A Lightweight Submission Frontend Toolkit - HepJob

Jiang Xiaowei

On behalf of System Group at IHEP

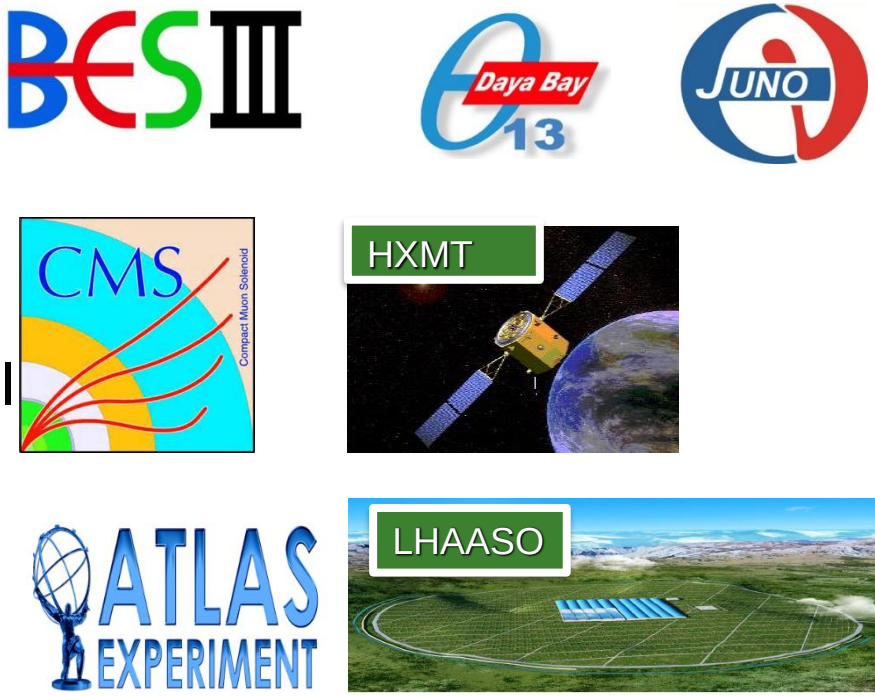
CHEP2019, 2019.11.04

Outlines

- Why to Implement HepJob
- Design & Structure
- Key Points
  - HepJob Commands
  - Routing job to the targeted destination
  - Pre-check for submission
  - Connect to SLURM through HepJob
- Test & Current Status

Background

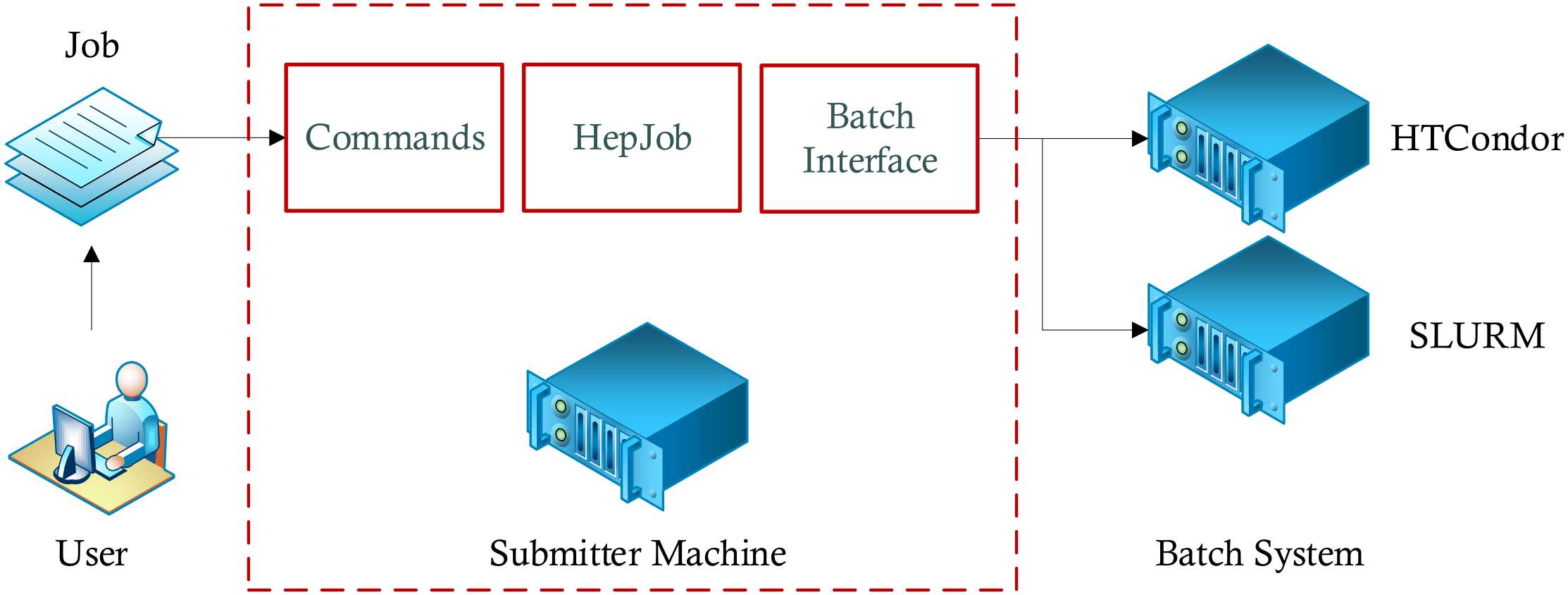
- Service Objects
  - 2813 users from >10 Experiments
- Seperated Computing Clusters
  - HTCondor for HTC
    - A main physical pool and an additional vm pool
    - 3 remote sites(chengdu, daocheng, ustc)
    - ~ 18,000 cpu cores in total
  - SLURM for HPC
    - 4000 CPU cores and 94 GPUs



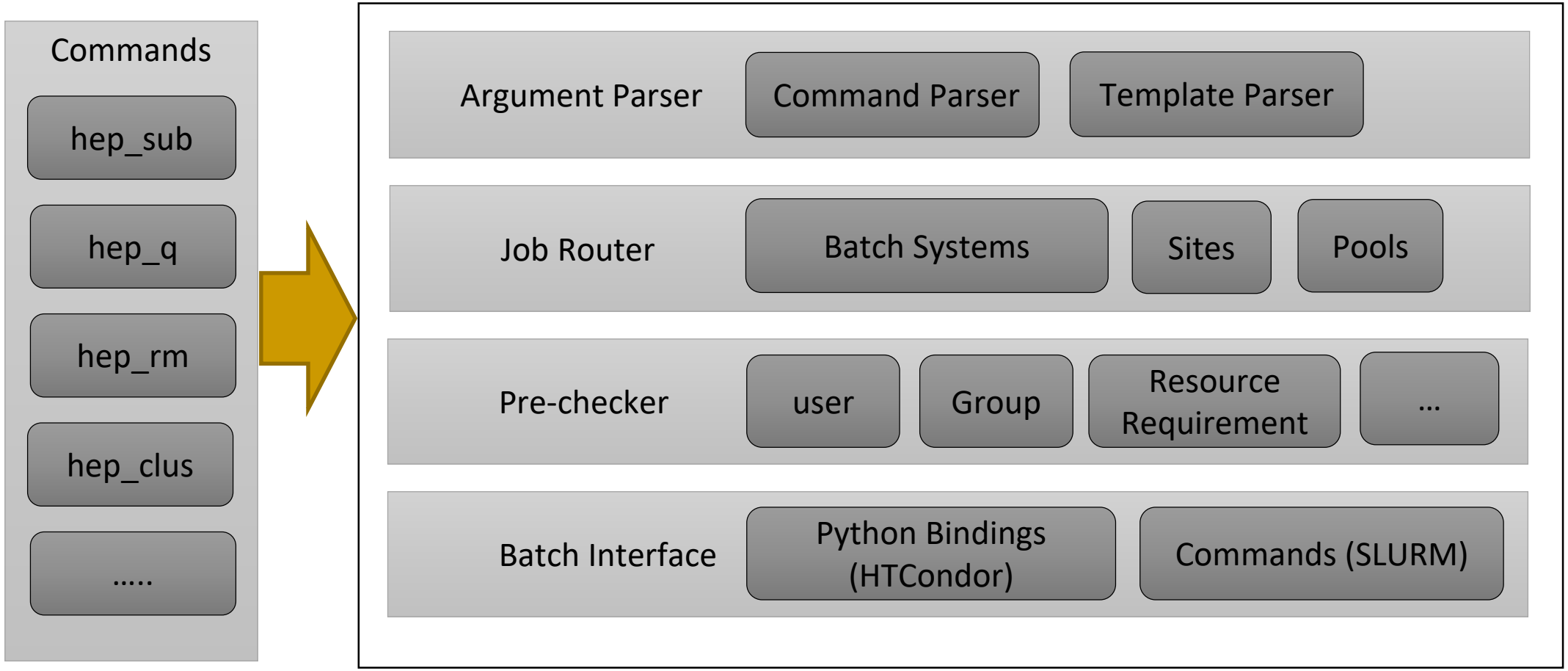
Motivation & Purpose

- Unify the submission entrance
  - different batch systems, separated pools and remote sites
- Pre-check submission conditions
  - username, group, resource
- Prevent some basic problems
  - file permission, file non-existence, …
- Simplify and localize the usage of computing cluster
  - Our users are used to the simple command-line usage of submission
  - Some specific situations of submission or interaction required

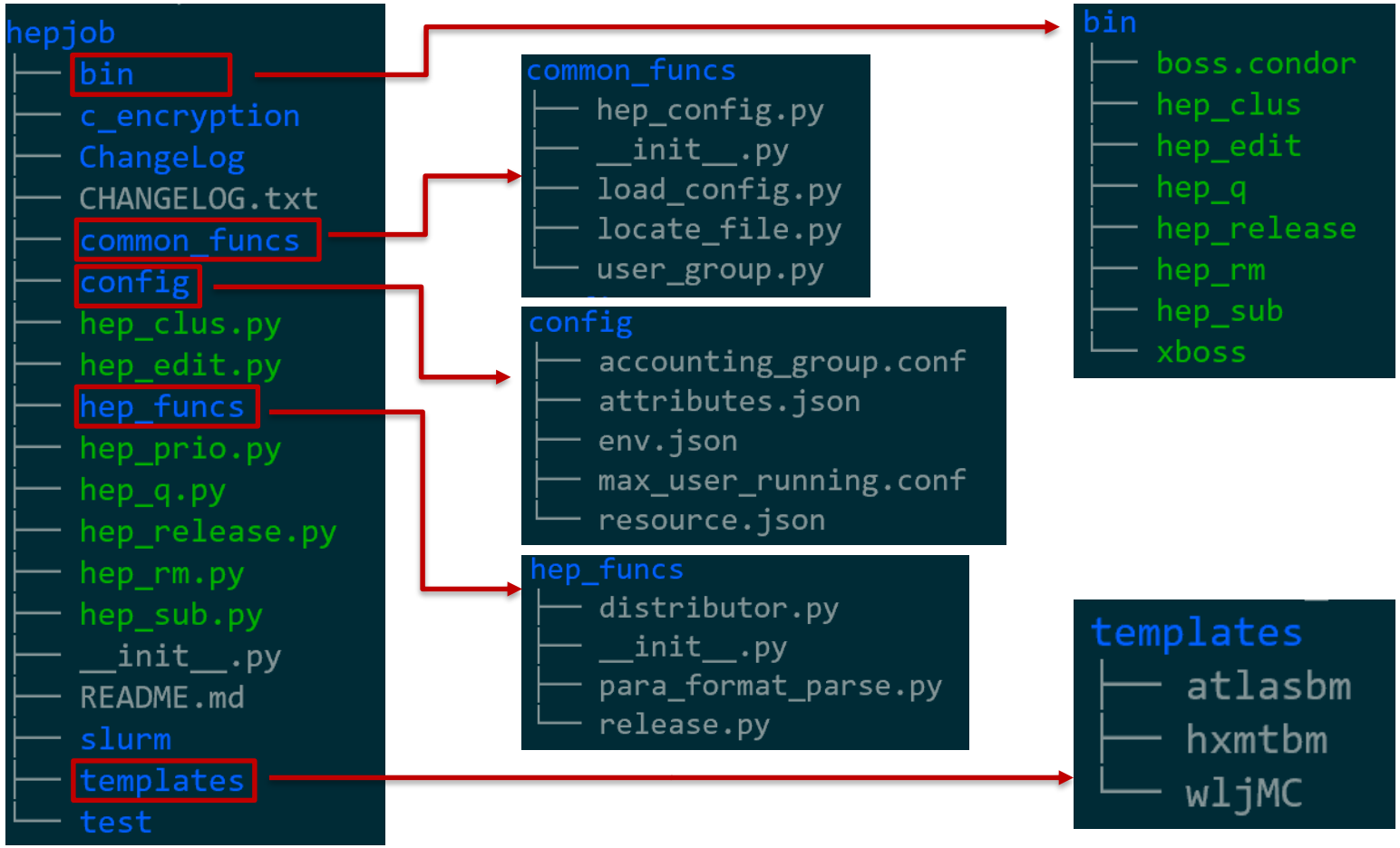
The role of HepJob



Structure of HepJob



Only deploy on the submitter machine



HepJob commands

- Commands
  - submit, query, remove,…
  - hep_sub, hep_q, hep_rm,
  - hep_clus, hep_edit, hep_release
  - …
- implemented with argparse module of Python
- Wrapping with C language

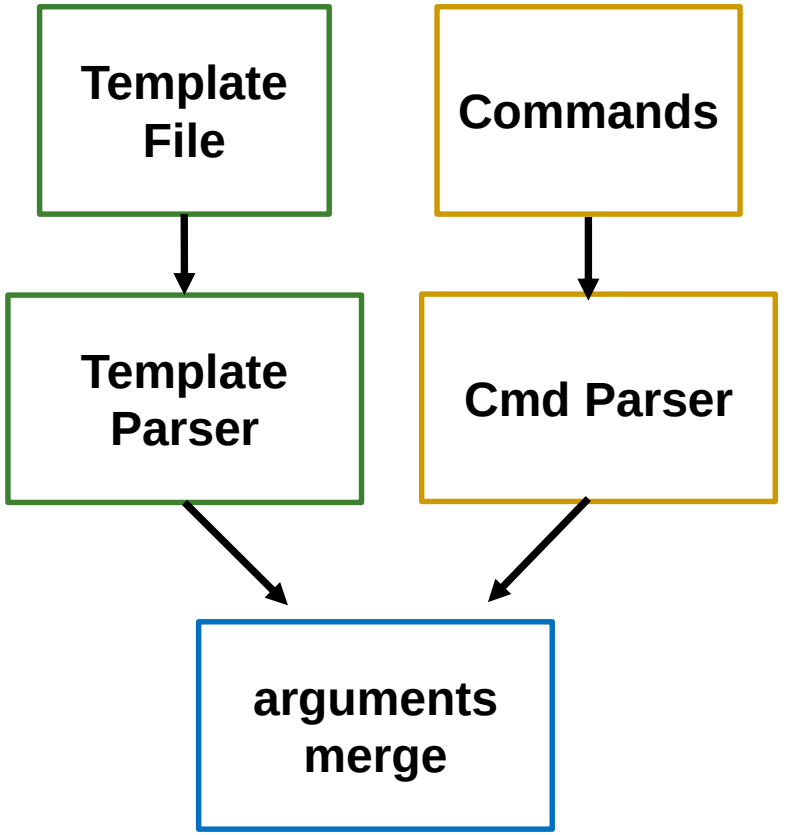
```
-bash-4.2$ hep_sub --help
usage: hep_sub [-h]
               [-g {u07,atlas,dyw,cms,physics,hxmt,offlinerun,simrun,higgs,juno,comet,cepcmpi}]
               [-p {virtual,physical}] [-u {vanilla,grid,docker}] [-o OUT]
               [-e ERROR] [-n NUMBER] [-os {SL5,SL6,SL7,ALL}]
               [-t {atlasbm,hxmtbm,wljMC}] [-prio PRIORITY]
               [-np NUMBERPROCESS] [-argu ARGUMENTS [ARGUMENTS ...]]
               [-dir DIRECTORY] [-mem MEMORY] [-quiet] [-part PARTITION]
               [-name NAME] [-slurm] [-site SITENAME] [-jf JOBFILE]
               [-tf TRANSFERFILE] [-wn WORKNODE] [-wt WALLTIME]
               [jobscript]

Submit job to the cluster of IHEP.

positional arguments:
  jobscript             set the job file

optional arguments:
  -h, --help            show this help message and exit
  -g {u07,atlas,dyw,cms,physics,hxmt,offlinerun,simrun,higgs,juno,comet,cepcmpi}, --group {u07,at
hxmt,offlinerun,simrun,higgs,juno,comet,cepcmpi}
                        write your groupname according to jobgroup.
  -p {virtual,physical}, --pool {virtual,physical}
                        set the pool you want submit jobs to
  -u {vanilla,grid,docker}, --universe {vanilla,grid,docker}
                        set the universe
  -o OUT, --out OUT     set the output file.
  -e ERROR, --error ERROR
                        set the error file
  -n NUMBER, --number NUMBER
                        set the number of jobs
  -os {SL5,SL6,SL7,ALL}, --OperatingSystem {SL5,SL6,SL7,ALL}
                        set the system version of resource you want.
  -t {atlasbm,hxmtbm,wljMC}, --template {atlasbm,hxmtbm,wljMC}
                        set the template of job submission you want.
  -prio PRIORITY, --priority PRIORITY
                        set the inner job priority of your own jobs.
  -np NUMBERPROCESS, --numberprocess NUMBERPROCESS
                        set the total cores required by your job.
  -argu ARGUMENTS [ARGUMENTS ...], --arguments ARGUMENTS [ARGUMENTS ...]
```

Templates for complex submission condition

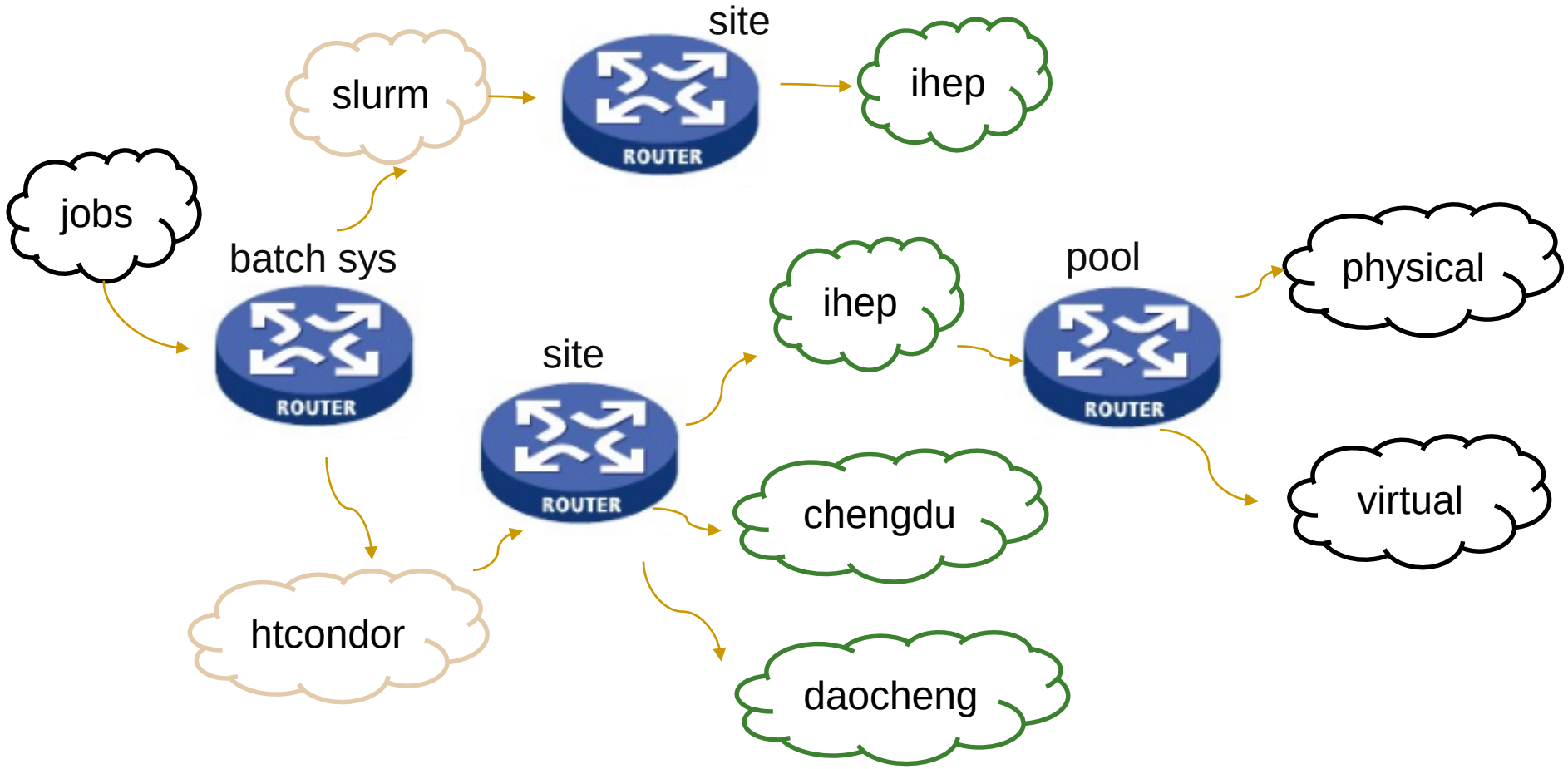
- Pre-define some specific requirements
- Simplify the submission command
- E.g., with the altlasbm template,

```
hep_sub job.sh -np 16 -mem 30000 -g atlas
```

```
hep_sub job.sh -t atlasbm
```



Routing Job to the targeted destination (1)



Routing Job to the targeted destination (2)

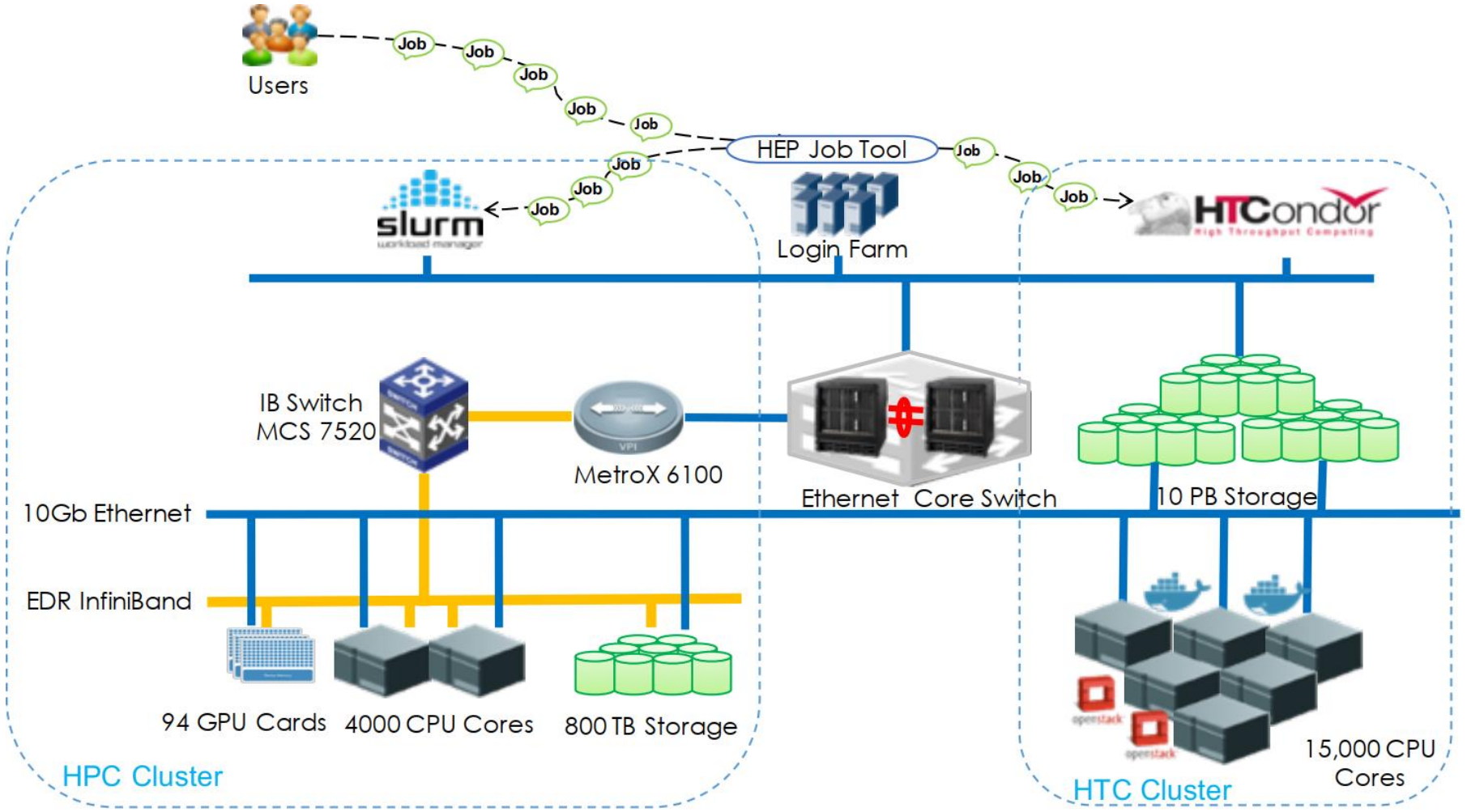
- Store the routing table in a json file
  - ../config/resource.json

```
{
        "condor_server":
        {
                "ihep":
                [
                        "physical":
                        [
                                "schedd": "job@xxxxxxx.ihep.ac.cn",
                                "cm": "xxxxxxxx.ihep.ac.cn"
                        ],
                        "virtual":
                        [
                                "schedd": "job@xxxxxxx.ihep.ac.cn",
                                "cm": "collector@vmxxxxxx.ihep.ac.cn"
                        ]

                ],
                "daocheng":
```

Pre-check for HTCondor

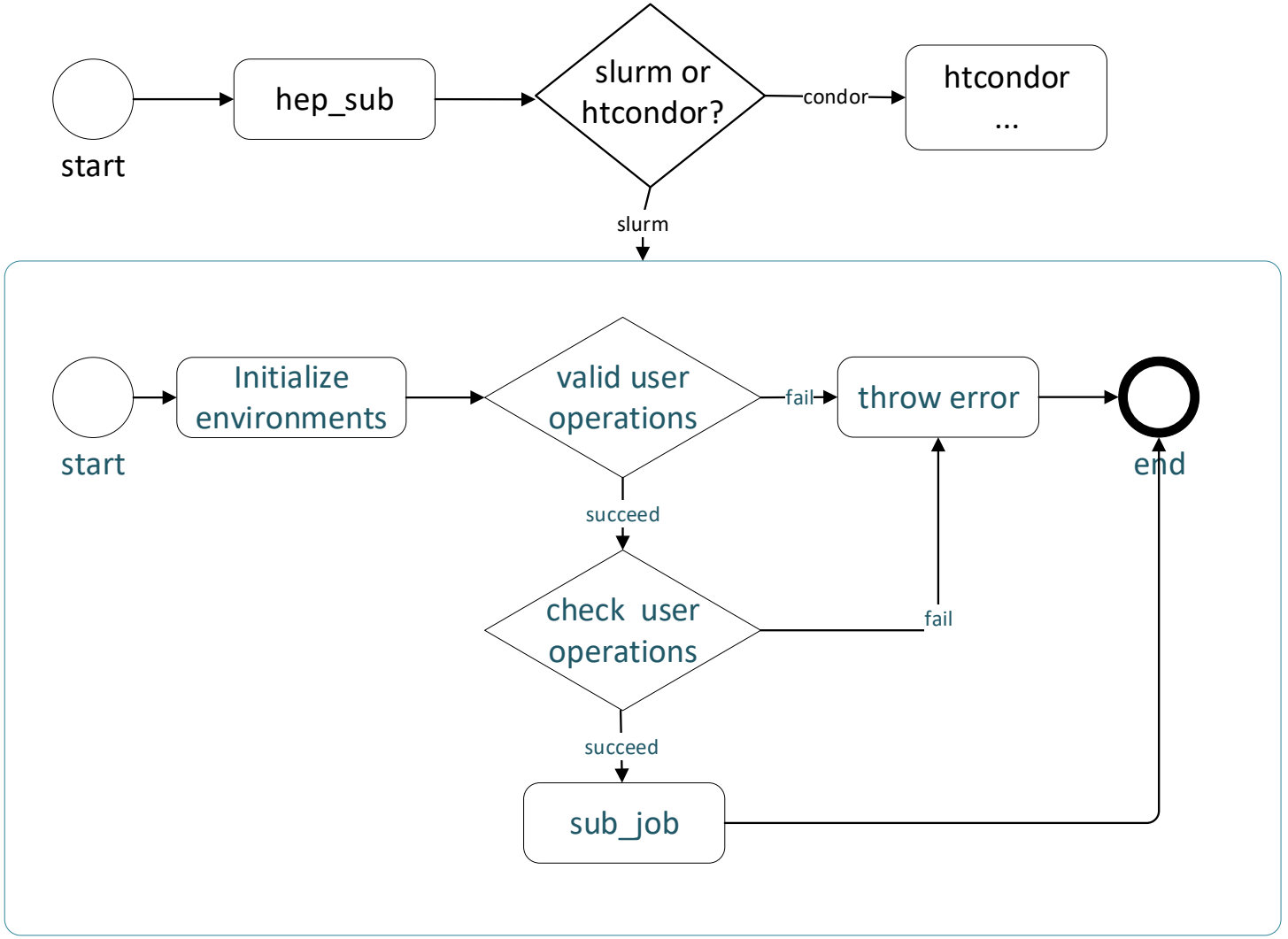
- Requirements Check:
  - accounting_group
  - max_memory
  - max_cpu
  - operating system
  - singularity
  - job universe
  - site
  - …
- Problem Check:
  - output/error file existence
  - job script file existence
  - job script executable permission
  - submission permission
  - job amount limit
  - correct environments
  - …

Connect to SLURM through HepJob



SLURM plugins for HepJob

```
slurm
├── slurm_clus_chk.py
├── slurm_clus.py
├── slurm_job.py
├── slurm_q_chk.py
├── slurm_q.py
├── slurm_rm_chk.py
├── slurm_rm.py
├── slurm_sub_chk.py
└── slurm_sub.py
```



Use Case

- HepJob helps user do a lot:
  - Normally, submission just need

```
-bash-4.1$ hep_sub job.sh
INFO: Set the primary group 'u07' as your job group.
1 job(s) submitted to cluster 49989114.
```

- Query job with hep_q:

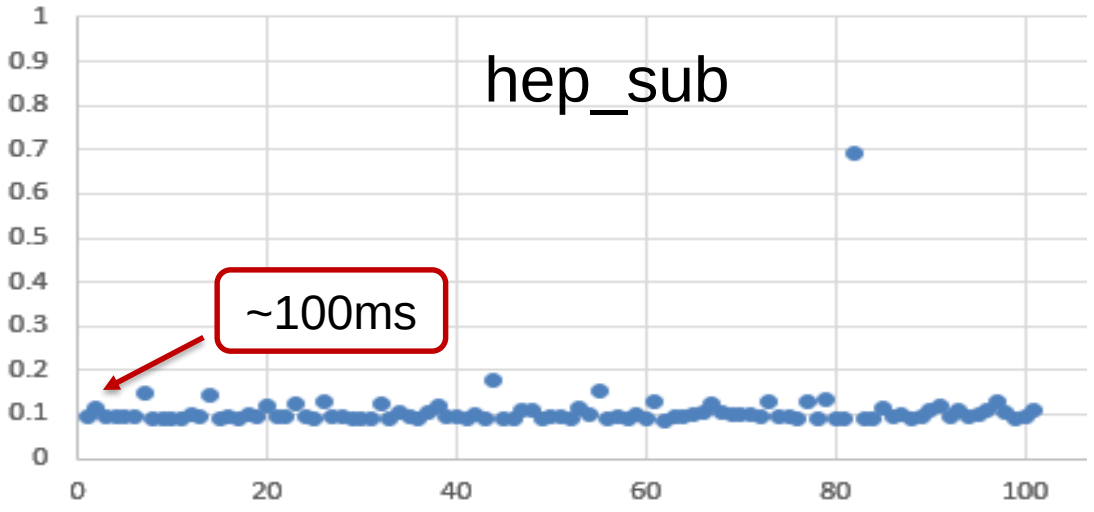
```
-bash-4.1$ hep_q -u

JOBID           OWNER           SUBMITTED       RUN_TIME        ST PRI SIZE CMD
47979976.0      jiangxw         10/15 14:28     0+00:00:00      I  0   0.0  job.sh
49989113.0      jiangxw         10/29 09:15     0+00:00:07      R  0   0.0  job.sh
49989114.0      jiangxw         10/29 09:15     0+00:00:00      I  0   0.0  job.sh

3 jobs; 0 completed, 0 removed, 2 idle, 1 running, 0 held, 0 suspended
```

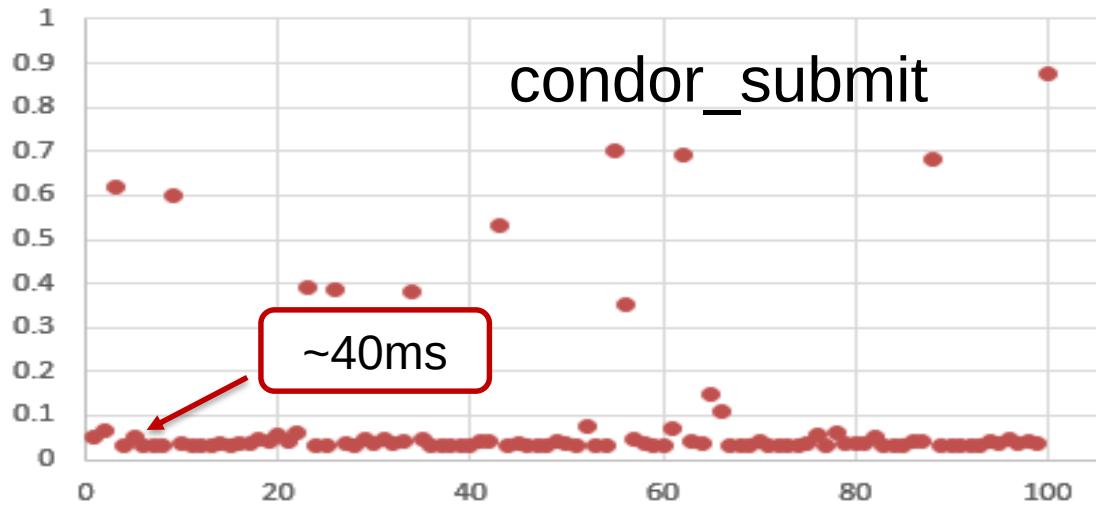
- Delete job with hep_rm:

```
-bash-4.1$ hep_rm 49989113.0
Job 49989113.0 has/have been deleted.
1 job(s) has/have been deleted.
-bash-4.1$ hep_q -u

JOBID           OWNER           SUBMITTED       RUN_TIME        ST PRI SIZE CMD
47979976.0      jiangxw         10/15 14:28     0+00:00:00      I  0   0.0  job.sh
49989114.0      jiangxw         10/29 09:15     0+00:00:04      R  0   0.0  job.sh

2 jobs; 0 completed, 0 removed, 1 idle, 1 running, 0 held, 0 suspended
```

Execution Time Consume Test

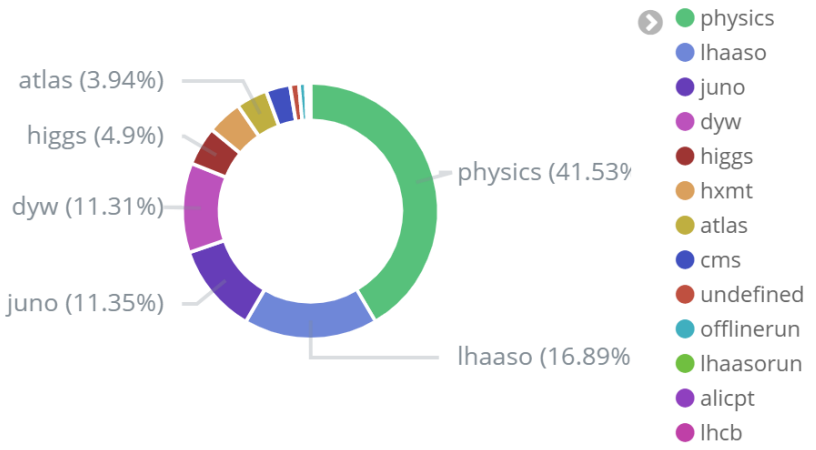
- A simple test to submit 100 jobs via hep_sub and condor_submit
- Time Consume:
  - hep_sub takes around 100 ms for each submission, while condor_submit takes around 40ms
  - The time consume of hep_sub for each submission can be acceptable





Current Application Status

- Serving since 2015
- Experiments supporting
  - BES, JUNO, DYW, CMS, LHAASO,..
- Batch system supporting
  - HTCondor, SLURM
- Pool supporting
  - physical, virtual, mpi
- Sites
  - IHEP, daocheng, chengdu,…



AcctGroupUser: Descending	Unique count of GlobalJobId
weilh	1,181,386
lhaasorec	910,181
hdpc	749,668
xuw	702,671
agilman	489,773
libh	421,681
huanggh	417,849
nikolaos	342,627
yaoyh	332,223
lizy	297,474

Summary

- HepJob is a lightweight submission frontend toolkit:
  - command-line usage
  - easy deployment
  - supporting HTCondor and SLURM
- At IHEP, using HepJob:
  - the multiple cluster entrances are unified
  - some basic problems can be pre-checked and avoided
  - the usage case is simplified and customized
- This toolkit is under preparing to open source

Thanks!