High-throughput calculation and workflow automation

Abhilash Patra, USC

We will use MongoDB for our database. As the cluster doesn't allow us to use it in the login node, we will be using the interactive way to use one compute node for our setup and database connection.

After logging in to the CARC, use the following command to access one compute node for the interactive session.

```
salloc --time=02:00:00 --cpus-per-task=4  --partition=sharada

or

salloc --time=00:30:00  -cpus-per-task=4 --partition=debug
```

Change the values and partition according to your needs. It will assign one compute node.

Setting up MongoDB and other Python packages

Procedure 1 (Using pip)

Step 1: Preparation

Before we begin, I assume you have access to the project folder. We'll be using the MongoDB executable to store Firetask information. I've placed this executable in the project directory for your convenience.

Step 2: Copy the MongoDB Executable

First, let's make sure you have a bin directory in your home folder and then copy the MongoDB executable there:

```
cd
mkdir bin
cd bin
cp /project/ssharada_52/patraa/bin/mongod .
```

Step 3: Update your PATH

Next, let's ensure your system knows where to find the MongoDB executable by adding your bin directory to the PATH.

Open your .bashrc file for editing:

```
vi ~/.bashrc
```

Then, add the following line:

```
export PATH="/home1/patraa/bin:$PATH"
```

Make sure to replace /home1/patraa/bin with the actual path to your bin directory. If you're unsure about the path, the pwd command will display it when you're inside the bin directory.

Remember to save any changes you make to .bashrc and then source the file or restart the terminal session for the changes to take effect.

```
source ~/.bashrc
```

Installation of FireWorks, Pymatgen, and Custodian

Step 4: To install Pymatgen, FireWorks, and Custodian to your local system, the following commands can be used, or you can follow Stephen's page. Before installing, change the directory to home (though not necessary, for convenience)

```
pip install FireWorks --user
pip install pymatgen --user
pip install custodian --user
```

Link to Stephen's page-

https://stephen-quiton.github.io/pages/FW1-PythonInst.html
https://stephen-quiton.github.io/pages/FW4-Advanced-Setups.html

Procedure 2 (Using Conda)

```
module purge
module load conda
mamba init bash
source ~/.bashrc
mamba create --name fwworkflow
mamba activate fwworkflow
```

To install any Python package in the conda environment, search for the package on the anaconda webpage https://anaconda.org/

```
mamba install conda-forge::mongodb
mamba install conda-forge::fireworks python=3.10
mamba install conda-forge::pymatgen
mamba install conda-forge::custodian
```

(mamba env list) — command to check conda environments

Pymatgen Parsers for Q-Chem input file generation and excited state output extraction

Step 5: I have kept all the developed parsers for Q-Chem and necessary files for FireWorks in the folder ‘/project/ssharada_52/patraa/my_files’. Copy this file to your home and add your pythonpath to this directory as follows,

```
cp -r /project/ssharada_52/patraa/my_files .
vi ~/.bashrc
```

Then, add the following line and save the file:

```
export PYTHONPATH=“/home1/patraa/my_files:$PYTHONPATH"
```

Make sure to replace /home1/patraa/my_files with the actual path to your my_files directory. If you're unsure about the path, the pwd command will display it when you're inside the my_files directory.

Remember to save any changes you make to .bashrc and then source the file or restart the terminal session for the changes to take effect.

```
source ~/.bashrc
```

Q-Chem Workflow Automation

I have kept examples of the following steps in the folder

```
/project/ssharada_52/patraa/qchem_workflow
```

Copy this folder to your home

```
cd
cp -r   /project/ssharada_52/patraa/qchem_workflow .
```

Input File Creation

Input file from a .xyz file

The .xyz format is a general way to represent molecules. This file can be created from Iqmol, or if you have the coordinates of the system of interest, you can create Q-chem input files.

```
cd qchem_workflow/qcinput_from_xyz
python xyz_to_inp..py
```

Other Input files from Q-Chem .out file

We often need to calculate further properties from the optimized geometry, or if one geometry is not converged due to time limit, the geometry from incomplete calculation needs to be parsed to a new input file. These examples will create new input files from previous calculations. Only need to run the Python file. One example in each case is also provided.

Example directory: *_/project/ssharada_52/patraa/qchem_workflow/qcinput_from_qcoutput_*

Q-Chem 5

/project/ssharada_52/patraa/qchem_workflow/qcinput_from_qcoutput/qc5
/project/ssharada_52/patraa/qchem_workflow/qcinput_from_qcoutput/qc5_es

Q-Chem 6

/project/ssharada_52/patraa/qchem_workflow/qcinput_from_qcoutput/qc6

Output parsing and Analysis

In each directory, I place a ‘geometry’ folder that contains one example output file, along with a Python script named ‘transfer_geom.py’. This script is designed to create the input file.

Mulliken population
/project/ssharada_52/patraa/qchem_workflow/output_analysis/charge_out

Emission
/project/ssharada_52/patraa/qchem_workflow/output_analysis/emission_output_cyclohexane

Excited state EDA
/project/ssharada_52/patraa/qchem_workflow/output_analysis/output_eda

Exciton
/project/ssharada_52/patraa/qchem_workflow/output_analysis/exciton_out

Convert output to .xyz
/project/ssharada_52/patraa/qchem_workflow/output_analysis/xyz_files

Each directory contains two sample output files. The Python script 'excited_state_info.py', located in the parent directory, is designed to extract all necessary information from the output files in each subdirectory and create a consolidated data file. To execute this process correctly, the paths for each directory must be updated in the Python script.

Pymatgen output parser for ground state Q-Chem calculation

/project/ssharada_52/patraa/qchem_workflow/pymatgen_original/qchem/outputs.py

For Q-Chem output file containing multiple calculations

I have placed an example of a Q-Chem output file, which encompasses multiple calculations, in the following directory. Since this output file consolidates all information into a single .out file, extracting specific data can be challenging. To facilitate this, pymatgen offers a feature to decompose the composite output file into multiple .out.n files, where 'n' represents the number of jobs specified in the Q-Chem input. This decomposition allows for the extraction of required data from each individual .out.n file.

/project/ssharada_52/patraa/qchem_workflow/multiple_calc_out

FireWorks Job submission (if used pip for installation)

References

https://materialsproject.github.io/fireworks/

https://stephen-quiton.github.io/pages/FW1-PythonInst.html

Step 1 - Create a folder named 'data' in your home directory. Inside the 'data' folder, create two subfolders named 'db' and 'fw_logs'. Remember to note down the path of these folders for future reference.
Step 2 - You will require the following five files:

1. my_fworker.yaml
2. my_launchpad.yaml
3. SLURM_template.txt

I have kept the above three files in the `/project/ssharada_52/patraa/my_files` directory. Since you have already copied the `my_files` folder to your home directory, you will only need the path to this directory for access.

4. my_qadapter.yaml
5. add_job.py

Let's understand what these files do by taking one example. The below path is for the example directory.

/project/ssharada_52/patraa/qchem_workflow/FireWorks_jobs

In this directory, there are Q-Chem input files and three other files 'my_qadapter.yaml', 'add_job.py', and 'webgui.py'.

Step 3 - Modify the 'my_qadapter.yaml' by changing paths to your launchpad, slurm_template, and log file.

Step 4 - Make sure the port number is same in your 'add_job.py', 'my_launchpad.yaml' and 'webgui.py'.

Step 5 - The following commands need to be run

Setting up mongodb to a path

```
mongod --dbpath /home1/patraa/data/db --nojournal --port 27017 >> /home1/patraa/data/db/db1.txt &
```

Reset jobs if required; it will erase jobs previously assigned, if any.

FireWorks Job submission (for conda environment)

We need one configuration file for MongoDB. I have placed one in the following directory.
/project/ssharada_52/patraa/bin

```
cp /project/ssharada_52/patraa/bin/mongod.conf .
```

Change the paths and direct or make folders in data directory

```
mkdir mongolog
mongod --config mongod.conf &
```

The following commands are common for all

```
lpad reset
```

Assign jobs to Fireworks. All available .inp files will be added as fire tasks.

```
python add_job.py
```

Following command launch all jobs. Fireworks will automatically submit the jobs individually. The number 12 can be changed accordingly. It is the maximum number of jobs allowed in the queue in HPC.

```
qlaunch -r rapidfire -m 12
```

To check the status of the jobs running

```
lpad -l my_launchpad.yaml get_wflows
```

The following command is necessary to see the updated status of submitted jobs.

```
lpad recover_offline
```

WebGUI

```
ssh -L 8080:127.0.0.1:8080 patraa@endeavour2.usc.edu
```

Before running webgui.py, the following command is required; otherwise, it throws the error "WARNING: This is a development server. Do not use #it in a production deployment. Use a production WSGI server instead."

```
export FLASK_ENV=development
```

Then

```
python webgui.py
```

or

```
python /location/of/webgui.py &> log.txt &
```

to write the output in a file

After running webgui.py in HPC, the address http://127.0.0.1:8080/ in any laptop or desktop browser will lead to the WebGUI window.

Path for source files if installed using pip

```
/home1/patraa/.local/lib/python3.11/site-packages/
```

Path for source files if installed using conda

```
/home1/yshi1583/.conda/envs/mongoenv/lib/python3.10/site-packages/
```

Discovery does not allow login nodes to be used as interactive jobs

```
export MKL_NUM_THREADS=1
export NUMEXPR_NUM_THREADS=1
export OMP_NUM_THREADS=1
```

Alternate way: Use of interactive mode in CARC

https://www.carc.usc.edu/user-information/user-guides/software-and-programming/matlab