一种高能物理计算任务调度的前端管理与实现

姜晓巍

中国科学院高能物理研究所

2017年07月04日

目录

- 背景
- hepjob架构和关键技术
- hepjob应用
- 总结和展望

背景

BESIII (Beijing Spectrometer III at BEPCII)
100TB raw data/year *19

DYB (Daya Bay Reactor Neutrino Experiment)
200TB/year* 9 years

JUNO (Jiangmen Underground Neutrino Observatory)
2PB/year*30 year

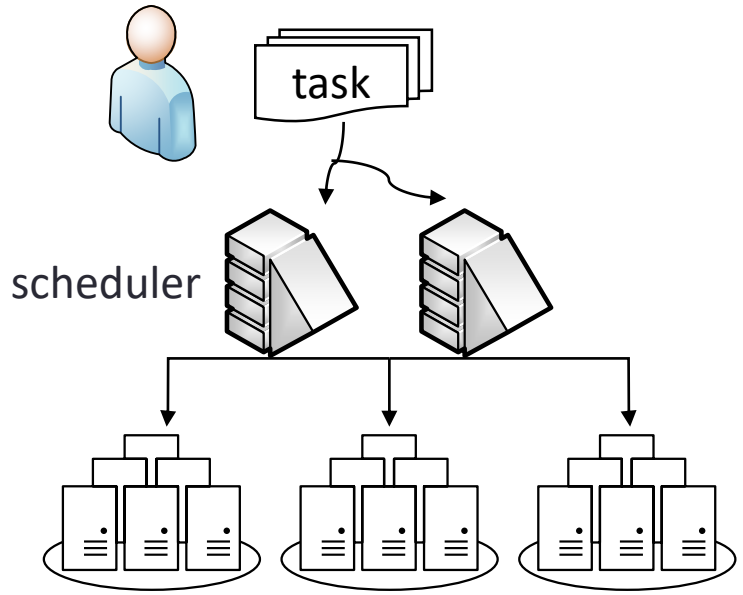
YBJ (Tibet-ASgamma ARGO-YBJ Experiments)

Large High Altitude Air Shower Observatory
1.2PB/year *10 year

Hard X-Ray Moderate Telescope

背景

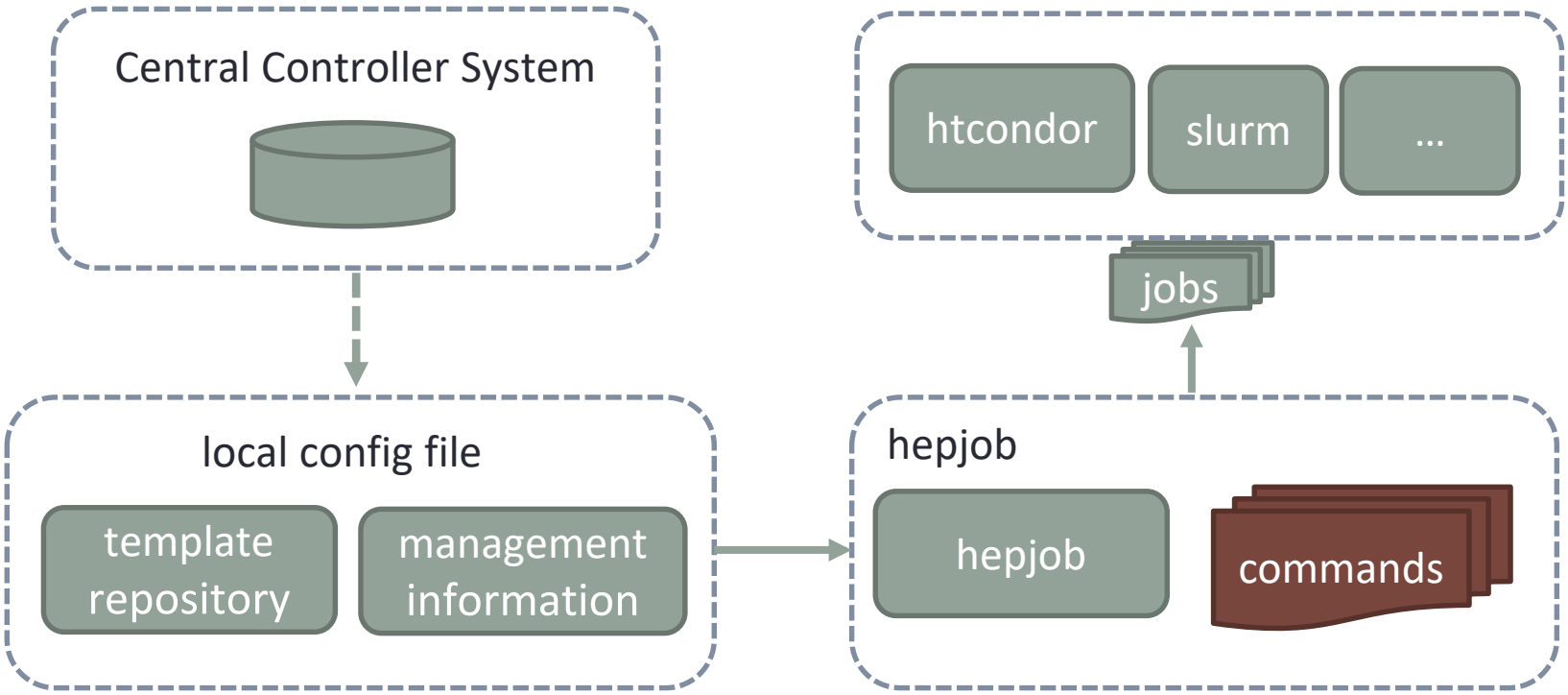
- 多调度系统共存
  - htcondor, 用于本地集群
  - slurm, 用于高性能集群
  - pbs, 部分实验使用
- 远程站点
  - buaa, chengdu, ustc
- 多集群共存
  - 物理集群
  - 虚拟集群
  - 高性能集群



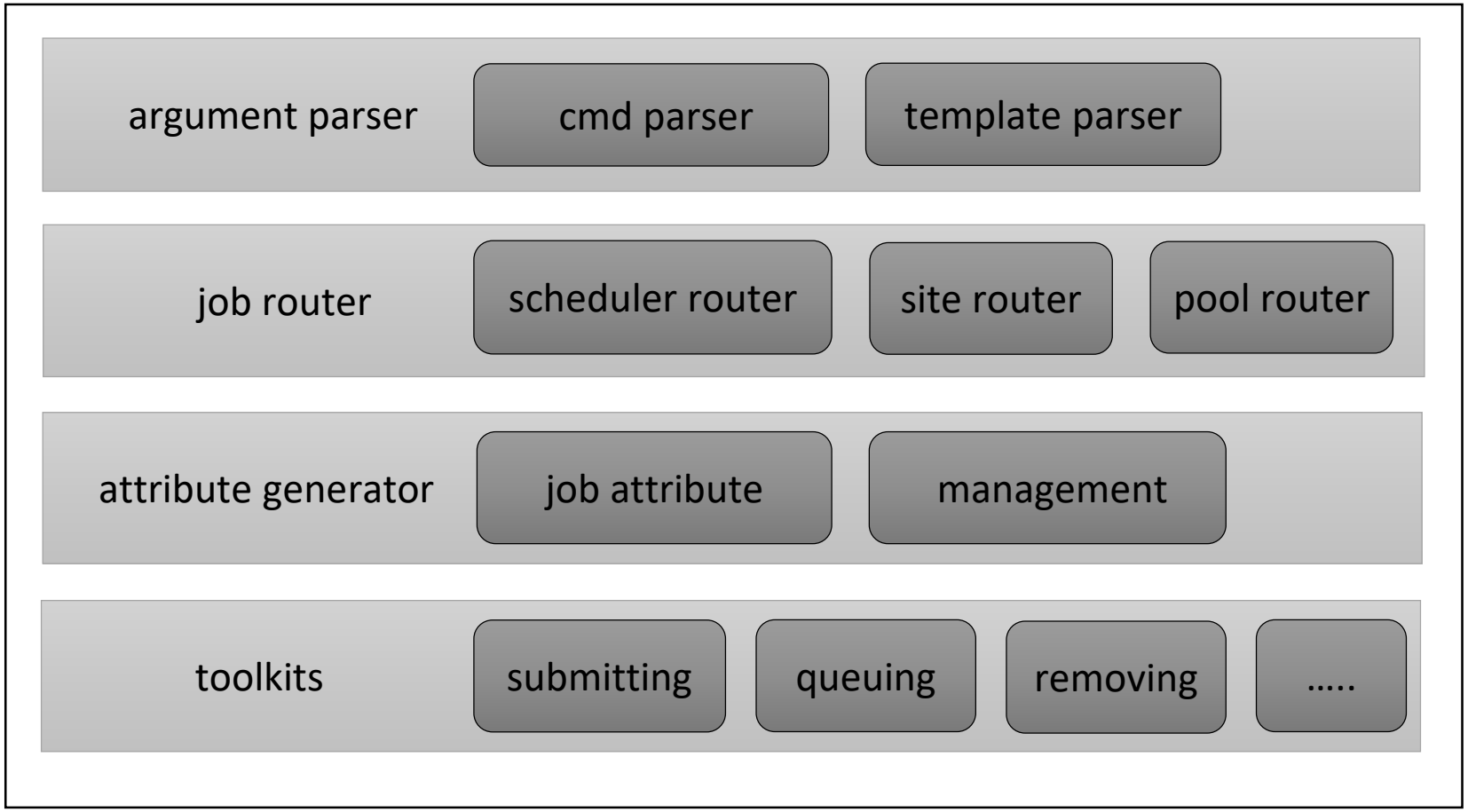
背景

- 调度控制分散
  - 多调度器、多站点、多集群没有统一调度入口，不易于协同调度
- 用户使用繁琐
  - 繁杂的功能使用户产生困扰
  - 用户花费大量学习使用
- 管理难于统一
  - 各调度系统的用户、组等信息管理各异
  - 作业管理和限制各异

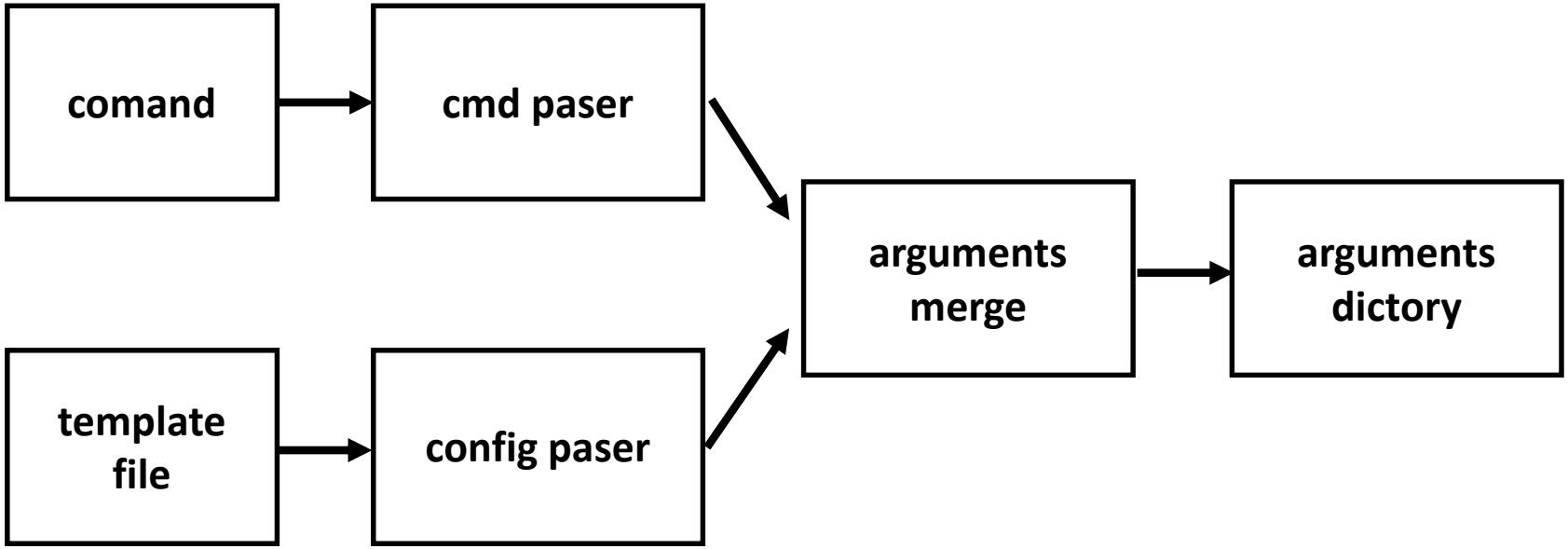
hepjob整体架构


Central Controller System
htcondor
slurm
...
jobs
local config file
template repository
management information
hepjob
hepjob
commands

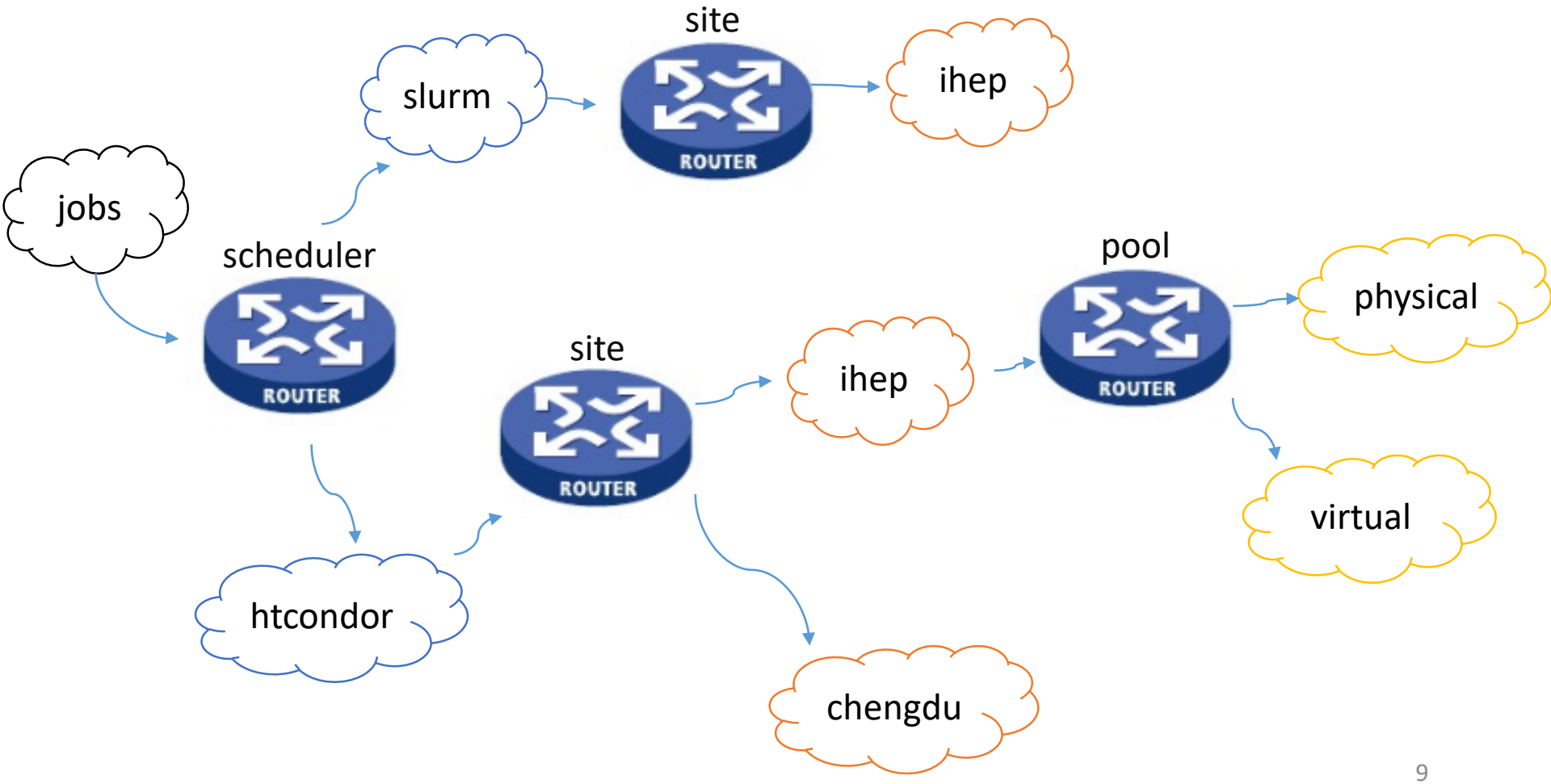
hepjob功能层次


argument parser
cmd parser
template parser
job router
scheduler router
site router
pool router
attribute generator
job attribute
management
toolkits
submitting
queuing
removing
.....

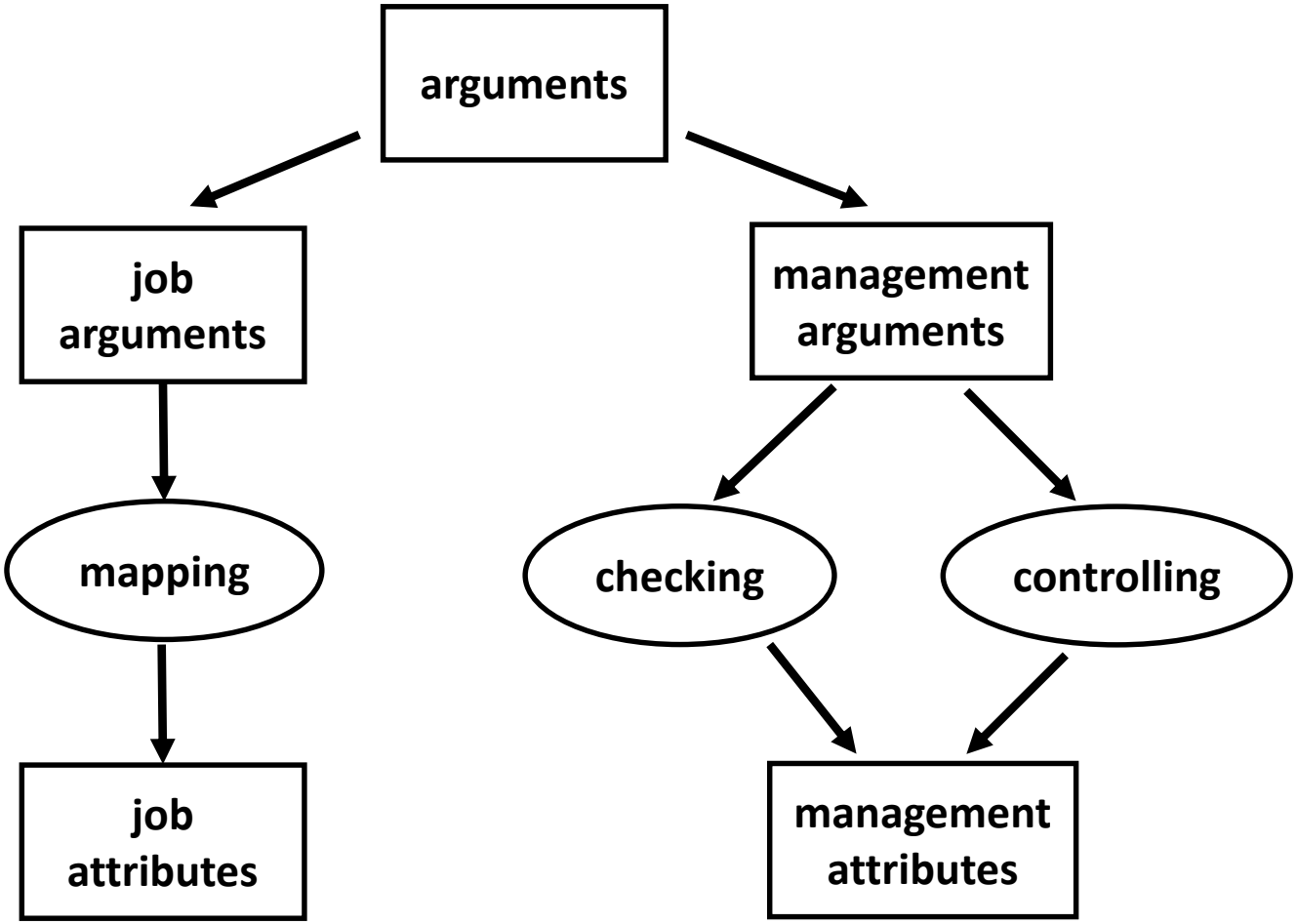
argument parser



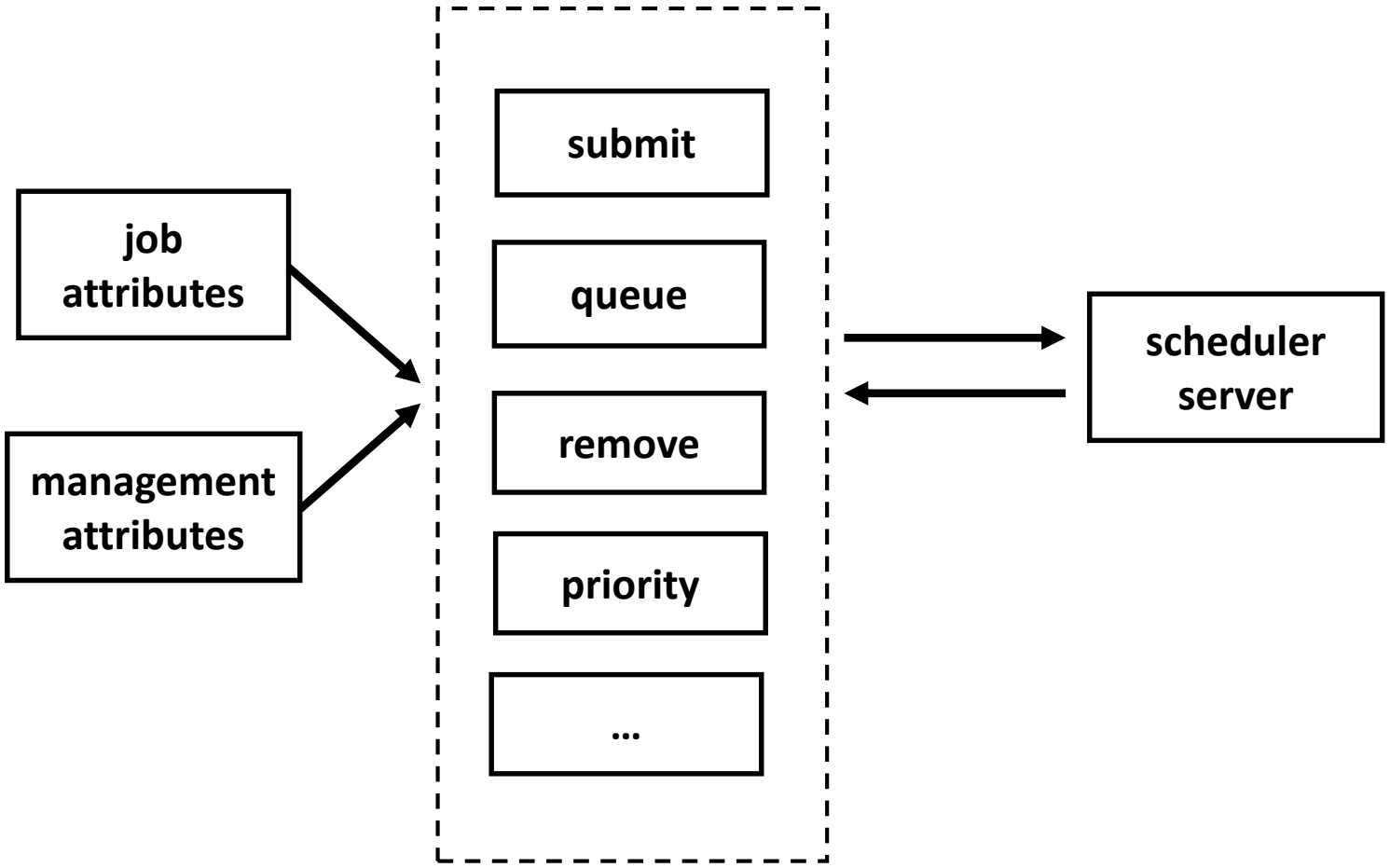
job router


site
slurm
ihep
jobs
scheduler
pool
physical
site
ihep
ROUTER
ROUTER
ROUTER
virtual
htcondor
chengdu

attribute generator



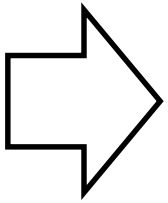
toolkits


job attributes
management attributes
submit
queue
remove
priority
…
scheduler server

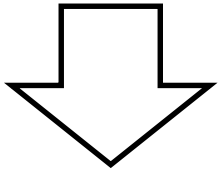
data flow

```
positional arguments:
  jobscript             set the job file

optional arguments:
  -h, --help            show this help message and exit
  -g {u07,hxmt,juno,atlas,dyw,offlinerun,cms,comet,higgs,physics,cepcmpi}
sics,cepcmpi}
                        write your groupname according to jobgroup.
  -p {virtual,local}, --pool {virtual,local}
                        set the pool you want submit jobs to
  -u {vanilla,grid}, --universe {vanilla,grid}
                        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 OPERATINGSYSTEM, --OperatingSystem OPERATINGSYSTEM
                        set the system version of resource you want.
  -t {atlasbm}, --template {atlasbm}
                        set the template of job submission you want.
  -prio PRIORITY, --priority PRIORITY
```

```
[request]
RequestCpus = 16
RequestMemory = 30000

[user info]
AccountingGroup = atlas
```



{'memory': 1500, 'arguments': ['3000'], 'partition': None, 'group': None, 'name': None, 'jobfile': None, 'universe': 'vanilla', 'numberprocess': None, 'number': 1, 'priority': None, 'jobscript': None, 'template': 'atlasbm', 'error': '.err', 'directory': None, 'OperatingSystem': 'SL6', 'pool': 'local', 'out': '.out'}



```
Name               OpSys      Arch   State     Activity LoadAv Mem   ActvtyTime

slot10@ams002.ihep LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:52
slot11@ams002.ihep LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:50
slot12@ams002.ihep LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:52
slot1@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:47
slot2@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:46
slot3@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:41
slot4@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:50
slot5@ams002.ihep. LINUX      X86_64 Owner     Idle      0.540 2001  0+02:36:03
slot6@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:49
slot7@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:40
slot8@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:36
slot9@ams002.ihep. LINUX      X86_64 Owner     Idle      0.000 2001  0+02:35:50

                     Machines Owner Claimed Unclaimed Matched Preempting

        X86_64/LINUX       12    12       0         0       0          0

               Total       12    12       0         0       0          0
```

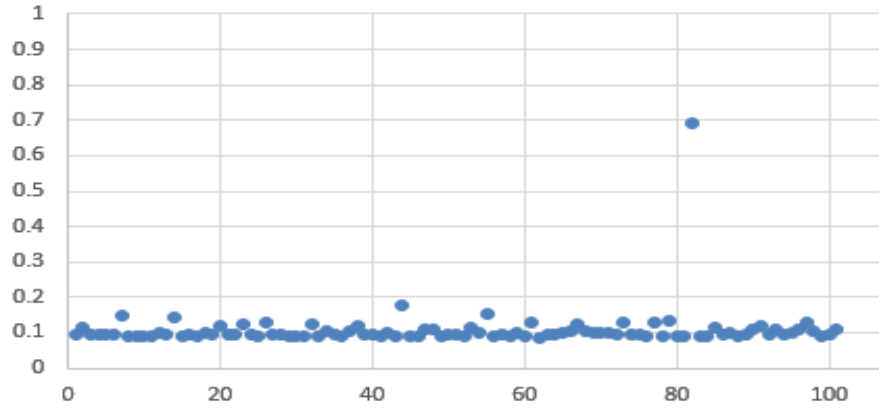
```
Last Priority Update:  7/4  01:03
Group                   Config   Use   Effective   Priority   Res   Total Usage  Time Since Requested
  User Name              Quota  Surplus  Priority    Factor   In Use (wghted-hrs) Last Usage Resources
------------------------------ ------- ---------- ---------- ----- ------------ ---------- ----------
cms                      0.05 ByQuota              1000.00     0     717935.81   0+02:37          8
  joshuha@ihep.ac.cn                      6657.68   1000.00     0      20027.31   0+02:37
  lizh@ihep.ac.cn                        41869.36   1000.00     0      49936.12   0+03:17
hxmt                     0.09 ByQuota              1000.00     0     762110.31   0+15:12          8
  lihancheng@ihep.ac.cn                   941.42   1000.00     0       6975.79   0+15:12
atlas                    0.05 ByQuota              1000.00     1     168250.70    <now>           1
  zhouws@ihep.ac.cn                       500.00   1000.00     0       4594.35   0+14:31
  sunxh@ihep.ac.cn                       3357.33   1000.00     1       2881.25    <now>
  xzxchen@ihep.ac.cn                     8411.85   1000.00     0      10043.20   0+13:36
higgs                    0.04 ByQuota              1000.00     2     985831.12    <now>           5
  cuizw@ihep.ac.cn                        500.00   1000.00     0      39631.74   0+02:56
```

```
-- Schedd: job@scheddd01.ihep.ac.cn : <192.168.51.33:28366>
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
23833135.0   huanghp         7/4  01:49   0+00:02:20 R  0   0.1  job_00001.txt.cond
23833137.0   huanghp         7/4  01:49   0+00:02:20 R  0   0.1  job_00002.txt.cond
23833139.0   huanghp         7/4  01:49   0+00:02:20 R  0   0.1  job_00003.txt.cond
23833142.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00004.txt.cond
23833144.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00005.txt.cond
23833146.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00006.txt.cond
23833149.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00007.txt.cond
23833151.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00008.txt.cond
23833153.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00009.txt.cond
23833156.0   huanghp         7/4  01:49   0+00:01:57 R  0   0.1  job_00010.txt.cond

10 jobs; 0 completed, 0 removed, 0 idle, 10 running, 0 held, 0 suspended
```

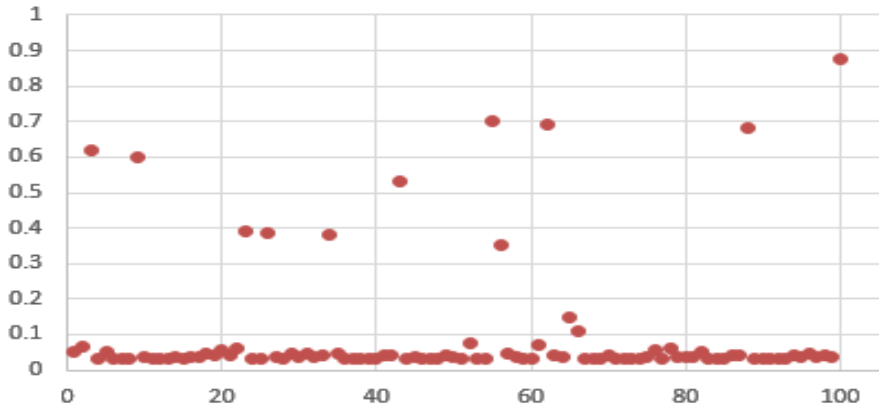
执行效率

- 对比htcondor与hepjob提交作业的执行效率

hepjob



htcondor



应用情况

- 支持实验
  - bes, juno,dyw, cms, lhaaso, hxmt等
- 支持调度器
  - htcondor, slurm
- 支持集群
  - physical, virtual, slurm, mpi

实验组	作业数量
bes	9881264
juno	7180867
dyw	1614148
cms	1520591
higgs	759667
lhaaso	500047
hxmt	448898
atlas	263520
cc	3004
comet	1314
总计	**22169002**

总结

- hepjob能够有效解决高能所的计算任务调度问题
  - 集中调度控制
  - 规范简化用户使用
  - 统一管理

- 后续与展望：
  - 优化中央管理系统，更好的支持多调度器
  - 优化提升hepjob的执行效率
  - 依据实际需求增加工具集、参数等

Thanks
Advice & Questions