Workflow Failures

- Soft Failures
  - Job is running for too long
  - Job is queued for too long
  - Rate of jobs entering various states changes significantly in a short time
  - Workflow is not making progress
  - Job fails repeatedly
  - Workflow has too many failed jobs
  - DAGMan lost track of a job
- Hard Failures
  - Job fails with non-zero exit code
  - Job runs for too long
  - Job is stuck in the queue
  - Job succeeds, but fails to produce one of its output files
  - An input file is missing
  - User's proxy is expired
  - Job does not produce an invocation record
  - Executable is missing
  - Out of space
  - Expired or missing X.509 certificate

Soft Failures

Job is running for too long

Symptoms

Based on previous results for the same job type, this job has been in the running state for an abnormally long period of time.

Steps to Reproduce

Create a workflow that has 99 jobs that all run for 10 seconds, and 1 job that runs for 100 seconds.

Job is queued for too long

Symptoms

Based on previous, recent history for jobs in this workflow, a particular job has been queued for an unusually long time.

Steps to Reproduce

Create a workflow that has a condor::requirements profile that requires the job to be in the queue for some long period before running.

Rate of jobs entering various states changes significantly in a short time

Symptoms

- Job failure rate increases significantly
- Job starting rate decreases significantly
- Job success rate decreases significantly

Steps to Reproduce

Create a workflow with multiple levels. All the jobs in the top level succeed, and all the jobs in lower levels fail.

Create a workflow that has a long chain of jobs, then add a level with a bunch of parallel jobs.

Something similar can be done for other states.

Workflow is not making progress

Symptoms

No events have been seen for a given workflow for an abnormally long period of time, or the rate of events changes significantly.

Steps to Reproduce

Create a workflow that has a bunch of short jobs followed by jobs that sleep forever.

Job fails repeatedly

Symptoms

Job is being retried, but has failed many times.

Steps to Reproduce

Create a workflow that has a job with retries = 100, and that always returns non-zero.

Workflow has too many failed jobs

Symptoms

A large percentage of jobs in the workflow has failed.

Steps to Reproduce

Create a workflow where most of the jobs fail and only a few succeed.

DAGMan lost track of a job

Symptoms

There are 2 cases:

1. DAGMan is running, but there are no jobs in the Condor queue. In this case DAGMan thinks a job is still running, but the job has finished.
2. DAGMan submits the same job twice and ignores the first submission. This usually happens when the first submission succeeded, but condor_submit tells DAGMan that it did not for whatever reason.

Steps to Reproduce

It is not clear what causes this problem, so it is not clear how we can reproduce it, but it usually occurs when you run a workflow and put a lot of load on the schedd. It might be possible to reproduce duplicate submissions by just waiting until a job is submitted, and then manually submitting it again. It might also be possible to reproduce this by putting a wrapper around condor_submit that randomly returns a non-zero exit code.

Hard Failures

Job fails with non-zero exit code

Symptoms

The return code of the job is not zero. There may also be some errors in the stdout/stderr of the job.

Steps to Reproduce

Add a job to the workflow that returns a non-zero exit code, or modify an existing job to return non-zero.

Job runs for too long

Symptoms

Job's status is 'Running' for too long. The definition of "too long" depends on the job type, arguments, and execution host.

Steps to Reproduce

Add a job to the workflow that runs /bin/sleep for longer than the expected time.

Job is stuck in the queue

Symptoms

The job's status is 'Idle' for too long, or the job's status is 'Held'. The definition of "too long" depends on the execution environment. This can be caused by many different things: bad requirements, requirements that do not match glideins, GRAM problems, busy site, etc.

Note: We just need some way to determine that the job is not making progress, not a diagnosis of the problem. It is sufficient to say that the job's behavior is anomalous.

Steps to Reproduce

Idle status can be achieved by submitting a job that has a Condor "requirements" expression that does not match any available resources. For example: "requirements = False".

Held status can be achieved using the "condor_hold" command.

Job succeeds, but fails to produce one of its output files

Symptoms

A job exits with return code 0, but does not produce one or more of its expected output files. Later, a job that depends on the missing output fails with "file not found" or some other error. For example, if there are two jobs, X and Y, and a dependency X->Y, but X does not produce any outputs, then Y should fail.

Steps to Reproduce

Wrap a job in an existing workflow with a script that deletes one of the job's outputs.

An input file is missing

Symptoms

A job fails because one of its input files could not be located. Typically there is some sort of "file not found" error in stdout/stderr. Missing workflow inputs should fail on the stage in transfer job, missing intermediate files will either fail on a normal job, or on a stage out transfer job.

Steps to Reproduce

For a missing workflow input, locate a file that is listed in the replica catalog for the workflow and move it to another location. For a missing intermediate file, wrap a job in the workflow with a script that deletes an output file.

User's proxy is expired

Symptoms

Grid jobs fail on submission and go into Held state. Transfer jobs fail with an error in stdout/stderr.

Steps to Reproduce

Start running a workflow that uses Globus for either job submissions or GridFTP for transfers, and then either delete the user's proxy, or create a new proxy with a very short lifetime.

Job does not produce an invocation record

Symptoms

Condor and DAGMan consider the job finished, but the stdout/stderr files for the job are empty, and there are no Kickstart invocation records.

Steps to Reproduce

Modify a job in a planned workflow so that a) it doesn't use a Kickstart wrapper, and b) it sends all of its stderr/stdout to /dev/null.

Executable is missing

Symptoms

The job fails with non-zero exit code and some sort of “file not found” error is present in stdout/stderr.

Steps to Reproduce

Modify the transformation catalog to point at the wrong executable.

Out of space

Symptoms

The job fails with non-zero exit code and there is a “no space left on device” error in the stdout/stderr.

Steps to Reproduce

Create a small file using dd, format it with a file system, and mount it via loopback. Then run a job that writes to the mounted file until it is full.

```
# dd if=/dev/zero of=/tmp/disk bs=1M count=10
# losetup /dev/loop0 /tmp/disk
# mkfs.ext3 /dev/loop0
# mkdir /tmp/mnt
# mount /dev/loop0 /tmp/mnt
# chmod 1777 /tmp/mnt

run job to write to /tmp/mnt/foo

# umount /tmp/mnt
# losetup --d /dev/loop0
# rm /tmp/disk
# rmdir /tmp/mnt
```

Expired or missing X.509 certificate

Symptoms

Transfer jobs fail with mysterious errors, GRAM jobs sit idle in the queue forever. GRAM jobs in the queue will have “detected down globus resource” errors in the job log.

Steps to Reproduce

Temporarily rename the grid certificate for the target site in /etc/grid-security/certificates.