Extreme Data Exploration

Dr Mark Wilkinson

Co-chair DiRAC HPC Facility
Project Lead: DiRAC-3 Data Service
A brief history of DiRAC

- 2011: £15M investment from BIS to support a national HPC facility for theoretical astrophysics, particle physics and cosmology
- Recurrent costs funded by STFC
- 2012: Five systems deployed:
  - 1.3 Pflop/s Blue Gene Q (Edinburgh)
  - 14.8TB RAM Shared Memory System (Cambridge)
  - 3 Clusters with various levels of interconnectivity and fast IO (Cambridge, Durham, Leicester)
- All systems designed around science requirements
  - Data-Centric Systems
- Common access to systems via SAFE
  - transparency
  - ease of access for users
- Service started on 1\textsuperscript{st} December 2012
Progress So Far

• Academic National Facility is up and running:
  – Academic-led design led to rapid uptake within community
  – Competitive, internationally peer reviewed allocation process
  – Currently 500 users and 37 Projects
  – A real step change in capability
  – More than 150 scientific papers generated since Dec 2012

• Now the focus is on:
  – Industrial and Public Sector Engagement
    • Identifying needs in these sectors
    • Assessing the services we can provide
  – Training our researchers
    • A training framework and progression
  – DiRAC-3: the next iteration of DiRAC (2015/16)
Heterogeneous Architectures
- matching hardware to problems

Many Core Chips and innovative networking power the BG/Q - used for high-precision lattice quantum chromodynamics calculations

Our Shared Memory and Xeon Phi System can be deployed in data fitting workflows - used for analysis of cosmological data
What do we use DiRAC for?

- We simulate very complex systems such as Galaxies
- We test our physical models against data
- Explore data for evolutionary changes, new effects and empirical relationships
- Assess the impact of model parameters
Data analysis with DiRAC-2

I. COSMOLOGY: Precision cosmology with the Planck satellite

II. MODELLING THE UNIVERSE: Observational consequences of inflation, early universe theory, and black hole physics

III. EXOPLANET SEARCHES: Detection, observation and characterisation of sub-stellar objects
The EAGLE Project

“In order to understand galaxy formation, we need a hydrodynamic simulation that matches the present-day galaxy mass function.”

constructors: Booth, Bower, Crain, Dalla Vecchia, Frenk, Furlong, Jenkins, Rosas-Guevara, Schaye, Theuns, McCarthy, Schaller, Springel, +++
The Eagle Universe at $z=0$
The Eagle Universe at $z=0$
The Eagle Universe at $z=0$
Zoom in by a factor 10....
Zoom in by a factor 10....
Zoom in by a factor 10....
How to survive Mega-data!

• EAGLE volume is 100 Mpc, 3 billion gas particles.
  – 28 full snapshots (all particle data)
  – 400 snapshots (high time resolution outputs)
• How is it possible to handle so much data?

• Our strategy:
  - Peano-Hilbert particle indexing
  - Extended particle data structures
  - Subfind: on-the-fly analysis
  - Halo particle output
  - Line-of-sight spectra
• Tree construction and data-base access
The ESA *Gaia* mission

- Launched on Dec 19th
- Measure the positions and 3D velocities of one billion stars
- A precise, 6-D map of the Milky Way
- European Space Agency (ESA) mission with significant UK involvement
- Science goal: build a complete model of the Milky Way
The ESA Gaia mission

- Launched on Dec 19th
- Measure the positions and 3D velocities of one billion stars
- A precise, 6-D map of the Milky Way
- European Space Agency (ESA) mission with significant UK involvement
- Science goal: build a complete model of the Milky Way
Observed data → Modelling Assumptions → Optimise model → Dynamical model → Compare model to data → Final model

Gaia modelling workflow
That sounds pretty standard/straightforward: what’s the problem?
What the Milky way looks like from outside.....
What the Milky way looks like from outside.....

You are here
Our view from inside the Milky Way
Gaia data challenges

- Raw data downlinked from satellite: 100-200 TB
- Augmented data set, including additional data and Bayesian parameter PDFs: ~1 PB
- Need to compare compute-intensive, multi-billion particle dynamical models to these data
- Large numbers of parameters in models have to be constrained by comparison with data
- Some features are time-dependent rather than intrinsic to the Galaxy: need to avoid modelling the noise (“trees”)
- Difficulty is that we don’t yet know which parts are trees and which are the forest (or even if there is a forest…..)
Building up the Milky Way
The Field of Streams
(Belokurov et al. 2006++)
Solutions

• Use additional prior information (physics) to identify “interesting” features
• Build smooth models of increasing complexity
• Models are computationally intensive
• Bayesian tools widely used (MCMC, Nested Sampling)
• Multiple modelling approaches being adopted - cross-comparison to reduce systematics
• Investigating novel tools (e.g. MapReduce)
Conclusions

• Data sets in all DiRAC science areas are entering, or have already entered, the PB scale

• Problems include both small numbers of large simulations and large numbers of moderate simulations

• Brute force approaches to analysis are not viable - we have to be innovative in software and hardware

• Use properties of the data, or prior knowledge (e.g. laws of physics) to direct analysis

• We are very interested in developing new partnerships in other fields and outside academia
How to contact us and find out more

- www.dirac.ac.uk
- DiRAC contacts:
  - Jeremy Yates (Project Director) jyates@star.ucl.ac.uk
  - Mark Wilkinson (Co-Chair) mark.wilkinson@le.ac.uk
  - Harpreet Dhanoa (Project Manager) hd@star.ucl.ac.uk
- STFC contact:
  - Rachel Reynolds: rachel.reynolds@stfc.ac.uk
- More details about science projects:
  - Wilkinson, Astronomy & Geophysics, (Dec 2013)