CONTAINERIZED BIOINFORMATICS APPLICATIONS FOR HPC

Yucheng Zhang, Senior Life Science Scientist

ACCESS | Advancing Innovation

NSF

PURDUE UNIVERSITY | Rosen Center for Advanced Computing

What to expect from this bioinformatics workshop?

Objectives

- What are containers and why should we use them
- How to use singularity to pull, run and build containers
- Containerized bioinformatics applications deployed on Anvil

Containerized Bioinformatics applications for HPC

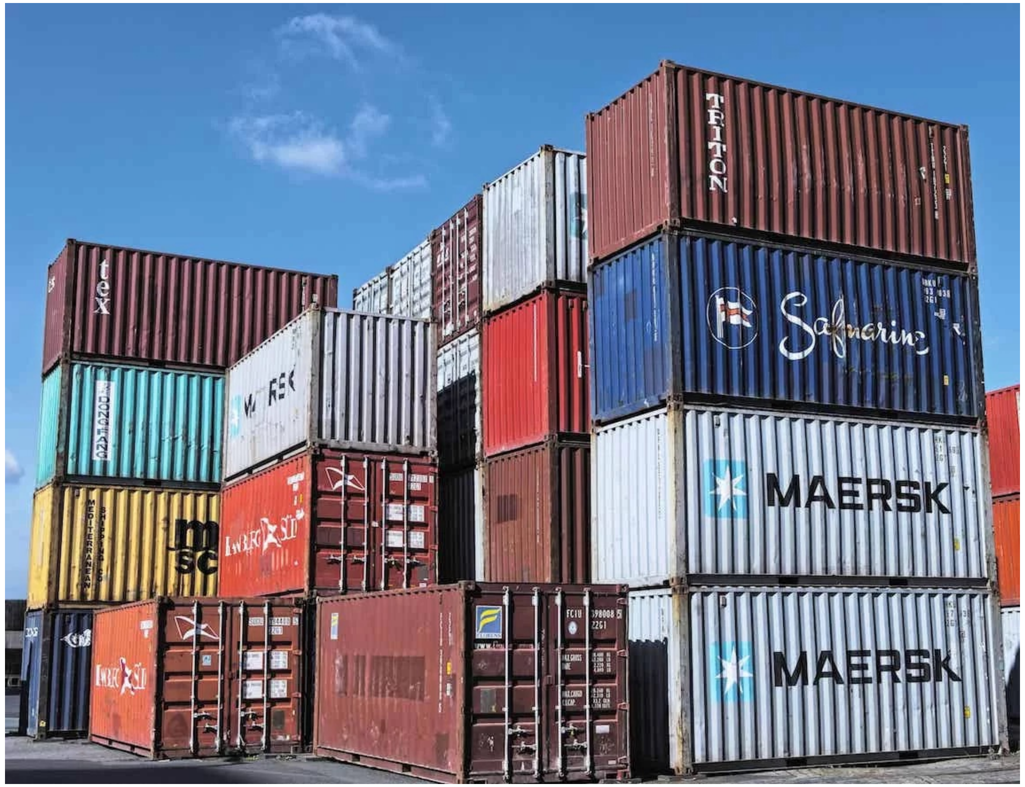
Containers

What are containers?

❖ The arrival of modern shipping containers changed our transportation industry.

❖ Container is a standardized way to package items together into one shipment.

1. Standard packaging
2. Isolation and efficiency
3. Separation of concerns
4. Portable



Containers

A **container** is an abstraction for a set of technologies that aim to solve the problem of how to get software to run reliably when moved from one computing environment to another.

A container **image** is simply a file (or collection of files) saved on disk that stores everything you need to run a target application or applications.

Registry: a place to store (and share) container images.

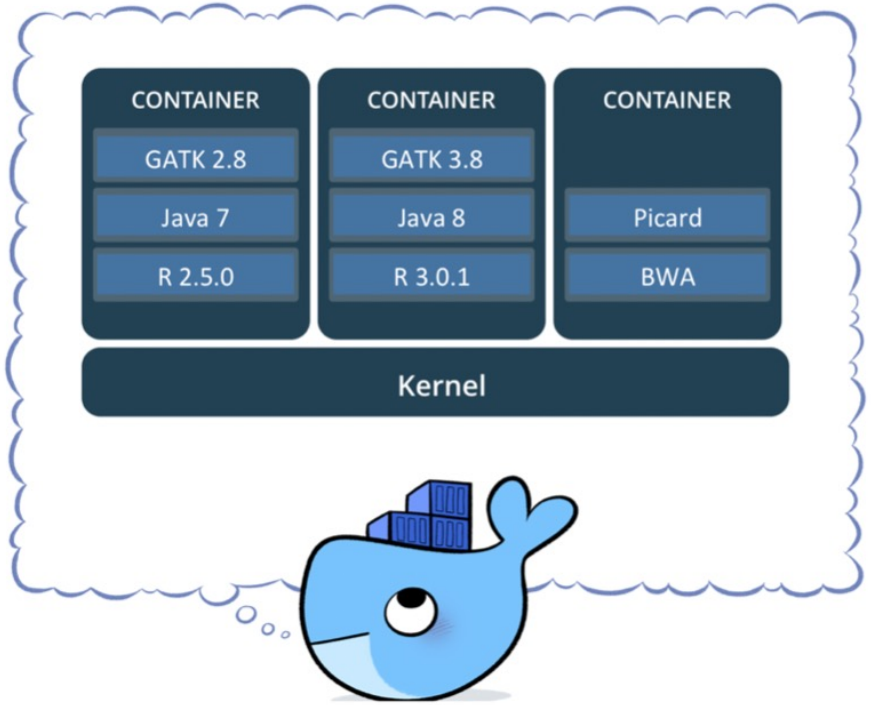
docker hub

GitHub Container Registry

BioContainers

Why use containers

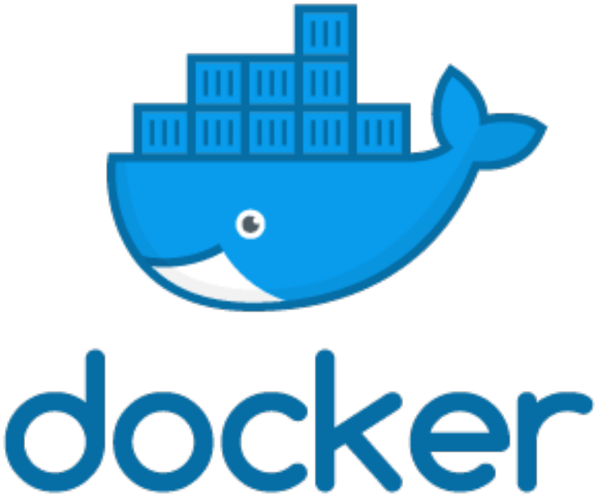
- **Getting organized:** containers keep things organized by isolating programs and their dependencies inside containers.
- **Build once, run almost anywhere:** containers allow us to package up our complete software environment and ship it to numerous operating systems.
- **Reproducibility:** containers can ensure identical versions of apps, libraries, compliers, etc.



Docker

The concept of containers emerged in 1970s, but they were not well known until the emergence of Docker containers in 2013.

Docker is an open source platform for building, deploying, and managing containerized applications.



Some concerns about the security of Docker containers in HPC: Docker gives superuser privileges, but we do not want users to have full, unrestricted admin/ root access.

Adobe
ARROW
AT&T
Blue Apron
INTEGRATE
Lucent Health
paloalto NETWORKS
Bitbucket_Full
PathFactory
PayPal
pipedrive
Segment
snowflake
spiceworks
splunk>
stripe
TARGET
The Container Store
UNIVERSITY OF CALGARY
verizon
VIRGINIA TECH
Yale

Singularity

- Singularity was developed in 2015 as an open-source project by researchers at Lawrence Berkeley National Laboratory led by Gregory Kurtzer.

- Singularity is emerging as the containerization framework of choice in HPC environments.
  1. Enable researchers to package entire scientific workflows, libraries, and even data.
  2. **Can use docker images.**
  3. **Does not require root privileges to run.**

apptainer

- Singularity was recently renamed to Apptainer and hosted by the Linux Foundation.
- Apptainer uses the command apptainer to replace the previous command singularity.
- The singularity command also works, because it is alias for the command apptainer.
- The variable prefixes are APPTAINER_ and APPTAINERENV_ instead of the previous SINGULARITY_ and SINGULARITYENV_.

Containerized Bioinformatics applications for HPC

Singularity basics

singularity pull

Download or build a container from a given URI.

singularity pull [output file] URI

Example: **singularity pull** blast_2.13.0.sif **docker://staphb/blast:2.13.0**

custom name *URL*

Supported URIs include:

❖ **library**: Pull an image from the currently configured library

library://user/collection/image[:tag]

❖ **docker**: Pull a Docker/OCI image from Docker Hub, or another OCI registry.

docker://user/image[:tag]

❖ **http, https**: Pull an image using the http(s?) protocol

https://depot.galaxyproject.org/singularity/hisat2%3A2.2.1--he1b5a44_2

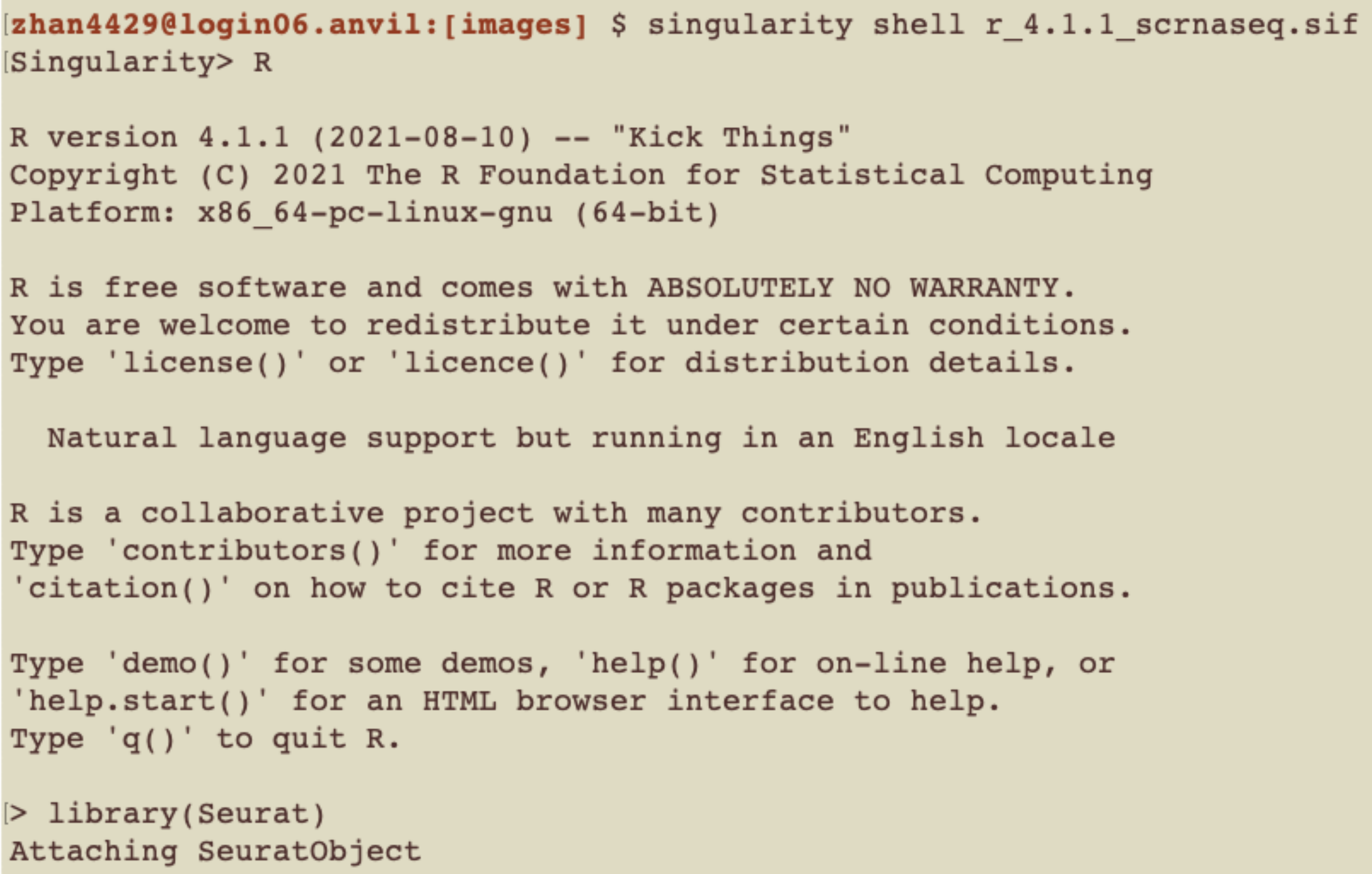
Recommended container registries for bioinformatics

- **DockerHub** (https://hub.docker.org)
  - The largest repository of Docker container images.

- **Biocontainers** (https://biocontainers.pro/registry)
  - A community-driven project for bioinformatics containers.
  - 10.6K tools,45.5K versions,228.5Kcontainers and packages (As of March, 2023).

singularity shell

Users can go inside the container to run interactive commands



```
[zhan4429@login06.anvil:[images] $ singularity shell r_4.1.1_scrnaseq.sif
[Singularity> R

R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[> library(Seurat)
Attaching SeuratObject
```

singularity exec

A container may contain many executables/scripts.

singularity exec can be used to select which executable/script to run.

singularity exec image.sif command

For example:

```
singularity exec blast.2.13.0.sif blastn \
      -query nucleotide.fasta \
      -db nt -out blastn.out

singularity exec blast.2.13.0.sif blastp \
     -query protein.fasta \
     -db nr -out blastp.out
```

singularity build

Build using a singularity definition file

```
sudo singularity build image.sif definition.def
```

If you really need write access to a container you can use a writable sandbox.

```
sudo singularity build --sandbox container-sand definition.def
sudo singularity shell --writable container-sand
Singularity> apt-get update
Singularity> apt-get install –y packageName
sudo singularity build container.sif container-sand
```

apptainer build

Regular users don't have need sudo privilege to build containers.

```
$ sudo singularity build
```

```
$ singularity --version
3.8.5-2.el8
```

Anvil, Brown, Bell, Gilbreth,
Scholar, Workbench

```
$ singularity build
$ apptainer build
```

```
$ singularity --version
apptainer version 1.1.6-1
```

Negishi

Definition files

A **definition** file, or **def** file, is a recipe to build a container image with singularity. It is divided into two parts:

1. **Header**: the Header describes the core operating system to build within the container.
   - Bootstrap
   - From
2. **Section**: each section is defined by a **%** character followed by the name of the particular section. Different sections add different content or execute commands at different times during the build process.
   - help
   - setup
   - files
   - labels
   - **environment**
   - **post**
   - runscript
   - …

Bootstrap

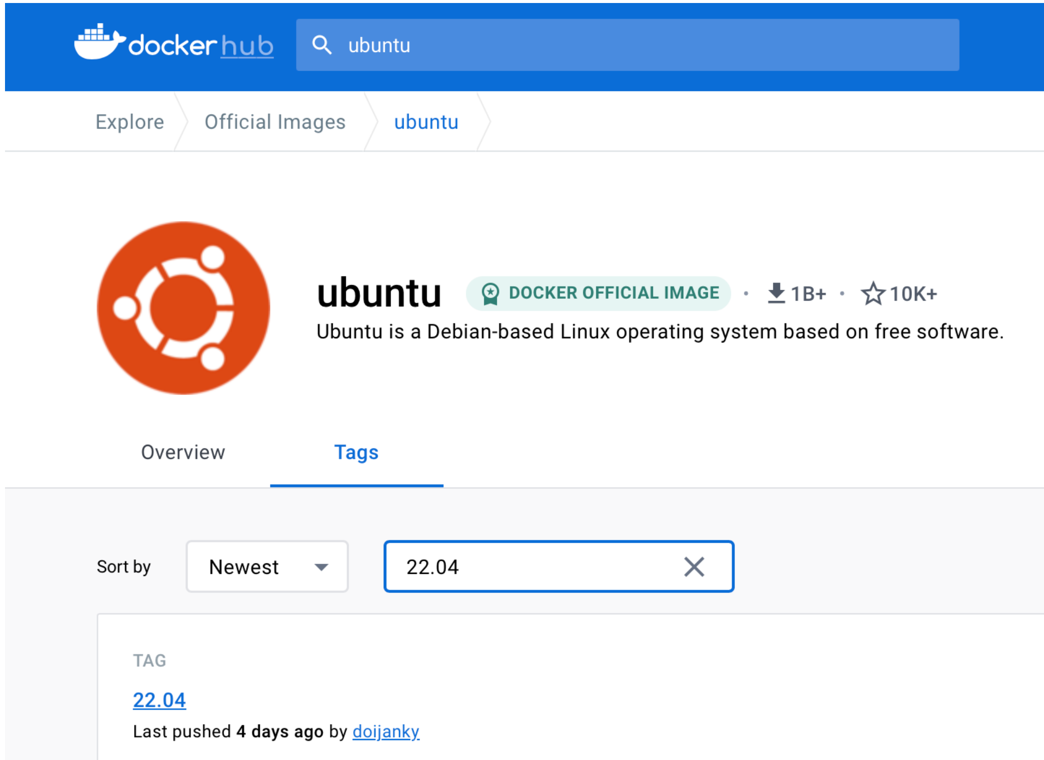
References the kind of base you want to use (e.g., docker, debootstrap, shub).

Images hosted on Docker Hub

```
Bootstrap: docker
From: ubuntu:22.04
```

Images saved on your machine

```
Bootstrap: localimage
From: /apps/biocontainers/images/mamba.sif
```



ubuntu DOCKER OFFICIAL IMAGE · 1B+ · 10K+

Ubuntu is a Debian-based Linux operating system based on free software.

Sort by: Newest

22.04

TAG

22.04

Last pushed **4 days ago** by doijanky

Anatomy of a definition file: Reciprocal Smallest Distance

```
Bootstrap: docker
From: ubuntu:18.04

%post
# Update and install system libraries
apt-get -y update
apt-get -y install --no-install-recommends --no-install-suggests build-essential libssl-dev wget

# KALIGN 2.0.4
cd /opt  && mkdir kalign2
cd kalign2 && wget http://msa.sbc.su.se/downloads/kalign/current.tar.gz
tar -xvf current.tar.gz && ./configure && make

# RSD
cd /opt && git clone https://github.com/todddeluca/reciprocal_smallest_distance
cd reciprocal_smallest_distance
python setup.py install

%environment
export PATH=/opt/kalign2:/opt/reciprocal_smallest_distance/bin:$PATH
```

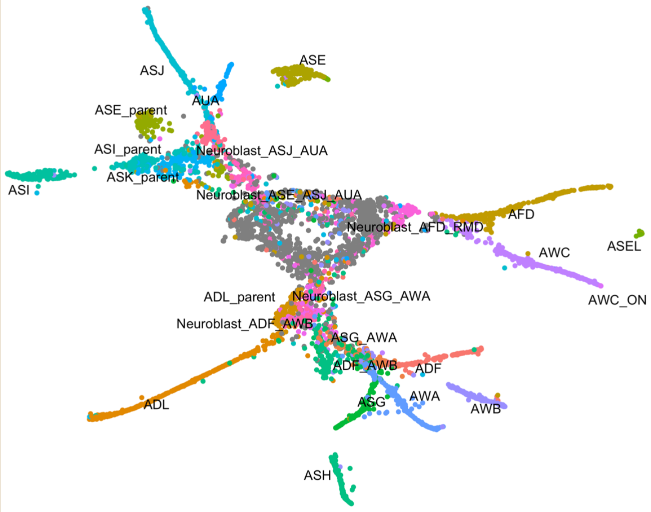
Anatomy of another definition file: phylofisher

```
Bootstrap: localimage
From: /apps/biocontainers/images/mamba.sif

%labels
Author "Yucheng Zhang <zhan4429@purdue.edu>"

%help
This container contains phylofisher version 1.2.5.

%post
mamba install -c bioconda -c conda-forge \
      python=3.7.10 phylofisher=1.2.5
```

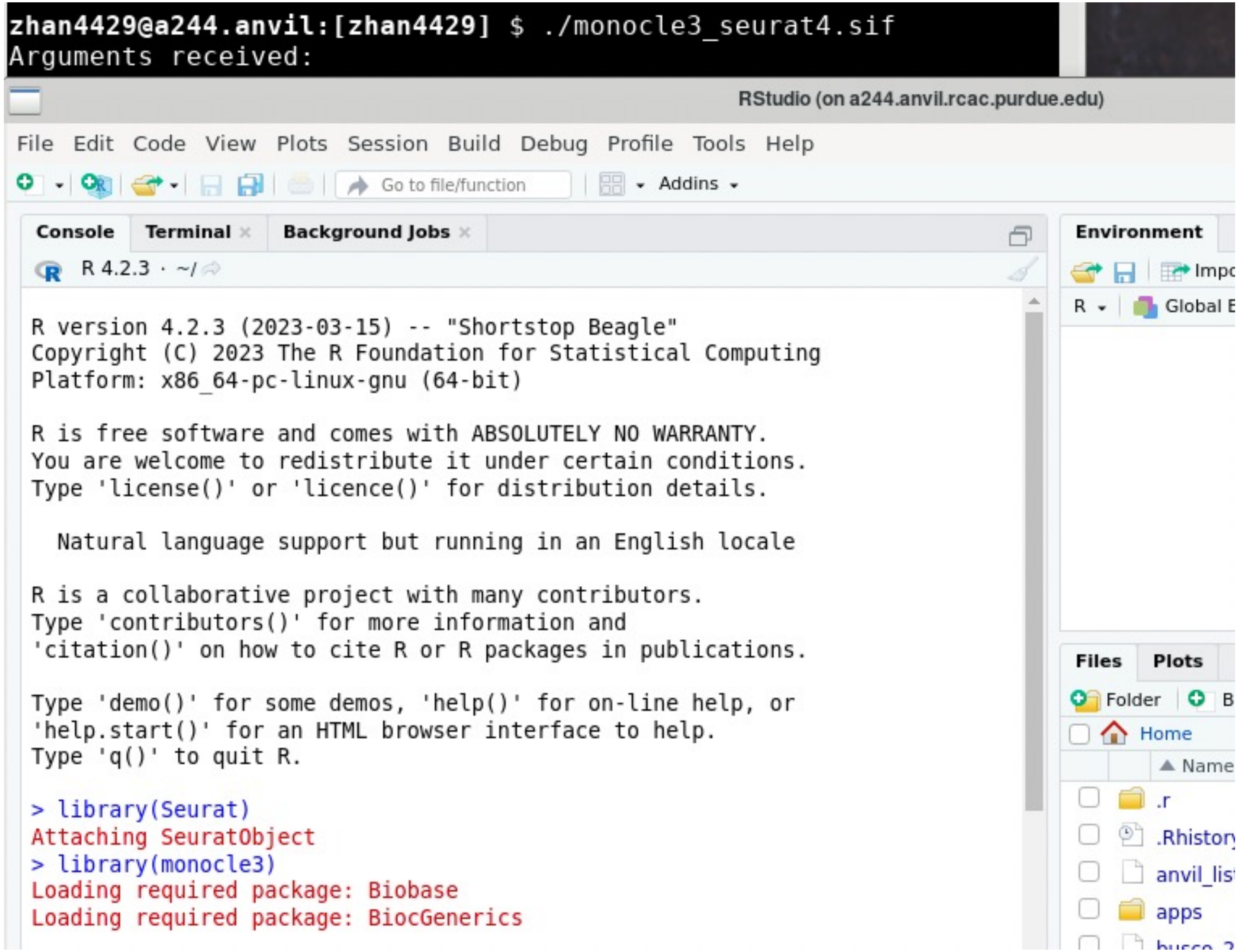
```
sudo singularity build phylofisher_1.2.5.sif phylofisher_1.2.5.def    #singularity
singularity build phylofisher_1.2.5.sif phylofisher_1.2.5.def       #apptainer
```

Anatomy of the 3rd definition file: R/Rstudio

```
Bootstrap: localimage
From: /apps/biocontainers/images/r4.2.3_rstudio2023.sif

%post
apt-get update
apt-get -y install libgdal-dev

## monocle3
Rscript -e "BiocManager::install(c('BiocGenerics', 'DelayedArray’, \
                  'DelayedMatrixStats’, 'limma', 'lme4', 'S4Vectors’, \
                  'SingleCellExperiment’, 'SummarizedExperiment’, \
                  'batchelor', 'HDF5Array’, 'terra', 'ggrastr'))"
Rscript -e "devtools::install_github('cole-trapnell-lab/monocle3’)”
## Seurat
Rscript -e "install.packages('Seurat’)”

%runscript
rstudio “$@”
```



customized Rstudio app for scRNAseq



```
zhan4429@a244.anvil:[zhan4429] $ ./monocle3_seurat4.sif
Arguments received:
```

```
R version 4.2.3 (2023-03-15) -- "Shortstop Beagle"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(Seurat)
Attaching SeuratObject
> library(monocle3)
Loading required package: Biobase
Loading required package: BiocGenerics
```

Bind host directories

- Programs running inside a container will not have access to directories and files outside of your home and the current directory.

- Singularity allows you to map directories on your host system to directories within your container using bind mounts.

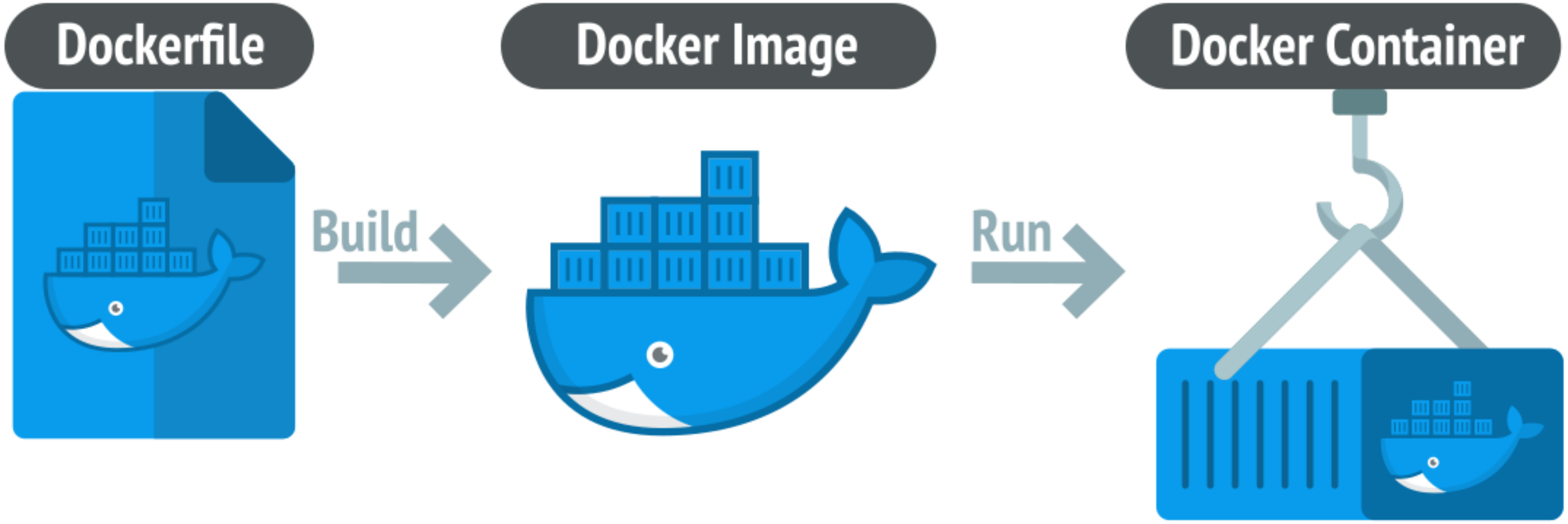
```
singularity shell/run/exec --bind hostdir:containerdir image.sif
```

Singularity binds several directories into the container image automatically. **$HOME**, **/tmp** and **$PWD** is the default list.
We also configured singularity to bind **/apps**, **/depot**, and **/scratch** on Purdue community clusters, to bind **/anvil**, and **/apps** on ACCESS Anvil.

Docker build

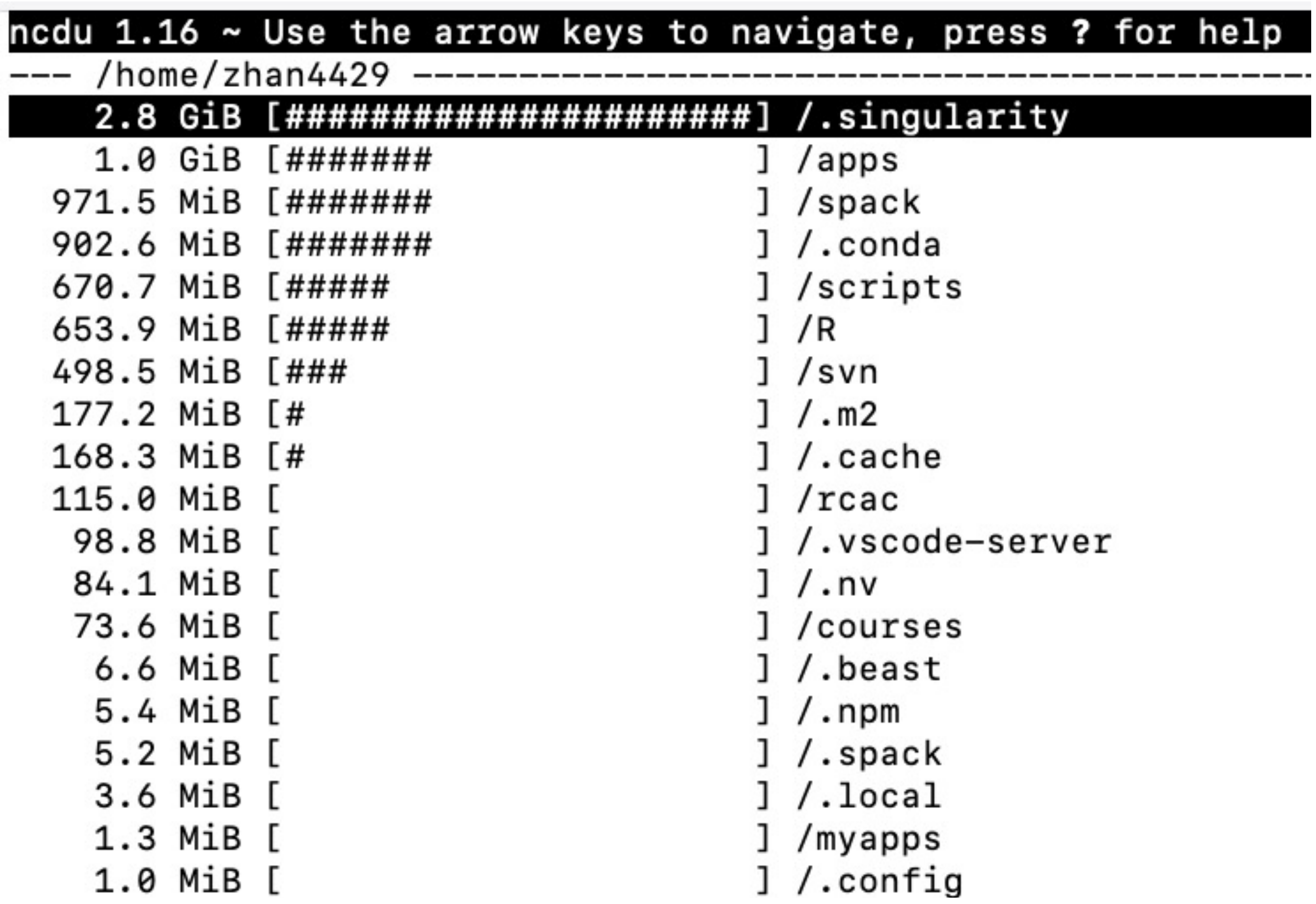
Using docker to build containers is another option:

1. Docker has a large, active, and stable ecosystems of container images.
2. Singularity can use docker images.



cache

$ ncdu $HOME



```
ncdu 1.16 ~ Use the arrow keys to navigate, press ? for help
--- /home/zhan4429 -------------------------------------------
    2.8 GiB [######################] /.singularity
    1.0 GiB [#######               ] /apps
  971.5 MiB [#######               ] /spack
  902.6 MiB [#######               ] /.conda
  670.7 MiB [#####                 ] /scripts
  653.9 MiB [#####                 ] /R
  498.5 MiB [###                   ] /svn
  177.2 MiB [#                     ] /.m2
  168.3 MiB [#                     ] /.cache
  115.0 MiB [                      ] /rcac
   98.8 MiB [                      ] /.vscode-server
   84.1 MiB [                      ] /.nv
   73.6 MiB [                      ] /courses
    6.6 MiB [                      ] /.beast
    5.4 MiB [                      ] /.npm
    5.2 MiB [                      ] /.spack
    3.6 MiB [                      ] /.local
    1.3 MiB [                      ] /myapps
    1.0 MiB [                      ] /.config
```

To mitigate this, users can either run the singularity pull command with argument --disable-cache or manually clean $HOME/.singularity/cache

singularity pull --disable-cache URI

GPU support

For many applications, CPU compute resources provide sufficient performance. However, for a certain class of applications, the massively parallel compute power offered by GPUs can speed up operations by orders of magnitude.

Run a container with GPU acceleration

For AMD GPUs:

```
singularity shell/run/exec --rocm myimage.sif [command] [argument]
```

For NVIDIA GPUs:

```
singularity shell/run/exec --nv myimage.sif [command] [argument]
```

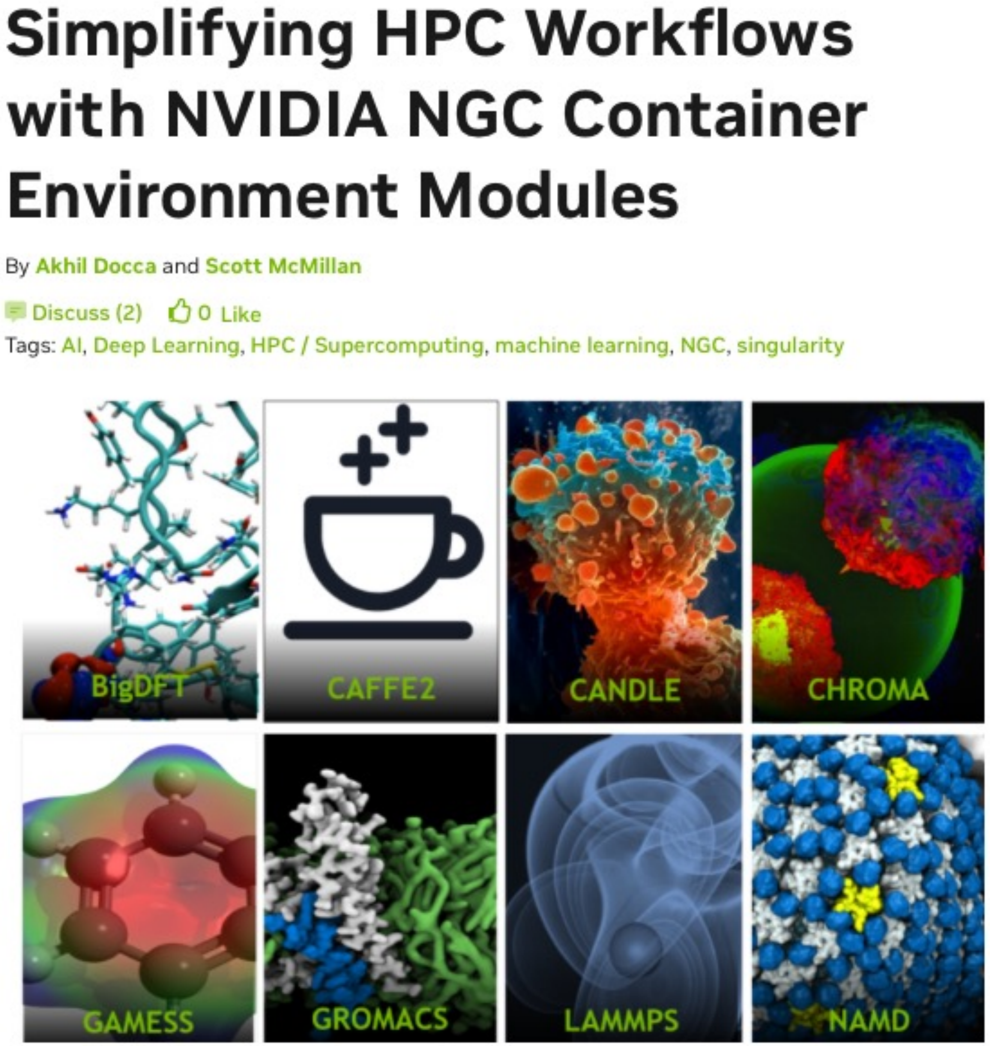
Containerized Bioinformatics applications for HPC

Biocontainers on RCAC clusters

NGC container environment modules

NGC container environment modules are lightweight wrappers that make it possible to transparently use NGC containers as environment modules.

1. Allow HPC users to utilize familiar environment module commands.
2. Leverage all the benefits of containers, including portability and reproducibility.



Simplifying HPC Workflows with NVIDIA NGC Container Environment Modules

By Akhil Docca and Scott McMillan

Discuss (2) 0 Like

Tags: AI, Deep Learning, HPC / Supercomputing, machine learning, NGC, singularity

https://github.com/NVIDIA/ngc-container-environment-modules

How to use biocontainers?

~800 modules for ~600 applications (As of March. 2023)

Load biocontainers

```
module load biocontainers
```

Check available applications

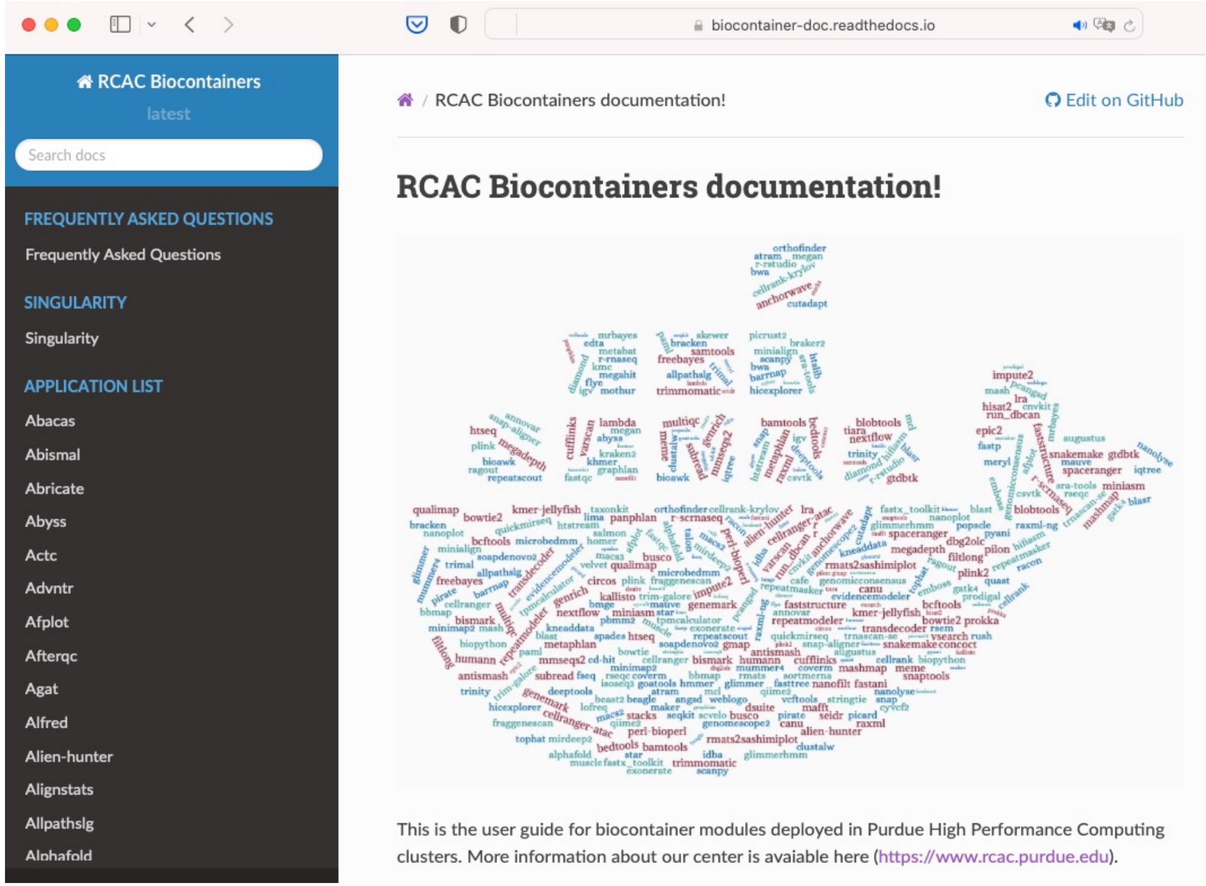
```
module avail
```

Load and run specific tools

```
module load samtools/1.16
samtools idxstats input.bam
```

Biocontainers documentation

$ module load biocontainers
User guides for each biocontainer module can be found in https://biocontainer-doc.readthedocs.io/en/latest



biocontainer-doc.readthedocs.io

RCAC Biocontainers
latest

Search docs

FREQUENTLY ASKED QUESTIONS

- Frequently Asked Questions

SINGULARITY

- Singularity

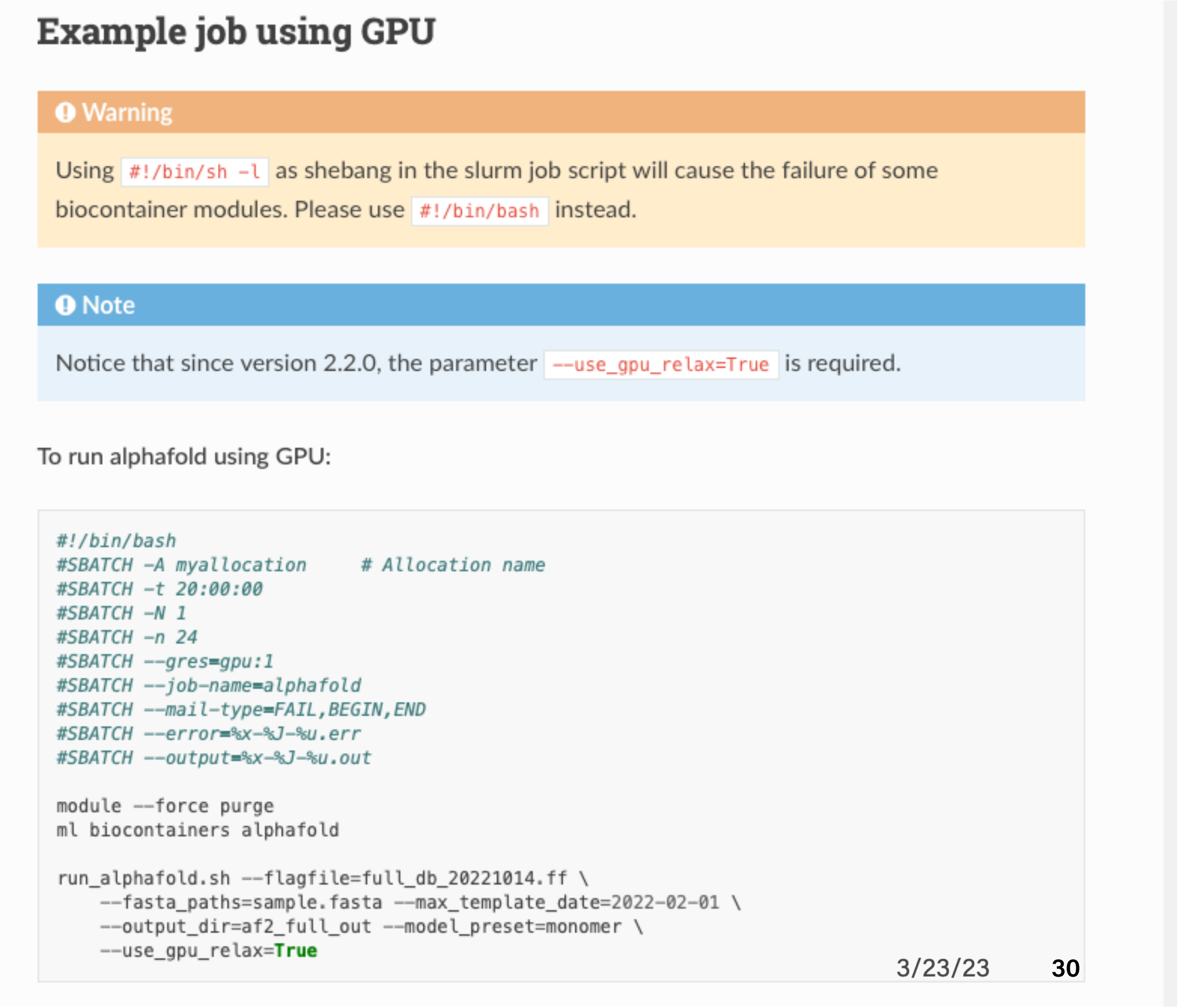
APPLICATION LIST

- Abacas
- Abismal
- Abricate
- Abyss
- Actc
- Advntr
- Afplot
- Afterqc
- Agat
- Alfred
- Alien-hunter
- Alignstats
- Allpathslg
- Alphafold

/ RCAC Biocontainers documentation! Edit on GitHub

RCAC Biocontainers documentation!

This is the user guide for biocontainer modules deployed in Purdue High Performance Computing clusters. More information about our center is avaiable here (https://www.rcac.purdue.edu).



Example job using GPU

Warning

Using `#!/bin/sh -l` as shebang in the slurm job script will cause the failure of some biocontainer modules. Please use `#!/bin/bash` instead.

Note

Notice that since version 2.2.0, the parameter `--use_gpu_relax=True` is required.

To run alphafold using GPU:

```
#!/bin/bash
#SBATCH -A myallocation     # Allocation name
#SBATCH -t 20:00:00
#SBATCH -N 1
#SBATCH -n 24
#SBATCH --gres=gpu:1
#SBATCH --job-name=alphafold
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%J-%u.err
#SBATCH --output=%x-%J-%u.out

module --force purge
ml biocontainers alphafold

run_alphafold.sh --flagfile=full_db_20221014.ff \
    --fasta_paths=sample.fasta --max_template_date=2022-02-01 \
    --output_dir=af2_full_out --model_preset=monomer \
    --use_gpu_relax=True
```

PURDUE UNIVERSITY
Rosen Center for Advanced Computing

r-rnaseq and r-scrnaseq

R-RNAseq

Customized R container for RNAseq analysis.

- ComplexHeatmap
- DESeq2
- DEXSeq
- edgeR
- ggrepel
- Limma
- pheatmap
- tidyverse

RStudio®

https://biocontainer-doc.readthedocs.io/en/latest/source/r-rnaseq/r-rnaseq.html

R-scRNAseq

Customized R container for scRNAseq analysis.

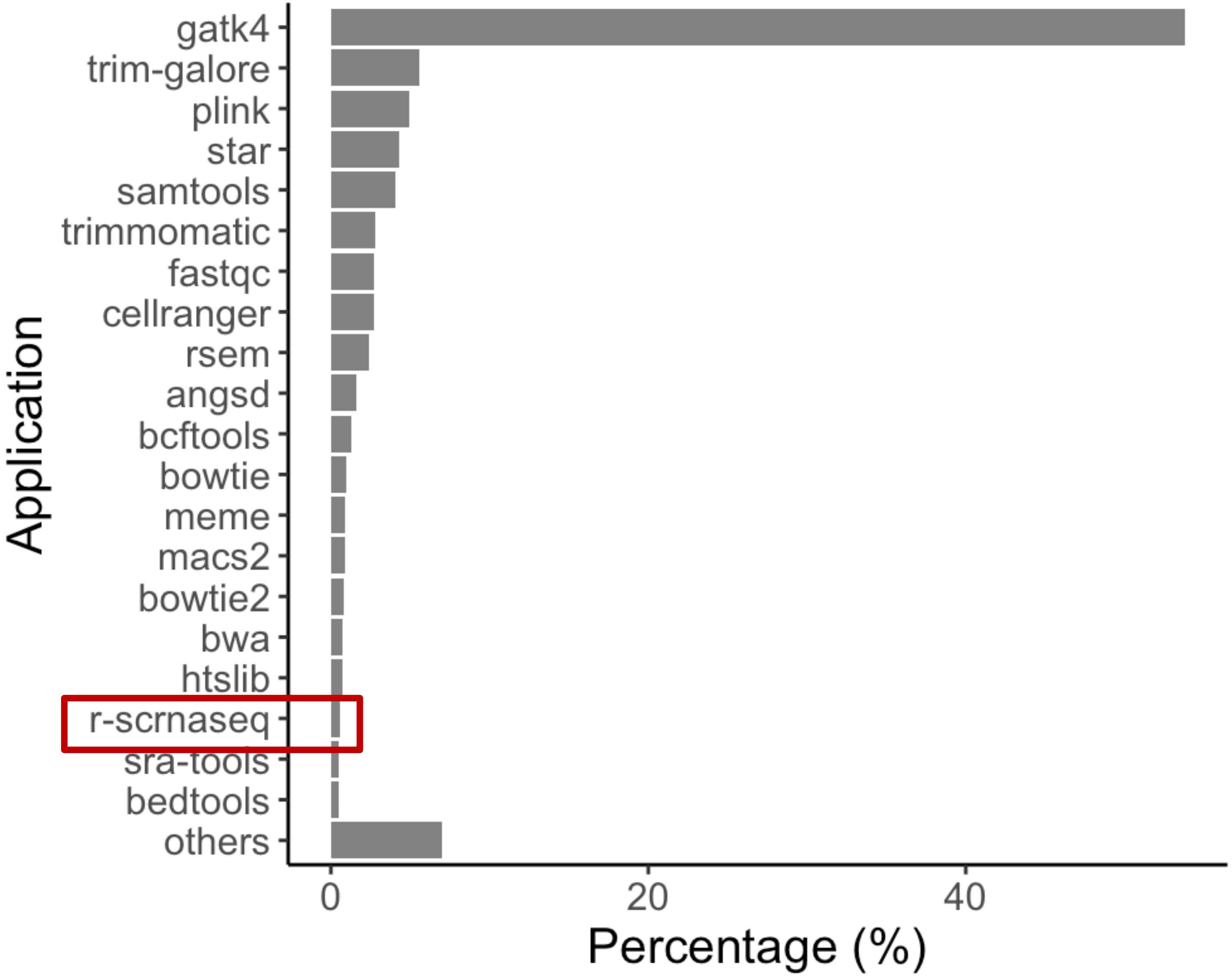
- CellChat
- CoGAPS
- DESeq2
- doSNOW
- DropletUtils
- edgeR
- Limma
- miQC
- monocle
- monocle3
- Nebulosa
- ProjecTILs
- rliger
- scCATCH
- scDblFinder
- SCHNAPPs
- scMappR
- seurat
- seurat-wrappers
- SingleR
- SnapATAC
- SoupX
- tidyverse
- tricycle
- velocyto.R

And more

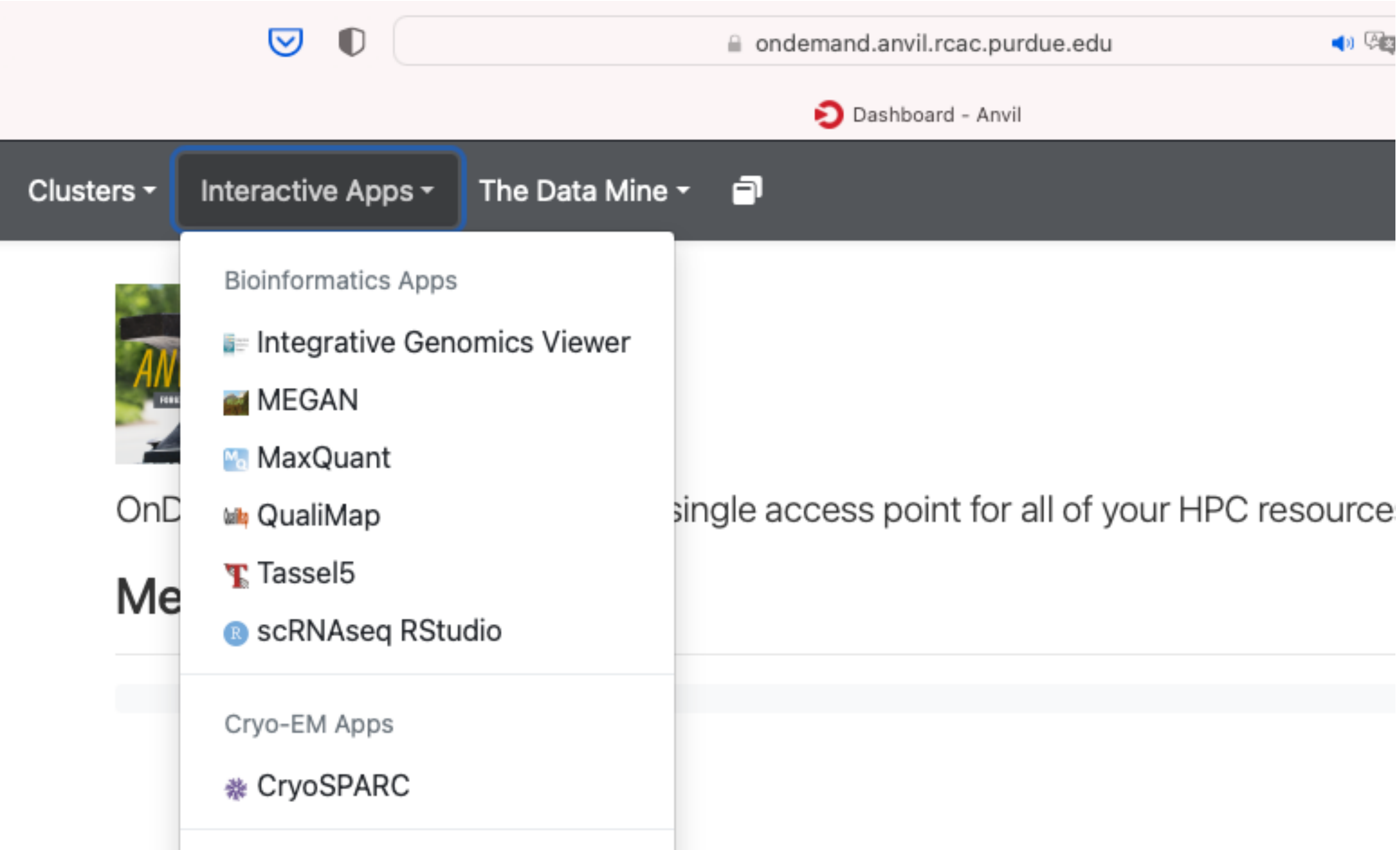
RStudio®

https://biocontainer-doc.readthedocs.io/en/latest/source/r-scrnaseq/r-scrnaseq.html

r-scrnaseq is one of the most popular biocontainer apps

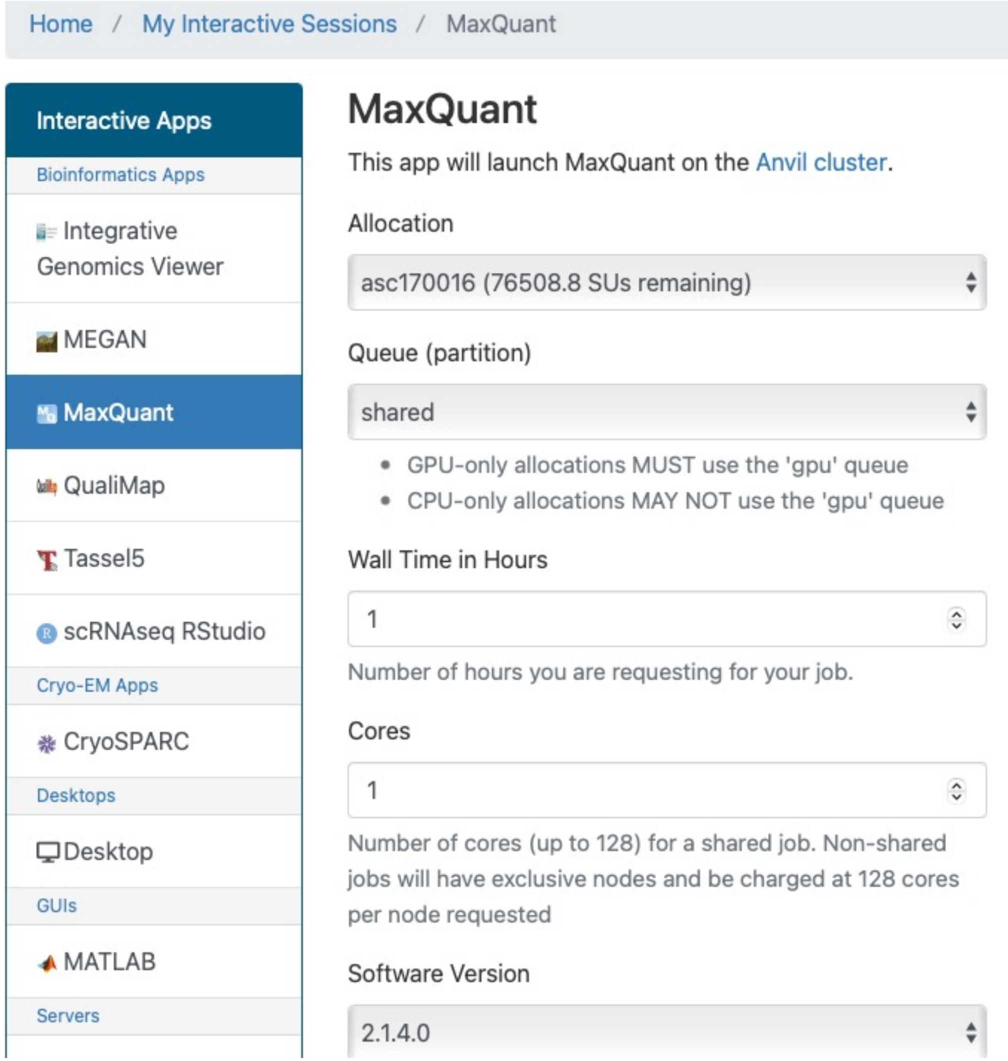


Open OnDemand Interactive Apps



ondemand.anvil.rcac.purdue.edu

Dashboard - Anvil



Home / My Interactive Sessions / MaxQuant

MaxQuant

This app will launch MaxQuant on the Anvil cluster.

Allocation

asc170016 (76508.8 SUs remaining)

Queue (partition)

shared

- GPU-only allocations MUST use the 'gpu' queue
- CPU-only allocations MAY NOT use the 'gpu' queue

Wall Time in Hours

1

Number of hours you are requesting for your job.

Cores

1

Number of cores (up to 128) for a shared job. Non-shared jobs will have exclusive nodes and be charged at 128 cores per node requested

Software Version

2.1.4.0

Containerized Bioinformatics applications for HPC

Interactive and batch Jobs

Interactive Jobs on Anvil

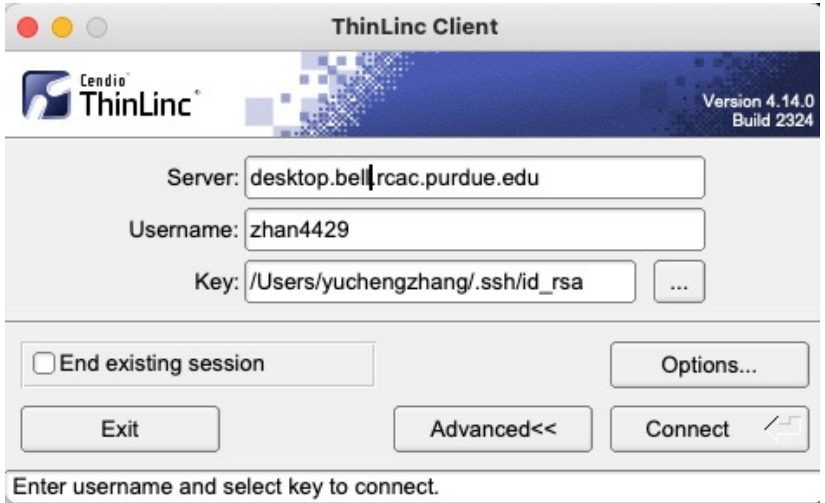
Interactive jobs are run on compute nodes, while giving you a shell to interact with. They give you the ability to type commands or use a graphical interface in the same way as if you were on a front-end login host.

1 node, 24 cores, and 10 hours walltime

```
sinteractive -N1 -n24 -p shared -t10:00:00 –A myallocation
```

1 node, 12 cores, 1 GPU and 4 hours walltime

```
sinteractive -N1 -n12 -p gpu --gpus-per-node=1 -t4:00:00 -A myGPUallocation
```



Server: desktop.bell.rcac.purdue.edu

Username: zhan4429

Key: /Users/yuchengzhang/.ssh/id_rsa

□ End existing session

Enter username and select key to connect.

https://www.cendio.com/thinlinc/download

Batch Jobs in Anvil: CPU

```
 #!/bin/bash

#SBATCH -A myallocation # Allocation name
#SBATCH -t 20:00:00
#SBATCH -N 1
#SBATCH -n 24
#SBATCH -p shared
#SBATCH --job-name=star
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%j-%u.err
#SBATCH --output=%x-%J-%u.out
#SBATCH --mail-user=useremailaddress

module --force purge
module load biocontainers star/2.7.10a

STAR  --runThreadN 24  --runMode genomeGenerate  \
          --genomeDir ref_genome  \
          --genomeFastaFiles ref_genome.fasta
```

%x: job name
%j: jobid
%u: userid

Batch Jobs in Anvil: GPU

```
 #!/bin/bash

#SBATCH -A myGPUallocation # GPU Allocation name
#SBATCH -t 20:00:00
#SBATCH -N 1
#SBATCH -n 24
#SBATCH -p gpu
#SBATCH --gpus-per-node=1
#SBATCH --job-name=parabricks
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%j-%u.err
#SBATCH --output=%x-%J-%u.out
#SBATCH --mail-user=useremailaddress

module --force purge
module load  biocontainers parabricks

pbrun haplotypecaller \
        --ref FVZG01.1.fsa_nt \
        --in-bam output.bam \
        --out-variants variants.vcf
```

THANK YOU

ACCESS Help Desk:
https://support.access-ci.org

ACCESS | Advancing Innovation

NSF

PURDUE UNIVERSITY | Rosen Center for Advanced Computing