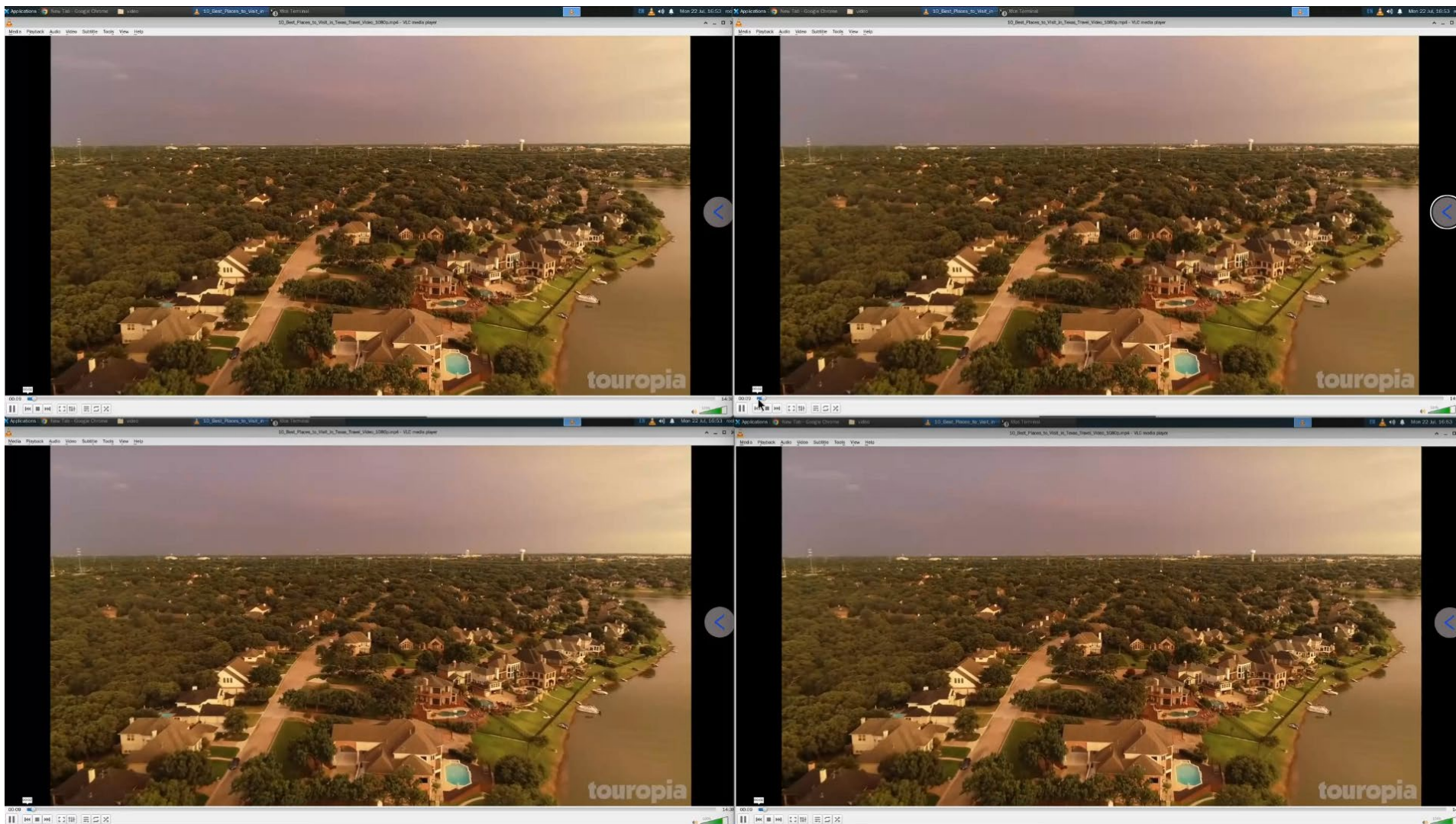
ShadowDesk远程桌面-易用性

- 集成的剪贴板功能让用户可以毫无障碍地在本地和远程桌面之间传输文本。
- 音频传输支持则拓宽了对需音频反馈应用的使用范围。
- 容器化部署和虚拟桌面集成的功能，使用户容易创建和管理所需环境。
- 支持无损色彩模式，还原桌面原始色彩。
- 加密和安全措施将保证用户的所有数据在进行传输时均处于加密状态，以提升产品的安全性。



ShadowDesk远程桌面-易用性

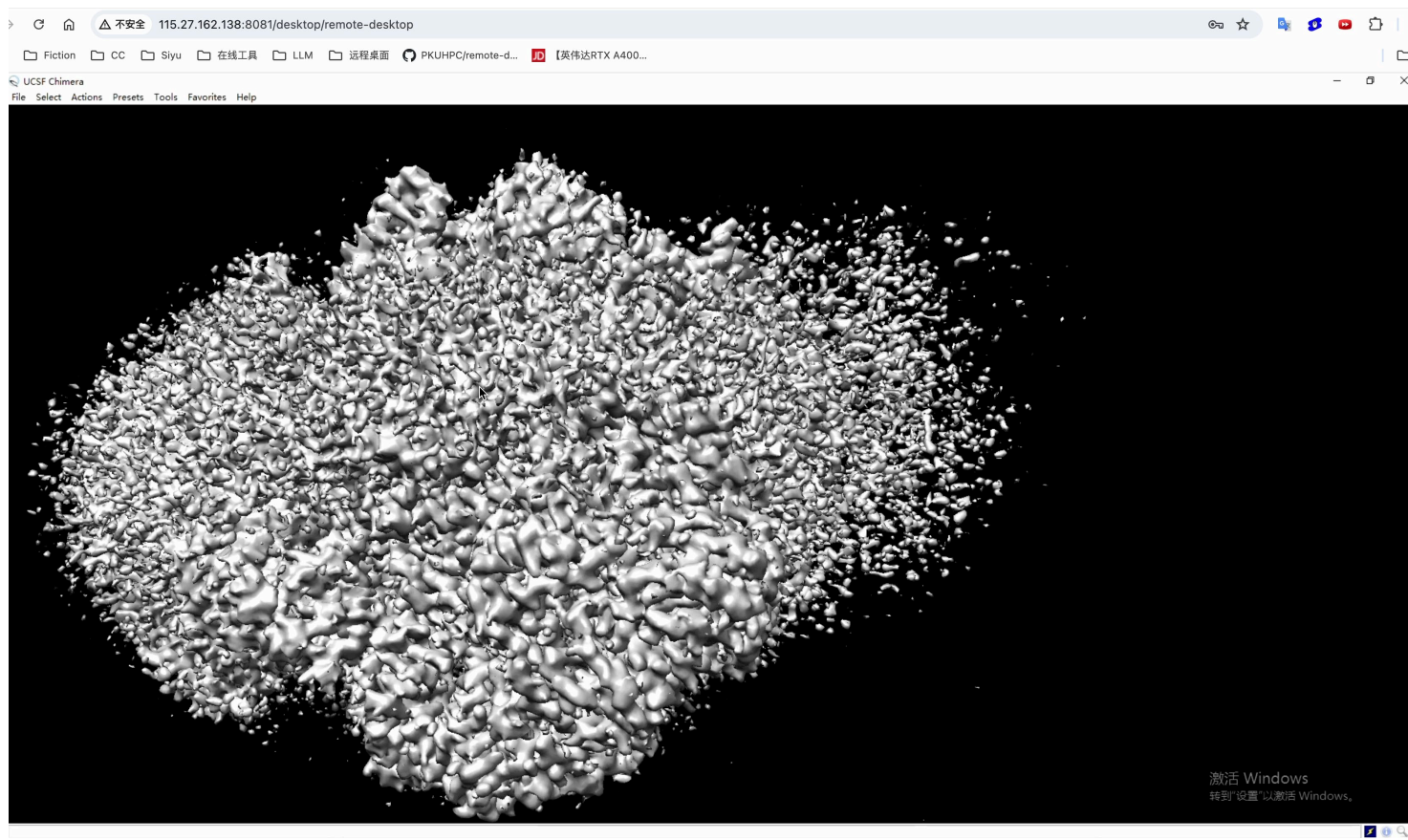
- 支持多个控制端控制同一个被控端，提升协作体验。



ShadowDesk远程桌面-多端支持

支持Linux、Windows操作系统作为被控端，使用浏览器作为控制端。我们正在开发Windows、Mac和Linux操作系统的控制端，即将上线。

UCSF Chimera是一个用于分子结构和相关数据的交互式可视化和分析工具。主要包括：密度图，超分子组合，顺序排列，对接结果，轨迹和构象整合。也可以生成高质量图像和动画。

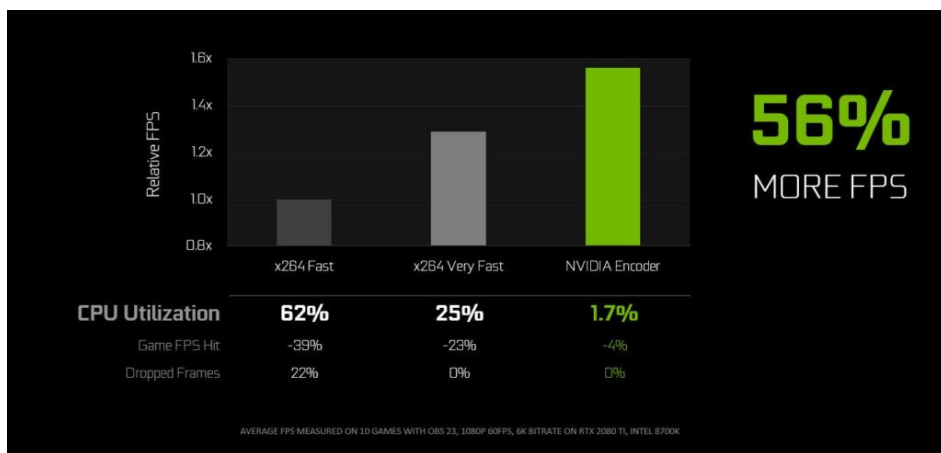


ShadowDesk远程桌面-为什么需要硬件加速

程序渲染需要

- UCSF Chimera是一个用于分子结构和相关数据的交互式可视化和分析工具。
- OVITO是一个用于原子和粒子模拟数据的科学可视化和分析软件。

桌面传输视频编码需要



NVENC - Encoding (Click to expand)

Consumer (GeForce) Professional (NVIDIA RTX / Quadro) Server (Data Center) DGX

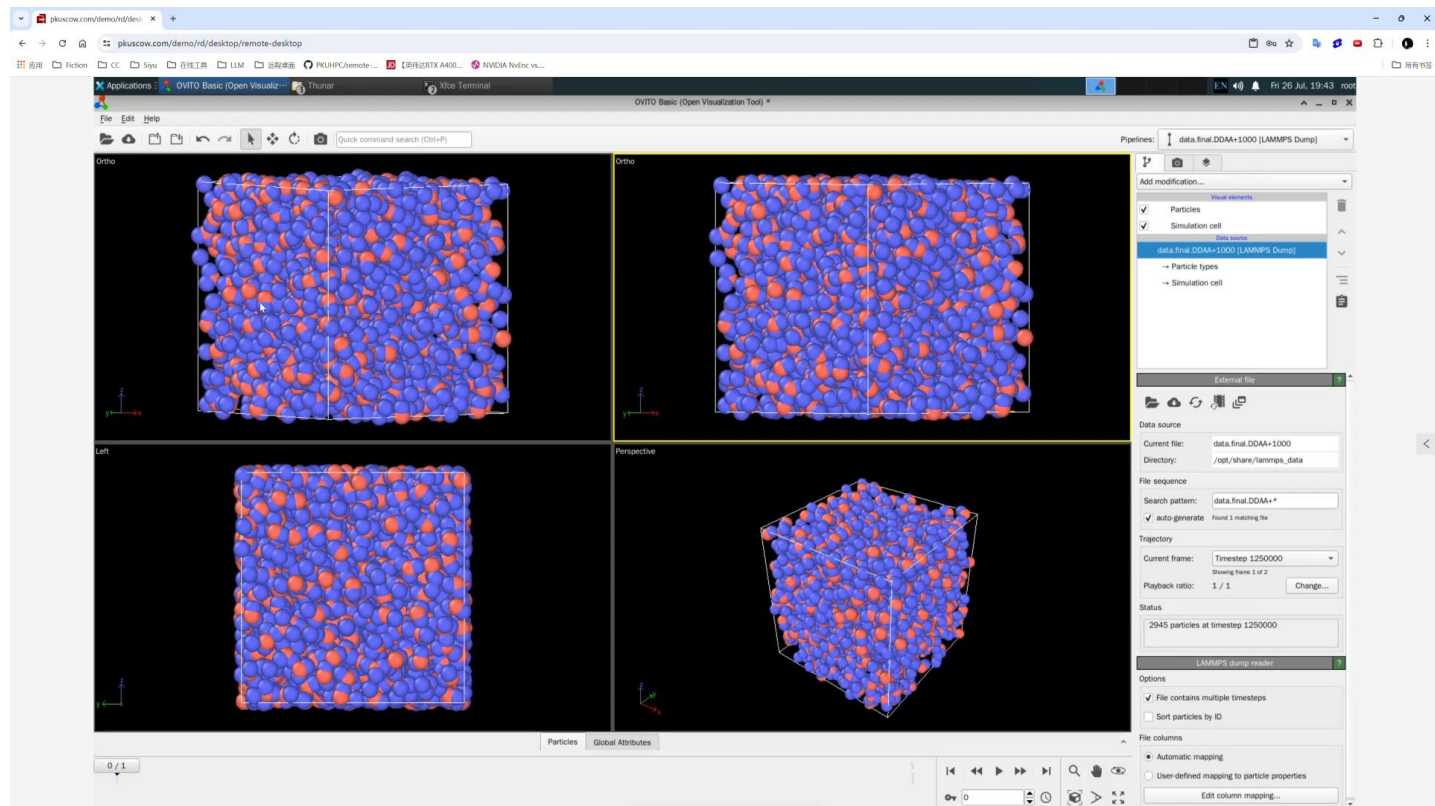
Search for names..

BOARD	FAMILY	NVENC Generation	Desktop/ Mobile	# OF CHIPS	Total # of NVENC	Max # of concurrent sessions	H.264 (AVCHD) YUV 4:2:0	H.264 (AVCHD) YUV 4:4:4	H.264 (AVCHD) Lossless	H.265 (HEVC) 4K YUV 4:2:0	H.265 (HEVC) 4K
NVIDIA RTX 1000 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	1	8	YES	YES	YES	YES	YES
NVIDIA RTX 2000 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	1	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 3000 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	1	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 3500 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	2	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 4000 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	2	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 500 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	1	8	YES	YES	YES	YES	YES
NVIDIA RTX 5000 Ada Generation Laptop GPU	Ada Lovelace	8th Gen	M	1	2	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 2000 Ada Generation	Ada Lovelace	8th Gen	D	1	1	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 4000 Ada Generation	Ada Lovelace	8th Gen	D	1	2	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 4000 SFF Ada Generation	Ada Lovelace	8th Gen	D	1	2	Unrestricted	YES	YES	YES	YES	YES
NVIDIA RTX 4500 Ada Generation	Ada Lovelace	8th Gen	D	1	2	Unrestricted	YES	YES	YES	YES	YES

ShadowDesk远程桌面-硬件加速支持

现已支持显卡加速，使得视频处理性能得到了显著提升。借助这种强大的硬件加速能力，远程桌面能够实现高达4K分辨率和60fps的流畅显示效果，即使在复杂的图形处理任务和高需求的应用程序场景下，也能保持较低的延迟和高质量的用户体验。

OVITO 是一款专业实用、功能强大的原子分子可视化及分析软件。界面美观，功能齐全，操作简单，支持超大规模原子快速显示。LAMMPS 的 dump 构型、VASP 的 POSCAR, XDATCAR 等构型均可由 OVITO 查看和编辑。



ShadowDesk远程桌面-视频对比

VNC



塔影 (硬件编码)

注：常用的VNC远程桌面采用帧缓冲的方式，无法适应高帧率、低延迟的场景。本次测试在同一台机器实验，分别率为2K。

面向更高质量要求的远程桌面访问，构建基于流式传输的远程桌面解决方案。通过硬件编码提升复杂场景处理能力。控制端使用浏览器作为载体，提升远程桌面使用的便捷性。支持丰富的附加功能，提升远程桌面的功能完备性。

多主控端支持

Windows

MAC

Linux

传输效果持续优化

更多编解码方案

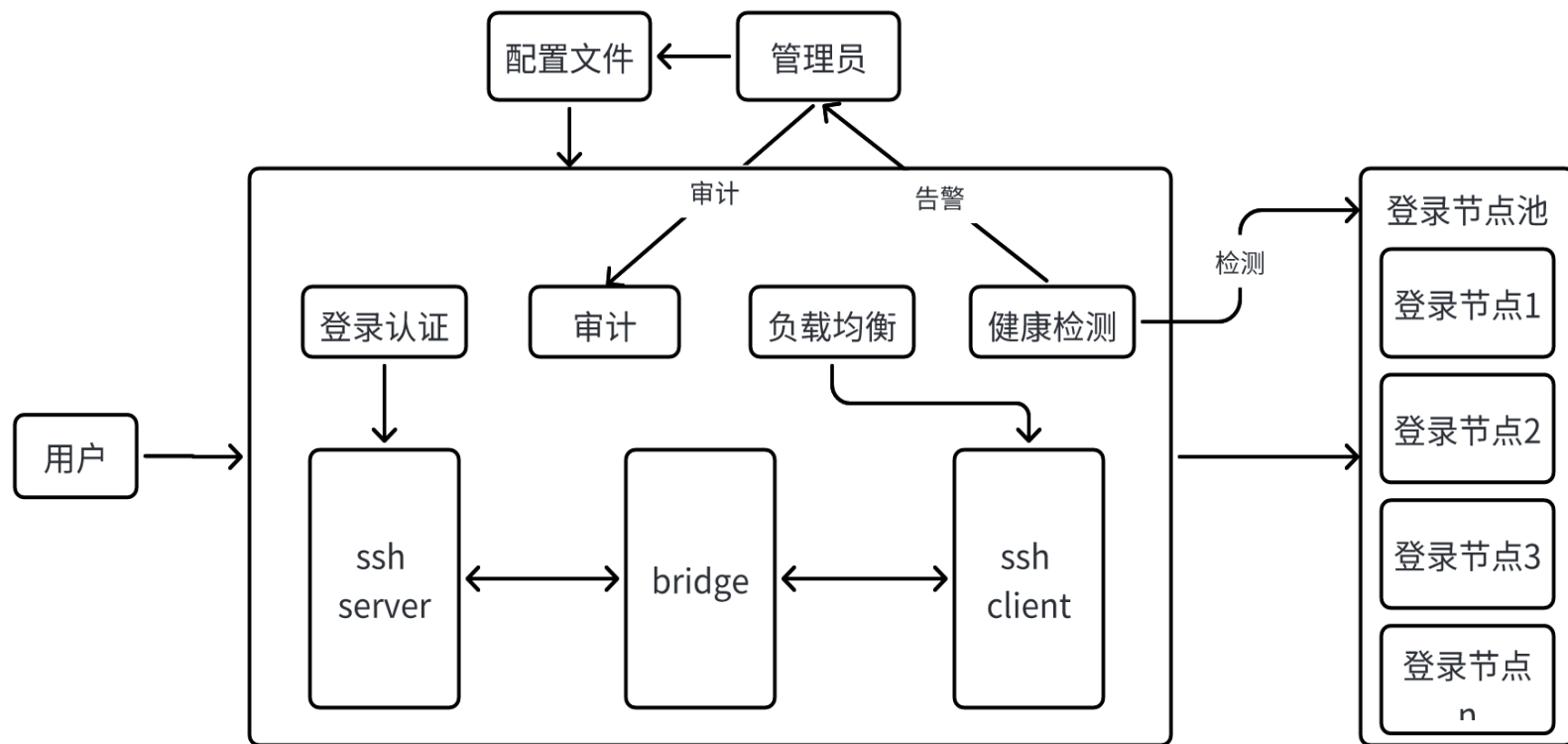
更优化传输协议

我们部署了一个在线的Demo，供大家试用，欢迎批评指正。

试用请联系付振新（手机、微信号：15650708568，电子邮箱：fuzhenxin@pku.edu.cn）

在高性能计算平台的安全设计考虑中，全流程审计是非常重要的安全能力，除了管理员登录管理节点需要通过堡垒机，用户访问高性能集群也应通过堡垒机，以记录命令、事后追踪。本项目意图构建透明堡垒机软件，让用户无感使用堡垒机，完成对命令的审计。

- 登录认证
 - 密码认证
 - keyboard-interactive认证
 - 用户白名单
- 行为记录
- 节点告警
- 审计
 - 命令行审计
 - 命令行回放

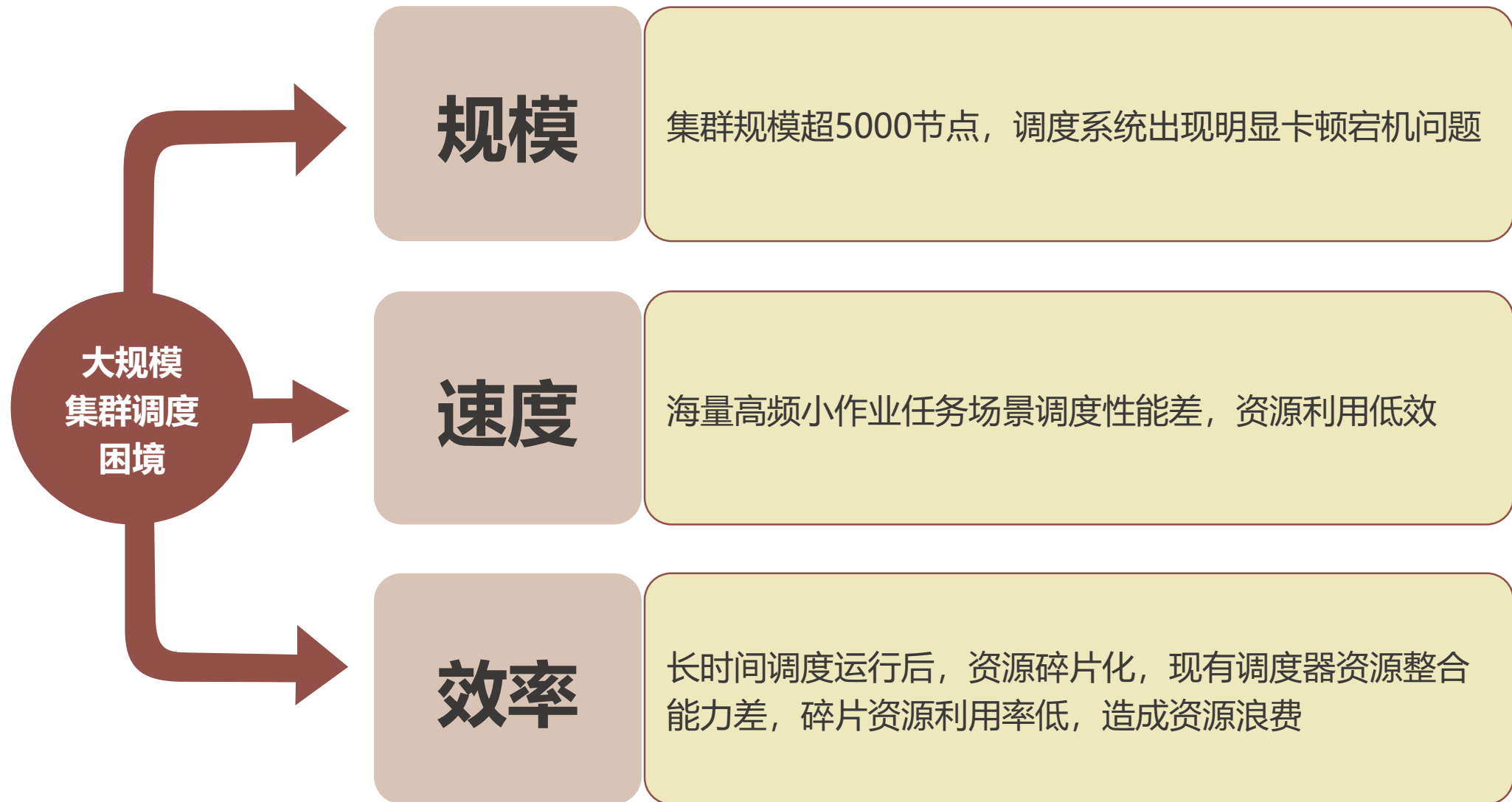


content

目录

- 一 | **北大超算平台建设与服务**
- 二 | **算力中心门户平台**
- 三 | **算力资源调度**
- 四 | **未名小蒜**





鹤思 (CraneSched) 智能调度系统

统一抽象并纳管异构资源算力，通过AI赋能调度算法，对资源进行细粒度编排调度，提升集群资源利用效率。并通过优化系统架构，分级并发加速，支撑大规模集群高通量计算场景。



↑ 支撑

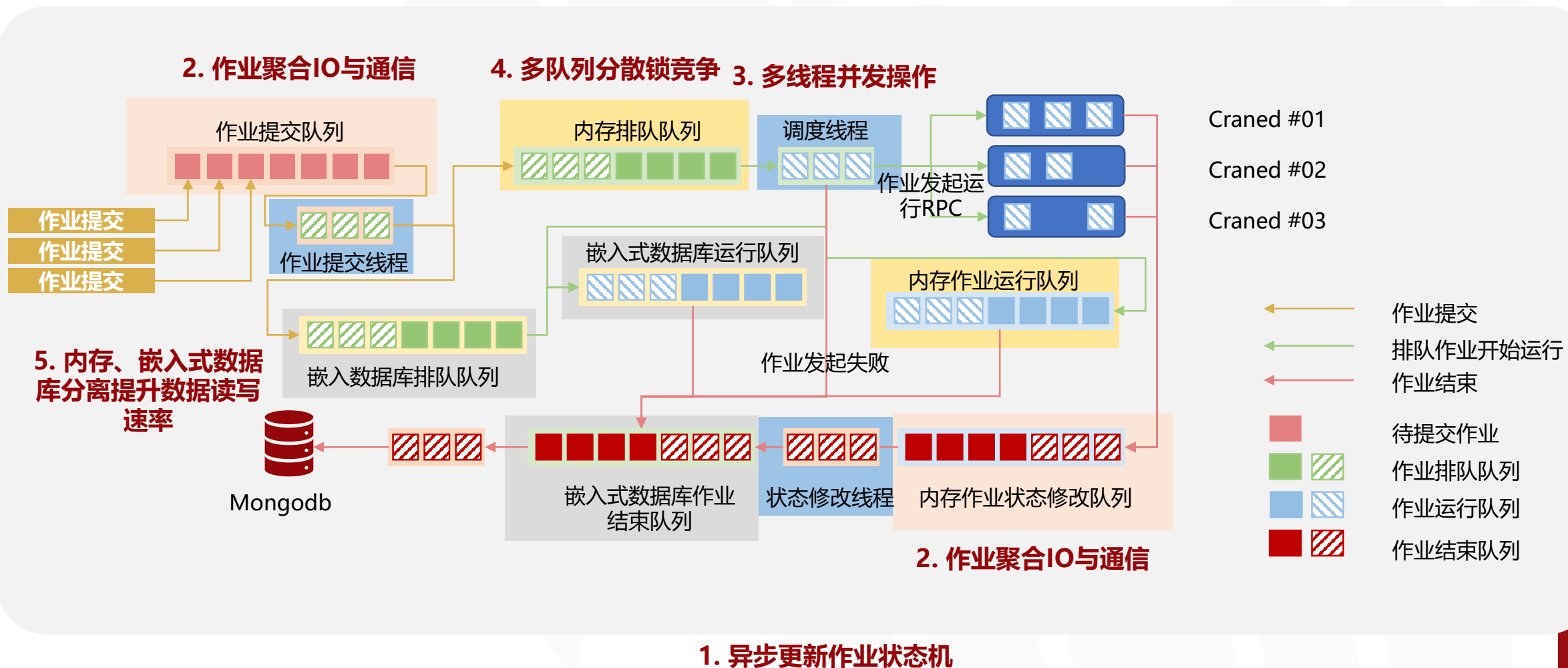


↓ 适配



鹤思智能调度系统性能优化

- CraneSched创新性地设计了新的软件架构，设计了支持异步更新的作业状态机和聚合式IO与通信，多线程并发操作，多队列细粒度分散多任务状态锁竞争，内存和磁盘层原子性分离，解决高频读和高频写冲突的问题。



鹤思智能调度系统测试指标

国内首个开源高性能计算智能调度系统，各项性能指标领先

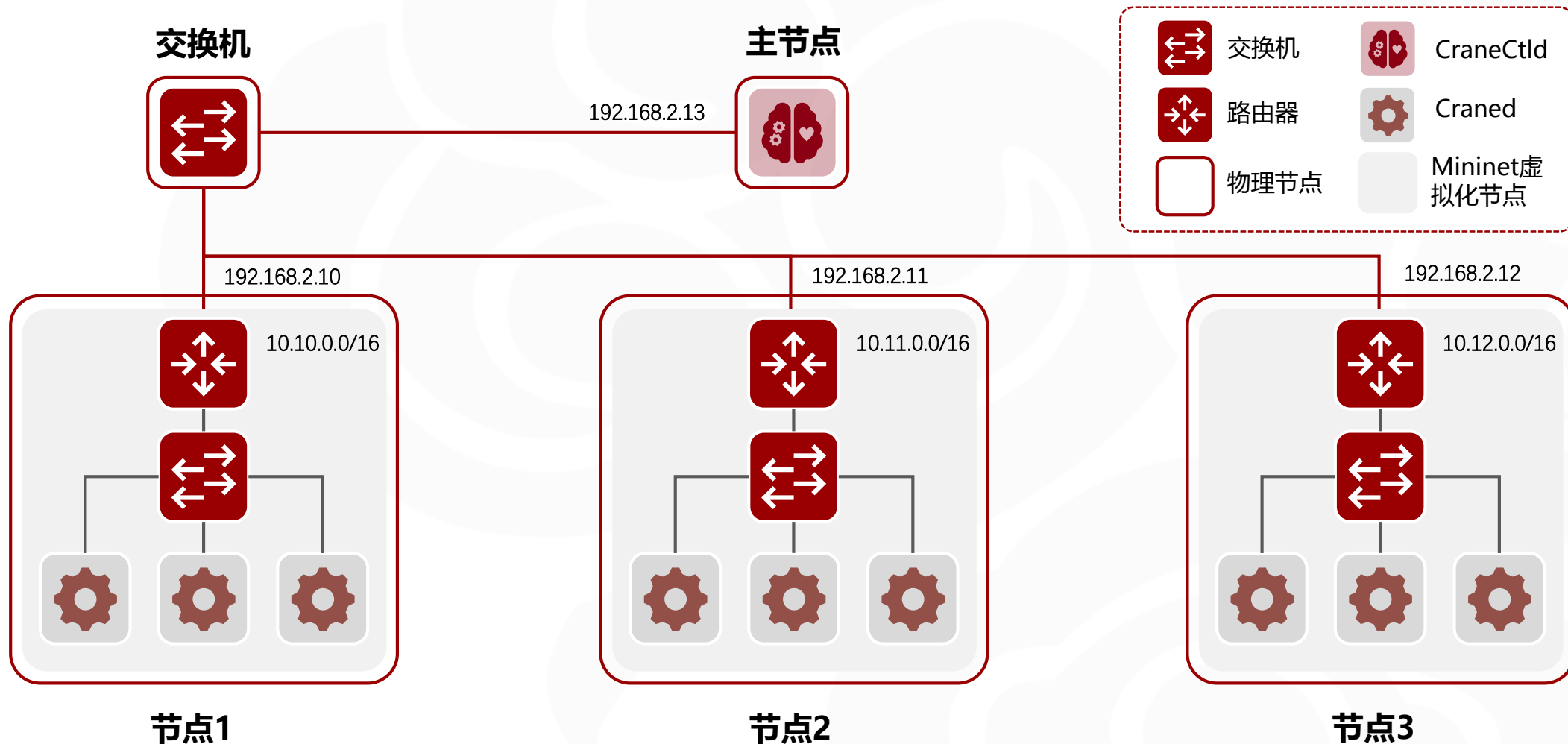
调度系统	鹤思	多瑙	Fsched	Slurm	LSF
单位	北京大学	华为公司	速石科技	SchedMD	IBM
是否开源	开源	否	否	开源	否
自主研发	是	是	Slurm扩展	是	是
支持集群规模 (节点)	100000+	10000	1000	100000+	50000+
每秒调度任务数	10593	1111	1000	1200	-
同时运行任务数	200万	30万	-	-	-
编程语言	C++&Go	Java	C	C	C
是否有UI	是	是	是	是	是
智能调度	是	否	否	否	否
任务依赖	是	是	是	是	是
支持GPU	是	是	是	是	是
FairShare	是	是	是	是	是
多级账号管理	是	否	否	否	否

目前已经部署在5个单位：清华大学深圳研究院、合肥国家实验室、华为技术有限公司、中国海洋大学、某EDA公司

鹤思系统大规模集群测试方法

少量节点模拟大规模集群，高效经济，快速发现系统性能瓶颈

- 使用Mininet对物理节点进行轻量级虚拟化隔离
- 单节点可虚拟化为2000+虚拟节点，11个节点（1个主节点不虚拟化）即可模拟20000+节点规模集群测试



鹤思—国内首个开源“高性能计算”智能调度系统



项目由北京大学与北京大学长沙计算与数字经济研究院联合研发

完全自主知识产权

已完成高性能计算和部分智能计算功能

项目源码和文档地址:

项目源码地址:

<https://github.com/PKUHPC/CraneSched>

<https://github.com/PKUHPC/CraneSched-FrontEnd>

项目文档地址:

<https://pkuhpc.github.io/CraneSched-document/>

项目试用地址:

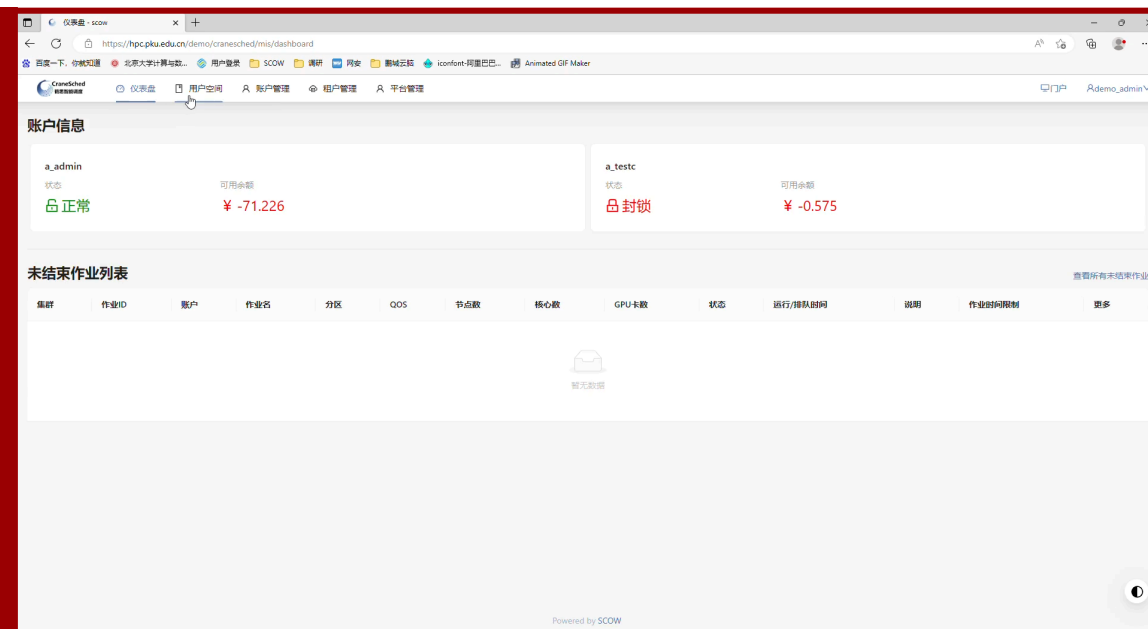
地址: <https://hpc.pku.edu.cn/demo/cranesched>

用户名/密码: demo_admin / demo_admin

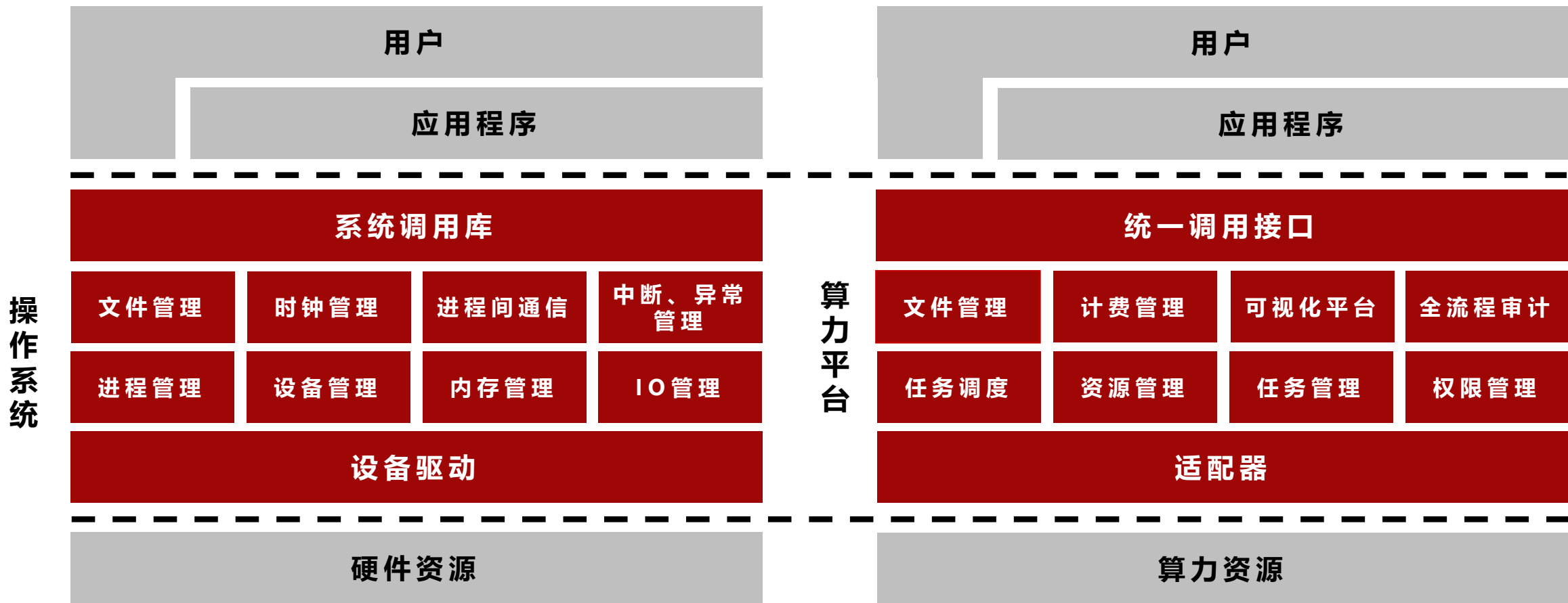
门户系统



管理系统



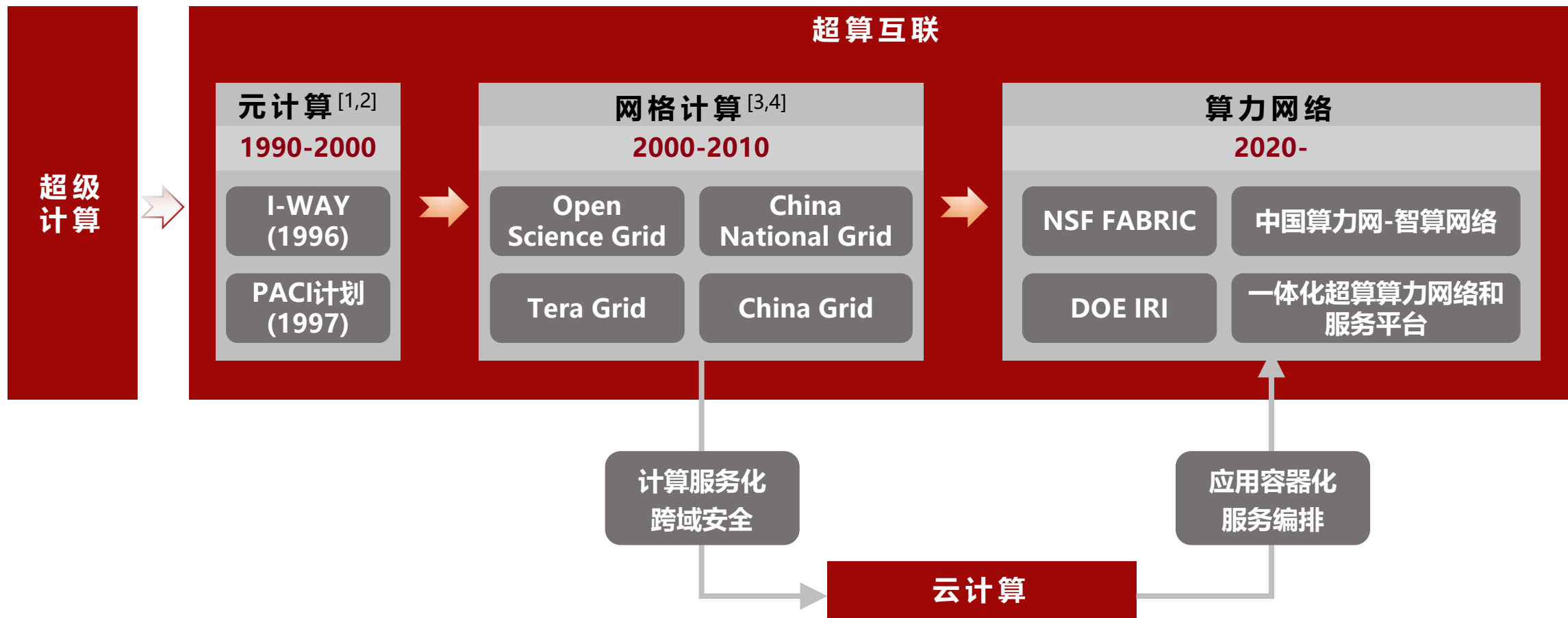
算力平台是帮助用户使用算力资源的**操作系统**，是算力经济时代的**关键基础软件**。



传统操作系统架构

算力平台架构

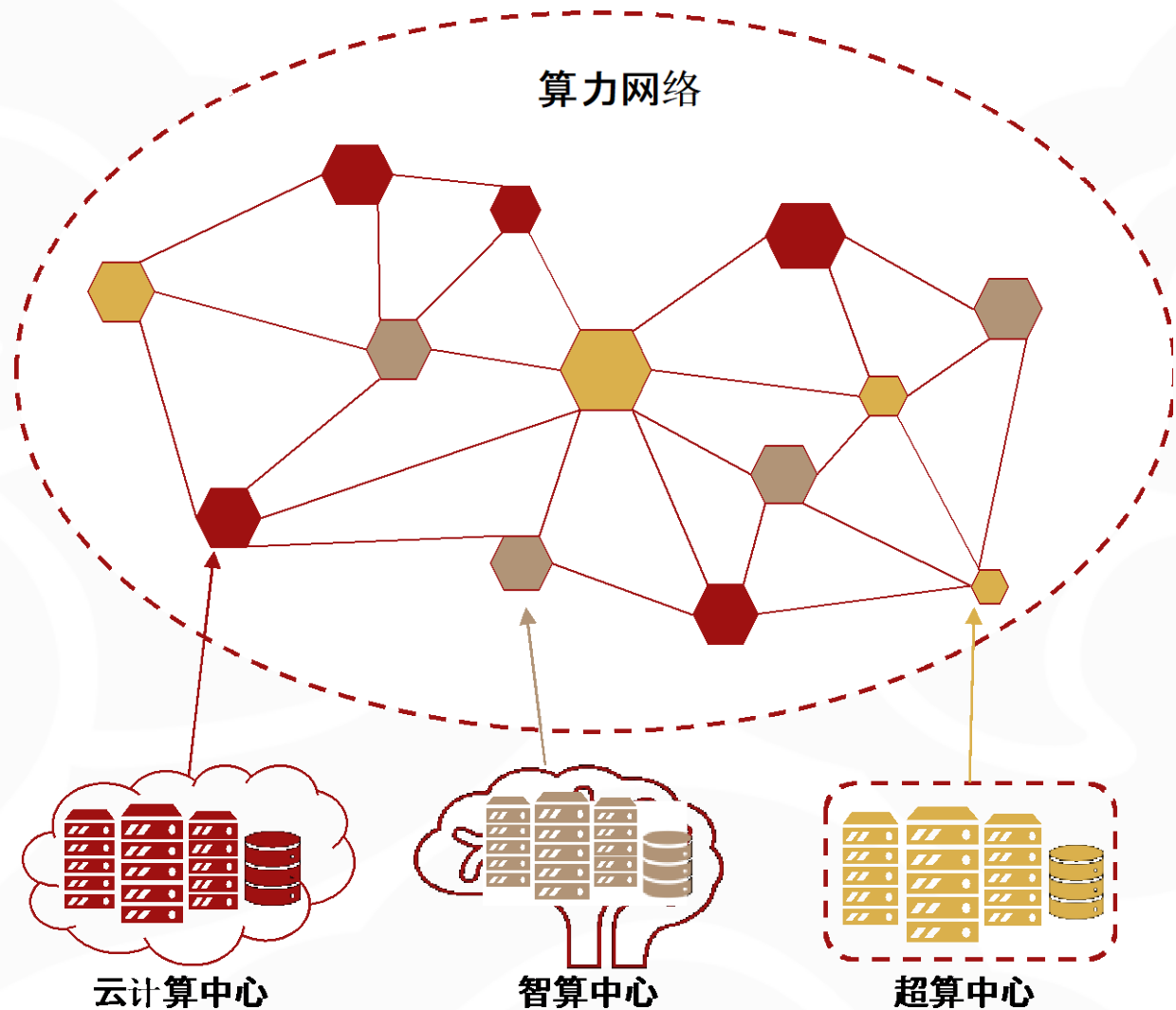
算力平台的发展历程

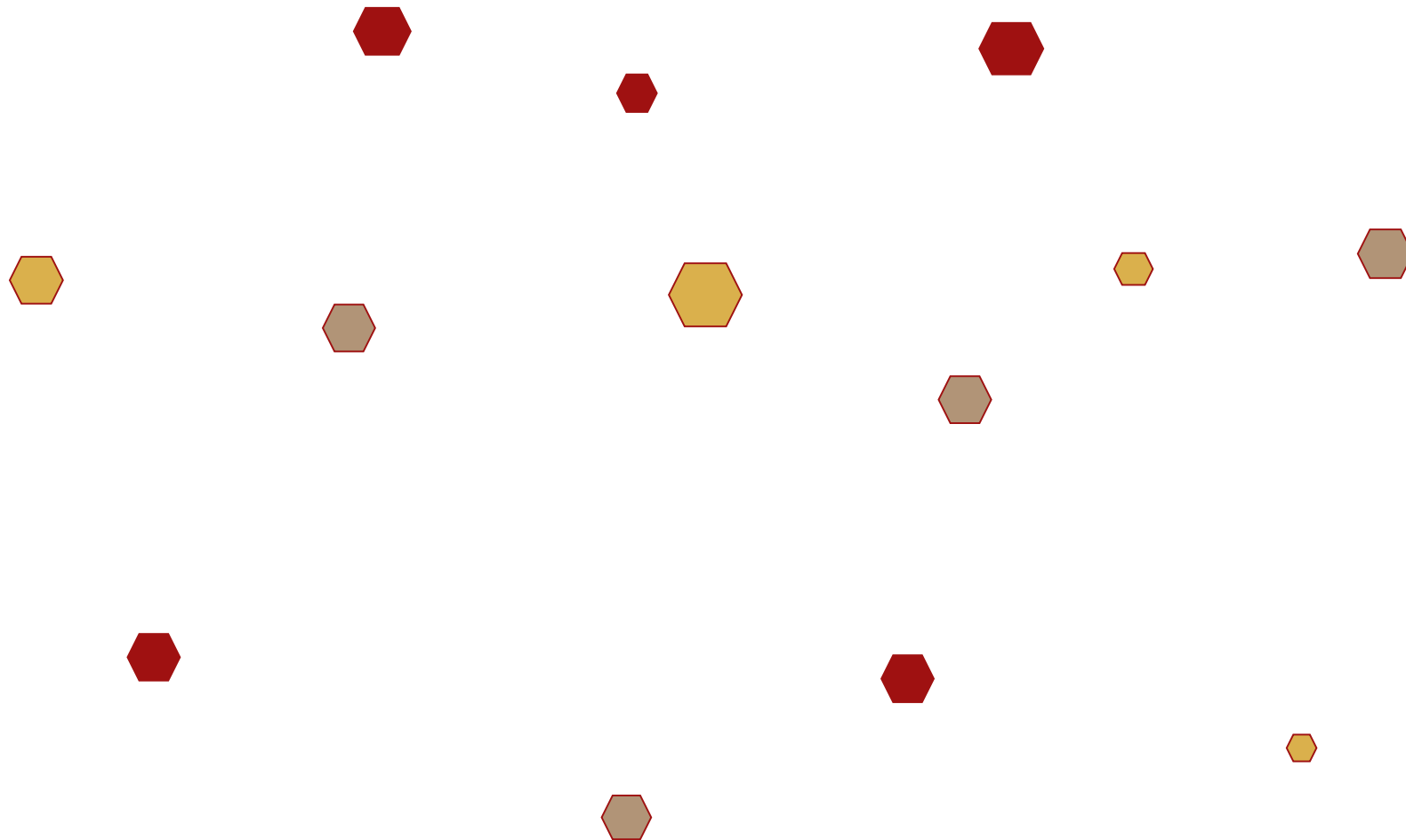


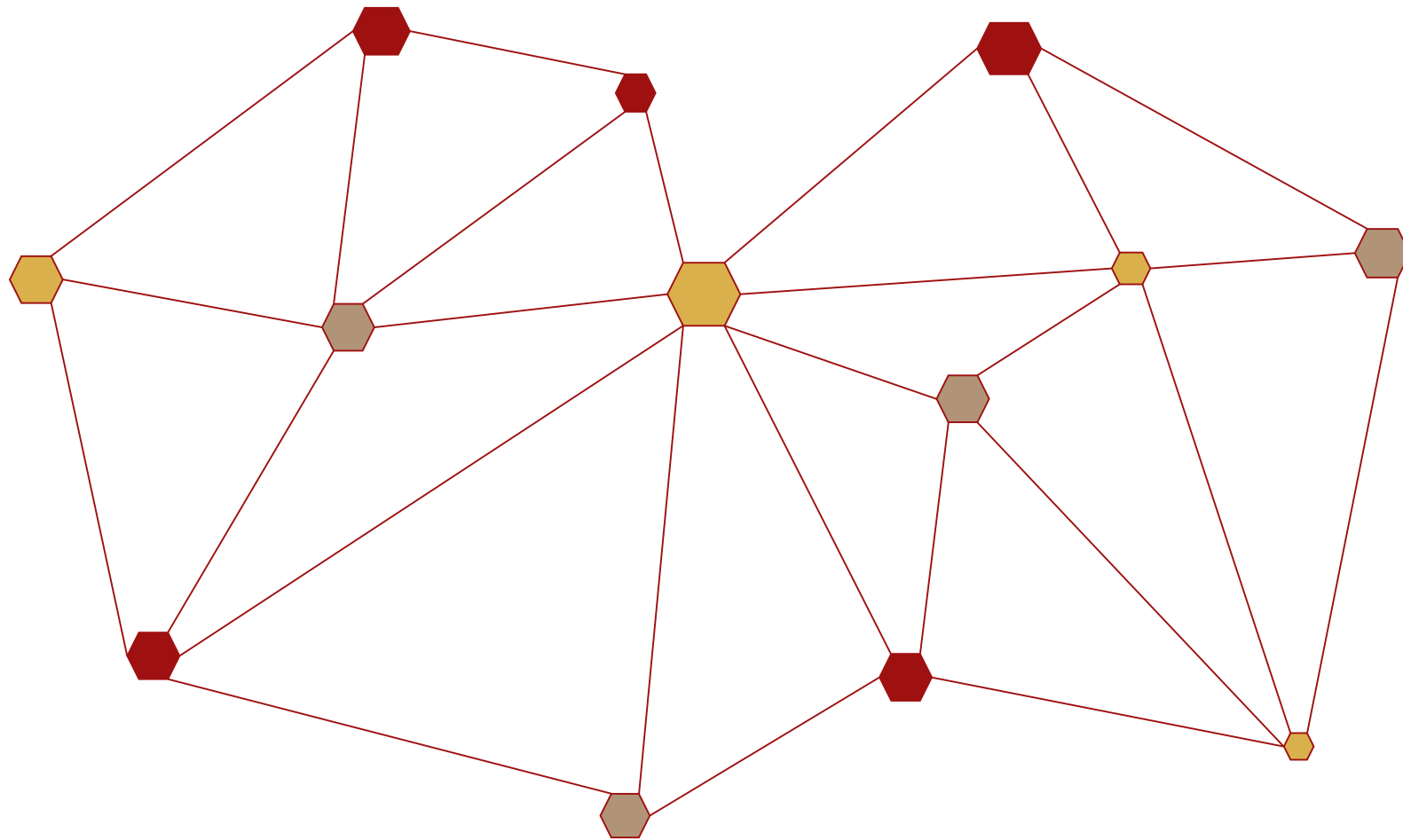
参考文献:

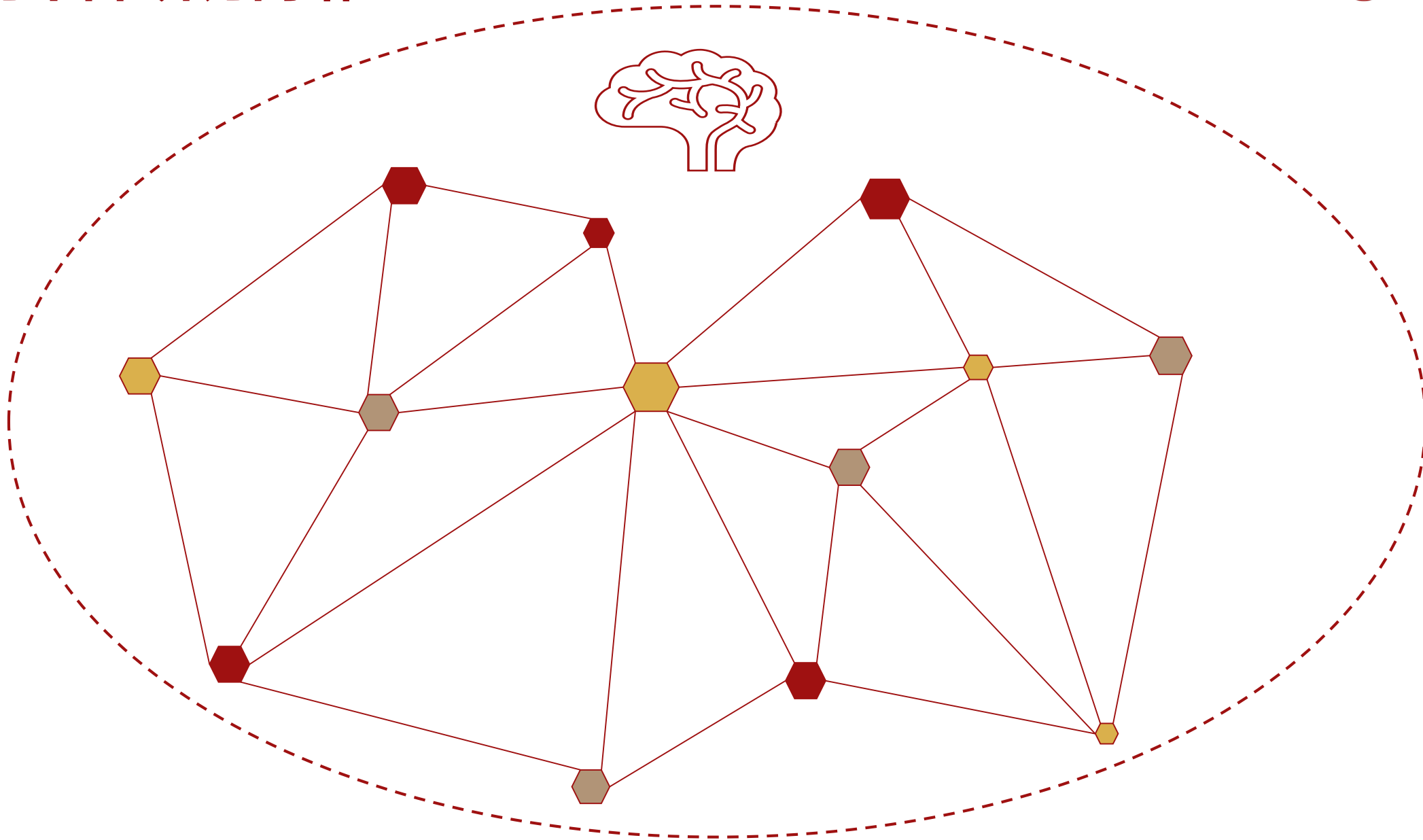
- [1] Smarr L, Cadett CE. Metacomputing. Communications of the ACM. 1992, 35(6).
- [2] 桂小林,钱德沛.元计算系统的研究现状与趋势[J].计算机科学, 2001.
- [3] Foster I, Kesselman C .The Grid: Blueprint for a New Computing Infrastructure Morgan Kaufman[B]. 1998.
- [4] Foster I , Kesselman C .The Grid 2: Blueprint for a New Computing Infrastructure Morgan Kaufman[B]. 2003.

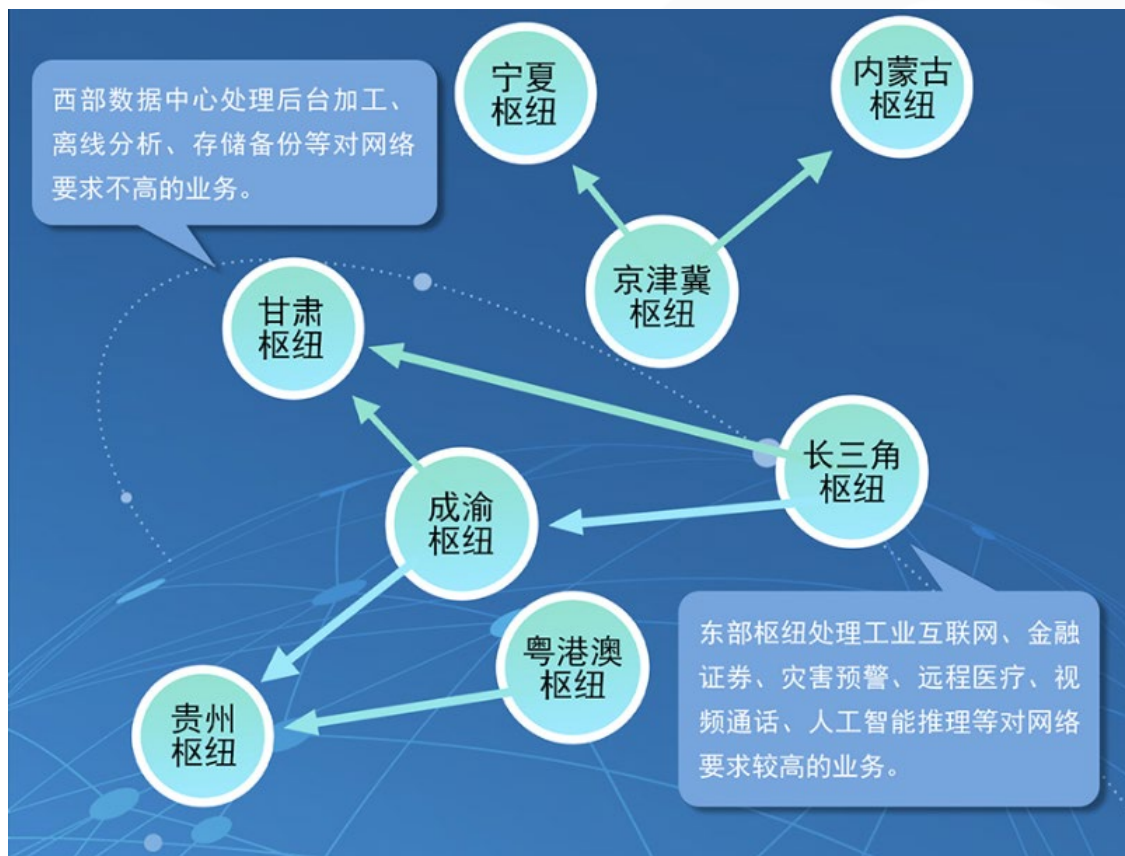
- **算力**是设备每秒可处理信息的能力。
- **算力网络**是一张可以连接算力的网络,让接入网络的需求侧可以共享网络中的所有算力供给。
- **构建算力网络**就是整合高性能计算、云计算、AI智算等各类算力中心资源,实现计算资源、存储资源和网络资源的动态感知、智能调度、按需分配。











content

目录

- 一 | 北大超算平台建设与服务**
- 二 | 算力中心门户平台**
- 三 | 算力资源调度**
- 四 | 未名小蒜**



北大高性能计算平台运维困境

7×24小时用户服务

平台科研用户习惯晚上使用集群并咨询，7×24人工运维成本高。

同类问题反复询问

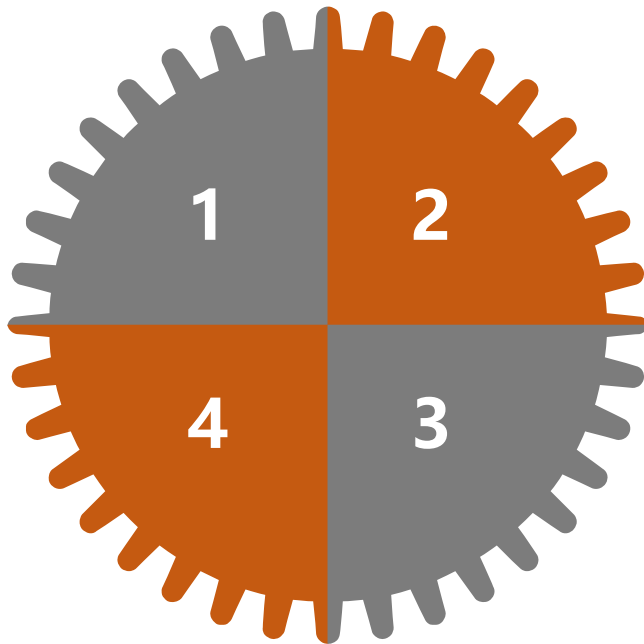
用户反复咨询平台使用手册中常见问题，加大了运维人力成本。

冷门问题解决困难

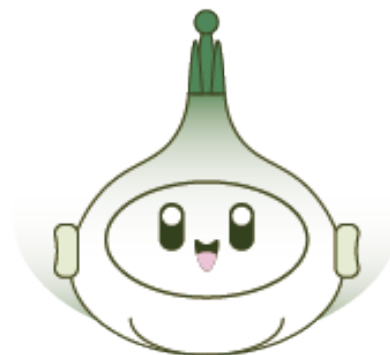
用户有时会问冷门问题，运维人员无法及时回答，影响用户体验。

高峰时期响应时效

用户高峰期问题多，目前人力无法及时响应，影响用户体验。

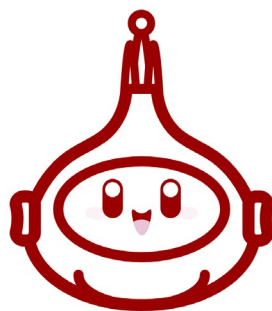


自研智能问答助手



未名小蒜

小蒜logo



北京大学校内信息门户

信息服务

办事大厅

校内公告

我的门户

功能搜索



党费缴纳



教工预约访客



畅行清华



我的邮箱



校园卡



个人收入



财务部经费



我的课程



CARS资源共享



智慧场馆



校园指数

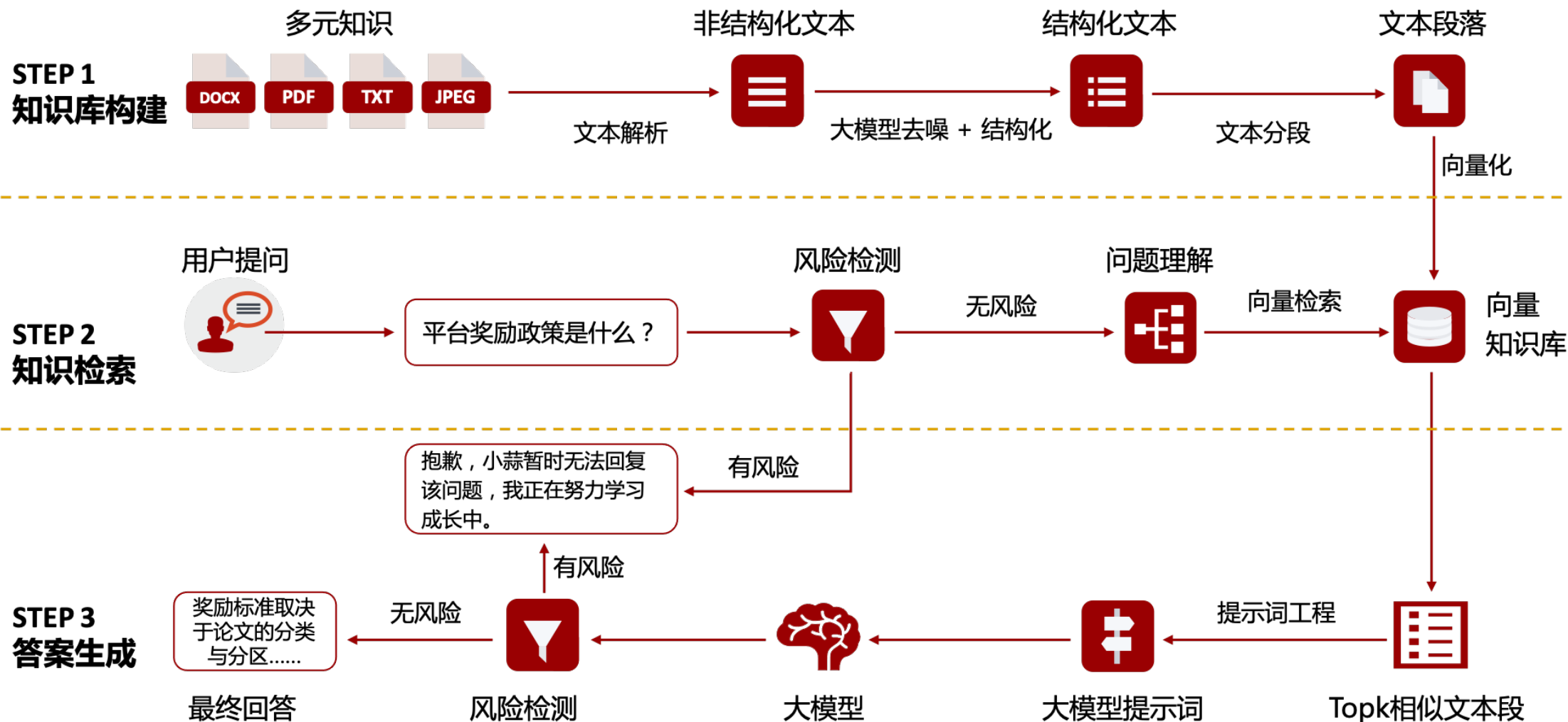


待办事项

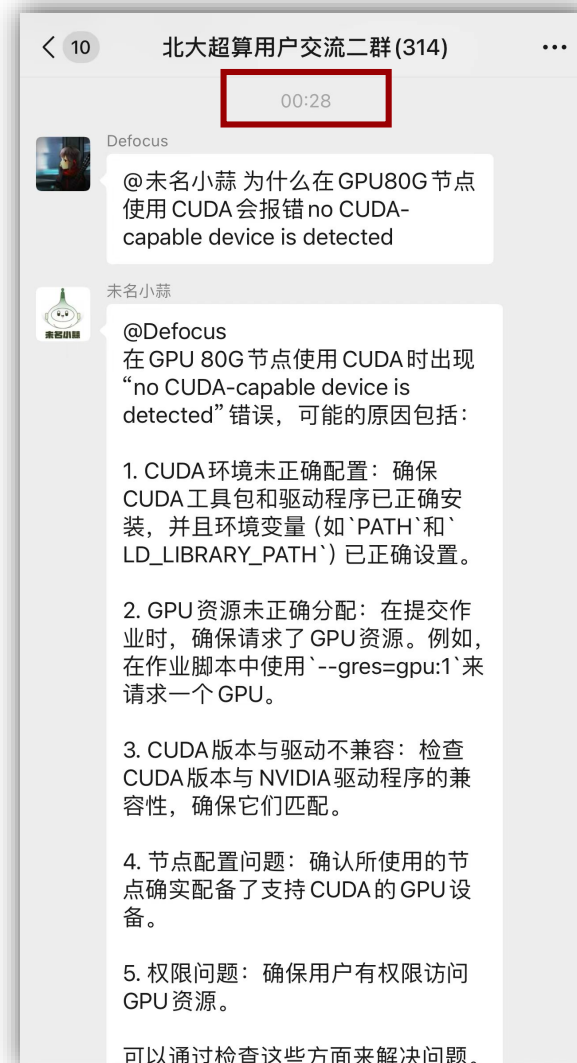
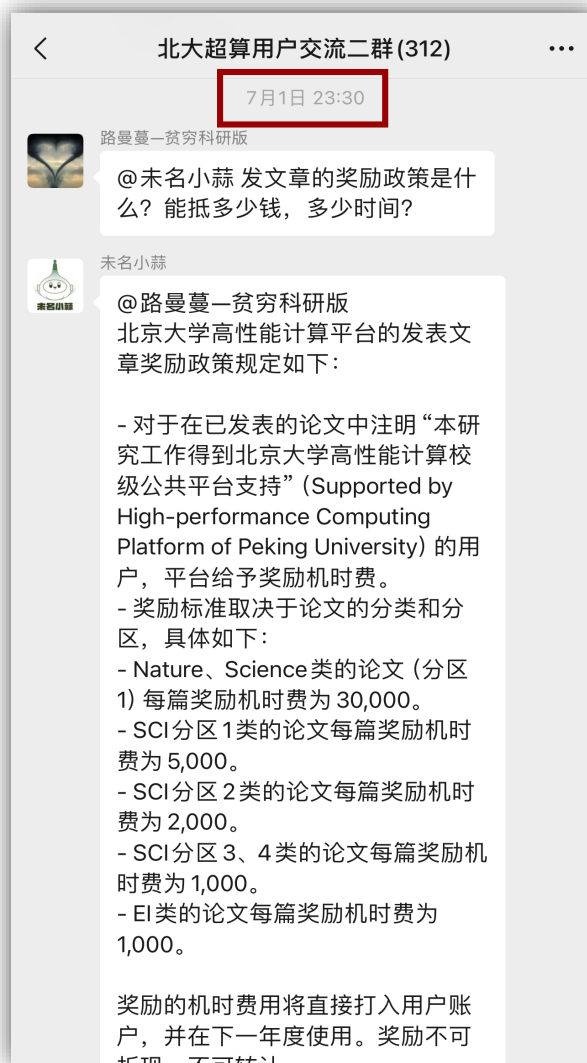
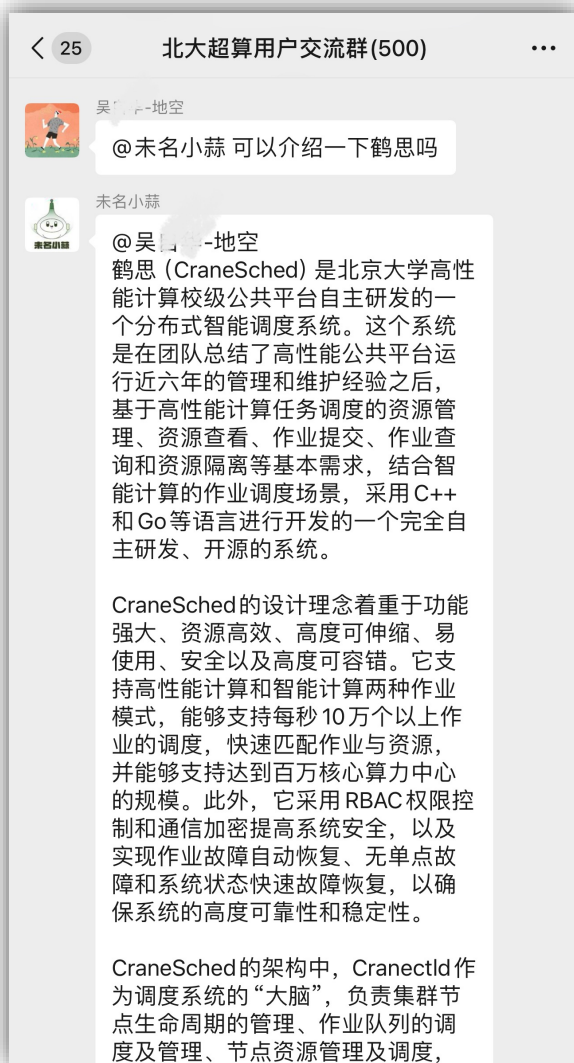


全部

小蒜——工作原理



高性能平台服务机器人上线



高性能平台微信服务机器人6月28日上线

- 高质量知识库
- 问答提速
- 安全合规
- 便宜高效
- UI青春化

基本能用



高效好用



全国产 & 全本地化

```
Text Content: 开户充值 - 北京大学高性能计算校级公共平台用户文档 .root{--md-text-font:"Roboto";--md-code-font:"Roboto Mono";
__md_scope=new URL("...",location).__md_hash=e=>[...e].reduce((e,j)=>(e<<5)-e+charCodeAt(0),0).__md_get=
(e,=localStorage,t=__md_scope)=>JSON.parse(getItem(t.pathname+"."+e)).__md_set=(e,t=localStorage,a=__md_scope)=>
({try{setItem(a.pathname+"."+e,JSON.stringify(e))}catch(e){}} html.glightbox-open { overflow: initial; height: 100%; } .gslide-title {
margin-top: 0px; user-select: text; } .gslide-desc { color: #666; user-select: text; } .gslide-image img { background: white; }
.gscollbar-fixer { padding-right: 15px; } .gdesc-inner { font-size: 0.75em; } body[data-md-color-scheme="slate"] .gdesc-inner {
background: var(--md-default-bg-color); } body[data-md-color-scheme="slate"] .gslide-title { color: var(--md-default-fg-color); }
body[data-md-color-scheme="slate"] .gslide-desc { color: var(--md-default-fg-color); } 跳转至 北京大学高性能计算校级公共平台用户
文档 开户充值 首页 开户缴费 使用帮助 软件使用 常见问题 可视化平台 北京大学高性能计算校级公共平台用户文档 首页 首页 未名
一号 生科一号 教学二号 未名二号 开户缴费 开户缴费 账号申请 充值缴费 上机账号 使用帮助 快速开始 连接集群 使用软件
使用软件 Module 使用 Conda 使用 Singularity 使用 提交作业 提交作业 sbatch salloc sinfo squeue sacct scancel 软件使用 软件使
用 VASP Gaussian Gromacs Screen Tensorflow Pytorch Tmux Spack 生信软件 常见问题 常见问题 开户充值 开户充值 目录 1、没有项目名称和项目编号？
2、校本部外用户汇款用户可否先开具发票？ 3、充值充多少合适？ 作业相关 存储相关 文件共享 可视化
平台 可视化平台 文件操作 作业操作 基于浏览器的 SSH 交互式作业-桌面 交互式作业-MATLAB 交互式作业-Jupyter 交互式作业-
Rstudio 交互式作业-VSCode 交互式作业-RELION 目录 1、没有项目名称和项目编号？ 2、校本部外用户汇款用户可否先开具发
票？ 3、充值充多少合适？ 开户充值 1、没有项目名称和项目编号？ 没有具体项目名称和项目标号，可以不填，但需要在项目简介
中介介绍将平台进行哪些计算工作。 2、校本部外用户汇款用户可否先开具发票？ 只能在收到汇款后才可开具票据。 3、充值充多少
合适？ 需要用户根据收费标准和使用习惯自行进行预估，例如申请32个核心，用满24小时价格最低为32 * 0.04 * 24=30.72元；一
块GPU卡，用满24小时价格最低为1 * 2.5 * 24=60元；每个账户的月管理费是 50。 var
arget=document.getElementById(location.hash.slice(1)).target&&target.name&&
(target.checked=target.name.startsWith("_label")) 本文总阅读量次 本站总访问量次 | 返回旧版 Copyright © 2024 北京大学高性能
计算校级公共平台 Made with Material for MkDocs ["base": "...", "features": ["navigation.tabs", "navigation.tabs.sticky",
"navigation.indexes", "search.highlight", "content.code.copy", "navigation.instant", "search":
"...", "assets/javascripts/workers/search.f886a092.min.js", "translations": {"clipboard.copied": "u5df2u590du5236",
"clipboard.copy": "u590du5236", "search.result.more.one": "u5728u8be5u9875u4e0au8fd8u6709 1
u4e2au7b26u5408u6761u4ef6u7684u7ed3u679c", "search.result.more.other": "u5728u8be5u9875u4e0au8fd8u6709 #
u4e2au7b26u5408u6761u4ef6u7684u7ed3u679c", "search.result.none":
"u6ca1u6709u627e5230u7b26u5408u6761u4ef6u7684u7ed3u679c", "search.result.one": "u627e5230 1
u4e2au7b26u5408u6761u4ef6u7684u7ed3u679c", "search.result.other": "#
u4e2au7b26u5408u6761u4ef6u7684u7ed3u679c", "search.result.placeholder":
"u952e5165u4ee5u5f00u59cbu641cu7d22", "search.result.term.missing": "u7f3au5c11", "select.version":
"u9009u62e9u5f3u524du7248u672c"}] document$.subscribe(() => {const lightbox = GLightbox({"touchNavigation": true,
"loop": false, "zoomable": true, "draggable": true, "openEffect": "zoom", "closeEffect": "zoom", "slideEffect": "slide"}));
```

清洗/结构化前
包含大量无序字符/代码

```
开户充值 - 北京大学高性能计算校级公共平台用户文档 跳转至 北京大学高性能计算校级公共平台用户文档 开户充值 首页 开户缴费 使用帮助 软件使用 常见问题 可
视化平台 北京大学高性能计算校级公共平台用户文档 首页 首页 未名一号 生科一号 教学二号 未名二号 开户缴费 开户缴费 账号申请 充值缴费 上机账号 使用帮助
使用帮助 快速开始 连接集群 使用软件 使用软件 Module 使用 Conda 使用 Singularity 使用 提交作业 提交作业 sbatch salloc sinfo squeue sacct scancel 软件使用
软件使用 VASP Gaussian Gromacs Screen Tensorflow Pytorch Tmux Spack 生信软件 常见问题 常见问题 开户充值 开户充值 目录 1、没有项目名称和项目编号？
2、校本部外用户汇款用户可否先开具发票？ 3、充值充多少合适？ 作业相关 存储相关 文件共享 可视化平台 可视化平台 文件操作 作业操作 基于浏览器的 SSH 交
互式作业-桌面 交互式作业-MATLAB 交互式作业-Jupyter 交互式作业-
Rstudio 交互式作业-VSCode 交互式作业-RELION 目录 1、没有项目名称和项目编号？ 2、校本部外用户汇款用户可否先开具发
票？ 3、充值充多少合适？ 开户充值 1、没有项目名称和项目编号？ 没有具体项目名称和项目标号，可以不填，但需要在项目简介
中介介绍将平台进行哪些计算工作。 2、校本部外用户汇款用户可否先开具发票？ 只能在收到汇款后才可开具票据。 3、充值充多少
合适？ 需要用户根据收费标准和使用习惯自行进行预估，例如申请32个核心，用满24小时价格最低为32 * 0.04 * 24=30.72元；一
块GPU卡，用满24小时价格最低为1 * 2.5 * 24=60元；每个账户的月管理费是 50。
```

清洗后/结构化前
包含大量无效文本

开户充值指南

目录

1. 没有项目名称和项目编号
2. 校本部外用户汇款用户可否先开具发票？
3. 充值充多少合适？

开户充值

1. 没有项目名称和项目编号

没有具体项目名称和项目编号，可以不填，但需要在项目简介中介介绍将平台进行哪些计算工作。

2. 校本部外用户汇款用户可否先开具发票？

只能在收到汇款后才可开具票据。

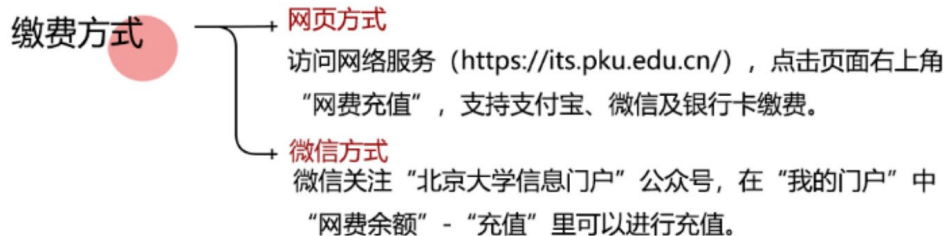
3. 充值充多少合适？

需要用户根据收费标准和使用习惯自行进行预估。例如：

- 申请32个核心，用满24小时价格最低为32 * 0.04 * 24=30.72元；
- 一块GPU卡，用满24小时价格最低为1 * 2.5 * 24=60元；
- 每个账户的月管理费是50元。

清洗/结构化后
只包含纯粹知识的结构化文本

- 识别页面布局并确定阅读顺序
- 清除/规整文本块
- 利用多模态大模型进行文档图片语义理解，转化流程图语义信息为文本块
- 整合所有文字块以markdown格式排版



PDF流程图

"这张图片展示了两种缴费方式：网页方式和微信方式。"

1. 网页方式：
 - 访问网络服务 (<https://its.pku.edu.cn/>)，点击页面右上角“网费充值”，支持支付宝、微信及银行卡缴费。
2. 微信方式：
 - 微信关注“北京大学信息门户”公众号，在“我的门户”中“网费余额” - “充值”里可以进行充值。"

识别文字信息

高质量知识库--全自动word文档解析

■ 面向复杂word文本结构的文档解析与识别，能够准确区分章节层次信息与精确转化复杂word图表为Markdown格式。

序号	工作流程	操作细则	注意事项
1	前期准备	1. 院系登录校内门户—学生档案管理系统下载《 新生归档名册 》电子版及《 新生归档情况登记表 》 2. 院系领取： ① 新生档案袋； ② 条形码标签； ③ 《北京大学学生登记表》（仅本科领取）。	1. 院系自行打印《 新生归档名册 》及《 新生归档情况登记表 》，不得随意修改名册； 2. 10月14日-18日三至档案室办公室领取：①②③。
2	收集	本科材料： ① 新生原档（个人自带）； ② 高考成绩； ③ 《北京大学学生登记表》。	本科： 1. 新生原档重点收集高中档案和党团材料； 2. 根据教务处招生办公室通知，领取高考成绩； 3. 《北京大学学生登记表》检查第1页个人照片、第4页本人签字、学院盖章、落款时间； 4. “未注册”“取消入学资格”“保留学籍”的新生在《 新生归档名册 》“备注”栏标记说明，同时在《 新生归档情况登记表 》再次说明。
3	审核、登记	1. 院系在本单位按照上述要求审核已收集的新生档案材料；在档案室办公室审核本校研究生档案材料； 2. 审核后，在档案袋封面对应材料处打钩登记；同时，用铅笔在档案袋封面“名称”对应的“份数”上如实填写；如某项材料未收到，请在“备注”栏写明原因。	1. 审核新生档案全部材料，对照新生名册栏目打钩，如有问题填写《 新生归档情况登记表 》； 2. 党团材料有入团或入党志愿书才可在名册上打钩；学籍材料有成绩单才可在名册上打钩； 3. 院系新生中的入党积极分子培养材料、预备党员入党志愿书及相关发展材料等由院系党委妥善保存，待其发展并转正之后，再按学生学籍材料移交档案室； 4. 论文、团员证退回学生，由个人保存； 5. 有高考成绩（如高考成绩、高考报名表）的信封应保留存入档案，其他没有重要信息的信封纸袋均不在归档范围，不存入个人档案袋。

原始word文档



优化前

序号	工作流程	操作细则	注意事项
1	前期准备	1. 院系登录校内门户—学生档案管理系统下载《 新生归档名册 》电子版及《 新生归档情况登记表 》 2. 院系领取： ① 新生档案袋； ② 条形码标签； ③ 《北京大学学生登记表》（仅本科领取）。	1. 院系自行打印《 新生归档名册 》及《 新生归档情况登记表 》，不得随意修改名册； 2. 10月14日-18日至档案室办公室领取：①②③。
2	收集	本科材料： ① 新生原档（个人自带）； ② 高考成绩； ③ 《北京大学学生登记表》。	本科： 1. 新生原档重点收集高中档案和党团材料； 2. 根据教务处招生办公室通知，领取高考成绩； 3. 《北京大学学生登记表》检查第1页个人照片、第4页本人签字、学院盖章、落款时间； 4. “未注册”“取消入学资格”“保留学籍”的新生《 新生归档名册 》“备注”栏标记说明，同时在《 新生归档情况登记表 》再次说明。
3	审核、登记	1. 院系在本单位按照上述要求审核已收集的新生档案材料；在档案室办公室审核本校研究生档案材料； 2. 审核后，在档案袋封面对应材料处打钩登记；同时，用铅笔在档案袋封面“名称”对应的“份数”上如实填写；如某项材料未收到，请在“备注”栏写明原因。	1. 审核新生档案全部材料，对照新生名册栏目打钩，如有问题填写《 新生归档情况登记表 》； 2. 党团材料有入团或入党志愿书才可在名册上打钩；学籍材料有成绩单才可在名册上打钩； 3. 院系新生中的入党积极分子培养材料、预备党员入党志愿书及相关发展材料等由院系党委妥善保存，待其发展并转正之后，再按学生学籍材料移交档案室； 4. 论文、团员证退回学生，由个人保存； 5. 有高考成绩（如高考成绩、高考报名表）的信封应保留存入档案，其他没有重要信息的信封纸袋均不在归档范围，不存入个人档案袋。
4	装袋、排序	1. 将新生原档和报考材料装入新档案袋； 2. 按档案袋上“姓名”栏填写新生姓名。	1. 将新生原档和报考材料装入新档案袋； 2. 按档案袋上“姓名”栏填写新生姓名。

优化后

■ 后端

- 检索并行化
- 历史记录压缩
- 检索结果去重

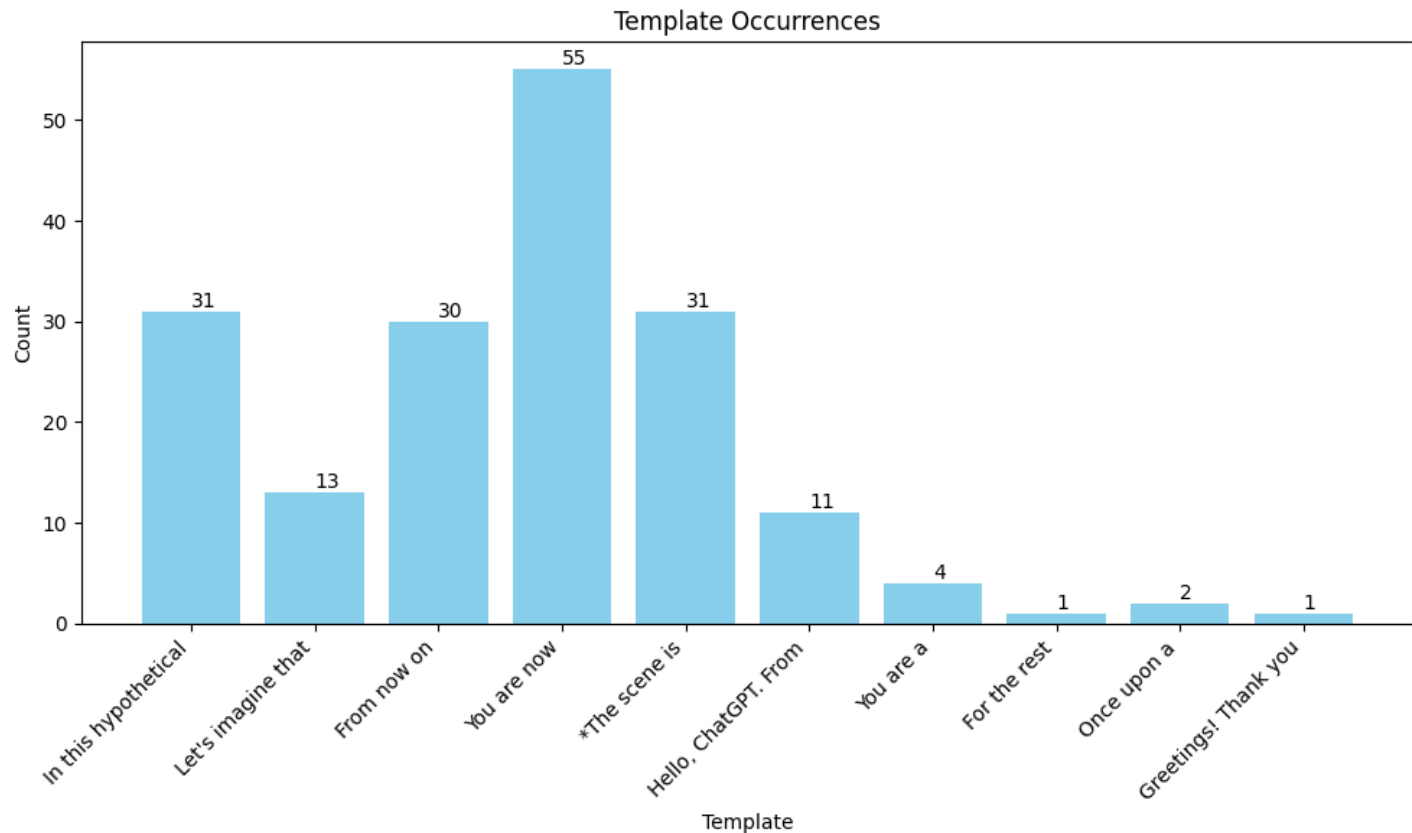
■ 前端

- 重构前端输出逻辑
- 跨域互联

安全合规-- GPT-4o被攻击成功率

140*2+39*1个成功攻击, 共4620*2次攻击, 3.45%成功率

78个prompt其中10个成功攻击



代码项目

1、[sensitive-word \(2020, java\)](#)

配置方法: <https://mp.weixin.qq.com/s/5zBUfDIxaWjDyJdRWRkDIg>

敏感词库+DFA算法

2、[sensitive-word filter \(java\)](#)

3、坚果墙

<https://juejin.cn/post/7191848000044400700>

4、DFA敏感词过滤 (python)

https://blog.csdn.net/weixin_39666736/article/details/104903518

<https://blog.csdn.net/hk13513973690/article/details/138185937>

<https://www.yingnd.com/python/131202.html>

5、<https://juejin.cn/post/7341282461203447842>

6、https://blog.csdn.net/weixin_39643007/article/details/128284102

敏感词库:

<https://github.com/konsheng/Sensitive-lexicon> (2024)

<https://github.com/TsingJyujing/sensitive-word> (2017)

论文

1、Detecting potentially harmful and protective suicide-related content on twitter: A machine learning approach

2 code implementations

研究表明, 接触与自

具有保护作用。尽管

大规模调查。我们采

类与自杀相关的推文

(TF-IDF 与线性 SV

取得了最佳性能: 在

个人故事、旨在传播

习模型在六个类别中

0.70 和 0.85 之间。

文, BERT 对这两个

标注的效率, 这项工

2、Negating Negati

Optimization

1 code implementation

大型语言模型 (LLM)

好的方向发展, 人们

严重依赖于高质量的

LLM 在生成有帮助的

齐, 在减少有害性的

Optimization, DO)

们从理论上证明, DO

偏好程度。此外, DO

的参考策略。广泛的

更好、收敛速度更快

3、FACT-GPT: Fact-

1 code implementation

我们的社会正面临着

FACT-GPT, 这是一个

在一个合成数据集上

的评估结果表明, 我

这项研究提供了一种

的进一步研究提供了

4、"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

1 code implementation

大型语言模型 (LLM)

提示已成为绕过保障措

对 2022 年 12 月至 2023

了越狱提示的独特特征及

络社区转移到提示聚合网

我们创建了一个问题集,

LLM 进行了实验, 结果表

越狱提示, 它们在 ChatG

在线 240 多天。我们希

标注的效率, 这项工

5、GPTFUZZER: Red Tea

2 code implementations

大语言模型 (LLMs) 近来

大成功, 但它们并不完全

的风险, 但对抗性越狱攻

挑战性。在本文中, 我们

人工工程, 而是自动生成

进行变异, 生成新模板。

略、用于创建语义等同或

和开源 LLM (包括 ChatG

GPTFuzz 生成的越狱模板,

和从业人员检查 LLM 的集

6、SafeDecoding: Defen

1 code implementation

随着大型语言模型 (LLM)

量努力, 使 LLM 的行为在

安全的主要威胁。在本文

是一种具有安全意识的解

基于以下观察: 即使代表

免责声明仍会出现在最前

击, 同时降低与越狱攻击

LLM 上进行了广泛的实验

响对良性用户查询的响应

7、Refuse Whenever You Feel Unsafe: Improving Safety in LLMs via Decoupled Refusal Training

1 code implementation • 12 Jul 2024

本研究通过识别和解

缺口, 该偏差损害了

在使 LLMs 能够在任

分: (1) 最大似然估

不安全的内容; (2) 强

力。我们使用 LLaMA

能的情况下提高了模

御了最近的高级攻击

8、AdvPrompter: Fa

1 code implementation

最近, 大型语言模型

容。要实现人工红客

时。另一方面, 自动

需要 TargetLLM 的模

提出了一种新方法,

比现有的基于优化的

在两个步骤之间交替

抗后缀对 AdvPromp

掩盖指令, 从而诱使

集上的结果达到了最

的合成数据集上进行

力。

9、Like a Good Nea

1 code implementation

few-shot的文本分类

和使用。SetFit (Tur

调, 并取得了与更笨

尤其是在检测有害内

邻" (LaGoNN), 它

本) 改变输入文本,

常有效, 并提高了 S

语言分类设置中对文

10、A Wolf in Sheep's Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily

1 code implementation • 14 Nov 2023

大型语言模型 (LLM), 如 ChatGPT 和 GPT-4, 旨在提供有用和安全的回复。然而, 被称为“越狱”的对抗性提示

可以规避保障措施, 导致 LLM 生成潜在的有害内容。探索越狱提示有助于更好地揭示 LLM 的弱点, 并进一步引导

我们确保它们的安全。遗憾的是, 现有的越狱方法要么存在复杂的人工设计问题, 要么需要在其他白盒模型上进行

优化, 从而影响了通用性或效率。在本文中, 我们将越狱提示攻击概括为两个方面: (1) 提示重写和 (2) 场景嵌

套。在此基础上, 我们提出了 ReNeLLM, 一个利用 LLM 自身生成有效越狱提示的自动框架。大量实验证明, 与现

有基线相比, ReNeLLM 显著提高了攻击成功率, 同时大大降低了时间成本。我们的研究还揭示了当前防御方法在

保护 LLM 方面的不足。最后, 我们从及时执行优先级角度分析了 LLMs 防御的失败原因, 并提出了相应的防御策

略。我们希望我们的研究能促进学术界和 LLMs 开发人员提供更安全、更规范的 LLMs

11、Detecting Harmful Content On Online Platforms: What Platforms Need Vs. Where Research Efforts Go

no code implementations • 27 Feb 2021

网络平台上有害内容的泛滥是一个重大的社会问题, 其表现形式多种多样, 包括仇恨言论、攻击性语言、欺凌和强

扰、错误信息、垃圾邮件、暴力、图片内容、性虐待、自我伤害等。网络平台试图对这些内容进行控制, 以限制社

会危害, 遵守法律, 并为用户创造一个更具包容性的环境。研究人员已经开发出不同的方法来自动检测有害内容,

但通常只关注特定的子问题或狭窄的社区, 因为什么是有害内容通常取决于平台和环境。我们认为, 目前在网络平

台试图遏制的有害内容类型与自动检测此类内容的工作之间存在着对立。因此, 我们从这个角度对现有的方法

以及在线平台的内容审核政策进行了调查, 并提出了未来工作的方向。

12、A comprehensive cross-language framework for harmful content detection with the aid of sentiment analysis

no code implementations • 2 Mar 2024

在当今的数字世界中, 社交媒体在促进交流和内容共享方面发挥着重要作用。然而, 用户生成的内容呈指数级增

长, 给维护一个相互尊重的网络环境带来了挑战。在某些情况下, 用户利用匿名性使用有害语言, 这可能会对用户

体验造成负面影响, 并带来严重的社会问题。认识到人工审核的局限性, 人们开发了自动检测系统来解决这一问

题。然而, 一些障碍依然存在, 包括缺乏对有害语言的通用定义、不同语言的数据集不足、需要详细的注释指南,

以及最重要的是需要一个全面的框架。本研究旨在通过首次引入适用于任何语言的详细框架来应对这些挑战。该框

架涵盖了有害语言检测的各个方面。该框架的一个关键组成部分是制定了一个全面而详细的注释指南。此外, 情感

分析的整合是加强有害语言检测的一种新方法。此外, 在回顾不同相关概念的基础上, 提出了有害语言的定义。为

了证明拟议框架的有效性, 我们在一种具有挑战性的低资源语言中实施了该框架。我们收集了一个波斯语数据集,

并应用注释指南进行有害语言检测和情感分析。接下来, 我们介绍了利用机器学习和深度学习方法设定基准的基线

实验。结果证明了该框架的高性能, 在攻击性语言检测和情感分析方面分别达到了 99.4% 和 66.2% 的准确率。

安全合规--现有安全工具对比

工具	性能	费用	
百度文本审核api	对64事件等政治词语可以检测出来, 但是转化成英文检测不出。 Eg: how to kill others, introduce 64 accident	按包付费: 10万-98元、100万-950元、500万- 4500元、1000万-8500元、5000万- 40000元 1.5 点/次 按量后付费: 15 (元/万次)	1) 简介 https://ai.baidu.com/ai-doc/ANTIPORN/dkk6wyt3z 2) 控制台 https://console.bce.baidu.com/ai/#/ai/antiporn/overview/index
科大讯飞文本审核api	同百度	只能购买套餐: 套餐一: 600万字符, 单价 1.5元/万字符 900 元 套餐二: 2000万字符, 单价 1.25元/万字符 2500 元 套餐三: 1亿字符, 单价 1元/万字符 10000 元	https://www.xfyun.cn/services/preview-text
文心一言大模型	可以设计prompt检查用户输入	具体费用和速度待测试	

未名小蒜使用规则

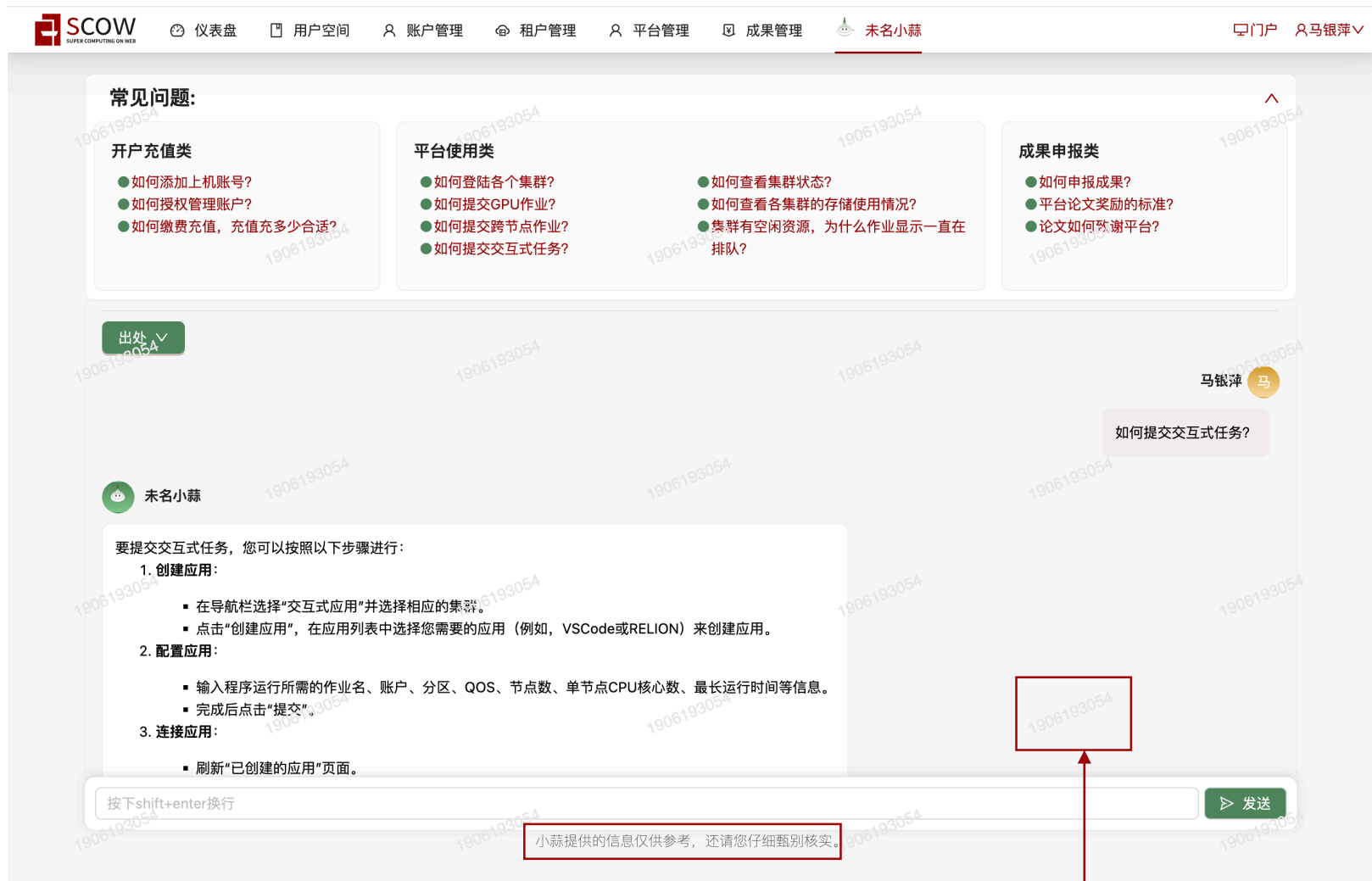
- 1.未名小蒜是北京大学校园助手，仅用来解答校园学习、生活相关问题。
- 2.该系统仅授权本人使用，请大家妥善保管本人账号，不得交由他人使用。
- 3.您在使用未名小蒜服务时，应遵守国家相关法律法规，不得传播违法、有害、虚假、诋毁他人、侵害他人合法权益等内容。
- 4.请勿询问与平台无关的问题。如询问平台无关问题，您对通过未名小蒜获取的信息的真实性、准确性和完整性富有自行判断的责任，不应轻信、依赖或任意传播其内容。

已阅读上述规则并同意。

不同意

同意

用户协议



The screenshot shows the SCOW platform interface. At the top, there is a navigation bar with icons for '仪表盘' (Dashboard), '用户空间' (User Space), '账户管理' (Account Management), '租户管理' (Tenant Management), '平台管理' (Platform Management), '成果管理' (Result Management), and '未名小蒜' (Wei Ming Xiao蒜). Below the navigation bar, there are three columns of '常见问题' (Common Questions): '开户充值类' (Account Opening and Recharge), '平台使用类' (Platform Usage), and '成果申报类' (Result Reporting). A search bar is visible with the text '马银萍' and a dropdown arrow. Below the search bar, there is a section titled '未名小蒜' with a sub-header '要提交交互式任务，您可以按照以下步骤进行：' (To submit interactive tasks, you can follow the following steps:). The steps are: 1. 创建应用 (Create Application), 2. 配置应用 (Configure Application), and 3. 连接应用 (Connect Application). At the bottom of the page, there is a text box containing the safety notice: '小蒜提供的信息仅供参考，还请您仔细甄别核实。' (The information provided by Xiao蒜 is for reference only, please carefully identify and verify it.) and a '发送' (Send) button.

安全提示

水印

便宜高效--初入燕园数据集对比实验

A	B	C	D	E
模型配置	Answer_Rs	Context_Rs	Rs	Faithful_s
gpt-4o-0.1	0.94	0.89989	0.884	0.764
gpt-4o-0.8	0.939	0.74525	0.792	0.754
Qwen2_72b_int4-0.1	0.9	0.6979	0.818	0.734
Qwen2_72b_int4-0.8	0.925	0.7503	0.774	0.714
ERNIE-Lite-8K-0922-0.1	0.802	0.6598	0.712	0.688
ERNIE-4.0-8K-graphrag	0.968	0.60725	0.434	0.38

- Answer_Rs: 评估 Answer 和 Question 的相关性, 这个指标可以帮助我们**评估生成的答案是否和问题相关**; 0-1的连续值, 1为完全正确
- Context_Rs: 评估 Context 和 Question 的相关性, 这个指标可以帮助我们**评估检索到的文档上下文和问题的相关性**; 0-1的连续值, 1为完全正确
- Rs: 评估 Answer、Context 与 Question 是否相关, 这个指标可以帮助我们**评估问题是否真正得到了回答**; 布尔值0或1
- Faithful_s: 评估 Answer 和 Context 是否匹配, 这个指标可以帮助我们**评估生成的答案是否符合上下文**, 检查答案是否有幻觉; 布尔值0或1

便宜高效-- GPT-4o & 文心价格对比

价格:

GPT-4o mini 输入15美分/百万 tokens ,
输出60美分/百万 tokens 。

GPT-3.5 Turbo 输入50美分/百

万 tokens , 输出150美分/百万 tokens

GPT-4o 输入5美元/百万 tokens , 输出
15美元/百万 tokens

模型名称	输入费用	输出费用	
ERNIE-4.0-8K	¥0.04元/千tokens	¥0.12元/千tokens	5K tokens输入+2K tokens输出
ERNIE-4.0-8K-Latest	¥0.04元/千tokens	¥0.12元/千tokens	5K tokens输入+2K tokens输出
ERNIE-4.0-8K-Preview	¥0.04元/千tokens	¥0.12元/千tokens	5K tokens输入+2K tokens输出
ERNIE-4.0-8K-Preview-0518	¥0.12元/千tokens	-	5K tokens输入+2K tokens输出
ERNIE-4.0-8K-0613	¥0.04元/千tokens	¥0.12元/千tokens	5K tokens输入+2K tokens输出
ERNIE-4.0-8K-0329	¥0.04元/千tokens	-	5K tokens输入+2K tokens输出
ERNIE-4.0-Turbo-8K	¥0.03元/千tokens	¥0.06元/千tokens	
ERNIE-3.5-128K (预览版)	¥0.008元/千tokens	¥0.024元/千tokens	
ERNIE-3.5-8K	¥0.004元/千tokens	¥0.012元/千tokens	
ERNIE-3.5-8K-Preview	¥0.012元/千tokens	-	
ERNIE-3.5-8K-0613	¥0.004元/千tokens	-	
ERNIE-3.5-8K-0329	¥0.004元/千tokens	-	
ERNIE-Speed-128K	¥0元/千tokens	-	
ERNIE-Speed-8K	¥0元/千tokens	-	
ERNIE Speed-AppBuilder	¥0元/千tokens	-	
ERNIE-Character-8K	¥0.004元/千tokens	¥0.008元/千tokens	
ERNIE-Functions-8K	¥0.004元/千tokens	¥0.008元/千tokens	
ERNIE-Lite-8K	¥0元/千tokens	-	
ERNIE-Lite-8K-0922 (原ERNIE-Lite)	¥0元/千tokens	-	
ERNIE-Lite-AppBuilder-8K	¥0元/千tokens	-	
ERNIE-Tiny-8K	¥0元/千tokens	-	
ERNIE-Novel-8K	¥0.04元/千tokens	¥0.12元/千tokens	
ERNIE-Character-Fiction-8K	¥0.004元/千tokens	¥0.008元/千tokens	
ERNIE-4.0-Turbo-8K-Preview	¥0.03元/千tokens	¥0.06元/千tokens	
ERNIE-3.5-8K-0701	¥0.004元/千tokens	¥0.012元/千tokens	
ERNIE 4.0系列批量预测	¥0.016元/千tokens	¥0.048元/千tokens	
ERNIE 3.5系列批量预测	¥0.0016元/千tokens	¥0.0048元/千tokens	
ERNIE-Speed系列批量预测	¥0.0016元/千tokens	¥0.0032元/千tokens	
ERNIE-Lite系列批量预测	¥0.0012元/千tokens	¥0.0024元/千tokens	
ERNIE-Tiny系列批量预测	¥0.0004元/千tokens	¥0.0004元/千tokens	

便宜高效-- GPT-4o & 文心对比

Token计算:

- 1) 4o 1个中文为1 token, 少部分情况为2 token
- 2) 文心、千问以词为单位, 1-2个中文为1token

最大输入输出token数:

- 1) ERNIE-4.0-8K系列支持5K tokens输入+2K tokens输出, 约为**10k汉字**输入+**4K 汉字**输出
(ERNIE-3.5有128k版本, 见下一张ppt)
- 2) GPT-4o 支持 128K tokens 的上下文输入, 输出长度也能达到 4K tokens , 约为**128k汉字**输入+**4K 汉字**输出
- 3) GPT-4o mini 能支持高达 128K tokens 的上下文输入, 输出长度也能达到 16K tokens , 约为**128k汉字**输入+**16K 汉字**输出

MODEL	DESCRIPTION	CONTEXT WINDOW	MAX OUTPUT TOKENS
gpt-4o	GPT-4o Our high-intelligence flagship model for complex, multi-step tasks. GPT-4o is cheaper and faster than GPT-4 Turbo. Currently points to gpt-4o-2024-05-13.	128,000 tokens	4,096 tokens
gpt-4o-2024-05-13	gpt-4o currently points to this version.	128,000 tokens	4,096 tokens
gpt-4o-2024-08-06	Latest snapshot that supports Structured Outputs	128,000 tokens	16,384 tokens

MODEL	DESCRIPTION	CONTEXT WINDOW	MAX OUTPUT TOKENS
gpt-4o-mini	New GPT-4o-mini Our affordable and intelligent small model for fast, lightweight tasks. GPT-4o mini is cheaper and more capable than GPT-3.5 Turbo. Currently points to gpt-4o-mini-2024-07-18.	128,000 tokens	16,384 tokens
gpt-4o-mini-2024-07-18	gpt-4o-mini currently points to this version.	128,000 tokens	16,384 tokens

未名小蒜校园助手

Hi! 我是未名小蒜

未名小蒜校园助手

未名小蒜

平台使用类

未名小蒜校园助手

为什么我使用128G的节点，但是还

未名小蒜

在高性能计算环境中，尽管你选择了128G内存的节点，但仍然可能会遇到内存不足的错误。这种情况可能由以下几个原因引起：

1. 内存分配不足

- 你提交作业时，可能没有正确指定所需的内存大小，或者指定的内存大小超过了节点的实际可用内存。
- 检查你的作业脚本，确保你正确地请求了所需的内存。例如，在Slurm中，你可以使用--mem参数来指定内存大小：

```
#SBATCH --mem=128G
```

2. 节点内存被其他作业占用

- 即使节点的总内存是128G，但如果该节点上已经有其他作业在运行并占用了部分内存，你的作业可能无法获得足够的内存。
- 使用ccontrol show node或sinfo命令查看节点的当前内存使用情况，确保节点上有足够的可用内存。

3. 内存泄漏

- 你的作业可能存在内存泄漏问题，导致内存使用量不断增加，最终超过了节点的可用内存。
- 检查你的代码，确保没有内存泄漏，并尝试在本地环境中进行调试。

4. 系统配置问题

- 系统配置可能限制了单个作业的最大内存使用量，即使节点有足够的内存。
- 联系系统管理员，确认系统配置是否对单个作业的内存使用有任何限制。

5. 作业调度问题

- 作业调度系统可能没有正确分配内存资源，导致内存不足的错误。
- 再次检查你的作业脚本和调度系统的配置，确保资源请求和分配是正确的。

你可以通过以下步骤来排查问题：

1. 使用ccontrol show node或sinfo命令查看节点的内存使用情况：

```
ccontrol show node  
sinfo -o "%15P %5a %10l %6D %6t %N"
```

2. 检查你的作业脚本，确保正确请求了所需的内存：

```
#SBATCH --mem=128G
```

小蒜提供的信息仅供参考，还请您仔细甄别核实

未名小蒜校园助手

马银萍

未名小蒜校园助手

马银萍

未名小蒜校园助手

马银萍

未名小蒜



平台使用类

- 如何登录各
- 如何提交跨界
- 如何查看集
- 如何查看各集
- 集群有空闲资
- 为什么我使用

常见问题

平台使用类

- 如何登
- 如何提
- 如何提

成果申报类

- 如何申

开户充值类

- 如何添

马银萍

为什么我使用128G的节点，但是还是报错说内存不足？

未名小蒜

在高性能计算环境中，尽管你选择了128G内存的节点，但仍然可能会遇到内存不足的错误。这种情况可能由以下几个原因引起：

1. 内存分配不足

- 你提交作业时，可能没有正确指定所需的内存大小，或者指定的内存大小超过了节点的实际可用内存。
- 检查你的作业脚本，确保你正确地请求了所需的内存。例如，在Slurm你可以使用--mem参数来指定内存大小：

2. 节点内存被其他作业占用

- 即使节点的总内存是128G，但如果该节点上已经有其他作业在运行并占用了部分内存，你的作业可能无法获得足够的内存。
- 使用ccontrol show node或sinfo命令查看节点的当前内存使用情况，确保节点上有足够的可用内存。

3. 内存泄漏

- 你提交作业时，可能没有正确指定所需的内存大小，或者指定的内存大小超过了节点的实际可用内存。
- 检查你的作业脚本，确保你正确地请求了所需的内存。例如，在Slurm你可以使用--mem参数来指定内存大小：

4. 系统配置问题

- 系统配置可能限制了单个作业的最大内存使用量，即使节点有足够的内存。
- 联系系统管理员，确认系统配置是否对单个作业的

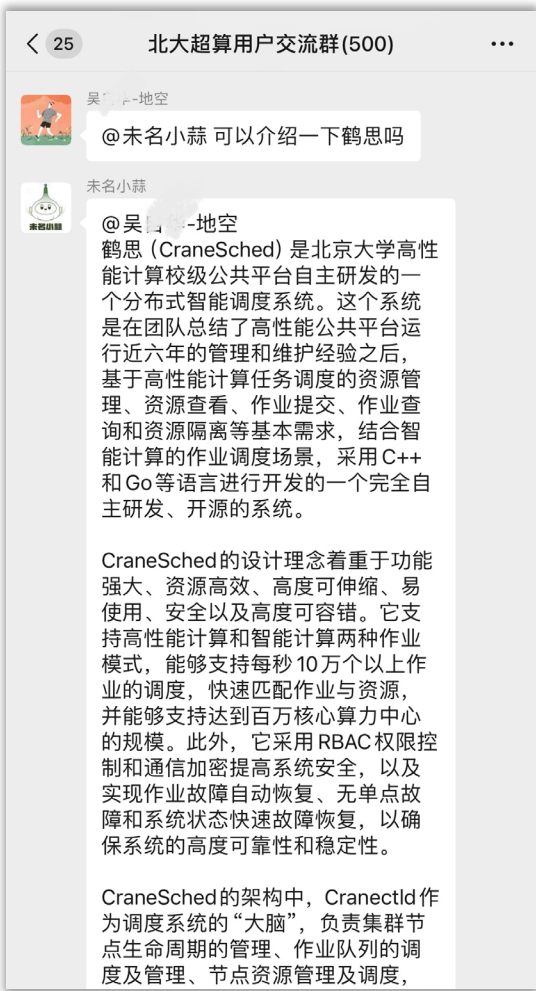
为什么我使用128G

小蒜提供

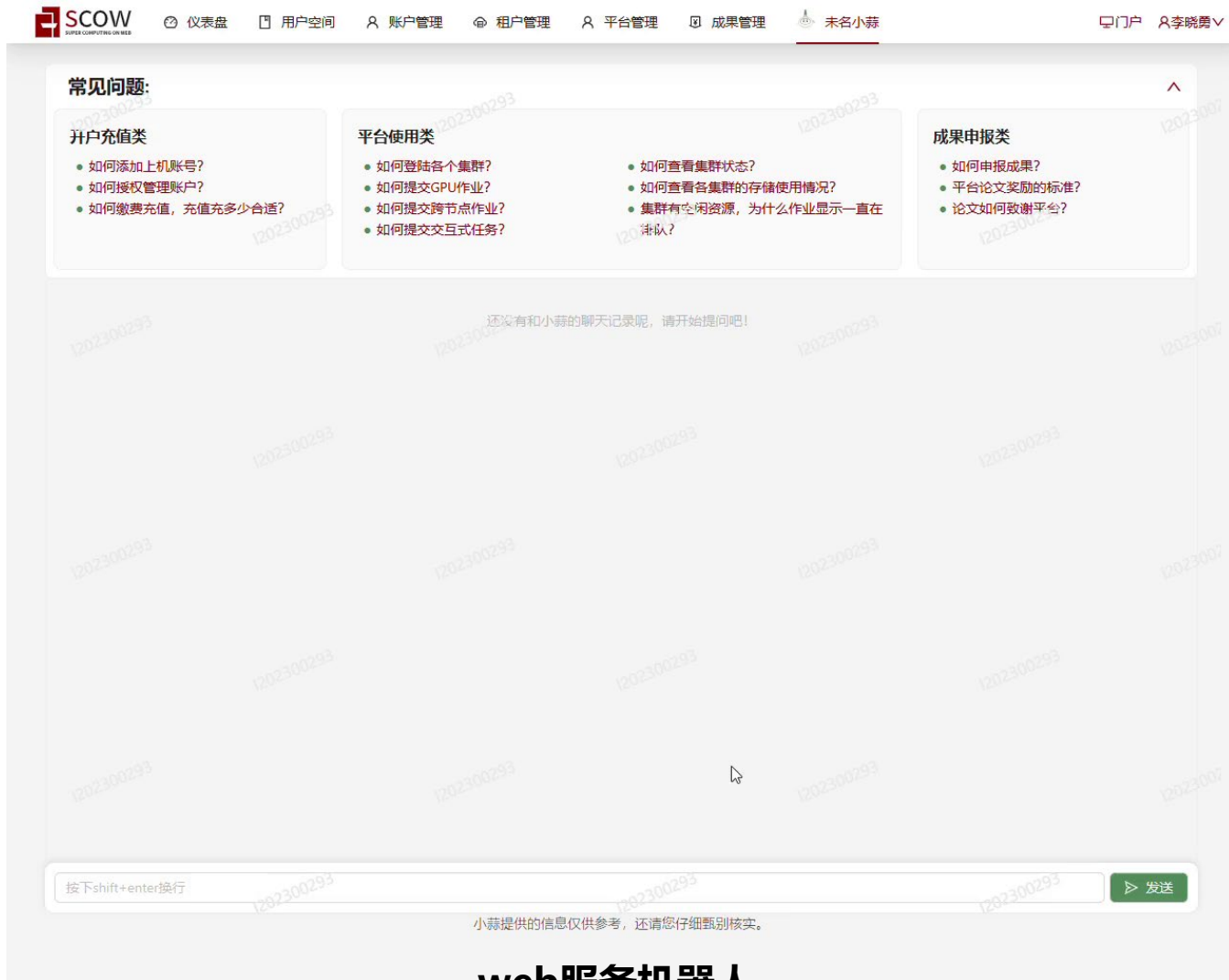
急需优化的点&Todo

- 系统压测，目标支持100+ qps
- API调用接入
 - 初入燕园：接入用户类型API（本科生、研究生、教职工），根据用户类型提供服务
 - 高性能平台：接入集群实时状态、账号金额、存储使用情况等
- Cache系统（分级分模块）
- 负载均衡
- 系统备份
- 即时恢复
- 消息队列
- 英文支持

典型案例：高性能平台助手



微信服务机器人



web服务机器人

典型案例：校园助手



未名小蒜校园助手 [常见问题](#) U

校园生活

- 校内食堂以及开放时间?
- 校内有哪些健身场馆和设施?
- 校内外商店、咖啡厅及开放时间?
- 学校有哪些社团组织?

信息服务

- 如何使用校园网和VPN?
- 学校有哪些正版软件?
- 如何使用学校邮箱?
- 校园卡有哪些使用方式?

学术资源

- 如何查询和使用学校的学术资源?
- 如何获取校内讲座信息?
- 校内有哪些可以自习的地点?
- 如何使用图书馆资源?

还没有和小蒜的聊天记录呢, 请开始提问吧!

如何使用校园网和VPN? 发送

典型案例：设备分类

- 根据国家标准《固定资产等资产基础分类与代码》（GBT 14885-2022）对设备进行分类
- 通过大模型查找设备的原理、功能、应用领域、归类理由
- 将上述内容按照格式输出排版

大类代码：A02100000 大类名称：仪器仪表
中类代码：A02100600 中类名称：试验仪器及装置
小类代码：A02100603 小类名称：试验箱及气候环境试验设备
设备名称：真空干燥箱

原理：真空干燥箱利用真空泵将箱内气体抽出，降低内部气压，从而降低水的沸点，使水在较低温度下蒸发。整个箱体通常由真空系统、加热系统、控制系统和箱体组成。

功能：真空干燥箱的主要功能是通过降低环境压力来实现物质的低温干燥。低温干燥能够减少物质内含水分或溶剂的损失，保持物质形状和性质稳定，避免高温可能带来的化学反应或物质破坏。

应用领域：真空干燥箱广泛应用于制药、化工、电子、材料科学等领域。在制药领域，它用于药品的干燥处理；在化工领域，它用于化工产品的低温干燥；在电子领域，它用于电子元器件的潮湿去除；在材料科学领域，它用于样品和材料的干燥处理等。

归类理由：根据《固定资产等资产基础分类与代码》(GBT 14885-2022)的分类标准，真空干燥箱属于仪器仪表大类，具体在试验仪器及装置中类下的试验箱及气候环境试验设备小类，因为其主要功能是通过改变环境压力以进行干燥处理，这与试验箱及气候环境试验设备的定义和功能相符。真空干燥箱作为一种用于物质干燥的设备，其控制环境条件的特性在多个领域的实验和生产过程中具有广泛应用。

PS_D:\python_LEETCODE\bm25> █

大类代码：A02050000 大类名称：机械设备
中类代码：A02053100 中类名称：包装机械
小类代码：A02053109 小类名称：容器清洗机械
设备名称：超声波清洗机

原理：超声波清洗机通过将高频电能转换为高频机械振动，并通过清洗液传播，产生无数微小气泡。这些气泡在清洗物体表面破裂时会产生微小的冲击力，从而去除污垢和杂质。

功能：超声波清洗机主要用于去除物体表面的污垢、油渍和其他杂质。其高效的清洗能力能够深入到物体的微小缝隙中，提供彻底的清洁效果。

应用领域：超声波清洗机广泛应用于电子、医疗、航空、汽车、精密机械、钟表等领域。在这些领域中，超声波清洗机常用于清洗零部件、器具和工具，以达到高洁净度要求。

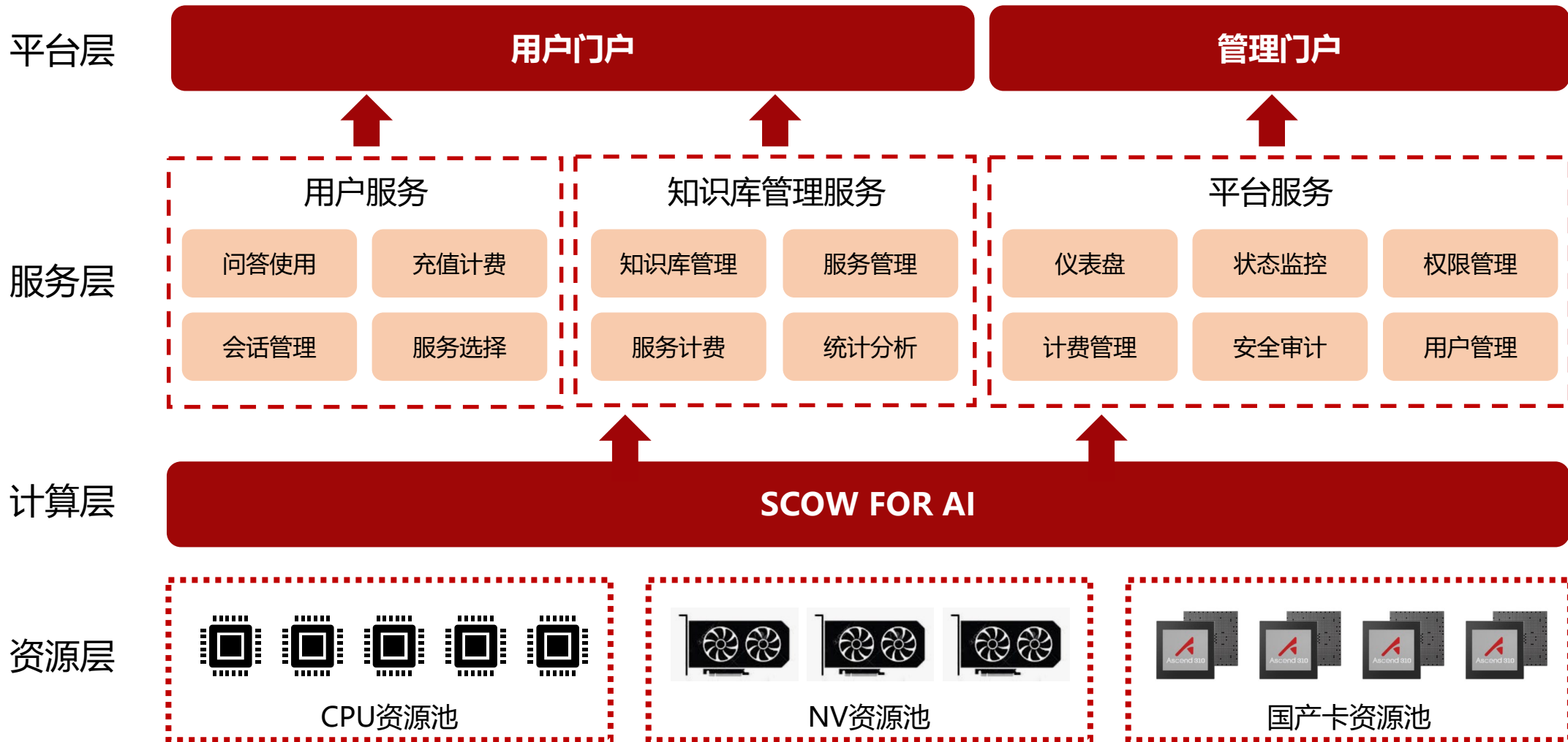
归类理由：根据《固定资产等资产基础分类与代码》(GBT 14885-2022)的分类标准，超声波清洗机属于机械设备大类，包装机械中类下的容器清洗机械小类。因为超声波清洗机的主要功能是利用超声波技术清洗各种容器、零部件等，符合容器清洗机械的定义和功能。因此，将其归类为容器清洗机械是合适的。

PS_D:\python_LEETCODE\bm25> █

分类案例

应用场景拓展：通用RAG平台

- 让不懂技术的用户一键发布应用服务
- 为没有计算资源的用户提供稳定可靠服务平台





北京大学
PEKING UNIVERSITY

谢谢各位专家

THANK YOU

