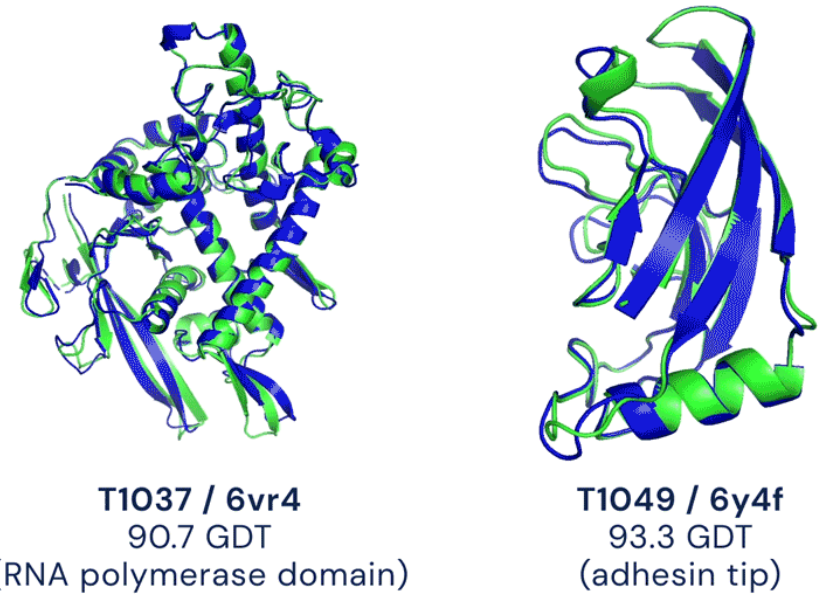
AlphaFold

在超算上的使用

苏小明

2021年9月

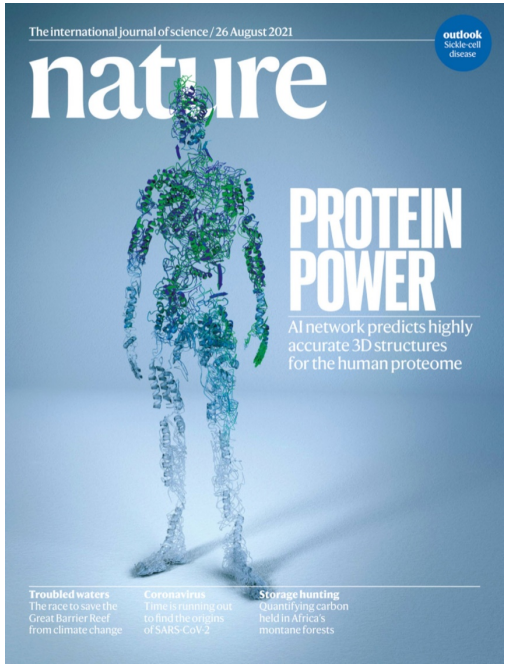
AlphaFold



AlphaFold 能够快速预测蛋白质结构，精度可媲美冷冻电镜

施一公评价：AlphaFold 蛋白结构预测是本世纪最重要的科学突破之一

Nature 评价：It will change everything

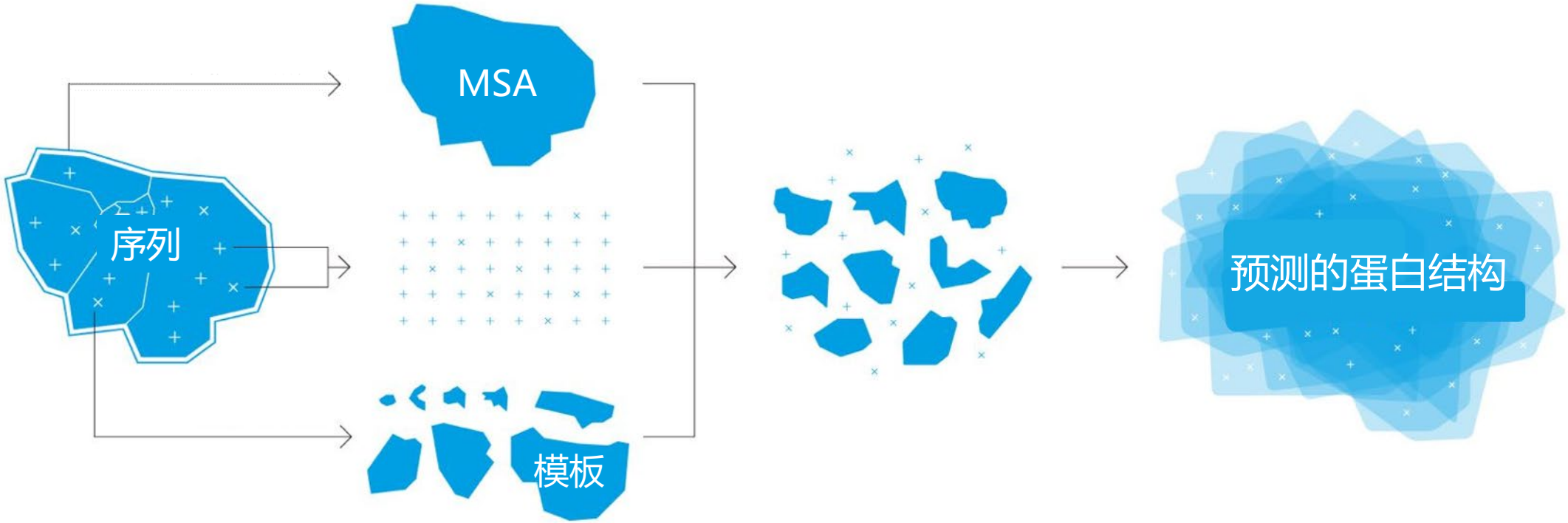


August 26, 2021

August 20, 2021

August 28, 2021

AlphaFold 框架



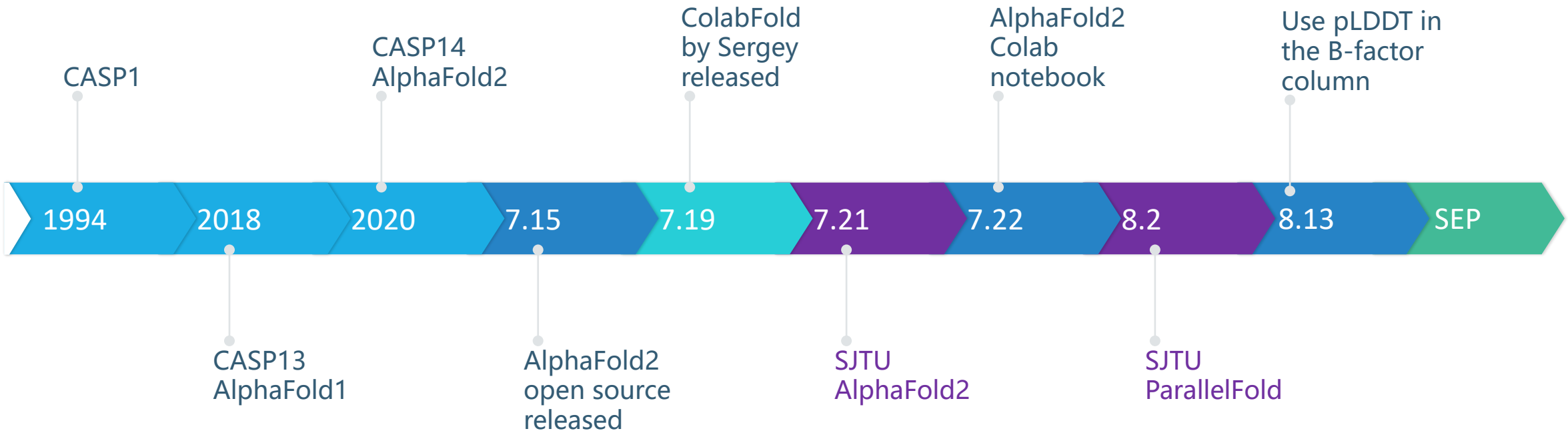
第一部分：序列特征生成

(使用 CPU)

第二部分：神经网络预测

(使用 GPU)

AlphaFold Timeline



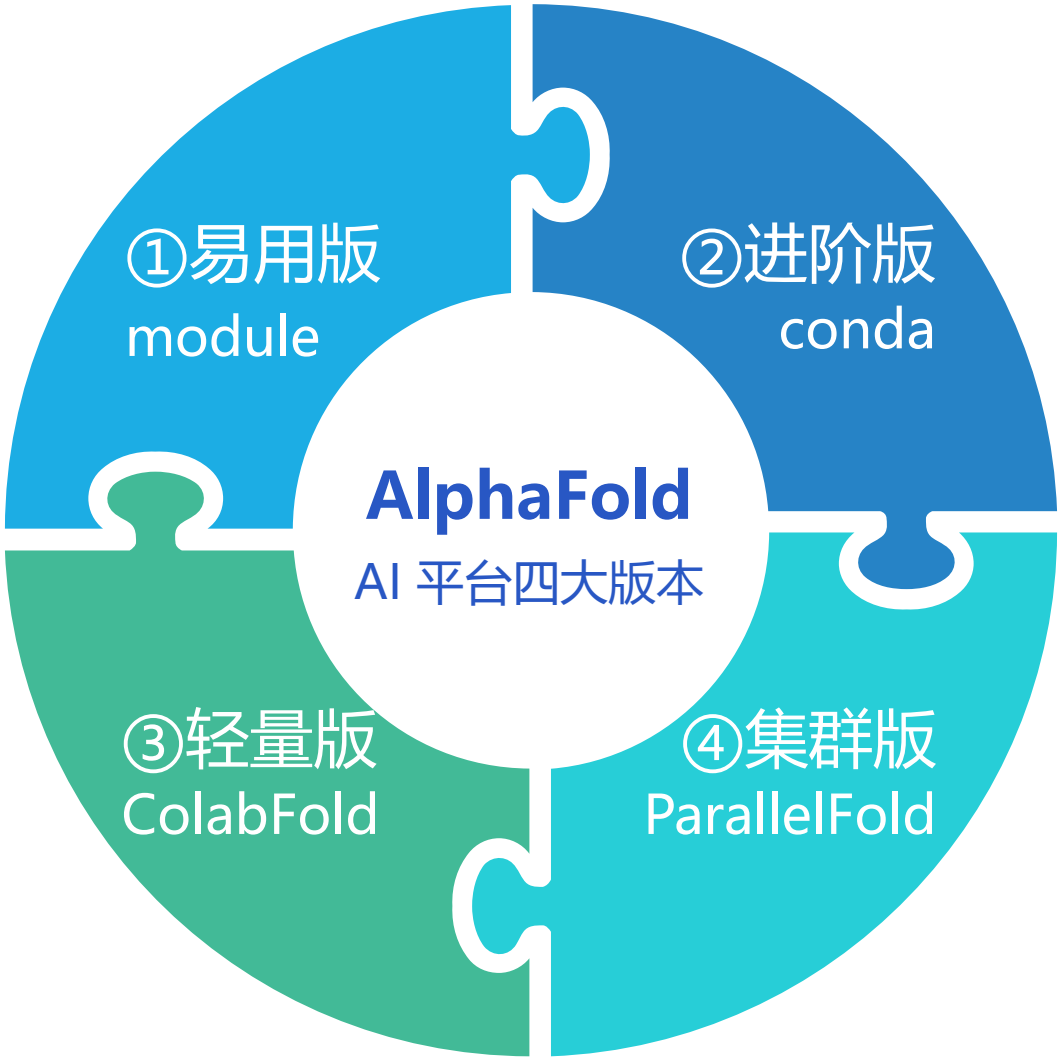
交我算平台 AlphaFold 四大版本

无需安装
满足大部分计算需求

module load alphafold

计算快速
几分钟算完短序列

使用 MMseq2 替代 JackHMMER
Google Colab 本地部署



自定义使用
参数、设置修改

可选择 pTM, Amber, Recycling 次数等

大规模计算
CPU与GPU分离计算

省时、省钱、一次算完几百几千个蛋白

AlphaFold 文档

超算平台用户手册 | Search | SJTU HPC Docs 36 Stars · 26 Forks

上海交大超算平台用户手册 文档 | 软件 | 生物信息计算

超算平台用户手册
快速上手
系统
帐号
可视化平台
登录
数据传输
作业
容器
软件
软件模块使用方法
编译器、库、工具
基准测试
AI 计算
原子分子工程计算
生物信息计算
AlphaFold2
AUGUSTUS
BASIL-ANISE
BatVI
BCFtools
BEDTOOLS2
BICseq2
BISMARK
Blast-plus
BOWTIE
BOWTIE2
BreakDancer
BWA
cd-hit
CDO
cdsapi

AlphaFold2

AlphaFold2 基于深度神经网络预测蛋白质形态，能够快速生成高精确度的蛋白质 3D 模型。以往花费几周时间预测的蛋白质结构，AlphaFold2 在几小时内就能完成。

我们对 AlphaFold 持续优化，可至 ParaFold 网站了解我们的工作：https://parafold.sjtu.edu.cn

我们将于 9 月 15 日（周三）在闵行校区图书信息楼 9 楼举办《AlphaFold 使用与优化》专题培训，欢迎大家参加：报名问卷

AlphaFold2 四大版本

交大 AI 平台提供 AlphaFold 四大版本：

- module 版，更新日期：2021 年 9 月 12 日。加载即用，免除安装困难。可满足大部分计算需求；
- conda 版，支持自主选择 model 模型、pTM 计算、Amber 优化、本地数据集、修改 Recycling 次数等；
- ColabFold 版，快速计算，含有多种功能，由 Sergey Ovchinnikov 开发。可在交大 DGX-2 上通过 conda 安装使用；
- ParallelFold 版，支持 CPU、GPU 分离计算，适合大规模批量计算。

版本一：module

module 版为全局部署的 `alphafold/2-python-3.8`，更新日期：2021 年 9 月 12 日

module 使用前准备

- 新建文件夹，如 `alphafold`。
- 在文件夹里放置一个 `fasta` 文件。例如 `test.fasta` 文件（内容如下）：

```
>sp|P68431|H31_HUMAN Histone H3.1 OS=Homo sapiens OX=9606 GN=H3C1 PE=1 SV=2
MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTE
LLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEACEAYLVGLFEDTNLCAIHAKRVTI
MPKDIQLARRIRGERA
```

Contents

1 版本一：module

2 版本二：conda

3 版本三：ColabFold

4 版本四：ParallelFold

https://docs.hpc.sjtu.edu.cn/app/bioinformatics/alphafold2.html

版本一：module（满足大部分计算需求）

slurm script

```
#!/bin/bash
#SBATCH --job-name=alphafold
#SBATCH --partition=dgx2
#SBATCH -N 1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=6
#SBATCH --gres=gpu:1
#SBATCH --output=%j.out
#SBATCH --error=%j.err

module load alphafold

run_af2  $PWD --preset=casp14  test.fasta  1
--max_template_date=2021-09-12  2
```

若使用 2 块 GPU，改为：

```
#SBATCH --cpus-per-task=12
#SBATCH --gres=gpu:2
```

output

<target_name>/

1. features.pkl
   ranked_{0,1,2,3,4}.pdb
   ranking_debug.json
2. relaxed_model_{1,2,3,4,5}.pdb
   result_model_{1,2,3,4,5}.pkl
3. timings.json
4. unrelaxed_model_{1,2,3,4,5}.pdb

msas/

- bfd_uniclust_hits.a3m
- mgnify_hits.sto
- uniref90_hits.sto

https://docs.hpc.sjtu.edu.cn/app/bioinformatics/alphafold2.html

版本二：Conda（自定义）

Conda clone

- 克隆集群上已备好的 conda af10 环境
- 直接使用 AlphaFold, ColabFold, ParallelFold
- 操作简易

Conda 完整安装

- 从头开始安装各软件、下载文件和补丁
- 逐个安装 AlphaFold, ColabFold, ParellelFold
- 操作有难度

https://docs.hpc.sjtu.edu.cn/app/bioinformatics/alphafold2.html

版本二：Conda（完整安装）

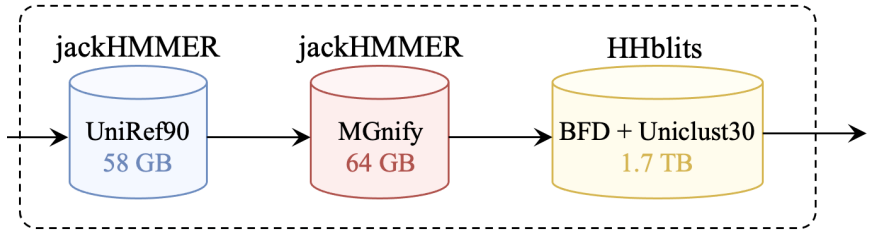
安装事项

版本选择

CUDA driver	11.0	10.2
python	3.8/3.7	3.8/3.7
CUDA toolkit	11.0.3	10.1
cuDNN	8.0.4	7.6
jaxlib	0.1.69	0.1.69
TensorFlow	2.5.0	2.3.0

Conda 版本的优点

可指定使用 DGX-2 本地数据集



(Lustre) /scratch/share/AlphaFold/data

VS.

(DGX-2 local) /home/share/AlphaFold/data

https://parafold.sjtu.edu.cn/docs/quick-start/

版本三：ColabFold（轻量、多功能版）

ColabFold screen output

ColabFold on Linux

homooligomer: '1'

total_length: '56'

running mmseqs2 1

207 Sequences Found in Total

model_1_ptm_seed_0 recycles:3 tol:0.65 pLDDT:58.39 pTMscore:0.40

model_2_ptm_seed_0 recycles:3 tol:1.11 pLDDT:89.54 pTMscore:0.67

model_3_ptm_seed_0 recycles:3 tol:0.56 pLDDT:79.10 pTMscore:0.54

model_4_ptm_seed_0 recycles:3 tol:0.81 pLDDT:64.31 pTMscore:0.49

model_5_ptm_seed_0 recycles:3 tol:1.70 pLDDT:90.48 pTMscore:0.70

model rank based on pLDDT

rank_1_model_5_ptm_seed_0 pLDDT:90.48

rank_2_model_2_ptm_seed_0 pLDDT:89.54

rank_3_model_3_ptm_seed_0 pLDDT:79.10

rank_4_model_4_ptm_seed_0 pLDDT:64.31

rank_5_model_1_ptm_seed_0 pLDDT:58.39

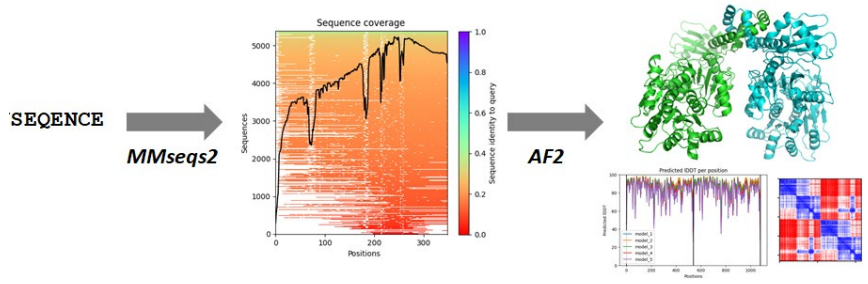
predicted alignment error 2

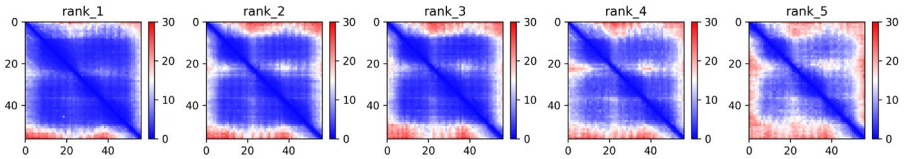
predicted contacts

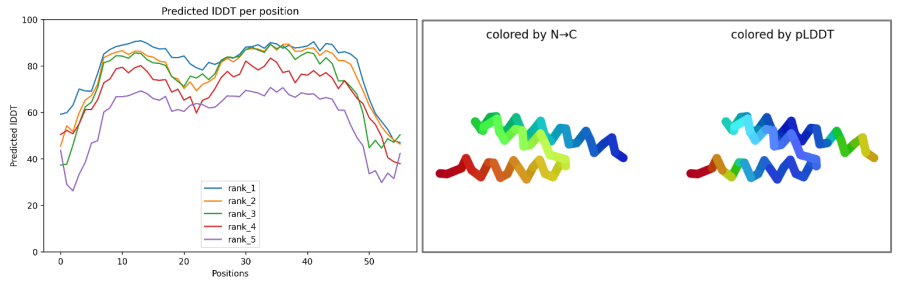
predicted distogram

predicted LDDT

ColabFold files generated



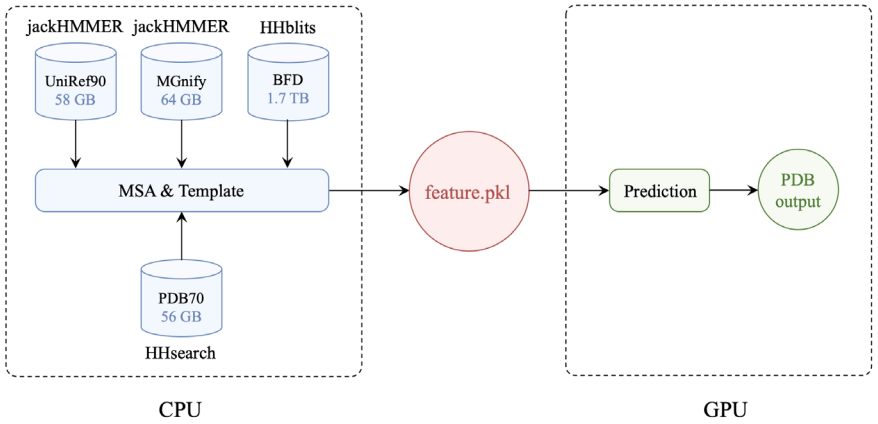




https://docs.hpc.sjtu.edu.cn/app/bioinformatics/alphafold2.html

版本四：ParallelFold 集群版

CPU 与 GPU 分离，支持集群上的大规模计算



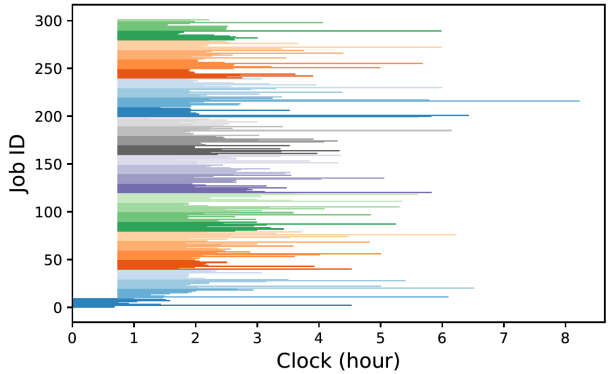
原理：自动检测 feature.pkl 文件是否存在

- 若存在，执行 GPU 计算
- 若不存在，从头开始计算
- 也可仅算 CPU 部分

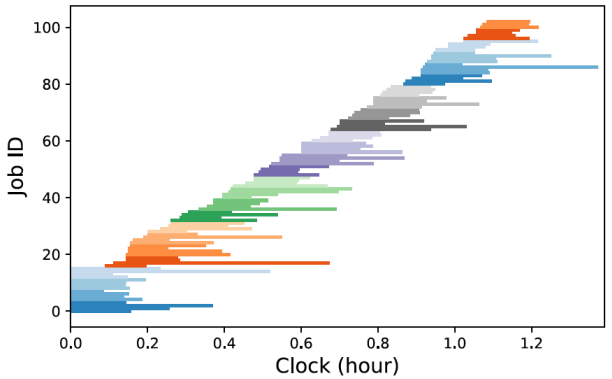
应用：大规模计算，CPU 至 cpu/small 队列

计算效果：16 卡1小时算出100个蛋白

6小时内算完300个蛋白的MSA



1小时算完100个蛋白的GPU部分（16卡）

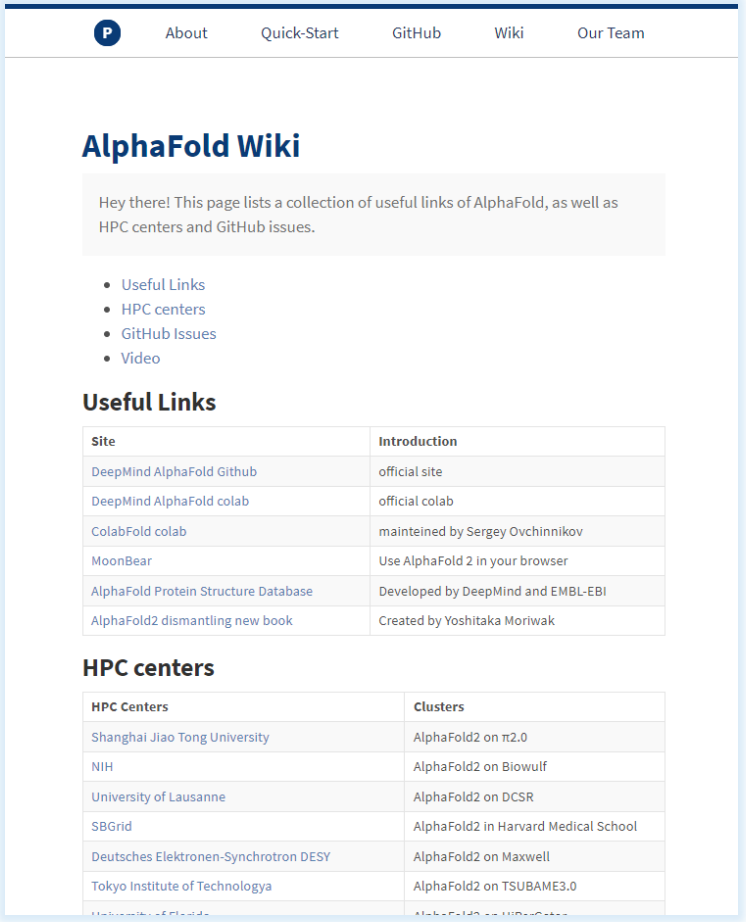


https://docs.hpc.sjtu.edu.cn/app/bioinformatics/alphafold2.html

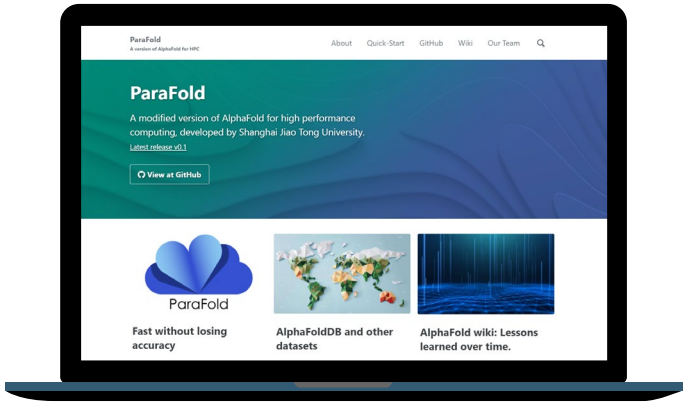
ParaFold

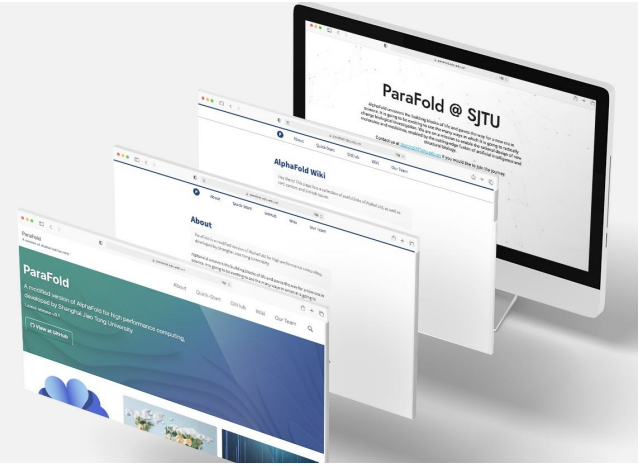
我们的优化网站 https://parafold.sjtu.edu.cn

优化总结



wiki





常见问题





常见问题 1：AlphaFold 运行速度

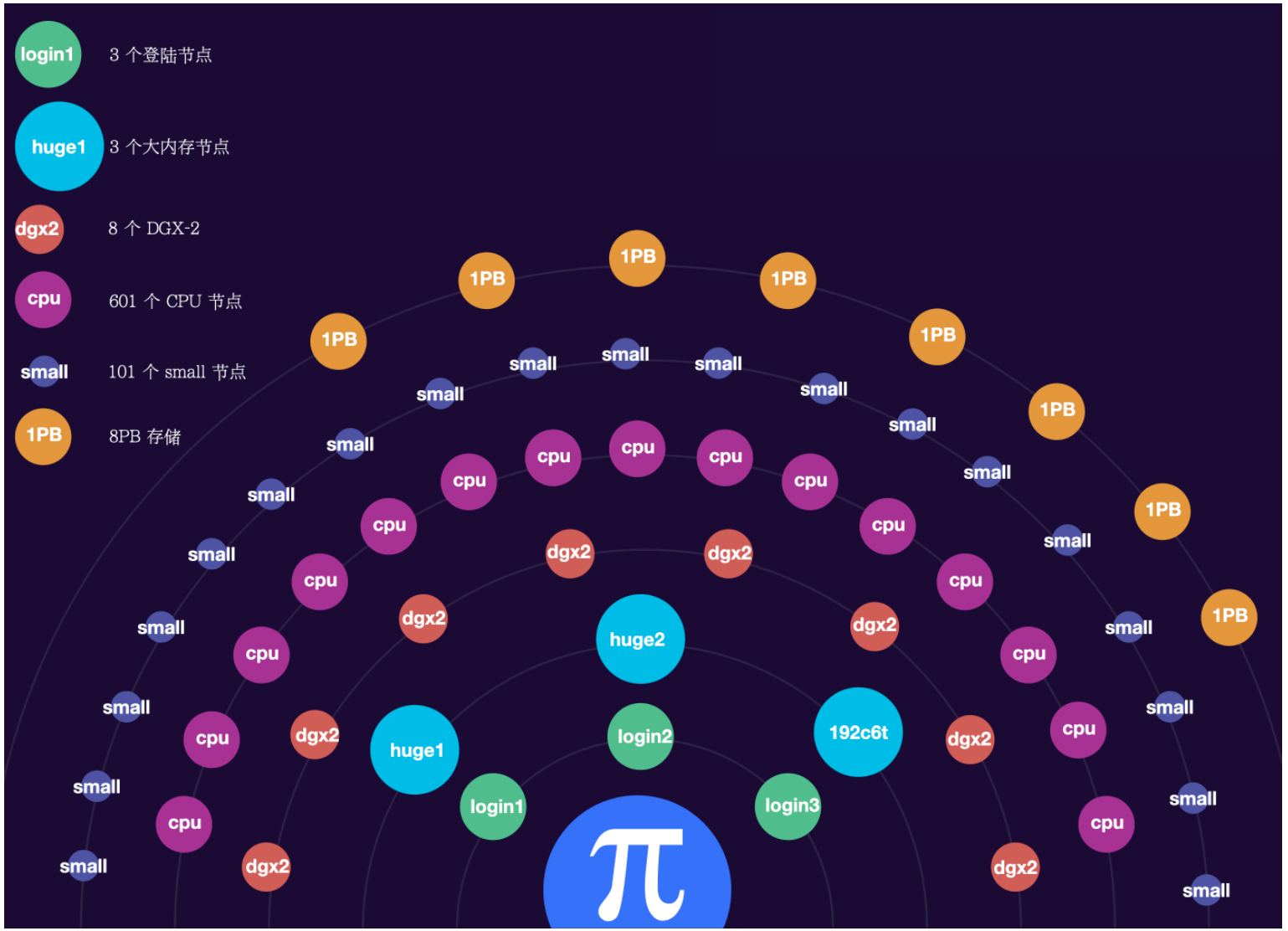
AlphaFold's speed

1. If you read the Nature paper, you'll see that AlphaFold 2 is more accurate, and the GPU times are in fact very fast: 0.6 minutes at 256 residues, 1.1 minutes at 384 residues, and 2.1 hours at 2,500 residues.

2. In general, you can expect the time to grow with the length of the protein and the MSA search taking up to a few hours with a slow disk / CPU.

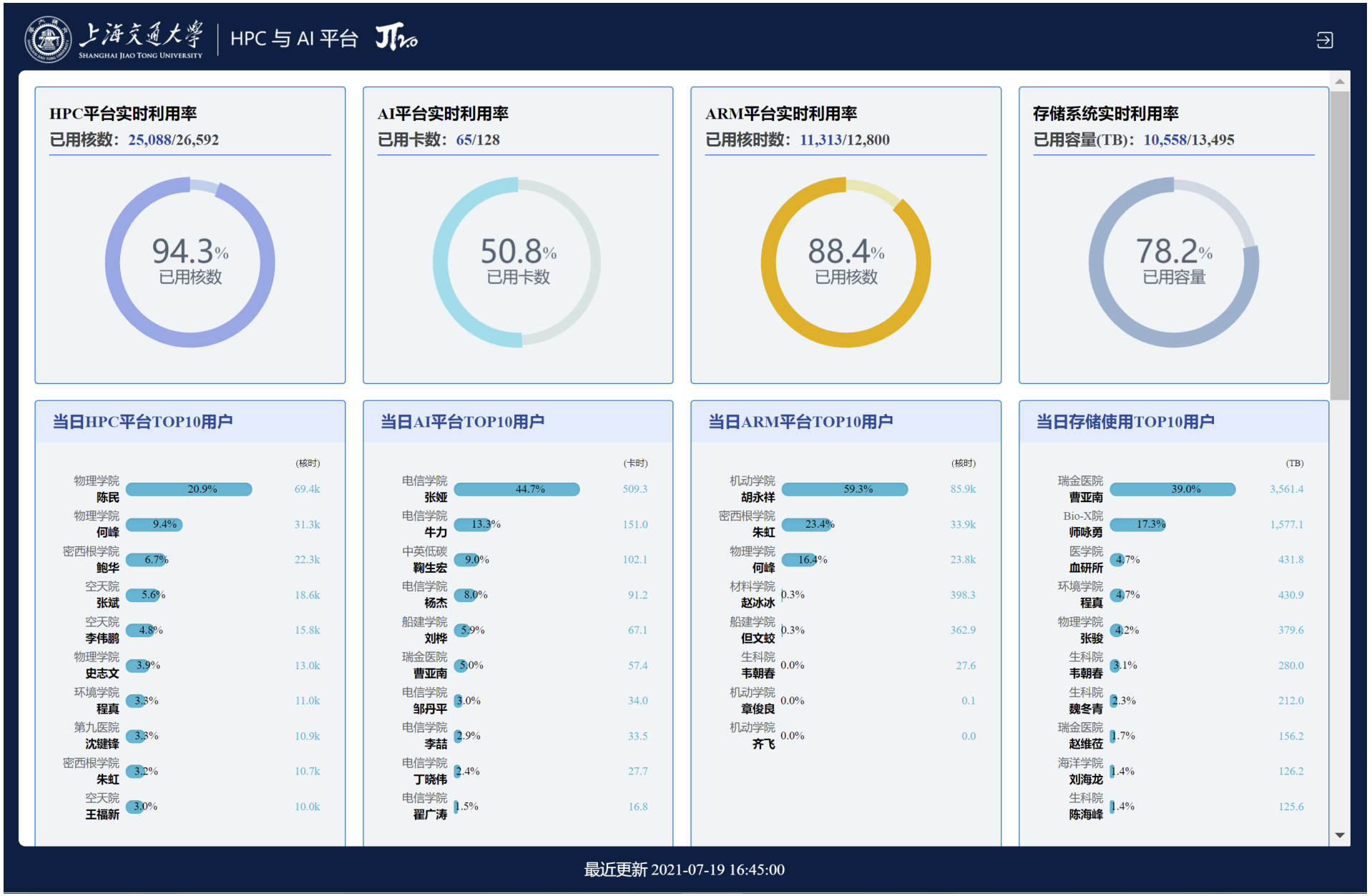
   For the actual folding (i.e. running the AlphaFold model), the disk speed doesn't matter anymore, what matters is whether you are using a GPU and its performance.

常见问题 2：排队



https://account.hpc.sjtu.edu.cn/top/

常见问题 2：集群利用率



https://account.hpc.sjtu.edu.cn/top/

交我算平台新服务

①科研软件工程

鼓励支持交大师生自研软件

我们将协助部分软件的优化

②科学大数据/AI平台

交我算平台存储能力将扩增至 100PB

欢迎课题组科研数据存放、分析与合作

服务①：科研软件工程

①代码现代化

以软件工程方法管科研软件开发

- 代码复杂度评估
- 项目开发进度管理
- 结果正确性验证
- 代码规范化

②性能优化

移植到异构平台，优化代码性能

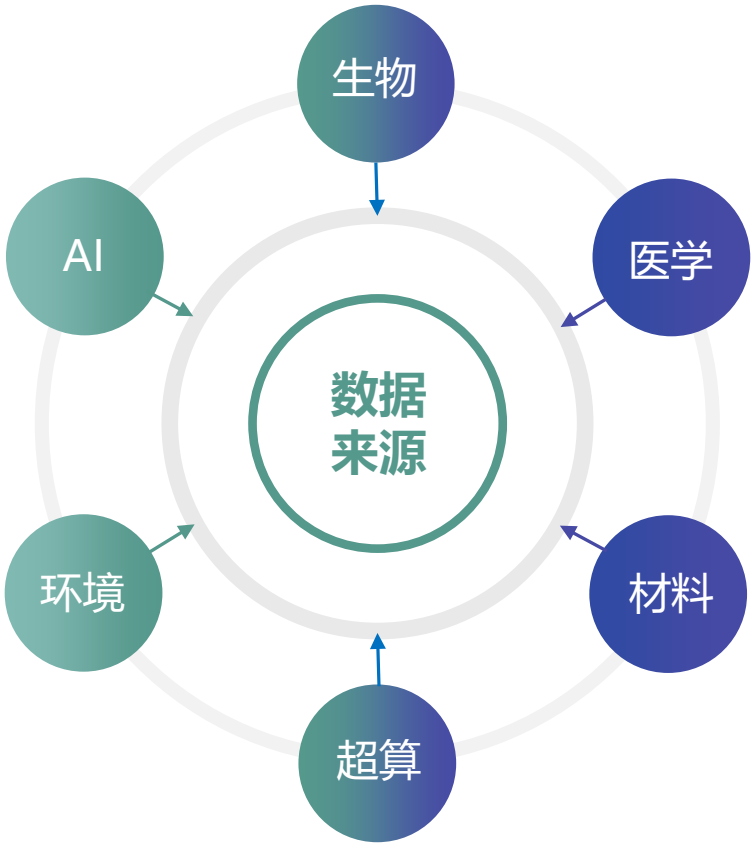
- 移植到国产平台
- 移植到GPU平台
- 架构-算法协同设计
- 并行优化

③应用推广

输出科研软件工程能力

- 建立软件站点
- 编写软件文档
- 向用户开放使用
- 代码合并到上游社区

服务②：科学大数据/AI平台



数据建模

从大数据视角看科学问题

- 评估分析需求
- 清理原始数据
- 设计数据流图
- 选择分析算法

数据托管

搭建数据上传下载网站

- 建立数据仓库
- 补全数据信息
- 分配访问授权
- 开放访问方法

算力提供

一站提供所需计算资源

- 对接云/超算/AI
- 数据互通
- 一键分析
- 按需组合

服务③：免费教学支撑

覆盖 60 多门课程

超 3000 名学生收益

分子模拟的理论与实践
生物信息分析
物理学的数学与数值方法
飞行器设计
计算物理
计算生物化学
自然语言理解
自然语言处理
基于生物医学统计的信号处理
GPU计算及深度学习
工程学导论
统计计算与机器学习
计算材料学
强化学习
电力系统暂态稳定
仿生微型机器人技术
计算机系统结构实验
多媒体通信系统与实现
最优化理论基础
……

上云前

费钱：学院要建机房、购买机器用于教学
费力：老师亲自指导学生遇到的计算问题
费时：学生需4～8小时进行下载、安装、调试软件
费心：受限计算能力，可选实验种类少

上云后

省钱：云平台免费提供教学所需资源
省力：中心安排专业技术人员做助教
省时：云上镜像，统一部署，开箱即用
仅软件部署，至少已节省师生 12,000 小时
省心：算力强大，老师方便选择各种实验

计算平台教学支撑

为全校师生提供免费教学上机资源

算力强大	标准环境	快速交付	支持服务
统一标准化的教学实践环境	一键点击分钟级获取	较个人电脑快数十倍	精准软硬件配置精准时间配置

交我算平台介绍

高性能计算平台π2.0

国内高校**前列**

26000核 Intel 6248
100G OPA高速网络

2019年建成

云计算平台jCloud2.0

国内高校**最大**

12000核 Intel 6148
118TB内存

2018年建成

人工智能计算平台

国内高校**唯一**

8台 DGX-2
128张 Tesla v100 GPU卡

2019年建成

融合存储平台

国内高校**前列**

现有存储30PB
计划扩容15PB

2019年建成

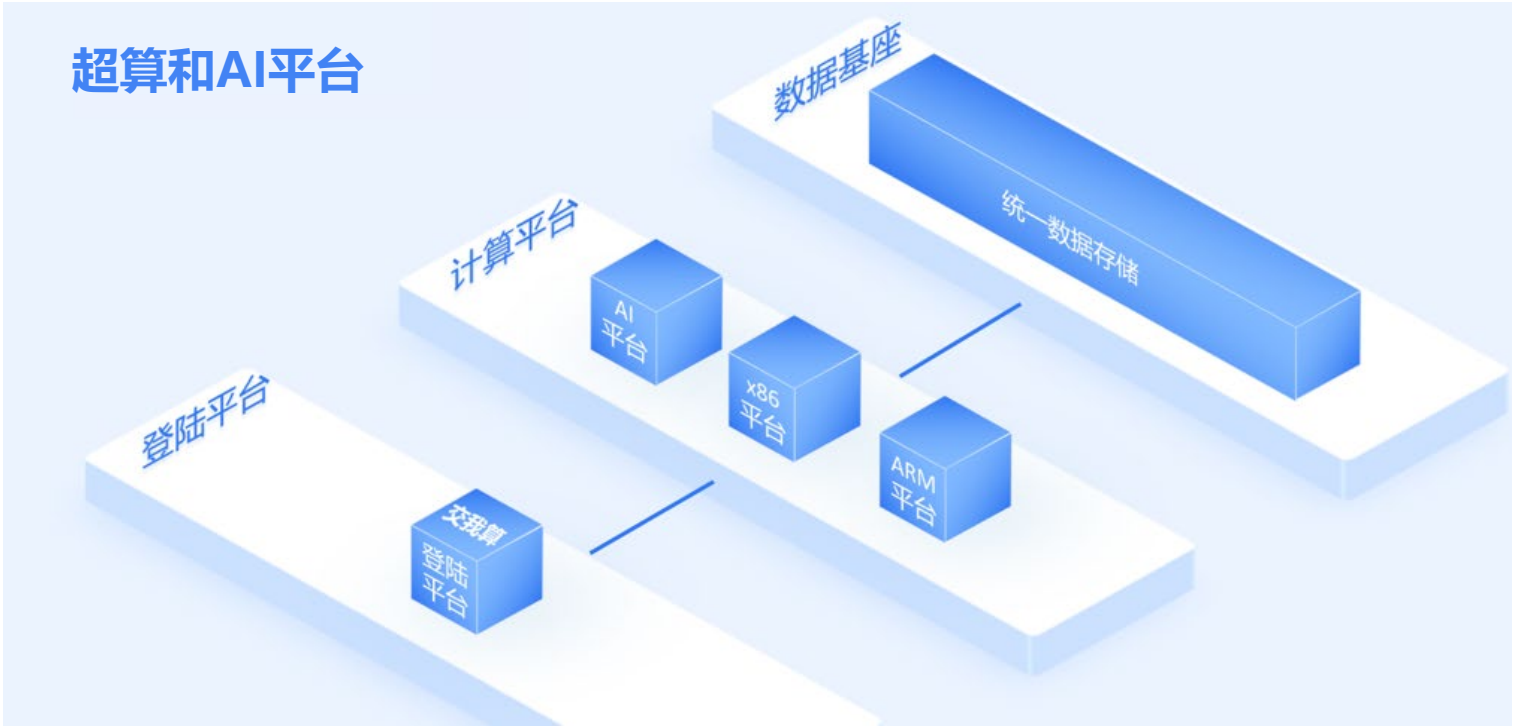
ARM高性能计算平台

国内高校**首个**

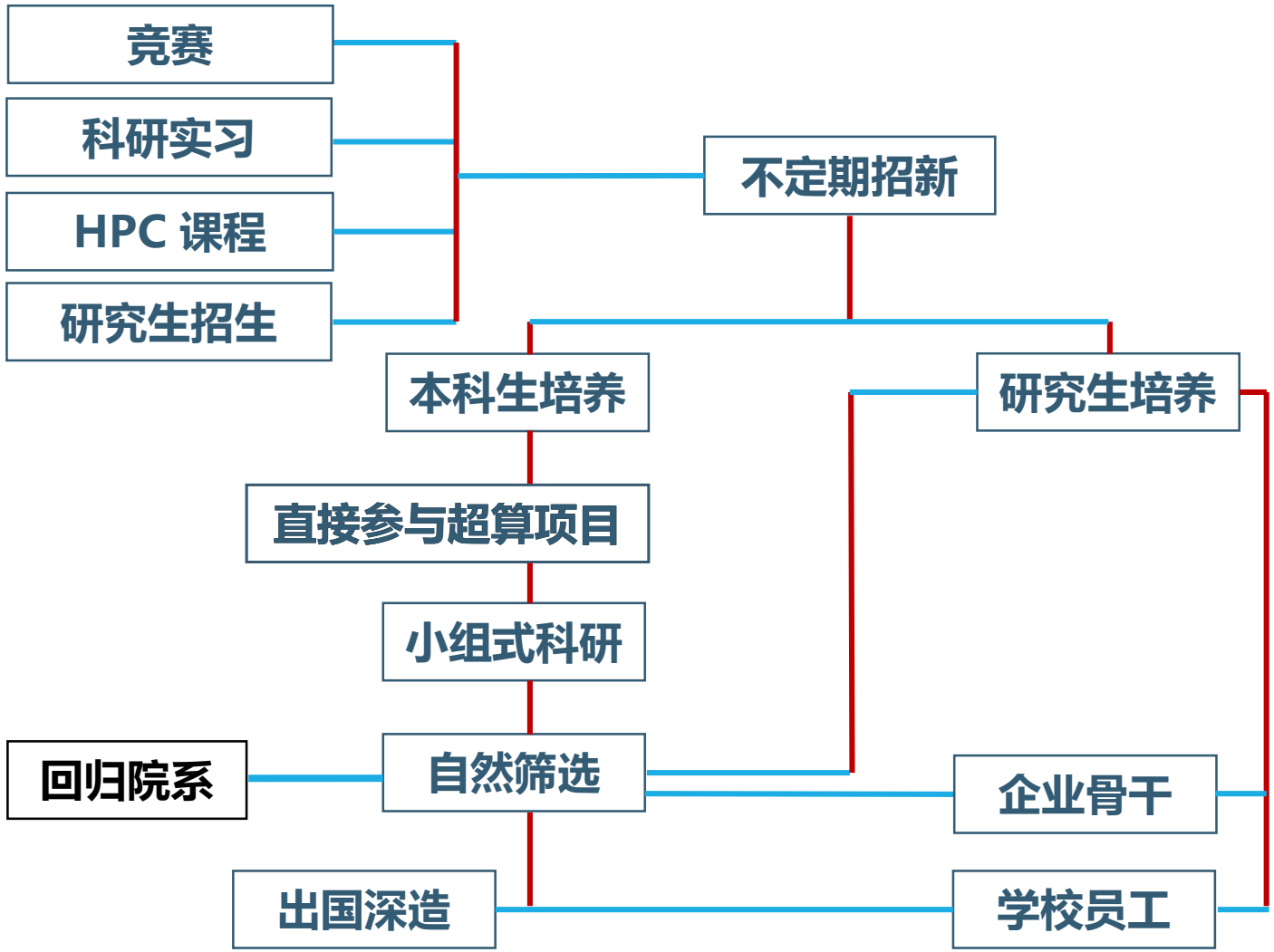
12800 CPU核

2021年建成

超算和AI平台

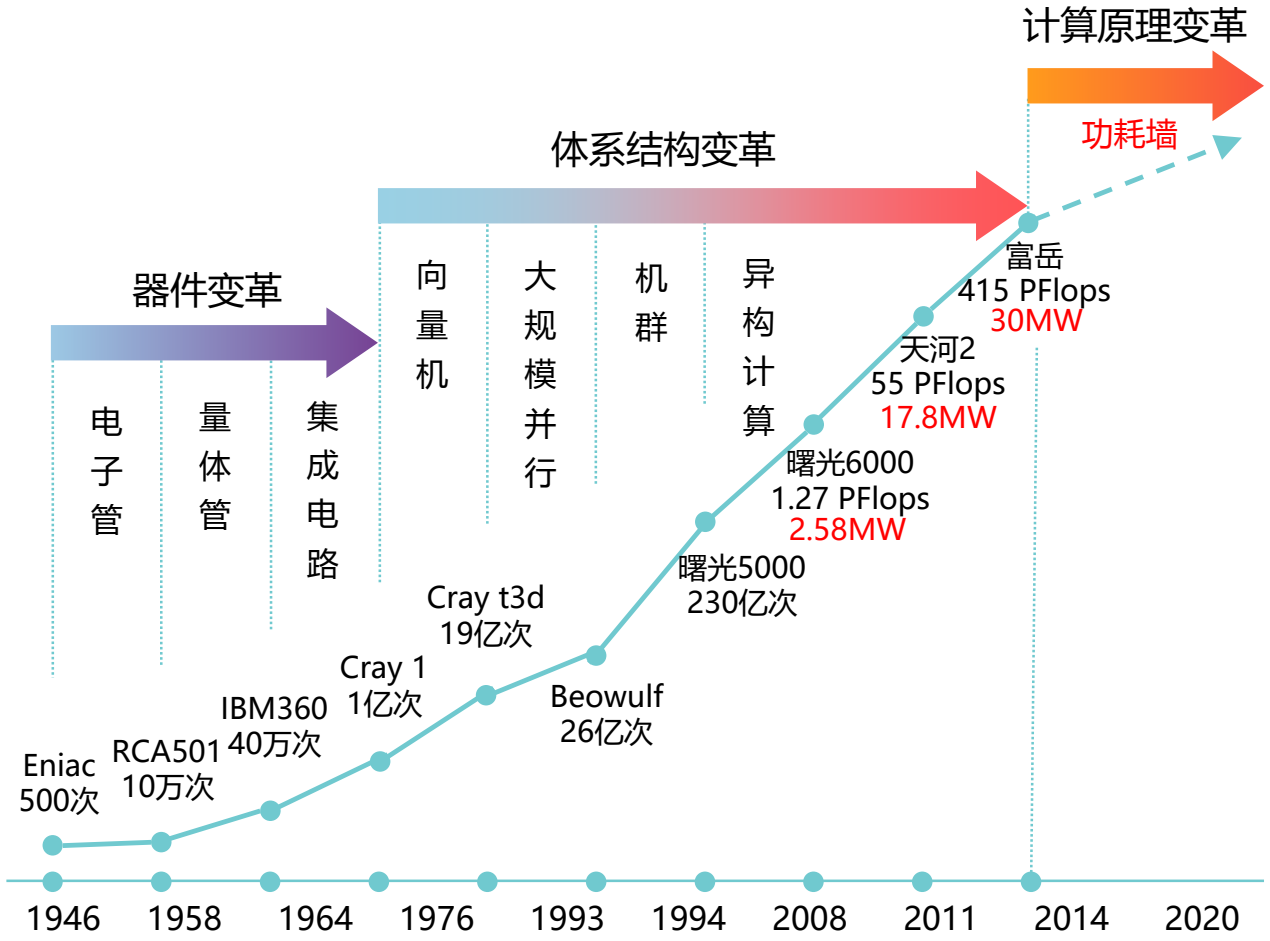


交我算平台招新


竞赛
科研实习
HPC 课程
研究生招生
不定期招新
本科生培养
研究生培养
直接参与超算项目
小组式科研
回归院系
自然筛选
企业骨干
出国深造
学校员工

算力发展

算力每 10 年提升 1000 倍



8 年将超算搬到老师桌上

2013 年

上海交大 π 1.0 超算

CPU 0.15 PFlops

2021 年

NVIDIA A100 工作站

0.16 PFlops

≈

Once Upon a HPC

太湖之光
2016
93Pflops

天河一号
2010
4.7Pflops

CRAY 1
1976
160Mflops

Japan Fugaku
2020
416Pflops

IBM SUMMIT
2018
200Pflops

天河二号
2013
33.86Pflops

Japan
EarthSimulator
2002
40Tflops

Intel ASCI Red
1997
1Tflops

交大超算建设

π 1.0 (2013年)

- TOP 500 排名第 158

π 2.0 (2019年)

- 2.1 Pflops

ARM (2021年)

- 100 个 ARM 节点

联想超算 (2021年底)

- 联想杨元庆捐赠 1 亿元
- 相当于 2.5 ~ 3 个 π **2.0**

计算