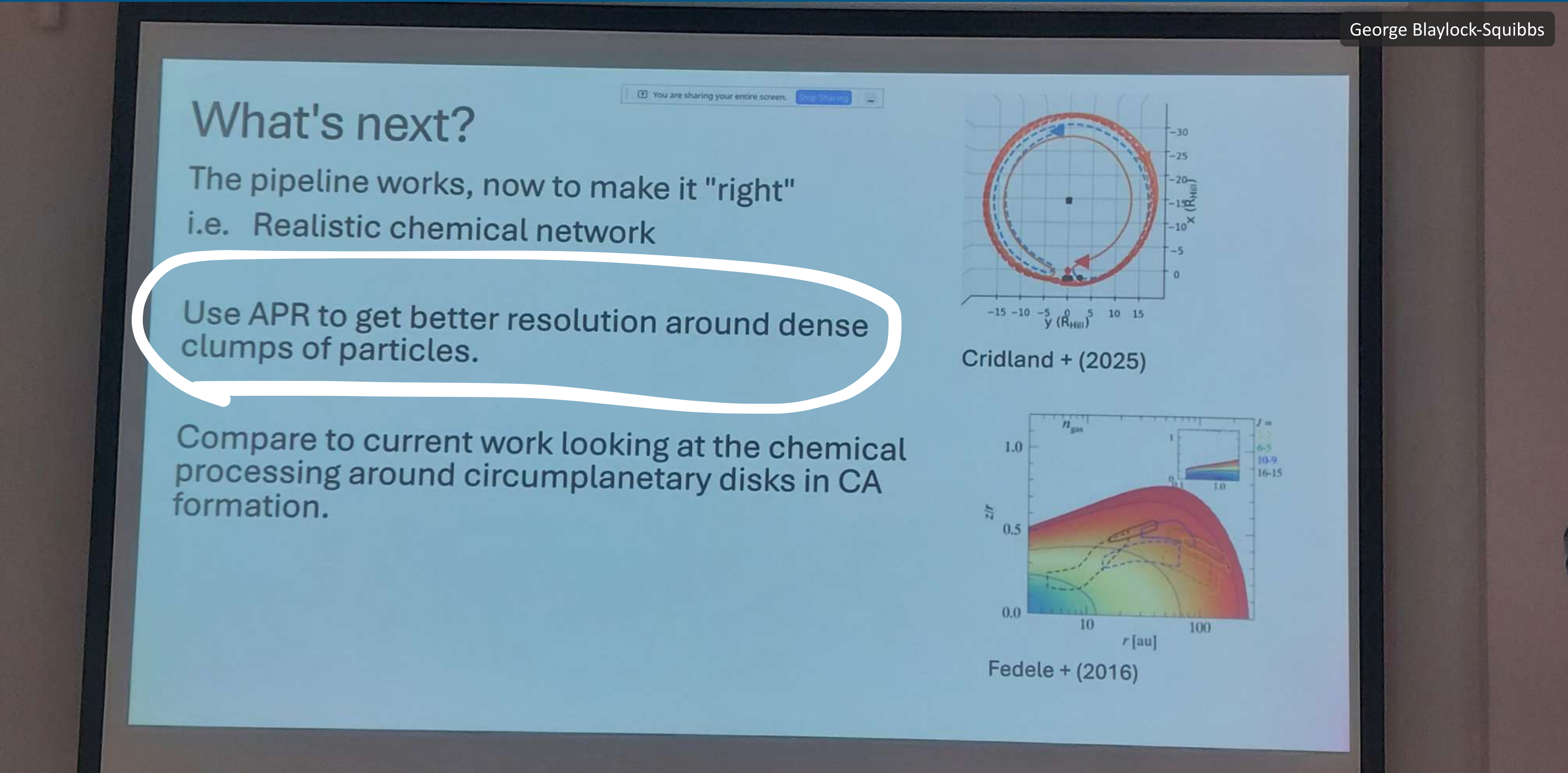
Parallelizing Multi-Physics SPH in Phantom

Terrence Tricco
Memorial University of Newfoundland

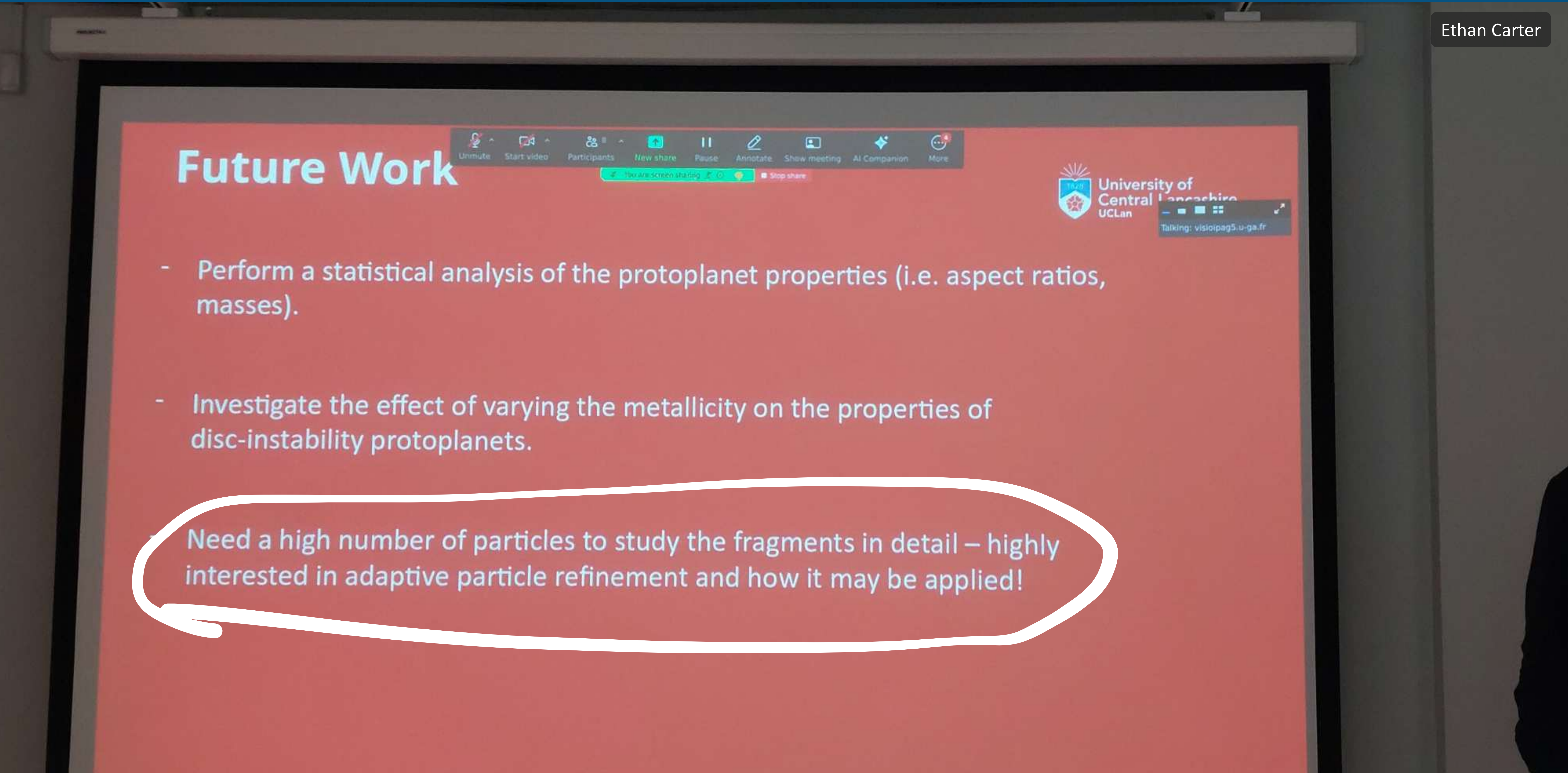
Do you use MPI in Phantom?

Need for Larger Simulations

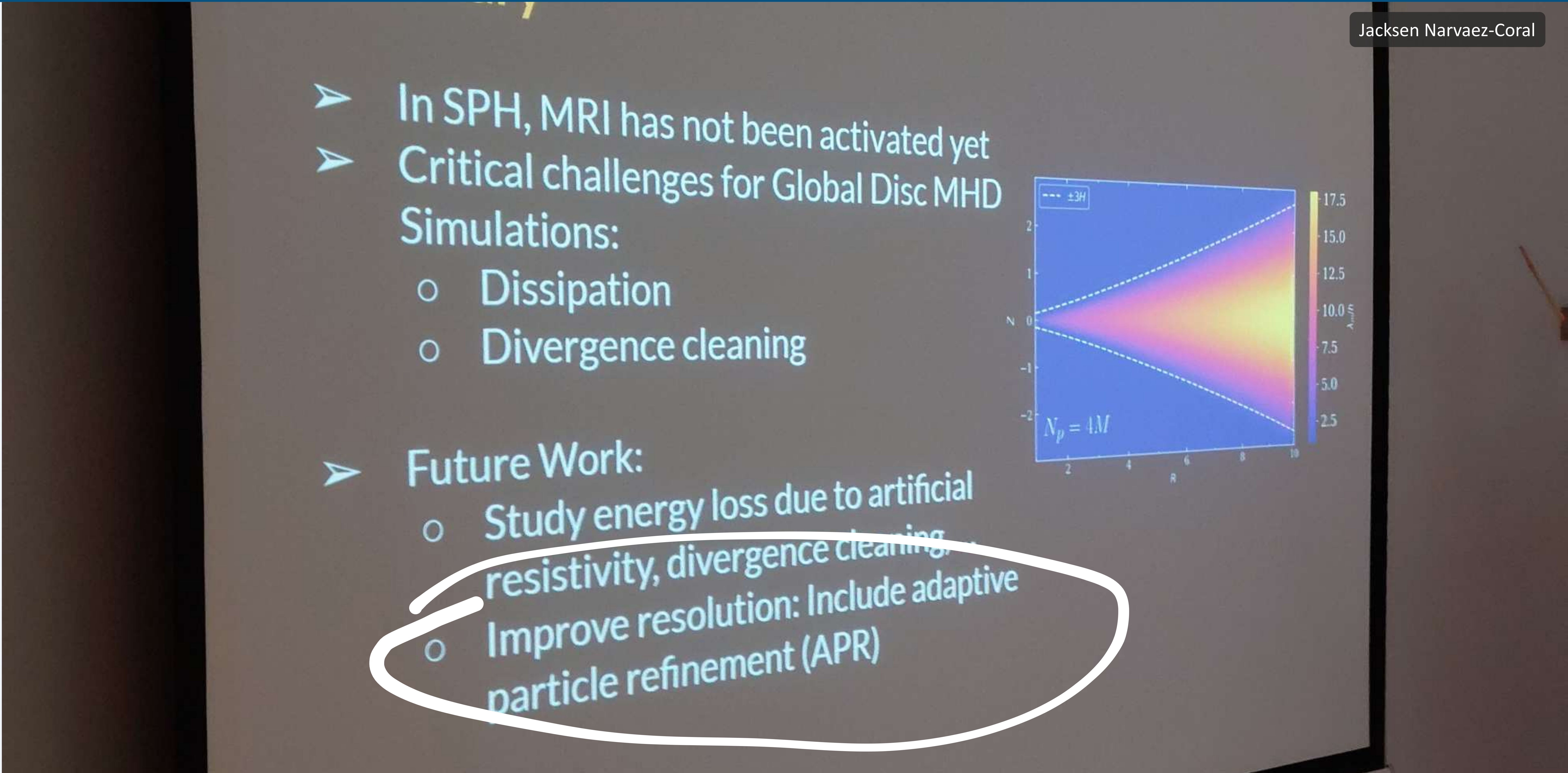
George Blaylock-Squibbs



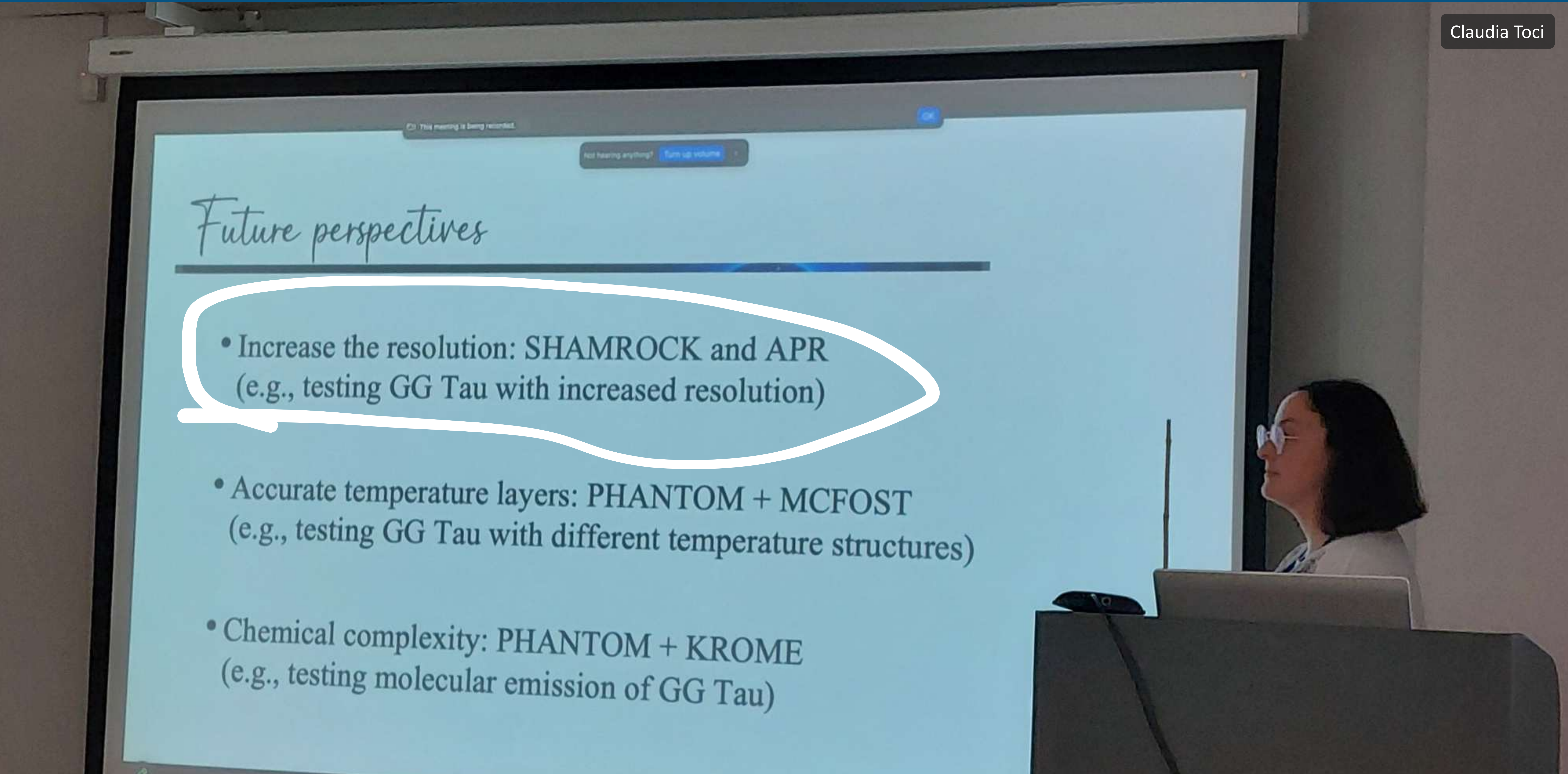
Need for Larger Simulations



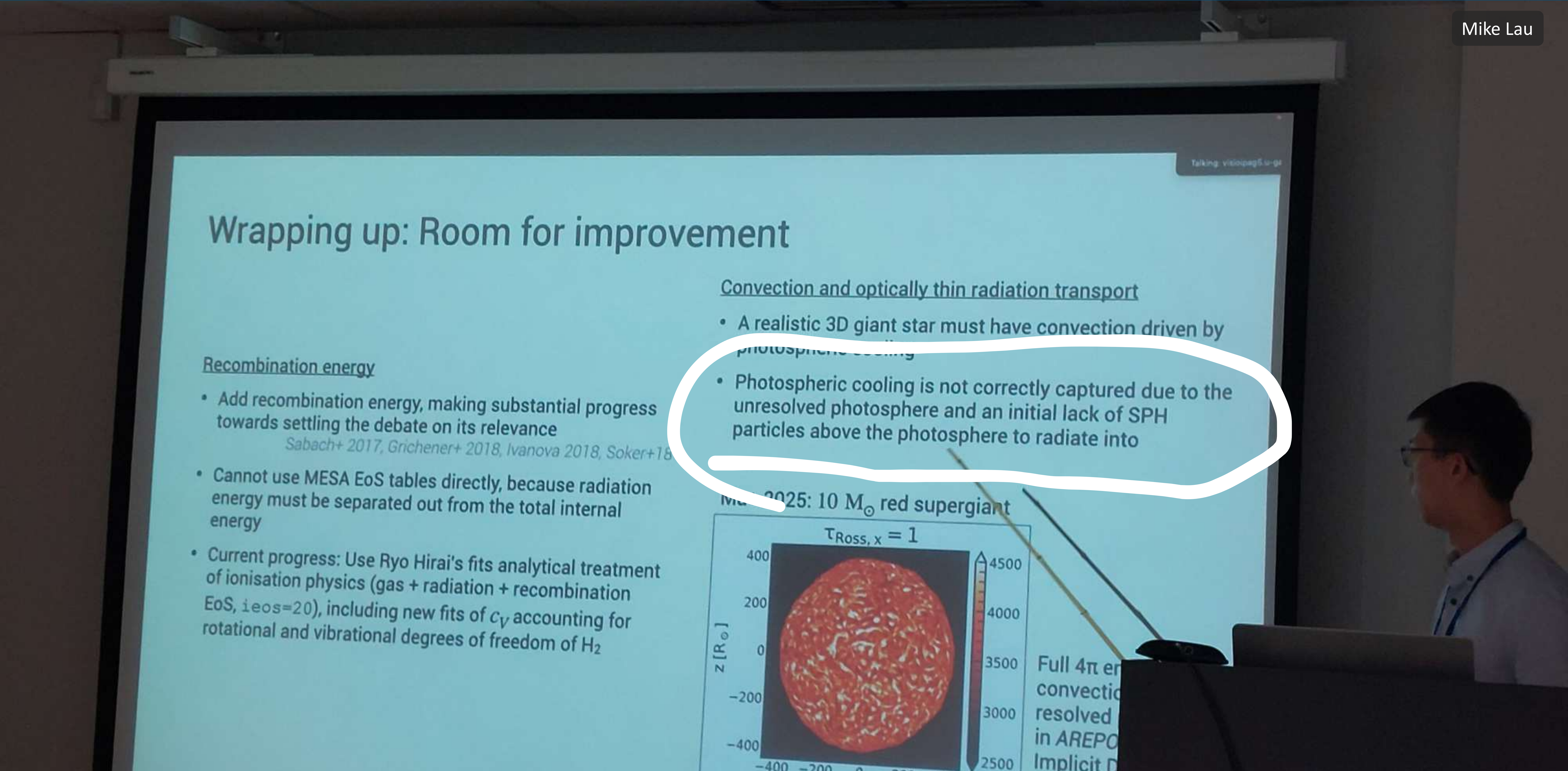
Need for Larger Simulations



Need for Larger Simulations

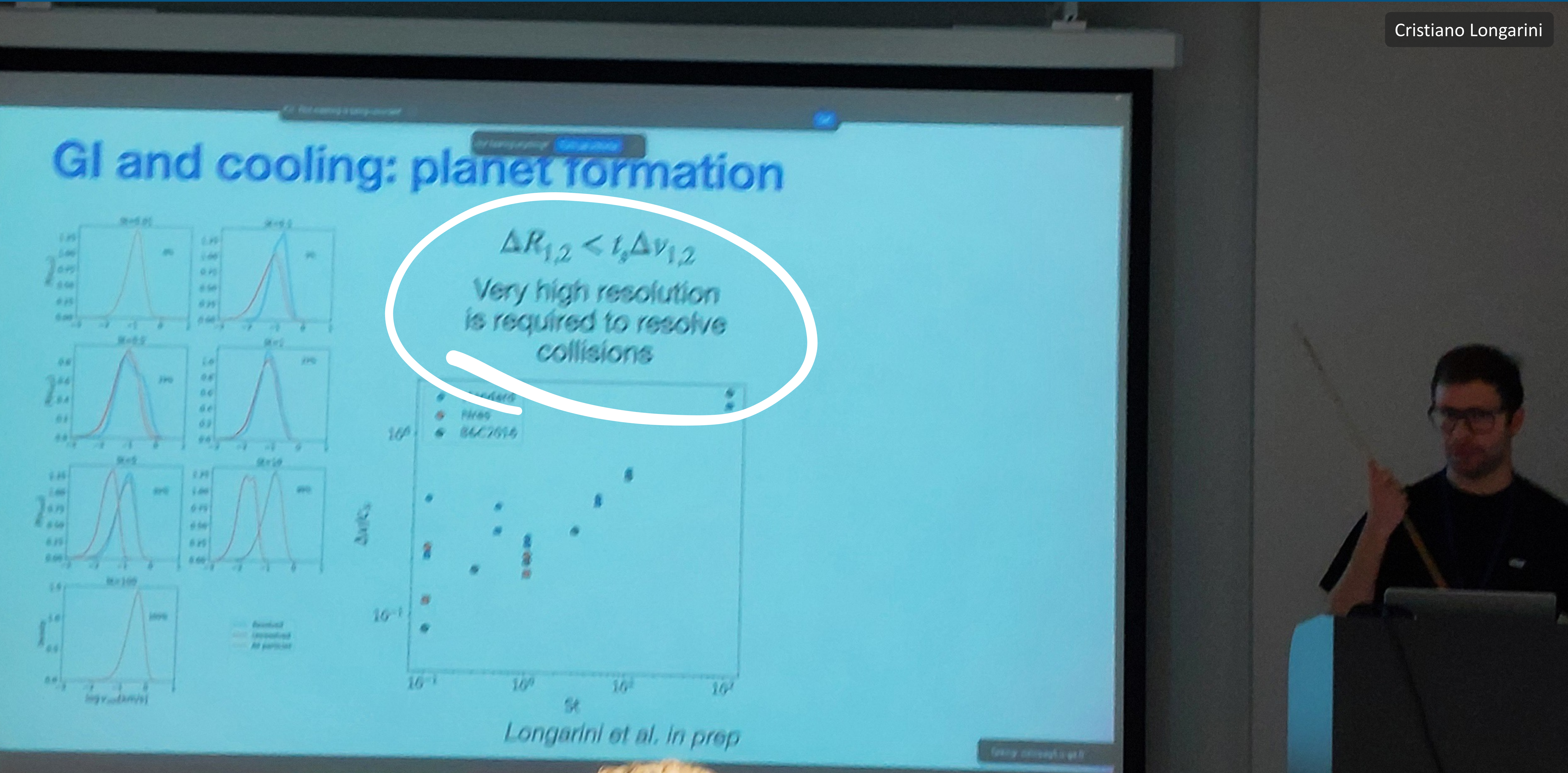


Need for Larger Simulations



Need for Larger Simulations

Cristiano Longarini



There is a clear need to run higher resolution simulations.

What is possible with Exascale

Guillaume Laibe

Resolution

e.g. resolve the radius of a planet in a disc simulations

Integration

e.g. form a planet after the protostar

Statistics

e.g. 1 simulation = all discs ever generated with RAMSES

Environnement

e.g. run tens of objets in interaction

To Higher Resolution

- How can we reach higher resolutions?

  1. Algorithmic optimizations.

  2. More compute.

To Higher Resolution

- How can we reach higher resolutions?

1. Algorithmic optimizations.
    - Adaptive particle refinement (APR) – ~6x speedup.
    - Individual particle timesteps – 100-1000x speedup.

2. More compute.

To Higher Resolution

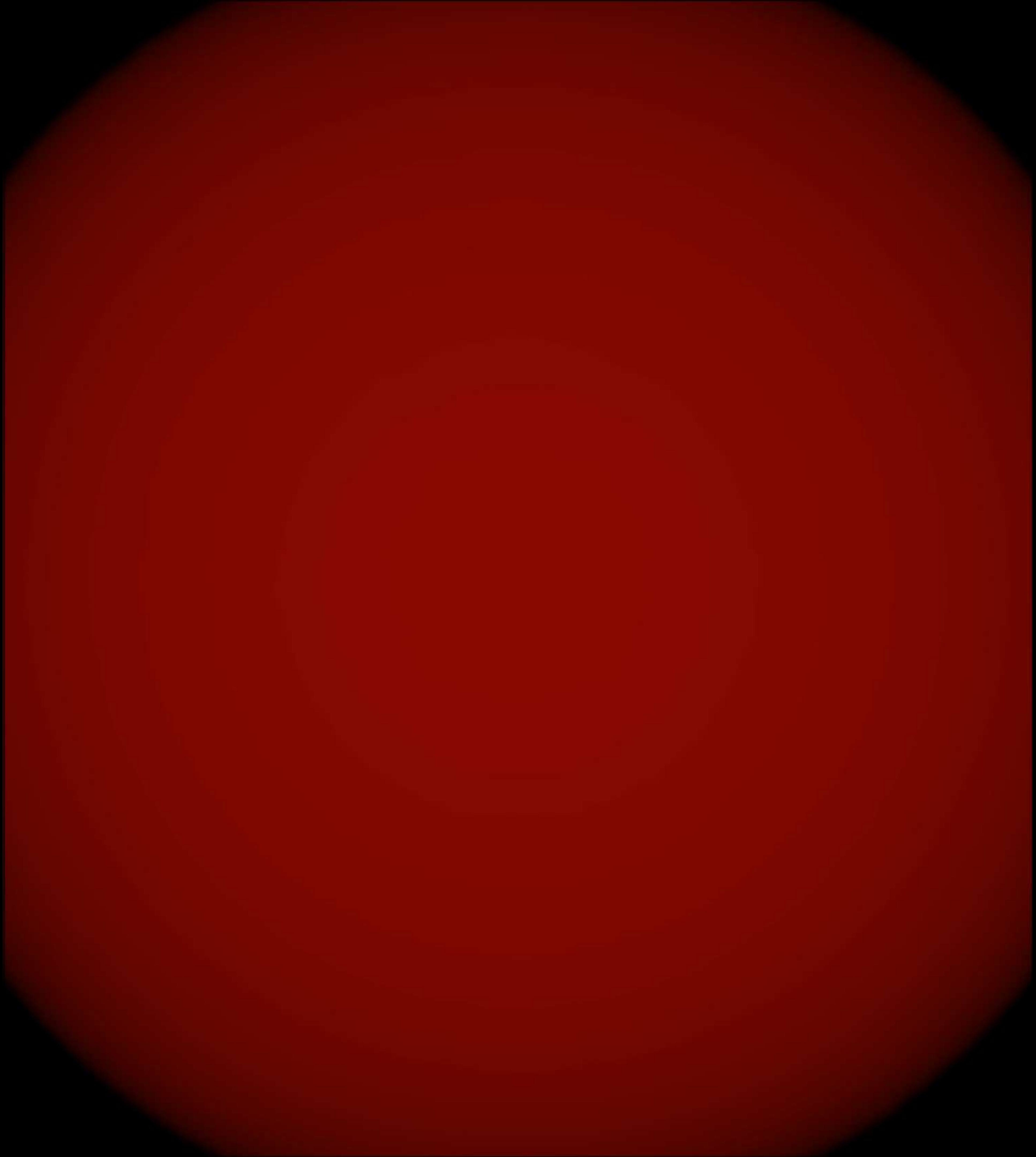
- How can we reach higher resolutions?

1. Algorithmic optimizations.
   - Adaptive particle refinement (APR) – ~6x speedup.
   - Individual particle timesteps – 100-1000x speedup.

2. More compute.
   - OpenMP – ~48-64x speedup.
   - MPI – ~2-1000x speedup?
   - GPUs – ~10x speedup?

How good is Phantom’s MPI scaling?

Gravitational Collapse



MPI in Phantom

64 cores (OpenMP)

```
-------->   TIME =     8.275     : full dump written to file jet1m_00097   <--------


 input file jet1m.in written successfully.
 Since code start: 746 timesteps, wall: 956s cpu: 5.09E+04s cpu/wall: 53
 Since last dump : 8 timesteps, wall: 12s cpu: 692s cpu/wall: 60

                           wall          cpu   cpu/wall   load bal       frac
 └step           :       11.62s      691.92s      59.52    100.00%     92.08%
   ├tree         :        2.88s      135.36s      47.08    100.00%     22.77%
   ├density      :        2.50s      163.60s      65.44    100.00%     19.80%
   │ ├local      :        2.50s      161.29s      64.51    100.00%     19.80%
   ├force        :        5.75s      370.66s      64.46    100.00%     45.54%
   │ ├local      :        4.62s      292.27s      63.19    100.00%     36.63%
   ├cons2prim    :        0.12s        5.43s      43.41    100.00%      0.99%
 └write_dump     :        1.00s        0.98s       0.98    100.00%      7.92%
-------------------------------------------------------------------------------
|** Number of steps since last summary:                      8              **|
|** Wall time since last summary:                       11.625 seconds      **|
-------------------------------------------------------------------------------
|* particles woken                                                           *|
| #steps   | mean # part/step |  max # part/step                              |
|        4|             13.00|                 17                             |
-------------------------------------------------------------------------------

-------------------------------------------------
| bin |      dt     |    npart  |    frac  | cpufrac |
-------------------------------------------------
|  0    8.886E-02          0     0.00%    20.38% |
|  1    4.443E-02     572589    37.45%    19.05% |
|  2    2.221E-02     666016    43.56%    26.99% |
|  3    1.111E-02     290346    18.99%    33.58% |
-------------------------------------------------
 0.51M particles per second, IND TIMESTEPS efficiency: 75.48%
```

128 cores (2 MPI nodes + OpenMP)

```
-------->   TIME =     8.264     : full dump written to file jet1m-2n_00093   <--------


 input file jet1m-2n.in written successfully.
 Since code start: 744 timesteps, wall: 3640s cpu: 3.61E+05s cpu/wall: 99
 Since last dump : 8 timesteps, wall: 50s cpu: 4983s cpu/wall: 100

                           wall          cpu   cpu/wall   load bal       frac
 └step           :       50.00s     4982.64s      99.65     99.84%     96.15%
   ├tree         :        9.88s      224.62s      22.75     88.00%     18.99%
   │ └balance    :        7.25s       14.16s       1.95    100.00%     13.94%
   ├density      :       27.25s     3259.12s     119.60     99.94%     52.40%
   │ ├local      :        7.12s      814.16s     114.27     95.52%     13.70%
   │ └remote     :       20.12s     2434.36s     120.96     98.39%     38.70%
   ├force        :       12.38s     1464.86s     118.37     99.29%     23.80%
   │ ├local      :        6.88s      815.92s     118.68     89.08%     13.22%
   │ └remote     :        4.62s      573.53s     124.01     86.59%      8.89%
   ├cons2prim    :        0.25s        6.25s      25.00     81.82%      0.48%
 └write_dump     :        2.00s        2.31s       1.16     43.69%      3.85%
-------------------------------------------------------------------------------
|** Number of steps since last summary:                      8              **|
|** Wall time since last summary:                       50.000 seconds      **|
-------------------------------------------------------------------------------
|* particles woken                                                           *|
| #steps   | mean # part/step |  max # part/step                              |
|        4|             42.25|                 55                             |
-------------------------------------------------------------------------------

-------------------------------------------------
| bin |      dt     |    npart  |    frac  | cpufrac |
-------------------------------------------------
|  0    8.886E-02          0     0.00%    14.87% |
|  1    4.443E-02     638476    41.76%    13.64% |
|  2    2.221E-02     643124    42.06%    27.32% |
|  3    1.111E-02     247351    16.18%    44.17% |
-------------------------------------------------
 0.11M particles per second, IND TIMESTEPS efficiency: 56.51%
```

MPI in Phantom

64 cores – 53 cpu/wall time

```
-------->   TIME =     8.275    : full dump written to file jet1m_00097   <--------

 input file jet1m.in written successfully.
 Since code start: 746 timesteps, wall: 954s cpu: 5.09E+04s cpu/wall: 53
 Since last dump : 8 timesteps, wall: 12s cpu: 692s cpu/wall: 60

                              wall         cpu   cpu/wall   load bal       frac
 └step               :      11.62s     691.92s     59.52    100.00%     92.08%
  ├tree              :       2.88s     135.36s     47.08    100.00%     22.77%
  ├density           :       2.50s     163.60s     65.44    100.00%     19.80%
  │ ├local           :       2.50s     161.29s     64.51    100.00%     19.80%
  ├force             :       5.75s     370.66s     64.46    100.00%     45.54%
  │ ├local           :       4.62s     292.27s     63.19    100.00%     36.63%
  ├cons2prim         :       0.12s       5.43s     43.41    100.00%      0.99%
 └write_dump         :       1.00s       0.98s      0.98    100.00%      7.92%
-------------------------------------------------------------------------------
|** Number of steps since last summary:                        8             **|
|** Wall time since last summary:                         11.625 seconds     **|
-------------------------------------------------------------------------------
|* particles woken                                                            *|
| #steps   | mean # part/step |  max # part/step                               |
|         4|             13.00|               17                               |
-------------------------------------------------------------------------------

-------------------------------------------------
| bin |     dt     |    npart   |    frac  | cpufrac |
-------------------------------------------------
|  0    8.886E-02           0    0.00%    20.38% |
|  1    4.443E-02      572589   37.45%    19.05% |
|  2    2.221E-02      666016   43.56%    26.99% |
|  3    1.111E-02      290346   18.99%    33.58% |
-------------------------------------------------
 0.51M particles per second, IND TIMESTEPS efficiency: 75.48%
```

128 cores – 99 cpu/wall time

```
-------->   TIME =     8.264    : full dump written to file jet1m-2n_00093   <--------

 input file jet1m-2n.in written successfully.
 Since code start: 744 timesteps, wall: 3640s cpu: 3.61E+05s cpu/wall: 99
 Since last dump : 8 timesteps, wall: 50s cpu: 4983s cpu/wall: 100

                              wall         cpu   cpu/wall   load bal       frac
 └step               :      50.00s    4982.64s     99.65     99.84%     96.15%
  ├tree              :       9.88s     224.62s     22.75     88.00%     18.99%
  │ └balance         :       7.25s      14.16s      1.95    100.00%     13.94%
  ├density           :      27.25s    3259.12s    119.60     99.94%     52.40%
  │ ├local           :       7.12s     814.16s    114.27     95.52%     13.70%
  │ └remote          :      20.12s    2434.36s    120.96     98.39%     38.70%
  ├force             :      12.38s    1464.86s    118.37     99.29%     23.80%
  │ ├local           :       6.88s     815.92s    118.68     89.08%     13.22%
  │ └remote          :       4.62s     573.53s    124.01     86.59%      8.89%
  ├cons2prim         :       0.25s       6.25s     25.00     81.82%      0.48%
 └write_dump         :       2.00s       2.31s      1.16     43.69%      3.85%
-------------------------------------------------------------------------------
|** Number of steps since last summary:                        8             **|
|** Wall time since last summary:                         50.000 seconds     **|
-------------------------------------------------------------------------------
|* particles woken                                                            *|
| #steps   | mean # part/step |  max # part/step                               |
|         4|             42.25|               55                               |
-------------------------------------------------------------------------------

-------------------------------------------------
| bin |     dt     |    npart   |    frac  | cpufrac |
-------------------------------------------------
|  0    8.886E-02           0    0.00%    14.87% |
|  1    4.443E-02      638476   41.76%    13.64% |
|  2    2.221E-02      643124   42.06%    27.32% |
|  3    1.111E-02      247351   16.18%    44.17% |
-------------------------------------------------
 0.11M particles per second, IND TIMESTEPS efficiency: 56.51%
```

MPI in Phantom

64 cores – 0.51M particles per second

```
-------->   TIME =     8.275     : full dump written to file jet1m_00097   <--------


 input file jet1m.in written successfully.
 Since code start: 746 timesteps, wall: 956s cpu: 5.09E+04s cpu/wall: 53
 Since last dump : 8 timesteps, wall: 12s cpu: 692s cpu/wall: 60

                             wall          cpu  cpu/wall  load bal       frac
 └step               :     11.62s      691.92s     59.52   100.00%     92.08%
   ├tree             :      2.88s      135.36s     47.08   100.00%     22.77%
   ├density          :      2.50s      163.60s     65.44   100.00%     19.80%
   │ ├local          :      2.50s      161.29s     64.51   100.00%     19.80%
   ├force            :      5.75s      370.66s     64.46   100.00%     45.54%
   │ ├local          :      4.62s      292.27s     63.19   100.00%     36.63%
   ├cons2prim        :      0.12s        5.43s     43.41   100.00%      0.99%
 └write_dump         :      1.00s        0.98s      0.98   100.00%      7.92%
------------------------------------------------------------------------------
|** Number of steps since last summary:                      8              **|
|** Wall time since last summary:                       11.625 seconds      **|
------------------------------------------------------------------------------
|* particles woken                                                           *|
| #steps   | mean # part/step |  max # part/step                             |
|        4|             13.00|                  17                          |
------------------------------------------------------------------------------

-------------------------------------------------
| bin |     dt     |    npart  |    frac  | cpufrac |
-------------------------------------------------
|  0    8.886E-02          0     0.00%    20.38% |
|  1    4.443E-02     572589    37.45%    19.05% |
|  2    2.221E-02     666016    43.56%    26.99% |
|  3    1.111E-02     290346    18.99%    33.58% |
-------------------------------------------------
 0.51M particles per second, IND TIMESTEPS efficiency: 75.48%
```

128 cores – 0.11M particles per second

```
-------->   TIME =     8.264     : full dump written to file jet1m-2n_00093   <--------


 input file jet1m-2n.in written successfully.
 Since code start: 744 timesteps, wall: 3640s cpu: 3.61E+05s cpu/wall: 99
 Since last dump : 8 timesteps, wall: 50s cpu: 4983s cpu/wall: 100

                             wall          cpu  cpu/wall  load bal       frac
 └step               :     50.00s     4982.64s     99.65    99.84%     96.15%
   ├tree             :      9.88s      224.62s     22.75    88.00%     18.99%
   │ └balance        :      7.25s       14.16s      1.95   100.00%     13.94%
   ├density          :     27.25s     3259.12s    119.60    99.94%     52.40%
   │ ├local          :      7.12s      814.16s    114.27    95.52%     13.70%
   │ └remote         :     20.12s     2434.36s    120.96    98.39%     38.70%
   ├force            :     12.38s     1464.86s    118.37    99.29%     23.80%
   │ ├local          :      6.88s      815.92s    118.68    89.08%     13.22%
   │ └remote         :      4.62s      573.53s    124.01    86.59%      8.89%
   ├cons2prim        :      0.25s        6.25s     25.00    81.82%      0.48%
 └write_dump         :      2.00s        2.31s      1.16    43.69%      3.85%
------------------------------------------------------------------------------
|** Number of steps since last summary:                      8              **|
|** Wall time since last summary:                       50.000 seconds      **|
------------------------------------------------------------------------------
|* particles woken                                                           *|
| #steps   | mean # part/step |  max # part/step                             |
|        4|             42.25|                  55                          |
------------------------------------------------------------------------------

-------------------------------------------------
| bin |     dt     |    npart  |    frac  | cpufrac |
-------------------------------------------------
|  0    8.886E-02          0     0.00%    14.87% |
|  1    4.443E-02     638476    41.76%    13.64% |
|  2    2.221E-02     643124    42.06%    27.32% |
|  3    1.111E-02     247351    16.18%    44.17% |
-------------------------------------------------
 0.11M particles per second, IND TIMESTEPS efficiency: 56.51%
```

MPI in Phantom

64 cores – 0.51M particles per second

```
-------->   TIME =     8.275     : full dump written to file jet1m_00097   <--------

 input file jet1m.in written successfully.
 Since code start: 746 timesteps, wall: 956s cpu: 5.09E+04s cpu/wall: 53
 Since last dump : 8 timesteps, wall: 12s cpu: 692s cpu/wall: 60

                         wall        cpu   cpu/wall   load bal      frac
 └step          :      11.62s    691.92s      59.52    100.00%    92.08%
   ├tree        :       2.88s    135.36s      47.08    100.00%    22.77%
   ├density     :       2.50s    163.60s      65.44    100.00%    19.80%
   │ ├local     :       2.50s    161.29s      64.51    100.00%    19.80%
   ├force       :       5.75s    370.66s      64.46    100.00%    45.54%
   │ ├local     :       4.62s    292.27s      63.19    100.0
   ├cons2prim   :       0.12s      5.43s      43.41    100.00%     0.99%
 └write_dump    :       1.00s      0.98s       0.98    100.00%     7.92%
 ------------------------------------------------------------------------------
|** Number of steps since last summary:                      8              **|
|** Wall time since last summary:                       11.625 seconds      **|
 ------------------------------------------------------------------------------
|* particles woken                                                           *|
| #steps   | mean # part/step |  max # part/step                              |
|        4|            13.00|              17                                |
 ------------------------------------------------------------------------------

 ------------------------------------------------
| bin |      dt     |   npart  |   frac  | cpufrac |
 ------------------------------------------------
|  0    8.886E-02          0     0.00%    20.38% |
|  1    4.443E-02     572589    37.45%    19.05% |
|  2    2.221E-02     666016    43.56%    26.99% |
|  3    1.111E-02     290346    18.99%    33.58% |
 ------------------------------------------------
 0.51M particles per second, IND TIMESTEPS efficiency: 75.48%
```

128 cores – 0.11M particles per second

```
-------->   TIME =     8.264     : full dump written to file jet1m-2n_00093   <--------

 input file jet1m-2n.in written successfully.
 Since code start: 744 timesteps, wall: 3640s cpu: 3.61E+05s cpu/wall: 99
 Since last dump : 8 timesteps, wall: 50s cpu: 4983s cpu/wall: 100

                         wall        cpu   cpu/wall   load bal      frac
 └step          :      50.00s   4982.64s      99.65     99.84%    96.15%
   ├tree        :       9.88s    224.62s      22.75     88.00%    18.99%
   │ └balance   :       7.25s     14.16s       1.95    100.00%    13.94%
   ├density     :      27.25s   3259.12s     119.60     99.94%    52.40%
   │ ├local     :       7.12s    814.16s     114.27     95.52%    13.70%
   │ └remote    :      20.12s   2434.36s     120.96     98.39%    38.70%
   ├force       :      12.38s   1464.86s     118.37     99.29%    23.80%
   │ ├local     :       6.88s    815.92s     118.68     89.08%    13.22%
   │ └remote    :       4.62s    573.53s     124.01     86.59%     8.89%
   ├cons2prim   :       0.25s      6.25s      25.00     81.82%     0.48%
 └write_dump    :       2.00s      2.31s       1.16     43.69%     3.85%
 ------------------------------------------------------------------------------
|** Number of steps since last summary:                      8              **|
|** Wall time since last summary:                       50.000 seconds      **|
 ------------------------------------------------------------------------------
|* particles woken                                                           *|
| #steps   | mean # part/step |  max # part/step                              |
|        4|            42.25|              55                                |
 ------------------------------------------------------------------------------

 ------------------------------------------------
| bin |      dt     |   npart  |   frac  | cpufrac |
 ------------------------------------------------
|  0    8.886E-02          0     0.00%    14.87% |
|  1    4.443E-02     638476    41.76%    13.64% |
|  2    2.221E-02     643124    42.06%    27.32% |
|  3    1.111E-02     247351    16.18%    44.17% |
 ------------------------------------------------
 0.11M particles per second, IND TIMESTEPS efficiency: 56.51%
```

MPI is 5x slower!!
(for this problem)

Multi-Physics SPH

- One challenge is including **multiple types of physics**.
- Shamrock has demonstrated that **hydrodynamics** can be scaled to **500 billion particles.**
- But how do we do that for gravity, dust (1-fluid, 2-fluid, multigrain, growth, etc), magnetic fields, relativity, radiation (FLD), sink particles, winds, chemistry evolution, etc etc?

Scaling Multi-Physics SPH

- I am very interested in **parallelizing multi-physics SPH** simulations.

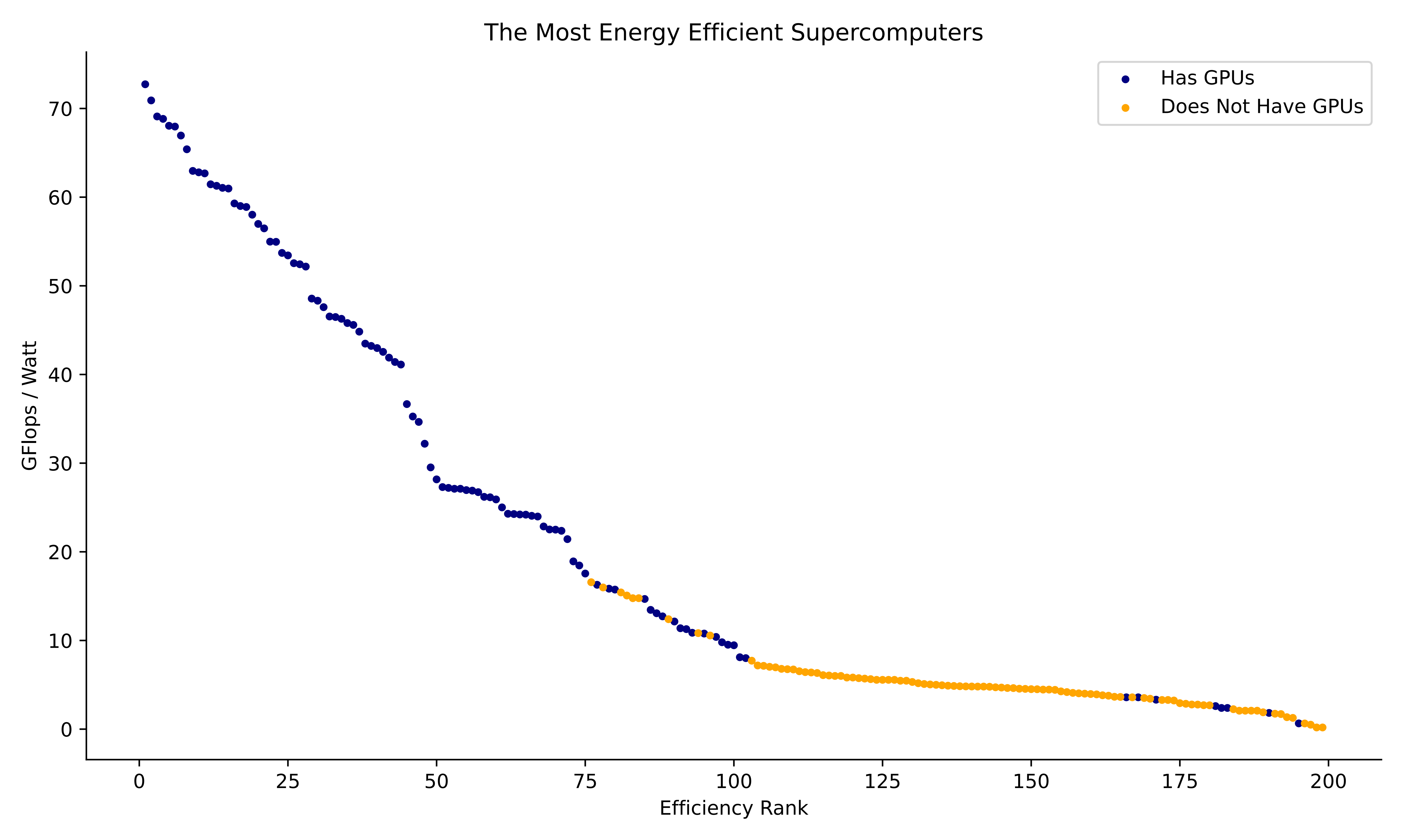
Scaling Multi-Physics SPH

- I am very interested in **parallelizing multi-physics SPH** simulations.

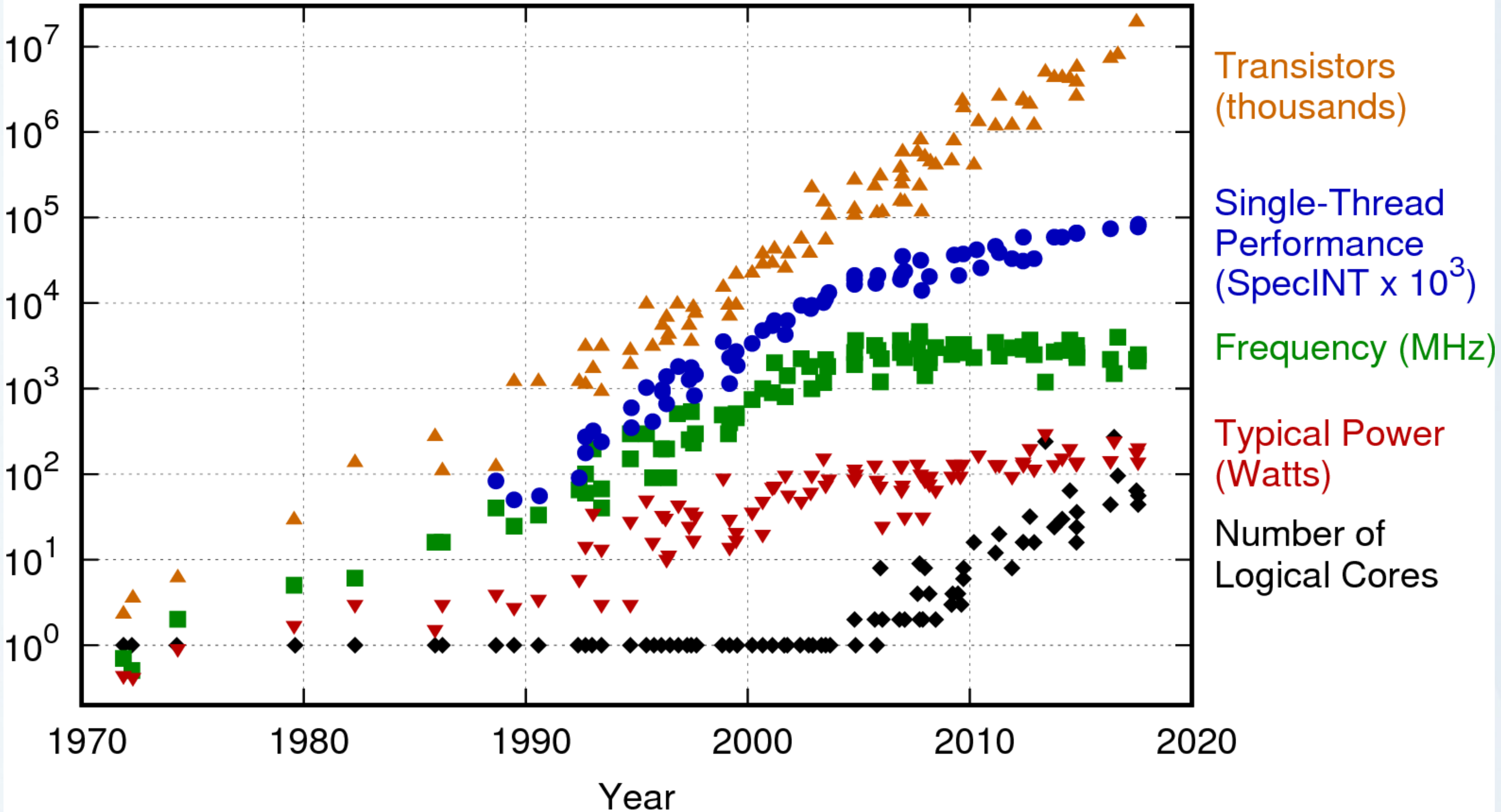
- Algorithmic optimizations are important!
- But we need to **combine** existing algorithmic optimizations with the ability to add more compute.

- One very important area will be GPUs.
  (see talks, e.g., by Timothée, Andrew and Tom)

Supercomputer Energy Performance

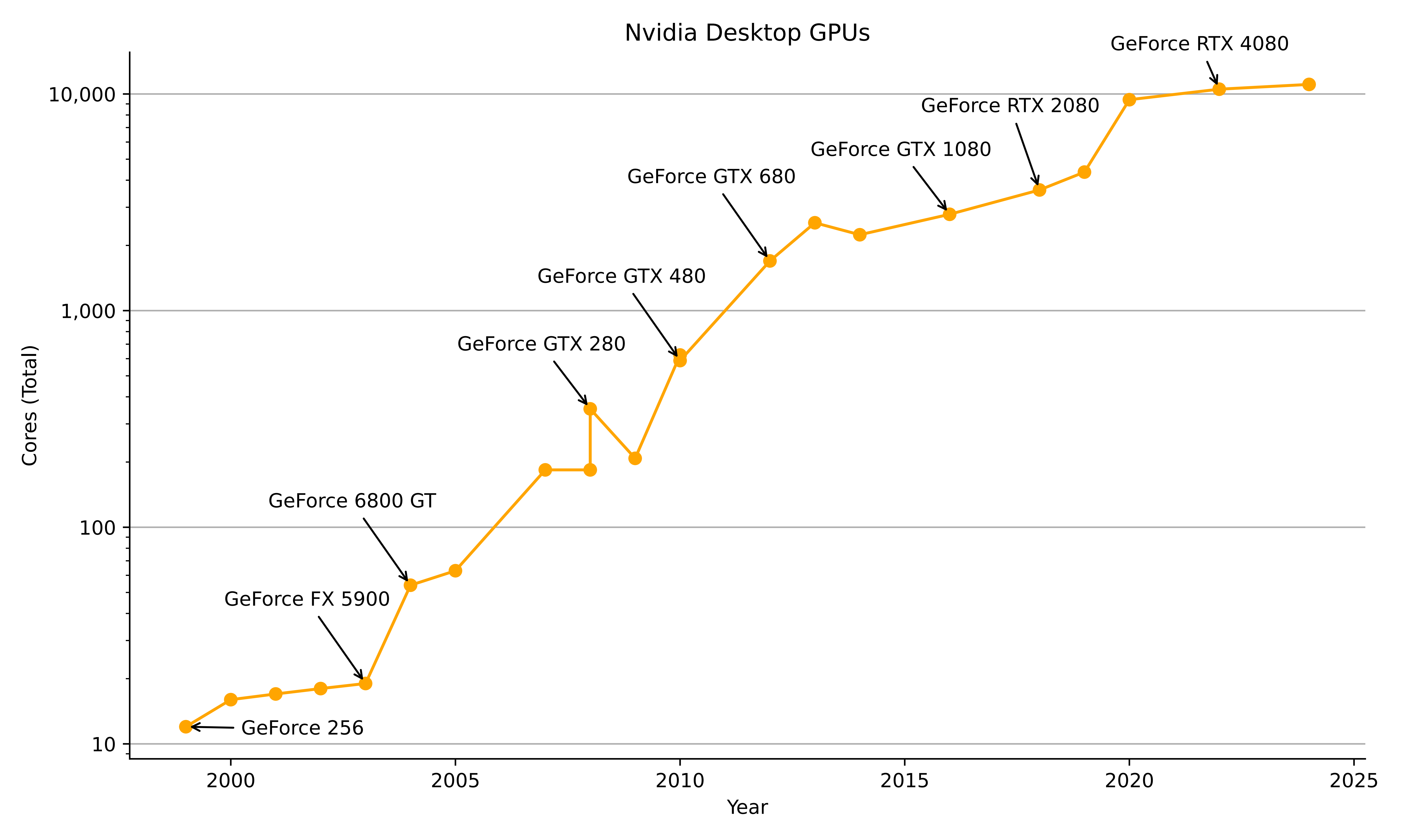


42 Years of CPU Trends



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

30 Years of Nvidia GPUs



Do I have answers?

Do I have answers?

No. But we are looking for solutions.

Phantom Parallelization Roadmap

- Short term goals for improving the parallel performance of Phantom:

1. Use 1-4 GPUs to increase compute.
   - Most HPC nodes have an attached GPU (or 4).
   - Needs to work with individual particle timesteps (Andrew Harris).
   - Key is for your simulation to not go slower when GPU=yes.

2. MPI scaling to 4-8 nodes.
   - Will be engaging with the HPC consortium in Canada to improve Phantom's current MPI implementation.

Summary

- We are working on **MPI** and **GPU** angles to increase **Phantom** parallelization.

- Important to retain algorithmic optimizations + all the physics.

- I did not speak about this, but am very interested in **MHD everything**.
  - MRI in global discs (Jacksen Narvaez).
  - Magnetized white dwarf mergers.