北京大学
PEKING UNIVERSITY

面向算力网络的智算调度系统Crane及算力中心门户和管理平台SCOW

樊春 2023.6.27 恩施

北京大学计算中心/北京大学长沙计算与数字经济研究院

2022年度国内高校校级算力中心建设情况

学校	项目总计(元)
复旦大学	2.4亿
华南理工大学	1.9亿
清华大学	1.2亿
北京师范大学	6800万
兰州大学	6000万
南京农业大学	5331万
西安电子科技大学	5180万
西北农林科技大学	4800万
电子科技大学	4800万
中山大学	4567万
天津大学	4220万
河海大学	3860万
北京邮电大学	3490万
北京师范大学(珠海)	3300万

算力中心面临的困难

截至2022年6月底，我国在用数据中心机架总规模超过590万标准机架，服务器规模约2000万台，算力总规模超过150EFlops，排名全球第二。

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

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

算力中心门户和管理平台SCOW

难以融合
别，造成
算力浪费

SCOW——Super Computing On Web

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特色——图形化界面，使用方便

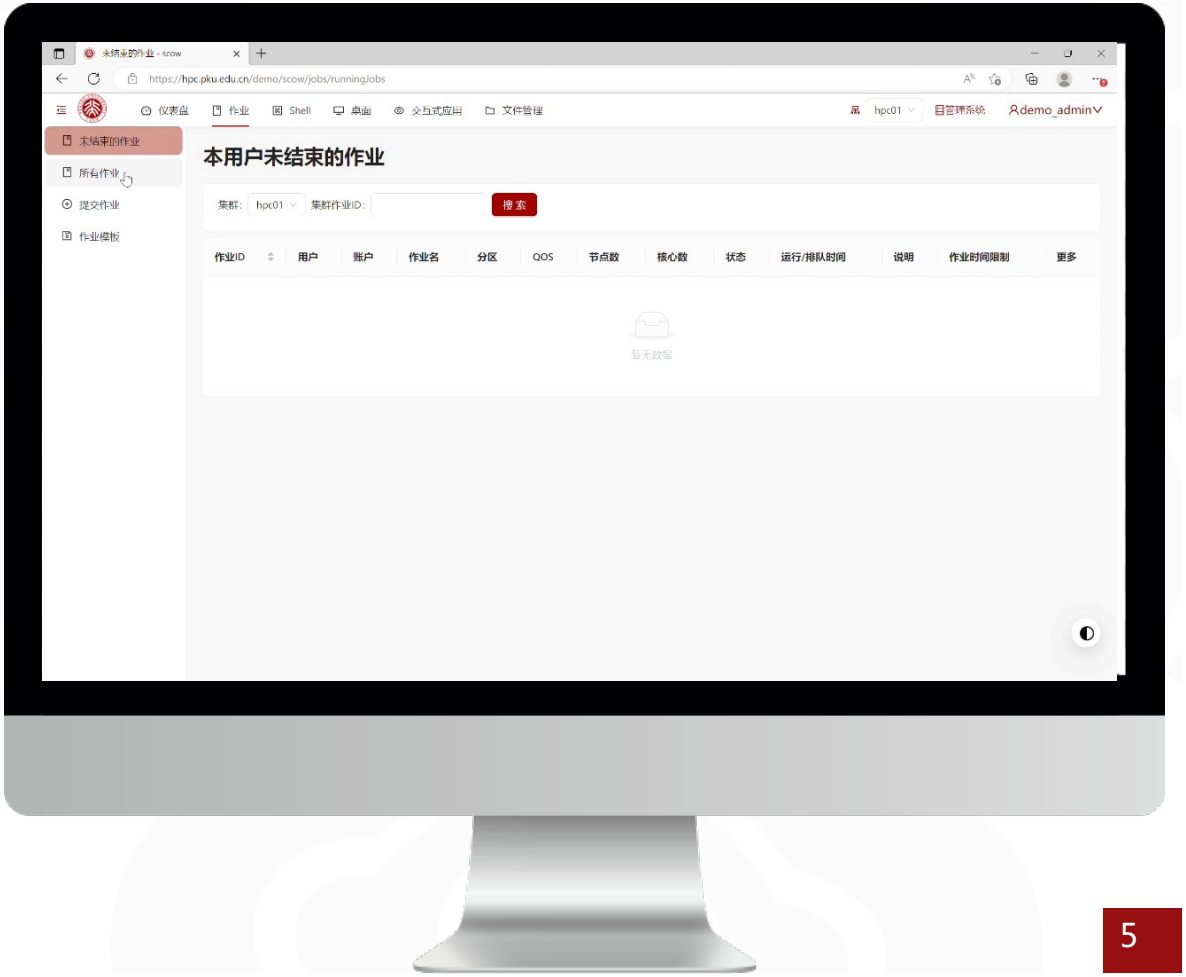
北京大学 PEKING UNIVERSITY

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 VNC连接
- Web 文件管理
- Web 交互式应用连接 (VSCODE)
- Web 启动交互式应用
- Web 作业管理



SCOW特色——功能丰富，管理简单

北京大学 PEKING UNIVERSITY

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。

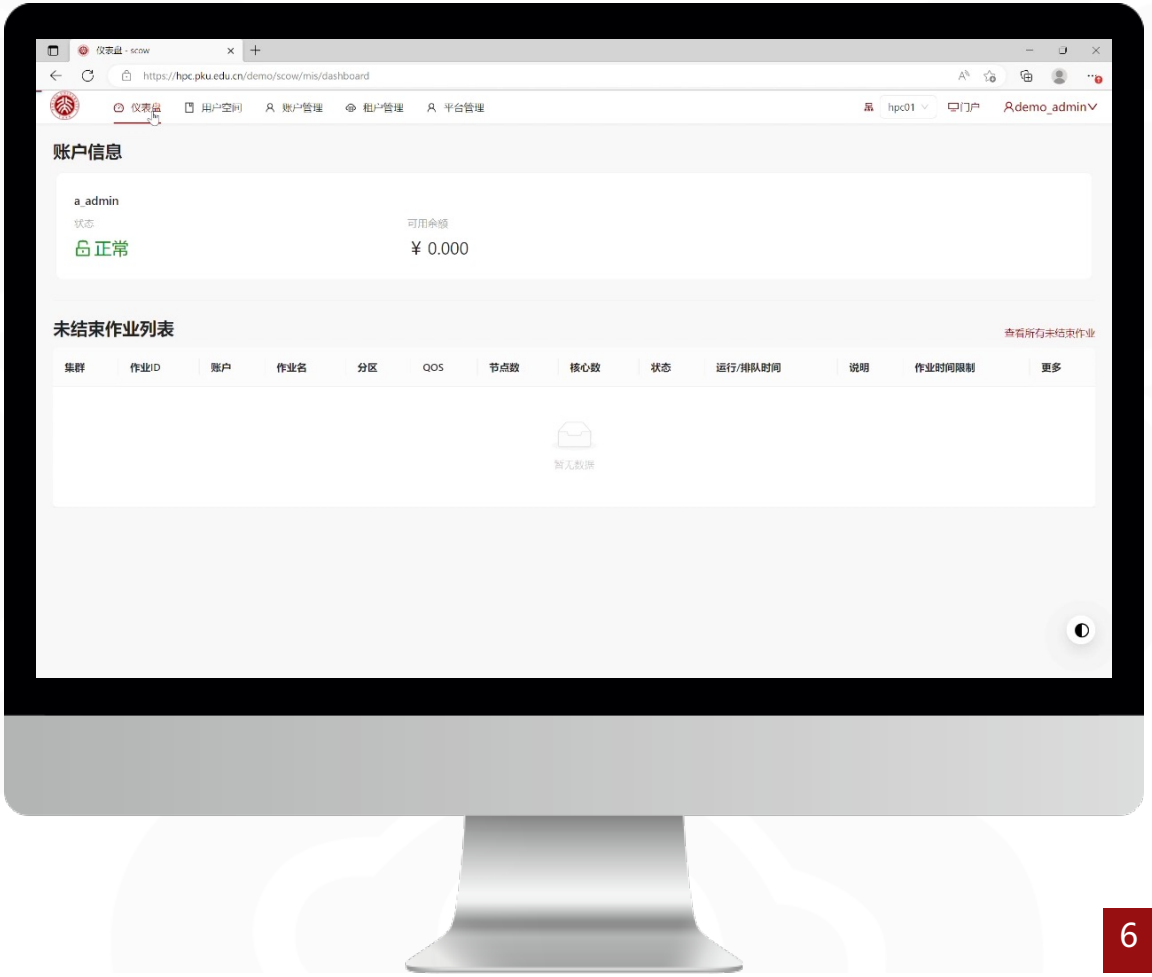
支持多集群管理

标准化模式、模型

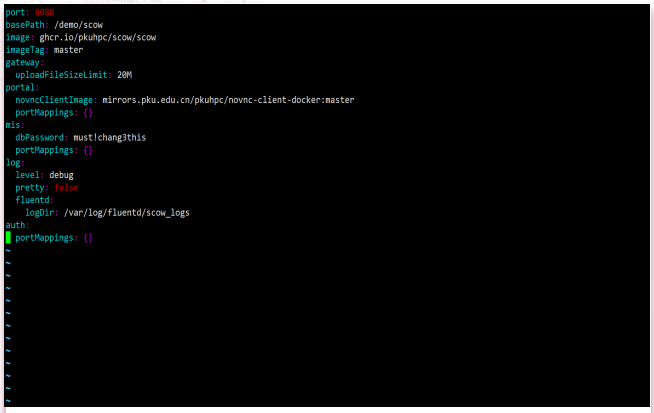
- 超算平台管理模式
- 用户账户模型
  （租户-账户-用户三级模型）
- 计费收费模型

管理系统多项功能

- 账户管理
- 租户管理
- 平台管理
- 财务管理

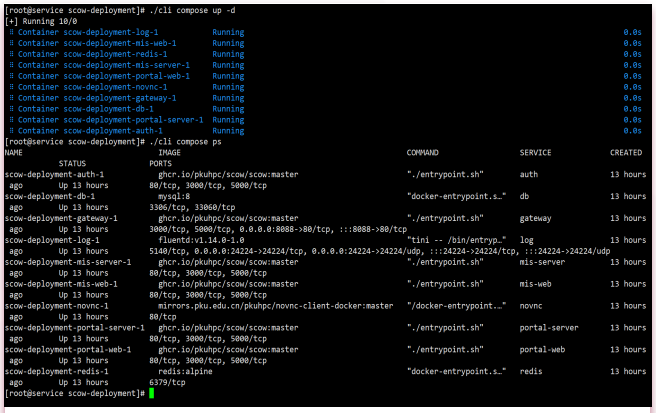


SCOW特色——一体化部署，开箱即用



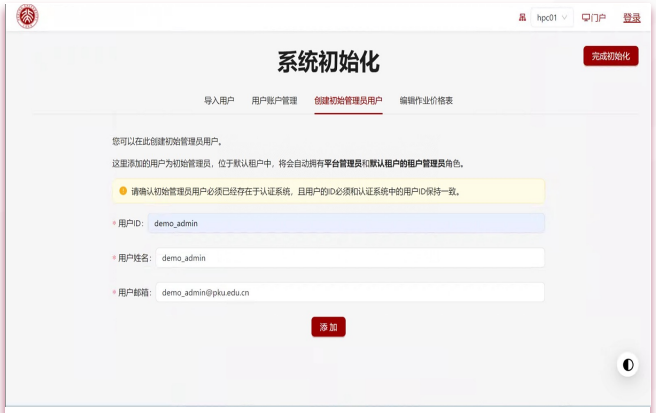
01 编辑配置文件

配置部署的模块、集群基础信息、认证系统、源作业、交互式应用、slurm数据库等信息。



02 编排与启动服务

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

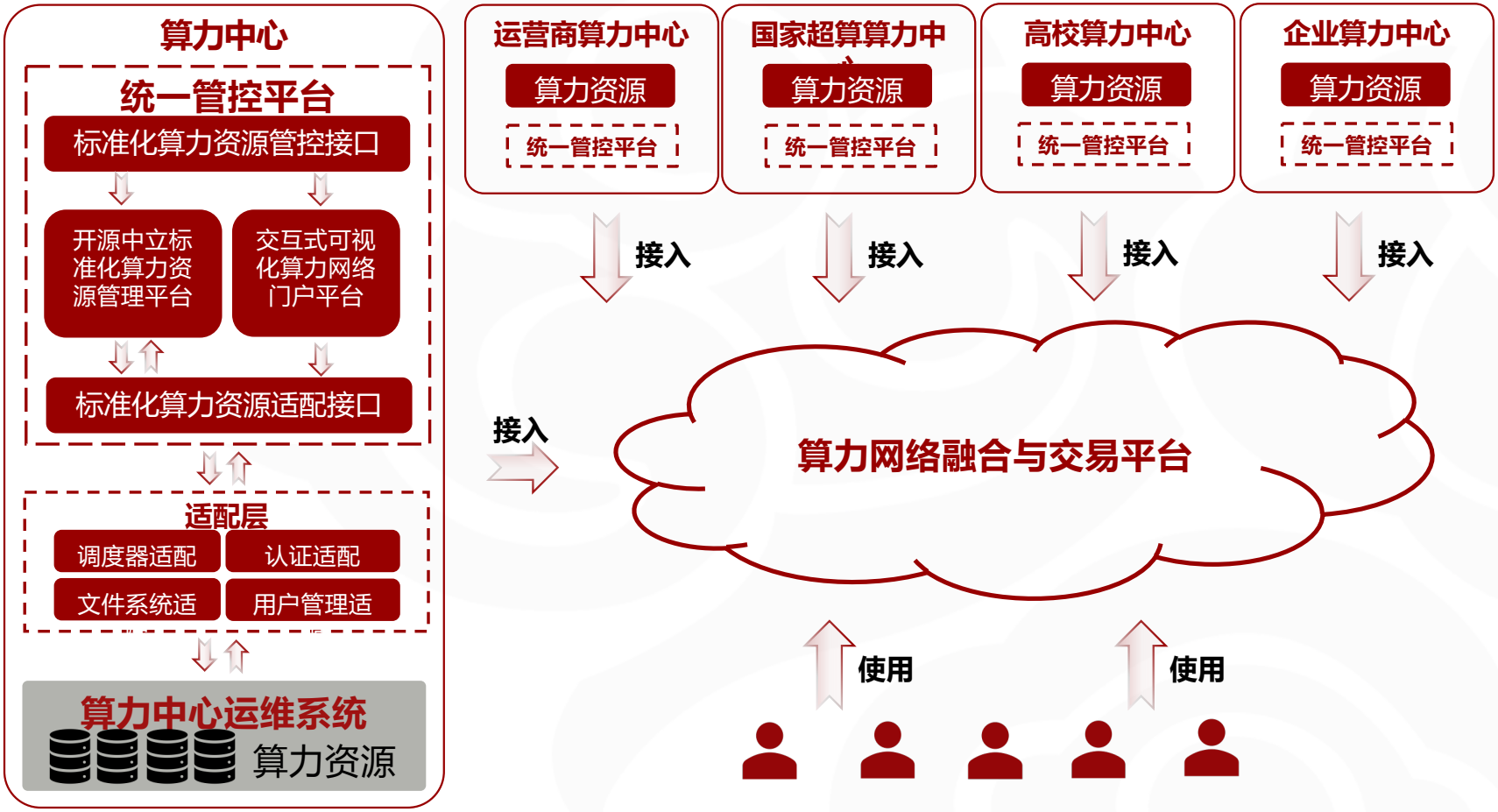


03 初始化管理系统

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。

SCOW特色——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。



SCOW——创新与突破

基于透明代理的全流程审计

面向CI/CD的自动化
安全检测机器人

实现基于文件元属性和网络自适应的跨集群智能文件传输

研发低延迟高画质的远程桌面连接技术

建立标准化算力中心管理模式

相关论文在2022年中国高性能计算学术年会中获得实践类最佳论文

SCOW——国内首个开源算力中心门户和管理平台

北京大学 PEKING UNIVERSITY

- **项目由北京大学高性能计算平台组织研发**

  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**

  项目源码地址：https://github.com/PKUHPC/SCOW

  项目文档地址：https://pkuhpc.github.io/SCOW/docs/common/deployment

  更多项目信息：https://icode.pku.edu.cn/info/1134/1137.htm

- **项目试用地址：**

  地址：https://hpc.pku.edu.cn/demo/scow

  管理员 用户名/密码 ：demo_admin / demo_admin

  普通用户 用户名/密码 ：demo_user / demo_user

北京大学高性能计算平台

北京大学 PEKING UNIVERSITY

5070

师生用户：
5070人

96

院系单位:
96个

19

两院院士:
19位

6

高校校长:
6位

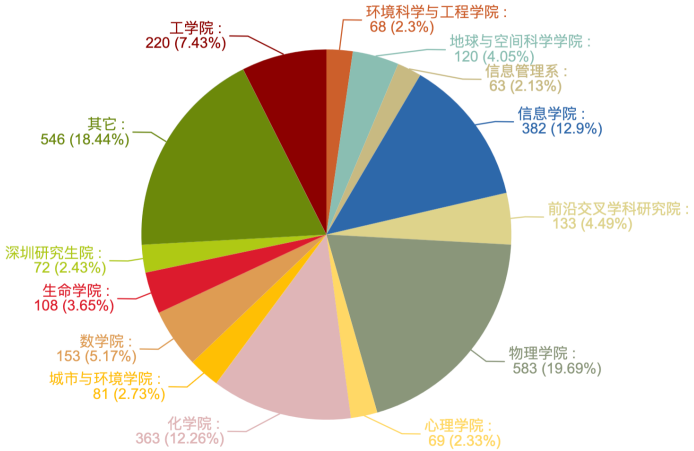
87.6%

未名一号年均使用率:
87.58%

81.3%

未名生科一号年均使用率:
81.3%

平台服务用户



用户来源分布

平台资源使用状况

未名一号（上）和未名生科一号（下）资源使用情况

北京大学 PEKING UNIVERSITY

支撑高水平论文

1400+篇获得用户致谢的论文

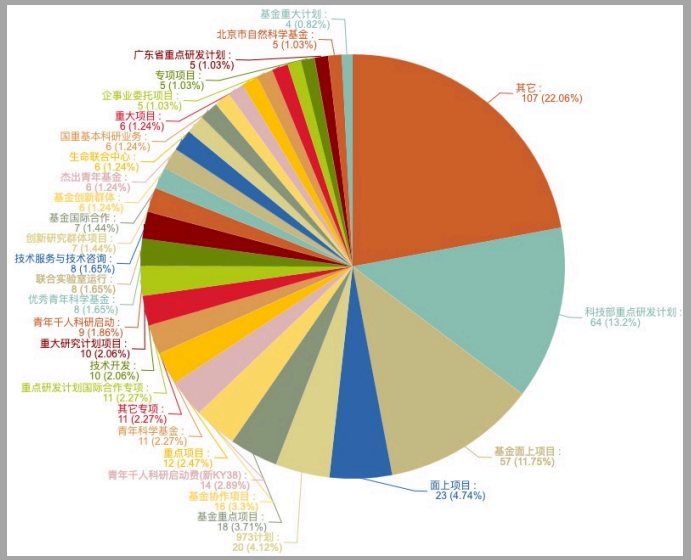
nature 15篇　Cell A Cell Press journal 2篇　Science AAAS 3篇

JACS 30篇　PRL 18篇　PNAS 4篇

nature International weekly journal of science Nature系列文章90+篇

支持科研项目

支持项目超545个，总金额达31.36亿元



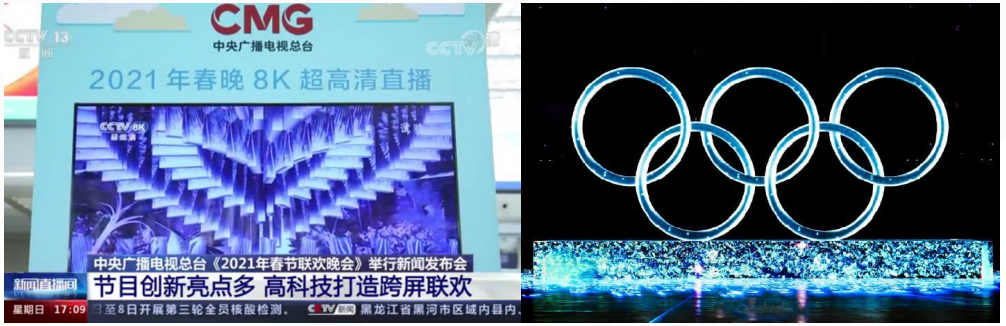
支持科研项目类型分布

申报论文发表情况

年份	论文数	Nature	Science	Cell	Nature子刊	PRL	PNAS	JACS
2017	**3**	0	0	0	0	0	0	0
2018	**99**	3	0	0	6	1	1	1
2019	**197**	3	1	1	9	3	0	5
2020	**254**	3	1	1	18	2	0	5
2021	**342**	3	0	0	19	6	0	7
2022	**390**	2	1	0	22	6	2	7
2023	**119**	1	0	0	1	0	1	5
合计	**1482**	**15**	**3**	**2**	**75**	**18**	**4**	**30**

2020年度国家技术发明**一等奖**

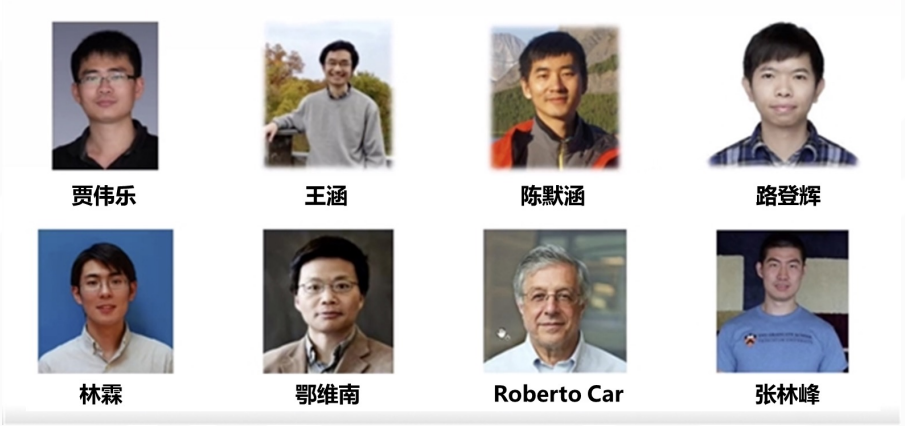
超高清视频多态基元编解码关键技术



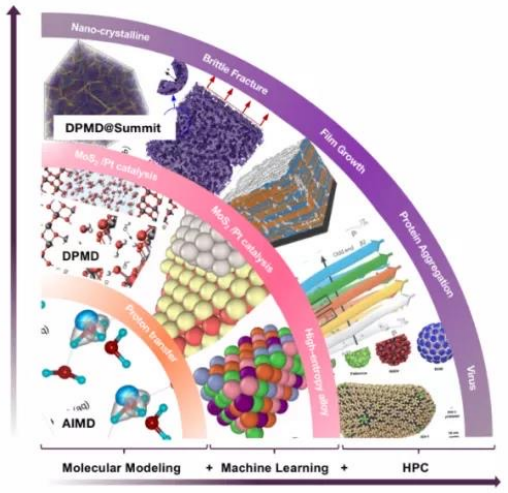
北京大学 PEKING UNIVERSITY

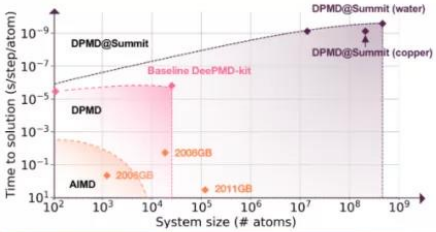
2020年戈登贝尔奖

基于深度学习的分子动力学模拟



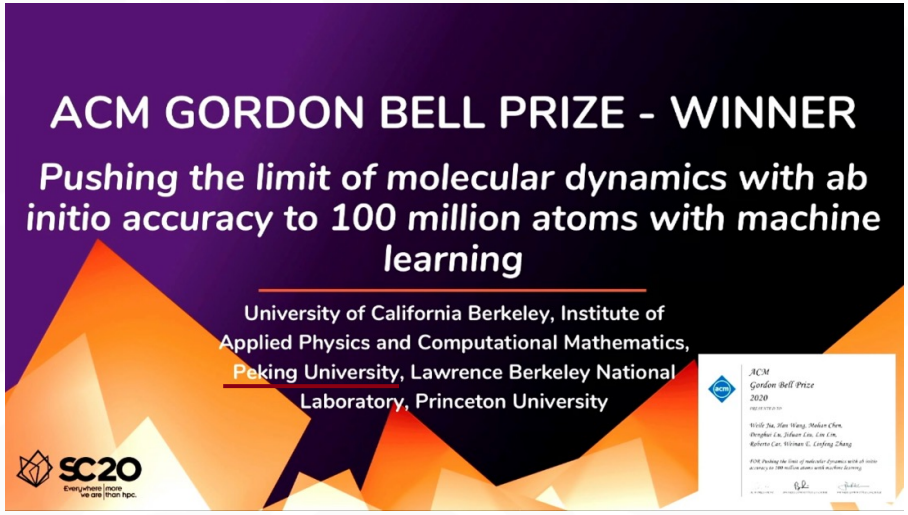
分子建模+人工智能+高性能计算





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



SCOW——人才培养

北京大学 PEKING UNIVERSITY

SCOW
SUPER COMPUTING ON WEB

+

HPC Game

第零届 北京大学高性能计算综合能力竞赛

SCOW——推广北大模式

北京大学
PEKING UNIVERSITY

已部署

北京大学、中国科学院国家天文台（测试）、中南大学、内蒙古香侬云科技发展有限公司、广州大学、湘潭大学、上海交通大学、国家蛋白质科学中心（上海）、中国科学技术大学、安徽汴水之畔超级计算中心、飞腾信息技术有限公司、西湖大学

部署推进中

山西省超级计算中心

达成部署意向：

中国科学院高能物理研究所、湖南师范大学、湖南中科曙光信息有限公司、华为人工智能创新中心、马栏山视频文创产业园

SCOW——未来

北京大学 PEKING UNIVERSITY

SCOW
SUPER COMPUTING ON WEB

服务教学科研

助力“双一流”建设

推广北大模式

在全球范围内推广

支撑国家战略

东数西算-> “双碳”战略

算力中心面临的困难

北京大学 PEKING UNIVERSITY

- **目前算力中心作业调度系统国外的软件占据90%以上的市场**
  - 在传统**高性能计算**中应用最广泛作业调度系统为**SLURM、LSF**
  - 在**智能计算服务领域**最广泛的容器式集群调度系统为**Kubernetes**

亟需一款国产高性能计算和智能计算调度系统

鸣鹤（Crane）智算调度规划

北京大学 PEKING UNIVERSITY

支持

应用领域

- 工业设计制造
- 航空航天
- 生物医药健康
- 智慧城市
- 新能源材料
- 大气海洋环境
- 智能驾驶

智算调度系统

跨平台调度

任务跨平台计算
数据跨平台传输
软件跨平台协同

软硬件资源管理

容器编排
软硬件资源编排
多中心网络拓扑感知

异构资源融合

超算资源
智算资源

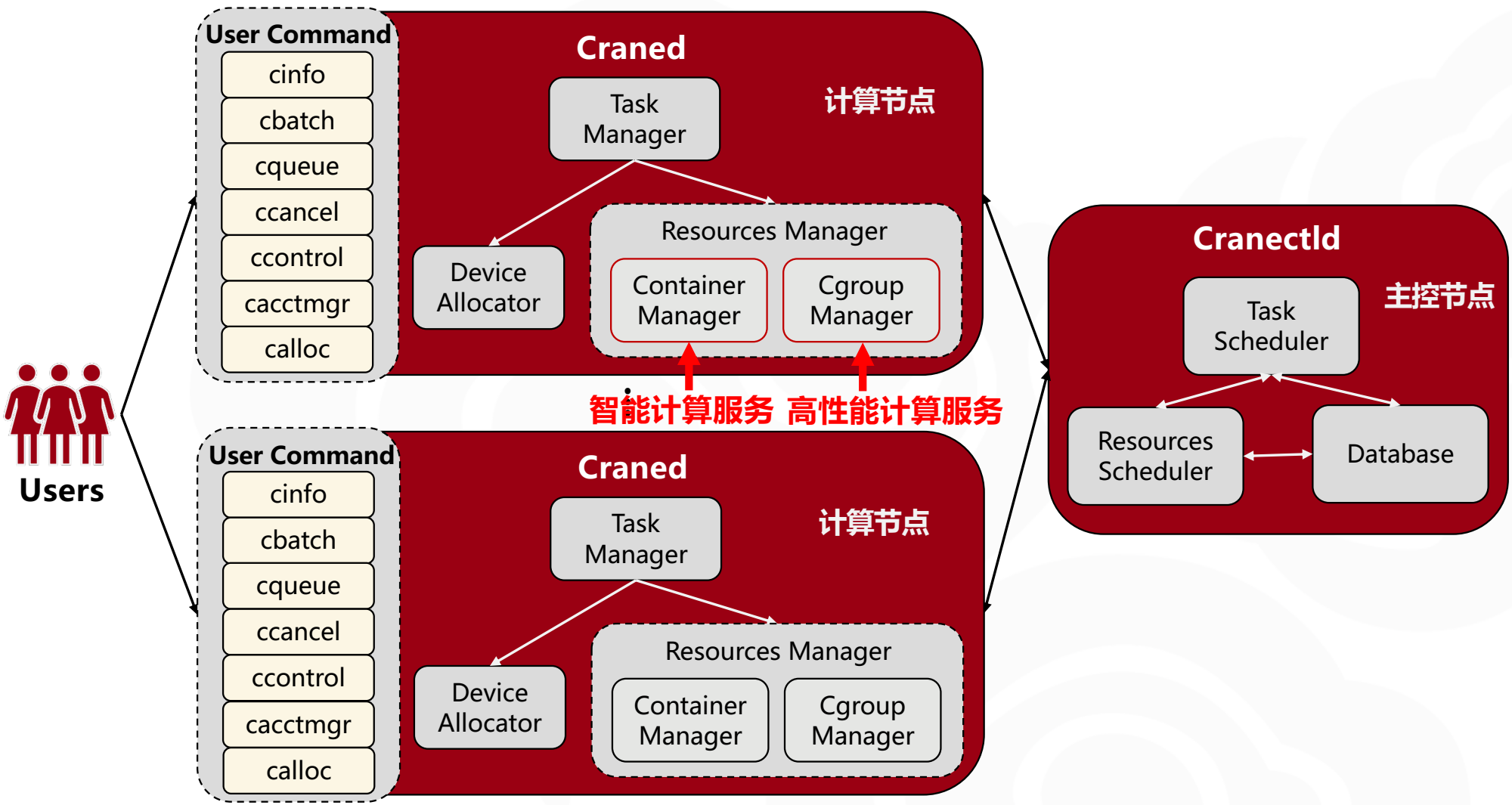
高效算法研究

多因素优先级回填算法
数据拓扑感知传输算法
软件跨平台匹配算法

硬件厂商

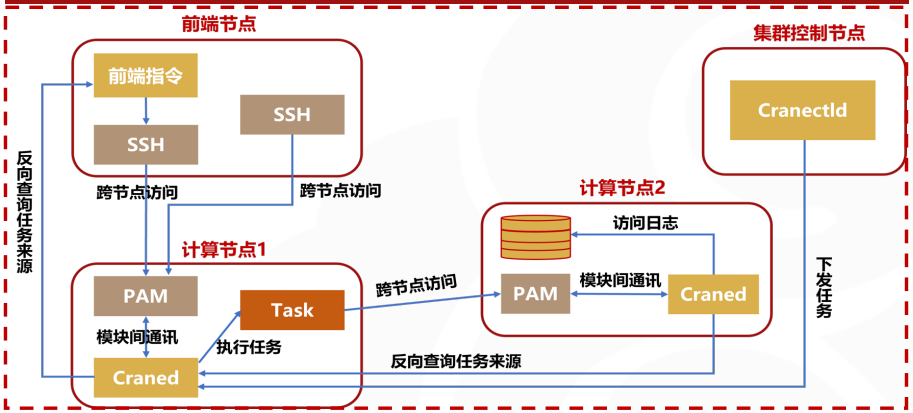
AMD
NVIDIA
intel
中科曙光 Sugon
inspur 浪潮
HUAWEI
Lenovo
DELL

Crane系统架构

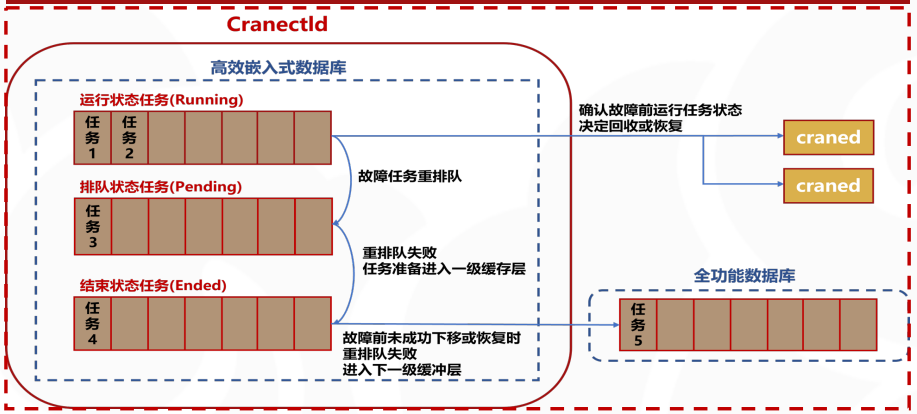

User Command
cinfo
cbatch
cqueue
ccancel
ccontrol
cacctmgr
calloc
Craned
计算节点
Task Manager
Device Allocator
Resources Manager
Container Manager
Cgroup Manager
智能计算服务
高性能计算服务
Users
User Command
cinfo
cbatch
cqueue
ccancel
ccontrol
cacctmgr
calloc
Craned
计算节点
Task Manager
Device Allocator
Resources Manager
Container Manager
Cgroup Manager
CranectId
主控节点
Task Scheduler
Resources Scheduler
Database

Crane关键技术研发

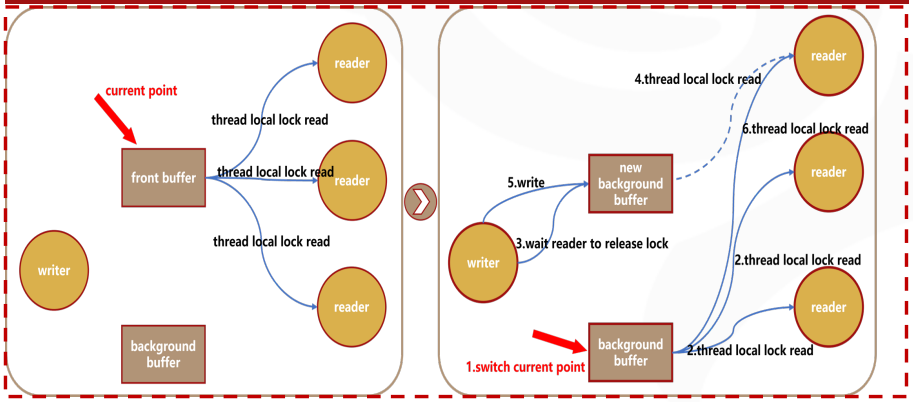
跨节点计算资源逃逸保护机制



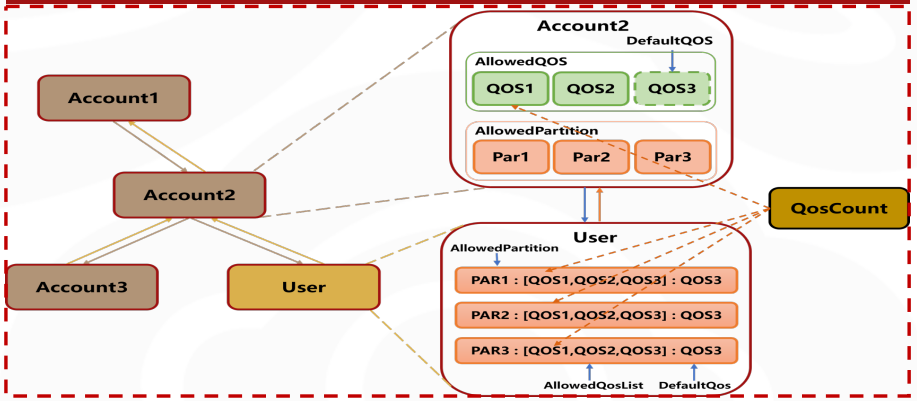
多级缓存故障恢复机制



大规模数据并发访问一致性机制



权限管理和安全通信机制



Crane常用用户指令

- cinfo: 查看节点与分区状态
- calloc：提交交互式作业
- cbatch：提交批处理作业
- cqueue：查看队列中的作业信息
- ccancel：取消队列中正在运行或者排队的作业
- ccontrol：查看/修改集群中的分区、节点、作业状态
- cacctmgr：管理账户/用户/QOS信息：添加账户/用户/QOS、删除账户/用户/QOS、查找账户/用户/QOS

```
#!/bin/bash
#CBATCH -p CPU
#CBATCH -N 2
#CBATCH -c 2
#CBATCH --mem 3400M
#CBATCH --ntasks-per-node 1
#CBATCH -J "relion_map3d"
#CBATCH --time 50:00:00
#CBATCH -o crane_relion_map3d_%j.out


echo "$CRANE_JOB_NODELIST" | tr ";" "\n" > crane.hosts

module load mpich/4.0 relion/3.0_beta_mpich_4.0
mkdir relion/relion_map3d

starfile="ribosome_500.star"
reffile="ribosome_10000_reference.mrc"
output="relion/relion_map3d/run"

mpirun -n 7 --machinefile crane.hosts relion_refine_mpi --o $output --i $starfile  --particle_diameter 340 \
 --angpix 2.82 --ref $reffile --offset_range 6 --offset_step 1 --ini_high 100 --iter 25 --tau2_fudge 4 --K 4 \
 --oversampling 1 --healpix_order 2 --sym C1 --j 1  --random_seed 1 --firstiter_cc --flatten_solvent \
 --zero_mask --norm --scale --ctf
```

批处理脚本样例

Crane ——国内首个开源智算调度系统

北京大学 PEKING UNIVERSITY

- **项目由北京大学高性能计算平台组织研发**

  已完成高性能计算功能，具有自主知识产权；

- **项目源码和文档地址：**

  项目源码地址：https://github.com/PKUHPC/Crane

  https://github.com/PKUHPC/Crane-FrontEnd

  项目文档地址：https://pkuhpc.github.io/Crane-document/

- **项目试用地址：**

  地址：https://hpc.pku.edu.cn/demo/crane

  用户名/密码 ：demo_admin / @icode2023

北京大学
PEKING UNIVERSITY

谢谢

THANK YOU

HPC/AI六大发展趋势

北京大学 PEKING UNIVERSITY

算力大众化

HPC/AI for Everyone.

算力网络

削峰填谷，实现算力资源利用效率最大化

“双碳”目标

碳达峰和碳中和，绿色发展

算力融合

超算中心，智算中心成为计算中心（算力中心）

RoCE

真的能替代IB吗？

浪费是必要的

chatGPT

算力大众化

HPC/AI for Everyone

算力网络

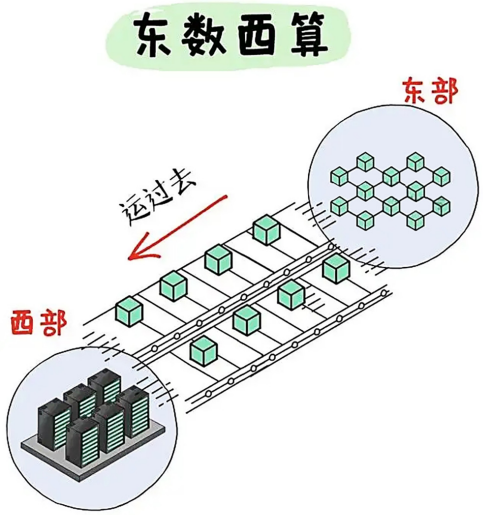
北京大学 PEKING UNIVERSITY

政策支持

2020.04 国家发展和改革委员会首次对**“新基建”**的具体含义进行了阐述，算力基础设施概念首次在国家政策层面提出。

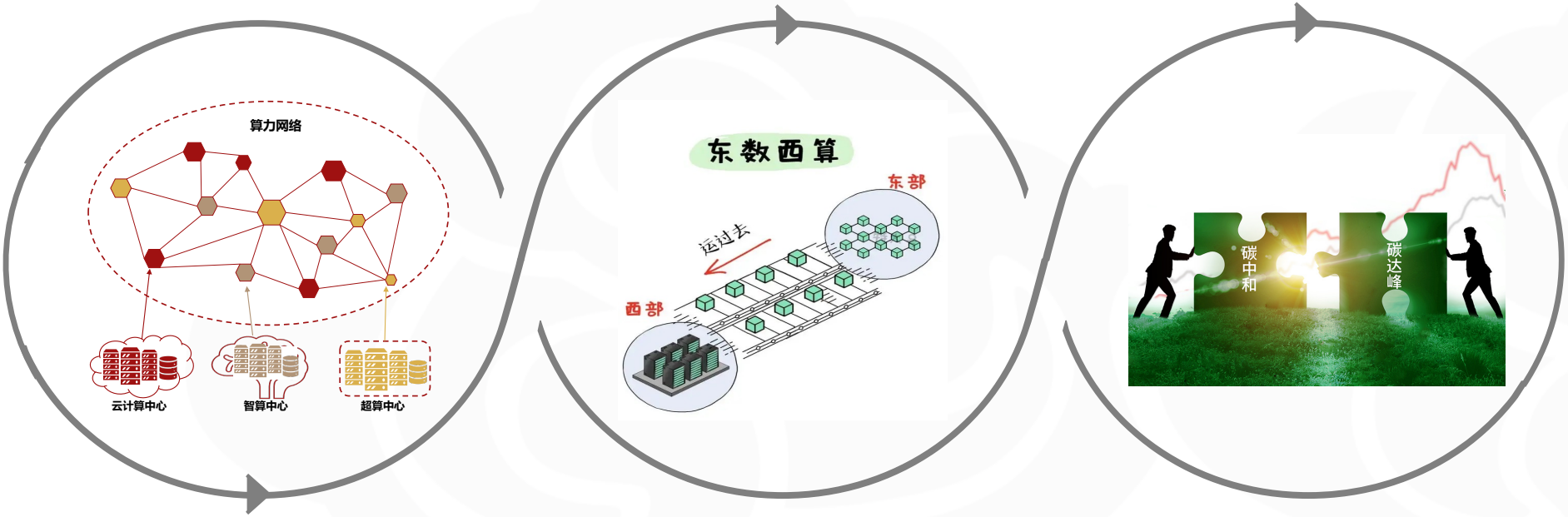
2021.05 发改委发布《全国一体化大数据中心协同创新体系**算力枢纽**实施方案》，**算力网络**概念第一次在国家政策层面的出现。

2022.03 提请十三届全国人大五次会议审查的计划报告提出，实施**“东数西算”**工程。



通过构建数据中心、云计算、大数据一体化的新型算力网络体系，将东部算力需求有序引导到西部，优化数据中心建设布局，促进东西部协同联动，让西部的算力资源更充分地支撑东部数据的运算，更好为数字化发展赋能。

碳中和



2023年4月17日
国家超算互联网工作启动，服务国家战略，构建算力网络

“东数西算”工程在全国布局了8个算力枢纽，引导大型、超大型数据中心向枢纽内集聚，形成数据中心集群

围绕实现碳达峰、碳中和目标采取有力措施，持续提升能源利用效率，加快能源消费方式转变

构建算力网络 服务国家战略

中华人民共和国中央人民政府
www.gov.cn

简 | 繁 | EN | 注册 | 登录

国务院 总理 新闻 政策 互动 服务 数据 国情 国家政务服务平台

首页 > 新闻 > 滚动

科技部启动"人工智能驱动的科学研究"专项部署工作

2023-03-27 20:09 来源： 新华社 【字体：大 中 小】 打印

新华社北京3月27日电（记者 宋晨）为贯彻落实国家《新一代人工智能发展规划》，科技部会同自然科学基金委近期启动“人工智能驱动的科学研究”（AI for Science）专项部署工作，紧密结合数学、物理、化学、天文等基础学科关键问题，围绕药物研发、基因研究、生物育种、新材料研发等重点领域科研需求展开，布局“人工智能驱动的科学研究”前沿科技研发体系。

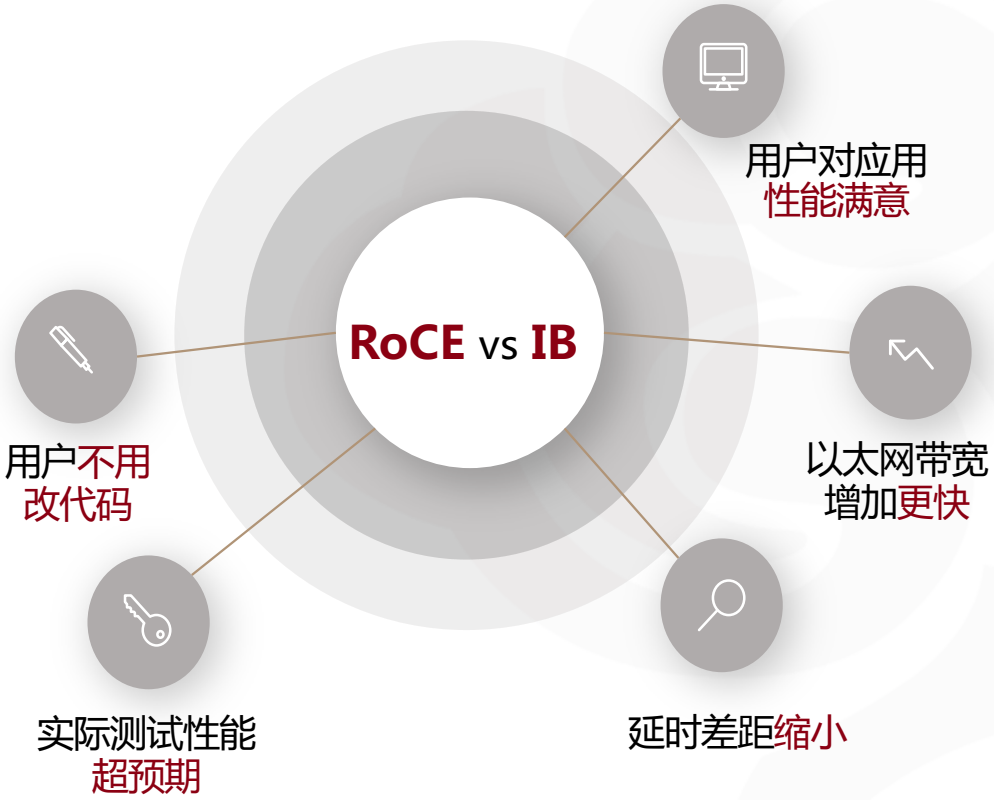
科技部将推进面向重大科学问题的人工智能模型和算法创新，发展一批针对典型科研领域的“人工智能驱动的科学研究”专用平台，加快推动国家新一代人工智能公共算力开放创新平台建设，支持高性能计算中心与智算中心异构融合发展，鼓励绿色能源和低碳化，推进软硬件计算技术升级，鼓励各类科研主体按照分类分级原则开放科学数据。

HPC/AI融合

https://slurm.schedmd.com/SC22/Accelerating-with-Slurm.pdf

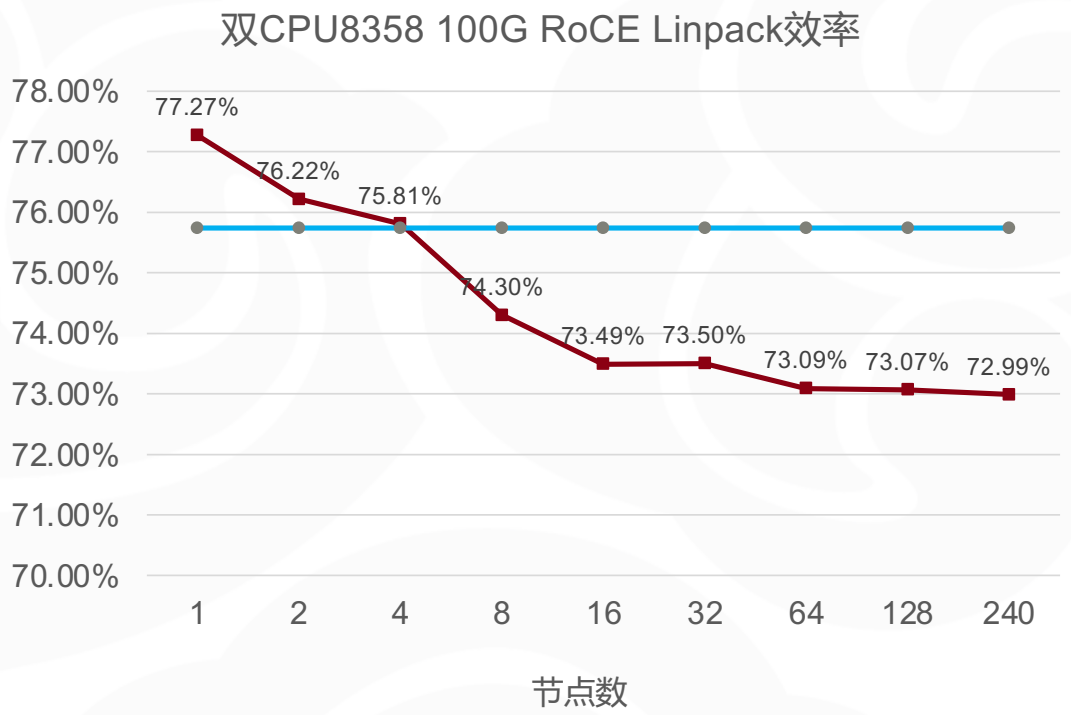
https://openai.com/research/scaling-kubernetes-to-7500-nodes

RoCE网络可以大规模用了



240节点HPL测试，线性度超预期

基于超融合以太RoCE



蓝线为集群中效率最低的单节点Linpack值

创建创新的土壤——”浪费”是必要的

ChatGPT的出现和火爆使我们再次看到美国科技企业作为国家技术创新主体所具有的强大创新能力。关注和坚持投入给人类社会带来深远影响的科技创新，从而抢占新兴产业市场。

2023年2月27日，创新型科技文化是创新型国家的灵魂因素。基础研究和科技创新是不能追求效率的，而我们对基础研究采用的评价基本还是以效率为重的工程性评价方法。

启示

——赵沁平院士，ChatGPT研讨会

必要的浪费

资源充足	国际交流	自由探索	方向指引