**MPI HPC**

1.. MW: I've been talking to Onno (cc'ed) and Pete about Firedrake and running

in parallel on ARC3 or ARC4. Will singularity allow us to run across multiple nodes and cores with MPI? What is the best solution for us to request on ARC as regards regular updates as well as parallel execution?

There are several of us who would want access to a solution - Onno's group are ahead in that they could be using it much sooner. If Singularity is the best solution is there any significant overhead?  The Firedrake team vaguely alluded to a performance overhead but equally I do not think they have much experience with singularity. If this is better as a general IT request please let me know and I'll put one in. I thought maybe if we asked informally first we could make a more specific request.

*OB: Where do I start with respect to ARC (have not logged on or used it yet but now have some FD code coming up for usage), please?*

*MW: John gave the basic outline of how he ran Firedrake on ARC4. I have tested it on one core interactively with a tutorial example and my code also ran. To run larger/longer jobs we should use the job scheduler and submit the runs to the queue.*

[*https://arcdocs.leeds.ac.uk/usage/batchjob.html*](https://arcdocs.leeds.ac.uk/usage/batchjob.html)

*I don’t know how many cores you expect to use and how long a run may be. Firedrake documentation suggests that about 50000 unknowns per core is OK for performance if you want an estimate. Jobs can run for up to 48 hrs I think.*

2.. JH: My thumb in air feeling on this (alongside papers I've read) is that performance issues of tools like Apptainer/Singularity are often greatly exaggerated.

3.. MW: I had seen that approach for MPI/Singularity in another example so it would be good to know if it works. Has Apptainer now superceded Singularity or do we have both? Are they different or is it just renamed?

*OB: What is Apptainer?*

*MW: Apptainer is the newest version of Singularity. It is simply a rename and I think they can be used interchangeably at present. At some point the ‘singularity’ command will be removed.*

4.. JH: Typically the best way to run software with MPI and containers is to have a compatible MPI runtime within the container, and start it up with mpirun outside of the container.  So something like:

mpirun apptainer run example.sif python somecode.py

Would need to have a play with firedrake to see whether that's feasible withe the existing images for it.

*OB: I am lost here; what is JH saying? From whether does one run FD or have the FD installation, on Arc or locally? Is there one FD install on Arc which several people use, which is updated or does all users fend for themselves? That was our original question, was it not?*

MW: We have to create the Firedrake container locally ourselves using singularity. This builds the environment within which we run Firedrake. When we do this we have to accept whatever version of MPI the Firedrake team used to create the Firdrake image (the Docker file). Ideally this should match the best version of MPI available on ARC4 (there are several, eg. OpenMPI, MPICH, … ). If the version is not the “best” it will still run in parallel. My opinion is that unless we need huge amounts of cores and absolute peak performance this should not be an issue.

Using Firedrake requires 2 stages:

1. We build the Firedrake application from the Docker image locally using Singularity to create a “.sif” file (Singularity image file). This only need doing once, and after that only if we want to update Firdrake. The Docker image has a date of release so we can track if it is updated.
2. We run Firedrake using the job scheduler and Singularity as per point 6. below.

5.. MW: I had seen that approach for MPI/Singularity in another example so it would be good to know if it works. Has Apptainer now superceded Singularity or do we have both? Are they different or is it just renamed?

6.. JH: Have you got a suitable example parallel firedrake script I could test with?  Doesn't need to be anything elaborate. I don't think this looks too painful.  I've just tried with what I think I told you to use last time, but rather than running firedrake in parallel, I built another hello world executable with the mpich within the container, and ran it with mvapich2 on the outside (as that has a compatible launcher):

#!/bin/bash

#$ -P feps-cpu

#$ -l nodes=2

#$ -cwd -V

#$ -l h\_rt=0:10:0

module swap openmpi mvapich2

module add apptainer

mpiexec singularity exec --env 'PATH=/home/firedrake/firedrake/bin:$PATH' -B /run -B /nobackup -B ~/.cache:/home/firedrake/firedrake/.cache firedrake\_latest.sif ./hello

That ran the MPI program within the container just fine, so I'd hope we can do the same thing with firedrake.

As to apptainer vs singularity... a mix of things, with politics involved.  Singularity was created at Lawrence Berkeley, but then was spun off into a company Sylabs.  That led to a commercial and an open source fork.  Then there was a falling out, with one of the core developers leaving Sylabs, and apptainer was created, which has been adopted by the Linux Foundation, with I think likely gives it a very healthy future.

Big feature that came with recent Apptainer is the ability to build containers as a regular user without sudo or any other trickery.  This means you can build and tweak containers as a normal user without needing to use VMs or have anything else setup by IT.  They left in all the old singularity commands, so you can carry on using it exactly as you were before and it should work just fine.  I'd expect to remove singularity from ARC in the near future, given Apptainer really does replace it.

7.. MW: This is a simple example from the Firedrake team

<https://www.firedrakeproject.org/demos/parprint.py>  
It gives some basic feedback on the distribution of data

to each core.

8.. MW: Was that test on ARC4?

9.. JH: Yes.  Would probably work on ARC3 with singularity instead of apptainer, but haven't tested.

8.. MW: Thanks Onno. The difference may not be that big but it is worth doing some testing and timing. It will depend on how many cores you want to run on.  John mentioned infiniband, which is important if you want to run on more than one node of ARC4. Each node is 40 cores so up to that number you are within a node and infiniband is not relevant.

**Extra:**

**From:** Jack Betteridge <notifications@github.com>  
**Sent:** Tuesday, January 3, 2023 11:37 AM  
**To:** firedrakeproject/firedrake <firedrake@noreply.github.com>  
**Cc:** Onno Bokhove <O.Bokhove@leeds.ac.uk>; Author <author@noreply.github.com>  
**Subject:** Re: [firedrakeproject/firedrake] pytest fails after FD seems installed (Discussion #2652)

There is a guide to using Singularity/Apptainer using the Docker image as a starting point here: [https://github.com/firedrakeproject/firedrake/wiki/singularity](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fgithub.com%2Ffiredrakeproject%2Ffiredrake%2Fwiki%2Fsingularity&data=05%7C01%7CO.Bokhove%40leeds.ac.uk%7Ca170a363b27d468caf1908daf875befd%7Cbdeaeda8c81d45ce863e5232a535b7cb%7C1%7C0%7C638095481993407626%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=vgo%2FknQhtYMziKYMOzlvaFYpZeZ6eF5qo3T6kVSYyjQ%3D&reserved=0)

It has been tested on ARCHER2, but due to various issues in scaling tests I don't have performance numbers. The biggest factor in performance will be getting the container to use the host MPI distribution. Feel free to contact me if you have questions about setting this up.