塔图算力解决方案

北京大学长沙计算与数字经济研究院

长沙尖山塔图科技有限公司

目录 CONTENTS

PART1 概述

PART2 产品

PART 1

概述

关于 北京大学长沙计算与数字经济研究院

北京大学长沙计算与数字经济研究院是北京大学、长沙市人民政府和湖南湘江新区（长沙高新区）管理委员会共建的长沙市属事业单位。研究院以建设高水平新型研发机构为目标，围绕先进计算、大数据、数字经济与数字化转型等领域的核心科学技术问题和重大应用难题，广泛集聚全球顶尖人才团队，开展基础研究、应用研究和政策研究，培养创新人才，孵化科技企业，构建具有国际影响力、引领国内先进计算和数字经济发展的人才聚集平台、前沿研究平台和成果转化平台。

关于 长沙尖山塔图科技有限公司

长沙尖山塔图科技有限公司，是由北京大学长沙计算与数字经济研究院算力网络研究中心孵化设立的实体企业。公司团队具有丰富的超算、智算资源运维、运营、管理经验。公司面向各行业提供超智算一体化解决方案。核心产品涵盖算力平台系统SCOW、算力网络融合平台XSCOW、算力调度系统鹤思、远程桌面ShadowDesk、大模型智能体平台小蒜。不仅如此，依托深厚的平台管理经验，公司还提供超智算平台运维服务。

尖山塔图始终坚守创新、卓越、服务的理念，以专业的技术能力和精准的市场洞察，致力于为客户打造高效、稳定、智能的算力服务体系，助力行业蓬勃发展，共铸数字经济的辉煌未来。

关于 团队

核心研发团队源自北京大学，运维实施团队拥有多个国家超算中心、万节点集群的完整建设与运维经验，技术底蕴深厚。

团队29人中，22人拥有硕士及以上学历，7位毕业于北京大学，多数毕业于双一流高校。

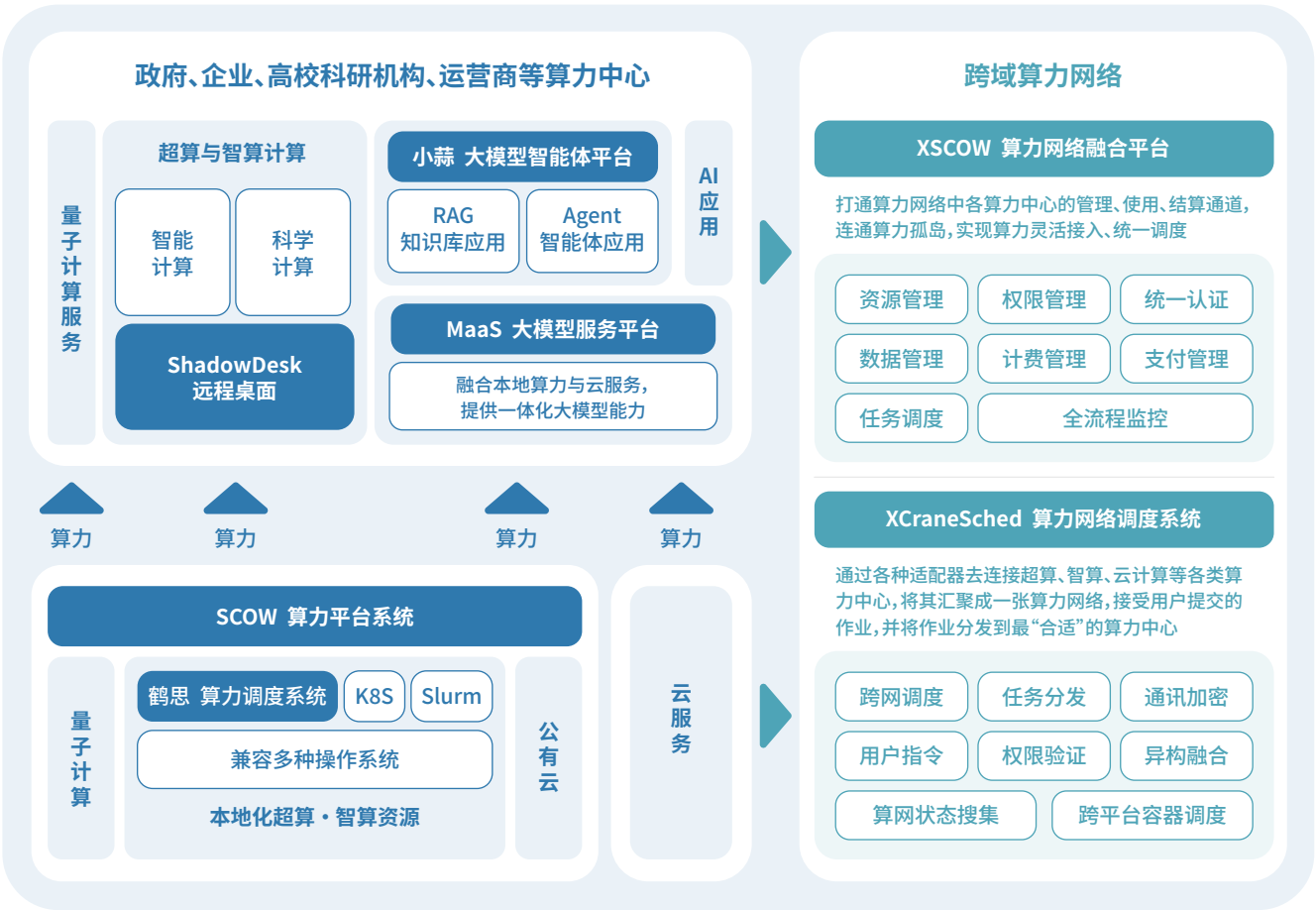
产品、开发、测试成员深度耕耘B端及C端市场，云计算与数据计算经验丰富，对复杂需求与高标准交付有深刻理解。

PART 2

产品

产品概览

尖山塔图立足北京大学校级算力平台近十年的技术积累与运营经验，精心打造了算力门户系统SCOW、算力网络融合平台XSCOW、算力调度系统鹤思、远程桌面 ShadowDesk以及大模型智能体平台小蒜，为政府、高校、企业、运营商提供算力平台运维一站式服务。

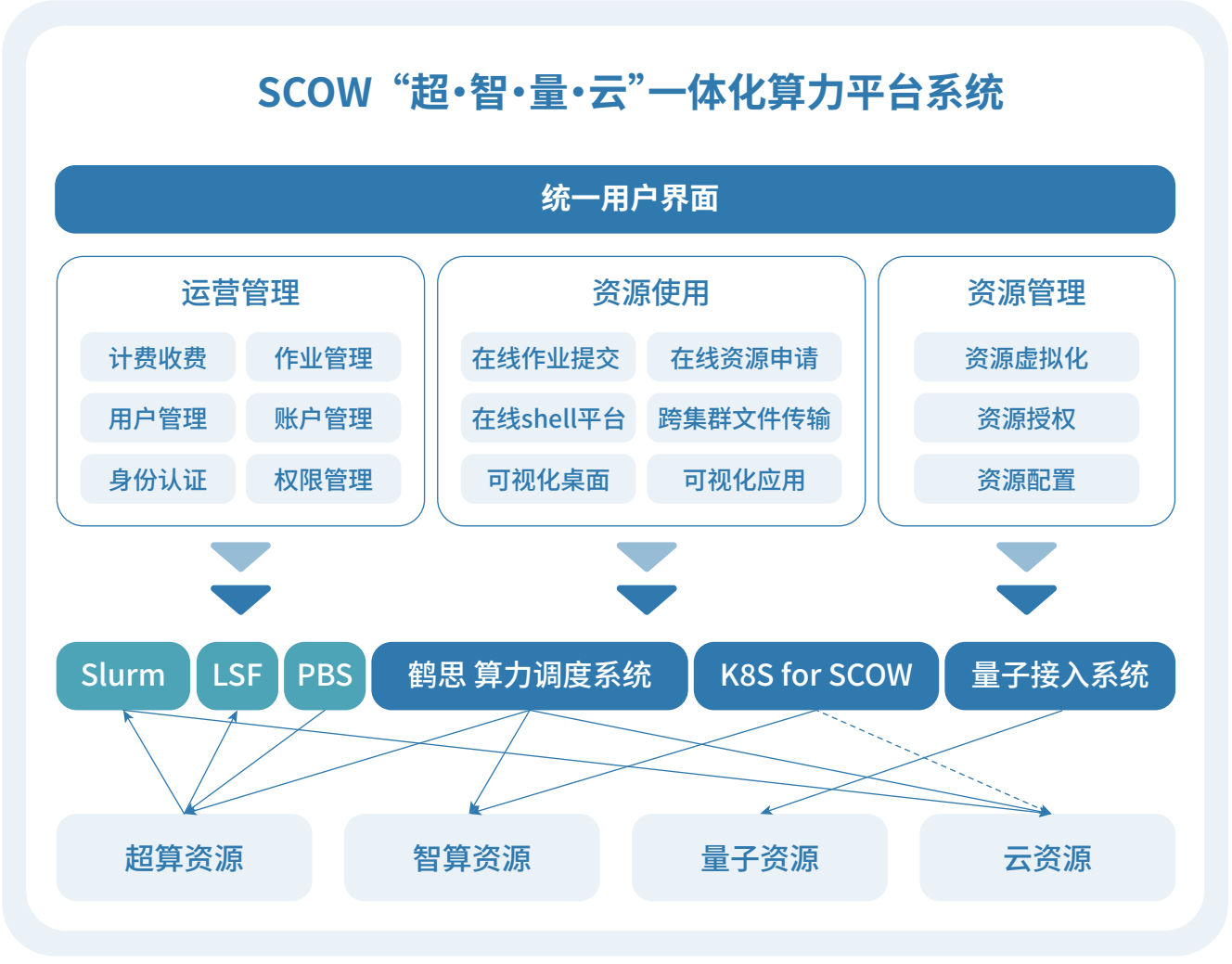


SCOW 算力平台系统

产品概述

为了解决算力中心在建设和运营中广泛存在的运营管理难、用户使用难、资源融合难等问题,降低集群平台软件的建设、管理、使用门槛,我们推出了算力平台系统SCOW(super Computing On Web)。SCOW可同时纳管基于不同硬件厂商、不同软件栈建设的包括HPC算力和AI算力在内的各类异构算力资源,向用户和管理员提供便捷完整的算力资源管理和使用功能;也可以接入算力网络,实现算力融合。支持基于HPC环境纳管虚拟机,让用户能够在无需管理员权限的情况下,在超算平台上灵活运行Windows等多种操作系统,并充分利用GPU等硬件加速资源。SCOW为算力中心提供了更开放、更便捷、更灵活、更高效的运营管理解决方案。

平台架构



功能亮点

图形化界面，使用方便

在算力资源使用方面,SCOW提供基于web页面的多个功能,降低了用户使用门槛，让linux小白用户也能顺利使用算力资源。

01

功能丰富,管理便捷

在算力中心管理运营方面,SCOW提供覆盖算力资源全生命周期的全流程管理能力，能够帮助算力中心快速建立管理和运营制度，快速投入使用。

02

标准化平台,支持算力融合

支持接入slurm、CraneSched、K8s等多种资源调度器,可纳管各类算力资源。同时面向算力网络提供标准化管控接口,支撑算力融合。

03

开放中立,支持开源

SCOW独立于各家厂商，帮助算力中心打破供应商锁定，能够接入各家供应商的各种软硬件设备，实现真正的开放中立。
发起并维护基于木兰宽松协议开源的社区项目OPENSCOW,助力算力生态发展。

开源仓库地址:https://github.com/PKUHPC/OPENSCOW

04

超·智·量·云 四算融合

在同一平台接入和管理超算、智算、量子计算和云计算资源，支持各种计算场景，在国内率先实现实现超智量云四算融合。

05

支持基于HPC环境的拟虚化功能

可在超算环境下运行虚拟机，突破传统超算系统的限制，在超算平台上高效运行Windows等多种操作系统，支持x86、ARM等架构，打造"超算即云"的使用体验。

06

快速部署，开箱即用

能够快速在新建集群部署上线或接入现有集群,部署几乎无侵入,可与其他管理平台共存。提供强大的运维工具,降低运维门槛。

07

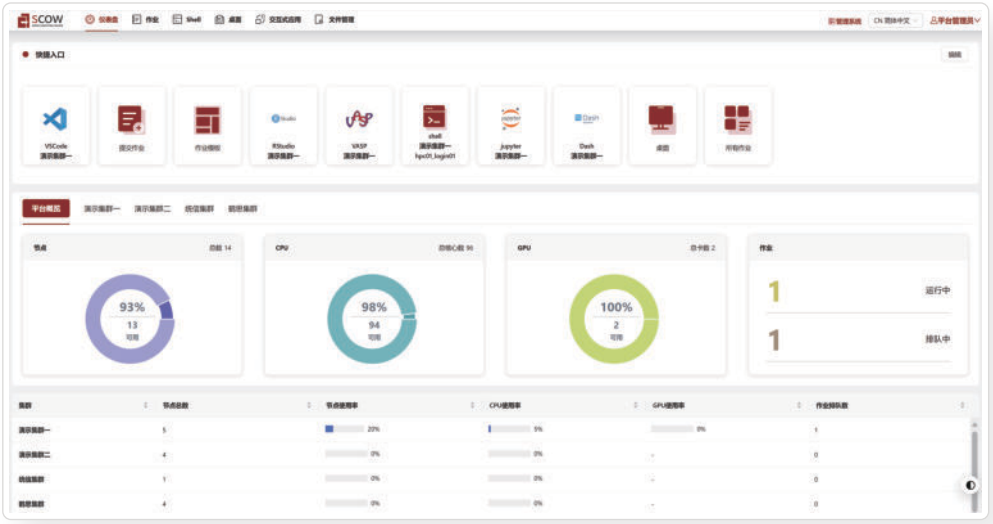
应用案例

部署案例 北京大学

基于SCOW运营校级公共平台，接入4个不同建设时间、不同供应商，分布在不同位置的算力集群，平台已有超过3000名校内师生用户，600多个账号。平均每天有超过70位用户登录SCOW，平台单个月支持用户执行的操作次数超过50000次，超过三分之一的用户通过SCOW使用集群资源。

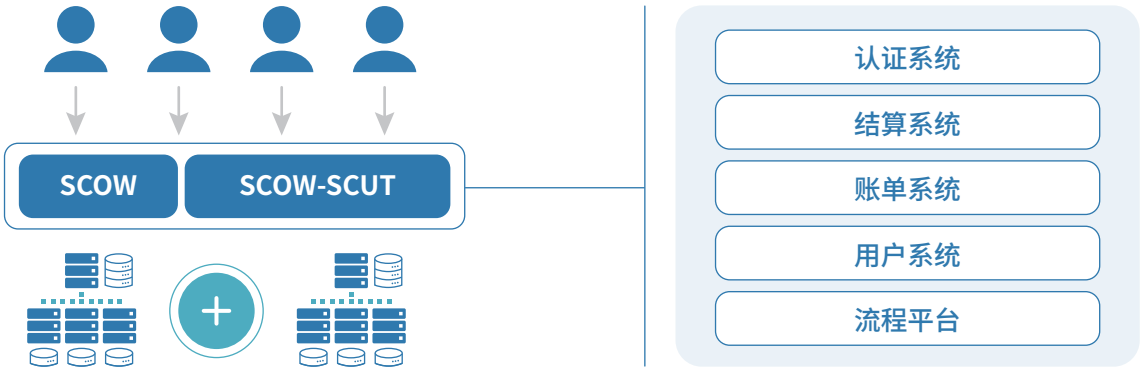
校内师生用户
3000+

账户
600+



部署案例 华南理工大学

2024年投入使用的新集群，集群由两家供应商分别建设，采用基于SCOW的算力平台运营方案，解决了多家供应商算力资源统一管理的问题。通过在SCOW上进行财务结算等功能的二次开发，满足学校平台特定功能需求。



推广部署 部署情况

已协助部署SCOW的算力中心

覆盖
23个省市

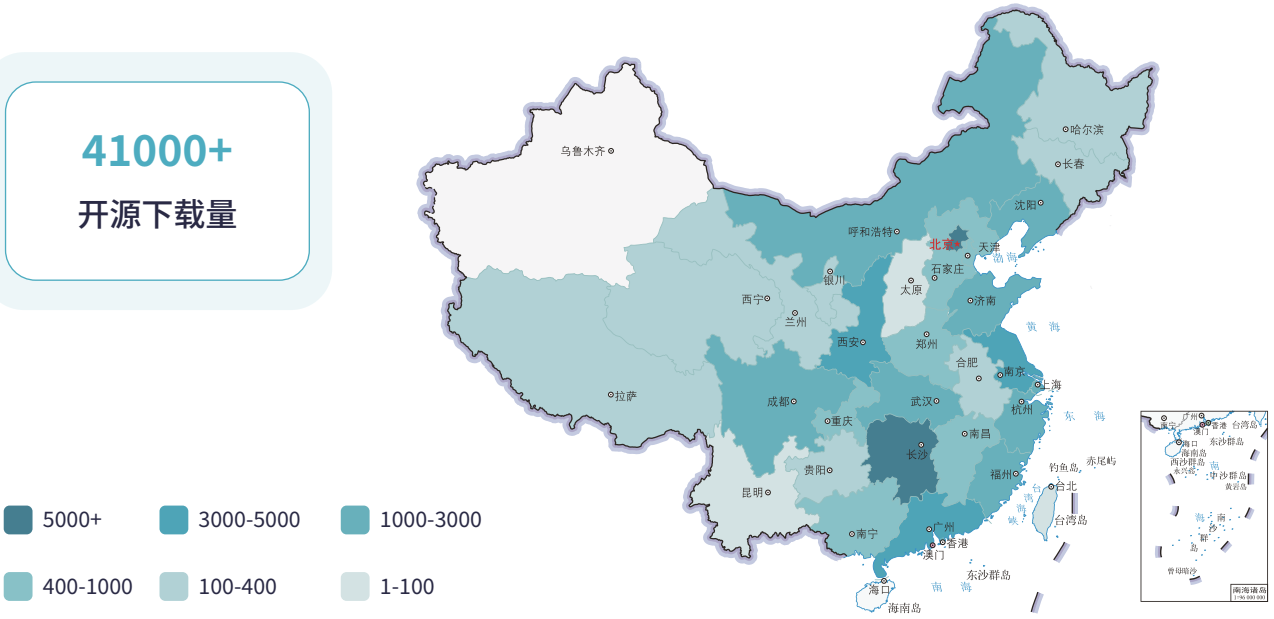
超过
70个算力中心

本图使用自然资源部监制地图制作(审图号:GS(2019)1825号)

推广部署 OPENSCOW社区版下载统计

自2023年7月以来,有国内30个省市以及美国、日本、新加坡、阿尔及利亚、印度尼西亚的用户下载了OPENSCOW社区版。

41000+
开源下载量



本图使用自然资源部监制地图制作(审图号:GS(2019)1825号)

XSCOW 算力网络融合平台

产品概述

XSCOW是基于SCOW平台，构建可实现产业化且兼容各类算力资源、对厂商中立的算力网络融合平台，以解决算力网络中使用、接入、交易三个关键问题。目标是建成能够覆盖超算资源、智算资源、云计算资源等各类算力资源的的算力网络统一大市场。

平台架构

功能亮点

快速接入

已经部署SCOW的超算中心，无需二次开发，可以快速接入到XSCOW平台。

01

统一调度和运营

平台连接供需侧，提供端到端一体化网络服务，用户在一个门户即可使用所有算力资源，无需在多个算力中心切换。

02

结算便利

用户只需一个账户支付所有算力中心的费用，平台按周期和算力中心结算。

03

应用案例

教育部高校智算融合共享平台

2023年11月13日，由教育部教育管理信息中心牵头成立高校智算融合共享平台，旨在推进整合各地高校超算中心，搭建供需平台，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。尖山塔图作为技术支撑方，进行平台整体搭建和软件研发。目前已有10所高校资源和16所高校用户接入平台 。

高校资源	高校用户
10所	16所

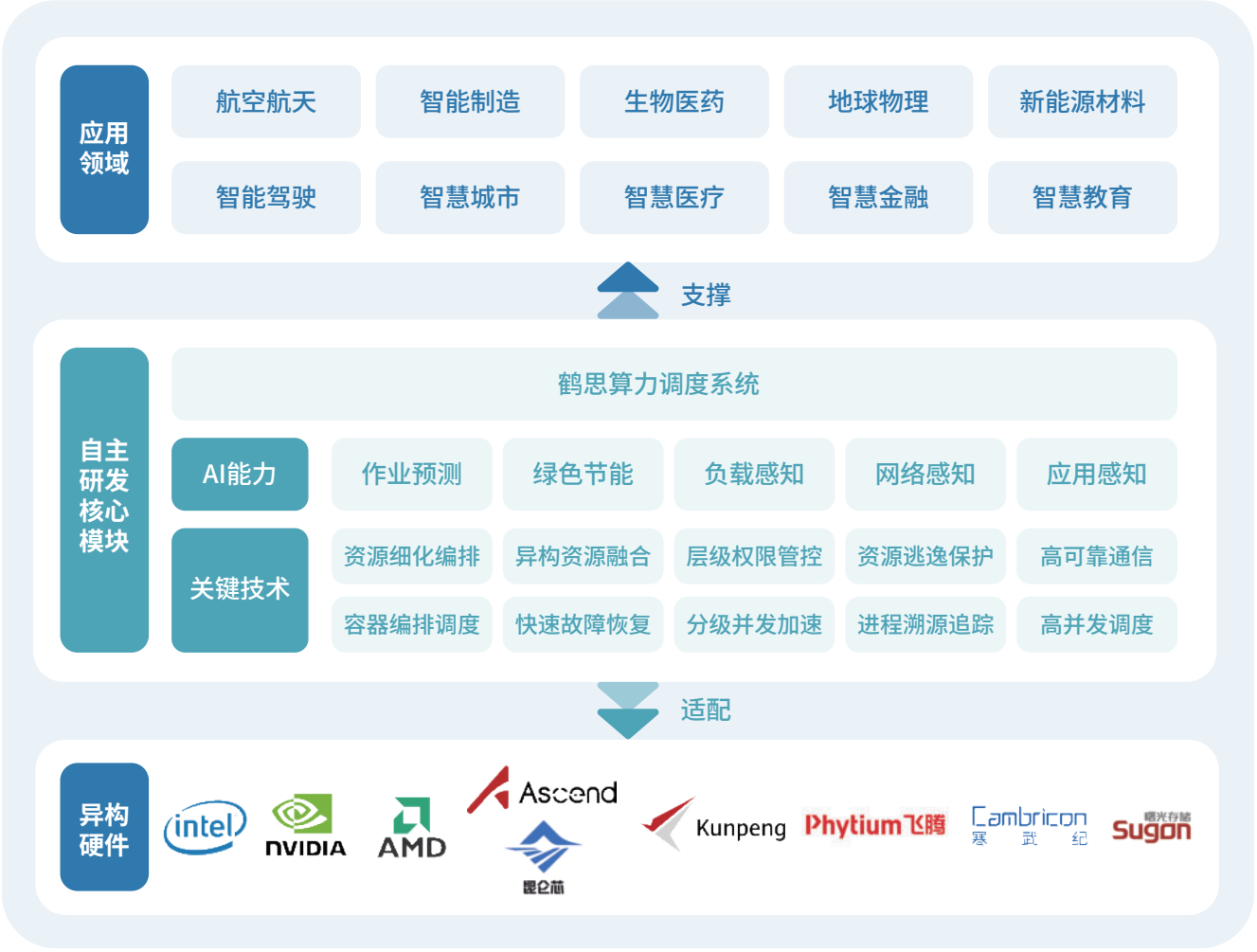
鹤思 算力调度系统

产品概述

鹤思(CraneSched)是北京大学计算中心与北京大学长沙计算与数字经济研究院联合研发的高效算力调度系统，支持高性能计算与智能计算，适配国产化操作系统及异构芯片服务器，并通过自研算法提升资源利用率、降低能耗与碳排放。系统兼容Slurm 和 LSF 指令，确保用户无缝迁移。同时，鹤思高效支持如DeepSeek、Qwen、Llama等大模型的训练与推理任务。

鹤思是国内首个支持超智算领域的信创开源算力调度系统，并入选　2024年工信部“典型应用案例”及“重点推荐应用案例”、“2024年教育信息技术应用创新优秀案例集”等多个奖项。

平台架构



功能亮点

性能指标国际领先

实测达Slurm等系统的5-20倍。支持超十万节点算力规模，小时级任务吞吐量超3800万，可并行处理任务数超200万，解决了大规模资源调度与高通量计算难题。

01

统一异构算力纳管

统一纳管华为鲲鹏、昇腾、飞腾、海光、寒武纪、昆仑芯、英伟达、AMD等国内外CPU及加速卡，支持OpenEuler、银河麒麟、Rocky、Ubuntu等国内外操作系统，攻克异构算力统一调度与国产硬件深度适配难题，降低异构集群管理成本，完善国产算力生态。

02

超算智算深度融合

突破资源壁垒，实现超算（气象、工业仿真、材料研发等，适配WRF、OpenFOAM等）与智算（大模型训练/微调/推理，支持PyTorch、DeepSeek、Qwen等）场景统一调度，消除资源孤岛，减少资源闲置浪费。

03

智能算法提效降耗

自研多项创新算法，全方位优化调度效率与能效：自研首个集成大语言模型的作业时间预测模型ORA算法（发表于CCF推荐B类会议ICS），预测准确率提升41%；自研TSMF算法（发表于CCF推荐中文B类期刊《计算机科学》，封面论文）显著改善用户体验与资源利用，算力90%高负载情况下，利用率可提升至97.3%，任务平均排队时间减少13.6分钟；自研节能算法实测低负载时集群能耗直降78.64%。

04

应用案例

案例一 北京大学

鹤思目前纳管北京大学未名教学二号和卓越中心集群两个校级集群，现有师生用户4500+，支撑科研任务40000+，支持大数据、生物医药设计、电池材料研究、医学大模型等科研场景。

01 北京大学未名教学二号集群

典型超算场景

用户场景：大数据 + 生物医药设计 + 电池材料研究

支撑师生实时在线教学科研

文科计算机小白用户快速上手

支撑教学课程200+课时

支撑师生科研任务27000+

系统稳定运行至今1年+

02 北京大学卓越中心集群

典型智算场景

用户场景：医学大模型 + 化学大模型 + 金融AI助手

算力集群软硬件全国产化典型案例

兼容全国产华为AI芯片及网络设备

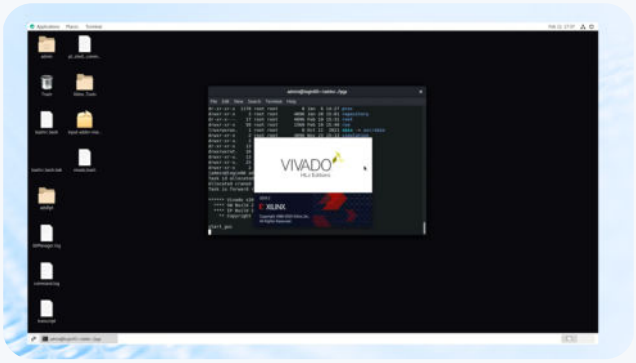
支撑大模型微调和推理任务

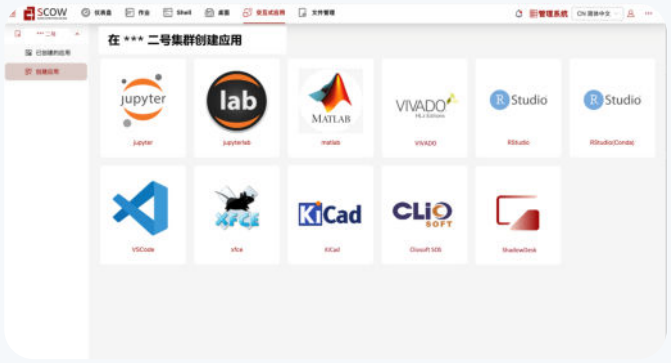
支撑师生科研任务12000+

系统稳定运行至今1年+

案例二 某芯片设计公司

自2024年起，某芯片设计公司（约200人规模）引入了“鹤思+SCOW”一体化解决方案管理其新建集群。该方案不仅成功解决了旧有商业软件因资源管控机制不合理而导致计算节点经常超载的问题，还通过深度适配主流EDA工具，为各类芯片设计任务提供了精准、高效的支持。





推广部署 部署情况

已协助众多高校科研院所和企业单位部署使用鹤思。

覆盖
全国8个省市

超过
10个算力中心

本图使用自然资源部监制地图制作（审图号：GS（2019）1825号）

ShadowDesk 远程桌面

产品概述

ShadowDesk是一款面向多种算力中心的高性能远程桌面控制软件，广泛适用于科研、工业、媒体、信创、办公、娱乐等领域。无论是EDA芯片设计、CAE仿真与工程设计，还是数据可视化、数字媒体创作、远程办公开发、云游戏等场景，ShadowDesk都能为用户提供安全、高效、流畅的远程桌面访问体验。ShadowDesk支持通过浏览器远程控制算力中心的Linux与Windows操作系统，实现仿真前后处理、芯片开发、数据建模与分析等多样化应用。产品兼容主流国产及国际处理器、操作系统和显卡，支持超算集群的独立部署或与云平台、超算平台无缝集成，灵活适配现有算力资源体系。

平台架构

功能亮点

高清晰低延迟

- 自研远程桌面传输协议，达到国际先进水平
- 支持4K分辨率、120FPS帧率桌面传输

01

硬件加速支持

- 控制端、被控端支持基于显卡的硬件加速，显著提升桌面端到端传输延迟
- 支持多会话共享显卡，最大化资源利用

02

功能丰富

- 音频传输、4:4:4彩模式、文件传输、剪贴板传输、手柄控制、数据端到端加密、容器化部署、虚拟桌面

03

多端支持

- 控制端支持Chorme、Firefox等主流浏览器
- 被控端支持Windows、Linux等操作系统

04

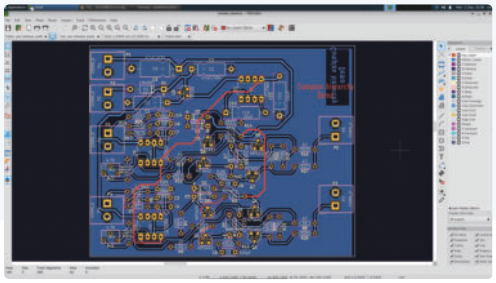
信创支持

- 支持华为鲲鹏、飞腾、海光等国产处理器
- 支持摩尔线程等国产显卡
- 支持OpenEuler等国产操作系统

05

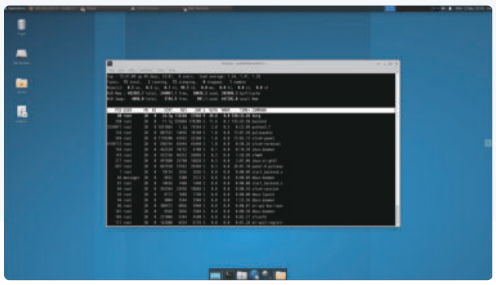
应用场景

电子设计自动化



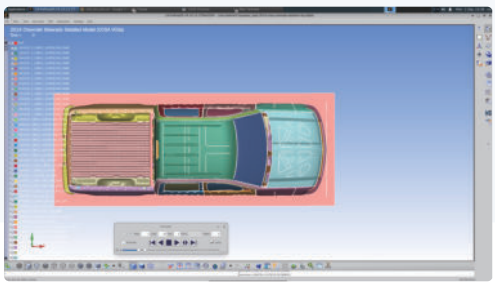
01

远程工作



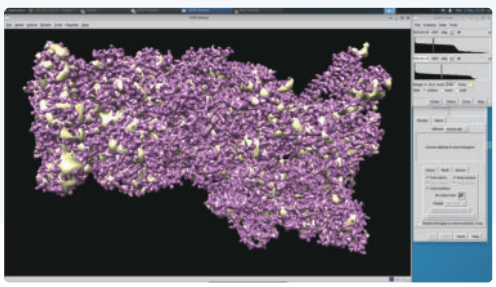
02

仿真前后处理



03

远程数据可视化与处理



04

云游戏

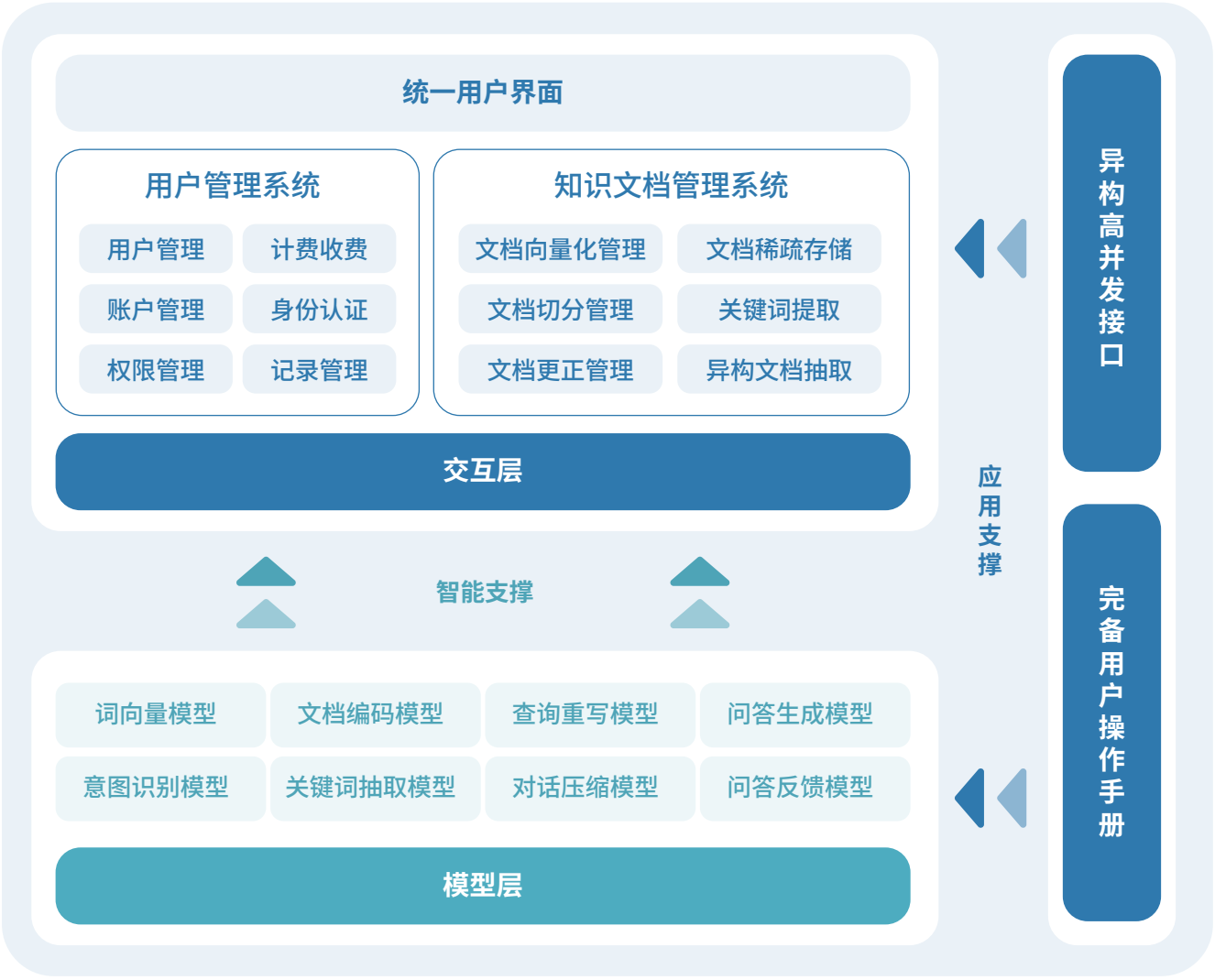
05

小蒜 大模型智能体平台

产品概述

小蒜是一个私有化部署的本地化大模型智能体系统，具备强大的智能体和知识库能力。相比起开源大模型方案，小蒜在生产级场景中具有显著优势，主要体现在数据安全、自主可控、性能优化、合规适配、长期服务、使用便捷度等方面。首先，本地化部署的小蒜实现了数据私有化，保障私有数据安全、服务安全。同时，小蒜允许不同的个人、部门、团队快速创建和修改属于自己的智能体和知识库，并提供精细化的增强型基于角色的访问控制模型，允许用户和群组自主调整智能体和知识库的使用权限，确保公共知识和私有知识之间的隔离。

平台架构



功能亮点

开箱即用，快速部署

平台针对技术小白用户设计，提供无代码化智能体配置功能，通过简洁直观的操作界面，实现一键配置、快速上线，大幅降低技术门槛，让智能问答应用触手可及。

01

多源知识融合

集成结构化数据库、非结构化文本和互联网资源，实现跨领域知识的智能关联与融合，提供全面准确的问答服务，满足多样化信息需求。

02

深度语义理解

精准把握用户意图和问题本质，实现上下文关联分析，提供智能、精准的回答和建议。

03

全生命周期学习优化

基于用户交互持续学习，自适应迭代知识体系和匹配算法，实现系统性能的自我进化，提供日益精准和个性化的问答服务。

04

私域定制匹配

深度整合企业专有知识库，可实现精准的领域适配，为多样化行业、组织提供专业智能问答解决方案。

05

可解释性输出

提供透明知识来源，呈现答案生成的逻辑链条，增强用户信任度，支持结果验证和深入探索，实现人机协作决策。

06

应用场景

企业智能助手

内部知识库问答，快速检索制度、流程、产品资料，保障数据不外泄。部门级智能体定制，如HR解答政策、IT支持故障排查，权限隔离灵活可控。

政务与金融合规场景

敏感数据本地处理，满足行业合规要求；提供可解释性输出，辅助审

教育科研机构

整合学术资源与私有研究数据，构建学科智能体，支持师生精准问答与知识关联。

医疗与专业服务

融合病例、药品库等结构化数据与文献，提供诊断辅助建议，同时保护患者隐私。

制造业与客服中心

关联产品手册、工单记录，实现故障自诊断；持续学习优化，提升客服响应准确率。

应用案例

基于小蒜的高性能计算领域智能问答模型已上线北京大学算力平台系统SCOW，该模型配备了专业高性能领域知识库、精准的自然语言处理能力，为高性能计算从业者。

基于小蒜的资产分类模型已成功在北京大学上线使用。该模型充分利用小蒜先进的语义检索技术，能够精准解析和分类固定资产信息。它可根据国家标准《固定资产等资产基础分类与代码》，从两千余个细分类别中快速定位并准确归类各类固定资产。

MaaS 大模型服务平台

产品概述

DeepSeek、通义千问、豆包、百度、OpenAI等厂商相继推出各具特色的大语言模型。随着开源模型能力的提升，大模型本地部署的需求也持续上升，本地部署、云端部署的大模型服务形式同时存在。然而，这种多样性也带来了使用上和管理上的问题，面临管理成本高、资源利用率低、本地算力和云端算力融合难、网络稳定性要求高等问题。为此，我们推出MaSS大模型服务平台，实现"一点接入、全网通达"的服务模式。向下兼容各类大模型API，向上提供统一接口，并集成多维调度、用户认证、权限管理、统一计费等能力，打造企业级的AI中台解决方案。

平台架构

功能亮点

支持多模型调度

通过多模型调度，实现了同一个模型在本地、云端多来源的融合调度，本地资源紧张时自动负载到云端资源，既能高效利用本地算力，又能弹性支持高并发使用。

01

支持批量推理调度

大模型使用有潮汐特性，往往白天使用量大，凌晨使用量小。通过批量推理调度，可以将用户批量提交的请求动态调度至闲时执行，充分利用本地硬件资源。

02

多模型、多用户、高并发

管理不同来源的模型，通过对接SCOW可进一步管理大规模推理集群，管理多用户与团队，涵盖权限、计费、预算等管理功能。面向大规模团队的高并发使用需求，系统经过详细测试，在高达5k的并发下仍能准确计费。

03

高效的对话记录与调试

支持开启对话记录，一键复现历史调用，便于高效调试，加快上层应用开发进度。

04

Email: luoya@icode.pku.edu.cn
Web: https://www.pkuscow.com

北京大学
长沙计算与数字经济研究院
PKU-Changsha Institute for Computing and Digital Economy

长沙尖山塔图
科技有限公司