HTCondor

姜晓巍

中国科学院高能物理研究所

2017年07月03日

目录

- HTCondor介绍
- HTCondor应用
- HTCondor作业
- HTCondor管理

HTCondor介绍

HTCondor

- HTCondor
  - 开源批处理系统，解决作业和资源的调度问题。
- 特点：
  - 高吞吐量
  - 大规模，可扩展性（20万slots）
- 现状：
  - 应用于CERN、LHC等大型高能物理研究中心
  - 高能所与HTCondor team保持良好交流

HTC与HPC

- HPC
  - 任务处理性能
- HTC
  - 任务处理吞吐

classad

- 核心策略（classad）
  - 依据作业classad与计算节点classad进行匹配

Pet Ad

```
Type  = "Dog"
Requirements =
    DogLover =?= True
Color = "Brown"
Price = 75
Sex = "Male"
AgeWeeks = 8
Breed = "Saint Bernard"
Size = "Very Large"
Weight = 27
```

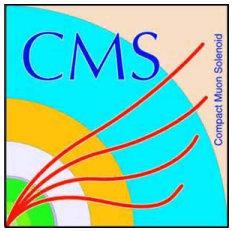
Buyer Ad

```
AcctBalance  = 100
DogLover = True
Requirements =
  (Type == "Dog")   &&
  (TARGET.Price <=
   MY.AcctBalance)  &&
  ( Size == "Large" ||
    Size == "Very Large" )
Rank =
 100* (Breed == "Saint
Bernard") - Price
```

HTCondor应用

高能物理计算

- 计算类型
  - 模拟
  - 重建
  - 分析
- 计算特点
  - 多以串行作业为主
  - 高吞吐量
- 其他
  - 用户群复杂
  - 资源隔离

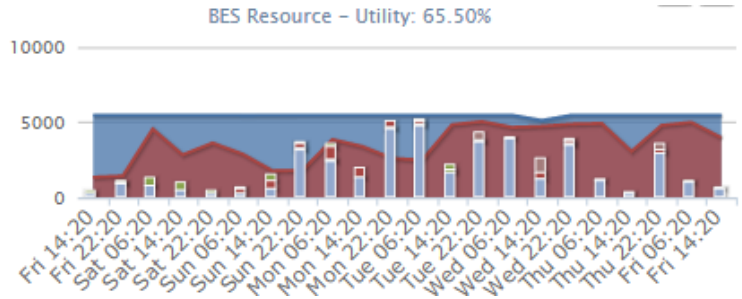


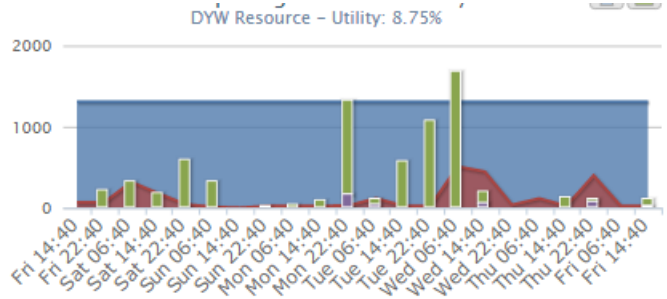
高能物理计算

- 设备状况
  - 28 提交节点
  - 3 队列服务器 (local cluster, virtual cluster, MPI cluster)
  - 3 中央管理服务器 (local cluster, virtual cluster, MPI cluster)
  - ~ 11000 物理 CPU cores和>1000动态的虚拟资源
- 作业状况
  - Avg 100,000 jobs/day
  - 主要为单核串行作业

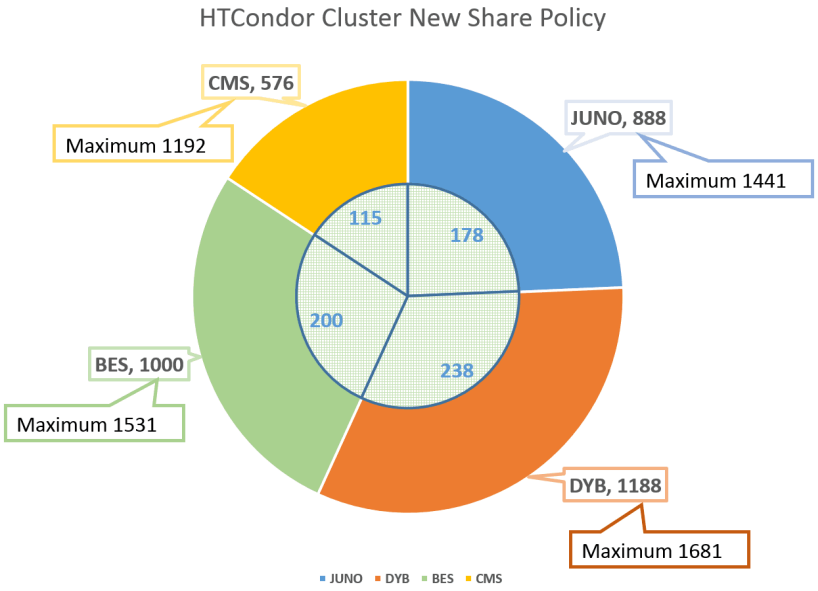
共享策略

- 资源隔离导致总体资源利用率不高



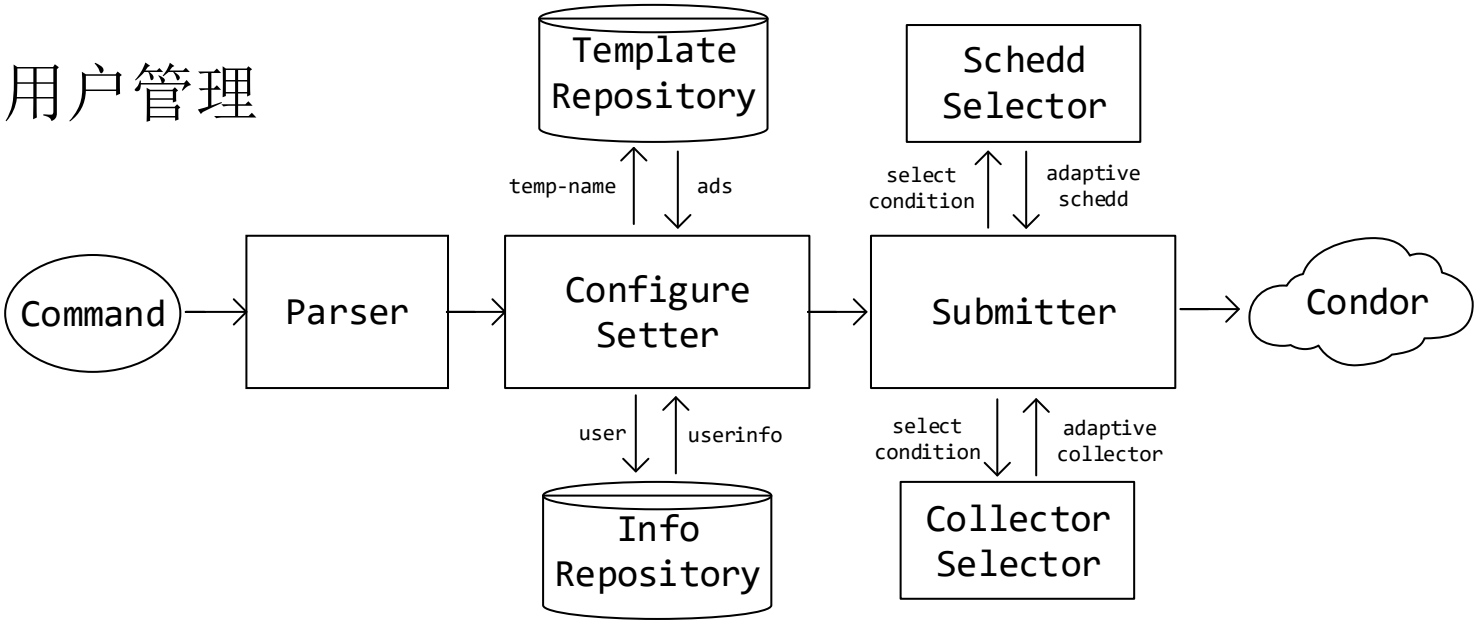


- 共享方式：
  - 各实验提供一定比例计算资源作为共享资源
  - 所有实验均可使用共享资源
- 公平共享
  - 用户使用资源少，优先级高

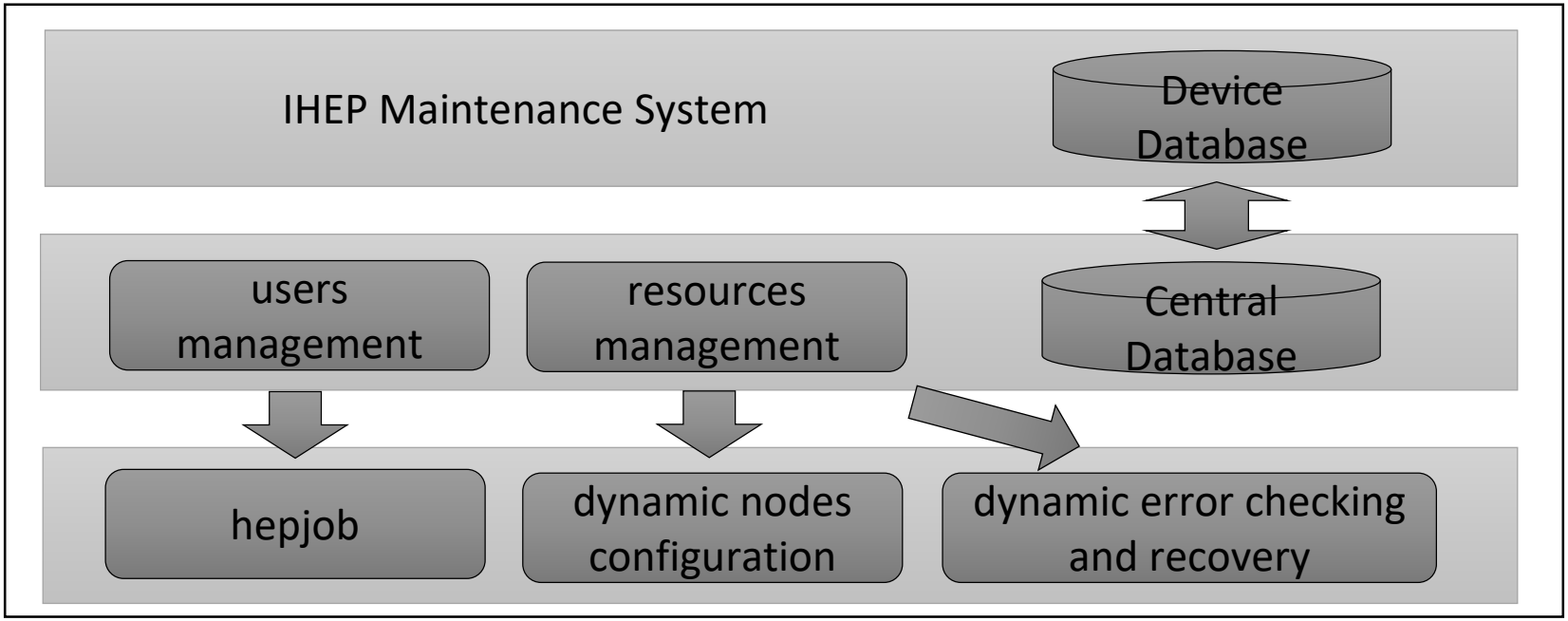


作业管理工具

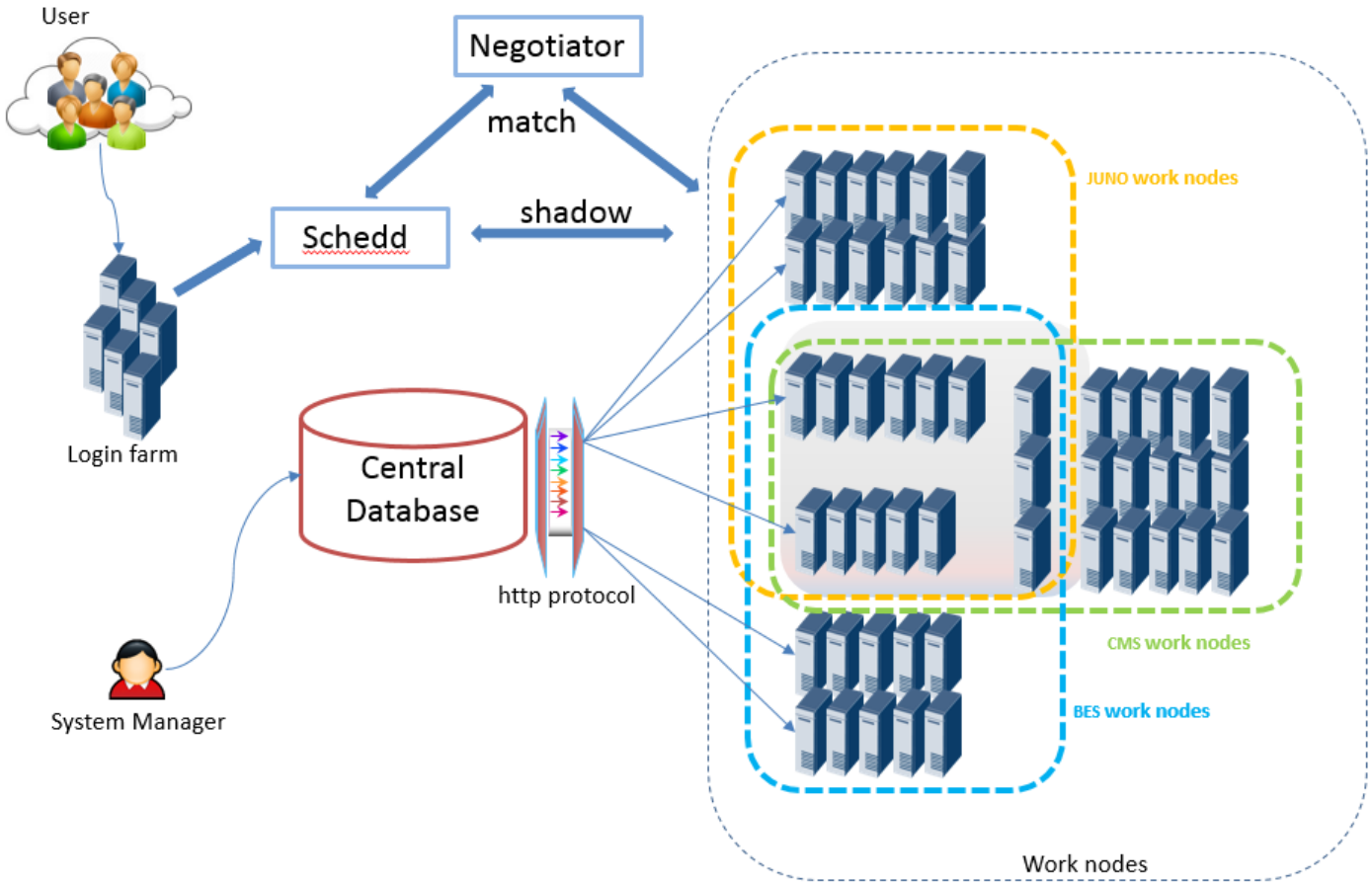
- hepjob
  - 规范和简化htcondor作业操作
  - 实现调度控制
  - 实现用户管理



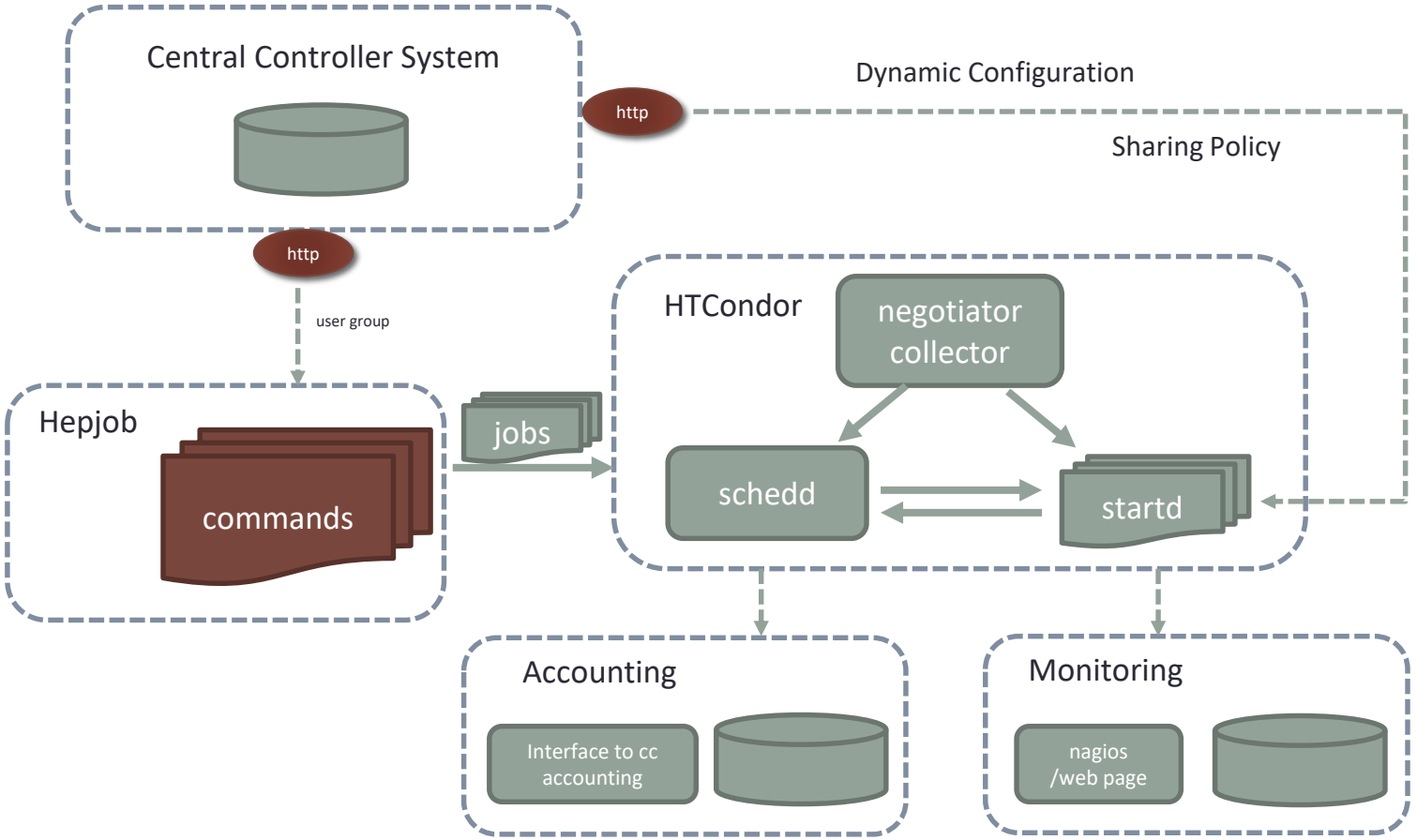
中央控制系统


IHEP Maintenance System
Device Database
users management
resources management
Central Database
hepjob
dynamic nodes configuration
dynamic error checking and recovery

集中部署


User
Negotiator
match
shadow
Schedd
Login farm
Central Database
http protocol
System Manager
JUNO work nodes
CMS work nodes
BES work nodes
Work nodes

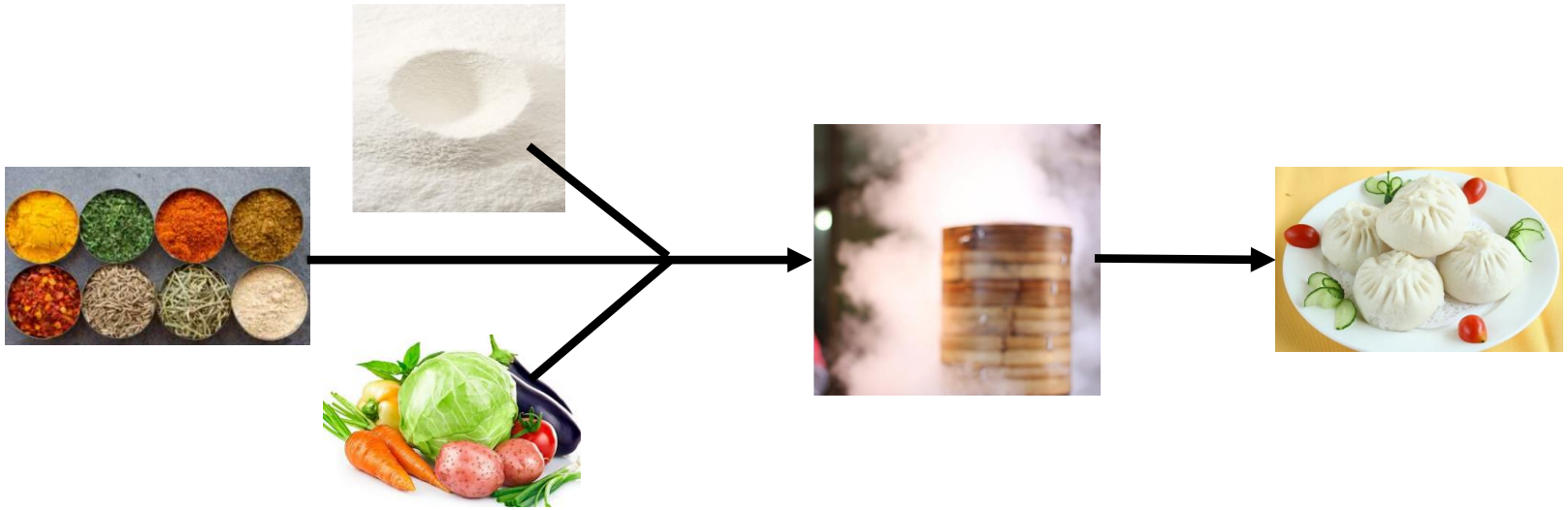
HTCondor集群架构



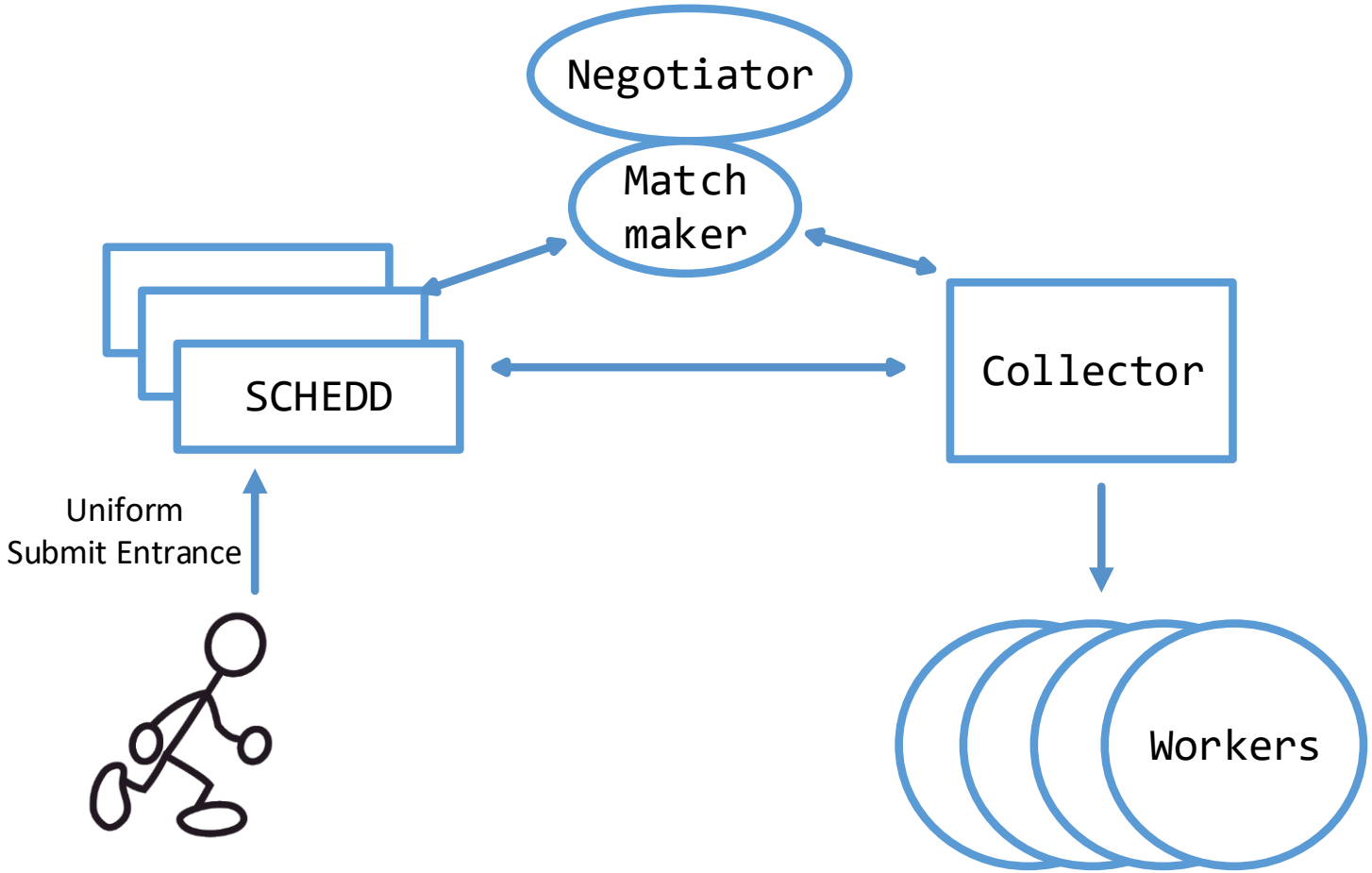
HTCondor使用

作业

- 作业即计算任务
- 作业主要包括三个部分，输入、执行程序和输出。



HTCondor作业


Negotiator
Match
maker
SCHEDD
Collector
Uniform
Submit Entrance
Workers

作业提交

- 1）编写作业执行程序
  - job.sh
- 2）编写作业描述文件：
  - submit.jdf
- 3）运行作业提交命令：
  - condor_submit

路径：
/eos/user/weih79

作业提交

- 作业执行程序

/eos/user/weih79/job.sh

```
#!/bin/bash

argu=$1
echo "hello," $1 "!"
for count in `seq 0 -1 -4`;
do
        time=$(/bin/date -d "$count day
ago" +%b' '%e', '%A)
        echo "have fun," $time "..."
        sleep 3
done
echo "bye," $1 "!"
```

作业提交

- 作业描述文件

/eos/user/weih79/submit.jdf

```
universe = vanilla
executable = job.sh
arguments = weihai

output = output/$(CLUSTER).$(PROCESS).out
error = error/$(CLUSTER).$(PROCESS).err
log = log/$(CLUSTER).$(PROCESS).log

should_transfer_files = yes
queue 1
```

作业描述文件

```
universe = vanilla

executable = job.sh
arguments = weihai

output = output/$(CLUSTER).$(PROCESS).out
error = error/$(CLUSTER).$(PROCESS).err
log = log/$(CLUSTER).$(PROCESS).log

should_transfer_files = yes
queue 1
```

- 声明作业类型

- Executable即作业执行程序
- Arguments即传入参数，列出所有参数以空格分隔

作业描述文件

```
universe = vanilla

executable = job.sh
arguments = weihai

output = output/$(CLUSTER).$(PROCESS).out
error = error/$(CLUSTER).$(PROCESS).err
log = log/$(CLUSTER).$(PROCESS).log

should_transfer_files = yes
queue 1
```

- output/error：作业的标准输出和标准错误
- log：HTCondor的跟踪作业处理过程。
- 注意：保证指定的路径存在

作业描述文件

```
universe = vanilla

executable = job.sh
arguments = weihai

output = output/$(CLUSTER).$(PROCESS).out
error = error/$(CLUSTER).$(PROCESS).err
log = log/$(CLUSTER).$(PROCESS).log

should_transfer_files = yes
queue 1
```

- 使用transfer方式，也可使用共享文件系统方式
- queue 1：代表提交一个作业

提交作业

- condor_submit：

```
condor_submit submit.jdf
```

提交作业

- 1) 编写相关文件
  - 作业执行程序、作业描述文件、输入输出及日志路径

```
cd /eos/user/weihxx
cp -rf /eos/user/weih79 *
```

- 2) 提交作业

```
condor_submit submit.jdf
```

查看作业状态

- 查看命令

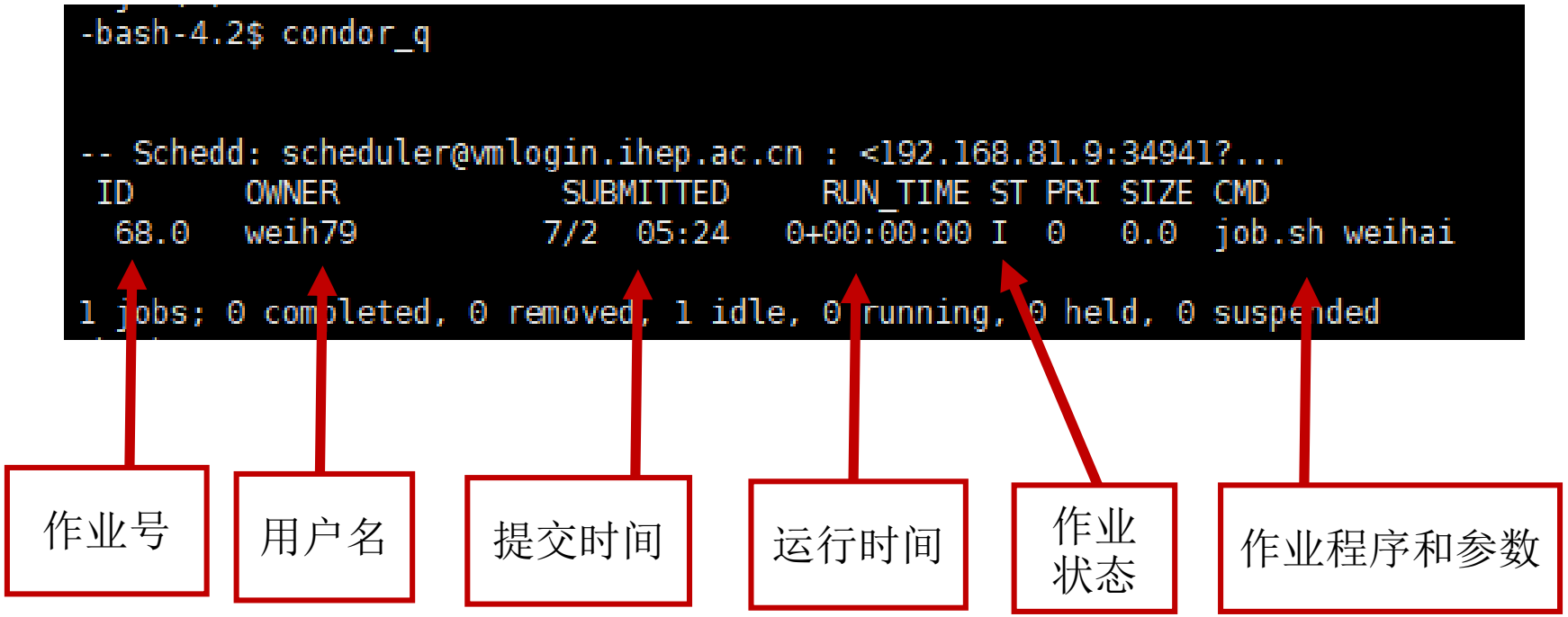
  condor_q

```
-bash-4.2$ condor_submit submit.jdf
Submitting job(s).
1 job(s) submitted to cluster 68.
```

```
-bash-4.2$ condor_q


-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
  68.0   weih79          7/2  05:24   0+00:00:00 I  0   0.0  job.sh weihai

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

作业查询信息

- 作业信息


-bash-4.2$ condor_q
-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
ID OWNER SUBMITTED RUN_TIME ST PRI SIZE CMD
68.0 weih79 7/2 05:24 0+00:00:00 I 0 0.0 job.sh weihai
1 jobs; 0 completed, 0 removed, 1 idle, 0 running, 0 held, 0 suspended
作业号
用户名
提交时间
运行时间
作业状态
作业程序和参数

作业排队

```
-bash-4.2$ condor_q


-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
  68.0   weih79          7/2  05:24   0+00:00:00 I  0   0.0  job.sh weihai

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

“I”即idle，作业在排队

```
weihxx/
     job.sh
     submit.jdf
     error/xx.err
     output/xx.out
     log/xx.log
```

作业运行

```
-bash-4.2$ condor_q


-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
  73.0   weih79          7/2  08:49   0+00:00:05 R  0   0.0  job.sh weihai

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

“R”即running，作业正在运行

```
weihxx/
     job.sh
     submit.jdf
     error/xx.err
     output/xx.out
     log/xx.log
```

```
execute/
     job.sh
     xx.err
     xx.out
     xx.log
```

作业结束

```
-bash-4.2$ condor_q 73


-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD

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

- 作业号为73的作业不在队列里

```
weihxx/
     job.sh
     submit.jdf
     error/xx.err
     output/xx.out
     log/xx.log
```

```
-bash-4.2$ cat output/73.0.out
hello, weihai !
have fun, Jul 2, Sunday ...
have fun, Jul 3, Monday ...
have fun, Jul 4, Tuesday ...
have fun, Jul 5, Wednesday ...
have fun, Jul 6, Thursday ...
bye, weihai !
```

作业删除

```
-bash-4.2$ condor_q


-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
  74.0   weih79          7/2  09:17   0+00:00:06 R  0   0.0  job.sh weihai
  75.0   weih79          7/2  09:17   0+00:00:00 I  0   0.0  job.sh weihai
  76.0   weih79          7/2  09:17   0+00:00:00 I  0   0.0  job.sh weihai
  77.0   weih79          7/2  09:17   0+00:00:00 I  0   0.0  job.sh weihai
  78.0   weih79          7/2  09:17   0+00:00:00 I  0   0.0  job.sh weihai
  79.0   weih79          7/2  09:17   0+00:00:00 I  0   0.0  job.sh weihai
  80.0   weih79          7/2  09:17   0+00:00:00 I  0   0.0  job.sh weihai

7 jobs; 0 completed, 0 removed, 6 idle, 1 running, 0 held, 0 suspended
```

- 删除作业号为80的作业

```
-bash-4.2$ condor_rm 80
All jobs in cluster 80 have been marked for removal
```

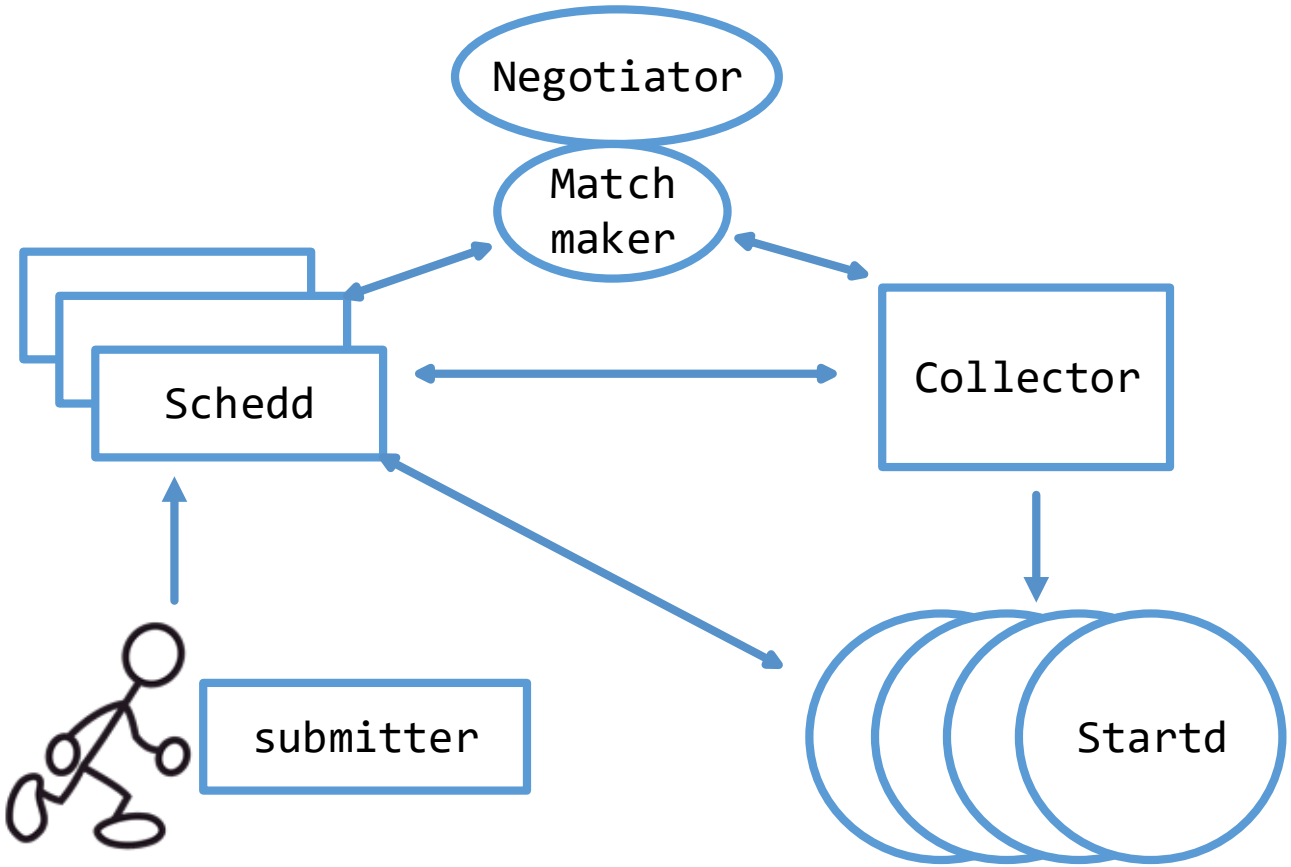
作业删除

```
-bash-4.2$ condor_q


-- Schedd: scheduler@vmlogin.ihep.ac.cn : <192.168.81.9:34941?...
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
  74.0   weih79          7/2  09:17   0+00:00:31 R  0   0.0  job.sh weihai
  75.0   weih79          7/2  09:17   0+00:00:11 R  0   0.0  job.sh weihai
  76.0   weih79          7/2  09:17   0+00:00:11 R  0   0.0  job.sh weihai
  77.0   weih79          7/2  09:17   0+00:00:11 R  0   0.0  job.sh weihai
  78.0   weih79          7/2  09:17   0+00:00:10 R  0   0.0  job.sh weihai
  79.0   weih79          7/2  09:17   0+00:00:11 R  0   0.0  job.sh weihai

6 jobs; 0 completed, 0 removed, 0 idle, 6 running, 0 held, 0 suspended
```

- 作业号为80的作业已移出队列

HTCondor管理

HTCondor系统结构



HTCondor安装

- yum安装

  https://research.cs.wisc.edu/htcondor/yum/

- rpm安装

  http://research.cs.wisc.edu/htcondor/yum/stable/rhel7/

```
-bash-4.2$ ls /eos/user/weih80/condor_packages/
condor-8.4.11-1.el7.x86_64.rpm
condor-classads-8.4.11-1.el7.x86_64.rpm
condor-external-libs-8.4.11-1.el7.x86_64.rpm
condor-procd-8.4.11-1.el7.x86_64.rpm
```

HTCondor安装

- 一般标准环境
  - 4类角色和配置
- login
  - 登陆节点，提交作业
- schedd
  - 维护队列
- cm
  - 调度和收集节点信息
- startd
  - 计算节点

Central manager

创建文件：/etc/condor/config.d/cm.conf

```
DAEMON_LIST = MASTER, COLLECTOR, NEGOTIATOR
COLLECTOR_NAME = collector
CONDOR_HOST = vmwn.ihep.ac.cn

ALLOW_WRITE = *
UID_DOMAIN = ihep.ac.cn
FILESYSTEM_DOMAIN = ihep.ac.cn


NEGOTIATOR_CONSIDER_PREEMPTION = false
```

启动condor服务

Schedd

创建文件：/etc/condor/config.d/schedd.conf

```
DAEMON_LIST = SCHEDD,MASTER
SCHEDD_NAME = scheduler
CONDOR_HOST = vmwn.ihep.ac.cn

ALLOW_WRITE = *.ihep.ac.cn

UID_DOMAIN = ihep.ac.cn
FILESYSTEM_DOMAIN = ihep.ac.cn

SEC_DEFAULT_AUTHENTICATION_METHODS = CLAIMTOBE
```

启动condor服务

Startd

创建文件：/etc/condor/config.d/startd.conf

```
DAEMON_LIST = MASTER, STARTD
CONDOR_HOST = vmwn.ihep.ac.cn
ALLOW_WRITE = *
UID_DOMAIN = ihep.ac.cn
FILESYSTEM_DOMAIN = ihep.ac.cn

SEC_DEFAULT_AUTHENTICATION_METHODS = CLAIMTOBE
```

启动condor服务

Login

创建文件：/etc/condor/config.d/login.conf

```
CONDOR_HOST = vmwn.ihep.ac.cn
SCHEDD_NAME = scheduler@vmlogin.ihep.ac.cn
UID_DOMAIN = ihep.ac.cn
FILESYSTEM_DOMAIN = ihep.ac.cn

SEC_DEFAULT_AUTHENTICATION_METHODS = CLAIMTOBE
```

注意：不启动condor服务

安装注意事项

- 1) 关闭服务器防火墙
  - cm/schedd/startd
- 2)condor服务
  - cm/schedd/startd启动condor服务
  - login不启动condor服务
- 3) 建议统一用户信息
  - login/schedd/startd

简易安装

- 1) 一台服务器身兼多个角色：
  - schedd/cm/startd

```
DAEMON_LIST = MASTER, COLLECTOR, NEGOTIATOR, SCHEDD, STARTD
COLLECTOR_NAME = collector
CONDOR_HOST = xxxx.ihep.ac.cn
SCHEDD_NAME = scheduler

ALLOW_WRITE = *
UID_DOMAIN = ihep.ac.cn
FILESYSTEM_DOMAIN = hep.ac.cn

SEC_DEFAULT_AUTHENTICATION_METHODS = CLAIMTOBE

NEGOTIATOR_CONSIDER_PREEMPTION = false
```

常用管理命令

- condor_q
- condor_status
- condor_history
- condor_userprio

condor_q

- condor_q -analyze
  - 分析作业处于等待状态的原因
- condor_q –l
  - 查看作业的所有信息
- condor_q –af +属性名
  - 定制输出作业属性

condor_status

- condor_status 查看资源状态

```
-bash-4.2$ condor_status
Name               OpSys      Arch   State     Activity LoadAv Mem   ActvtyTime

vm095085.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+00:00:03
vm095094.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:36:57
vm095098.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:37:53
vm095100.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:04:37
vm095101.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:37
vm095102.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:39
vm095103.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:37:58
vm095104.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:38:00
vm095105.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:37
vm095106.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:37
vm095107.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:39
vm095108.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:37
vm095109.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:37
vm095111.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:39
vm095112.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:39
vm095113.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:37
vm095114.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:36:57
vm095115.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:36
vm095120.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:42
vm095122.ihep.ac.c LINUX      X86_64 Unclaimed Idle      0.000 3695  0+01:54:39
                     Machines Owner Claimed Unclaimed Matched Preempting

        X86_64/LINUX       20     0       0        20       0          0

               Total       20     0       0        20       0          0
```

condor_userprio

- 查看用户优先级

```
-bash-4.2$ condor_userprio
Last Priority Update:  7/2  11:18
                    Effective   Priority   Res   Total Usage  Time Since
User Name            Priority    Factor   In Use (wghted-hrs) Last Usage
------------------ ----------- --------- ------ ------------ ----------
weih79@ihep.ac.cn        506.94   1000.00      0         1.53    0+00:02
------------------ ----------- --------- ------ ------------ ----------
Number of users: 1                             0         1.53    0+23:59
```

condor_history

- 作业历史情况

```
[root@vmlogin eos]# condor_history weih79
 ID     OWNER          SUBMITTED   RUN_TIME     ST COMPLETED   CMD
  86.0   weih79          7/2  09:35   0+00:02:04 C   7/2  09:38 /eos/user/weih79/job.sh weihai
  83.0   weih79          7/2  09:35   0+00:02:04 C   7/2  09:38 /eos/user/weih79/job.sh weihai
  84.0   weih79          7/2  09:35   0+00:01:43 C   7/2  09:37 /eos/user/weih79/job.sh weihai
  85.0   weih79          7/2  09:35   0+00:01:43 C   7/2  09:37 /eos/user/weih79/job.sh weihai
  82.0   weih79          7/2  09:35   0+00:01:42 C   7/2  09:37 /eos/user/weih79/job.sh weihai
  81.0   weih79          7/2  09:35   0+00:01:42 C   7/2  09:36 /eos/user/weih79/job.sh weihai
  78.0   weih79          7/2  09:17   0+00:01:42 C   7/2  09:19 /eos/user/weih79/job.sh weihai
  79.0   weih79          7/2  09:17   0+00:01:43 C   7/2  09:19 /eos/user/weih79/job.sh weihai
  77.0   weih79          7/2  09:17   0+00:01:43 C   7/2  09:19 /eos/user/weih79/job.sh weihai
  76.0   weih79          7/2  09:17   0+00:01:43 C   7/2  09:19 /eos/user/weih79/job.sh weihai
  75.0   weih79          7/2  09:17   0+00:01:43 C   7/2  09:19 /eos/user/weih79/job.sh weihai
  74.0   weih79          7/2  09:17   0+00:01:42 C   7/2  09:19 /eos/user/weih79/job.sh weihai
```

Thanks & Question

姜晓巍

jiangxw@ihep.ac.cn