Building Digital Twins of the Universe with the DiRAC HPC facility

Mark Wilkinson
Director, STFC DiRAC HPC Facility

RIKEN-CCS Café Presentation
25th February 2025

DiRAC

UKRI Science and Technology Facilities Council

The DiRAC HPC Facility

Science and Technology Facilities Council

DiRAC

Memory Intensive “COSMA8” (Durham)

DELL EMC

- 528 TB RAM
- Large-scale cosmological simulations

Extreme Scaling “Tursa” (Edinburgh)

EVIDEN
an atos business

- 704 Nvidia A100 GPUs
- Large lattice-QCD simulations

Data Intensive “DIaL” (Leicester)

Hewlett Packard Enterprise

- Heterogeneous architecture for complex simulation and modelling workflows

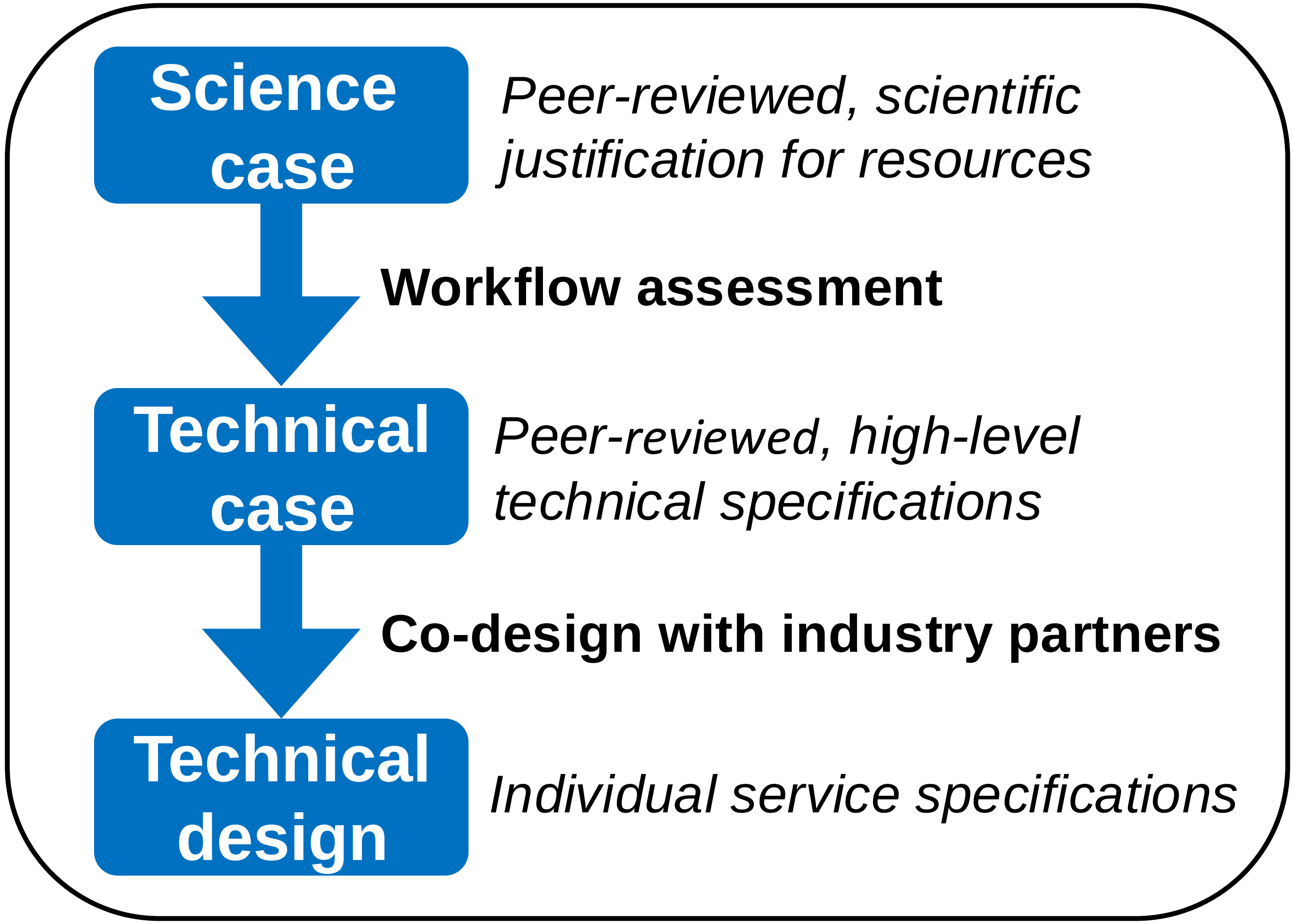
Data Intensive “CSD3” (Cambridge)

DELL EMC

- Heterogeneous architecture for complex simulation and modelling workflows

Project Office (UCL)

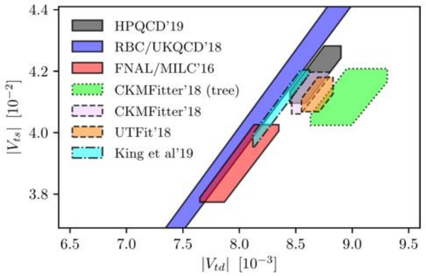
Applying the scientific method to HPC/AI service design

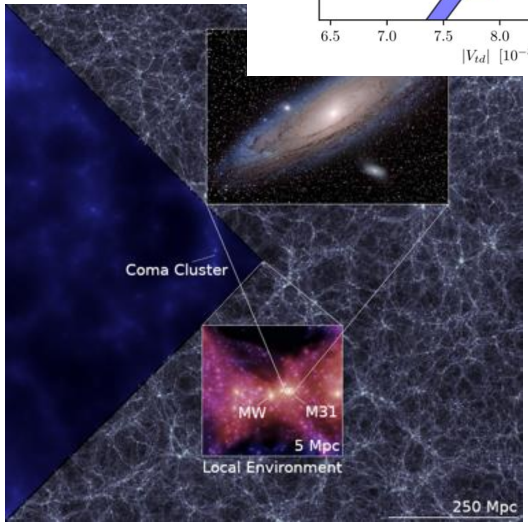


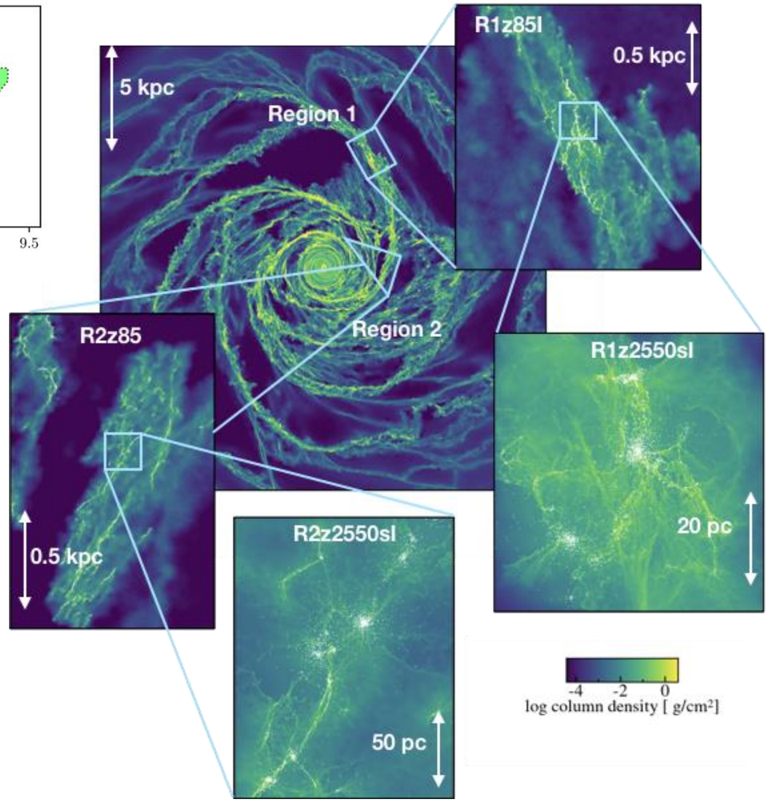
- Science case determines *both* scale and design of DiRAC services

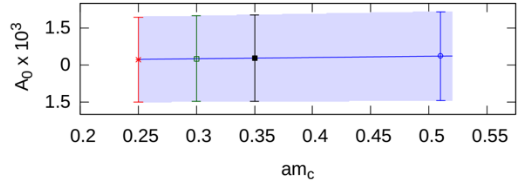
DiRAC Science Programme 2024-28

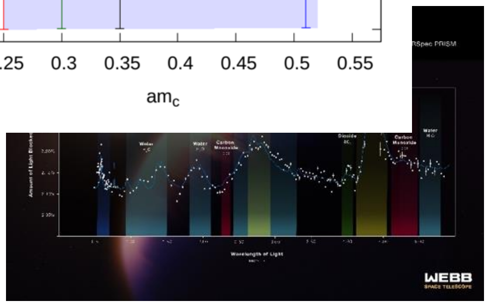
- Capability calculations include:
  - Galaxy formation
  - Lattice Quantum Field Theory
- Data Intensive calculations:
  - Gravitational waves
  - Gaia modelling
  - Precision cosmology
  - Planetary atmospheres
- Data challenges growing rapidly
  - Individual simulations generate 10Pb+
- Increasing use of AI/ML techniques to enhance simulation methods
  - At least 50% of fields are using or exploring AI over next 4 years.
  - DiRAC simulation data can be used to train AI models.
- STFC facilities also use AI extensively for data acquisition, data processing, data analysis











Co-design: the importance of people

DiRAC

- Investment in people is vital for productive HPC services
- DiRAC services require specialist technical support for hardware and users
- RSE team supports code improvement and re-factoring, energy efficiency, co-design, procurement & training

Evolution of Grid code (Boyle et al.) performance on Tursa relative to Tesseract

Stage	1 node	% inc.	16 nodes	% inc.	speed up 512 tess
Measured	9.2	-	5.3	-	1.1
Committed	9.2	-	5.83	10%	1.22
Acceptance	9.65	5%	6.15	16%	1.28
Commissioning	12	30%	8.8	66%	1.83
Peak	12.9	40%	9.9	87%	2.06

James Richings et al.

- Tursa Extreme Scaling service (DiRAC@Edinburgh) provides 5x the performance of its CPU-based predecessor for lattice QCD codes but uses just 50% of the power.
- Cosma8 Memory Intensive service (DiRAC@Durham) is 4x more efficient for cosmological simulations than comparable systems in Europe
- Clocking down of A100 GPUs on Tursa: ~5% performance loss for Grid code with ~15% energy saving

Combining AI and simulation in cosmology

Craig Bower, Corentin Houpert, Azam Khan, Shiqi Su, Ali Zahir
Ashiq Anjum, Martin Bourne, Debora Sijacki, Mark Wilkinson

BASE-II

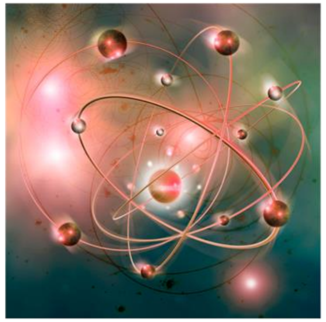
Blueprinting AI For Science at Exascale

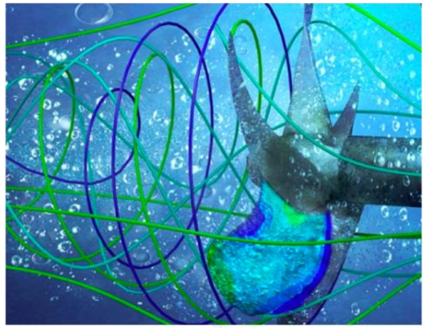
Jeyan Thiyagalingam
Paul Calleja, Marion Samler, Mark Wilkinson, Jeremy Yates

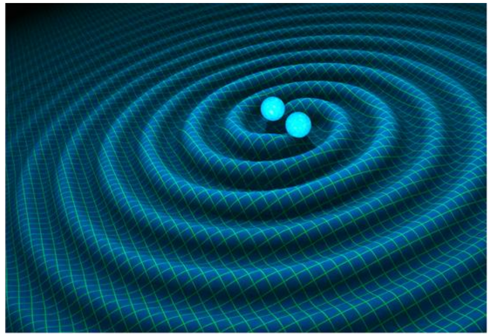
UCL
UNIVERSITY OF LEICESTER
UKRI Science and Technology Facilities Council
Scientific Computing
UNIVERSITY OF CAMBRIDGE
ExCALIBUR 10
DiRAC

Surrogate models for cosmology

- Our world is governed by PDEs at all scales







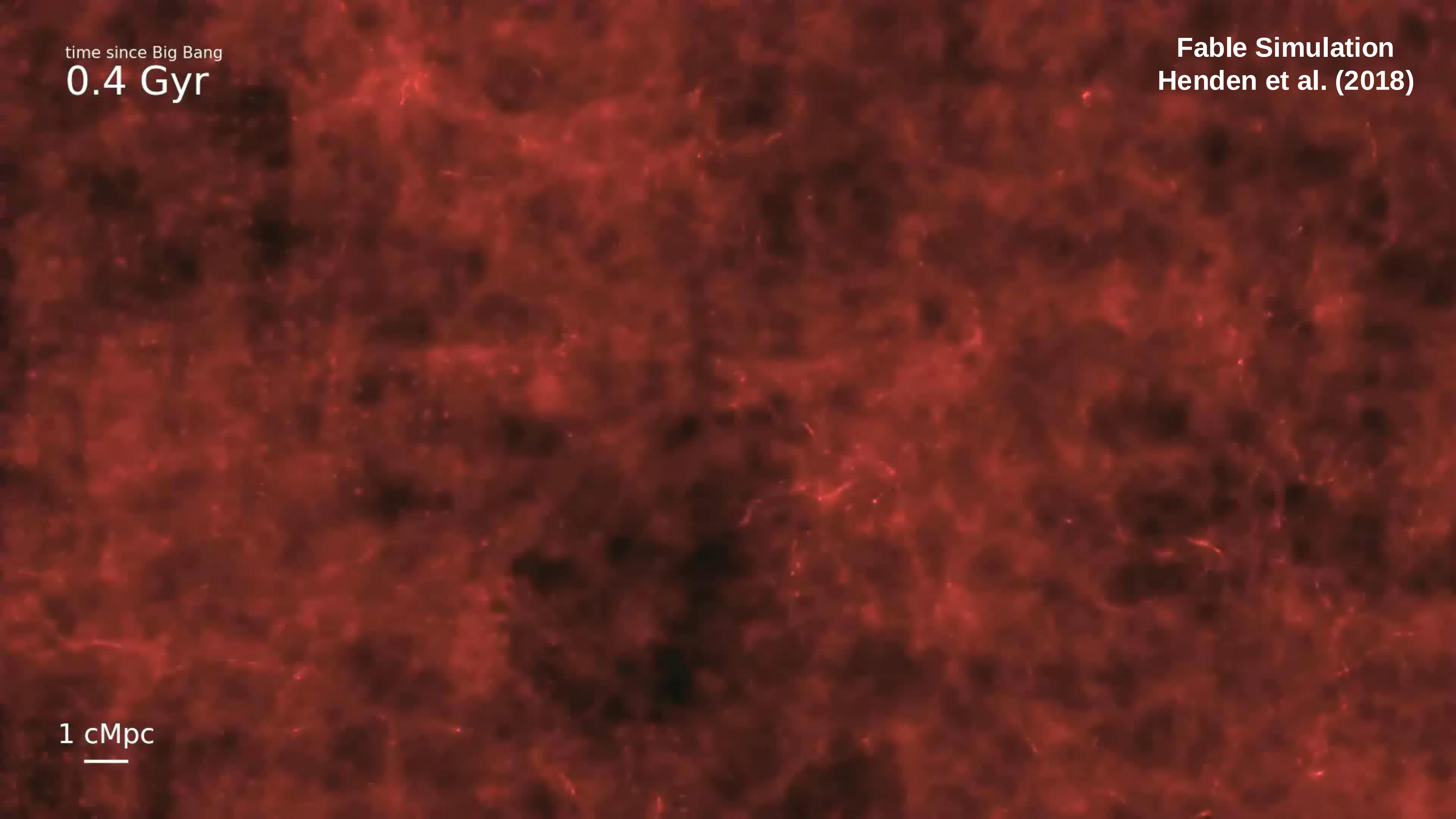
$$i\hbar \frac{\partial}{\partial t}\Psi = \hat{H}\Psi$$

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) = -p + \nabla\cdot\mathrm{T} + \mathbf{f}$$

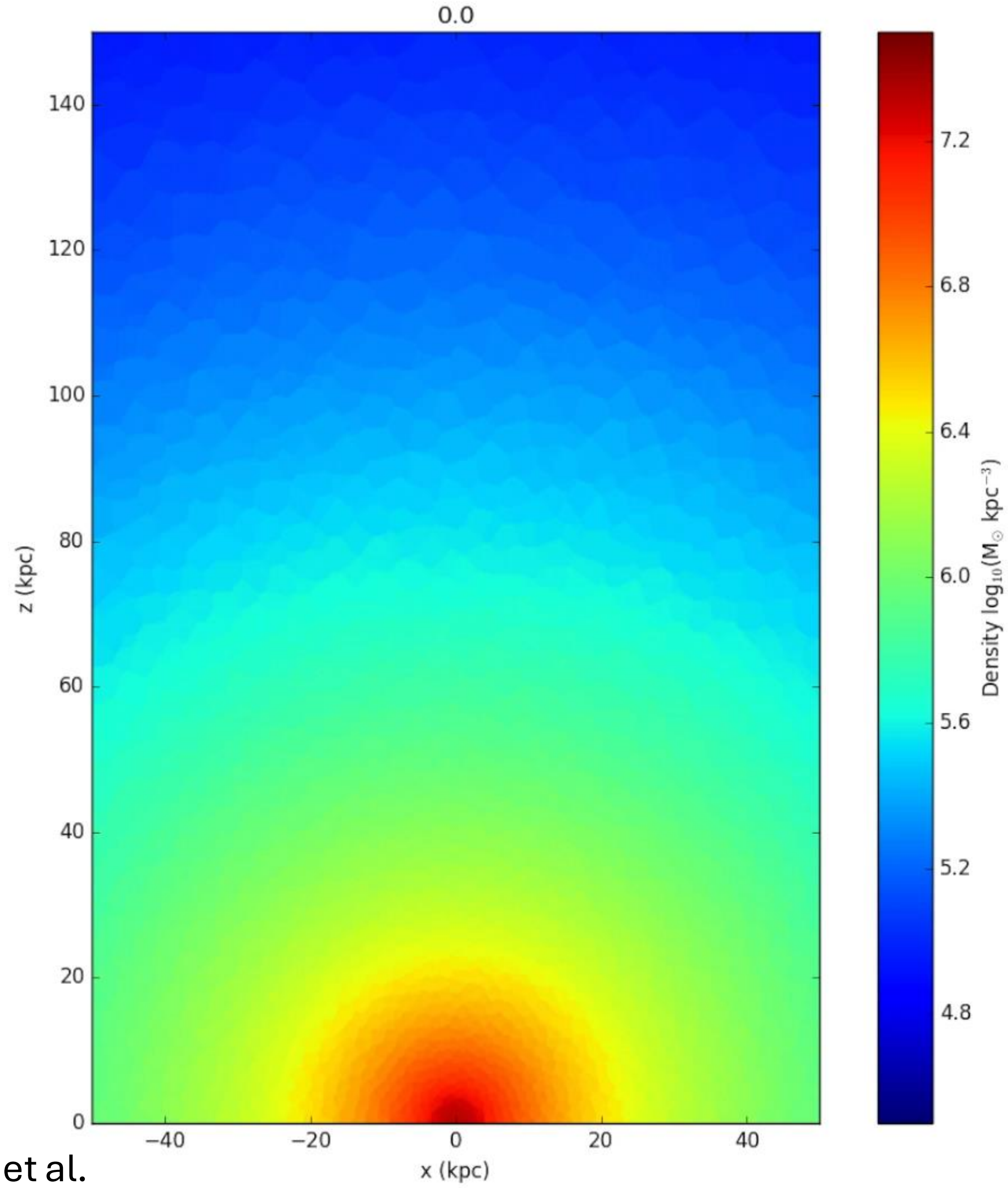
$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Planck Scale — Human Scale — Universe Scale

UNIVERSITY OF LEICESTER
BASE-II
Blueprinting AI For Science at Exascale
DiRAC



time since Big Bang
0.4 Gyr
Fable Simulation
Henden et al. (2018)
1 cMpc



Bourne et al.

High-energy jets in galaxies

Physical processes:

- Radiative transfer
- Magnetohydrodynamics
- Relativistic particles
- Transport processes
- General Relativity

Challenges:

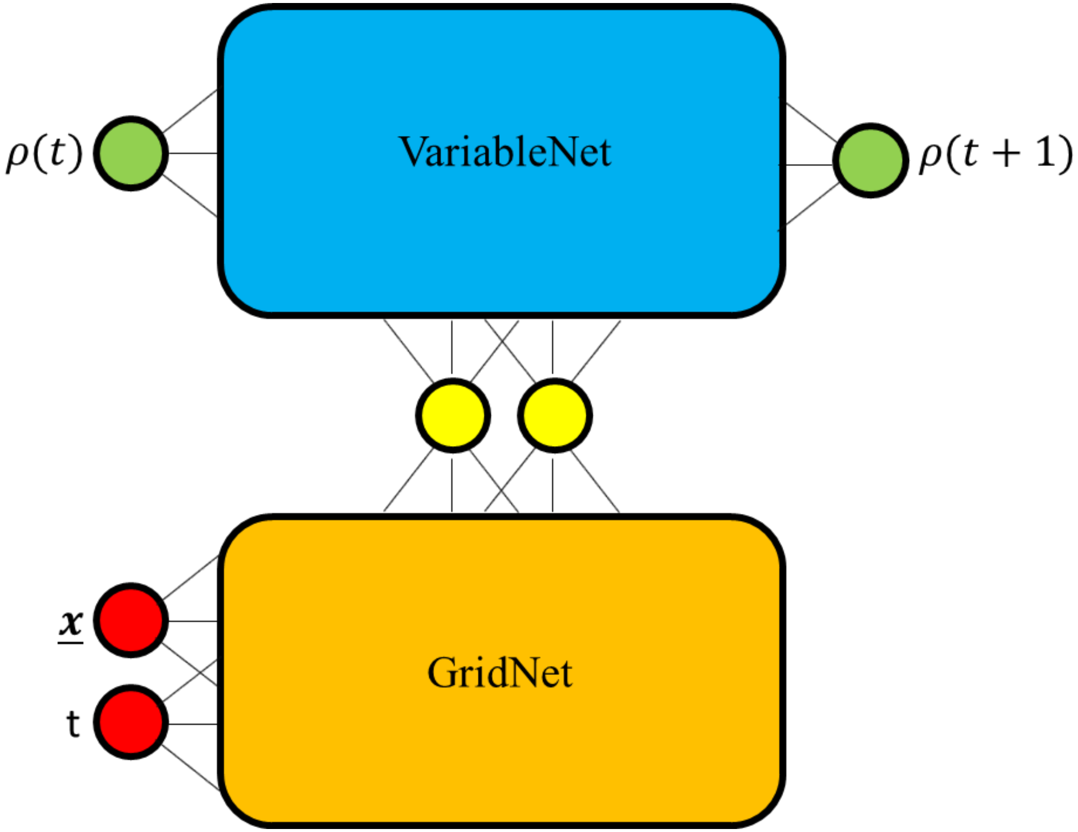
- 15 orders of magnitude range in spatial and temporal scales
- “Sub-grid” physics

DeepOJet

Bower et al., in prep

Deep Operator Network for mesh-agnostic upscaling and downscaling of AGN Jet Simulations

- Deep Operator Networks are trained to learn operators, mapping input functions to output functions
- Variables and associated coordinate geometry are separately encoded into two fully-connected MLP networks



UNIVERSITY OF LEICESTER

BASE-II

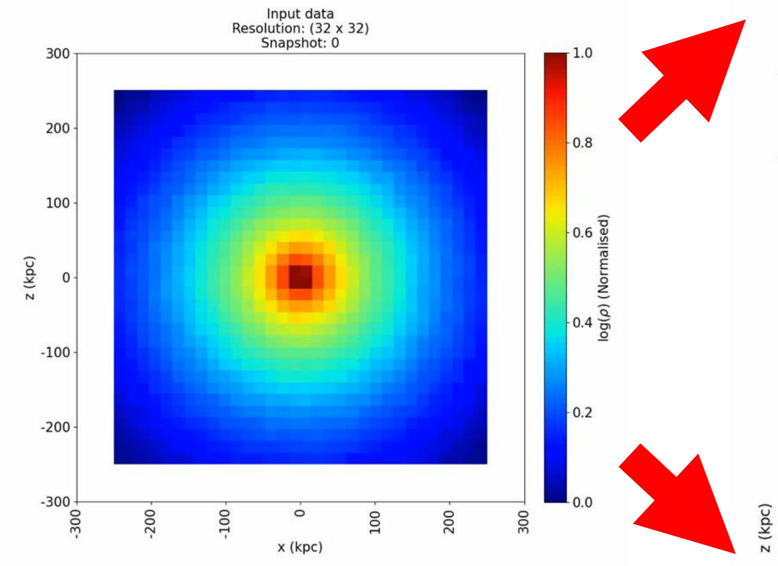
Blueprinting AI For Science at Exascale

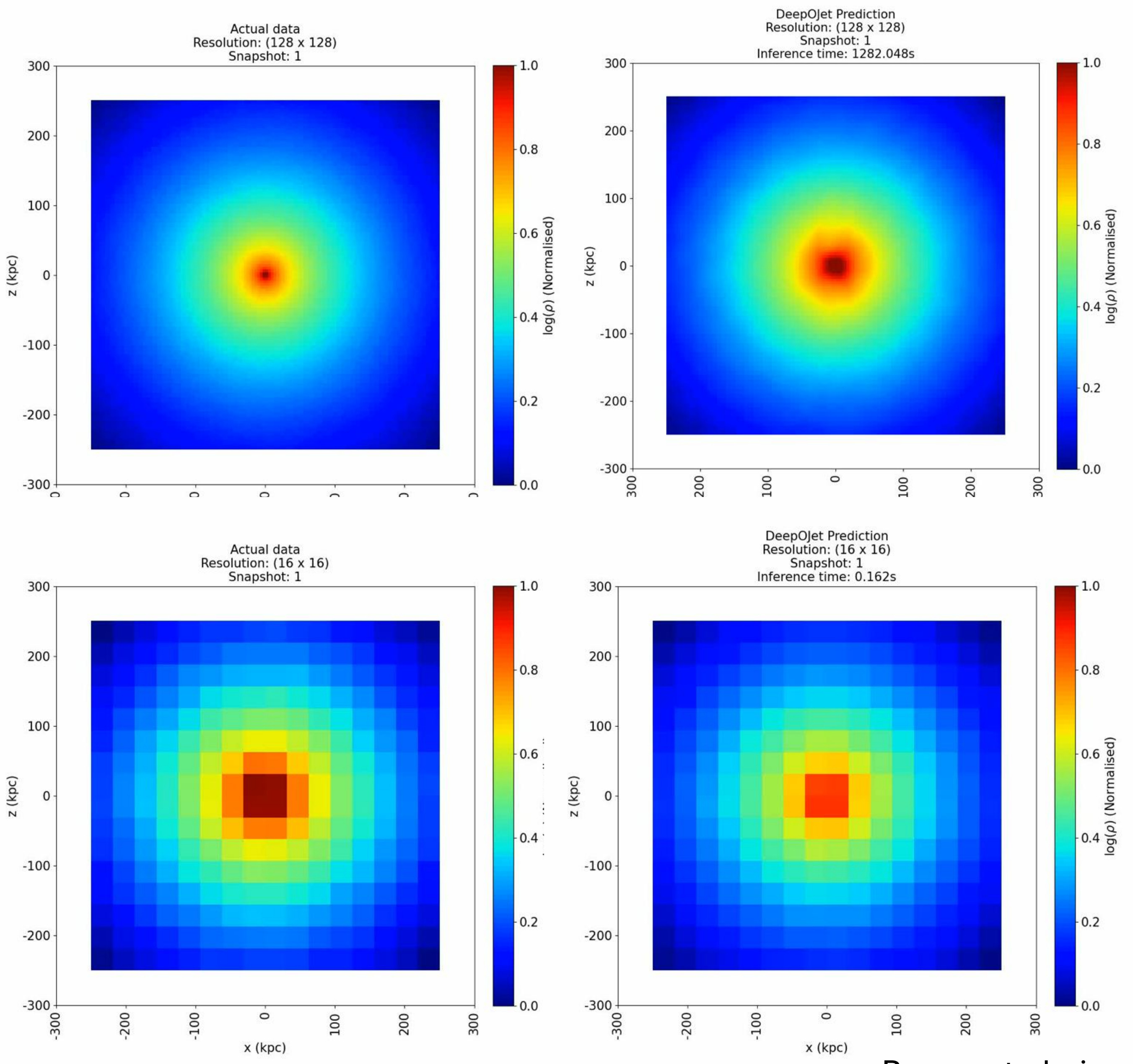
DiRAC

DeepOJet

Bower et al., in prep

Deep Operator Network for mesh-agnostic upscaling and downscaling of AGN Jet Simulations





UNIVERSITY OF LEICESTER

BASE-II
Blueprinting AI For Science at Exascale

DiRAC

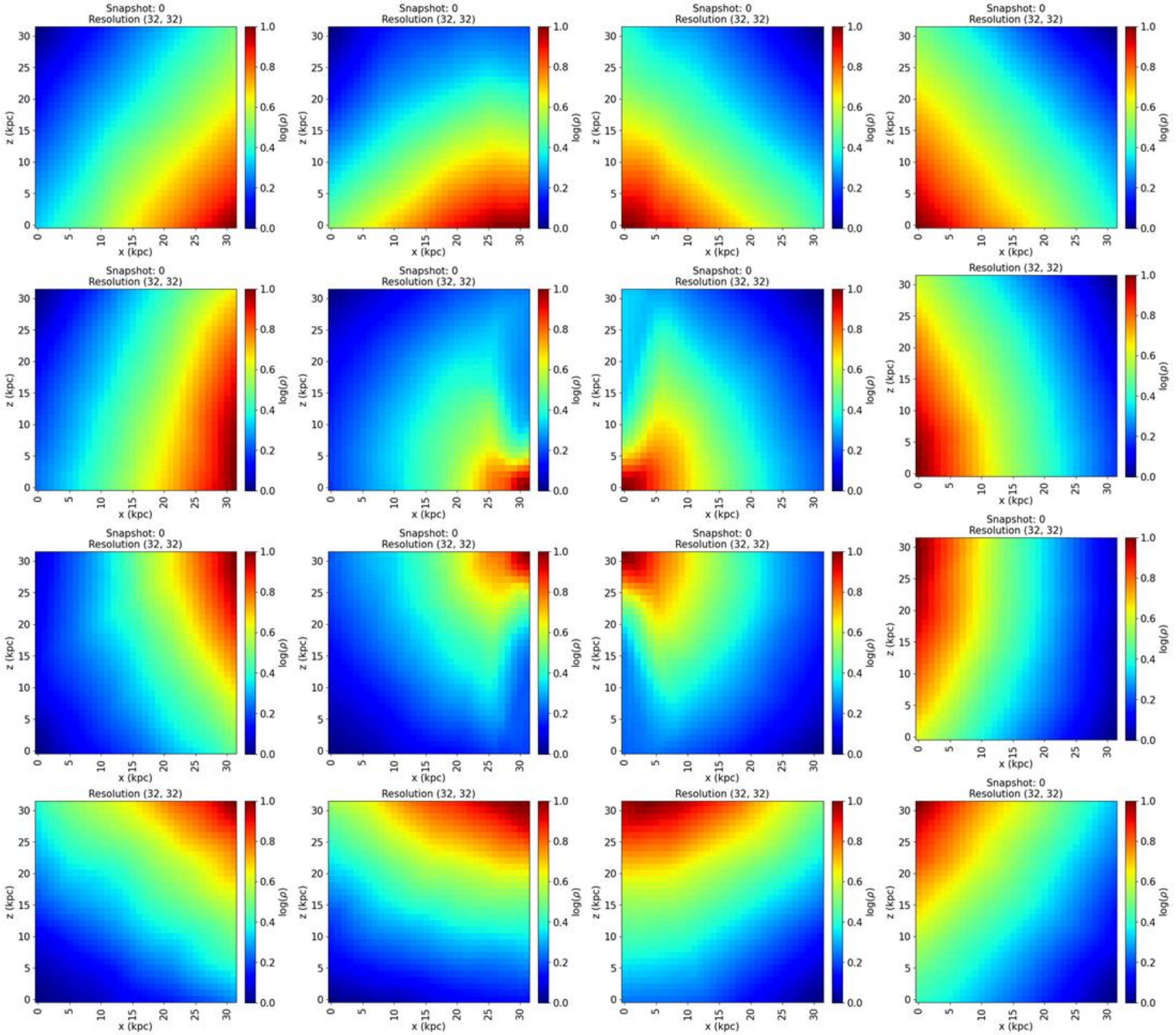
Bower et al., in prep

DeepOJet

Bower et al., in prep

- Apply DeepOJet to subgrids of measurements with co-ordinate systems corresponding to the entire grid
  - Delivers super-resolution prediction without any retraining.

- Evidence of *spectral bias*, where deep learning model over-generalise.
  - Train multiple DeepOJet models for each subgrid (in parallel) to reduce this



UNIVERSITY OF LEICESTER

BASE-II
Blueprinting AI For Science at Exascale

DiRAC

Surrogate models for radiative transfer

Shiqi Su (Leicester)

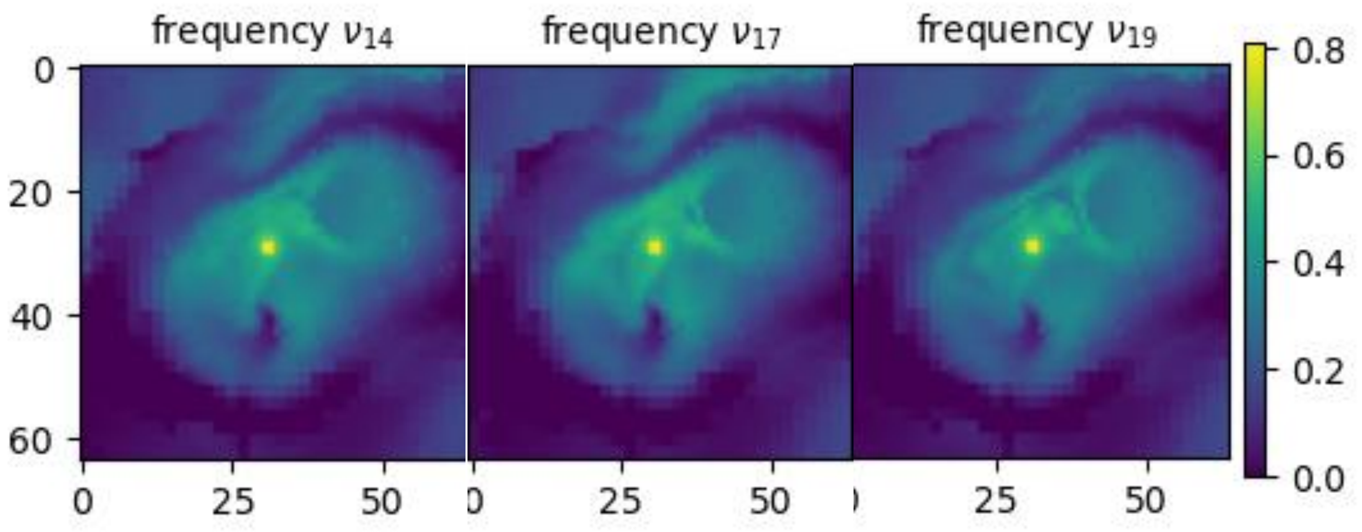
Using a 3D Residual Neural Network to build a surrogate model for the radiative transfer equation

$$\hat{\boldsymbol{n}} \cdot \nabla I_\nu(\hat{\boldsymbol{n}}) = \eta_\nu - (\chi_\nu + \chi_\nu^{\text{sca}}(\hat{\boldsymbol{n}}))I_\nu(\hat{\boldsymbol{n}}) + \oint d\Omega' \int_0^\infty d\nu' \, \Phi_{\nu\nu'}(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}') \, I_{\nu'}(\hat{\boldsymbol{n}}')$$

Initial application: stellar winds in AGB star binaries

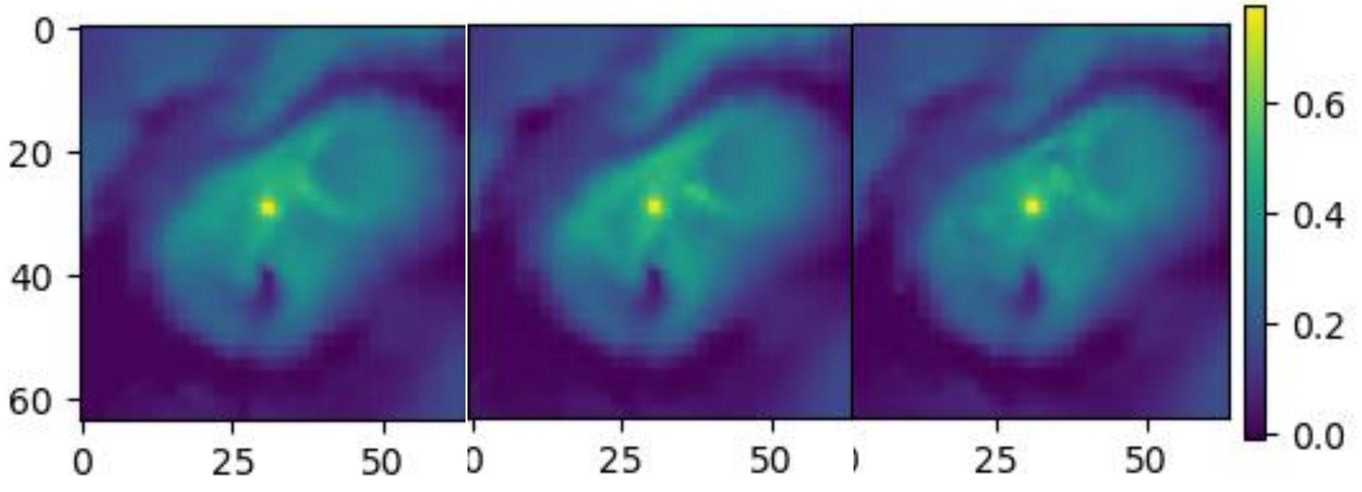
Surrogate model is 1000x faster than direct numerical calculation

Target



Su et al., in prep

Predicted

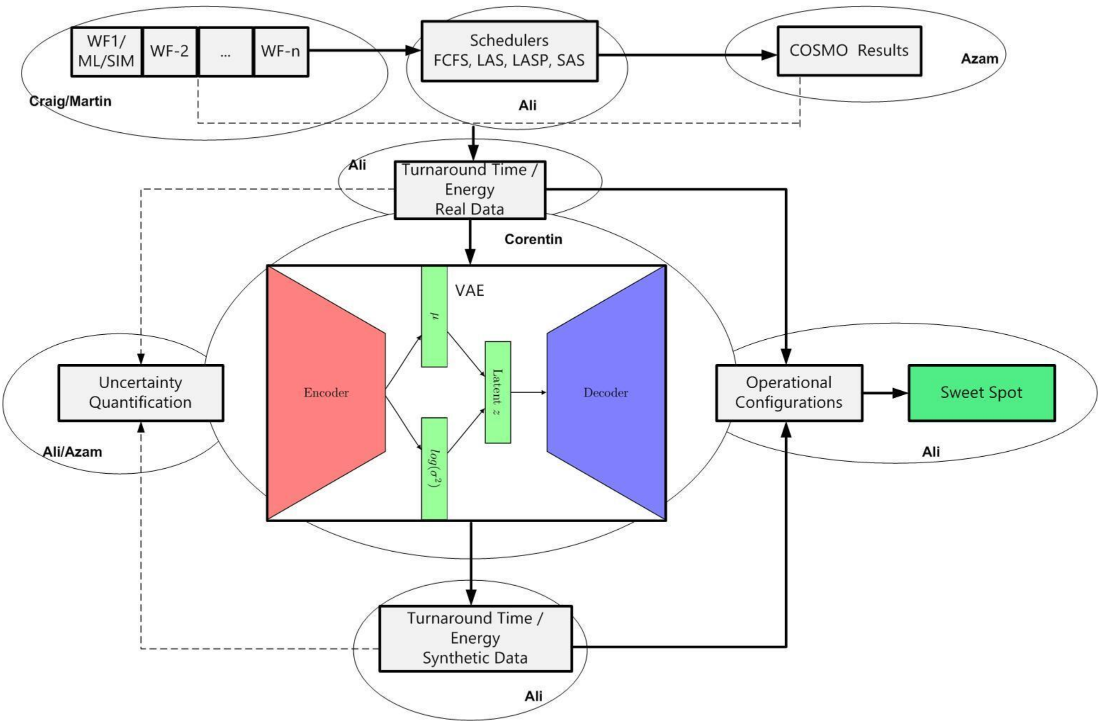


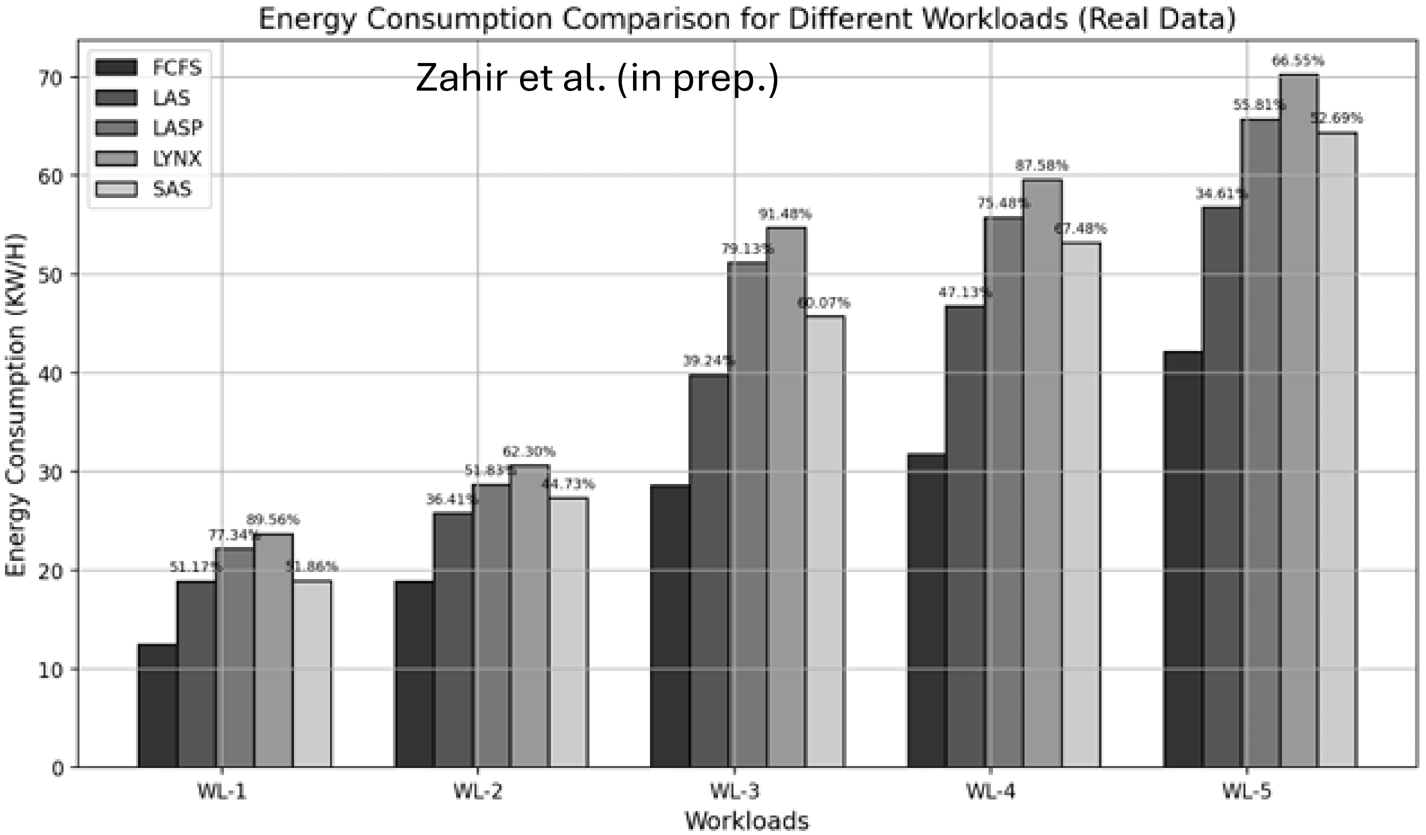
UNIVERSITY OF LEICESTER BASE-II Blueprinting AI For Science at Exascale DiRAC

Energy-aware scheduling – digital twins of compute clusters

Ali Zahir

- AI-based, physics-informed model to represent compute system
- Trained on data from particular workflows
- Use to explore potential scheduler configurations in terms of energy and turnaround time



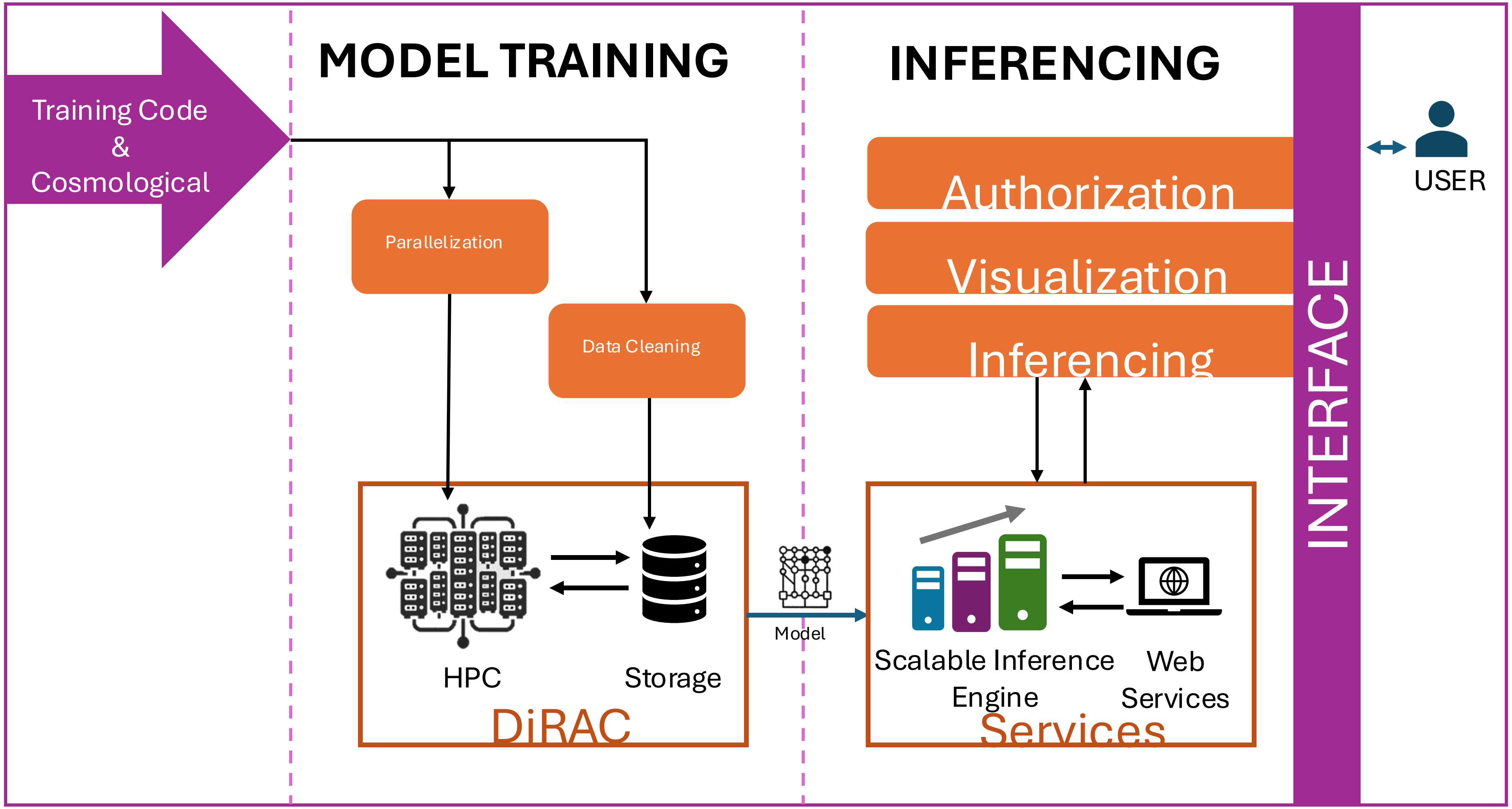


WF-1: 15% clock-down + SAS ⇒ EC down by 10% + TAT up by 5%

UNIVERSITY OF LEICESTER BASE-II Blueprinting AI For Science at Exascale DiRAC

Putting it all together

Azam Khan



UNIVERSITY OF LEICESTER BASE-II Blueprinting AI For Science at Exascale DiRAC

Co-designing HPC/AI services

- AI models require data
  - Must embed data requirements in system design and service planning
  - Also a requirement for simulation and other HPC workflows
  - Growing number/scale of data sources has network implications
- Avoid siloed AI services:
  - computing requirements are shared with other types of workload
  - AI is increasingly important in general HPC workflows
- Investment in people is key
  - RTPs, skills training programmes for researchers, etc.

Conclusions

- Co-design of HPC and AI services delivers increased productivity, cost-effectiveness and energy efficiency
  - Dependent on investments in people
- AI is becoming embedded throughout simulation workflows
- Surrogate models of sub-grid physics in cosmological simulations are essential to make the next generation of calculations possible

UNIVERSITY OF LEICESTER BASE-II Blueprinting AI For Science at Exascale DiRAC