Studies on Job Queue Health and Problem Recovery

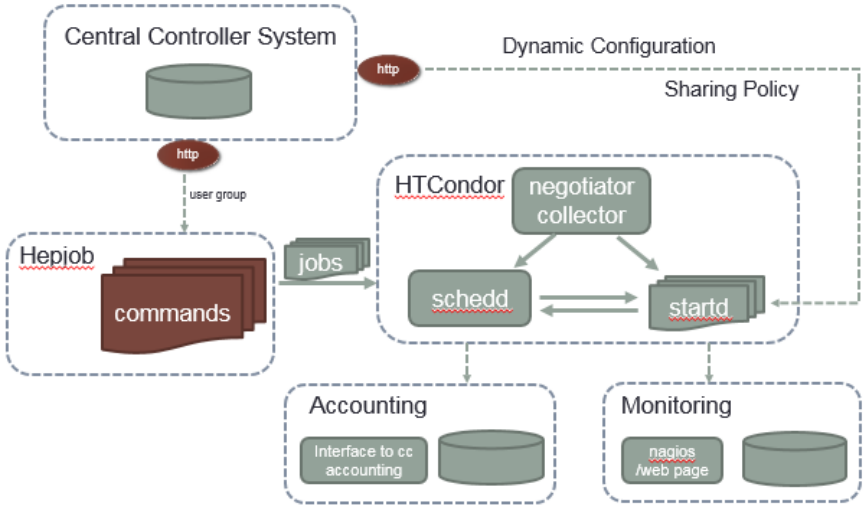
Xiaowei, JIANG
jiangxw@ihep.ac.cn
Scheduling Group, CC of IHEP
ISGC 2018

Outline

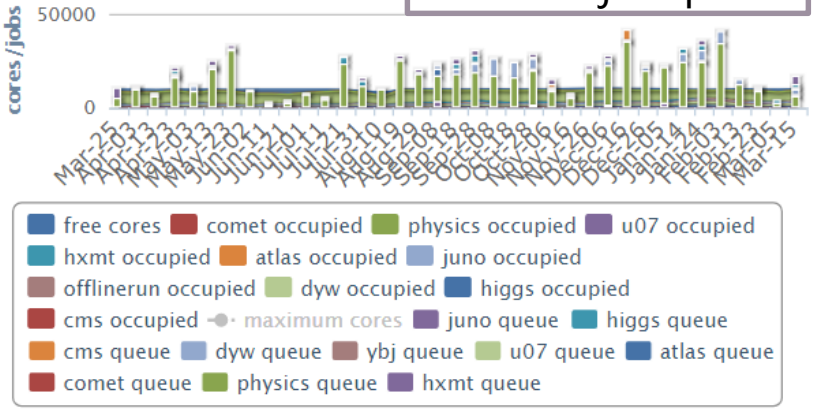
- Jobs and Resources at IHEP
- Introduction of Unhealthy Jobs
- Unhealthy Job Prevention and Check
- Problem Alarm and Handling

Job Scheduling at IHEP

- Now, htc computing cluster owns 11,000 CPU cores, support for BESIII, JUNO, DYW, CMS, ATLAS, HXMT,…
- After applying share policy, resource usage improves to 85%, nearly 100% at busy time.
- In 2017, >50 millions jobs are completed, average 150,000 jobs per day
- Local batch system has changed to HTCondor, which is managing the HTC cluster.



Always large number of jobs are stored in job queue

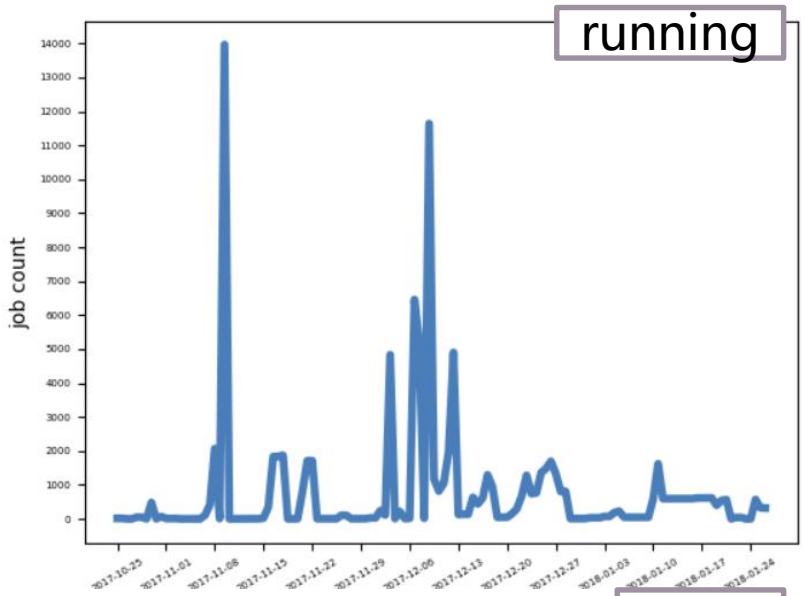


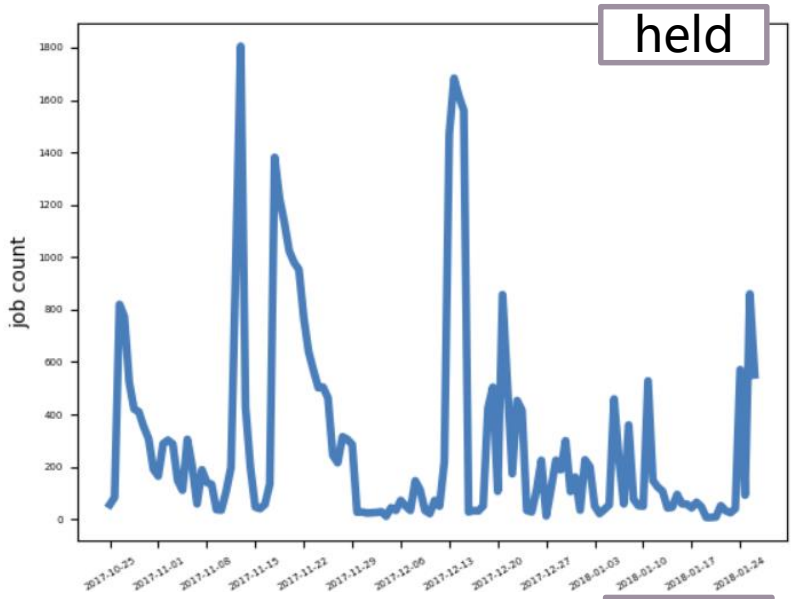
Unhealthy Job

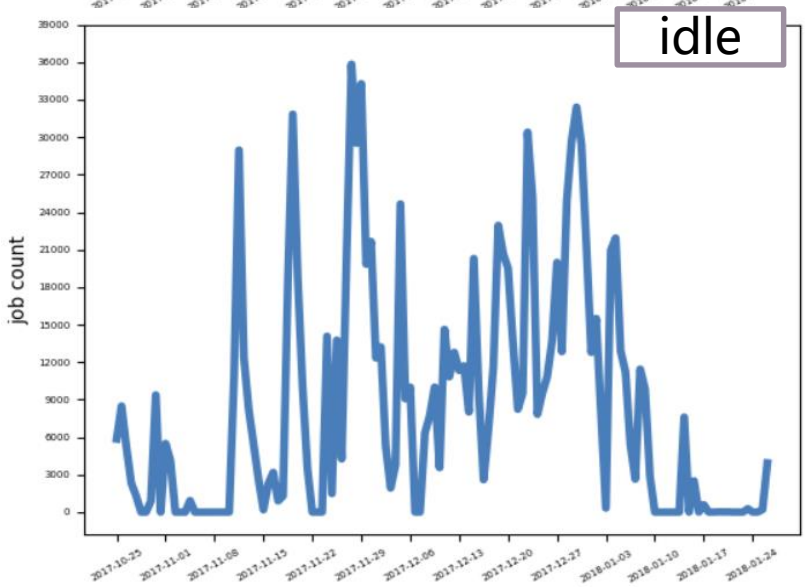
- What makes job unhealthy
  - Held
    - Problems of cluster： device errors, storage or network system problems, …
    - Misuse by users: incorrect or nonexistent path, exceeding disk quota, …
  - Idle over limitation time: low priority, wrong requirement conditions
  - Run over limitation time: occupied resources over limitation time
  - Removed unsuccessfully
- Motivations
  - Make unhealthy jobs handled in time
  - Save time spending on job maintenance for users and acquire job results as fast as possible.

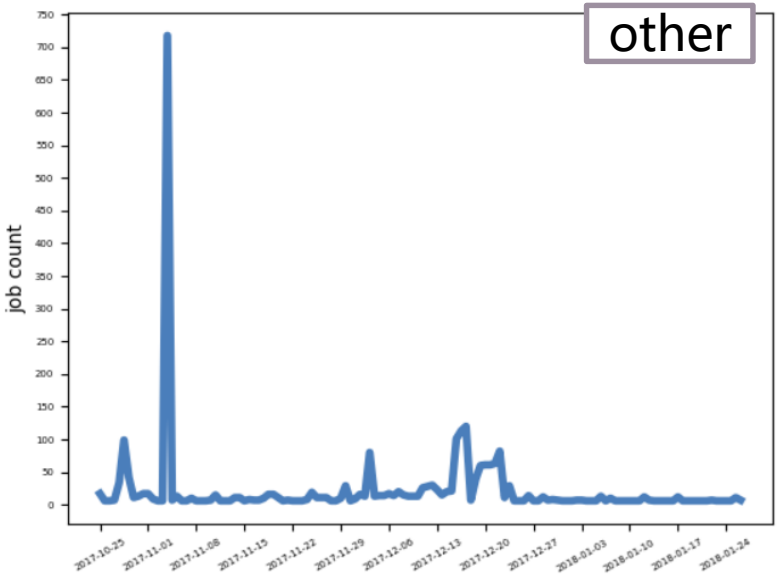
Unhealthy Job in Local Cluster

during Oct. 2017 to Jan. 2018



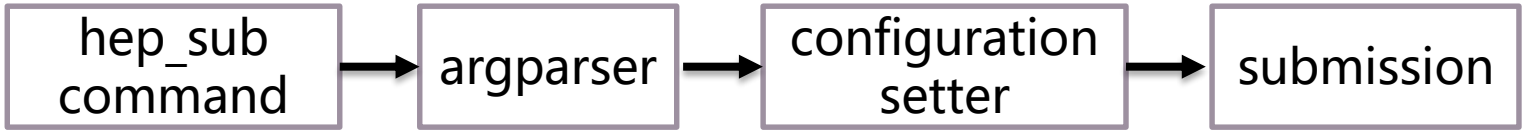






Pre-defined Job Classads

- One front-end scheduling tool add a predefined check in the submission process, helping users to config correct and precise matching conditions.



- 'accounting_group' is the most major attribute because the resources are divided into different division as groups, users only can request the resources in his/her groups.
- Other two cases:
  - Memory

```
## Maximum memory provided by cluster
MAX_MEMORY = 4800
```

  - Jobscript permission

```
permission = int(oct(os.stat(args.jobscript).st_mode)[-3:])
if not ((permission/100==7) or (permission/100)==5):
```

Self-check of Job Queue

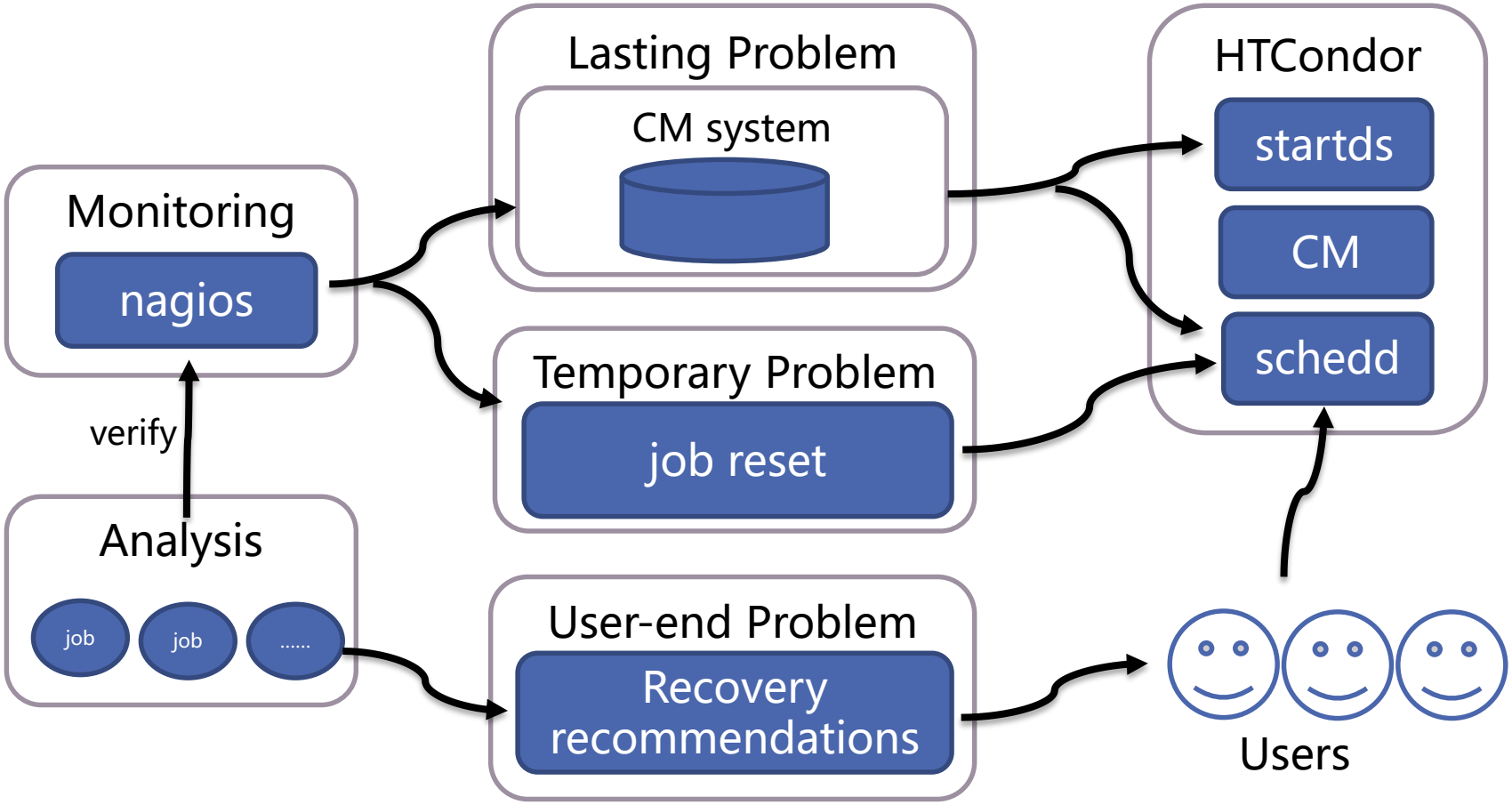
- Periodically scan over the job queue(schedd) with python-bindings and collect the unhealthy jobs as predefined conditions
- Conditions

Job status	Unhealthy features	Conditions
Held		JobStatus==5
Idle	Excessive idle time	JobStatus==1 && idle time > 86400
Running	Excessive run time	JobStatus==2 && WallTime > 100 hours
Removing	Excessive removing time	JobStatus==3 && remove time > 5 mins
other		jobStatus != 1/2/3/5

Categories of Unhealthy Jobs and Problems

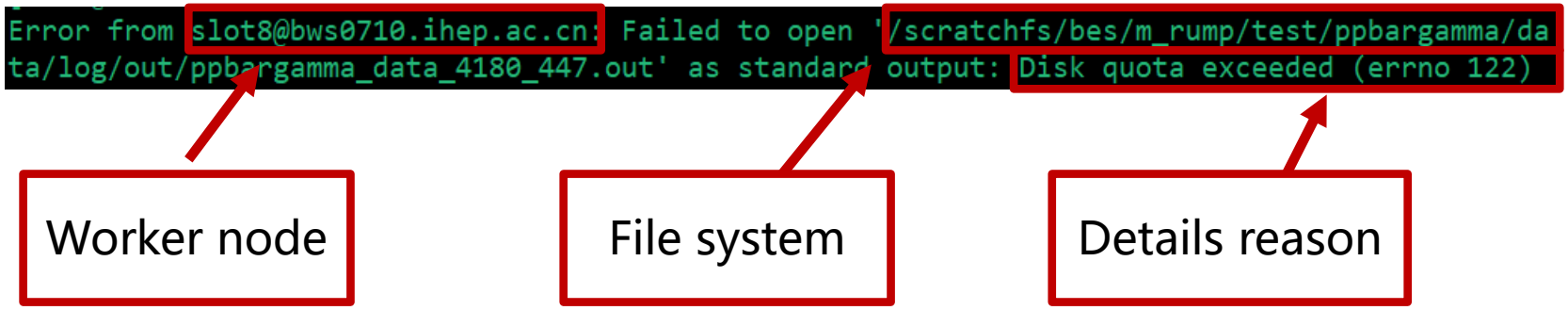
- Problems
  - Lasting resource problems
  - Temporary resource problems
  - User-end problems
- Jobs
  - Recovery by admin-end
  - Recovery by user-end

Problem Handling



Case of Held Job Analyzation(1)

- HTCondor provides an attribute “HoldReason” in job classads, which is pointing the reason caused job held as an defined format.
- One held case is shown as follow:



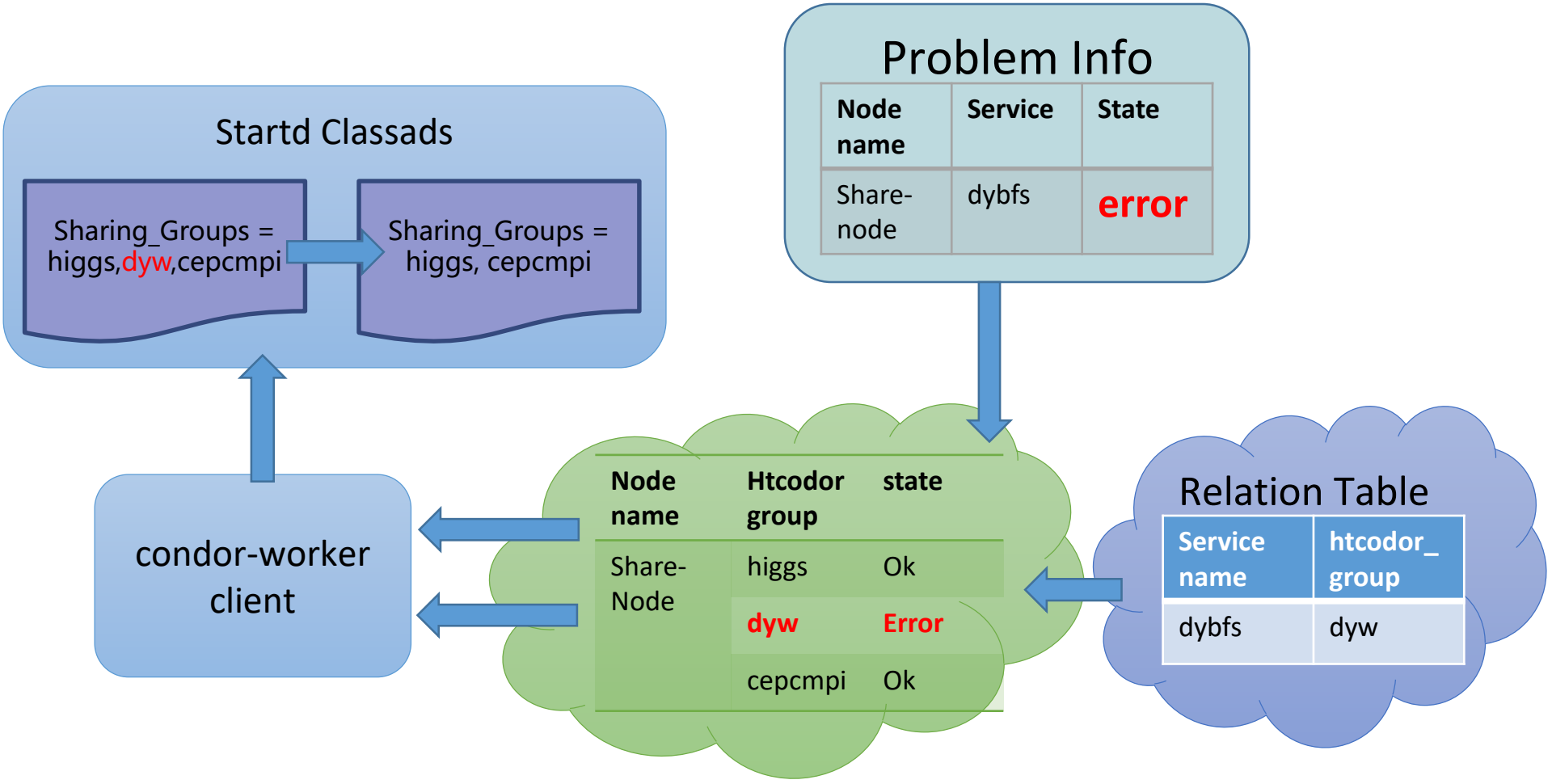
- In accordance with worker node, file system and detailed reasons, problem can be located and targeted recovery solution can be found.

Case of Held Job Analyzation(2)

- The following table shows some of the held problem which happened most frequently

status	Type	Type content	Reason
held	iwd	Cannot access initial working directory	No such file or directory
held	startd	Error from %: failed to open % as standard output	Input/output error (errno 5)
held	startd	Error from %: failed to open % as standard output	Permission denied (errno 13)
held	startd	Error from %: failed to open % as standard output	Disk quota exceeded (errno 122)
held	startd	Error from %: failed to open % as standard output	Transport endpoint is not connected (errno 107)
held	startd	Error from %: Cannot start container	invalid image name

Problems Recovery at Worker Node

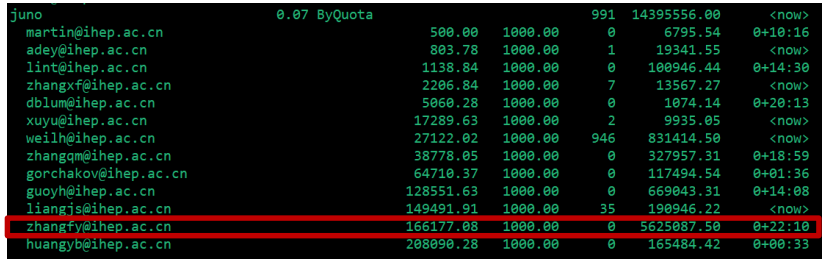


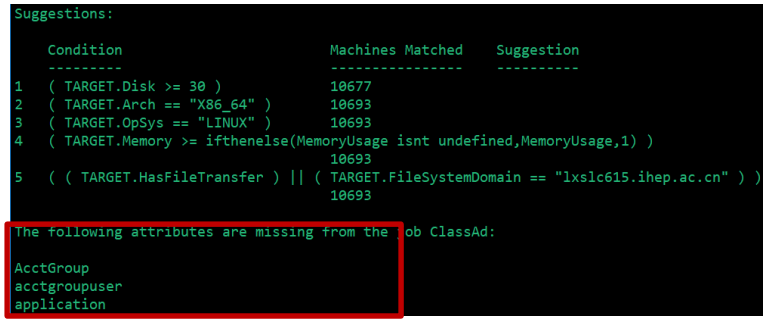
Problems Recovery at Job Queue

- Via HTCondor schedd operations, unhealthy jobs would be released or removed in time as the results of unhealthy job check and analysis.
  - Part of held jobs which are able to be covered by admin
  - Idle jobs which are waiting for matching for too long time
  - Running jobs which are claimed work nodes for too long time
  - Jobs in other status which should be not remaining unchanged

Problems Recovery at User-end

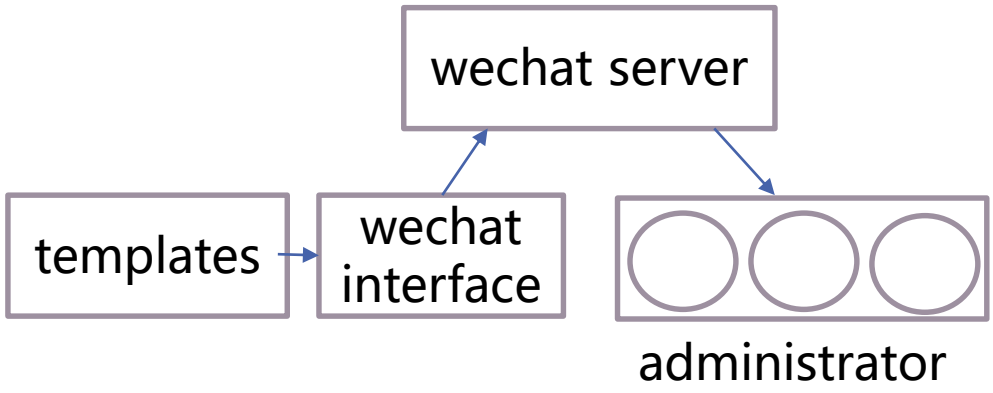
- Just give alarms to users with no recommendations
- But some effective recommendations could be provided to help users recovering their job issues in time
- Situations:
  - Lower Priority
    - Example of zhangfy
    - PRIORITY_HALFLIFE = 86400
    - Estimate the next match as the priority increasing
  - Wrong requirements classad
    - condor_q -analyze



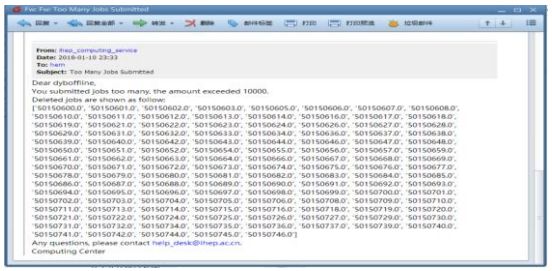


Alarm

- Wechat alarm
  - Wechat is a Chinese multi-purpose messaging and social media app developed by Tencent.
  - Wechat provides a public accounting for companies or communities.



- Mail alarm
  - Send alarm info to users



Summary

- Unhealthy jobs often appear in job queue, especially in htc cluster, which lead to delaying to get the results of experiment data processing.
- Handling unhealthy jobs in time can help users to reduce the time spending on job maintenance.
- A basic handling process of unhealthy jobs are shared.

Thanks !