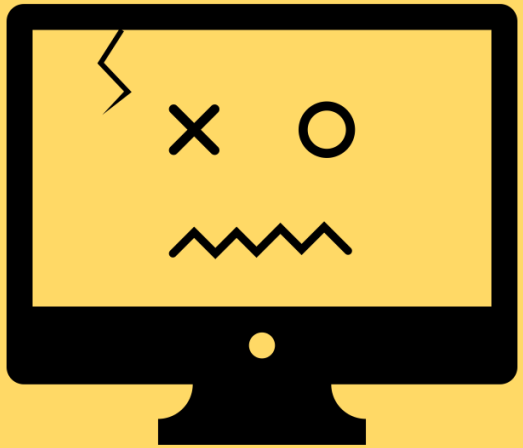
Troubleshooting Jobs on OSPool

Showmic Islam
Research Computing Facilitator@ OSG
HPC Application Specialist
Holland Computing Center
University of Nebraska-Lincoln

Outline

Job Failures

- Why a job may fail?
- What can go wrong?
- Reviewing failed jobs
- Job holds



Diagnosis

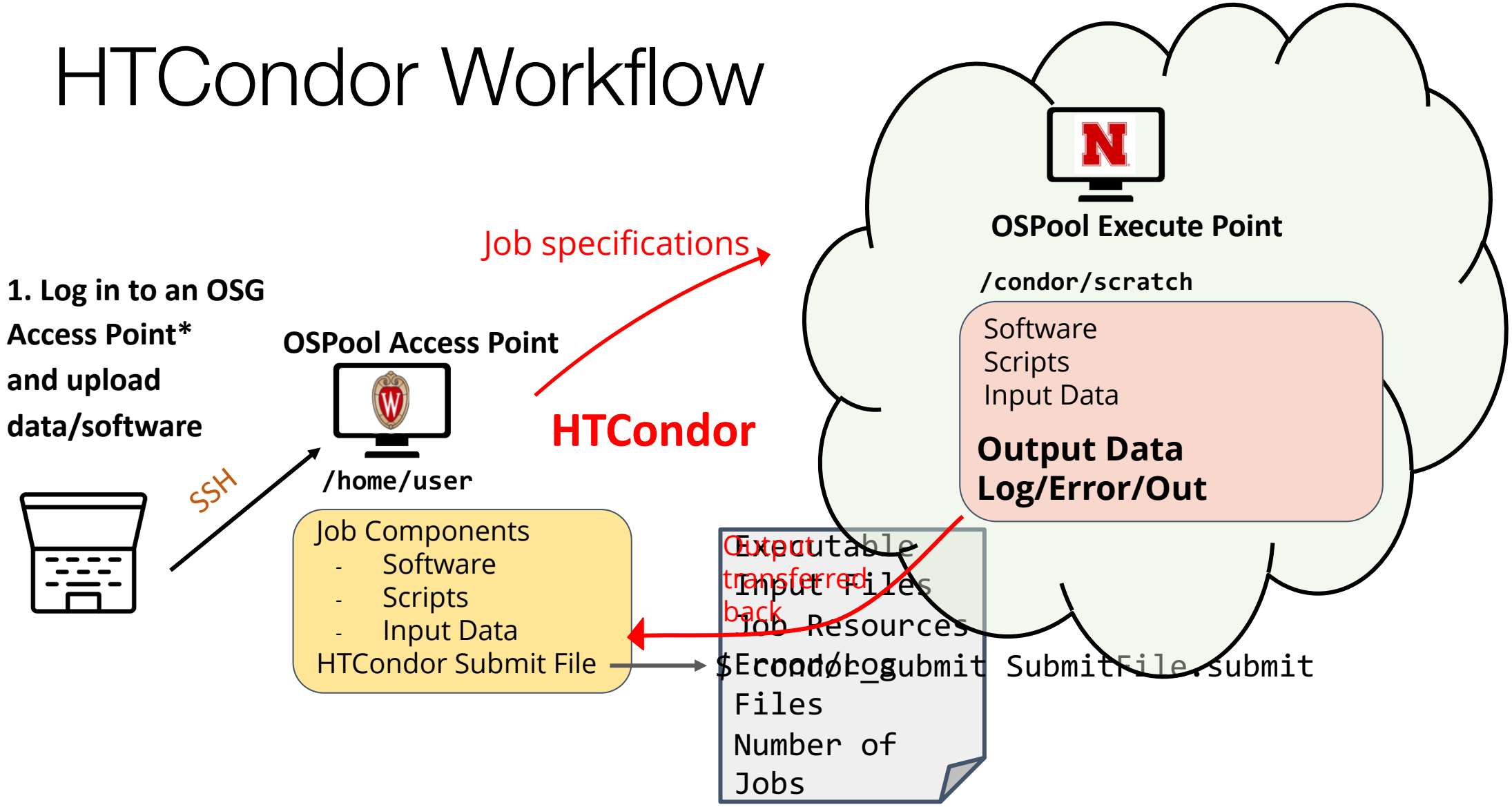
- Tips for troubleshooting
- Diagnosing Holds

Common Issues

- Examples of typos
- Badput

Broken computer by akash k from NounProject.com
Stethoscope by wyasa design from NounProject.com
Issues by sugiyarto from NounProject.com

HTCondor Workflow

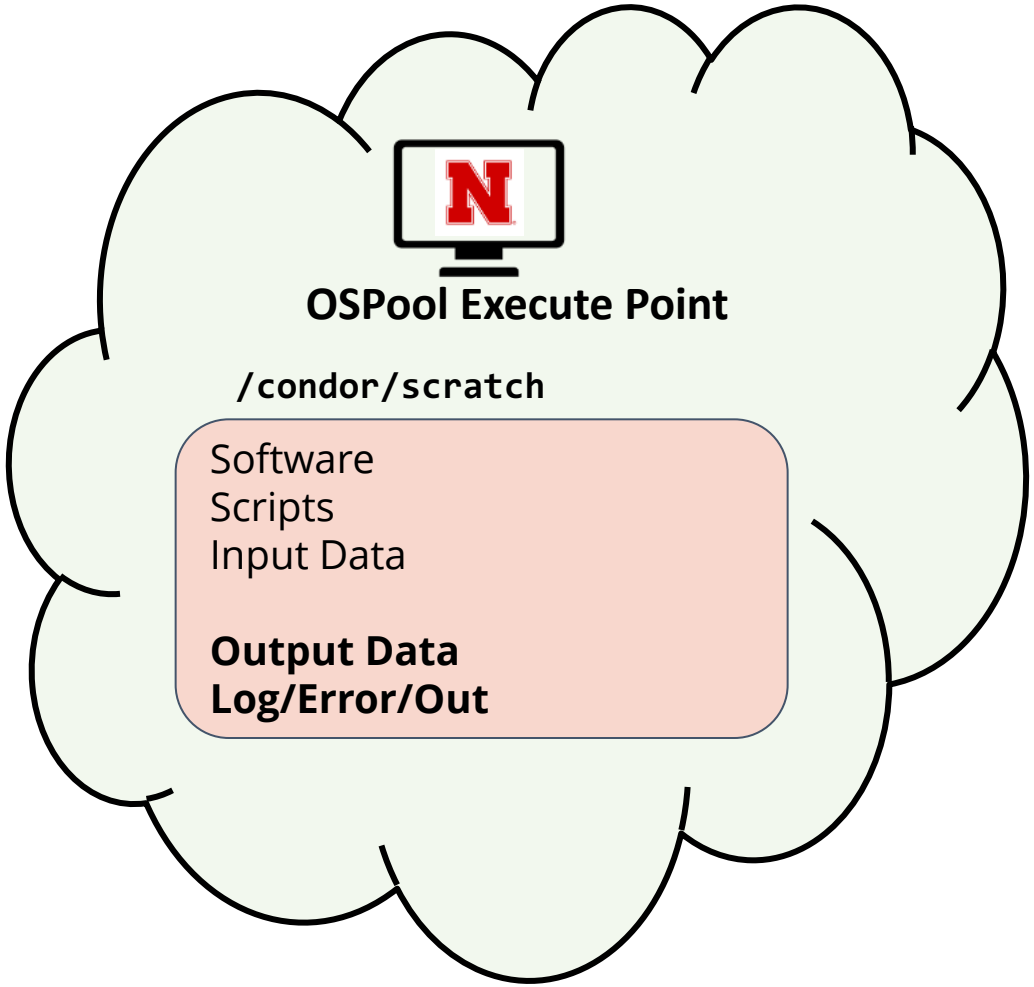


Job Failures

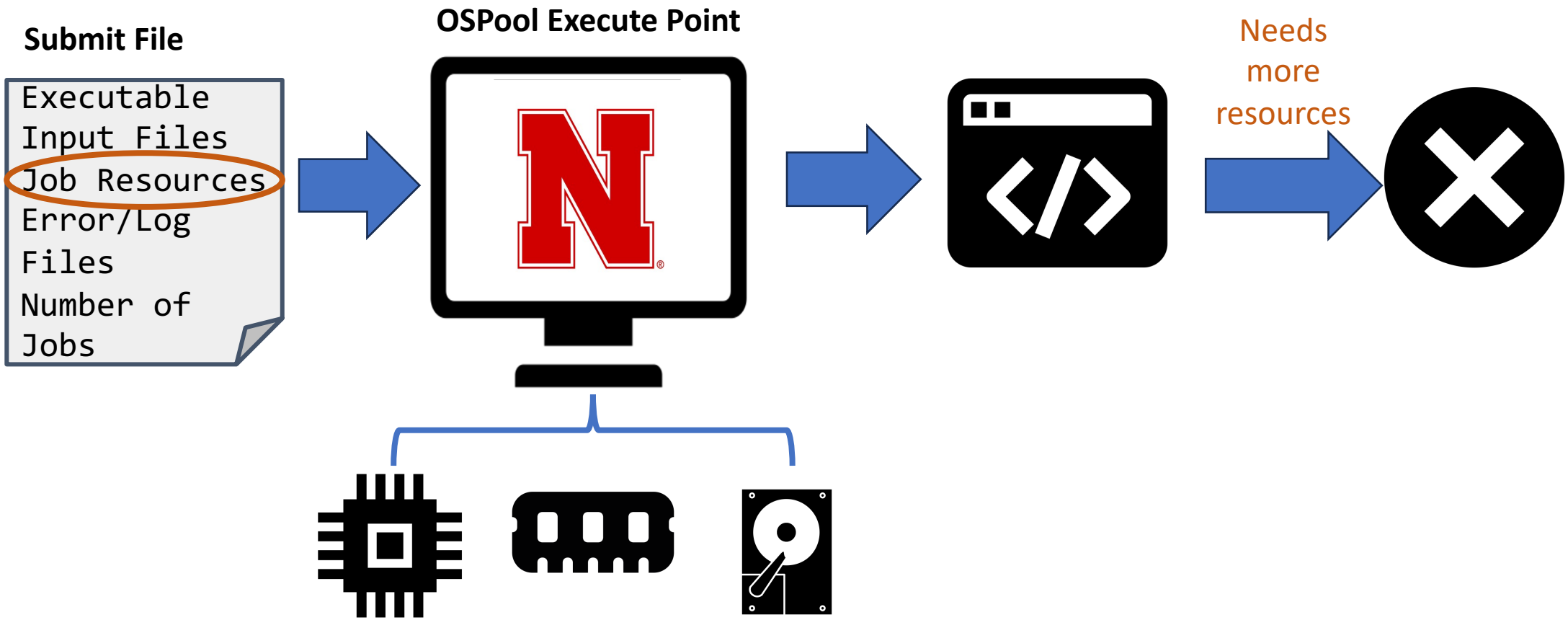
What can go wrong?

Let's start with the issues that we are all familiar with

- ❑ The code tries to run and fails
  - Script has typos (e.g. : misspelled input arguments)
  - Path names to a file or data are misspelled/wrong
  - Software does not have the required libraries (e.g. : mismatch between libraries in Window/linux)



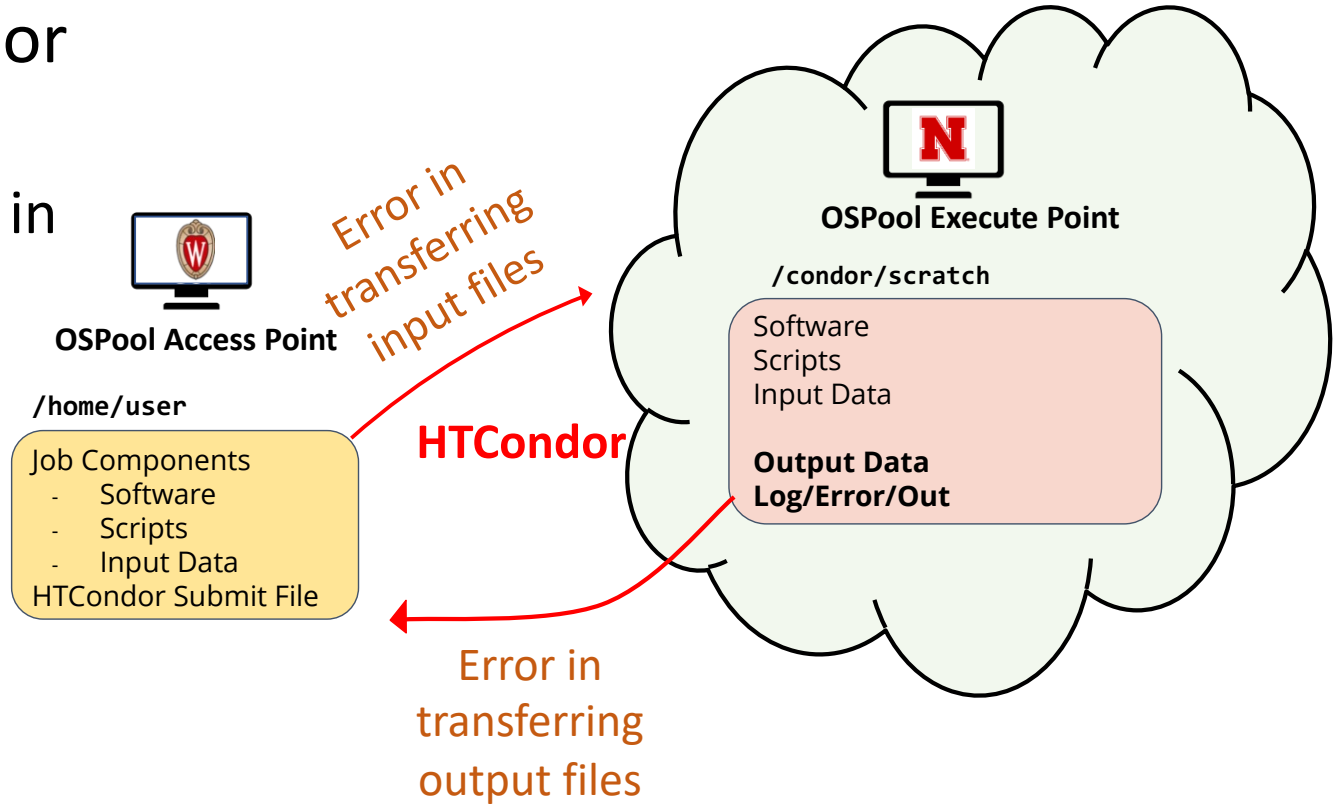
What else can go wrong?



Computer by miracle from NounProject.com
CPU by Prithvi from NounProject.com
RAM by George Ianta from NounProject.com
Code by Lagot Design from NounProject.com
Hard drive by Perilisima Shoeder from NounProject.com
Fail by Bluetip Design from NounProject.com

What other things can go wrong?

- Jobs can go wrong in the HTCondor workflow.
  - A job can't be matched(no machine in the pool to accommodate user's request)
  - Files not found for transfer
  - Job used too much memory
  - Job used too much disk space
  - Badly-formatted executable
  - And many more



Job Holds

- HTCondor will *hold* your job if there's a *logistical* issue that YOU (or maybe an admin) need to fix.
  - files not found for transfer, over memory, etc.
- A job that goes on hold is interrupted (all progress is lost) but remains in the queue in the "**H**" state until removed, or (fixed and) released.

```
$ condor_q
OWNER    BATCH_NAME     SUBMITTED   DONE   RUN    IDLE   HOLD  TOTAL JOB_IDS
cat      ID: 123456   7/11 11:23      _      _      _      1      1 123456.0
```

Diagnosis

General Troubleshooting Tips

- Comparing **expectations vs. what happened**: Either might be wrong!
- **Read messages carefully** — even if some parts make no sense, what hints can you get?
- Search **online** … but evaluate what you find
- Collect links and other resources that help
- Ask for help! And provide key details: versions, commands, files, messages, logs, etc.
- Always keep the log, error and condor output file

Reviewing Failed Jobs

- Job log, output and error files can provide valuable troubleshooting details:

Log	Output	Error
• when jobs were submitted, started, held, or stopped • where job ran • resources used • interruption reasons • **exit status**	• stdout (or other output files) may contain errors from the executable	• stderr captures errors from the operating system, or reported by the executable, itself.

Diagnosing Holds: Hold Reasons

If HTCondor puts a job on hold, it provides a hold reason, which can be viewed in the log file, with

`condor_q –hold <Job.ID>`, or with **`<username>`**:

*Failed to initialize user log to **/path***

- ❑ Could not create log file, check ***/path*** carefully

Error from …: memory usage exceeded request_memory

Job in status 2 put on hold by SYSTEM_PERIODIC_HOLD due to memory usage BBB.

- ❑ Job used too much memory
- ❑ Request more memory atleast *BBB* megabytes

*Transfer input files failure at **access point ap40** while sending files to the execution point. Details: reading from file **`/path`**: (errno 2) No such file or directory*

- ❑ Job can not find the files in **`/path`** *to transfer to execute point*
- ❑ Jargon: **SHADOW** is Access Point, **STARTER** is Execute Point

Example of common hold reasons

*Error from …: STARTER at … failed to send file(s) to <…>: error reading from **/path**: (errno 2) No such file or directory; SHADOW failed to receive file(s) from <…>*

- Job specified **transfer_output_files** but **/path** on execute point was not found

The job exceeded allowed execute duration of 20:00:00

- Job ran for too long

Error from ….: Starter failed to upload checkpoint

- Job failed to checkpoint (more on Thursday)

*Transfer output files failure at access point… while receiving files from the execution point. Details: Error from ….STARTER at … failed to send file(s) to …; SHADOW at … failed to create directory **/path** Disk quota exceeded*

- File **transfer error** due to exceeding disk space

What To Do About Held Jobs

1. If the situation can be fixed while job is held (e.g., you forgot to create directory for output):
   a. Fix the situation: `condor_qedit`
   b. Release the job(s): `condor_release` ***JOB_IDs***
      `condor_release <username>`
2. Otherwise (and this is common):
   a. Remove the held jobs: `condor_rm` ***JOB_IDs***
   b. Fix the problems
   c. Re-submit

DEMO

Common Issues

Issue: Failed to Parse

```
$ condor_submit job.sh
Submitting job(s)
ERROR: on Line 6 of submit file:
ERROR: Failed to parse command file (line 6).
```

- Completely failed to submit!
- **Notice:** `Failed to parse`
- **Why:** You tried to submit your executable (or other file), not an HTCondor submit file
- **Fix:** Submit an HTCondor submit file (e.g., .sub)

Issue: Typos in Submit File

```
$ condor_submit sleep.sub
Submitting job(s)
• ERROR: No 'executable' parameter was provided
• ERROR: Parse error in expression:
         RequestMemory = 1BG
• ERROR: Executable file /bin/slep does not exist
```

- Also failed to submit (missing job(s) submitted)
- **Why:** Typos in your submit file (e.g., BG for GB)
- **Fix:** Correct typos!

Issue: Jobs Idle for a Long Time

```
$ condor_q
OWNER    BATCH_NAME     SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
cat      ID: 123456    6/30 12:34      _      _       1      1 123456.0
```

Jobs are `idle` for a **long** time – *can be hard to judge!*

condor_q -analyze <JobId>

condor_q –better-analyze <JobId>

```
$ condor_q -better-analyze 123456.0
...
          Slots
Step    Matched  Condition
-----  --------  ---------
[0]       13033  TARGET.PoolName == "OSPool"
[9]       13656  TARGET.Disk >= RequestDisk
[11]          0  TARGET.Memory >= RequestMemory
```

Issue: Missing or Unexpected Results

- Job runs … but something does not seem right
  - Short or zero-length output file(s)
  - Very short runtime (almost instant)
- May be problems with app, inputs, arguments, …
  - Check log files for unexpected exit codes, etc.
  - Check output and error files for messages from app
  - Can't find anything? Add more debugging output

Issue: Badput

- What is *badput*?
  - Basically, wasted computing
    - Job runs for *97 minutes*, gets kicked off, starts over on another server
    - Job runs for *97 minutes*, is removed
  - Not jobs that must be re-run due to code changes! (that's just part of science, right?)
- Badput uses resources that others could have used
- If contacted, help us help you and others!

Tips for Avoiding Badput

- Always test with a **small set of jobs** before scaling up. (This practice applies to any modifications made to a **tried and tested** code as well)
- Monitor your jobs memory and disk usage

  condor_q <jobid> -af RequestMemory MemoryUsage |sort |uniq –c

  condor_q <jobid> -af RequestDisk DiskUsage |sort |uniq –c
- Have an idea/expectation about the software/code's limit- e.g. **Segfault.**
- Have a general idea about the inner workings of the software and libraries.

DEMO 2

More Troubleshooting Resources

- Tim Cartwright's OSG User School 2021 talk
- OSPool Documentation-https://portal.osg-htc.org/documentation/
- Can't solve issues-email us: support@osg-htc.org

Acknowledgements

Christina Koch; Tim Cartwright

This work was supported by NSF grants MPS-1148698, OAC-1836650, and OAC-2030508