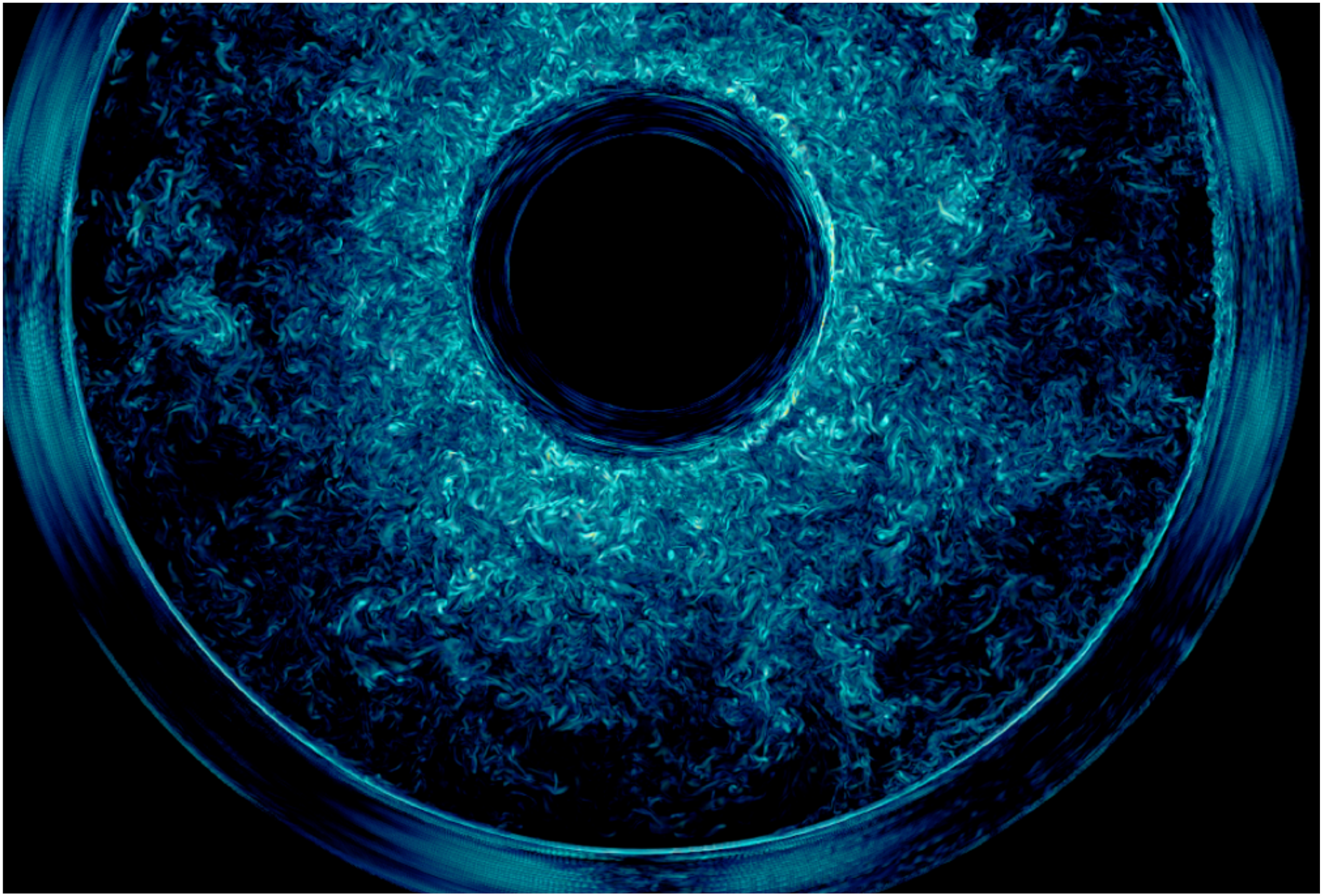
Case study: Large-scale 3D simulations of convection in a star

Falk Herwig

Vorticity of a slice He-flash convection



Large parallel runs on Niagara

https://docs.computecanada.ca/wiki/Niagara

compute canada
calcul canada

Page | Discussion | Read | View s

Wiki Main Page

Support

- Getting started
- Getting help
- Running jobs
- Known issues
- System status

Resources

Niagara

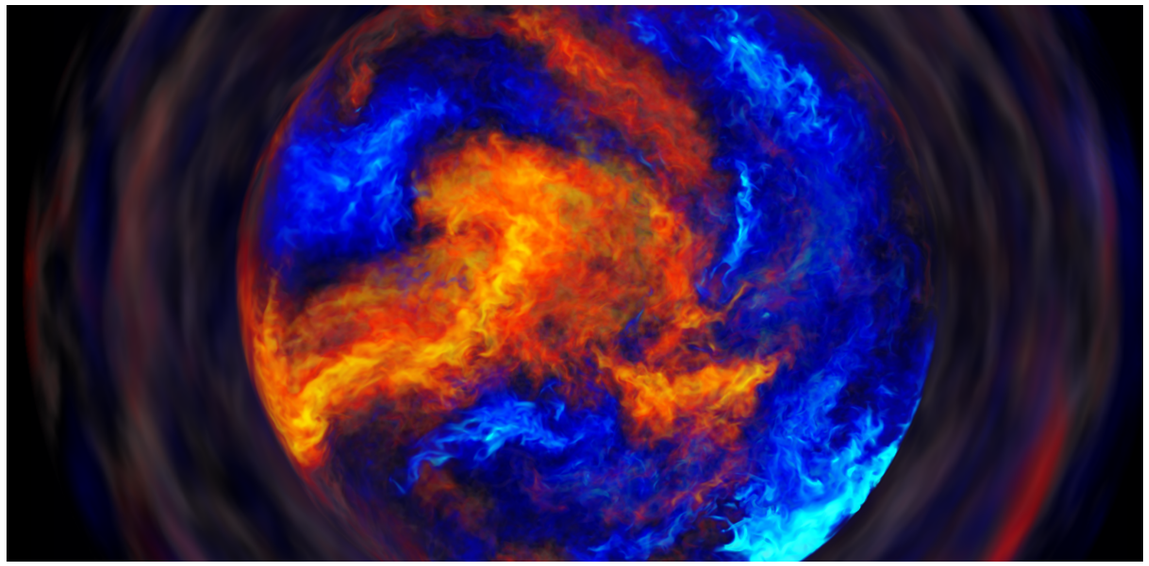
Other languages:	English ⊞ • français ⊞

Availability: In production since April 2018
Login node: **niagara.computecanada.ca**
Globus endpoint: **computecanada#niagara**
Data mover nodes (rsync, scp, ...): **nia-dm2, nia-dm2**, see Moving data
System Status Page: **https://docs.scinet.utoronto.ca**

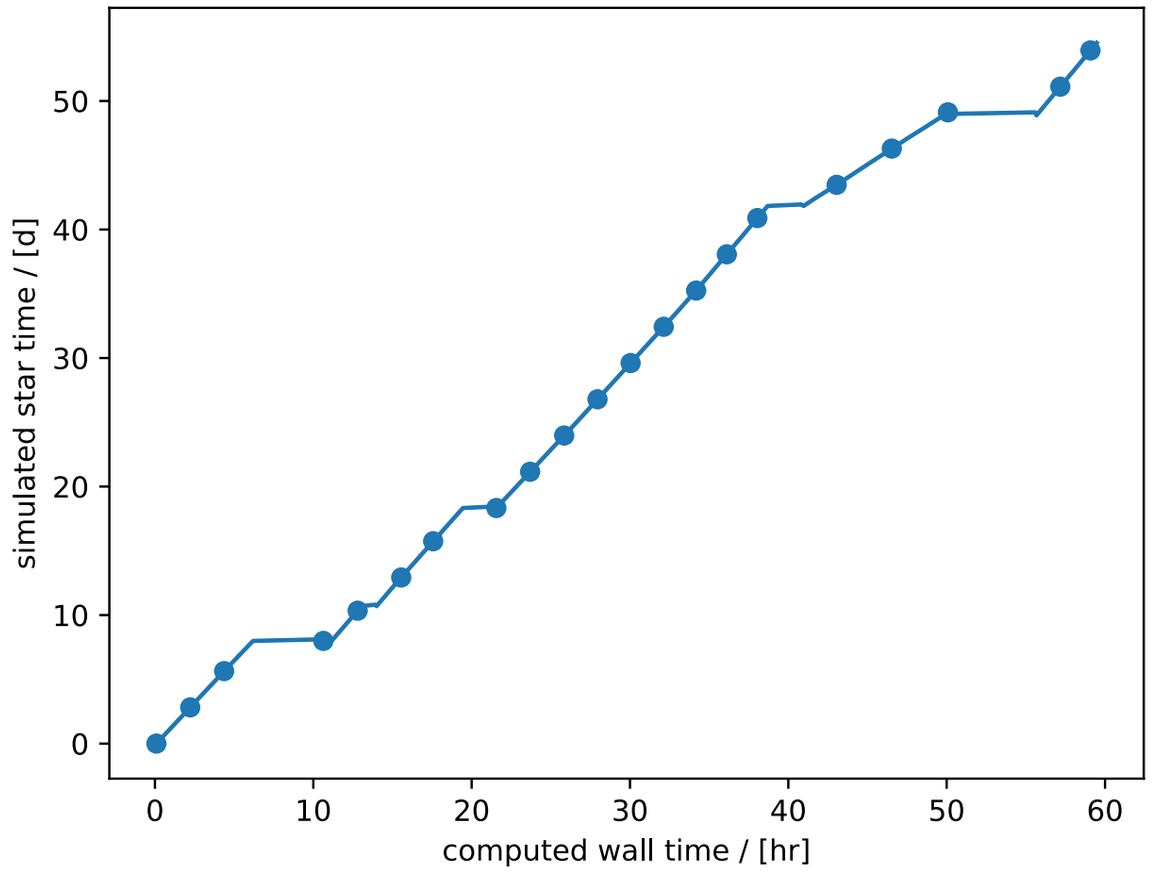
Niagara hardware specifications [edit]

- 1548 nodes, each with 40 Intel Skylake cores at 2.4GHz, for a total of 61,920 cores.
- 202 GB (188 GiB) of RAM per node.
- EDR Infiniband network in a so-called 'Dragonfly+' topology.
- 7PB of scratch, 2PB of project space (parallel filesystem: IBM Spectrum Scale, formerly known as GPFS).
- 256 TB burst buffer (Excelero + IBM Spectrum Scale).
- No local disks.
- No GPUs.
- Theoretical peak performance ("Rpeak") of 4.75 PF.
- Measured delivered performance ("Rmax") of 3.07 PF.
- 685 kW power consumption.

Running on Niagara



- 1536^3 grid, 3 billion grid zones
- 2.02M time steps
- 1088 nodes on Niagara
- 40 cores per node, Intel Skylake
- 80 threads per node
- 16 time steps per second
- 50 hours of wall clock time
- 60 days of star time



Simulated star time vs. time passed during the calculation according to the clock on the wall. Most parts were run on 1088 nodes, except a short section from hour 40 – 50 when we ran on 544 nodes and thus the rate of progress was reduced by a factor 2. Horizontal portions around 9, 20, 40 and 50-57 hrs indicate periods when the job was waiting in the queue between 8 hr jobs to proceed. Dots mark every 24th dump, which are written every 4126 time steps.

Where to find images, movies and more about these simulations?

- http://astrowww.phys.uvic.ca/~fherwig/Niagara

- YouTube playlist core convection: https://bit.ly/2HzTKtw

- Vimeo compilation Core convection: https://vimeo.com/album/5138161

About effective large-scale parallel simulations

- Key challenge: processing on cores is fast but getting data there is slow
- Overlap of communication and computation
- Set aside team leaders to deal with communications while workers continue to work
- Compress messages into smaller number of large messages, requires finding balance, depends on hardware
- Processors get faster but the network is not getting faster at the same rate

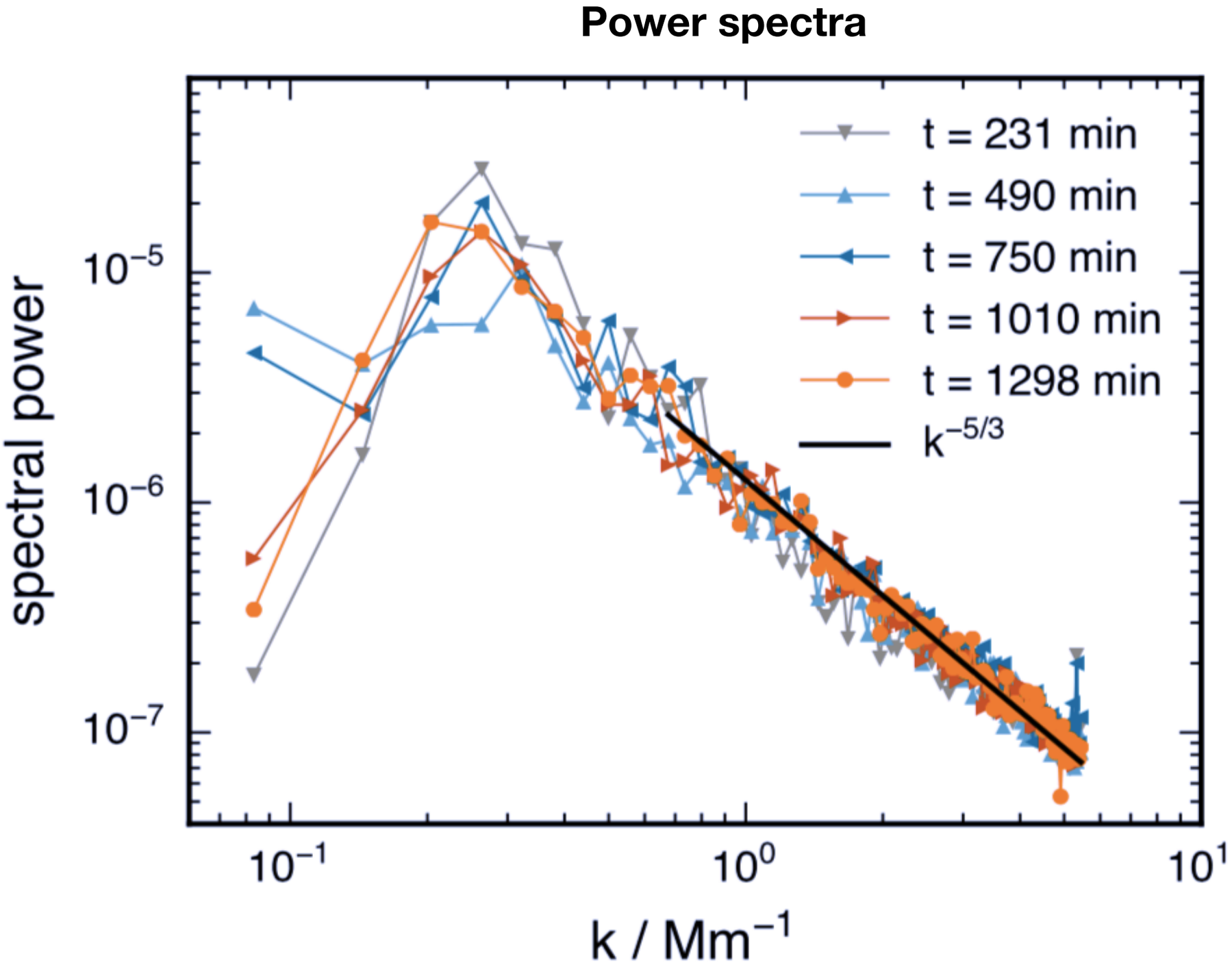
- Dedicated code to do just this type of problem, not general purpose code
- $1.2^{10}=6.2$ → pick up each 10-20% effect, for example
- Beware of the different flop cost of different operations, there is a logical operation for addition or multiplication (1 flop each) but not for sin, cos or divide (14 flops each)
- Solve problems the way computers like them, often not how mathematicians would do it

References:

Woodward PR, Herwig F, Wetherbee T. Comput. Sci. Eng. 20(5):8–17
Woodward PR, Lin P-H, Mao H, Andrassy R, Herwig F. 2018. eprint arXiv: 1810.13416 (ASTRONUM conference paper)

A possible analysis we would like to perform:

Spherical harmonic power spectra of the radial velocity component at $r = 17$ Mm and at 5 different points in time in the F4 run. Only the first 100 wave numbers are considered. The Kolmogorov scaling $\propto k^{-5/3}$ is shown for comparison.



Cyber challenges

- Data deluge naive approach: one dump 4 quantities 1536^3 x 8 byte x 4 quantities = 115GB - disk read (100MB/s) 20min, network transport (20MB/s) 1.6hrs
- 500 time steps = 57.5TB - disk read 6.6d, network transport 33d
- Large data sets require remote access and smart data management strategies

- Multi-scale, multi-physics, multi-method
- Complex data interactions and collaborative cyber- research environments, data fusion

- Research data management - in reality means combining access to data, data-specific analytics tools and capability to execute analytic tools to data
- Sharing work flows
- Legacy software, tools, and workflows
- Reproducibility of science

Data management for PPMstar code

- In-code generation of science-ready analysis data and images (this is a substantial departure from common practice)
  - instead of just writing out *raw* data we generate three types of algorithmically compressed file types
  - bobs - bricks of bytes algorithmically compressed
  - briquette data - spatially filtered (by factor 4 in each direction, 4 bytes, but 32 primary and derived quantities)
  - rprofs - radial profiles of lots of things
  - as well as one image for each of 10 quantities for each output dump

- Example: one 768^3-grid run with 1M time steps writes out 466 dumps
  - data held on project for immediate analysis (briquette & rprofs): 260GB
  - data on nearline: 3 restart dumps and bobs of 10 quantities (631GB)
  - instead of:

```
d = 466 *8*10* 768**3 /1.e12
print("Uncompressed data volume: 466 * 10 * 8 * 768**3 /1.e12 =  {:.2f}TB".format(d))
```

```
Uncompressed data volume: 466 * 10 * 8 * 768**3 /1.e12 =  16.89TB
```

Example for algorithmic data compression