HTCondor Transfers Scale Test

Edgar Fajardo and Ilan Scheinkman

October 15 2019

Original motivation

GLUEX experiment wanted to know if they could use the HTCondor Transfer File mechanism to bring the output of their jobs back.

GLUEX Expected numbers

After talking with Frank W and settling on some approximated numbers. This was the expected scenario:

Parameter	Value
# Parallel running jobs	20000
Output sandbox	10-100 MB
Input Sandbox	1-10 MB
Job length	8h - 9h

GLUEX Expected numbers

- The numbers on the last slide imply an expected rate of as follows:

$$O(n,l,s) = \frac{nJobs * size}{length} = \frac{20000 * 90}{9 * 3600} \approx 55.5 \frac{\text{MB}}{\text{sec}}$$

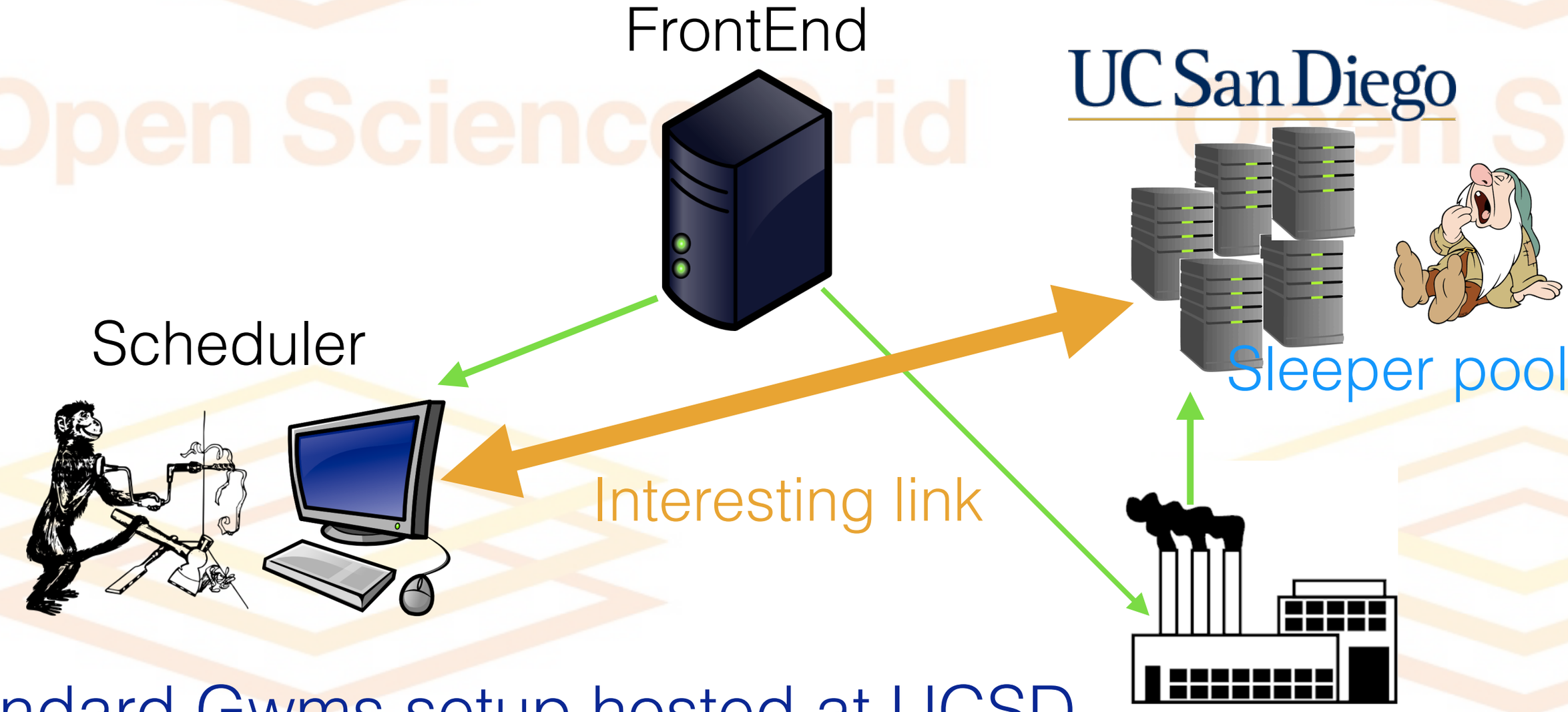
From worker node to scheduler

Can HTCondor handle this?

This question can be broken into more:

- Can HTCondor transfer saturate the network interface?
- How much CPU time is wasted on HTCondor Transfers?
- Does latency between Schedd and worker node has any effect on throughput?
- What is the consistency percentage of files transferred

How does the testbed look like?



Standard Gwms setup hosted at UCSD

How does the scheduler look like

Hardware	Spec
Memory	128 GB
Network Card	10 Gbit Full duplex
Core count	24
Disk Setup	Two SSD

Can HTCondor transfer saturate the network?

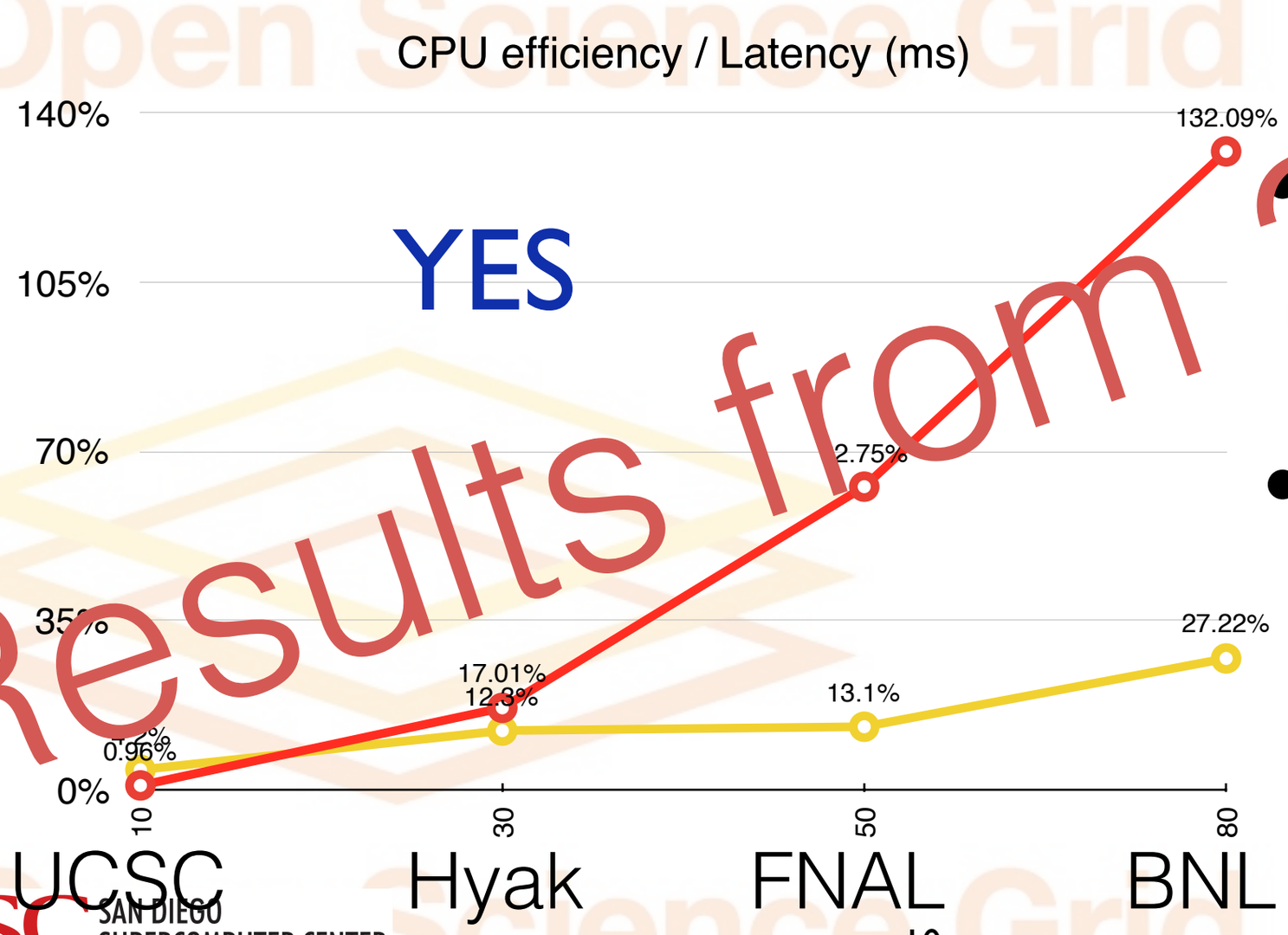
YES over two SSD:

Parameter	Value
# jobs	4000
out Size	250 MB
job length	30 min
Achieved throughput	~800 MB/s

This is **ten** times GLUEX needs

What is the cpu time lost due to transfers?

Does latency between Schedd and worker node has any effect on throughput?



YES

Results from 2107

- The red line is at two times the expected rate.
- The yellow line is at the expected rate.

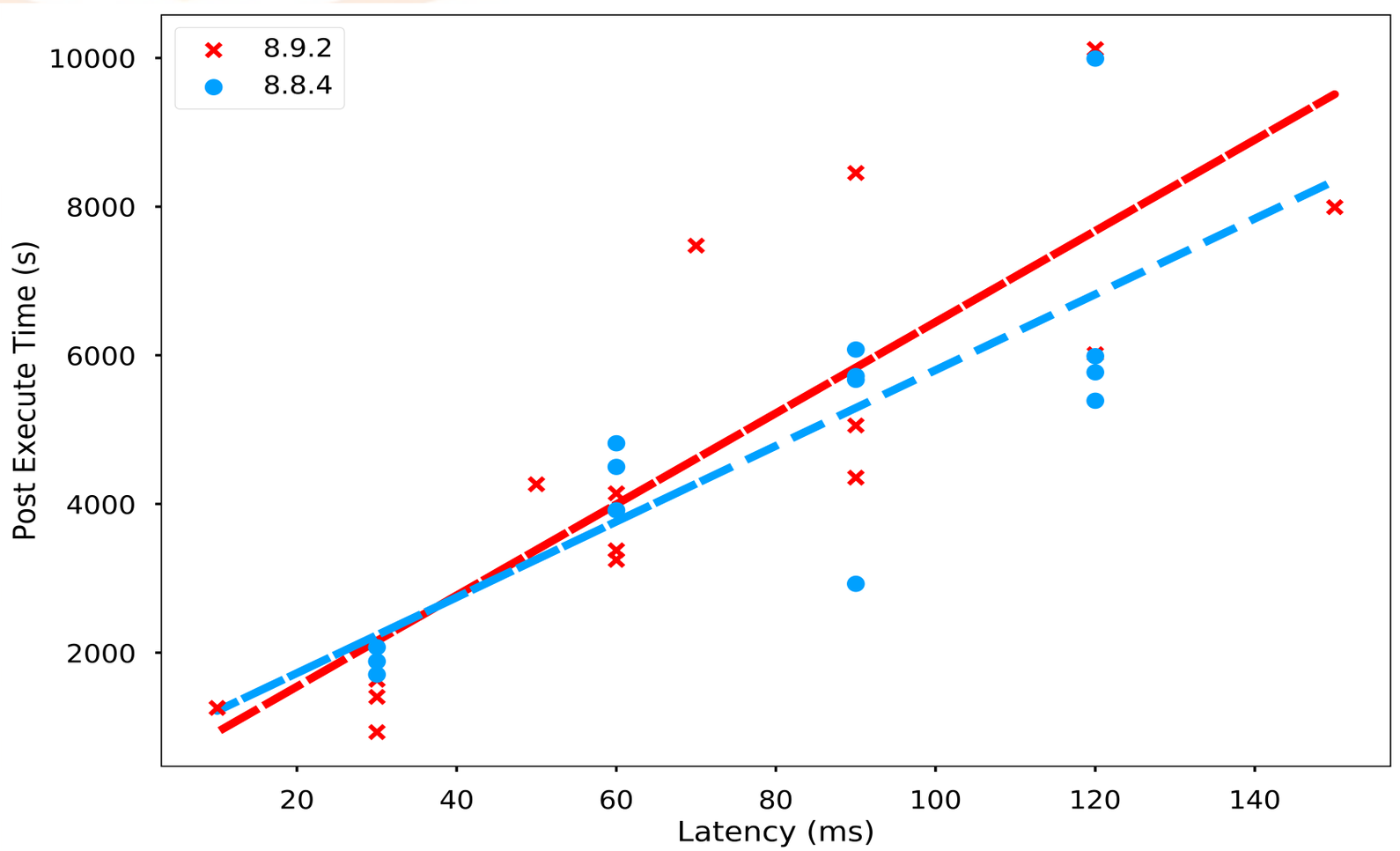
SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

So, What is new?

- During summer 2019 Ilan Scheinkman obtains the UCSD Physical Sciences Undergraduate Research Reward scholarship to redo tests more in-depth via GlideTester
  - This means a more precise measurements since we can control all jobs start at the same time and thus and more or less end at the same time.
  - We wanted to test new versions of HTCondor (8.8 and 8.9)
  - We could test transfer consistency

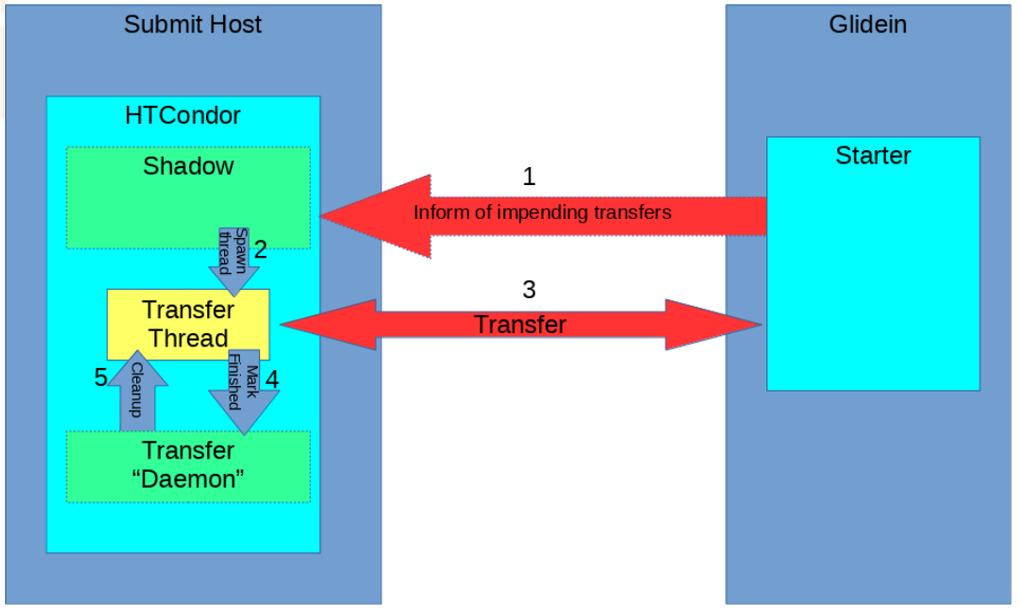
Results



Latency and Post execution (transfer) time are linearly correlated

Hyphoteses

- Starter informs Shadow of filenames and sizes needed to be transferred.
- Shadow creates new thread for each file's download via the File Transfer daemon
- The files are transferred 64 KB at a time (with no syncing)
- If the Glidein is disconnected mid-transfer, the Shadow marks the download thread for reaping and attempts a reconnect
- On reconnect, a new download thread is created
- When finished, the thread is marked for reaping and both Shadow and Glidein ads are updated



Why are starters disconnected?

StarterLog

```
12:27:23 Initializing a VANILLA shadow for job 141.121
12:27:24 (141.121) (1959362): Request to run on glidein_32837_33783684@cabinet-0-0-2.t2.ucsd.edu <169.228.
12:27:25 (141.121) (1959362): File transfer completed successfully.
12:48:37 (141.121) (1959362): condor_read() failed: recv(fd=7) returned -1, errno = 104 Connection reset b
12:48:37 (141.121) (1959362): IO: Failed to read packet header
12:48:37 (141.121) (1959362): Can no longer talk to condor_starter <169.228.131.252:34901>
12:48:37 (141.121) (1959362): Trying to reconnect to disconnected job
12:48:37 (141.121) (1959362): LastJobLeaseRenewal: 1565034132 Mon Aug  5 12:42:12 2019
12:48:37 (141.121) (1959362): JobLeaseDuration: 2400 seconds
12:48:37 (141.121) (1959362): JobLeaseDuration remaining: 2015
12:48:37 (141.121) (1959362): Attempting to locate disconnected starter
12:48:37 (141.121) (1959362): Found starter: <169.228.131.252:33588?CCBID=169.228.130.47:9630%3faddrs%3d16
12:48:37 (141.121) (1959362): Attempting to reconnect to starter <169.228.131.252:33588?CCBID=169.228.130.
12:49:08 (141.121) (1959362): CCBClient: Timed out waiting for response after requesting reversed connecti
12:49:08 (141.121) (1959362): Failed to reverse connect to starter at <169.228.131.252:33588> via CCB.
12:49:08 (141.121) (1959362): Attempt to reconnect failed: Failed to connect to starter <169.228.131.252:3
12:49:08 (141.121) (1959362): JobLeaseDuration remaining: 2369
12:49:08 (141.121) (1959362): Scheduling another attempt to reconnect in 8 seconds
12:49:16 (141.121) (1959362): Attempting to locate disconnected starter
12:49:19 (141.121) (1959362): Found starter: <169.228.131.252:33588?CCBID=169.228.130.47:9630%3faddrs%3d16
12:49:19 (141.121) (1959362): Attempting to reconnect to starter <169.228.131.252:33588?CCBID=169.228.130.
12:49:49 (141.121) (1959362): CCBClient: Timed out waiting for response after requesting reversed connecti
12:49:49 (141.121) (1959362): Failed to reverse connect to starter at <169.228.131.252:33588> via CCB.
12:49:49 (141.121) (1959362): Attempt to reconnect failed: Failed to connect to starter <169.228.131.252:3
12:49:49 (141.121) (1959362): JobLeaseDuration remaining: 2328
12:49:49 (141.121) (1959362): Scheduling another attempt to reconnect in 16 seconds
---- (Extra reconnect attempts cut for space) ----
12:55:56 (141.121) (1959362): Scheduling another attempt to reconnect in 256 seconds
13:00:12 (141.121) (1959362): Attempting to locate disconnected starter
13:00:12 (141.121) (1959362): Found starter: <169.228.131.252:33588?CCBID=169.228.130.47:9630%3faddrs%3d16
13:00:12 (141.121) (1959362): Attempting to reconnect to starter <169.228.131.252:33588?CCBID=169.228.130.
13:00:12 (141.121) (1959362): Reconnect SUCCESS: connection re-established
13:00:12 (141.121) (1959362): ERROR "FileTransfer::Download called during active transfer!" at line 1663 i
13:00:12 Initializing a VANILLA shadow for job 141.121
13:00:12 (141.121) (1964759): Request to run on glidein_32837_33783684@cabinet-0-0-2.t2.ucsd.edu <169.228.
13:00:12 (141.121) (1964759): Job 141.121 is being evicted from glidein_32837_33783684@cabinet-0-0-2.t2.uc
13:00:12 (141.121) (1964759): logEvictEvent with unknown reason (108), not logging.
13:00:12 (141.121) (1964759): **** condor_shadow (condor_SHADOW) pid 1964759 EXITING WITH STATUS 108
13:00:17 Initializing a VANILLA shadow for job 141.121
13:00:17 (141.121) (1964765): Request to run on glidein_49306_396784080@sdsc-9.t2.ucsd.edu <169.228.132.10
13:00:17 (141.121) (1964765): Job 141.121 is being removed: The job attribute PeriodicRemove expression '(
13:00:17 (141.121) (1964765): Job 141.121 is being removed: The job attribute PeriodicRemove expression '(
13:00:17 (141.121) (1964765): DoUpload: SHADOW at 169.228.132.175 failed to send file(s) to <169.228.132.1
13:00:17 (141.121) (1964765): File transfer failed (status=0).
13:00:18 (141.121) (1964765): Job 141.121 is being evicted from glidein_49306_396784080@sdsc-9.t2.ucsd.edu
13:00:18 (141.121) (1964765): logEvictEvent with unknown reason (113), not logging.
13:00:18 (141.121) (1964765): **** condor_shadow (condor_SHADOW) pid 1964765 EXITING WITH STATUS 113
13:00:34 (141.121) (1959362): condor_read() failed: recv(fd=5) returned -1, errno = 104 Connection reset b
13:00:34 (141.121) (1959362): ReliSock::get_bytes_nobuffer: Failed to receive file.
13:00:34 (141.121) (1959362): get_file(): ERROR: received 786432 bytes, expected 104857600!
13:00:34 (141.121) (1959362): DoDownload: SHADOW at 169.228.132.175 failed to receive file /data2/ilan_gli
13:00:34 (141.121) (1959362): condor_write() failed: send() 251 bytes to <169.228.131.252:46189> returned
13:00:34 (141.121) (1959362): Buf::write(): condor_write() failed
13:00:34 (141.121) (1959362): Failed to send download failure report to <169.228.131.252:46189>.
```

```
022 (141.121.000) 08/05 12:48:37 Job disconnected, attempting to reconnect
    Socket between submit and execute hosts closed unexpectedly
    Trying to reconnect to glidein_32837_33783684@cabinet-0-0-2.t2.ucsd.edu <169.228.131.252:41860?CCBID=
...
028 (141.121.000) 08/05 12:48:37 Job ad information event triggered.

023 (141.121.000) 08/05 13:00:12 Job reconnected to glidein_32837_33783684@cabinet-0-0-2.t2.ucsd.edu
    startd address: <169.228.131.252:41860?CCBID=169.228.130.47:9621%3faddrs%3d169.228.130.47-9621#560298
    starter address: <169.228.131.252:33588?CCBID=169.228.130.47:9630%3faddrs%3d169.228.130.47-9630#55761
...
028 (141.121.000) 08/05 13:00:12 Job ad information event triggered.
...
007 (141.121.000) 08/05 13:00:12 Shadow exception!
  FileTransfer::Download called during active transfer!
  0  -  Run Bytes Sent By Job
  2700  -  Run Bytes Received By Job
```

ShadowLog

Other Hypotheses

- We see from the logs the TCP disconnects but we do not know why. It only happens when adding both latency and high concurrency.
- We add the artificial latency to the submit host using the tc command but we could be adding latency to HTCondor daemons on the same machine which talk on the public interface.
- Does anyone know how to add latency but only IP based?

Conclusions

- At the proposed rates GLUEX will efficiently use its resources bringing their output through the HTCondor transfer mechanism
- It also holds true if they double the rates
- Latency greatly influences the efficiency of the HTCondor transfer mechanism.
- HTCondor never missed a single checksum on file transfer even under heavy stress.

Future Work

- We want to test comparing same transfers using HTTP.
- One of LIGO's frameworks (bayeswave) already running on the grid via OSG is aiming to use HTCondor transfer system for brining their output data. For O3 and O4 that could multiple times ten the needs of Gluex.

Acknowledgements

- HTCondor Dev team for looking at our data and very eager to solve and test different issues.
- The different funding agencies that sponsored this work
- To the UCSD Physics division for the summer grant.
- UCSD T2 Admins for providing the infrastructure for the tests.

Questions

1-800-Condor-Masters