高能所计算平台培训

徐吉平　计算中心
中国科学院高能物理研究所

主要内容

- IHEP计算平台简介
- 账号及登录
- 文件存储系统
- 计算作业系统
- 常见FAQ
- 典型使用示例

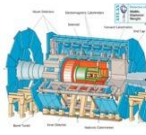
主要内容

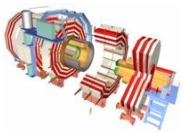
- **IHEP计算平台简介**
- 账号及登录
- 文件存储系统
- 作业系统
- 常见FAQ
- 典型使用示例

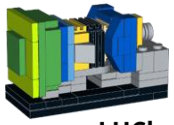
高能物理实验及计算平台概况

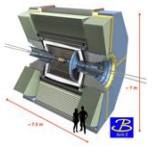
- 实验粒子物理学
- 理论粒子物理学
- 天体物理学和宇宙射线
- 加速器技术
- 核分析技术
- 光源、中子源技术
- 科学计算和存储技术
- 网络技术
- 人工智能
- 量子计算
- …

国际合作

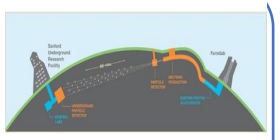

ATLAS


CMS


LHCb

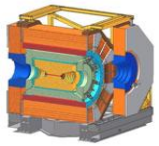

BELLE II

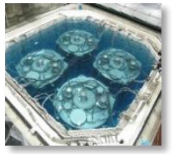

AMS02

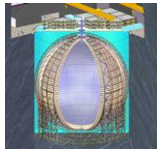

DUNE

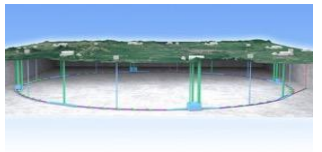
粒子物理实验

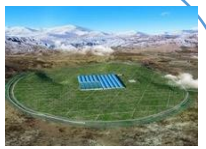
IHEP主导


BESIII


DYB


JUNO


CEPC


LHAASO

粒子物理实验


AliCPT


ASγ


HXMT


GECAM

天体物理实验

光源、中子源实验

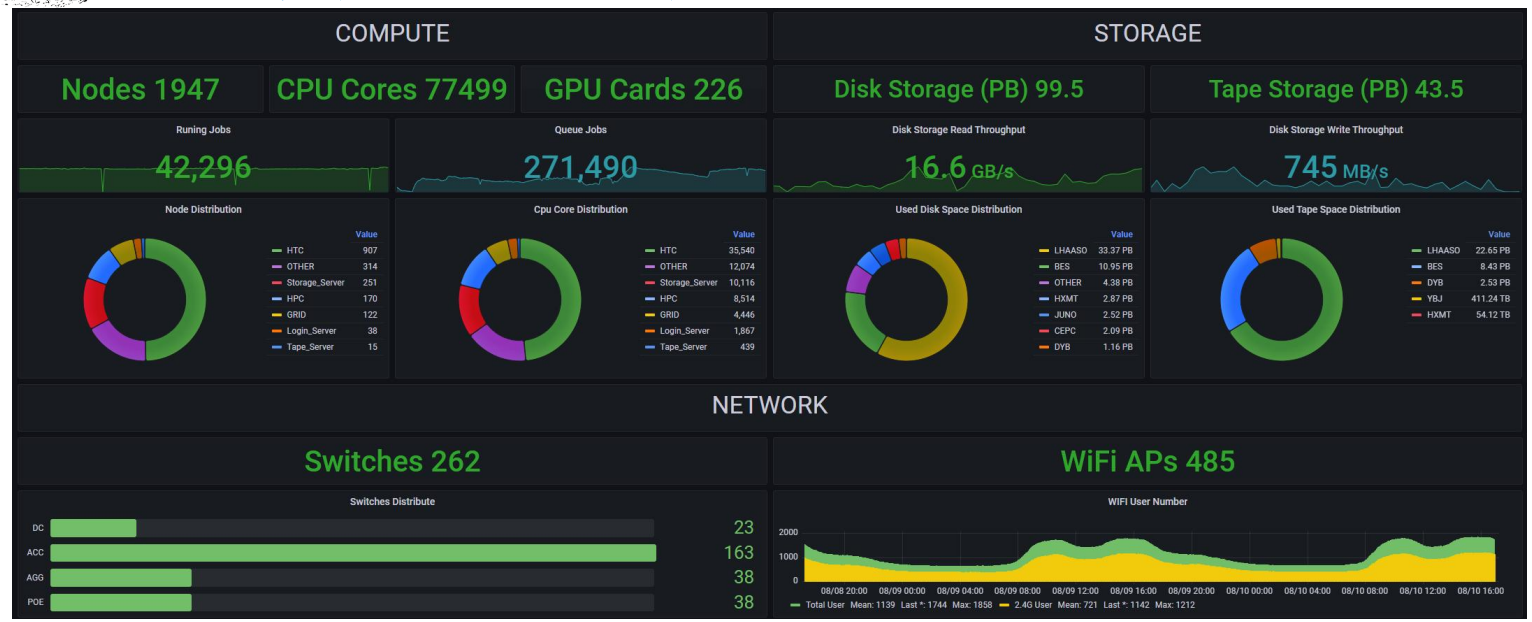
新挑战


BSRF


CSNS


HEPS

计算平台概况



当今大科学装置实验快速发展，实验数据存储和计算需求急速上升，集群规模也在持续扩大中

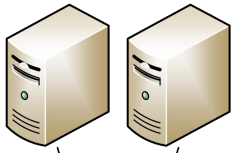
- 超过 7.7万 CPU 核和 200+ GPU卡
  - 支持高通量计算（HTC）和高性能计算（HPC）两种本地计算模式
  - 支持国际物理网格计算（WLCG二级站点）、分布式计算（Dirac & dHTC）、虚拟云计算（容器编排&虚拟机）
- 拥有100 PB 的磁盘空间和 40 PB 的磁带存储空间 （相当于存储上百万部高清电影）
  - 磁盘存储：Lustre & EOS
  - 磁带存储：EOS CTA (Castor)

计算平台架构

分布式计算系统

科学计算系统

前端登录集群

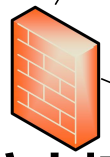


本地计算集群
HTCondor
Slurm

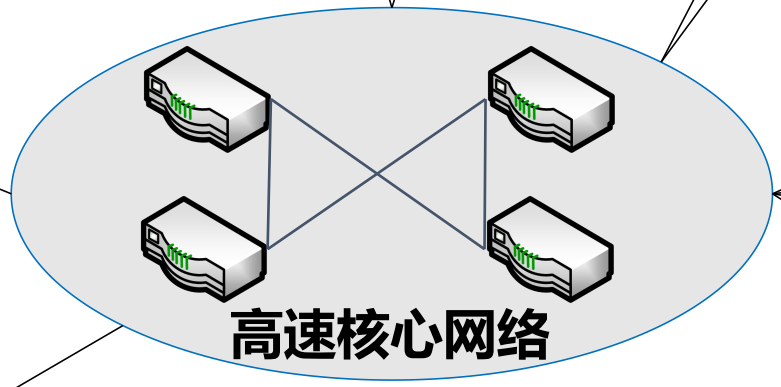
云计算集群

kubernetes 容器编排

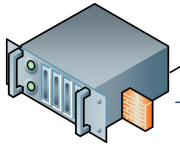
openstack. 虚拟机



防火墙

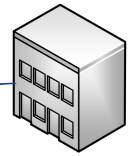


高速核心网络



数据传输

专线链路

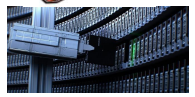


实验站

高速网络系统

SSD全闪存

磁盘存储



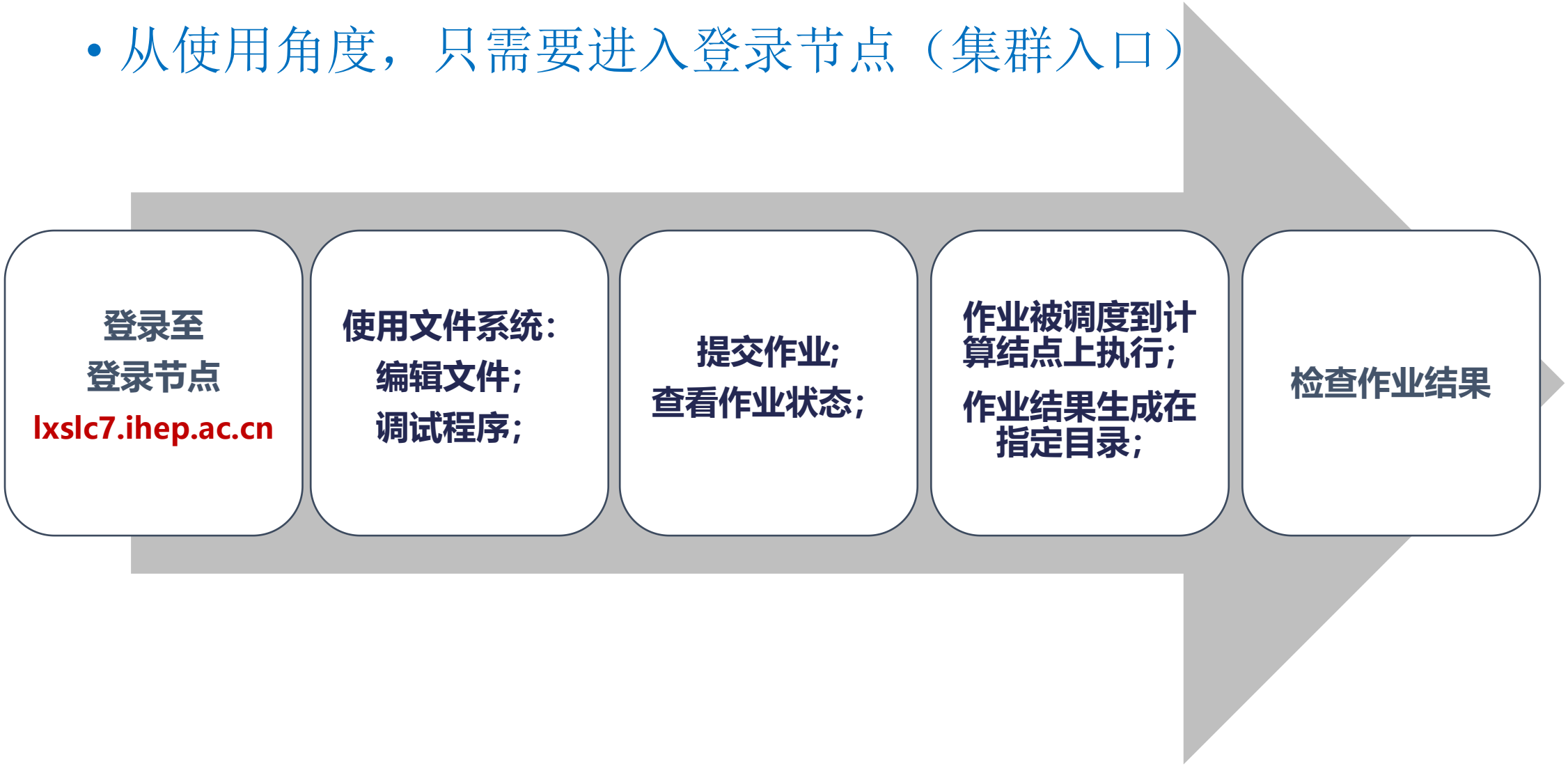
磁带存储

数据备份

海量存储系统

基本使用流程

- 从使用角度，只需要进入登录节点（集群入口）

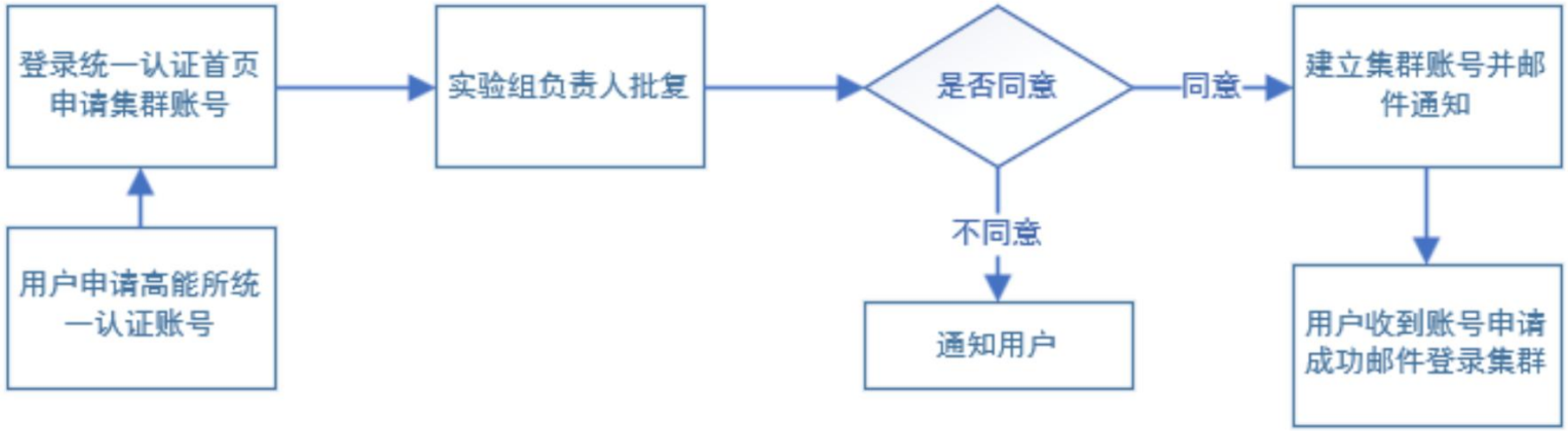


主要内容

- IHEP计算平台简介
- **账号及登录**
- 文件存储系统
- 作业系统
- 常见FAQ
- 典型使用示例

计算平台账号管理

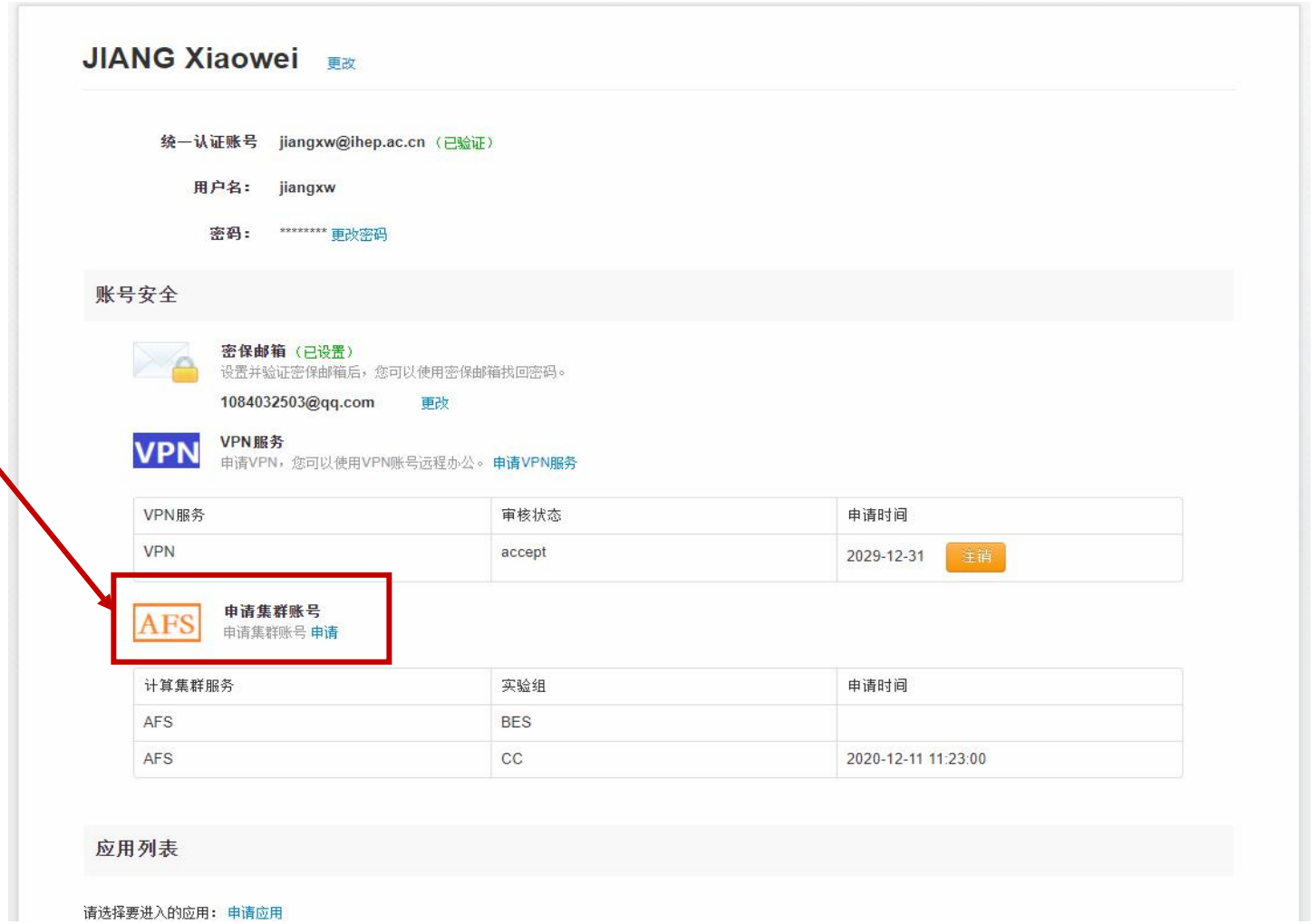
- 计算平台用户隶属于一个或多个实验，得到各个实验计算负责人批准后，该用户才能拥有计算平台的个人账号（AFS账号），使用各个实验的计算和存储资源。



账号申请（1）

IHEP统一认证用户页面包含申请集群账号的入口

https://login.ihep.ac.cn



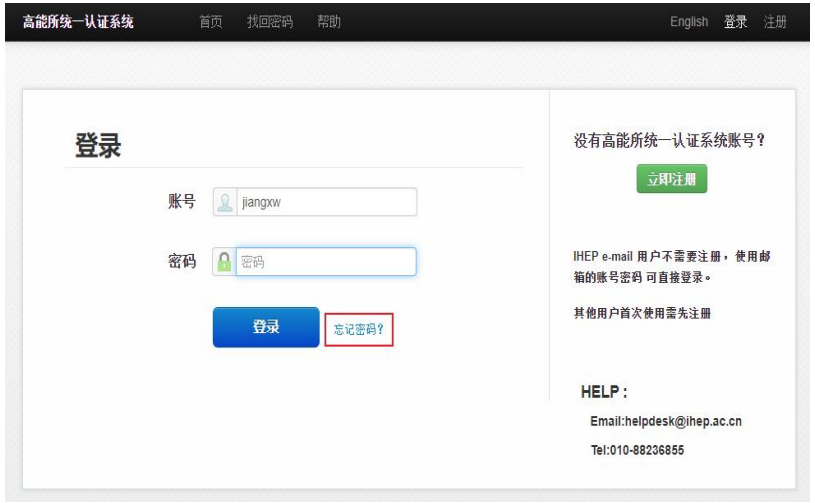
账号申请（2）

填写注册信息：

- 个人信息
  - 部门
  - 隶属应用
  - 用户组
- 导师/课题组长的信息

密码管理

- 计算集群账号密码与统一认证账号密码一致
- 密码遗忘&重置密码
- 密码修改
  - https://login.ihep.ac.cn/user/password.do?act=showChangePassword



注意事项

- **信息真实：**用户申请账号填写的手机号码、电子邮件地址必需真实有效
- **密码复杂度：**为了个人账号安全，请使用强密码格式。计算平台规定用户密码长度不得少于10位，且必须包含大写字母、小写字母、数字和特殊字符。不符合上述要求的密码将不被系统接受。
- 账号创建成功后，用户会收到通知邮件
- **注意密码有效期提醒：**在密码到期前的30天、7天和2天，用户将会分别收到三次邮件提醒
- 用户账号的信息如果发生改变，请及时与计算中心联系更新

容器使用（登录节点）

- **因安全问题,不再提供SL5等系统登录节点**
- **提供SL5/6/7容器供用户调试软件**
- 镜像查看

```
$ hep_container images
```

- 支持用户组查看 – 容器中仅可访问用户组相关存储目录

```
$ hep_container groups
```

- 进入容器环境

```
$ hep_container shell SL6
```

- 直接使用容器执行操作

```
$ hep_container exec SL6 cat /etc/redhat-release
```

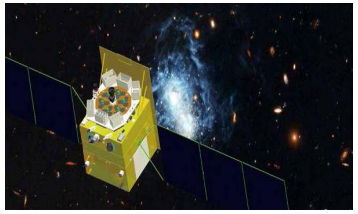
主要内容

- IHEP计算平台简介
- 账号及登录
- **文件存储系统**
- 作业系统
- 常见FAQ
- 典型使用示例

高能物理实验数据规模

当前大科学实验产生的数据在PB级别快速增长

- 北京正负电子对撞机BECPII/BESIII
  - 每年~1PB raw data，已经积累10PB+
- 江门中微子实验
  - 每年将产生3PB数据
- 高海拔宇宙线实验LHAASO
  - 每年产生10PB以上的数据
- 高能同步辐射光源HEPS
  - 每年产生超过150PB的原始数据，
    单个实验数据产生速率> 400Gbps
- 高能空间天文实验HXMT/HERD/eXTP/GeCAM等
  - 全部运行后预计每年将产生10PB数据
- 参与的国际高能物理实验的数据存储
  - CMS、ALTAS、LHCb、Belle II等

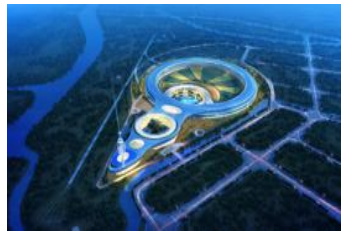



~550KM

空间天文卫星
（HXMT, GeCAME等）

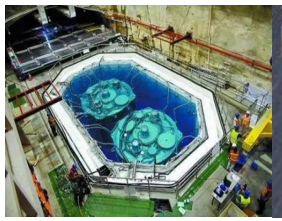
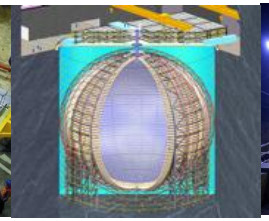

~4400M

高海拔宇宙线观测站
（LHAASO, YBJ等）

~-5M

粒子对撞机
（BEPCII, HEPS,CSNS等）

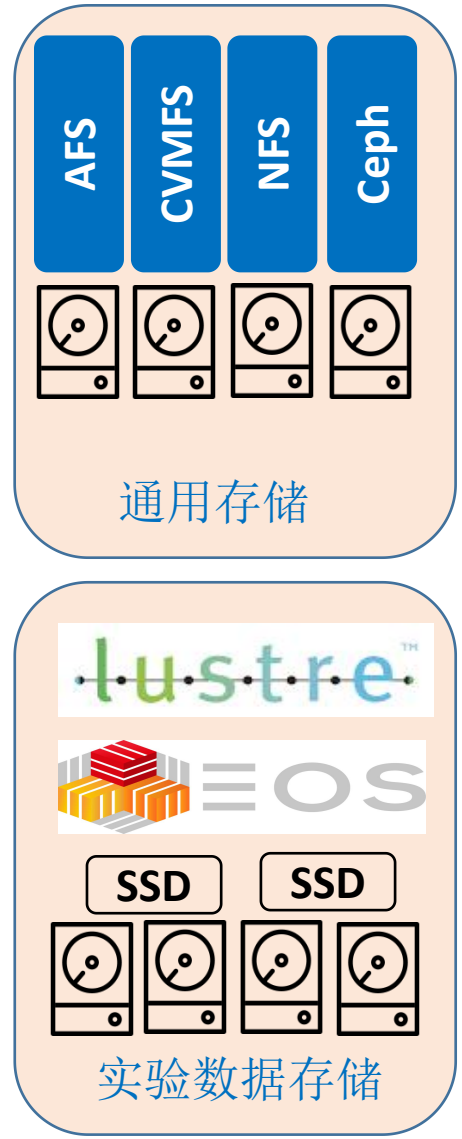

-2500M
~-300M

地下中微子实验
（JUNO, Dayabay等）

文件存储系统

为了满足个人和实验数据的存储需求，计算中心为各个实验组和用户提供了多种级别的文件存储服务

- HOME目录
  - AFS文件存储系统：个人文件存放
- 软件存储
  - CVMFS文件存储系统：主要用于存放实验软件
- 实验数据存储
  - LUSTRE文件存储系统：主要用于存放海量实验数据；个人数据目录；临时目录
  - EOS文件存储系统：主要用于存放海量实验数据（LHAASO,JUNO,HXMT）



AFS存储系统

- 用户卷
  - 用户成功登录集群后，默认进入用户afs home目录
  - 每个用户在个人afs home目录下拥有500MB空间
  - 路径：
    - /afs/ihep.ac.cn/users/a-z/username
- 如果tokens过期，将无法读写；需要更新tokens，执行kinit和aklog命令，获得权限
- 登录节点可读写，作业在计算节点运行对afs无写权限
  - 尽量不使用用户home目录作为作业提交目录

AFS存储常用操作

AFS存储系统有一系列特殊的操作命令（实操练习）

- 设置访问权限

```
$ fs setacl -dir /afs/ihep.ac.cn/users/x/xujp/mydir -acl huqb all
```

  - 只有afs命令设置的访问权限有效
  - Linux的访问权限设置（例如chmod 400)对AFS文件无效

- 查看目录Quota

```
$ fs listquota /afs/ihep.ac.cn/users/x/xujp/mydir
```

- 查看目录空间使用情况

```
$ fs quota /afs/ihep.ac.cn/users/x/xujp
```

- 用户tokens获取及延期

```
$ kinit username
Password:
$ aklog
```

LUSTRE存储系统（1）

Lustre是世界Top500计算机中使用最广泛的分布式文件系统

- 开源可定制，兼容多种底层网络、IO性能横向性能扩展等特点
- 自2008年开始使用，**目前容量40PB，提供BES，DYB，JUNO**等多个实验海量数据存储服务
- **/scratchfs** 存放临时文件
- **/workfs2** 提供用户交互使用，登录节点可读写，计算节点无写权限

	总空间（TB）	用途	用户限额	备份情况
/besfs5	1800	BESIII生产数据，BESIII group数据，用户数据	users目录每人50GB使用空间，group目录各组不同	原始数据有磁带备份，其它无备份
/bes3fs	1900	BESIII 生产数据		原始数据有磁带备份，其它无备份
/bes3fs	1900	BESIII 生产数据		原始数据有磁带备份，其它无备份
/besfs3	2600	BESIII 生产数据		原始数据有磁带备份，其它无备份
/besfs4	2500	BESIII 生产数据		原始数据有磁带备份，其它无备份

LUSTRE存储系统（2）

/publicfs	3000	ATLAS,CMS,LHCB,UCAS分池共享数据盘	每人5TB使用空间，30万文件数	各组资源完全隔离使用，无备份
/sharefs	1500	Alicpt,BES,HBKG,HEPS,MBH分池共享数据盘		各组资源完全隔离使用，无备份
/scratchfs	572	用户临时文件	每人500GB使用空间，20万文件数	无备份
/workfs2	22	用户个人文件（推荐保存重要文件或结果）	每人5GB使用空间，5万文件数	全盘备份
/cefs	2500	CEPC 实验数据，用户数据		无备份
/junofs	3100	JUNO 实验数据	每人500GB使用空间，30万文件数	无备份
/dybfs	2500	DYB 实验数据，用户数据	每人1TB使用空间，30万文件数	原始数据有磁带备份，其它无备份
/dybfs2	1400	DYB 实验数据，用户数据	每人1TB使用空间，30万文件数	原始数据有磁带备份，其它无备份
/gecamfs	1500	GECAM 实验数据		无备份
/hxmtfs	1300	HXMT 实验数据		无备份
/hpcfs	1800	Slurm集群GPU应用数据		无备份
/lhaasofs	610	LHAASO 用户数据	每人200GB使用空间，50万文件数	无备份

LUSTRE常用操作（1）

Lustre存储系统有一系列特殊的操作命令（实操练习）

- 查看用户资源配额

```
$ lfs quota -u zhangsan -h /publicfsDisk
# 命令输出如下:  (已用空间) (软空间配额) (硬空间配额) (已存文件数) (软文件数配额)
quotas for user zhangsan (uid XXXX): Filesystem kbytes quota limit grace files quota limit grace
/publicfs 3.3G 5G 5G - 232010 300000 300100
```

- Project空间配额（只适用于/lhaasofs/user与/besfs5）
  - 限制指定目录空间使用
  - 运行命令查看目录配额使用情况

```
# lfs quota -p 1097 -h /besfs5
    Filesystem    used   quota   limit   grace   files   quota   limit   grace
       /besfs5  1.362T*    50G     50G       -  293249       0       0       -
```

LUSTRE常用操作（2）

- 设置目录的访问控制（ACL）
  - 设置访问控制

```
$ setfacl -m user:wanglu:rwx /besfs4/wanglutest
```

  - 查看acl权限

```
$ getfacl /besfs4/wanglutest
```

  - 删除acl权限

```
$ setfacl -x user:wanglu /besfs4/wanglutest
```

  - 恢复linux原来的权限设置

```
$ setfacl -b /besfs4/wanglutest
```

注意：

1. 只有目录的属主可以操作目录的ACL权限
2. 设置过ACL权限后，ls -l 目录会多一个”+”，此时，Linux原来的permission规则会失效
3. 可以对一个组添加ACL setfacl -m group:xxx /xxx/xxx

EOS存储系统

EOS是CERN开发的面向EB级的磁盘文件存储系统，目前容量50PB

- **目前提供LHAASO、HXMT、JUNO等实验的海量数据存储服务。**

实例名	挂载点	实例服务器地址	总空间	用途
LHAASO实验	/eos	eos01.ihep.ac.cn	14.18 PB	LHAASO本地实验数据
LHAASO稻城	/eos/daocheng	lhmteos01.ihep.ac.cn	2.54 PB	LHAASO稻城快速重建
HXMT实验	/mnt/hxmt	hxmteos01.ihep.ac.cn	806.22 TB	HXMT实验数据
JUNO实验	/eos/juno	junoeos01.ihep.ac.cn	1.33 PB	JUNO实验

EOS存储常用操作（1）

EOS存储系统有Fuse和Xrootd两种访问方式

- FUSE方式在登录节点交互使用
- 推荐使用 xrootd方式访问实验数据

```
$ xrdfs root://eos01.ihep.ac.cn  ls /eos/user/file.txt
```

```
$ eos root://eos01.ihep.ac.cn  ls /eos/user/file.txt
```

- 使用EOS前需设置实例服务器地址环境变量：EOS_MGM_URL
  - 在北京集群查看：

```
$ echo $EOS_MGM_URL
root://eos01.ihep.ac.cn
```

  - 如果不存在，可自行设置：

```
$ export EOS_MGM_URL=root://eos01.ihep.ac.cn
```

EOS存储常用操作（1）

- 访问方式(xrootd方式)
  - 作业程序（C++）：使用xrootd方式打开.root格式文件

```
TFile *filein = TFile::Open(“root://eos01.ihep.ac.cn//eos_absolute_path_filein_name.root”)
 或
TFile *fileout = TFile::Open(“root://eos01.ihep.ac.cn//eos_absolute_path_fileout_name.root”)
```

  - 非root格式文件参考：
    - http://afsapply.ihep.ac.cn/cchelp/zh/local-cluster/storage/EOS/

注意：打开的文件使用后，应使用TFile:Close()及时关闭

EOS存储常用操作（2）

- 查看资源配额情况

```
$ eos quota /eos/user/z/zhangsan
```

- 输出

```
$ eos quota /eos/user/z/zhangsan
By user ...
# ____________________________________________________________________________________________________
# ==> Quota Node: /eos/user/z/zhangsan/
# ____________________________________________________________________________________________________
user       used bytes logi bytes used files aval bytes aval logib aval files filled[%]  vol-status
ino-status
zhangsan  800 GB  800 GB  10 k-   1.00 TB    200GB    25M -  9.08       ok         ok
        （已用空间）          （已使用文件数）（空间配额）（文件数配额）
By group ...
# ____________________________________________________________________________________________________
# ==> Quota Node: /eos/user/z/zhangsan/
# ____________________________________________________________________________________________________
# ....................................................................................................
group      used bytes logi bytes used files aval bytes aval logib aval files filled[%]  vol-status
ino-status
u07        800 GB  800 GB  10 k-   0 B        0 B        0 -        100.00     ignored    ignored
```

EOS存储常用操作（3）

- 查看回收站文件

```
$ eos recycle ls
```

- 清空回收站中的文件

```
$ eos recycle purge
```

- 恢复回收站中的某个文件

```
$ eos recycle restore 0000000008b0f7bf
```

注意：目前/eos回收站中的文件只保留3天时间。

CVMFS存储系统

- 使用CVMFS文件系统提供可伸缩、可靠和低维护的软件分发服务
- 13个软件卷，为不同实验作业提供软件环境
- 1个公共卷，提供计算平台上用户的公共软件存储
- CVMFS文件系统对用户只读
- 如果需要安装应用软件，需要联系计算中心安装

路径	用途
/cvmfs/bes.ihep.ac.cn	提供bes所需的软件库
/cvmfs/bes3.ihep.ac.cn	提供bes3所需的软件库
/cvmfs/cepc.ihep.ac.cn	提供cepc实验所需的数据分析软件
/cvmfs/exo.ihep.ac.cn	提供exo实验所需的数据分析软件
/cvmfs/dcomputing.ihep.ac.cn	提供分布式计算实验所需的数据分析软件
/cvmfs/gluex.ihep.ac.cn	提供gluex实验所需的数据数据分析软件
/cvmfs/hxmt.ihep.ac.cn	提供HXMT实验所需的数据数据分析软件
/cvmfs/heps_ap.ihep.ac.cn	提供HEPS实验所需的数据数据分析软件
/cvmfs/juno.ihep.ac.cn	提供juno实验所需的数据分析软件
/cvmfs/lhaaso.ihep.ac.cn	提供lhaaso实验所需的数据分析软件
/cvmfs/lqcd.ihep.ac.cn	提供lqcd实验所需的数据分析软件
/cvmfs/mlgpu.ihep.ac.cn	提供gpu机器学习相关软件
/cvmfs/raq.ihep.ac.cn	提供raq实验所需的数据分析软件

路径	
/cvmfs/common.ihep.ac.cn	存储系统组公共脚本和软件

数据备份与恢复

- 如需恢复数据：发送具体的目录名、文件名及恢复的日期到 helpdesk@ihep.ac.cn

应用	目录	备份策略
BES	/home/bes	每天一次备份，可恢复一个月之内的数据
BSRF	/home/bsrf	每天一次备份，可恢复两周之内的数据
LHC	/home/lhc	每天一次备份，可恢复两周之内的数据
CC	/home/cc	每天一次备份，可恢复两周之内的数据
ATLAS	/publicfs/atlas/codesbackup	每天一次备份，可恢复两周之内的数据
ATLAS	/afs/ihep.ac.cn/soft/atlas	每周一次备份，可恢复一个月之内的数据
CMS	/afs/ihep.ac.cn/soft/CMS	每周一次备份，可恢复一个月之内的数据
公共目录	/afs/ihep.ac.cn/users	每天一次备份，可恢复两周之内的数据
公共目录	/workfs2	每天一次备份，可恢复一个月之内的数据

存储目录组织

- 针对使用文件系统的几点建议
  - **单一目录下不要有过多(几万以上)数据或脚本等文件**，应按照一定规律创建子目录，将文件放在子目录下，**单目录文件数量控制在3000以内**。
  - **作业中避免使用ls *或rm *等含通配符的操作；**如果只需查看文件名信息，可以使用/bin/ls代替ls命令，可以加快速度；如果需要查看/eos目录，则使用“eos ls 目录绝对路径”，速度会更快。
  - 可以直接将**生成的数据文件存放在 eos 上，建议使用xrootd方式读写文件**

主要内容

- IHEP计算平台简介
- 账号及登录
- 文件存储系统
- **作业系统**
- 常见FAQ
- 典型使用示例

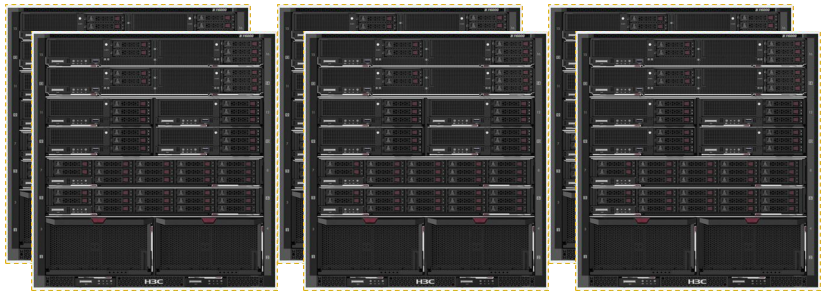
计算作业系统

HTC集群

HTC集群支持高通量计算作业
(High Throughput Computing)

采用HTCondor作为负载管理系统

作业大多是单核或单节点作业



HPC集群

HPC集群支持高性能计算作业
(High Performance Computing)

采用Slurm作为负载管理系统

作业大多为多核并行、GPU作业

HTCondor作业

- **高能所计算集群针对HTC作业开发了一个作业前端工具hepJob，进行了HTC集群的优化封装，推荐使用hepJob命令**
- 环境设置作业准备
  - 加载HepJob环境 （已添加到~/.profile）

```
# bash用户
$ export PATH=/afs/ihep.ac.cn/soft/common/sysgroup/hep_job/bin:$PATH
# tcsh用户
$ setenv PATH /afs/ihep.ac.cn/soft/common/sysgroup/hep_job/bin:$PATH
```

  - 作业脚本需有可执行权限

```
# 查看作业脚本是否有可执行权限
$ /bin/ls -l job.sh
-rw-r--r-- 1 jiangxw u07 85 Aug 29 18:23 job.sh

# 赋予作业脚本可执行权限
$ /bin/chmod +x job.sh
```

HTCondor作业常用命令（1）

- 作业提交

```
$ hep_sub job.sh
```

- 作业查询

```
$ hep_q -u <username>
```

- 作业删除

```
$ hep_rm 3745232 3745233.0
```

- 挂起作业释放

```
$ hep_release 3745233.0
```

- 修改作业需求

```
$ hep_edit 3745233.0 -m 8000
```

HTCondor作业常用命令（2）

- 按组查询作业时长限制

```
$ hep_clus -g juno --walltime
```

实验	短作业(short)时长限制(小时)	普通作业时长限制(小时)	mid作业时长限制(小时)及资源使用量限制(百分比)
BES	<0.5	<40	<100:10%
JUNO	<0.5	<20	<100:10%
DYW	<0.5	<10	<100:10%
CEPC	<0.5	<10	<100:10%
ATLAS	<0.5	<10	<100:10%
CMS	<0.5	<10	<100:10%
HXMT	<0.5	<14	<100:10%
GECAM	<0.5	<24	<100:10%
LHCb	<0.5	<100	
LHAASO	<0.5	<15	<100:10%

注意，未设置mid作业的实验，默认提交mid作业时，资源使用量限制为1.

在作业中获取作业信息

- 获取作业ID

```
#!/bin/bash
JobId=$_CONDOR_IHEP_JOB_ID
```

- 获取运行节点

```
#!/bin/bash
ExecWorkNode=$_CONDOR_IHEP_REMOTE_HOST
```

- 获取作业提交时间

```
#!/bin/bash
SubmissionTime=$_CONDOR_IHEP_SUBMISSION_TIME
```

提交大内存作业

- 默认情况下，作业分配资源不会考虑内存大小（随机分配）
- **如果作业有大内存特殊需要，使用-mem参数指定内存：**

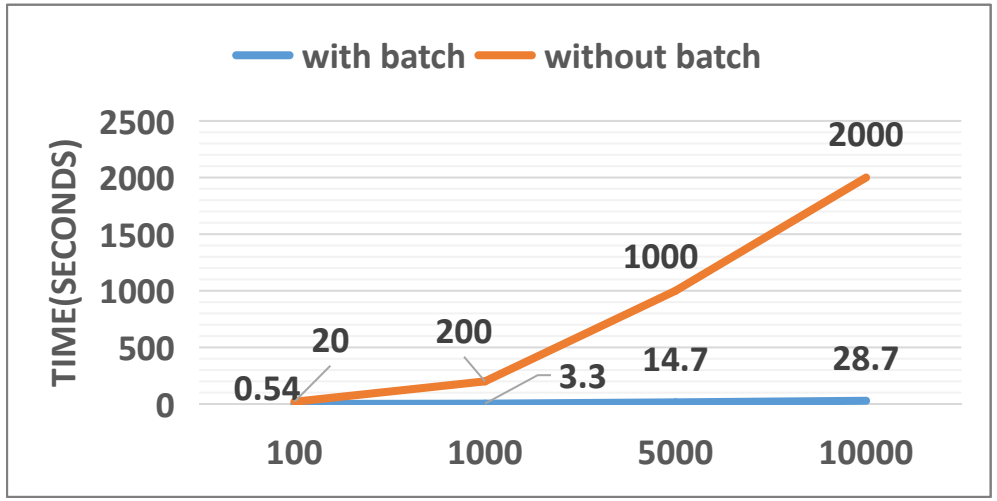
```
$ hep_sub -mem 3000 job.sh
```

- 其中，-mem参数值单位为MB，实例中3000表示3GB内存
- 注意，大内存节点相对较少，尽量控制单作业的内存使用

批量作业提交(1)

- 作业脚本名格式相同，可以批量提交作业，参数-n指定作业数量
- 示例1：

```
real_job_20191204_1.sh
real_job_20191204_2.sh
real_job_20191204_3.sh
real_job_20191204_4.sh
real_job_20191204_5.sh
real_job_20191204_6.sh
real_job_20191204_7.sh
real_job_20191204_8.sh
real_job_20191204_9.sh
```



- 作业脚本名格式相同：real_job_20191204_*.sh，且关键字符是从0开始递增的数字，提交这些作业只需要运行：

```
$ hep_sub real_job_20191204_"%{ProcId}".sh -n 11
```

批量作业提交(2)

- 以宏模式批量提交作业
- 将多个作业的提交过程封装在一个单独的命令中，从而简化批量作业的管理和提交过程

- 示例2：

1.已有脚本

```
real_job_20191201.sh
real_job_20191202.sh
real_job_20191203.sh
real_job_20191204.sh
real_job_20191205.sh
real_job_20191206.sh
real_job_20191207.sh
...
real_job_20191230.sh
real_job_20191231.sh
```

2.额外准备脚本

```
#!/bin/bash

# get procid from command line
procid=$1

# map 0,1,2,...,30 to 1,2,3,...,31
sub_name_number=`expr $procid + 1`

# format 1,2,3,...,31 to 01,02,03,...,31
sub_name=`printf "%02d\n" $sub_name_number`

# run the real job script by the formatted file name
bash real_job_201912"${sub_name}".sh
```

- 3.提交作业

```
$ hep_sub real_job_parent.sh -argu "%{ProcId}" -n 31
```

HPC作业：SLURM

- 申请权限
  - 提交作业前，**用户需申请AFS账号**：申请页面
  - 账号申请成功后，上述组别的成员请分别**向各组计算负责人发送邮件申请slurm集群使用授权**。未经授权，报错如下：

```
sbatch: error: Batch job submission failed: Invalid account or account/partition
combination specified
```

  - 经计算负责人与集群管理员授权后，用户方可提交作业至slurm集群中运行。

SLURM作业常用命令

- 作业提交

```
$ sbatch slurm_sample_script_1.sh
```

- 作业查询

```
$ squeue 或 sacct -u <user_name>
```

- 作业删除

```
$ scancel <job_id>
```

- 查看集群状态

```
$ sinfo
```

SLURM作业准备

- 样例：/cvmfs/slurm.ihep.ac.cn/slurm_sample_script

GPU作业

```
#! /bin/bash
######## Part 1 ######
#SBATCH --partition=gpu
#SBATCH --qos=normal
#SBATCH --account=lqcd
#SBATCH --job-name=gres_test
#SBATCH --output=job-%j.out
#SBATCH --ntasks=2
#SBATCH --mem-per-cpu=2048
#SBATCH --gres=gpu:v100:2
######## Part 2 ######
echo "hello world!"
```

队列信息

资源信息

GPU作业

CPU作业

```
#! /bin/bash
######## Part 1 ######
#SBATCH --partition=mbh
#SBATCH --qos=regular
#SBATCH --account=mbh
#SBATCH --job-name=mbh_test
#SBATCH --output=job-%j.out
#SBATCH --ntasks=20
#SBATCH --mem-per-cpu=2048
######## Part 2 ######
echo "hello world!"
```

作业脚本通常由两部分组成：

- 作业运行参数，以**#SBATCH**为开头，指明作业运行的参数
- 作业运行内容，通常为可执行程序，如可执行脚本、**MPI**程序等

作业运行参数

- **partition**:作业所属队列
- **account**:作业所属账户组
- **qos**:作业的服务质量(优先级)

为必选项，必须指明

SLURM-GPU作业

- 应用组
  - lqcd,gpupwa,junogpu,mlgpu,higgs,bldesign
- 各组的资源分区（partition）、作业队列（QOS）、计算节点

partition（节点分区）	qos（队列）	group	资源限制	节点资源
lgpu	long	lqcd	QOS long - 作业运行时间不超过30天 - 每组作业数量（运行+排队）不超过64个 - 每个作业每CPU核最大可使用40GB内存	- 1个节点，每个节点384GB 内存 - 共8张GPU卡，36个CPU核
gpu	normal, debug	lqcd, gpupwa, junogpu, mlgpu,higgs	QOS normal - 每个作业运行时间不超过48小时 - 每组作业数量（运行+排队）不超过512个，每组可使用的GPU卡数量不超过128张 - 每个用户作业数量（运行＋排队）不超过96个，每用户可使用的GPU卡数量不超过64张 - 每个作业每CPU核最大可使用40GB内存 QOS debug - 作业运行时间不超过15分钟 - 每组作业数量（运行＋排队）不超过256个，每组可使用的GPU卡不超过64张 - 每个用户作业数量（运行＋排队）不超过24个，每个用户可使用的GPU卡不超过16张 - 每个作业每CPU核最大可使用40GB内存 - QOS debug 优先级高于 QOS normal优先级	- 23个节点，每个节点384GB 内存 - 共182张GPU卡, 840个CPU 核

SLURM-CPU作业

- 应用组
  - mbh,bio,cac,nano,heps,cepcmpi,alicpt,bldesign,raq
- 各组的资源分区（partition）、作业队列（QOS）、计算节点

partition(节点分区)	QOS（作业队列）	account / group（组别）	worker nodes（计算节点）
mbh,mbh16	regular	mbh	16个节点，共256个CPU核
cac	regular	cac	8个节点，共384个CPU核
nano	regular	nano	7个节点，共336个CPU核
bioq	regular	bio	16个节点，共256个CPU核
biofastq	regular	bio	12个节点，共288个CPU核
heps	regular,advanced	heps	34个节点，共1224个CPU核
hepsdebug	hepsdebug	heps	1个节点，共36个CPU核
cepcmpi	regular	cepcmpi	36个节点，共1696个CPU核
ali	regular	alicpt	16个节点，共576个CPU核
bldesign	blregular	bldesign	3个节点，共108个CPU核
raq	regular	raq	12个节点，共672个CPU核

SLURM-CPU作业

- 队列资源使用限制

QOS	作业最大运行时间	可提交的最大作业数量	优先级
regular	60天	每个用户4000个，每个组8000个	低
advanced	60天	-，-	高
hepsdebug	30分钟	每个用户10个，-	中
blregular	30天	每个用户200个，每个组1000个	低

主要内容

- IHEP计算平台简介
- 账号及登录
- 文件存储系统
- 作业系统
- **常见FAQ**
- 典型使用示例

常见FAQ（实操手册）

- 高能所计算环境使用手册：
  http://afsapply.ihep.ac.cn/cchelp/zh/
- 如有网格计算、分布式计算、虚拟云计算等方面的需求，也可参考该手册


Thanks!
• 更多深度技术和知识请关注后续课程
• 实际操作部分
• https://note.ihep.ac.cn/s/ao4ywfO76