DiRAC-3 Memory Intensive Service

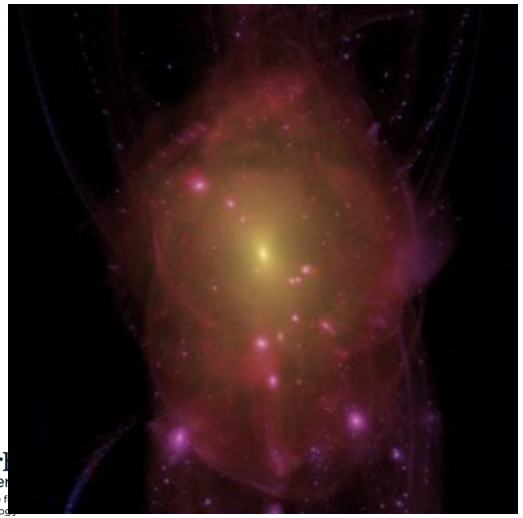
DiRAC@Durham: COSMA

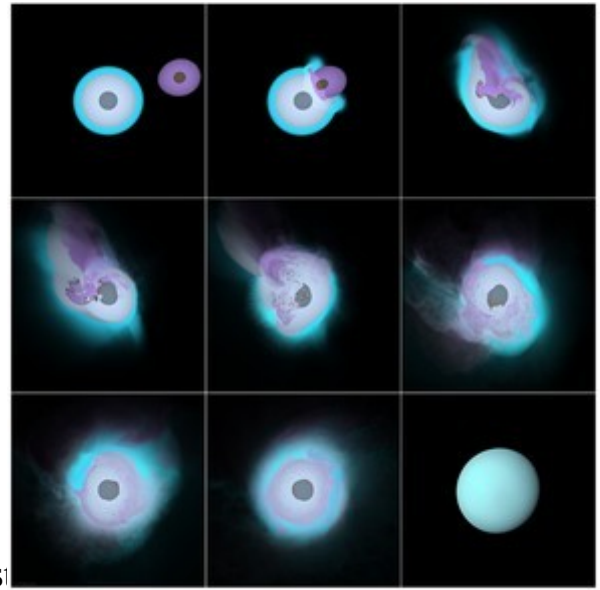
DiRAC Day
10th September 2020
Alastair Basden

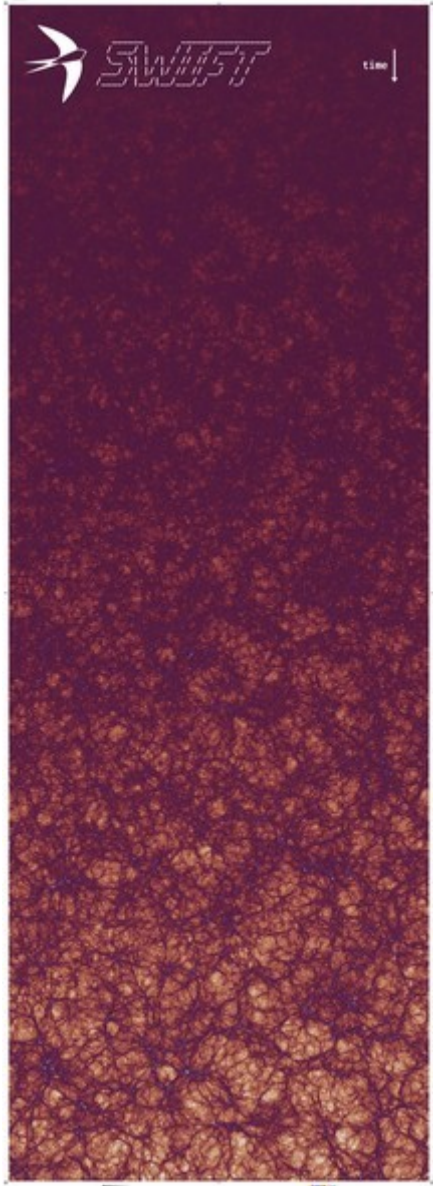
Durham University
Institute for Computational Cosmology

Large memory simulations

- Cosmological simulations
- Smoothed particle hydrodynamics
- Planetary modelling







Considerations

- Typical footprint:
  - Large jobs, many nodes
    - Good interconnect
  - High memory requirement
  - Long run times
    - Fast checkpointing essential

MI system focus

- Large amounts of RAM
  - Significantly more than is typical for HPC systems
    - ~50% of the cost!
    - COSMA5 - 2012: 8GB / core, 128GB / node
    - COSMA7 - 2018: 18GB / core, 512GB / node
    - DiRAC3 - 2020: 768 - 1024 GB / node

Network Fabric

- 100 GBit / s minimum
- Latency is important
  - Typically small transfers of non-contiguous memory
    - e.g. “particles”
  - Currently rules out Ethernet
    - (though Slingshot and newer 200G switches look promising)

Restart file storage

- Restart files typically written at regular intervals
  - Large, up to the RAM used by the simulation
  - Requires fast parallel storage
    - Written by all nodes simultaneously
- Currently:
  - Fast >200GB/s Lustre file system, 480TB
- DiRAC-3:
  - Increased to ~1PB

Bulk storage

- Currently:
  - 3.5PB Lustre parallel file system
- DiRAC-3:
  - ~10PB Lustre parallel file system

Extras

- Small number of Fat nodes
  - 3-4TB RAM
- Small number of GPU nodes
  - We hope to use a composable infrastructure making use of rCUDA

Testbeds

- BlueField cluster
  - Intelligent NIC studies
  - Communication offloading
- Distributed memory systems
  - Remote memory accessible on local nodes
  - e.g. Gen-Z, Inception

Timescales

- T0: Funding announced
- T2: Funding arrives
- T3: Procurement finalised
- T5: All delivered
- T6: Service opened for early adopters
- T7: General availability