LArSoft - Support #24185

Unusually long run time for first lar command compared with subsequent commands.

03/13/2020 02:26 PM - Patrick Green

Status:	Assigned	**Start date:**	03/13/2020
Priority:	Normal	**Due date:**	
Assignee:	Patrick Gartung	**% Done:**	0%
Category:	Usability	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.50 hour
Experiment:	-	**Co-Assignees:**	

Description

I am seeing an odd pattern in run times when running the same larsoft command multiple times sequentially - the first lar command takes significantly longer than all subsequent ones. For example, using sbndcode v08_36_01_3_MCP2_0 and running that standard sbnd prodcorsika_cosmics_proton.fcl on a gpvm the first lar command takes ~23s, then all subsequent calls take ~6 seconds. This pattern is then repeated if I exit and ssh into the gpvm again. The extra runtime is immediately after the lar command is called before any other couts, when it is presumably loading in the libraries. A similar effect has also been seen running other fcls in e.g. uboonecode, so this appears more general than just sbnd.

This also occurs and is significantly more impactful when running on Theta (at ALCF). The plot attached shows the runtimes of jobs running the same sbnd corsika fcl (now in a singularity container) running as many individual event jobs as possible on 448 cores for an hour. The runtimes of the first job on each core (the first 448) are much longer (~470s per job) than all the subsequent jobs (~70s per job). These are all entirely separate jobs encased in a new singularity environment each time, so should not effect each other in any way. Since our primary way of scaling on Theta is to run individual events across as many cores as possible this scales very poorly since almost all of our jobs are the first lar command on that core and have this long load time.

What is larsoft doing during this period? Is this something that could be either avoided or sped up?

History

#1 - 03/13/2020 04:32 PM - Gianluca Petrillo

How long does it need to pass before this long time is back?
I mean: you run lar (1'), lar (30"), lar (30"), lar (30")...
How long do you have to wait before a lar command, to get it back to 1'?

I am asking because the behaviour you describe is compatible with some cache being filled (the best candidate is CVMFS, where the libraries are stored), and the turnover time may exclude some of these.

Also, chances are that running lar with strace give you hints of what is happening.

#2 - 03/13/2020 05:53 PM - Patrick Green

- File first_run.txt added

- File second_run.txt added

"How long does it need to pass before this long time is back?"
This is not something I have looked at, I will test whether it reverts back to the slow running after some time. Previously it has always been closing the ssh session and
reconnecting.

In the case of running on Theta we are not using CVMFS, instead we have all the binaries copied onto Theta via pullproducts and are running these using singularity
container of Fermilabs sl7.

I have just tried running the strace, however the output of this doesn't mean much to me unfortunately. I have attach text files containing the output from running this twice (a slow case, "first_run.txt" and a subsequent fast case, "second_run.txt"). Could you take a look and see if this tells us anything about what is happening?

#3 - 03/16/2020 10:02 AM - Patrick Green

Hi Gianluca, below are the results from running prodcorsika_cosmics_proton.fcl in sbndcode repeatedly with increasing sleeps between the commands. The run time reverted back to the slower speed after a 6 hour sleep, but had not reverted after the previous 2 hour sleep - so it takes somewhere between 2-6 hours to reset, I can further refine that if that is useful.

initial lar command:

19 s
sleep 30s:
7 s
sleep 2m:
7 s
sleep 5m:
7 s
sleep 15m:
8 s
sleep 30m:
7 s
sleep 2h:
8 s
sleep 6h:
16 s

#4 - 03/16/2020 10:36 AM - Christopher Green

- Assignee set to Patrick Gartung

- Status changed from New to Assigned

- Category set to Usability

- Tracker changed from Bug to Support

#5 - 03/16/2020 12:59 PM - Patrick Gartung

A naive examination of the strace files does not shows a significant difference is the files accessed. My initial guess that ROOT was reconstructing the PCH files is ruled out by this.

My educated guess is that you are hitting the initial caching of the filesystem reads or in the case of Theta the first access occurs without the benefit of the on chip HBM cache.

Files

runtime_starttime_withserialmodeaffinity.png	32.2 KB	03/13/2020	Patrick Green
second_run.txt	11.8 MB	03/13/2020	Patrick Green
first_run.txt	11.8 MB	03/13/2020	Patrick Green