The n-Body Problem

Project 1 – Due Friday Mar 15

You will make a program to efficiently simulate n "bodies" in a 3D world. Each body has a mass, position, and velocity. Each body is gravitationally drawn towards every other body (and thus all draws every other body towards it). See wikipedia.org/wiki/N-body_problem for more details.

The Physics

The simulation will use Newtown's Law of Universal Gravitation, the Superposition Principle, and Newton's Second Law of Motion to determine how the bodies move within the 3D space. In the equations, for the i-th body m_i is its mass in kg and x_i, y_i, and z_i are its position in m.

- ***Newton's Law of Universal Gravitation***: the strength of the gravitational force F_{ij} between two bodies i and j is

$$F_{ij} = G\frac{m_i m_j}{r_{ij}^2}$$

where $G = 6.6743015\mathrm{e}{-}11\ \frac{\mathrm{Nm}^2}{\mathrm{kg}^2}$ is the gravitational constant, $r_{ij} = \sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2 + \Delta z_{ij}^2 + \varepsilon}$ is the Euclidian distance, $\Delta x_{ij} = x_j - x_i$, $\Delta y_{ij} = y_j - y_i$, and $\Delta z_{ij} = z_j - z_i$ are the differences in the positions of the two bodies, and $\varepsilon = 10^{-9}$ is a "softening" parameter to avoid division-by-zero.

- ***Superposition Principle:*** the net forces on a body in the x or y directions are the sum of all forces acting on the body in the direction. Thus, the overall forces on body i in the x and y directions are

$$F_x = \sum_{j=1}^{n} F_{ij}\frac{\Delta x_{ij}}{r_{ij}} \qquad F_y = \sum_{j=1}^{n} F_{ij}\frac{\Delta y_{ij}}{r_{ij}} \qquad F_z = \sum_{j=1}^{n} F_{ij}\frac{\Delta z_{ij}}{r_{ij}}$$

(note: the order of subtraction in Δx_{ij}, Δy_{ij}, and Δz_{ij} is important in the sums).

- ***Newton's Second Law of Motion:*** the acceleration (in $\mathrm{m/s}^2$) of a body in a direction is proportional to the force applied to it in that direction and its mass:

$$a_x = \frac{F_x}{m_i} \qquad a_y = \frac{F_y}{m_i} \qquad a_z = \frac{F_z}{m_i}$$

The naïve approach to computing all of the F_{ij} would be to compute them for every single i and j (except when $i = j$ when it would be undefined). ***Newton's Third Law of Motion*** states that for every force there is an equal and opposite force, i.e. $F_{ij} = F_{ji}$ (the gravitational force that body i applies to j is the same as the force applied from j to i). This means that you only have to compute F_{ij} for $j < i$ which is half of the number of computations.

Numerical Integration

Once you compute the acceleration for a body, you can then compute its new velocity and position for the next time. Every time step you will have to compute a new acceleration for every single body since acceleration is dependent on the current position and causes the current position to change. The repetitive evaluation of the values is called numerical integration. We cannot solve the integrals mathematically, but we can evaluate them over-and-over again to get an estimate of their true values. For this particular problem, the best numerical integration technique is called leapfrog integration.

In leapfrog integration, we compute the acceleration (as above) for the current time t (s), use that to compute the velocity halfway to the next time step ($t + {\Delta t}/{2}$), and finally use that to compute the position for the next time step ($t + \Delta t$). Thus, all of the velocities are a half-step off from the positions – they are "leapfrogging" with the positions.

Overall, for each body:

- Compute its current accelerations a_x, a_y, and a_z as above, using the current positions
- Compute its velocities v_x, v_y, and v_z for time $t + {\Delta t}/{2}$ by adding $a_x \Delta t$, $a_y \Delta t$, and $a_z \Delta t$ to its velocities from time $t - {\Delta t}/{2}$ (respectively)
- Compute its positions x, y, and z for time $t + \Delta t$ by adding $v_x \Delta t$, $v_y \Delta t$, and $v_z \Delta t$ to its positions from time t (respectively)

And then repeat for all time steps. As Δt is decreased the quality of the results increases since you the approximate values apply to a smaller range. However, decreasing Δt greatly increases the amount of computation required.

NOTE: You will never actually "do" anything with the halving of timesteps, instead this only influences when you are using "current" vs "future" values during computations. Be careful that you do not overwrite current values with future values before everyone has used them.

Example overall loop and outputting (pay careful attention to the loop bounds):

```
// TODO: allocate output matrix as num_outputs x 3*n
// TODO: save positions to row `0` of output

// Run simulation for each time step
for (size_t t = 1; t < num_steps; t++) {
    // TODO: compute time step...

    // Periodically copy the positions to the output data
    if (t % output_steps == 0) {
        // TODO: save positions to row `t/output_steps` of output
    }
}

// Save the final set of data if necessary
if (num_steps % output_steps != 0) {
    // TODO: save positions to row `num_outputs-1` of output
}
```

Programs

For this project, you must write several variations of this program. All of the programs take 5 or 6 command line arguments:

- time-step: a floating-point number > 0 in seconds that corresponds to Δt above
- total-time: a floating-point number $\geq \Delta t$ in seconds for total time to be simulated (T)
- outputs-per-body: number of positions to output per body
- input.npy: a file to load a matrix that contains the initial data
- output.npy: a file to write a matrix to that contains the output data
    - ***NOTE:*** first row is initial positions and last row is final positions computed regardless of the spacing of the remainder of the output positions
- num-threads *optional*: the number of threads to parallelize over
    - serial programs: don't have this argument at all
    - parallel programs: defaults to a reasonable value for the hardware/affinity

The input matrix npy file is an n-by-7 matrix with one row per body and columns:

- mass m_i (in kg)
- initial x, y, z position x_i, y_i, and z_i (in m)
- initial x, y, z velocity v_x, v_y, and v_z (in $\mathrm{m/s}$)

The code provided to you already reads these arguments and validates them. But that is all that is done. Look at the code for a list of variables established from the command-line arguments.

The only printed output (besides errors) is the time taken to compute everything in seconds (print with the `%g` format specifier). Your timing must include all computations, any operations you do to convert the data to a format you like better, and allocating memory. You should not include parsing the command-line arguments, loading the input file (just loading it, anything with its data must be timed), saving the output file (just saving it, anything with its data must be timed), computing/reporting the time taken, and cleanup (freeing memory/destroying objects).

Note: when `matrix_from_npy_path()` or `matrix_from_npy()` is used, the `Matrix` object is connected to the file, any changes to the `Matrix` you make will also change the file itself. You will either need to copy the `Matrix` (which must be timed) or never write to it.

You will make four programs:

- `nbody-s` – computes the results in serial using the naïve approach
- `nbody-s3` – computes the results in serial using Newton's 3rd Law of Motion
- `nbody-p` – computes the results in parallel using the naïve approach
- `nbody-p3` – computes the results in parallel using Newton's 3rd Law of Motion

Example Data and Scripts

On Canvas I have provided several example data sets along with their settings to run them. I also provide the expected output data so you can check if you code is operating properly. **All of these examples have 1000 output values per body.**

- sun-earth: 2-body setup with just the Sun and the Earth
  - expected results use $\Delta t = 1 \text{ min} = 60 \text{ s}$ and $T = 1 \text{ year} = 31557600 \text{ s}$
- solar-system-inner: 5-body setup with the Sun and Mercury, Venus, Earth, and Mars
  - expected results use $\Delta t = 1 \text{ min} = 60$, and $T = 1 \text{ year} = 31557600 \text{ s}$
- solar-system: 9-body setup with the Sun and all (official) planets
  - expected results use $\Delta t = 1 \text{ min} = 60 \text{ s}$ and $T = 1 \text{ year} = 31557600 \text{ s}$ or $T = 100 \text{ years} = 3155760000 \text{ s}$
- pluto-charon: 2-body setup that has is stable with the barycenter outside of both bodies
  - expected results use $\Delta t = 1 \text{ s}$ and $T = 3.5 \text{ days} = 302400 \text{ s}$[1]
- figure8: 3-body setup that has a stable figure-8 pattern
  - expected results use $\Delta t = 1e\text{-}4 \text{ s}$, and $T = 2.1 \text{ s}$
- figure8-rotate: figure8 example with an initial angular momentum
  - expected results use $\Delta t = 1\text{e-}4 \text{ s}$, and $T = 75 \text{ s}$
- random25: 25-body setup with random values
  - expected results use $\Delta t = 0.1 \text{ s}$ and $T = 500000 \text{ s}$
- random100, random1000, and random10000: large setups with random values
  - expected results use $\Delta t = 0.01 \text{ s}$ and $T = 1000 \text{ s}$

I am providing you with several Python scripts for checking the results, visualizing the results, and generating new inputs. You run them like `python3 file_name.py args`…. For example, to check if 2 npy files are (almost) equal:

```
python3 compare_npy.py my_earth_out.npy earth_exp_out.npy
```

Running any of the scripts with the `-h` argument will show helpful information. Feel free to modify these files for new features and share them (including with me).

Benchmarks

You ***must*** reach these on Expanse when compiling with `-Ofast` (instead of `-O3`) using the above parameters except that $T = 10s$. Your minimum speed over a few runs must be less than these values (i.e. you must be faster). Parallel ones can use as many threads as you want.

	random100	**random1000**	**random10000**
`nbody-s` and `s3`	0.1 secs	8 secs	750 secs
`nbody-p` and `p3`	0.1 secs	1.5 secs	20 secs

Hitting the following targets will be extra credit.

	random100	**random1000**	**random10000**
`nbody-s` and `s3`	0.05 secs	4.5 secs	450 secs
`nbody-p` and `p3`	0.015 secs	0.4 secs	15 secs

[1] About half of Pluto's day, not sure why I didn't do 1 Pluto day (551856 seconds)

Hints and Guidelines

- When using a shared node on Expanse, `get_num_cores_affinity()` in util.h will give you the number of cores actually allocated for your task, otherwise it returns the number of ***logical*** cores.
- Start with testing your programs with the smaller examples and checking the results.
- Start with the serial program in both cases and make sure it works.
- Don't commit any of the npy files (I don't need or want them).
- Organize your code to reduce redundancies, for example, make a function in a separate file that just performs a single F_{ij} and use that function in all of your programs.
    - Saving to the output is repeated a lot (3 times in each version), it should be its own function as well.
- Don't make `main()` do everything.
- You must use `double`s and not `float`s.

Speed Hints

- Do not use the `pow()` function to compute powers.
- There is a lot of math that is redundant and can be simplified/moved around to reduce the computational effort. One clear example of this is the mass variables that cancel each other out (avoids a multiplication and a division for each body pair).
    - This means you may need to do some algebra by hand!
- Do not use `malloc`/`free` in computational loops, they are expensive functions. Instead, allocate once before and re-use for each iteration of the loop.
- Explore AoS, SoA, and AoSoA versions of structs.
    - In particular, is there a difference between using position[3] and x, y, and z?
- Individual formula functions should probably be in a header (so it can be shared), but not in a .c file and be marked with `inline static`. This allows the compiler to more aggressively remove the function call overhead.
- When doing the 3rd-law programs make sure the threads are actually doing equal work (since $j < i$ is not the same for all values of i).

Analysis

You will also submit an analysis of how your programs function, answering the following questions. This is done in the analysis.md file. You may have to read up on Markdown format to make a good document (e.g. using headers, tables, and plot images).

Remember that when recording time to take the minimum of about 3 runs. Also, we are mostly caring about high-performance for large examples here. While I did give you lots of small examples, that was just for testing. You should be doing your analysis on huge samples (e.g. hundreds of thousands). I am expecting a few plots or similar to convey all of the information in your analysis.

1. What is the speedup of the naïve parallel program compared to the serial program?
    - Evaluate at several different sizes and talk about the trend
    - You may need to try various number of threads to get the optimal number, report the number of threads used by each different size
    - Evaluate the percent parallelism of the code (using Amdahl's Law) for each size
    - At what sizes is your serial program faster than your parallel program? (make sure to find a decent crossover point when they are approximately equal)
2. Repeat the above but for the 3rd-law program.
3. Compare your naïve program to your serial program. For what sizes is the naïve program faster and when is the 3rd-law program faster? What is the crossover point?

Rubric

10 pts for `nbody-s`
10 pts for `nbody-s3`
10 pts for `nbody-p`
10 pts for `nbody-p3`
25 pts for sufficient speed in all cases (limits will be published)
20 pts analysis
15 pts code quality