DiRAC, ExCALIBUR and UK Exascale computing

UKLFT meeting

28th March 2023

Alastair Basden
Durham University / DiRAC

Introduction

- Alastair Basden, Durham University
  - Co-chair, DiRAC Technical Directorate
  - COSMA Manager (DiRAC Memory Intensive Service)
  - Background in astronomy/instrumentation
- DiRAC, ExCALIBUR, Exascale

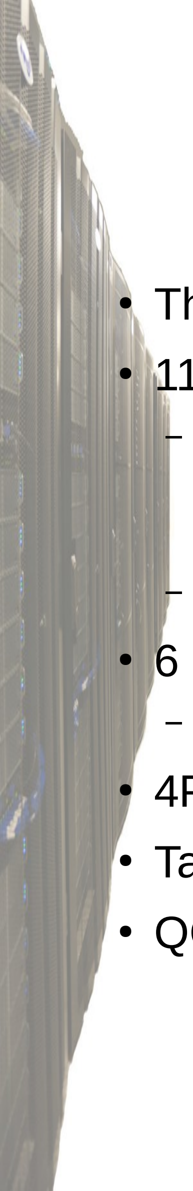
DiRAC: Introduction

- UK national HPC service for STFC researchers
  - Tier-1 facility
  - At times, largest provider of UK compute
- 4 sites:
  - Extreme Scaling @ Edinburgh
  - Data Intensite @ Leicester and Cambridge
  - Memory Intensive @ Durham
- Bespoke systems for the associated science
  - More cost effective than a single large system
  - Focus on Capability systems
    - For pushing the boundaries of what can be achieved

DiRAC: Current status

- DiRAC-2.5x: 2018
  - Leicester and Durham systems still operational
- DiRAC-3: October 2021
  - phase 2 part A: October 2023 (or sooner)
  - phase 2 part B
- DiRAC-4: The future

DiRAC @ Edinburgh: TURSA



- The Extreme Scaling service - a focus on massively parallel
- 114 GPU nodes (expanding to 178)
  - 4 NVIDIA A100 GPUs per node
    - 4 InfiniBand HDR (200GBit/s) per node (non-blocking)
      - One per GPU (Bespoke design to meet science goals)
  - 24 cores, 1TB RAM per node (Bespoke design to meet science goals)
- 6 CPU nodes:
  - 128 cores, 256GB
- 4PB Lustre
- Tape archival service
- QCD-focused

DiRAC @ Leicester: DiAL-2,3

- DiAL-3:
  - 200 nodes of 128 core, 512GB RAM
  - HDR200 InfiniBand (3:1)
  - 4PB storage
- DiAL-2:
  - 408 nodes of 36 core, 192GB RAM
  - EDR InfiniBand (2:1)
  - 3.5PB storage

DiRAC @ Cambridge: CSD3

- Peta-4 (CPU service):
  - 544 nodes with 76 cores, 256GB RAM *(267 nodes for DiRAC)*
    - HDR200 InfiniBand, (3:1)
  - 672 nodes with 56 cores, 192GB RAM *(119 nodes for DiRAC)*
    - HDR100 InfiniBand (3:1)
- Wilkes-3 (GPU service):
  - 80 nodes with: *(approx 10 nodes for DiRAC)*
    - 4x A100 GPUs
    - 128 cores, 1TB RAM
    - HDR200
- 23PB Lustre

DiRAC @ Durham: COSMA

- Memory Intensive service - a focus on large memory jobs
  - As capable as Archer-2 for some workloads
- COSMA8: 528 nodes with:
  - 128 cores, 1TB RAM (67k cores) (Bespoke design to meet science goals)
  - 13PB Lustre
  - HDR200 InfiniBand (non-blocking) (Bespoke design to meet science goals)
  - 1.2PB Fast scratch storage (Bespoke design to meet science goals)
- COSMA7: 448 nodes with:
  - 28 cores, 512GB RAM
  - 4PB Lustre
  - EDR InfiniBand and Rockport switchless Ethernet
- Tape archival service
- Primary workload: Cosmology

DiRAC: Coming soon

- DiRAC-3 phase-2
  - 64 extra nodes for Tursa (>50%)
    - Available from October
  - 168 extra nodes for COSMA (~50%)
    - Probably available from July
  - Funding being sought for Leicester and Cambridge expansions
    - Systems hopefully available from October 2024

DiRAC: Data curation

- Data Curation service coming soon
  - Long-term storage of data (10-15 years)
    - In an accessible manner
- Why?
  - Published data should remain available
    - Sometimes long after a researcher has left the field
  - Data should be FAIR (Findable, Accessible…)
  - Research data required for DiRAC researchers
    - after their projects end
- What should it look like?
  - Partly the reason for the delay!
  - Please feed in ideas!

DiRAC: Coming soon

- Work related to the DiRAC Federation project
  - Multiple sub-projects funded from 2021-2023
  - ~£1m solar panels to help power COSMA
  - HPC management software study
  - Data curation prototypes
  - AI utilisation optimisation
  - Training materials

DiRAC-4

- Probably GPU-heavy
  - But not necessarily discrete-GPU
- Design work starting this year
  - Get involved in the science case
  - And converting this into the technical case
  - Your input is important
  - Continue GPU porting

DiRAC4: possibilities

- May not be STFC-only
  - Other communities with similar needs
    - Please let us know if you can identify communities with similar compute needs
    - More capable systems

UK Exascale preparation

- £900m announcement for an Exascale system
  - Details not yet worked out
  - Unlikely to be a single large Exascale system
    - An Exascale ecosystem
    - Community science input
    - Industry co-design

DiRAC Innovation

- System co-design
  - Bespoke systems designed and built in-house
  - Focus on the science
- Silicon-level co-design
- Deployment of test systems
  - Including ExCALIBUR
- Software development
- Interdisciplinary knowledge transfer
- Metrics and user feedback
- DiRAC User Communities are a key part of this

ExCALIBUR

- UK Exascale preparation fund, £45m
  - 2019-2024
  - 10% for hardware exploration
    - Multiple ExCALIBUR test beds
      - RISC-V processors *
      - Rockport switchless 6D torus network **
      - AMD GPUs **
      - FPGA test bed
      - DPU test bed **
      - Visualisation test bed
      - Graphcore *
      - Cerebas
      - Storage
      - ARM+GPU test bed
      - CXL memory test bed ***

* - available for use

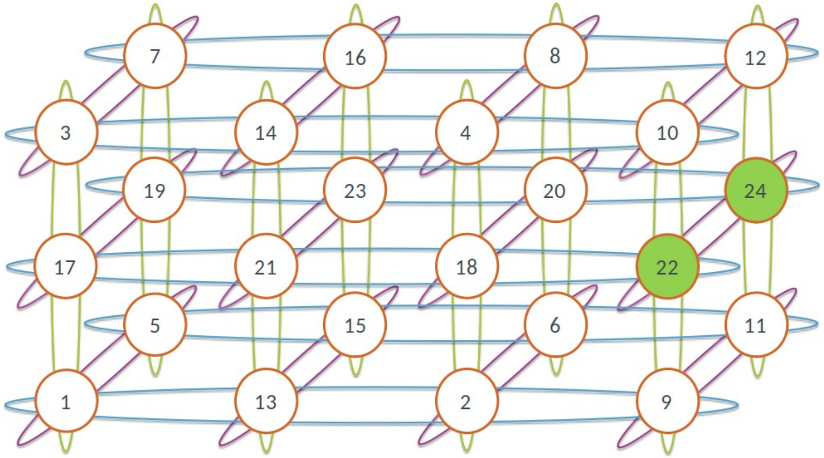
** - production system

*** - coming soon

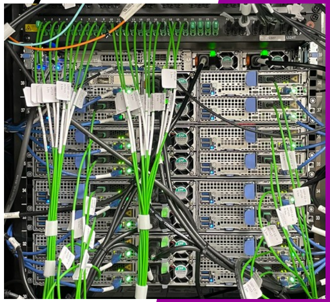
ExCALIBUR: Testbeds

- RISC-V - low performance RISC-V processors
  - Code testing
  - Processor design insight
  - Open-source processor
- Graphcore - AI-specific processor aimed at AI model training
- AMD GPU - MI100 and MI200 GPUs
  - available for performance comparisons
  - GPU workloads
- Quantum annealing credits (DWAVE)

Rockport ExCALIBUR testbed



- 6D torus network
- Switchless
- Consistently low latency
  - even with congestion
- Technology may underpin future composable fabrics
- 50% of COSMA7 converted to Rockport (224 nodes)
  - Direct comparison with InfiniBand



DPU ExCALIBUR testbed

- Data Processing Unit evaluation
- 24 nodes with BlueField-2 cards (200GBit/s)
  - PCIe card with 8 arm cores, 16GB RAM, network processing capability
  - Offloading MPI from host
  - MPI Task stealing
  - MPI progression

DiRAC: The RAC process

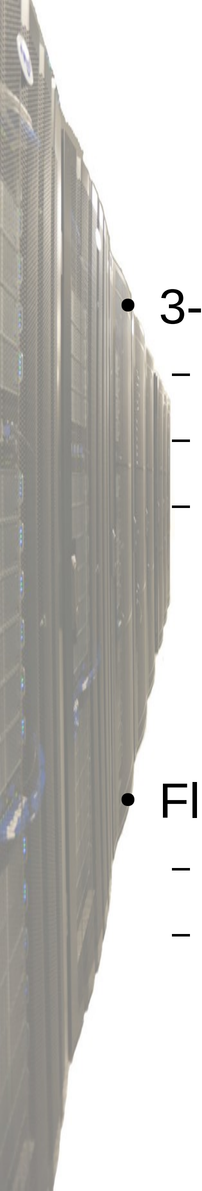
- Annual calls
- Not restricted to large projects
  - Supportive of early career researchers
- Seedcorn application at any time
  - New codes, new communities
- Director’s discretionary time

Compute eXpress Link

- Coming soon via ExCALIBUR - CXL testbed
  - Composable memory, GPUs, storage
    - Define via software how much memory (etc) a node has
    - Memory will be higher latency, lower bandwidth than node-native RAM
    - Codes will need to be NUMA aware to get best performance
      - Keep regularly accessed memory locally
  - Large future systems may include a central pool of RAM to be allocated upon demand
- ExCALIBUR testbed to provide early access to composable RAM
  - Allow codes to prepare for future systems.

DiRAC: Hackathons

- Typically 4-5 per year
  - Aimed at small teams
  - Working together with industry experts
    - Implementing improvements with given hardware/software
    - Compilers
    - GPUs
    - DPUs
    - Quantum
    - AI/ML
    - Performance analysis



DiRAC: Innovation placements

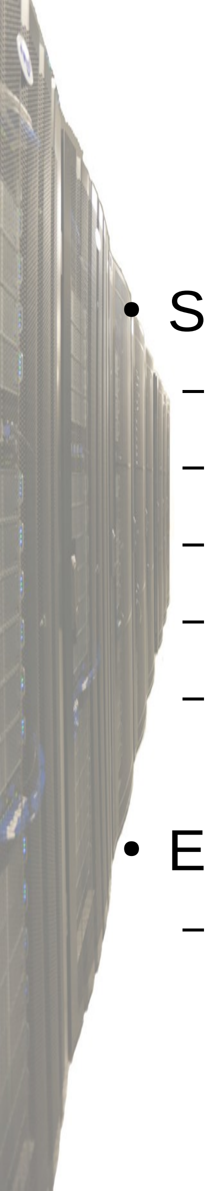
- 3-6 month paid placements in industry
  - PhD or Postdoc
  - Several per year
  - Current openings include:
    - IBM Quantum
    - Google Quantum
    - Save the Children
    - Epistemic AI
- Flow of skills in both directions
  - Provide academic skills to industry
  - and industry methods back to DiRAC

DiRAC: Training

- Focus on activities with mutual interest for DiRAC community and hardware providers
  - Key route for impact
  - Benefits capital investments
- Hackathons
- DiRAC Essentials training
  - For new users - about to be replaced
- Training for RSEs
- No generic HPC skills courses
  - Plenty of these around

DiRAC: RSE support

- DiRAC have a pool of ~5 RSE FTEs
  - Different areas of expertise
  - Please make use of these
    - If you have an immediate need, ask DiRAC Director
    - Or apply during the annual RAC calls
  - If applying for compute time, an RSE request is viewed favourably
- Code profiling and optimisation
- Porting to new platforms
- Improving efficiency of codes
- Improving parallelisation and scaling
- Facilitating discussions between research communities



DiRAC: Account management

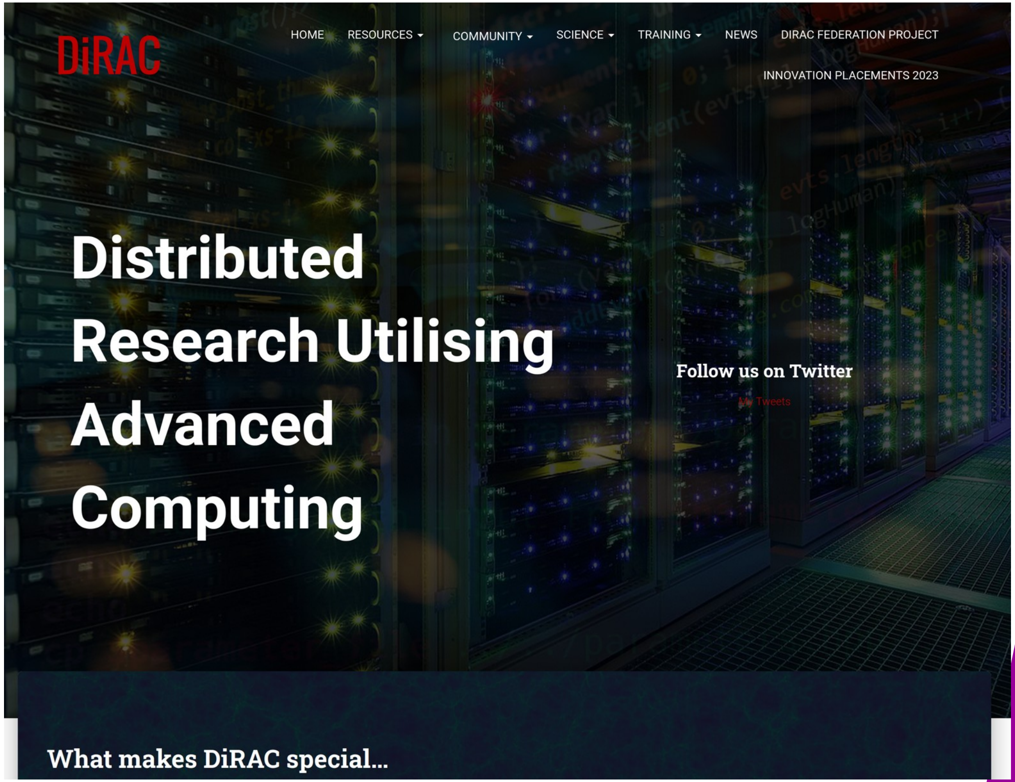
- SAFE: safe.epcc.ed.ac.uk/dirac
  - Create accounts
  - Apply to join projects
  - Manage credentials
  - Manage personal data (email address)
  - Annual check is required
    - Otherwise account locked
- Each site also has it’s own policies
  - Primarily for security

Feedback to communities

- Some systems report quarterly energy use back to users
  - To put it in context
  - To aid decisions about time requests if we start allocating by kW·hr
- Projects that under use their allocation by 50-80% receive a quarterly email
  - Not to point a finger
  - To make the PI aware - in the hope that better use can be made
  - And to provide a way of reporting problems
    - Queue lengths, lack of support, problems with file systems, etc

DiRAC: Website

- A new website is coming…
  - At some point



Distributed Research Utilising Advanced Computing

Follow us on Twitter

My Tweets

What makes DiRAC special...

Finally...

- If you use DiRAC facilities, please remember to add the appropriate acknowledgements!
- And do get in touch if you have problems
  - It should be a positive experience
    - Let Mark or me know if not
- Better systems + better software = better research!
  - Your input is important